

**DETECTION OF FAKE NEWS WITH NATURAL LANGUAGE
PROCESSING**

BY

**FAHMIDA HOSSAIN
ID: 193-25-825**

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

Supervised By

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



DAFFODIL INTERNATIONAL UNIVERSITY


DHAKA, BANGLADESH

JANUARY 22, 2022

APPROVAL

This thesis titled “**Detection of Fake News with Natural Language Processing**”, submitted by Fahmida Hossain, ID No: 193-25-825 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 22-01-2022.

BOARD OF EXAMINERS



Chairman

Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

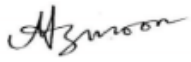
Md. Zahid Hasan

Associate Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Nazmun Nessa Moon

Associate Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

Dr. Mohammad Shorif Uddin

Professor

Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION

I hereby declare that the work entitled “**Detection of Fake News with Natural Language Processing**” submitted to the Daffodil International University, is a record of original work done by me. Except as acknowledged in the text and that the material has not been submitted, either in whole or in part, for a degree at this or any other university.

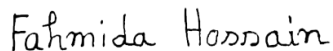
Supervised by:



Dr. Sheak Rashed Haider Noori

Associate Professor and Associate Head,
Department of Computer Science and Engineering,
Faculty of Science and Information Technology,
Daffodil International University

Submitted by:



Fahmida Hossain

ID: 193-25-825
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to almighty Allah for His divine blessing that makes me possible to complete this research work study successfully.

I'm really grateful and wish my profound and indebtedness to Supervisor **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of Computer Science to carry out this research work. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this study.

I would like to express my heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head, Department of CSE, for his kind help to finish my thesis and also to other faculty member and the staff of CSE department of Daffodil International University for their kind help to finish my work successfully

I would like to express my heartiest gratitude to other faculty member and the staff of CSE department of Daffodil International University for his kind help to finish my work successfully.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

In recent years sharing information through internet and various social platform has increased rapidly. It is very hard to find the credibility of those information. People are sharing news without knowing the proper fact. Considering the harmful effect that can be caused by this, detection of fake news has become a new phenomenon of the society. The research process of fake news detection is still in early stage. This paper aims to detect fake news with the help of Natural Language processing (NLP). For this proposed method, three different classifier- logistic regression, decision tree and random forest have been used. The labeled dataset has been collected from a public domain. Use of TF-IDF vectorizer has been made. In this paper the challenge, task formulation and all the steps of the NLP solution has been discussed. The comparison among the classifiers is also shown in this paper.

TABLE OF CONTENTS

DECLARATION.....	iii
ACKNOWLEDGEMENT.....	iv
ABSTRACT.....	v
LIST OF FIGURES.....	viii
LIST OF TABLES.....	x
CHAPTER 1.....	1
INTRODUCTION.....	1
1.1 Aim of the Study.....	2
1.2 Objectives.....	2
1.3 Rationale of this study.....	2
1.4 Report layout.....	3
CHAPTER 2.....	4
LITERATURE REVIEW.....	4
2.1 Related Works.....	4
2.2 Fake News.....	6
2.3 Spreading Process of Fake News.....	6
2.4 Manual Fake News Detection.....	7
2.5 Automated Fake News Detection.....	7
CHAPTER 3.....	9
RESEARCH METHODOLOGY.....	9
3.1 Workflow.....	9
3.2 Data Exploration.....	10
3.3 Data Preprocessing.....	10
3.4 Feature Extraction.....	11
3.4.1 TF-IDF.....	11
3.5 Classification Model.....	12

3.5.1	Logistic Regression.....	12
3.5.2	Decision Tree.....	13
3.5.3	Random Forest.....	14
CHAPTER 4		16
Coding and Implementation		16
4.1	Library Function	16
4.2	Read Dataset from CSV file	16
4.3	Vectorization.....	18
4.4	Modeling.....	18
4.5	Using Classification Model.....	19
4.5.1	Python implementation of Logistic Regression	19
4.5.2	Python Implementation of Decision Tree	19
4.5.3	Python Implementation of Random Forest	20
4.6	Building a Predictive System.....	20
CHAPTER 5		22
RESULT AND CONCLUSION		22
5.1	Performance and Discussion.....	22
5.2	Conclusion	24
CHAPTER 6		26
FUTURE SCOPE		26
6.1	Future Work & Scope	26
APPENDIX		27
Appendix A:	List of Stop Words	27
REFERENCE		29

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Data cleaning steps from dataset to machine learning algorithm	9
Figure 3.3: Using NLTK Libraries for Downloading Stopwords	11
Figure 3.5.1: Logistic Regression Model	13
Figure 3.5.2: Decision Tree Classifier	14
Figure 3.5.3: Random Forest Classifier	15
Figure 4.1: Library Function	16
Figure 4.2.1: Read Dataset from CSV File	17
Figure 4.2.2: Checking for Null Values	17
Figure 4.2.3: Replacing Missing Values with Empty String	18
Figure 4.3: Python Code for Using Tfidf Vectorizer	18
Figure 4.5.2: Python Code for Decision Tree	19
Figure 4.5.3: Python Code for Random Forest	20
Figure 4.6.1: Building A Predictive System	20
Figure 4.6.2: Result of The Predictive System	21

Figure 5.1.1: Accuracy Score of Logistic Regression	22
Figure 5.1.2: Accuracy Score of Decision Tree	22
Figure 5.1.3: Accuracy Score of Random Forest	23

LIST OF TABLES

TABLE	PAGE NO
Table 5.1: Performance of the Classifiers	24

CHAPTER 1

INTRODUCTION

Everyday via several mediums we receive various news from all over the world. But it's really difficult to decide if the news is valid or a propaganda to deceive people. Fake news is considered as a type of yellow journalism that spread hoaxes through print media and social media. Consuming this wrong information can affect society in a very harmful way.

In today's world anyone can publish anything intentionally or unintentionally. People without knowing the proper fact share this information all over the internet. Asking journalists or professionals to verify those claims is a conventional solution which is very time consuming as well as expensive. As the speed of spreading news over the internet is growing rapidly, now it's become very necessary to detect the authenticity of the news in a very short time. With the use of machine learning and artificial intelligence there is a chance that this issue can be resolved or overcome. [11]

The detection of fake news is a very challenging task which requires tremendous efforts. For reducing human time and pressure now a days automated fake news detection has gained interest. This will help to detect and stop spreading the fake news. The task of detecting fake news has been researched from various aspects in the subareas of computer science such as Machine Learning, Data Mining and NLP. [12]

This paper we will mainly focus about the automated fake news detection of text content in NLP. We have used different classifiers in this process. The accuracy score of these classifiers will be counted to determine the best performed model. We will also build a predictive system which will be able to predict the news as fake or true.

1.1 Aim of the Study

The problem of spreading fake news is very difficult to tackle now a days. In today's digital world where there are many information sharing platforms and with the advancement of AI which brings artificial bots that can be used in creating hoax and spreading fake news, it has become a great issue. So, gaining enough knowledge to acknowledge this problem and work on this to analyze, create and solve this issue now has become a necessity. Detecting fake news with NLP along with the knowledge of machine learning will definitely be a great help in this matter.

1.2 Objectives

Everyday huge number of news and information are publishing all over the world. All of these news and information are not true. Knowingly or unknowingly people are sharing this rumor all over the world with the help of internet in a shortest period of time. This kind of fake news can create a lot of problems and damage. Its high time to work on this matter and found a solution about it. Already lots of research are going on about detecting fake news. The objective of this paper is to build a machine learning model to detect fake news using NLP techniques. Different classifier model will be used and their accuracy score will be compared to decide which of them are giving best result.

1.3 Rationale of this study

Fake news has been there even before the era of internet. Fake news is a type of content which provides fabricated article and hoaxes. It can create a lot of misunderstanding and problem in the society. Recently at the time of global pandemic lots of rumors were spread throughout the internet. Rumors about public health is really a very sensitive issue. In today's world fake news is one of those problems which needs to be taken in control immediately. Researcher from all over the world are working on this matter. To prevent the horrific outcome of spreading fake news and rumors, at first, we need to focus on the process of detecting fake news. In this paper the detection process of fake news with the help of NLP has been discussed. A predictive model has been built which will be able to detect the news as 'TRUE' or 'FAKE'.

1.4 Report layout

There are six chapters in this research paper. They are: Introduction, Literature review, Research methodology, Coding and implementation, Result and conclusion, Future Scope.

Chapter One: Introduction; Aim of the study, Objectives, Rationale of this study, Report Layout

Chapter Two: Literature review; Fake news, Spreading process of fake news, Manual fake news detection, Automated fake news detection.

Chapter Three: Research methodology; Work flow, Data exploration, Data preprocessing, Feature extraction, Modeling, Classification model.

Chapter Four: Coding and implementation; library function, Read data from CSV file, Vectorization, Modeling, Using classification model, Building a predictive system

Chapter Five: Result and conclusion; Performance and discussion, Conclusion

Chapter Six: Future scope

CHAPTER 2

LITERATURE REVIEW

Literature review of some great works regarding to fake news detection is reviewing below:

2.1 Related Works

- i. “Fake News Detection Using Machine Learning Ensemble Methods” by Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf and Muhammad Ovais Ahmad**

This research paper discussed about the problem of classifying fake news by using ensemble techniques. They have identified the pattern in text that show the difference between fake news and true news. LIWC tool have been used to extract different features from real news. Then this extracted feature has been used as input to the models. They trained the model to obtain highest accuracy. Multiple performance metrics has also been used to compare the results for each algorithm. The authors also suggest to use graph theory and machine learning algorithm to identify key source of spreading fake news.

- ii. “Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges” by Nicollas R. de Oliveira, Pedro S. Pisa, Martin Andreoni Lopez, Dianne Scherly V. de Medeiros and Diogo M. F. Mattos**

In this paper they analyze the methods for data preprocessing in natural language, machine learning, vectorization and quality assessment of data retrieval. They review the most prevalent algorithms for detecting fake news. They also present quality metrics that was used in extraction of data. This paper mainly focusses on automatic detection by using computational apparatus. Contextualize the identification of fake news and future research initiatives are also discussed on this paper.

- iii. “Development of Fake News Model Using Machine Learning through Natural Language Processing” by Sajjad Ahmed, Knut Hinkelmann and Flavio Corradini**

In this study they use machine learning algorithms for detecting fake news. They applied three classifiers for their task; Naïve Bayes, Support Vector Machine and Passive Aggressive. They tried to show that simple classifier is not enough to detect fake news. In their study they combine text classification with machine learning techniques. Two publicly available dataset has been used for this process. They developed a system which gives 93% accuracy. The issues and challenges are also presented in this paper.[18]

iv. “Techniques of Fake News Detection” by Harshit Garg, Ms. Alisha Goyal, Ms. Ankita Joshi

In this paper the techniques of detecting fake news have been described in details. They introduced the basic concepts of fake news in both print media and social media. They mainly discuss about the approaches to detect fake news. Artificial intelligence, micro blog text and machine learning process has been described in this paper. Different classifier like – random forest, stochastic gradient descent, support vector machine, K-nearest neighbor and decision tree classifier has been broadly discussed here. They also further analyze the datasets, evaluation metrics and future direction to other applications.[10]

v. “A Survey on Natural Language Processing for Fake News Detection” by Ray Oshikawa, Jing Qian and William Yang Wang

In this paper they broadly describe about the automated fake news detection and discuss the challenges involved in detecting fake news. They review and compare all the process, task formulation, dataset and NLP techniques that have been developed for this task. This paper also highlights the difference between fake news detection and other related tasks. They compare and discuss the experimental results of different methods. They also discuss a final guideline for future work.

2.2 Fake News

The term “Fake News” refers to false information which is circulated under the guise of authentic news. It is defined as any information which is misleading and incorrect. Fake news is intentionally designed to mislead the people into believing it to be true. In spite of the lacking of clear understanding on the concept of fake news, the most accepted term is, news which is intentionally false and there is no authenticity of the source.[2]

There are several types of fake news which are listed below:

1. **False connection:** When the content is shared with false information. For example, when caption of a news does not reflect the content.
2. **False context:** A genuine content that is shared with false contextual data.
3. **Manipulated Content:** Misrepresentation of genuine information or imagery to deceive. For example, a news can be popularized by clickbait when its headline is sensationalist.
4. **Satire:** Not usually categorized as fake news and cause no harm but present in a humorous way that may fool readers.
5. **Misleading Content:** Misleading use of information.
6. **Imposter Content:** When genuine information is impersonated with false made-up stories.
7. **Fabricated content:** When a content is completely false and designed to do harm. [20]

2.3 Spreading Process of Fake News

On social network several entities and individuals are responsible to circulate, moderate and spread fake news. As the number of actors are involved in spreading fake news, the problem of identifying it becomes more complicated. The circulation of fake news mostly relies on the social media due to its large scale as well as the reach of social media. Most importantly the ability to share content repetitively helps to spread it immensely. Due to the increasing number of computer communication and easy internet access social media has become the most famous form of fake news dissemination. In print media the journalists and their organizations are mainly take responsibility for publishing any content. But the moderation works differently on social networks.

Each social media has different rules and regulation. The information which grows through social media increases the risk of damage caused by fake news. [2]

There are three different actors that are responsible for spreading fake news. They are – the adversary, the fact checker and the susceptible user. Malicious individuals faking as ordinary social network users using fake accounts or bots are known as adversary. They can work as a source or a promoter of fake news. The fact checker consists of various organization that confirm the news which has doubts about its authenticity. Checking authenticity of a news relies on fact checking journalism that are verified by human. So, this fact checking is not really very reliable. However, there are some automated solutions which use artificial intelligence to detect fake news for companies and consumers. Finally, the susceptible users are the social network users who are not able to differentiate between the fake news and authentic news and end up sharing the fake news on their social network account. They do it unknowingly without having any bad intention of spreading fraudulent content. [2]

2.4 Manual Fake News Detection

To detect fake news manually all the techniques and procedures that involved, causes a lot of time and struggle. The amount of data that generates daily on online is huge. Manual fact checking involves visiting various sites and news sources to compare the real news with the fake ones. On the other hand, as the technology has become very upgraded, all the information spread very fast through online. And if the information is fake, then it can create a lot of chaos in a short period of time. These reasons make manual fact checking very ineffective. Because after using tremendous amount of time and manpower the result we get, doesn't add much value. Hence the reason behind the use of automated fake news detection gets highlighted.

2.5 Automated Fake News Detection

Automated fake news detection mainly follows two steps. One is fact extraction and another one is fact checking. In fact extraction datasets are collected from internet as raw facts. Then this dataset is being preprocessed and after following several techniques, data is labeled as fake or real. There are many approaches and techniques that are used for automated fake news detection. One of them are machine learning approach. Machine learning algorithms are used in this approach to detect misinformation. Some of these algorithms are

- Naïve Bayes
- Decision Tree
- Random Forest
- Support Vector Machine
- Logistic Regression
- K-nearest-neighbor

In machine learning techniques NLP is used as a method for detecting fake news. NLP enables computer to understand human language. The accuracy score of automated fake news detection depends on the combination of models that are used to train the data. [32]

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Workflow

For detecting fake news, several steps need to be followed. In this study we will use NLP as a python computational tool. PANDAS which is an open-source library for data structure and data analysis, will be used. After collecting dataset, the first step is preprocessing the test and train data. Then with this preprocessed dataset, we will continue to our next step which is feature extraction. For feature extraction we will use NLTK (Natural Language Toolkit) library. TF-IDF vectorizer will be used for vectorization.

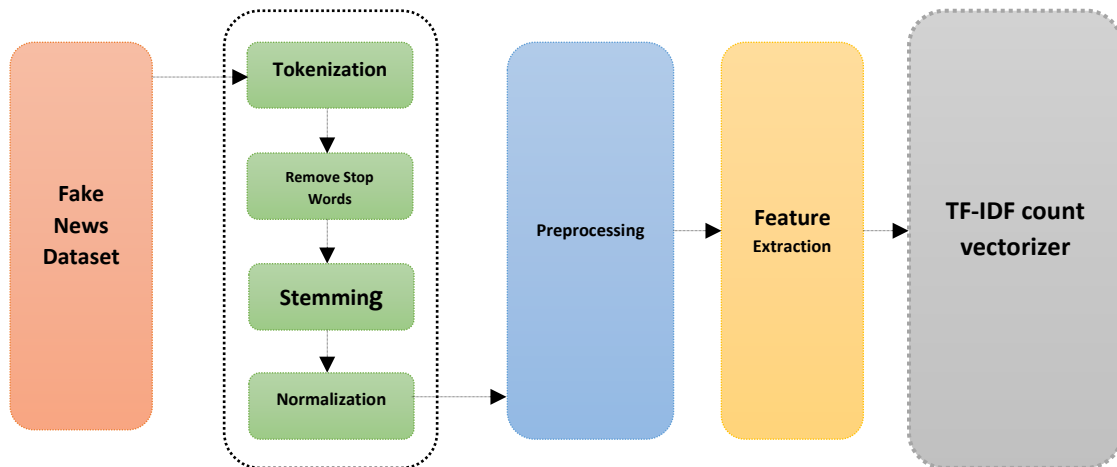


Figure 3.1: Data cleaning steps from dataset to machine learning algorithm

After that the extracted feature will be applied in machine learning algorithm. In this paper we will use logistic regression, decision tree and random forest classifier algorithm.

For programming part, we will use NUMPY library and the program will be run on Jupyter Notebook.

The goal is to build an NLP model to detect fake news. We will also test the efficiency of the classifiers using accuracy.

3.2 Data Exploration

The dataset which we are using for classification is collected from an open source called “Kaggle”. It is a CSV dataset. The dataset has collection of 20,800 news articles. It has five columns – id, title, author, text and label. The label has two binary value 0 and 1. The real value is labeled as 0 and fake value is labeled as 1. News articles of the dataset are collected from different news organization.

3.3 Data Preprocessing

Data preprocessing is a mandatory step in building machine learning model. The objective of preprocessing is to reduce the size of the data by removing information which is not necessary. The performance of a classifier depends on the size and quality of the text data. It’s very important to preprocess the data before using any classification model. Various python libraries like NLTK, NumPy and pandas are used for data preprocessing. Data preprocessing includes tokenization, stemming, lemmatization or weighting words and stop word removal. [26]

- **Tokenization:** Tokenization breaks the text data into individual words. It converts the words to their base form. It splits the entire text document in smaller unit. These smaller units are called token. Tokens are the elementary step of stemming and lemmatization. Tokenization is the extremely important step in NLP.
- **Stemming:** Stemming decreases the number of words. As an example, if there are three similar words in the text data like “Chocolates”, “Chocolatey” and “Choco”, it will be reduced and changed to the word “chocolate”.
- **Lemmatization:** Understanding the context lemmatization converts word into its meaningful base form. Lemmatization groups the different form of similar word together and analyze it as a single word.
- **Stop word removal:** Stop words are the words that are used to connect words bur doesn’t add much value in the context or meaning of the sentence. Stop words removal helps to remove the common words, preposition and conjunction.

```
In [2]: import nltk
        nltk.download('stopwords')
```

Figure 3.3: Using NLTK Libraries for Downloading Stopwords

3.4 Feature Extraction

Feature extraction is necessary to reduce the amount of redundant data from the text data set. It helps to build the model to increase the generalization steps in the machine learning. It also increases the speed of learning. As machine learning algorithms can't work on raw data, some feature extraction technique has been used to convert the text into a matrix of features. Some popular feature extraction techniques are:

- TF-IDF
- Bag of words
- Count vectorizer
- N-Gram Model

In this paper TF-IDF has been used for feature extraction purpose.

3.4.1 TF-IDF

TF-IDF stands for term frequency-inverse document frequency. It is used to measure a term in a document over dataset. The TF-IDF value increases or decreases proportionally according to the number of times a word appears in the document or number of documents in the corpus that contain the word. [7]

TF-IDF has most important use in automated text analysis. It is very helpful for scoring words in machine learning algorithms for NLP.

It has two sub parts:

- Term Frequency (TF)

- Inverse Document Frequency (IDF)

Term Frequency (TF)

Term frequency measures how often a term appears in a corpus. It calculates how many times a word occurs with respect to the total number of words. It is formulated as

$$TF(t) = \frac{\text{Number of time the terms } t \text{ appears in a document}}{\text{Total number of words in that document}}$$

There are other ways too for calculating term frequencies. Such as using maximum term frequency as well as average term frequency in a document. [8]

Inverse Document Frequency (IDF)

Inverse document frequency is used to measure the importance of a word in a document. It specifies how common or rare a word is in a corpus. The more frequent the word uses, the lower it scores. IDF is formulated as

$$IDF = \log \frac{N}{DFt}$$

Here N is the total number of documents and DFt is the number of documents that contains the term t. [22]

3.5 Classification Model

Different classifier can be used for modeling the dataset. The classifier that we have been used are described below:

3.5.1 Logistic Regression

Logistic regression is a supervised learning model. It is one of the most popular machine learning algorithms. This algorithm is mainly used to predict probability of dependent variables. Therefore, the output must be a discrete value and binary in nature. So, the result either would be 0 or 1, yes or no and true or false etc. Logistic regression is used for solving classification problem. It can find probabilities and can also classify new data by using continuous and discrete dataset. For initial classification logistic regression requires large sample size. [15]

The equation of logistic regression is-

$$y = \frac{e^{(b_0 + b_1 \cdot x)}}{1 + e^{(b_0 + b_1 \cdot x)}} \quad [14]$$

Here y is the output, b_0 is the intercept term and b_1 is the coefficient for the input value x .

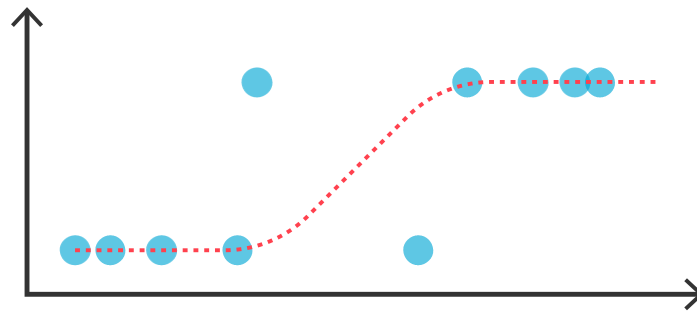


Figure 3.5.1: Logistic Regression Model

Logistic regression is used in machine learning to make accurate prediction. It is similar to linear regression but its target variable is binary.

3.5.2 Decision Tree

Decision tree is a supervised learning model which is used for both classification and regression problem. It splits dataset by selecting feature. The features can be in nominal or continuous form. It's a tree structured classifier. Internal nodes of this classifier represent the feature of the dataset, branches represent decision rules and the leaf nodes represent the outcome.[19]

The decision tree algorithm starts from the root node of the tree. It compares the values of the root node with real dataset. Based on this comparison it follows the brunch and jumps to the next node. This process continues till it reaches the leaf node.[21]

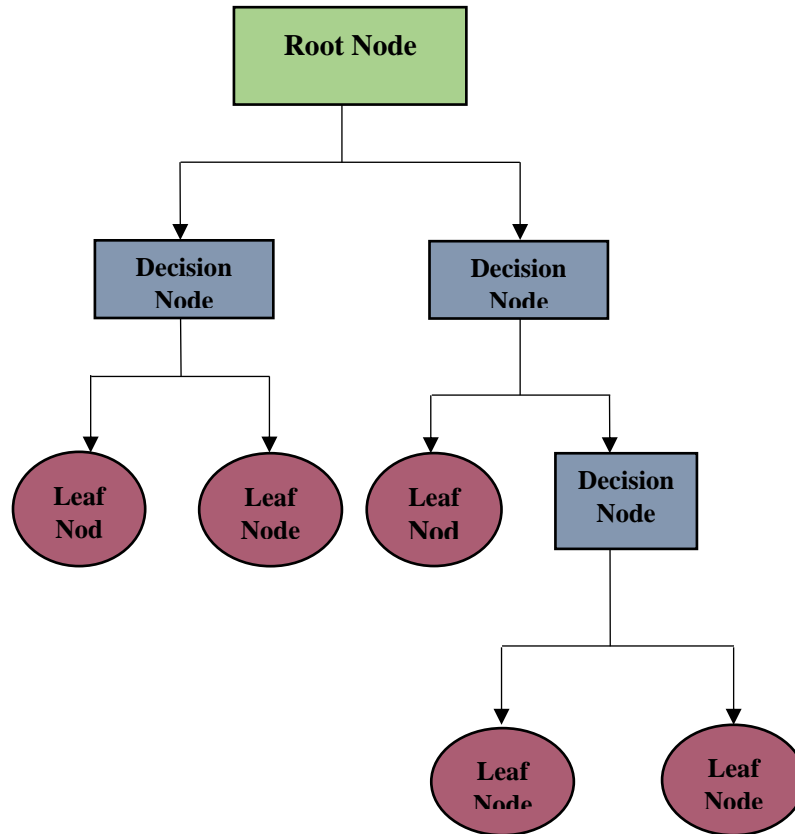


Figure 3.5.2: Decision Tree Classifier

3.5.3 Random Forest

Random forest is a regulated AI method. It is a popular classification model which is supervised in nature. It is an advanced form of decision tree. The forest it builds is an ensemble of decision trees. The increasing number of trees in the forest leads to the limited number of error rate. Multiple decision trees are built during training process while operating this classifier. The nodes of this classifier are split based on Gini index.[4]

Random forest applies a technique call bagging to individual trees in ensemble. Bagging frequently selects sample from training set and fits trees to this sample. The error rate in random forest is low compared to other classifier. It is capable of work on big dataset with high dimensionality.[5]

One biggest advantage of random forest is it can be used for both classification and regression problem.[4]

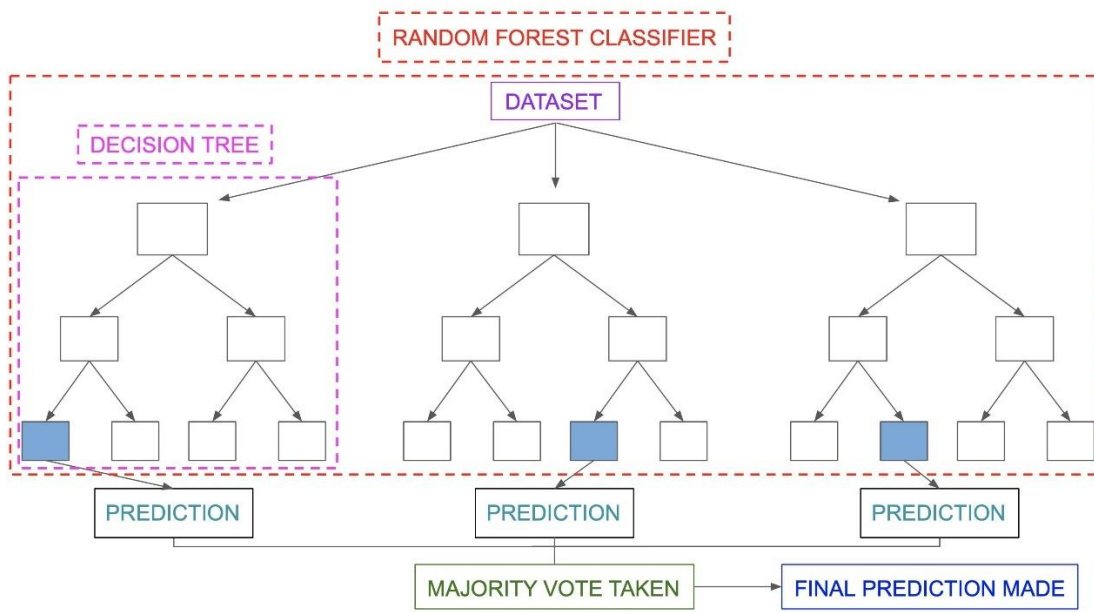


Figure 3.5.3: Random Forest Classifier [26]

CHAPTER 4

Coding and Implementation

4.1 Library Function

In this study we have used Python libraries such as NumPy, pandas and sklearn. First, we need to import all those libraries and the classifier we will use in this process. We have imported logistic regression, decision tree and random forest classifier from sklearn.

```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
```

Figure 4.1: Library Function

4.2 Read Dataset from CSV file

After importing all the libraries, we need to read the dataset from CSV file.

```
dataset = pd.read_csv('train.csv')

dataset.shape

(20800, 5)

dataset.head()
```

	id	title	author	text	label
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucas	House Dem Aide: We Didn't Even See Comey's Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Figure 4.2.1: Reading Dataset from CSV File

Now we need to check if there is any null value.

```
In [7]: # counting missing values
dataset.isnull().sum()

id          0
title      558
author     1957
text        39
label       0
dtype: int64
```

Figure 4.2.2: Checking for Null Values

As we can see there are some null values in Fig 4.2.2 now, we need to replace these missing values with empty string. After these the dataset is ready to use for preprocessing.

```
In [8]: # replacing missing values with empty string
dataset=dataset.fillna('')
```

Figure 4.2.3: Replacing Missing Values with Empty String

4.3 Vectorization

Vectorization is used in NLP to map words which is used to find word prediction and similarities in words. Vectorization converts word into numbers. Here we use Tfidf vectorizer for this process.

```
In [22]: vectorizer = TfidfVectorizer()
vectorizer.fit(X)

X = vectorizer.transform(X)
```

Figure 4.3: Python Code for Using Tfidf Vectorizer

4.4 Modeling

After vectorization we are splitting the data into test and train data.

```
Splitting the dataset to train and test data

In [24]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, stratify=Y, random_state=2)
```

Figure 4.4: Splitting the Dataset

We use three models into the data, logistic regression, decision tree and random forest. After that we will use these models to detect fake news and also calculate their accuracy score.

4.5 Using Classification Model

In this part we will implement our classifier by using python.

4.5.1 Python implementation of Logistic Regression

```
LR = LogisticRegression()  
  
LR.fit(X_train, Y_train)  
  
LogisticRegression()
```

Figure 4.5.1: Python Code for Logistic Regression

4.5.2 Python Implementation of Decision Tree

```
DT = DecisionTreeClassifier()  
DT.fit(X_train, Y_train)  
  
DecisionTreeClassifier()
```

Figure 4.5.2: Python Code for Decision Tree

4.5.3 Python Implementation of Random Forest

```
RFC = RandomForestClassifier(random_state=0)
RFC.fit(X_train, Y_train)

RandomForestClassifier(random_state=0)
```

Figure 4.5.3: Python Code for Random Forest

4.6 Building a Predictive System

Now we will build a predictive system by which we will get to check if the data is real or fake. If the data is true then it will show the output as 'Real News'. Otherwise, it will show 'Fake News'.

```
X_new = X_test[3]

prediction_LR = LR.predict(X_new)
prediction_DT = DT.predict(X_new)
prediction_RFC = RFC.predict(X_new)

if (prediction_LR[0]==0):
    print('Logistic Regression: Real News')
else:
    print('Logistic Regression: Fake News')

if (prediction_DT[0]==0):
    print('Decision Tree: Real News')
else:
    print('Decision Tree: Fake News')

if (prediction_RFC[0]==0):
    print('Random Forest: Real News')
else:
    print('Random Forest: Fake News')
```

Figure 4.6.1: Building a Predictive System

```
Logistic Regression: Real News  
Decision Tree: Real News  
Random Forest: Real News
```

Figure 4.6.2: Result of the Predictive System

We can see all of the three classifier is giving the same result. We have checked the 4th news of the dataset and our predictive system is showing that the news is true. As all of our classifier has given the same result so it is confirmed that the system is working accurately.

CHAPTER 5

RESULT AND CONCLUSION

5.1 Performance and Discussion

We have followed all the steps and procedure described in chapter 3 and 4 and got the accuracy score from those procedure.

Accuracy score of logistic regression:

```
#Test Data
X_test_prediction = LR.predict(X_test)
testdata_accuracy = accuracy_score(X_test_prediction, Y_test)

print('Accuracy score of the test data : ', testdata_accuracy)

Accuracy score of the test data : 0.9790865384615385
```

Figure 5.1.1: Accuracy Score of Logistic Regression

Accuracy score of decision tree:

```
#Test Data
X_test_prediction = DT.predict(X_test)
testdata_accuracy = accuracy_score(X_test_prediction, Y_test)

print('Accuracy score of the test data : ', testdata_accuracy)

Accuracy score of the test data : 0.9915865384615384
```

Figure 5.1.2: Accuracy Score of Decision Tree

Accuracy score of Random Forest:

```
#Test Data
X_test_prediction = RFC.predict(X_test)
testdata_accuracy = accuracy_score(X_test_prediction, Y_test)

print('Accuracy score of the test data : ', testdata_accuracy)

Accuracy score of the test data : 0.9942307692307693
```

Figure 5.1.3: Accuracy Score of Random Forest

For logistic regression we have got the accuracy of 0.98. For decision tree and random forest, the accuracy score is 0.99.

But the accuracy score is not a very good measure to understand the performance. So, after getting the accuracy score, we also check the result through precision and recall. The definition of precision and recall are as follows:

Precision: The ratio of positive data that are correctly predicted by the classifier to the total number of data which are predicted positive.[4]

The mathematical form of precision is,

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP}$$

Recall: The ratio of total number of true positive and the actual number of data that are positive.[4]

The mathematical form of recall is,

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN}$$

Table 1 shows the performance of our proposed classifier.

Table 5.1: Performance of the Classifiers:

CLASSIFIER	ACCURACY	PRECISION	RECALL
LOGISTIC REGRESSION	0.98	0.97	0.99
DECISION TREE	0.99	0.99	0.99
RANDOM FOREST	0.99	0.99	1.00

After implementing the classifier and comparing their performance we can observe that the random forest classifier has performed better than the other classifier with the accuracy of 0.99, precision of 0.99 and recall of 1.00.

5.2 Conclusion

With the growing number of using social media and various online contents, the spreading of fake news is increasing in a very alarming rate. Now a days people are more engaged with online news portal more than printed one. Most of the information that are on internet is unverified and people assumed it to be true. So, it has become very necessary to find a solution for this and to help people to determine if the news is authentic or a hoax. In this paper we have presented a detection process of fake news with the help of natural language processing (NLP). We have used a dataset collected from a public domain called “Kaggle”.

For building the model we have used Python. Libraries like PANDAS, NumPy and various NLP toolkits are also used for this purpose.

In this study we have used three classification model- logistic regression, decision tree and random forest. Among these the random forest stands out with an accuracy of 0.99.

The result shows that the approach we have used is highly favorable since it has showed a very good accuracy score. We have also calculated the precision and recall score of the performance of the classifier. We built a predictive system to show if the result is fake or real.

There are also various classification models which can be used for this purpose. Decision tree and logistic regression has also performed very well in this specific dataset.

CHAPTER 6

FUTURE SCOPE

6.1 Future Work & Scope

Fake news detection has become a very important issue now a days. Several researches are going on in this matter. There are few limitations in our proposed method that can be worked on. This solution only marks the news as authentic or unauthentic. But to find the source or credibility of real news is also necessary as it helps people to believe the news more. The detection process is only a small step of bigger problem.

In this study the classifier that has been used, none of them give the accuracy of 100%. This method doesn't show proper result if the train and test dataset don't contain the data of same categories.

Only one dataset is being used in this proposed method. Uses of different dataset will help more to measure the performance of this method.

There are many scopes to work on this field. It has many open issues that needs attention of researchers. For reducing the spread of fake news, identifying the main elements is very important. Deep learning techniques and graph theory can be used to identify the main source involved in this matter. Real time fake news detection in video is another direction to work on.[5]

Investigation on neural network models to see if the hand-crafted features can be combined with this and proper use of non-textual data can also be great scope to work on.

In recent years Misinformation Detection (MID) has also become a great research topic.

So, as we can see there are many scopes and field regarding fake news detection. Hope we will get a perfect solution of this matter in upcoming years.

APPENDIX

Appendix A: List of Stop Words

A	About	Above	After	Again
Against	All	Am	An	And
Any	Are	Aren't	As	At
Be	Because	Been	Before	Being
Below	Both	But	By	Can't
Could	Couldn't	Did	Didn't	Do
Does	Doesn't	Doing	Don't	Each
Few	For	From	Had	Hadn't
Has	Hasn't	Have	Haven't	He
He'll	He's	Her	Herself	Him
Himself	How	How's	I	I'd
I'll	I'm	I've	In	Into
Is	Isn't	It	It's	Itself
Let's	Me	Most	Mustn't	My
Myself	No	Not	Of	Off

On	Only	Or	Other	Our
Ourselves	Out	Over	Same	She
She'd	She'll	She's	Should	Shouldn't
So	Some	That	The	Their
Them	Themselves	Then	There	These
They	They'd	They'll	They're	They've
This	Those	To	Too	Under
Until	Up	Very	Was	Wasn't
We	We'd	We'll	We're	We've
Were	Weren't	What	What's	Which
While	Who	Who's	Whom	Why
Why's	With	Would	Wouldn't	You
You'll	You'd	You'll	You're	You've
Your	Yours	Yourself	Yourselves	

REFERENCE

- [1] Ray Oshikawa, Jing Qian, William Yang Wang, *A Survey on Natural Language Processing for Fake News Detection*, May 2020. [Online]. Available: <https://aclanthology.org/2020.lrec-1.747/>
[Accessed: 20-Jan-2022]
- [2] Ahmed Sa, Knut Hinkelmann, *Development of Fake News Model Using Machine Learning through Natural Language Processing*, December 2020.[Online].
Available:
https://www.researchgate.net/publication/344328246_Development_of_Fake_News_Model_using_Machine_Learning_through_Natural_Language_Processing
[Accessed: 20-Jan-2022]
- [3] Z Khanam, B N Alwasel, H Sirafi and M Rashid, *Fake News Detection Using Machine Learning Approaches*, December 2020. [Online]. Available:
<https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf>
[Accessed: 20-Jan-2022]
- [4] Vasu Agarwal, H. Parveen Sultana, Srijan Malhotra, Amitrajit Sarkar, *Analysis of Classifier for Fake News Detection*, October 2020. [Online]. Available: <https://pdf.sciencedirectassets.com/>
[Accessed: 20-Jan-2022]
- [5] U Mertoğlu, *Automated Fake News Detection in The Age of Digital Libraries*, November 2020. [Online]. Available: <https://ejournals.bc.edu/index.php/ital/article/view/12483>
[Accessed: 20-Jan-2022]
- [6] Iftikhar Ahmad , 1 Muhammad Yousaf, 1 Suhail Yousaf , 1 and Muhammad Ovais Ahmad, *Fake News Detection Using Machine Learning Ensemble Method*, October 2020. [Online]. Available: <https://www.hindawi.com/journals/complexity/2020/8885861/>
[Accessed: 20-Jan-2022]
- [7] Ray Oshikawa, Jing Qian, William Yang Wang, *A Survey on Natural Language Processing for Fake News Detection*, May 2020. [Online]. Available: <https://arxiv.org/abs/1811.00770>
[Accessed: 20-Jan-2022]

- [8] Fathima Nada¹, Bariya Firdous Khan², Aroofa Maryam³, Nooruz-Zuha⁴, Zameer Ahmed, *Fake News Detection Using Logistic Regression*, May 2019. [Online]. Available: <https://www.irjet.net/archives/V6/i5/IRJET-V6I5733.pdf>
[Accessed: 20-Jan-2022]
- [9] Pranav Bharti, Mohak Bakshi, R.Annie Uthra, *Fake News Detection Using Logistic Regression, Sentiment Analysis and Web Scraping*, May 2020. [Online]. Available: <http://sersc.org/journals/index.php/IJAST/article/view/15097>
[Accessed: 20-Jan-2022]
- [10] Resham N. Waykole, Anuradha D. Thakare, *A Review of Feature Extraction Methods For Text Classification*, April 2018. [Online]. Available: http://ijaerd.com/papers/finished_papers/A_review_of_feature_extraction_methods_for_text_classification-IJAERDV05I0489982.pdf
[Accessed: 20-Jan-2022]
- [11] Noman Islam, Asadullah Shaikh, Asma Qaiser, Yousef Asiri, Sultan Almakdi, Adel Sulaiman, Verdah Moazzam and Syeda Aiman Babar, *Ternion: An Autonomous Model for Fake News Detection*, September 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/19/9292>
[Accessed: 20-Jan-2022]
- [12] Veronica P ´erez-Rosas ´, Bennett Kleinberg, Alexandra Lefevre and Rada Mihalcea, *Automatic Detection of Fake News*, August 2017. [Online]. Available: <https://arxiv.org/abs/1708.07104>
[Accessed: 20-Jan-2022]

Detection_of_Fake_News_with_NLP.pdf

ORIGINALITY REPORT

25%

SIMILARITY INDEX

16%

INTERNET SOURCES

16%

PUBLICATIONS

14%

STUDENT PAPERS

PRIMARY SOURCES

1

dspace.daffodilvarsity.edu.bd:8080

Internet Source

4%

2

Nicollas R. de Oliveira, Pedro S. Pisa, Martin Andreoni Lopez, Dianne Scherly V. de Medeiros, Diogo M. F. Mattos. "Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges", Information, 2021

Publication

2%

3

Submitted to Daffodil International University

Student Paper

2%

4

Noman Islam, Asadullah Shaikh, Asma Qaiser, Yousef Asiri, Sultan Almakdi, Adel Sulaiman, Verdah Moazzam, Syeda Aiman Babar. "Ternion: An Autonomous Model for Fake News Detection", Applied Sciences, 2021

Publication

2%

5

www.hindawi.com

Internet Source

1%

6

www.geeksforgeeks.org

Internet Source

1%