PREMIER UNIVERSITY CHITTAGONG

Department of Computer Science & Engineering

Course Code : CSE-337

Course Title : Artificial Intelligence Laboratory.

Report Name: Fake News Detection.

Date of Submission : 06/09/2022

SU BMITTED BY

Mohiuddin Faysal ID:1903610201765
Priya Ghosh ID: ID:1903610201757
Prattay Bhowmik ID:1903610201759
Department : CSE
Semester: 6th
Section: C
Session : Spring_ 2022

SU BMITTED TO

Faisal Ahmed
Assistant Professor
Department of CSE

Author's Declaration of Originality

We declare that the report work entitled "Fake news Detection" submitted to the Premier University, is a record of an original work done by us under the guidance of Mr. Faisal Ahmed, Lecturer, Department of Computer Science & Engineering, Premier University, Chittagong. We can assure that the result of this report has not been submitted to any other university.

____

Mohiuddin Faysal
ID: 1903610201765

____

Priya Ghosh
ID: 1903610201757

____

Prattay Bhowmick
ID: 1903610201759

CERTIFICATION

The report entitled “Fake News Detection” submitted by, Mohiuddin Faysal, ID: 1903610201765, Priya Ghosh, ID: 19036102021757, Prattay Bhawmick, ID: 1903610201759 has been accepted as satisfactory in fulfillment of the course Artificial Intelligence Lab.

Faisal Ahmed
Lecturer
Department of Computer Science & Engineering
Premier University, Chattogram

Table of Contents

1. Introduction

1.1 Background . . . . ……………………………...6
1.2 Motivation . . . . ……………………………...6
1.3 Objective . . . . ……………………………….7
1.4 Summary . . . . ……………………………….7

2. Literature Review

2.1 Detecting Fake News in Social Media Networks …………………………………….8
2.2 A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection…......8
2.3 Fake News Detection Using Machine Learning Approaches…………………………..9
2.4 Convolutional Neural Networks for Fake News Detection……………………………9
2.5 A smart System for Fake News Detection Using Machine Learning…………………..10
2.6 Learning Hierarchical Discourse-level Structure for Fake News Detection……….......10
2.7 Fake news detection on Hindi news…………………………………………………..11
2.8 Transformer based Automatic COVID-19 Fake News Detection System……………….11
2.9 "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection……..12
2.10 Machine Learning Fake News Classification with Optimal Feature Selection……...12

3. Methodology/Proposed Method

3.1 Dataset Description . . . . ………………………….13
3.2 Preprocessing . . . . ………………………….15
3.3 Feature Exraction . . . . ………………………….16
3.4 Classification . . . . ………………………….16
3.4.1 Naive Bayes Classifier . . . . ……………………… 16
3.4.2 K-Nearest Neighbour Classifier . . . . ……………………….17
3.4.3 Decision Tree Classifier . . . . ……………………….17
3.4.4 Logistic Regression Classifier . . . . ……………………….17
3.5 Summary ……………………………………………………………………………18

4. Experimental Results and Discussion

4.1 Performance Measure . . . . …………………………...19
4.1.1 Accuracy . . . . ……………………………...19
4.1.2 Recall . . . . ………………………………19
4.1.3 Precision . . . . ……………………………...19
4.1.4 F1 Score . . . . ……………………………...19
4.2 Result . . . . ………………………………20
4.3 Summary . . . . ………………………………22

5. Conclusion & Future Work ………………………………………….23

Abstract

The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it. This report makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data,

Here, we have collected Fake news data as well as used five kinds of machine learning classifiers to analyze these data. Hence, the considered classifiers are Naive Bayes (NB), Decision Tree (DE), Logistic Regression (LR) as well as K-Nearest Neighbours (KNN). According to our analysis, DT displays higher precision when compared to others, while LR achieves better accuracy, precision, and F-score.

Keywords: Fake news detection, Machine Learning, Natural Language Processing, Tokenization, Stemming, True news, Fake news , Classification, Naive Bayes, k-Nearest Neighbors, Decision tree, Logistic Regression.

Chapter 01

Introduction

1.1 Background

News is the fastest means of spreading information around and forming views about the world. Fake news is a piece of information that is blown out of proportion and widely disseminated among the general public. It spreads faster than authentic, verified news and is more widely believed. As important as it is to get such information, it is also important to separate authentic news from basic hearsay. Distorted pieces of information lead to biased decisions and even societal chaos due to its pervasive nature: all important decisions are based on information. We form an idea about people or a situation by obtaining information. If the information we see on the Web is inauthentic or distorted, good decisions will be an impossibility. Fake news or distorted news spreads faster than regular news, leading to inefficiency in the market.

Documents are categorized fake or real news. Text mining, Natural Language Processing (NLP), and other computational methods are used in Fake news detection. Tokenization, word filtering, stemming, and classifications are all part of the Fake news detection methodology. Tokenization requires the division of text into discrete elements like words, integers, or punctuation. The next stage is stemming, which is the act of eliminating prefixes and suffixes to reveal a word’s stem. After preprocessing, we perform classification on the dataset using Nave Bayes, KNN classification, Decision Trees, and Logistic Regression. Here, we choose the most accurate model. As a result, we evaluate and research the aspects that have an impact on the ratings of our review text before classifying the news as Fake or Real.

1.2 Motivation

The widespread problem of fake news is very difficult to tackle in today’s digital world where there are thousands of information sharing platforms through which fake news or misinformation may propagate. It has become a greater issue because of the advancements in AI which brings along artificial bots that may be used to create and spread fake news. The situation is dire because many people believe anything they read on the internet and the ones who are amateur or are new to the digital technology may be easily fooled. A similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough acknowledge this problem take on this challenge to control the rates of crime, political unrest, grief, and thwart the attempts of spreading fake news.

1.3 Objective

We aim to analyze a repository of fake-news articles to better identify textual language patterns and differentiate fake news from the truth. The profundity of our work may not be supremely efficacious given our limited scope; however, it is a small step in the right direction to give the public a clearer perspective on the news that they consume.

1.4 Summary

In this report, we present a model to detect the Fake news. Where we applied some tools to detect fake and real news. Here many stages will be followed, which are tokenization, stemming and after preprocessing we performed classification on the dataset. Afterwards, we were able to choose the most accurate model.

Chapter-02

Literature Review

2.1 Detecting Fake News in Social Media Networks.

Research Authors: Mother Adware, Ali Al Waheeda.
College of Technological Innovation, Zayed University, Abu Dhabi 144534, UAE

The research paper Source:
https://www.sciencedirect.com/science/article/pii/S1877050918318210

They work for identifying a solution that could be used to detect and filter out sites containing fake news for purposes of helping users to avoid being lured by clickbait's. They analyze the prevalence of fake news in light of the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information.

They use simple and carefully selected features of the title and post to accurately identify fake posts. In this paper they use some methods like Bayes Net, Logistic, Naïve Bayes, Random trees. And their best experimental results show 99.4% accuracy using logistic classifier.

Furthermore, higher classification accuracy can be achieved by employing the ensemble classifiers or deep learning approaches.

2.2 A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection.

Research Authors: Soorya Dipti Das, Ayan Bask, and Sai kat Dutta.

Research paper Source: https://link.springer.com/chapter/10.1007/978-3-030-73696-5_16

The dataset contains 10,700 social media news items, the vocabulary size of which is 37,505 with 5141 words in common to both fake and real news.

In this paper, they describe Fake News Detection system that automatically identifies whether a tweet related to COVID-19 is "real" or "fake", as a part of CONSTRAINT COVID19 Fake News Detection in English challenge. They have used an ensemble model consisting of pre-trained models that has helped them achieve a joint 8th position on the leader board.

They have achieved an F1-score of 0.9831 against a top score of 0.9869. Post completion of the competition, they have been able to drastically improve their system by incorporating a novel heuristic algorithm based on username handles and link domains in tweets fetching an F1-score of 0.9883 and achieving state-of-the art results on the given dataset.

Their proposed method consists of five parts: (a) Text Preprocessing, (b) Tokenization, (c) Backbone Model Architectures, (d) Ensemble, and (e) Heuristic Post Processing

2.3 Fake News Detection Using Machine Learning Approaches.

Research Authors: Z Khanam, B N Al weasel, H Sirafi1 and M Rashid.

The research paper Source: https://ieeexplore.ieee.org/document/8862770

This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis.

They used Boost, Random Forests, Naive Bayes-Nearest Neighbors, Decision Trees Classifiers. the XGBOOST is depicting the highest accuracy with more than 75%, next is SVM and Random Forest with approximately 73% accuracy.

A superior accuracy can be attained by using further data preconditioning techniques. Furthermore, higher classification accuracy can be achieved by employing the ensemble classifiers or deep learning approaches.

2.4 Convolutional Neural Networks for Fake News Detection.

Research Authors: Yang yang, Lei Zheng, Jiawei Zhang and onigiri Cui.

The research paper Source: https://arxiv.org/abs/1806.00749

The dataset in this paper focuses on the news about American presidential election.
The dataset in this paper contains 20,015 news, where 11,941 fake news and 8,074 real news. For fake news, the real news is crawled from the well-known authoritative news websites, the New York Times, Washington Post, etc. The dataset contains multiple information, such as the title, text, image, author and website. To reveal the intrinsic differences between real and fake news, they solely use the title, text and image information.

They use 80% of the data for training, 10% of the data for validation and 10% of the data for testing.

They compare their model with several competitive baseline methods including CNN-image, LR text-1000, CNN-text-1000, CNN-text-1000, LSTM-text-400, GRU-text 400, TI-CNN-1000. With text and image information, TI-CNN outperforms all the baseline methods significantly.

In this paper, they propose a unified model TI-CNN, which can combine the text and image information with the corresponding explicit and latent features. The proposed model has strong expandability, which can easily absorb other features of news.

2.5 A smart System for Fake News Detection Using Machine Learning

Research Authors: Anjali Jain, Harsh Chatter, Avinashi Shakya
Dr. APJ Abdul Kalam University, Lucknow, India

Research paper Source: https://rb.gy/nvkqlc

For the fake news detection, they have collected the input data from Kaggle. They used various techniques and tool to detect fake news like NLP techniques, machine learning, and artificial intelligence.

In this paper, they used Naïve Bayes classifier, CNN classifier, SVM Classifier. They got 93.50% best accuracy from SVM classifier.
In future, ensuing algorithm may provide better results with hybrid approaches for the same purpose fulfilment.

2.6 Learning Hierarchical Discourse-level Structure for Fake News Detection.

Research Authors: Hamid Karimi, Jillian Tang.

The research paper Source: https://rb.gy/nvkqlc

In this paper, they looked into fake news detection from a new perspective. They hypothesized that hierarchical discourse-level structure of news documents offers a discriminatory power for fake news detection.

They utilize available online fake news datasets provided by kaggle.com. The Dataset contains 3360 fake and 3360 real documents.

In this paper they used some methods including N-grams, LIWC, RST, Bornn-CNN, LSTM [w+s], LSTM[s], HDSF. The proposed framework HDSF significantly outperforms all other methods. They got best accuracy 82.19% form HDSF.

2.7 Fake news detection on Hindi news

Research Authors: Sudhanshu Kumar & Doren Singh from National Institute of Technology Silchar, Silchar 788010, India.

The research paper Source: https://doi.org/10.1016/j.gltp.2022.03.014

In this work, they collected Hindi news from various types of sources. In this dataset, there are more than 2100 news article handpicked from different mainstream news channels.

They used Different machine learning algorithms such as Naïve Bayes, logistic regression and Long Short-Term Memory (LSTM) . Term frequency inverse document frequency (TF-IDF) is used for feature extraction. Naïve Bayes, logistic regression and LSTM classifiers are used and compared for fake news detection with probability of truth.

And it is observed that among these three classifiers, LSTM achieved best accuracy of 92.36%.

2.8 Transformer based Automatic COVID-19 Fake News Detection System.

Research Authors: Sunil Gundu and Radhika Mamudi.

The research paper Source: https://arxiv.org/abs/2101.00180

in this paper, they report a methodology to analyze the reliability of information shared on social media pertaining to the COVID-19 pandemic. The dataset containing 10,700 data points collected from various online social networks such as Twitter, Facebook, and Instagram, etc.

From the total dataset, 6,420 data points are reserved for training, 2,140 data points are used for hyperparameter tuning as a part of the validation phase, and the remaining 2,140 social media posts are kept aside for testing.

Their best approach is based on an ensemble of three transformer models (BERT, ALBERT, and XLNET) to detecting fake news. And they got best accuracy 0.98% in ensemble model.

In future, they will conduct a comparison of the performance analysis of deep learning algorithms and that might result in the improvement of the accuracy of the model.

2.9 “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection

Research Authors: William Yang Wang, Department of Computer Science University of California, Santa Barbara

The research paper Source: https://arxiv.org/abs/2110.10457

In this paper, they present a new available dataset for fake news detection. They collected a decade-long, 12.8K manually labeled short statements in various contexts from POLITIFACT.COM

They used four baselines: a majority baseline, a regularized logistic regression classifier (LR), a support vector machine classifier (SVM), a bi-directional long short-term memory networks model (Bi-LSTMs), and a convolutional neural network model (CNNs).

They compare the predictions from the CNN model with SVMs via a two-tailed paired t-test, and CNN was significantly better. When considering all meta-data and text, the model achieved the best result on the test data.

2.10 Machine Learning Fake News Classification with Optimal Feature Selection

Research Authors: Muhammad Fayyaz, Atif Khan, Muhammad Bilal, Sanaullah Khan.

The research paper Source: https://rb.gy/lmarhb

A total of 44,919 fake and real news was used in this research, including 23502 fake news and 2147 real news assessments using multiple machine learning models. They use sentence segmentation, tokenization, stop words removal and word stemming With NLP process.

They used some methods such as Machine-Human (MH), Naïve Bayes, CNN, LSTM, Bi-LSTM, C-LSTM, Heterogeneous Graph Neural Network (HAN), Cov-HAS, Char-level, C-LSTM and They got the best results with Random Forest classifier having 97.25 percent accuracy.

In future, Deep Ensembling models can be used for getting best result in fake news detection.

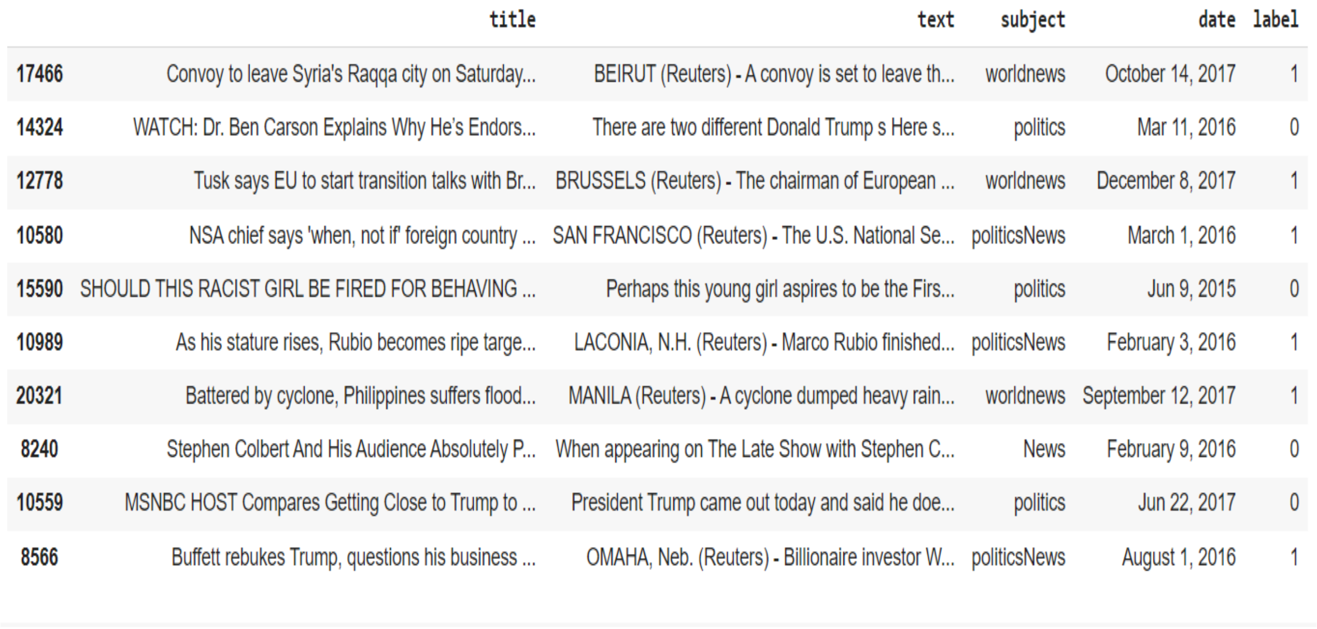
Chapter-03

Methodology/PROPOSED METHOD

3.1 Data Description

The dataset contains two types of articles fake and real News. This dataset was collected from real world sources; the truthful articles were obtained by crawling articles from Reuters.com (News website). As for the fake news articles, they were collected from different sources. The fake news articles were collected from unreliable websites that were flagged by PolitiFact (a fact-checking organization in the USA) and Wikipedia.
The dataset contains different types of articles on different topics; however, the majority of articles focus on political and World news topics. The dataset consists of two CSV files. The first file named “True.csv” contains more than 12,600 articles from reuter.com. The second file named “Fake.csv” contains more than 12,600 articles from different fake news outlet resources.

Each article contains the following information: article title, text, type and the date the article was published on. To match the fake news data collected for kaggle.com. The data collected were cleaned and processed, however, the punctuations and mistakes that existed in the fake news were kept in the text. The following table gives a breakdown of the categories and number of articles per category.



	title	text	subject	date	label
17466	Convoy to leave Syria's Raqqa city on Saturday...	BEIRUT (Reuters) - A convoy is set to leave th...	worldnews	October 14, 2017	1
14324	WATCH: Dr. Ben Carson Explains Why He's Endors...	There are two different Donald Trump s Here s...	politics	Mar 11, 2016	0
12778	Tusk says EU to start transition talks with Br...	BRUSSELS (Reuters) - The chairman of European ...	worldnews	December 8, 2017	1
10580	NSA chief says 'when, not if' foreign country ...	SAN FRANCISCO (Reuters) - The U.S. National Se...	politicsNews	March 1, 2016	1
15590	SHOULD THIS RACIST GIRL BE FIRED FOR BEHAVING ...	Perhaps this young girl aspires to be the Firs...	politics	Jun 9, 2015	0
10989	As his stature rises, Rubio becomes ripe targe...	LACONIA, N.H. (Reuters) - Marco Rubio finished...	politicsNews	February 3, 2016	1
20321	Battered by cyclone, Philippines suffers flood...	MANILA (Reuters) - A cyclone dumped heavy rain...	worldnews	September 12, 2017	1
8240	Stephen Colbert And His Audience Absolutely P...	When appearing on The Late Show with Stephen C...	News	February 9, 2016	0
10559	MSNBC HOST Compares Getting Close to Trump to ...	President Trump came out today and said he doe...	politics	Jun 22, 2017	0
8566	Buffett rebukes Trump, questions his business ...	OMAHA, Neb. (Reuters) - Billionaire investor W...	politicsNews	August 1, 2016	1

Table 3.1. Glimpse of the dataset

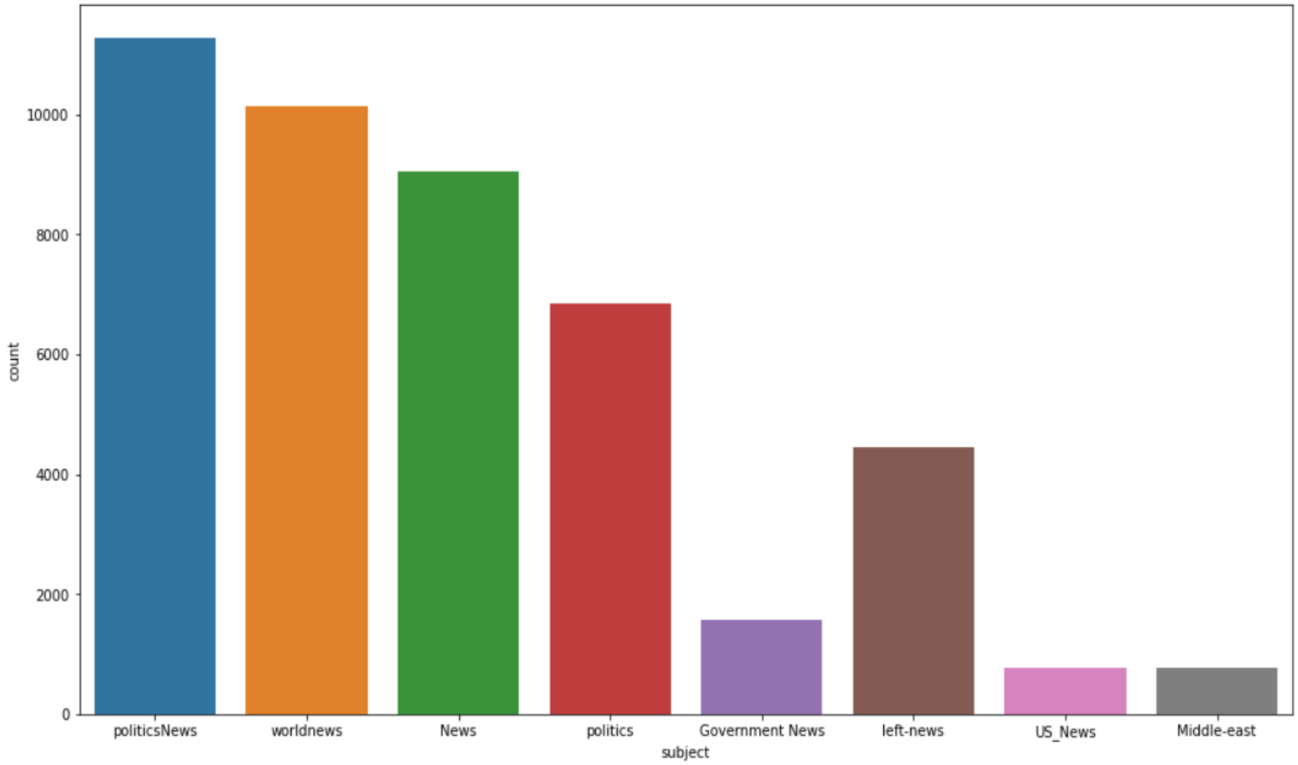


Figure 3.1. Graphical view of Dataset subject

News	**Size** **(Number of articles)**	***Subjects***	
Real-News	21417	**Type**	**Articles size**
		World-News	*10145*
		Politics-News	*11272*
Fake-News	23481	**Type**	**Articles size**
		Government-News	*1570*
		Middle-east	*778*
		US News	*783*
		left-news	*4459*
		politics	*6841*
		News	*9050*

Table 3.2. Fake & Real news size

In the dataset we acquired two types of news, Real and Fake. The Real news are 21417 in number and Fake news are 23481 in number. The graphical view of both the news are shown in Fig 3.2. Besides Article size number is shown in Fig 3.1.

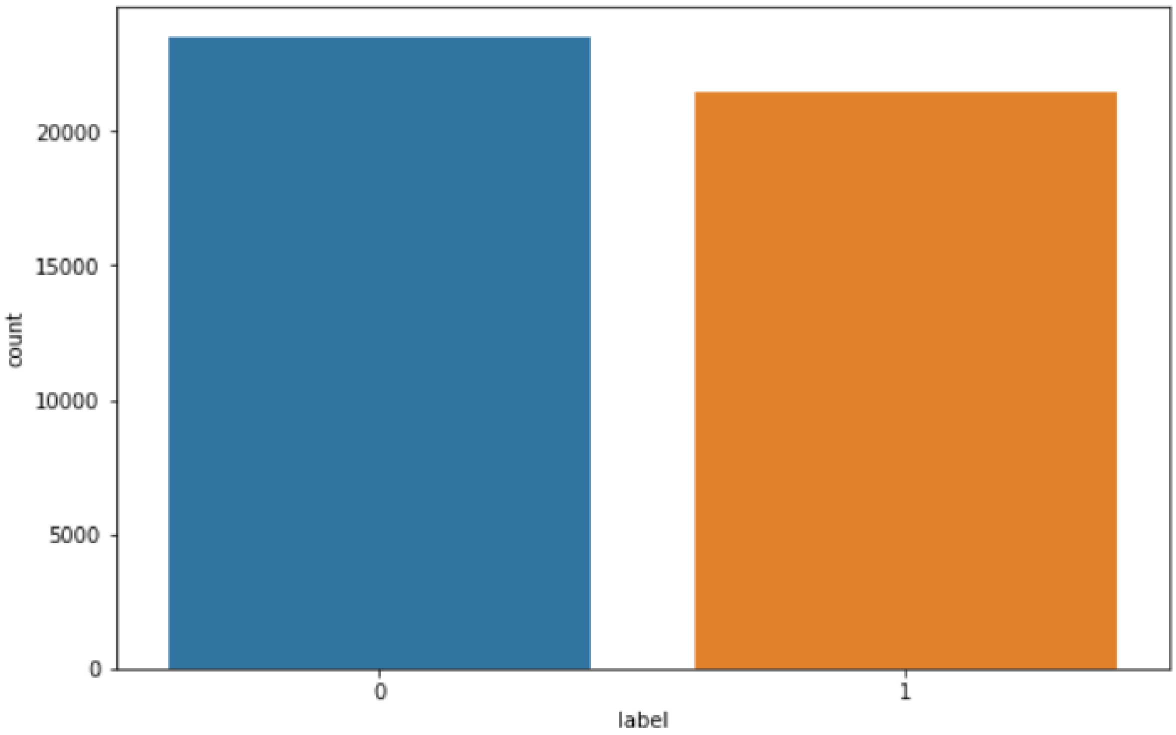


Fig 3.2: Graphical view of Real and fake news size.

3.2 Preprocessing

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So, for this, we have used the following data preprocessing tasks.

- **Punctuation Marks Removal:** An important NLP preprocessing step is punctuation marks removal. It is used to divide text into sentences, paragraphs and phrases - affects the results of any text processing approach, especially what depends on the occurrence frequencies of words and phrases, since the punctuation marks are used frequently in text. In this section, we have removed English punctuations by using removal punctuation functions.

- **Stop words Removal:** Stop word removal is one of the most commonly used preprocessing steps across different NLP applications. The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.

- **Stemming:** Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

Stemming is important in natural language understanding (NLU) and natural language processing (NLP).

- **Tokenization**: Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. We have used white space tokenization which is the simplest tokenization technique. It tokenizes into words by splitting the input whenever a white space is encountered. This is the fastest tokenization technique but will work for languages in which the white space breaks apart the sentence into meaningful words.

3.3 Feature Extraction:

TF-IDF: Term Frequency- Inverse Document Frequency is a widely used statistical method in natural language processing and information retrieval. It measures how important a term is within a document relative to a collection of documents. Words within a text document are transformed into importance numbers by a text vectorization process.

As its name implies, TF-IDF vectorizes a word by multiplying the word's Term Frequency (TF) with the Inverse Document Frequency (IDF).

TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

TF = (number of times the term appears in the document) / (total number of terms in the document) IDF of a term reflects the proportion of documents in the corpus that contain the term.

IDF = log * ((number of the documents in the corpus) / (number of documents in the corpus contain the term)) The TF-IDF of a term is calculated by multiplying TF and IDF scores

TF-IDF = TF * IDF

3.4 Classification:

We used Naïve Bayes classifier, KNN Classifier, Logistic Regression Classifier, And Decision Tree Classifier in our Experiment.

3.4.1 Naive Bayes:

3.4.1 Naive Bayes: Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems. It is a probabilistic classifier that predicts on the basis of the probability of an object. The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

P(A|B) is Posterior probability and P(B|A) is Likelihood probability. In our dataset we used the Bernoulli Naive Bayes classifier. It is used when characteristic values are continuous in nature then an assumption is made that the values linked with each class are dispersed according to Gaussian, that is Normal Distribution.

3.4.2 Logistic Regression: Logistic regression is a Machine Learning classification algorithm which is used to predict the probability of certain classes based on some dependent variables. The outcome of the Logistic Regression is always between (0 and 1).

$$h_{\Theta}(X) = \frac{1}{1 + e^{-\Theta X}}$$

In the above formula θ is the parameter that we want to train and X is the input data. The output is the prediction value when the value is closer to 1, which means the instance is more likely to be a positive sample(y=1). If the value is closer to 0, this means the instance is more likely to be a negative sample(y=0). The formula for loss function is:

$$J(\Theta) = -\frac{1}{m}\sum_{m}^{i=1}(y^{i}\log(p^{i}) + (1 - y^{i})\log(1 - p^{i}))$$

Here, m is the number of samples in the training data. y^{i} is the label of the itch sample, p^{i} I am the prediction value of the itch sample? We use this loss function to optimize our work.

3.4.3 Decision Tree: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier. There are two nodes, **Decision node** and **Leaf node**. Decision nodes represent the features of a dataset, branches represent the decision rules and whereas Leaf nodes represent the outcome.

3.4.4 K-Nearest Neighbor Classifier: A straightforward technique called K-Nearest Neighbor categorizes incoming data or cases based on a similarity metric after storing all of the previous examples. A data point is often categorized using the classification of its neighbors. As a result, it forecasts value using a lazy learner. However, because of the relationship between classification time and data size, KNN is regarded as the slowest classifier. It makes predictions by using Euclidean distance to calculate the distance between the query point and the context in the samples. The texts are categorized using training samples and attributes. Finally, the KNN predicts Where news’s are Real or Fake.

3.5 Summary

In this Part, we selected about 44k news from the dataset. the news is divided into real and fake. Firstly, we prepare the data where we cleanup the unused or confusing information which will not aid in the training.

Then we use TF-IDF to determine the most crucial words. It is usually used to find keywords or to rapidly summarize articles. Afterwards, we use different classifier to determine whether Fake news detection accuracy is high.

The commonly used Fake news detection is presented in Fig.3.3. As illustrated the model consists of five main blocks along with few minor components integrated in the system.

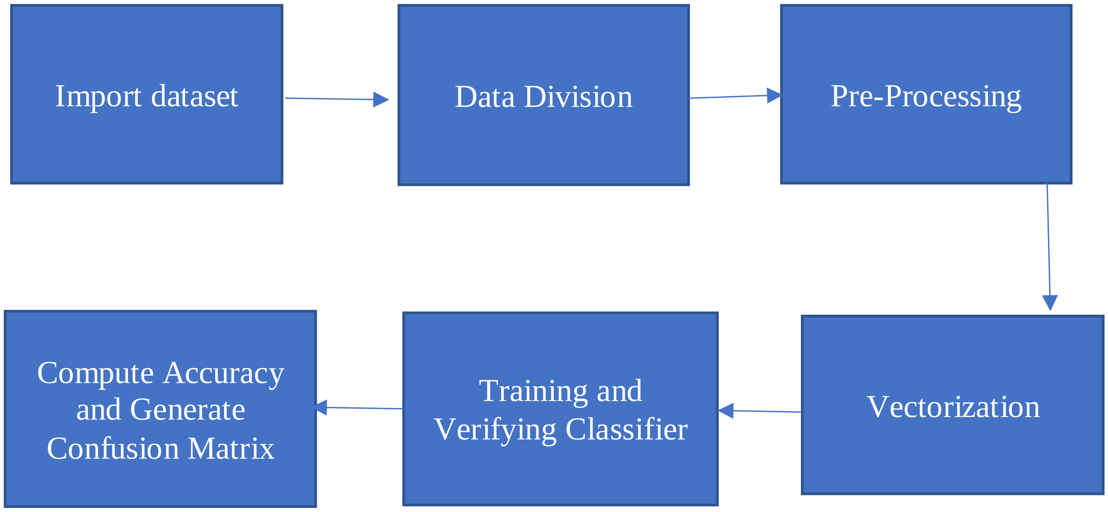

Import dataset
Data Division
Pre-Processing
Vectorization
Training and Verifying Classifier
Compute Accuracy and Generate Confusion Matrix

Fig 3.3: Graphical view for text based Fake news Detection

Chapter-04

Experiment and Results

4.1 Performance Measurement

Performance metrics are a part of every machine learning pipeline. They tell us if we're making progress, and put a number on it. All machine learning models, whether it's linear regression, or a SOTA technique like BRET, need a metric to judge performance.
Every machine learning task can be broken down to either Regression or Classification, just like the performance metrics. There are dozens of metrics for both problems, but we're going to discuss popular ones along with what information they provide about model performance. It's important to know how our model sees our data!

4.1.1 Accuracy: Accuracy is a metric that generally describes how the model performs across all classes. It is useful when all classes are of equal importance. It is calculated as the ratio between the numbers of correct predictions to the total number of predictions.

Accuracy = (TP+TN) / (TP+FP+FN+TN)

- True Negatives (TN): These are the accurately predicted negative values, meaning the current class value is no, and the expected class value is no as well;
- False Positives (FP): When the real class is negative, and the class is expected to be positive;
- False Negatives (FN): The real class is positive, but the class is expected to be negative.

4.1.2 Precision: The precision is calculated as the ratio between the numbers of Positive samples correctly classified to the total number of samples classified as Positive. The precision measures the model's accuracy in classifying a sample as positive.

Precision = TP/ (TP+FP)

4.1.3 Recall: The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect Positive samples. The higher the recall, the more positive samples detected.

Recall = TP / (TP+FN)

4.1.4 F1-score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

F1-score = 2*(Recall * Precision) / (Recall + Precision)

4.2 Result

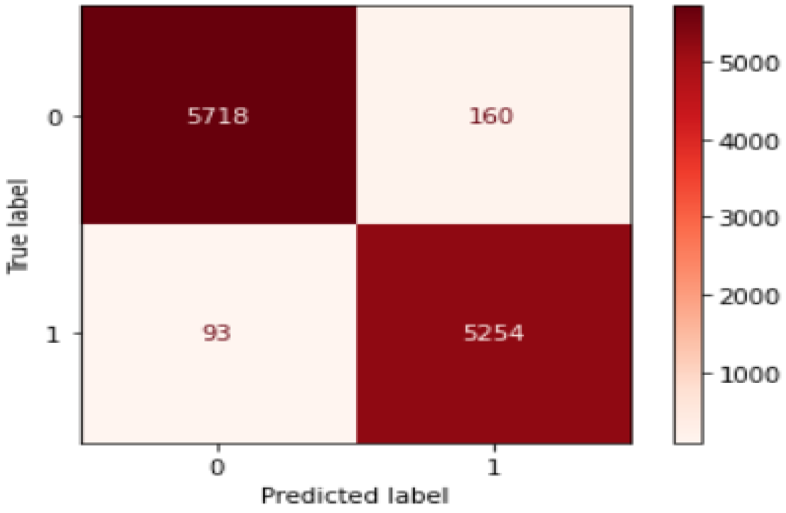


Figure 4.1. Confusion matrix plot for Naive Bayes Classifier.

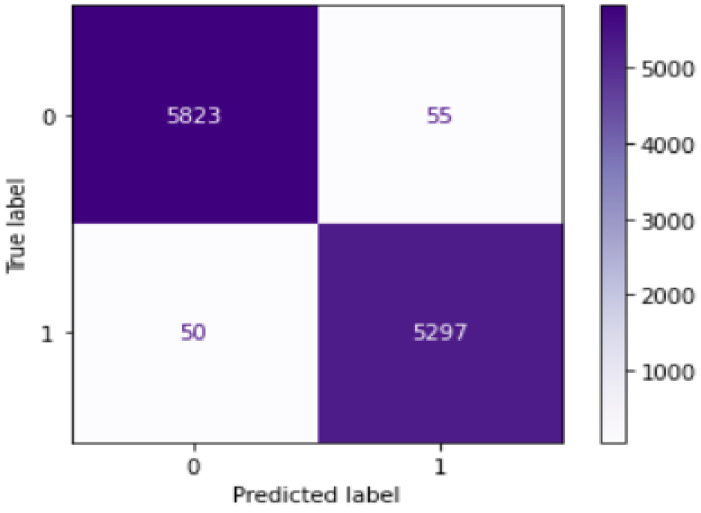


Figure 4.2. Confusion matrix plot for Logistic Regression Classifier.

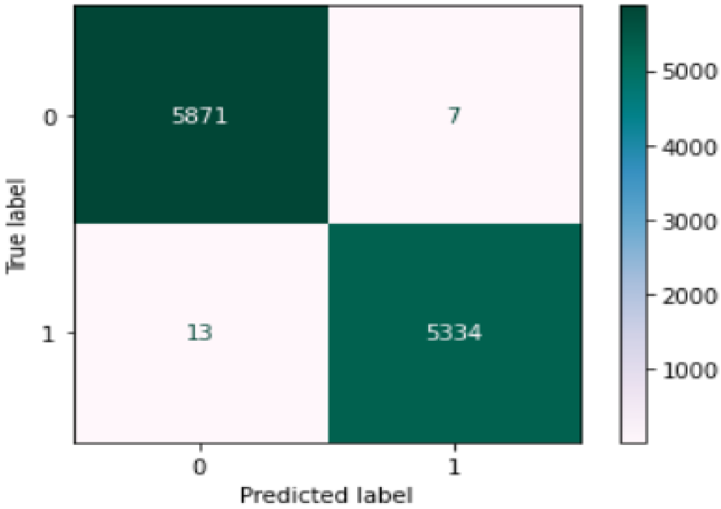


Figure 4.3. Confusion matrix plot for Decision Tree Classifier.

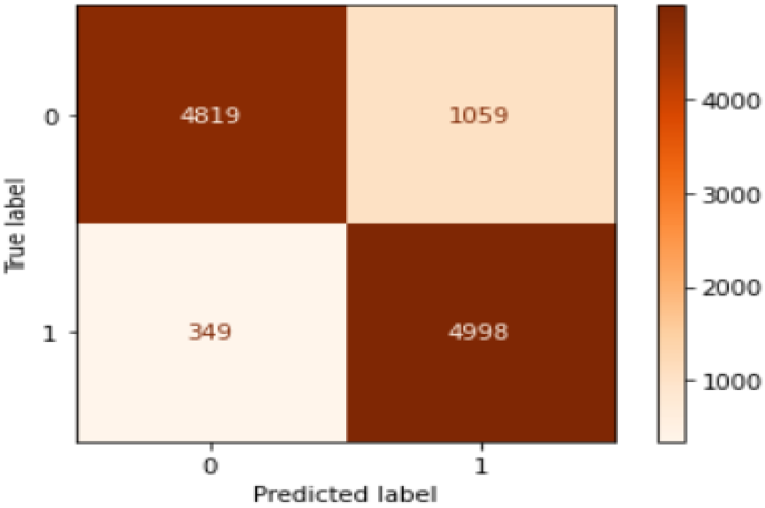


Figure 4.4. Confusion matrix plot for KNN Classifier

Classifier	News	Precision	Recall	F1-score	Accuracy
Naive Bayes	Fake Real	0.98 0.97	0.97 0.98	0.98 0.98	0.98
Logistic Regression	Fake Real	0.99 0.99	0.99 0.99	0.99 0.99	0.99
Decision Tree	Fake Real	1.00 1.00	1.00 1.00	1.00 1.00	1.00
KNN	Fake Real	0.93 0.83	0.82 0.93	0.87 0.88	0.87

Table 4.1. Generated result from all the classifiers

The table-4.1 shows precision, recall, f1-score, and accuracy of all the classifiers for both Real and Fake values. We see that, Decision Tree gives the accuracy of 1.00 percent which is better than other classifiers.

4.3 Summary

We have compared the result of the classification model based on their accuracy. We observed that Decision Tree gave the best accuracy than the others. We went through the steps where we could understand about the accuracy, recall, precision, f1-score and matriculated the ways to calculate those. Afterwards, we generated all the plot for confusion matrix and also showed the generated result table of all the classifiers.

Chapter-05

Conclusion and Future Work

In this Report we used natural language processing to classify fake news. To improve accuracy, NLP methods such as Tokenize and TF-IDF were used.

We did not use bag of words or Word2Vec for vectorizing the dataset, so
we do not know whether it will give better results than TF-IDF.

Traditional ML model algorithms such as naive Bayes classifier, logistic regression, decision tree, and KNN Classifier were implemented. Decision Tree Gives us best accuracy 1.00 percent.

In the future, various models and methods, such as naive bayes and logistic regression, KNN, can be improved by replacing them with other suitable classifiers that work well with the rest of the classifiers, yielding even better results. we wish to conduct a similar study on different languages specially on Bangla.