

**A STUDY ON FAKE NEWS DETECTION USING SUPERVISED MACHINE  
LEARNING CLASSIFICATION ALGORITHMS  
SOFTWARE ENGINEERING**

Submitted by:

**AKANKSHA SINGH  
2K20/SWE/02**

Under the supervision of  
**Mr. Sanjay Patidar**  
(Assistant Professor)



**DEPARTMENT OF SOFTWARE ENGINEERING  
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)**

Bawana Road, Delhi-110042

**MAY, 2022**


**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi - 110042

**DECLARATION**

I, Akanksha Singh, Roll No. 2K20/SWE/02 student of M. Tech (Software Engineering), hereby declare that the project Dissertation titled “A Study on Fake News Detection using Supervised Machine Learning Classification Algorithms” which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of and Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: Delhi

Date: 31<sup>st</sup> May 2022

  
Akanksha Singh  
(2K20/SWE/02)

**DEPARTMENT OF SOFTWARE ENGINEERING**  
**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Bawana Road, Delhi – 110042

**CERTIFICATE**

I hereby certify that the Project Dissertation titled “A Study on Fake News Detection using Supervised Machine Learning Classification Algorithms” which is submitted by Akanksha Singh (2K20/SWE/02) Department of Software Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 31<sup>st</sup> May, 2022

  
Mr. Sanjay Patidar

(Assistant Professor)


Department of Software Engineering

## ACKNOWLEDGMENT

First and foremost, I would like to convey my heartfelt sincere gratitude to **Mr. Sanjay Patidar**, Assistant Professor, for his continuous encouragement and support throughout my master's degree. He has not only been my supervisor in academics, but also a source of inspiration during my master's studies. I would also like to acknowledge all the faculty at DTU for providing the right academic resources and environment for this work to be carried out.

I would also like to express my gratitude to the University for providing us with the laboratories, infrastructure, testing facilities, and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors, and our peer group who aided us with all the knowledge they had regarding various topics.

  
Akanksha Singh

2K20/SWE/02

## **ABSTRACT**

The Online world has become an essential part of everybody's life in today's society. Almost everyone is using social media to gain information about all around the world. It is an easily accessible source of day-to-day news for almost everyone throughout the globe due to its relatively low cost, ease of access, and rapid dissemination. However, this news comes with a risk of being faulty or fake to mislead the readers. On one hand, it is the most inexpensive, easy, and convenient way of getting information in no time, on the other hand, understanding the background from a headline is vital, the much more difficult task would be predicting the user's purpose; this prediction would be a springboard in the field of natural language processing to identify fake news. False news/information has a tremendous impact on our social lives, in fact, in all fields, particularly politics and education, and organization. The propagation of false information has the potential to create significant social and emotional harm, as well as have potentially dangerous consequences. As a response, automated Fake news detection has become essential for maintaining a sustainable online and social media presence. This study includes the findings of various research being carried out to identify fake news and the possibility of fake content in a particular news article being predicted. In this study, the main motive is carrying out the experiment for the detection of false information by using NLP techniques and five supervised machine learning classification algorithms: Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and Logistic Regression and selecting the best algorithm.

# CONTENTS

<b>Declaration</b>	<b>i</b>
<b>Certificate</b>	<b>ii</b>
<b>Acknowledgement</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>Contents</b>	<b>v</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Symbols, Abbreviations</b>	<b>x</b>
<b>CHAPTER 1 INTRODUCTION</b>	<b>1-4</b>
1.1 Natural language processing and Machine Learning	1
1.2 Motivation	2
1.3 Objective	3
1.4 Thesis structure	3
<b>CHAPTER 2 LITERATURE REVIEW</b>	<b>5-8</b>
<b>CHAPTER 3 THEORETICAL CONCEPTS</b>	<b>9-11</b>
3.1 Supervised Machine Learning	9
3.2 Unsupervised Machine Learning	10
3.3 Features	10
3.4 Text Preprocessing	10
3.5 Feature Extraction	11
3.6 Python Libraries	11

<b>CHAPTER 4 RESEARCH METHODOLOGY</b>	<b>12-22</b>
4.1 Proposed Model	12
4.2 Data Collection	13
4.3 Data Preprocessing	14
4.3.1 Removing Unnecessary Columns	14
4.3.2 Lowercase Conversion	14
4.3.3 Punctuation Removal	14
4.3.4 Stopword Removal	14
4.4 Train and Test Split	15
4.5 Feature Extraction	15
4.5.1 Count Vectorizer	15
4.5.2 TF-IDF	16
4.5.2.1 Term Frequency	16
4.5.2.2 Inverse Document Frequency	17
4.6 Machine Learning Supervised Classification Techniques	17
4.6.1 Logistic Regression	17
4.6.2 Naïve Bayes	18
4.6.3 Support Vector Machine	19
4.6.4 Decision Tree	19
4.6.5 Random Forest	20
4.7 Model Evaluation	21
4.7.1 Accuracy	22
4.7.2 Precision	22
4.7.3 Recall	22
4.7.4 F1-Measure	21

<b>CHAPTER 5 EXPERIMENTAL RESULTS</b>	<b>22-30</b>
5.1 Data Exploration	23
5.1.1 Number of Fake and True Articles	24
5.1.2 Articles per Subject	24
5.1.3 Words appearing frequently in Fake and True News Articles	24
5.1.4 Word Cloud for Fake and True News	26
5.2 Confusion Matrix Obtained from ML Classifiers	27
5.3 Accuracy and Performance Measures	29
<b>CHAPTER 6 CONCLUSION AND FUTURE SCOPE</b>	<b>31</b>
<b>REFERENCES</b>	<b>32-35</b>
<b>PUBLICATION</b>	<b>36</b>



## LIST OF FIGURES

1.1	Venn Diagram for AI, ML, DL and NLP	2
3.1	Supervised Machine Learning	9
3.2	Unsupervised Machine Learning	10
4.1	Proposed Model	12
4.2	Head of the data	13
4.3	Support Vector Machine	19
4.4	Decision Tree	20
4.5	Random Forest	21
4.6	Confusion Matrix	21
5.1	Number of Fake and True Articles	23
5.2	Articles Per Subject	24
5.3	Frequent Word Count for Fake News Articles	25
5.4	Frequent Word Count for Real News Articles	25
5.5	Word Cloud for Fake News Articles	26
5.6	Word Cloud for Real News Articles	26
5.7	CM using Logistic Regression	27
5.8	CM using Naïve Bayes	27
5.9	CM using Support Vector Machine	28
5.10	CM using Decision Tree	28
5.11	Confusion Matrix using Random Forest	28
5.12	Graph for Comparison of all the Classifiers	30

## **LIST OF TABLES**

4.2 D <sub>x</sub> N Matrix	15
4.2 Count Vectorizer Matrix Representation	16
5.1 TP, TN, FP and FN values for Classifiers	30
5.2 Accuracy and Performance Measures	31

## **LIST OF ABBREVIATIONS**

AI: Artificial Intelligence

ML: Machine Learning

DL: Deep Learning

NLP: Natural Language Processing

LR: Logistic Regression

NB: Naïve Bayes

SVM: Support Vector Machine

DT: Decision Tree

RF: Random Forest

NLTK: Natural Language Toolkit

CM: Confusion Matrix

TP: True Positive

FP: False Positive

TN: True Negative

FN: False Negative

# CHAPTER 1

## INTRODUCTION

For a long time, social media has taken over a meaningful place in people's life. Fake news primarily prevails via social media and articles available online. Fake news indulges politics, democracy, education as well as finance and business at risk. Even while false news is not a new issue, people these days place a larger focus on social media, which leads to the acceptance of deceitful remarks and the subsequent propagation of the same wrong information. It is getting harder to tell the difference between accurate and misleading news these days, which leads to confusion and complications. Manually recognizing fake news is tough; it is only achievable when the individual identifying the news has extensive expertise in the subject. Fake news can destroy someone's career and if it is political and harm the nation and citizens of that country as well as it can also affect businesses, products, and reputations.

### 1.1 NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

Both NLP and ML are Artificial Intelligence subsets. A system can be automated without being programmed explicitly with the help of ML. It allows systems to learn and develop from experience.

Humans can communicate, exchange ideas, and grasp each other's perspectives, but machine cannot. Humans can communicate with one another using natural language. To function, the machine needs instructions in a methodical manner. We must teach the computer to understand natural language so that people and machines can communicate. Natural language processing aids in the completion of these activities. Artificial has many areas and NLP is one of them that allows people and machines to converse. It teaches the computer or gadget how to communicate with humans through speech or text. It enables machines to read, comprehend, and interpret messages from human language. Voice search, comprehension, and generation are all part of natural language processing.

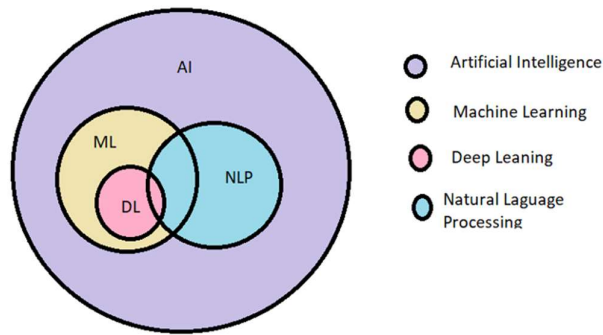


Figure 1.1. Venn Diagram for AI, ML, DL and NLP

The impact of false information has grown, often spilling over into the digital realm and endangering law enforcement. Considering the enormous volume of Web material, automating false information identification is a significant NLP problem that will benefit all digital content producers by reducing the amount of time and energy helps to identify and slow the transmission of false news.

## 1.2 MOTIVATION

The subterfuge film “Plandemic” which was released online in May 2020 caused a huge number of fake news to prevail in public that misguided the people that wearing a mask could increase the activation of coronavirus. Not only this there exists much misinformation related to deaths caused by coronavirus as well as the symptoms, the number of people getting infected, and vaccine availability and its effectiveness. Times Now and Aaj Tak released a film depicting a burial place in Xinjiang for Chinese troops who perished during the 1962 Indo-China conflict, claiming that it showed the grave of soldiers killed in the June 20 fight. When the Nation was going through the demonetization, there appeared a piece of false news saying that the notes of five hundred and two thousand are containing some kind of chip in them.

These illustrations demonstrate how misleading news shapes our intellectual, religious, political, and other views and relationships. These misleading information are poisonous for any human kind and should be eliminated or atleast reviewed and reduced at early stages to save the harm that they could possibly cause worldwide. The main question is why false information originates, why they're so likely to occur, and also why the vast

majority of the people are victims of these frauds. The answer of these questions are very easy to understand i.e. if someone wants the media spotlight or the attention of people around the world in order to trend over internet then prevailing such information fulfill their intentions. That indicates false news provides you fame, whether it is positive or negative reputation, which people must realize. Captivating attention may be carried out for a number of reasons, notably giving false information or deceiving, or simply to make things interesting for chatting. Whereas digitalization has returned with an increased volume of bogus news and information. The primary worry with fake news is that it retains the capacity to divide people and their viewpoints, which leads to damaged relationships both inside and outside of a country. Other negative outcomes of bogus news include economic losses, ruined social interactions, mocking public suffering, and, in rare cases, cybercrime.

### **1.3 OBJECTIVE**

This research includes a thorough examination of fake substance as well as multiple classification methods for distinguishing news assertions before we are influenced or propagate the deception. Our projected strategy can achieve ideal engagement of the situation created by false news as it arises presently. While being in the educational arena, or general knowledge to be a little more precise, everything is being contaminated by bogus stuff, therefore we need to be more worried about the problem that is growing with time. Bringing forth the enhancement skills of fake article categorization by text mining researchers using natural language processing on the urgent requirement for performance improvement. Natural language processing and classification techniques based on machine learning are employed in this study in order to detect the misinformation.

### **1.4 THESIS OUTLINE**

The thesis is well organized and divided into 5 chapters. The synopsis of the thesis is mentioned below precisely.

Chapter 1 encloses a brief description of ML and NLP, a basic introduction about fake news, the aims of the current study, and the motivation of the present work.

Chapter 2 presents the recent developments related to the work done by various eminent scientists around the globe. It focuses on the detection of false news using various

machine learning techniques and their outcomes.

Chapter 3 contains the key theoretical concepts used in this study which includes the basic definitions of concepts we will be encountering.

In chapter 4, we have given the methodology adopted throughout this study and the key steps followed in order to carry out the research. This includes dataset collection, preprocessing techniques, various classification techniques used along with the model evaluation approach and definitions.

In chapter 5, the experimental results have been shown which contains the graphs obtained after data exploration and results obtained after applying all five supervised ML classifiers.

Chapter 6 envelopes the current research study's conclusion and future scope.

## CHAPTER 2

### LITERATURE REVIEW

A significant amount of prior and ongoing studies is based on fake news detection. The misleading information has always been a serious concern worldwide due to its bad influence on social, religious, educational and civilization and many more fields. Inside this chapter, we provide a concise review of the previous work as well as the various machine learning approaches employed in the research. We studied research papers in order to carry out this work and filtered below papers for extensive literature review and summarized few of them in this section. The proposed model in this study has been drawn from various research being done and the algorithms that have been applied in research papers to propose a better way for detection of fake news.

Ahmed *et al.* [1] identified false news, by developing an AI model using n-gram analysis. They have used five machine learning classification methods i.e., K-Nearest Neighbour, SVM, Stochastic Gradient Descent, K-Nearest Neighbour, Decision Trees, and Linear Support Vector Machines. And found that Linear Support Vector Machine(LSVM) outperforms others with the precision of 92 percent.

Granik and Mesyura [2] presented a methodology for detecting bogus news utilizing naive bays on news posts on Facebook and got a 74 percent accuracy rate. And concluded that AI techniques could be used successfully to handle these kinds of problems.

Wang *et al.* [3] in his publication, says that the epidemic of misleading information not only produces distrust in the mainstream press but also volatility in the realm of politics. Fake news has an impact on folk's political choices throughout campaigns. False information was popular and disseminated swiftly in the walk to the 2016 U.s. Election, according to experts at the Oxford Internet Institute.

Zubiaga *et al.* [4] proposed sequential classifiers to classify rumour positions on social media platforms. They leverage Twitter as primary social media application, categorizing tweets into four categories. On eight data sets, they employed four sequential classifiers namely hawkes algorithms, along with linear CRF, long short term memory (LSTM), and tree CRF. They discovered that sequential classifiers that exploit the reciting property



outperformed non sequential classifiers in social media engagement, and that LSTM outperforms other sequential classifiers.

Campan, Cuzzocrea and Truta [5] in their study proposed a model how fake news spread on social media and how the internet affects the diffusion of false information in creating and spreading. They also discussed the solutions to reduce the dissemination of false information and provided the future research aspects in this area.

Kotteti, Na Li, and Qian [6] worked on using data imputation to improve the discovery of false information in their study. A unique data preparation strategy were applied by them in order to fill the missing value in the proposed dataset for enhancing the performance. For quantitative and hierarchical variables they used data modeling for missing values by choosing most frequent values for columns. To make up for the missing values, three things were performed: 1. Eliminate columns with missing values, 2. Replaced missing values with blank text, and 3. applied missing values using data imputation techniques. Finally they discovered and concluded that Multi Layer Perceptron(MLP) classes enhanced the accuracy by 16 percent.

Aphiwongsophon and Chongstitvatana [7] have a study on using naïve Bayes, SVM, and neural networks to detect the fake news and calculated the performance measures they have found that naïve Bayes has 96.08% and neural network and SVM 99.90% accuracy. Through this experiment, they found out that neural networks and support vector machines are having significant accuracy and high confidence.

Akshay and Amey [8] worked on detecting false news, and they suggested a strategy that we may use on Facebook. For predicting, he employed Naive Bayes. They utilised a dataset with 11000 articles sorted by categories (index, text, title and label). The dataset comprises news on science and industry in addition to politics. They used the title and content as their major source for implementation, as well as some n-gram references. And finally they discovered that naive bayes had accuracy of 93.10 percent and presented the way to improve this.

Ajao *et al.* [9] suggested a methodology which utilizes a mixture of (CNN) and (RNN) models to detect phony news tweets from Twitter posts. They included rumour stories namely Hebdo Charli, Ferguson and Ottawa Shooting, Sydney Siege, Germanwing Crash. Their suggested work on a hybridization of CNN-RNN intuitively recognises crucial features connected to misleading news articles without any foreknowledge of the news and achieves an accuracy of over 80%.

Riece *et al.* [10] are working on looking for a range of elements in news articles, postings,

and stories that might assist identify false news with increased precision. He demonstrated the significance of these new qualities in evaluating bogus news. Discrimination, integrity, involvement, domain location, and temporal patterns are some of these characteristics. They used 2282 Buzzfeed items in their analysis (news articles). Using KNN, Nave Bayes, Random Forest, XGBoost, SVM, and they analysed and described the strengths and limits of this technology, and discovered that XGBoost performed better when compared with other with an accuracy of 0.86. they employed many customized features in order to build this system such as for 31 total features they used POS tagging, bag of words. They divided the features into lexical features, language features, semantic, psychological and engagement features.

Perez Rosas *et al.* [11] suggested a model that automatically detects fake news for online resources. They developed a computer algorithm and tools to detect bogus news. They work with two different datasets. The first came via the internet, the second resulted from a mix of human data collecting and internet assistance. Their model is almost same as Reice in [10] but they have used SVM as a model and worked on a different dataset here. Tacchini *et al.* [12] produced a dataset using two different groups having news from conspiracy and science. To enhance the accuracy of their model they have leveraged the social component. They used logistic regression and harmonic method to classify news into false and reliable content. Harmonic algorithm exchanges the info between individuals who liked similar context.

Yuan, Q. Ma, and W. Zhou [13] suggested a model to identify false news. They proposed a structure-aware multi-head attention network (SMAN) based technique in their model. This strategy is based on the publications' and users' trustworthiness. Real-world datasets were employed in this technique. They tested this model against three distinct datasets and discovered that it had a high level of accuracy.

Ozbay and Alatas [14] have used AI techniques for detecting fake news. In the first phase, they preprocessed the dataset to transform unstructured data into structured data, and then they used text mining to construct about twenty- three supervised AI algorithms. They applied these algorithms to about three real-world data sets and found the accuracy and performance measures accordingly. The best average value they got was by using a decision tree, ZeroR, CVPS, and WIHW algorithms.

Ankit, Sudakar, and Anil [15] demonstrated a basic strategy for detecting false news on social media using a K- nearest neighbor classifier, which obtained an accuracy of roughly 79 percent when evaluated against a sample of Facebook news articles.

Nagaraja *et al.* [16] showed in their study that false information mostly circulates through social media and is propagated further without investigating the true data. They applied various NLP techniques and two ML algorithms i.e., naïve Bayes and SVM which gives 63% and 75% accuracy respectively.

Shifath, S. Islam, and F. Khan [17] have suggested a strategy that is transformer-based for identifying COVID-19 false information. They experimented using CNN and conventional lexicons. The dataset consists of COVID-19-related social media postings with classifications indicating whether they are false or legitimate. They also explored other hyper settings and explored various transformer-based models. RoBERTa shows that the maximum accuracy is 0.979.

# CHAPTER 3

## THEORETICAL CONCEPTS

This section introduces the fundamental theoretical ideas needed to comprehend the essential processes and operations of the study in consideration under this project. This section includes information of the concepts about data preprocessing, feature extraction, and the supervised machine learning classifiers and would help us to understand the proposed working for the fake news detection project.

We can assign any ML problem to one of the two broad classification as described following:

### 3.1 SUPERVISED MACHINE LEARNING

The ML model receives data input and accurate output data using supervised learning and seeks to discover a mappings that will map an input parameter to its output value [18]. We can say that. In this type of learning machines learn using labeled data and on the basis of that predicts the correct output. Linear Regression, Regression Trees, Support Vector Machine, Logistic Regression, Random Forest, and Nave Bayes and Decision Tree etc. all the algorithms fall under the supervised learning. Figure 3.1 depicts the working of Supervised ML algorithms.

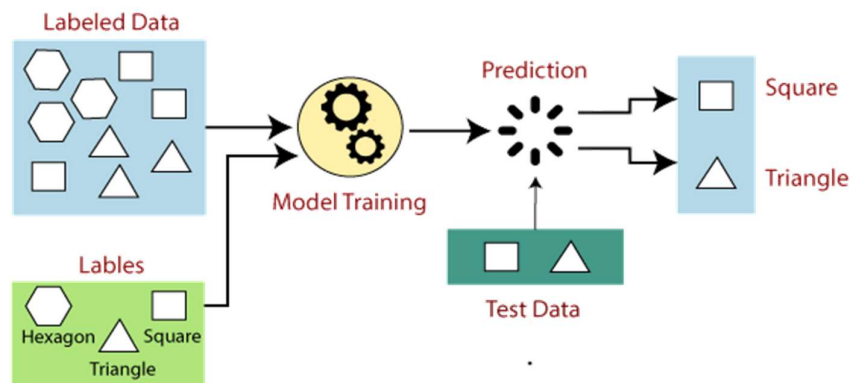


Figure 3.1. Supervised Machine Learning[27]

### 3.2 UNSUPERVISED MACHINE LEARNING

Here ML models are not made to learn using labelled data. In unsupervised learning models are trained using unlabeled datasets and then expected to function on them without guidance [18]. K-Means clustering, Hierarchical clustering, Apriori Algorithm all fall under the unsupervised learning. Figure 3.2. depicts the working of unsupervised ML algorithms.

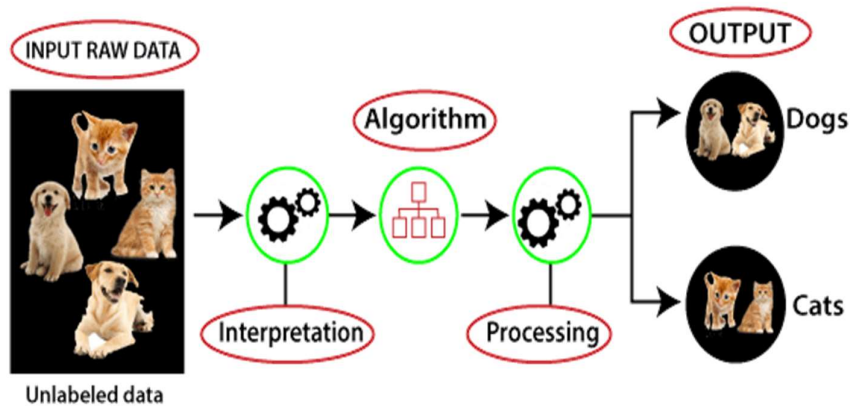


Figure 3.2. Unsupervised Machine Learning[28]

### 3.3 FEATURES

A feature is a quantifiable attribute of the object under consideration. Features exist as columns in datasets. Many characteristics characterize data items, each of which conveys the core of the entity under study. Variables, attributes, and characteristics are other terms for features.

### 3.4 TEXT PREPROCESSING

NLP or natural language text processing is a machine learning subfield that works handling textual data. Text preprocessing is the process of preparing text data for modelling. Every NLP project begins with this step. Aside from numeric values, text data is widely accessible and is utilized to assess and recommend solutions. However, before using the data for analysis or prediction, it must be processed.

### **3.5 FEATURE EXTRACTION**

The underlying problem when working with text comprehension is that the raw textual content can not be directly fed into ML algorithms/model. ML approaches provide outputs by learning from a pre-defined set of characteristics in the training examples for testing data. As a result, feature extraction algorithms are required to turn text into a matrix (or vector) of features.

### **3.6 PYTHON LIBRARIES**

Python has many different and easily available libraries. Using the python libraries we can just pick the set of functions and use it without having to start the code from scratch [19]. We have used various libraries in our experiment such as NLTK, Pandas, Numpy, Matplotlib and NLTK(Natural Language Tool Kit) that is used to work with NLP here we have used this library to deal with stop words and their removal. Pandas and Numpy are the key packages to load the data. Pandas was used to do database operations and data analysis. For computing, Numpy is utilised. For visualisation, Matplotlib was utilised. All of the algorithms' libraries are imported from Scikit Learn.

# CHAPTER 4

## RESEARCH METHODOLOGY

### 4.1 PROPOSED MODEL

Here we have given the model that would be followed in this project. This section introduces the model's architecture.

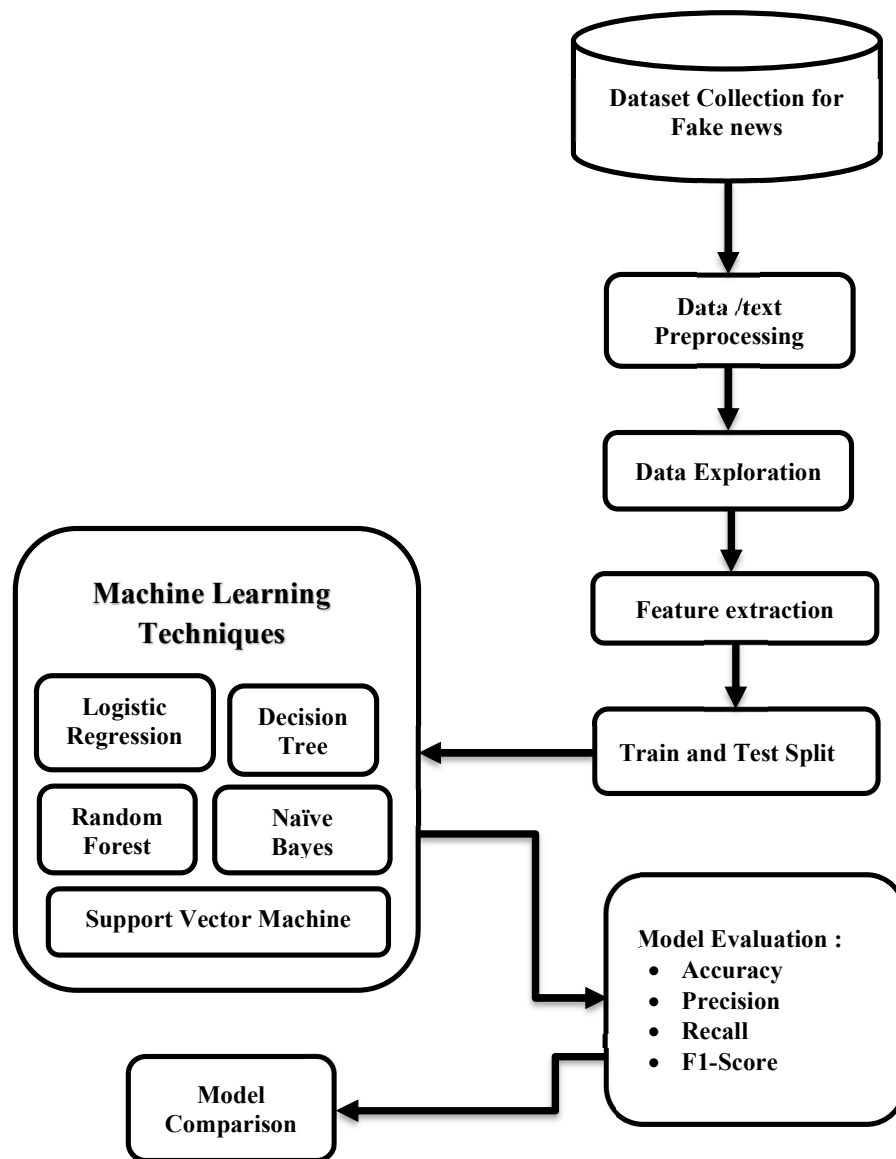


Figure 4.1. Proposed Model

In our Proposed Model the implementation has been carried out in following different steps:

1. In the first step the dataset collection is being performed.
2. Next on the collected data we would be performing text preprocessing such as removal of stop words, punctuation removal, lower casing and so on.
3. To convert the text into matrix we would perform feature extraction.
4. After collecting the data we have done data analysis to visualize and understand the pattern in data.
5. Then we would train the data to the machine learning model. Here we have used five ML classification techniques as given in Figure 4.1.
6. At the end we have to check how each model is performing so in order to assess a model we would calculate various evaluation measures to know the performance of each model.

Further we are giving stepwise analysis of all the processing and techniques performed in this experiment.

## 4.2 DATA COLLECTION

We obtained the dataset for this study from the Kaggle website [26] which contains two files. Out of these two one contains real news articles and another one contains fake news articles. Real articles are around 21417 and fake articles are around 23481 with a total of 23481. To further proceed with both the data set we had combined the data frames that contains the combination of both fake as well as real. We have also shuffled the data in order to prevent bias. Figure 4.2 shows the first five rows of our data:

	title	text	subject	target
0	Trump Thinks Rush Limbaugh Is Real News And T...	Rush Limbaugh praised Donald Trump s insanely ...	News	fake
1	Romney: It's 'disqualifying' for Trump not to ...	WASHINGTON (Reuters) - Former U.S. Republican ...	politicsNews	true
2	Irish government says significant progress on ...	DUBLIN (Reuters) - Ireland has made signfica...	worldnews	true
3	Former RNC Chair Michael Steele Refuses To Vo...	Donald Trump has dragged the Republican Party ...	News	fake
4	New Zealand foreign minister praises China tie...	WELLINGTON (Reuters) - New Zealand will develo...	worldnews	true

Figure 4.2. Head of the data



### **4.3 DATA PREPROCESSING**

Textual data is amongst highly unorganized sorts of data. It is quite difficult to deal with human language. Cleansing or preprocessing the data is as critical as building the model in any Machine Learning activity.

Textual data contains clutter in several forms, including sentiments, punctuation, and text in multiple cases [30]. Text preprocessing is a method of cleaning up text data before it is fed into a model. As the data we have used here is available in text as we know that text data requires preprocessing in order to be changed into a suitable structure for information display. In data preprocessing we have performed following steps:

#### **4.3.1 REMOVING UNNECESSARY COLUMNS**

The dataset we are using is having five features i.e. title, text, subject, date and target. So we dropped the unnecessary columns i.e. date and title as we will be working with only text here.

#### **4.3.2 LOWERCASE CONVERSION**

Since the machine treats lowercase and uppercase in different manner, it is easy for a machine to interpret a textual data provided in the same case either lower or upper. Hence we have made our text data in the lower case.

#### **4.3.3 PUNCTUATIONS REMOVAL**

Punctuation like comma(,), full stop(.), colon(:), exclamation(!) and so upto 32 punctuations are there. These punctuations don't add much importance to text so it is removed from the text data.

#### **4.3.4 STOPWORD REMOVAL**

There are many words in a text that occur very frequently in a document and have no much information. Such as 'a', 'is', 'the', and 'am'. To improve the accuracy of our analysis these words are ignored using NLTK(natural language toolkit) library for stop word removal.

#### 4.4 TRAIN AND TEST SPLIT

We split the data into train and test before feeding it into the machine learning model. We have separated the 30% data into test set and 70% into train set.

A subset of the dataset used to train the model has already revealed the outcome.

The detection model is tested on a subset of the dataset, and the test set is utilised to forecast the outcome.

#### 4.5 FEATURE EXTRACTION

We have used two feature extraction techniques here. First we vectorized the corpus using Count Vectorizer then applied Tf-idf. Below we have given a brief overview of both the techniques:

##### 4.5.1 COUNT VECTORIZER

It constructs a matrix, known as document term matrix which is a collection of dummy variables that represent whether or not there is no specific term in the document.. Count vectorizer is known to be one of the simplest techniques for vectorization. The count vectorizer would aim to generate a document term matrix where the individual cells will show the frequency of that word in a specific document and the columns would represent each word in the corpus. We are giving an example of how count vectorizer works:

**Document 1:** This is my thesis work. It contains information.

**Document 2:** The information is relevant.

Now if we create a list of tokens of unique words present in the above corpus:

**Unique words:** ['thesis', 'work', 'contains', 'information', 'relevant']

we can convert it into a matrix of  $D \times N$  where  $D=2$ (number of documents) and  $N=5$ (Number of Unique words):

Table 4.1.  $D \times N$  Matrix

	thesis	work	contains	information	relevant
D1	1	1	1	1	0
D2	0	0	0	1	1

As we can see from Table 4.1 that there are 5 unique words from the corpus(after stopword removal and converting to lowercase) represented in columns and 2 documents/text samples represented in rows in the above example. Count vectorizer would not process these words stored in string so it would be converting this to index value such as thesis at index 0, work at index 1, contains at index 2 and so on. Here we will represent the actual Table 4.2 as given below:

Table 4.2. Count Vectorizer Matrix Representation

	0	1	2	3	4
0	1	1	1	1	0
1	0	0	0	1	1

#### 4.5.2 TF-IDF

After vectorization we have applied tf-idf also in our corpus. (Term frequency inverse document frequency) basically gives an idea of how frequently a particular word appears in a corpus. It determines the significance of a term. [20].

Here are two terms separately:

TF: Term Frequency

IDF: Inverse Document Frequency

Let's understand these two terms one by one:

##### 4.5.2.1 TERM FREQUENCY

It may be described as a word's frequency or a ratio of the number of times a word occurs in a document to the overall number of words in the document. It can be expressed as follows:

$$TF(term) = \frac{\text{term 't' frequency occurring in document 'd'}}{\text{Total number of term 't' in a document 'd'}} \quad (4.1)$$

Where TF is term frequency of the particular term

For example:

**Document :** He loves to play cricket

If we calculate  $tf(\text{play}) = 1/5$

#### 4.5.2.2 INVERSE DOCUMENT FREQUENCY

It computes the importance of a term in the corpus. We can know if a certain term is rare or common in given corpus and accordingly make the decision of keeping that term in rare occurrences. It basically computes of how much a word is common or uncommon among all documents present in the corpus. It can be formulated as:

$$IDF(term) = \log\left(\frac{\text{Total number of documents 'd'}}{\text{Number of documents 'd' which has the term in it}}\right) \quad (4.2)$$

Where IDF is inverse document frequency of term 't'

So we can give the formula for computing TF-IDF in a corpus as below:

$$TFIDF(term) = TF(term) * IDF(term) \quad (4.3)$$

The product of a word's TF and IDF reflects how frequently the word appears in the document and how unique the token is over the whole corpus of documents.

### 4.6 MACHINE LEARNING SUPERVISED CLASSIFICATION TECHNIQUES

In this study, we used 5 supervised ML classification method to identify false news: LR, NB, SVM, DT, and RF, and determined which approach performed best based on accuracy. The description algorithms used are given as follows:

#### 4.6.1 LOGISTIC REGRESSION

It's a tool for categorizing binary data. For binary classification usually, Linear regression is used to create the best bit line. When two classes can be separated linearly, logistic regression is used. It is within the supervised machine learning algorithm category. It's a machine-learning-based categorization problem-solving approach. In logistic regression, a type of predictive analysis, the probability assumptions are applied. To complete a binary classification job, a linear equation is used as input, and the logistic function and log odds are used in the logistic regression model. It employs a more complicated function when compared to linear regression.

$$y = d0 + d1 * x \quad (4.4)$$

$$P = 1/(1 + e^{-y}) \quad (4.5)$$

$$\ln(P/(1 - P)) = d_0 + d_1 * x \quad (4.6)$$

Where In equation(4.6)  $d_0$  is the slope,  $d_1$  is intercept and  $x$  is a data point.

Equation(2) is a sigmoid function where  $P$  has been used to eliminate the outlier's effect.

#### 4.6.2 NAÏVE BAYES

The Naive Bayes approach, a supervised machine learning methodology based on the well-known Bayes theorem, is used to tackle classification issues. It's most commonly used for text classification with a big training dataset. One of the most simple and effective classification methods is the Naive Bayes Classifier. It allows for the rapid building of machine learning models as well as effective training and testing to make speedy predictions [23]. It's a probabilistic classifier, which implies the algorithm's whole basis is built on probabilities that have been computed, and it predicts based on an item's likelihood.

Naïve Bayes has three types of models:

1. Gaussian NB which assumes that a feature follows a normal distribution and is applied for classification.
2. Multinomial NB is a discrete count algorithm that is commonly used in text categorization problems. With the help of this we can know what are the possible outcomes of a particular word in document. Multinomial Nave Bayes was used in our experiment.
3. Bernoulli NB which is used in a binomial model where feature vectors are binary.

Naïve Bayes Equation:

$$P(G|H) = P(H|G) * P(G) / P(H) \quad (4.7)$$

Where:

G, H are events

$P(G|H)$  is the posterior Probability i.e. probability of occurring event G given H is true.

$P(H|G)$  is the Likelihood or the probability of event H given G is true.

$P(G)$  is the Class Prior Probability i.e. the independent probability of G

$P(H)$  is the Predictor of Prior Probability i.e. the independent probability of G

### 4.6.3 SUPPORT VECTOR MACHINE

The Support Vector Machine approach aims to find a hyperplane (where  $N$  is the number of characteristics) that clearly arranges the principal elements in an  $N$ -dimensional space. There is an assortment of hyperplanes from which to isolate the two kinds of informative items. Our point is to track down the plane with the biggest edge, or distance between relevant items from the two classes [21].

Boosting the edge distance gives some support, making it more straightforward to arrange ensuing information points. There are two types of SVM:

1. Linear Support Vector Machine which is used when a data can be classified in just two classes and data is linearly separable. In our project we are implementing linear SVM as a classifier because we have two classes one is fake and other one is true.
2. Nonlinear SupportVectorMachine is used when a straight line can't divide data points means data is non linearly separable and it can not be classified in just two classes.

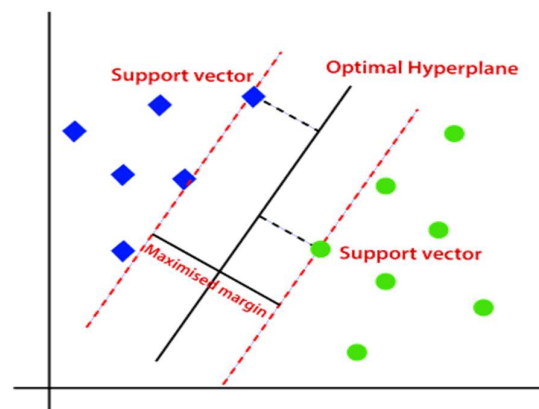


Figure 4.3. Support Vector Machine[29]

### 4.6.4 DECISION TREE

It is a Machine Learning supervised classification algorithm which means we have to clarify what the information is and what the relating yield is in the preparation information. It is a tree-like construction where the information is consistently parted by a specific boundary[22]. The elements of a dataset are addressed by the inner nodes and branches that address the decision rules and each leaf address the ultimate results or

choices. Figure 4.4 depicts the working of decision tree.

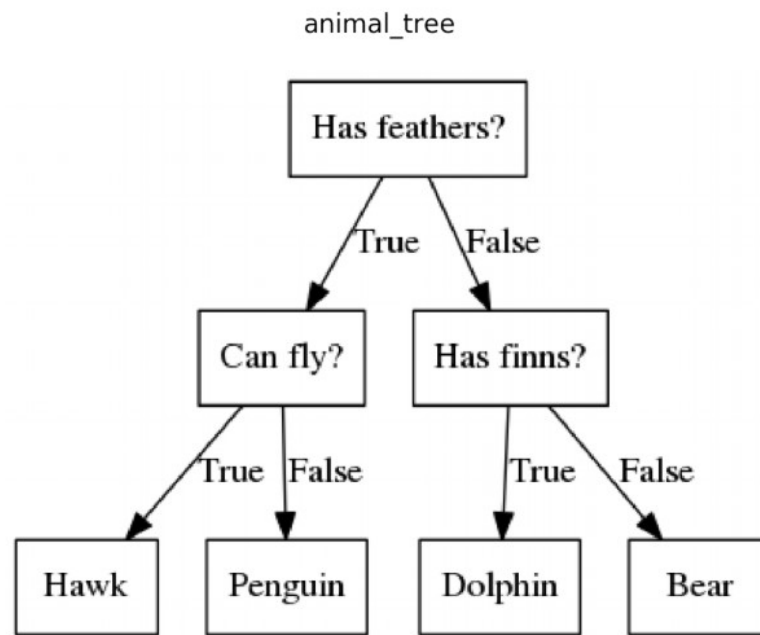


Figure 4.4. Decision Tree[31]

In the given figure the decision tree has root node based on which it is dividing the data into subsets and further after dividing it we have decision nodes and the terminal nodes or leaf nodes. After leaf nodes the tree can not be exceeded. In given example the tree is being branched on the attribute if the animal is having feathers or not. If yes then we check if it can fly or can not. If the answer of Can Fly? Is yes then it would be hawk otherwise penguin. if it has no feathers then we check if it has finns or not if the answer for finns is yes then we make the conclusion that it is dolphin otherwise we conclude that it is bear.

#### 4.6.5 RANDOM FOREST

Random forest is a supervised ML technique. It is basically established on the outfit learning techniques where different classifiers are united to deal with an issue and to chip away at the display of the presentation of the model. Random Forest is a classifier that calculates the dataset's predicted precision by averaging the results of many decision trees applied to different subsets of the dataset. As given in Figure 4.3 we can say that Random forest is combines two or more decision trees and gives the results.

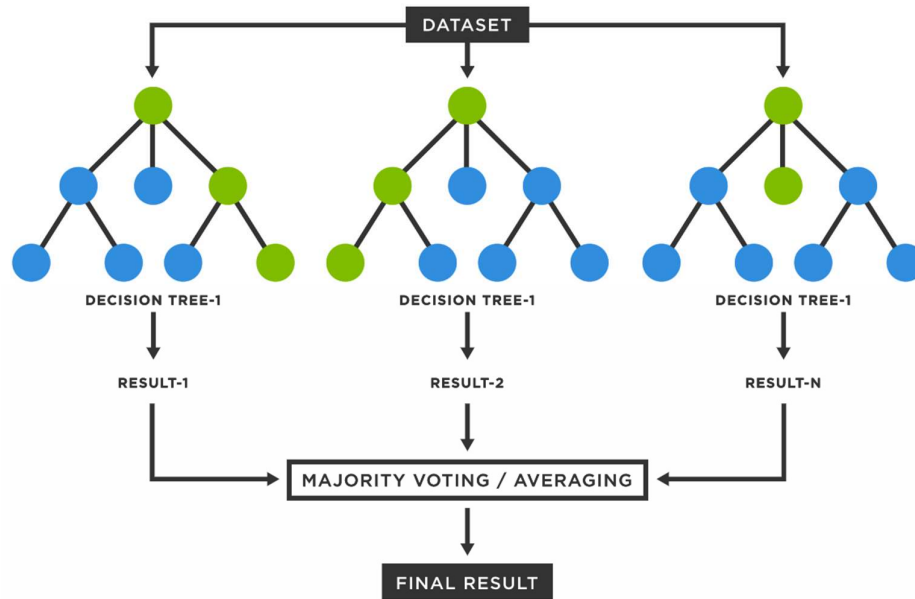


Figure 4.5. Random Forest [32]

#### 4.7 MODEL EVALUATION

Evaluation metrics are frequently used to assess categorization performance. As a result, the performance measurements are the most common. We utilized the following metrics from the confusion matrix, which is depicted in the image below, to analyze the model's performance:

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	TP (True Positive)	FP (False Positive)
	Negative (0)	FN (False Negative)	TN (True Negative)

Figure 4.6. Confusion Matrix

The four terms in the confusion Matrix can be stated as follows from the Figure 4.4: **TP** (True Positives) belongs to a class it actually belongs. For example, in our experiment, TP represents that the news which was actually fake is also predicted as fake. Where fake is class1 and true is class 0.



**FP** (False Positives) belongs to a class it actually doesn't belong. For example, in our experiment, FP represents that the news which was actually true is predicted as fake. Where fake is class1 and true is class 0.

**FN** (False Negative) doesn't belong to a class it actually should belong. For example, in our experiment, FN represents that the news which was actually fake but is predicted as true. Where fake is class1 and true is class 0.

**TN** (True Negatives) doesn't belong to a class it actually doesn't belong. For example, in our experiment, TN represents that the news which was actually true is also predicted as true. Where fake is class1 and true is class 0.

#### **4.7.1 ACCURACY**

The accuracy metric examines the trained model using test samples as input. The number of samples properly identified out of the total number of samples is known as accuracy. Mathematically, the following equation is used to express accuracy.

$$Accuracy = (TN + TP)/(TP + FN + FP + TN) \quad (4.7)$$

#### **4.7.2 PRECISION**

Precision is a proportion to tell how exact a classifier is performing. The number of false news samples projected as false news samples is the precision Precision P can be formulated as the ratio of total true positives to total predicted instances:

$$P = TP/(TP + FP) \quad (4.8)$$

#### **4.7.3 RECALL**

The Recall metric is the number of true fake news samples accurately predicted out of all the true fake news samples. It tells what percentage of positive instances were successfully identified and it's formula is:

$$R = TP/(TP + FN) \quad (4.9)$$

**4.7.4 F-Measure.** It is represented as a harmonic mean of Precision and recall and can be formulated as:

$$F = 2P.R/(P + R) \quad (4.10)$$

# CHAPTER 5

## EXPERIMENTAL RESULTS

In this section, we have given an analysis of the results we obtained as we performed the proposed experiment. Various graphs in data exploration step are obtained and helped in understanding the patterns in given textual data. We have implemented all five supervised machine learning classification algorithms and calculated the accuracy and other performance measures.

### 5.1 DATA EXPLORATION

In data exploratory analysis we try to visualize our data and find the meaningful insights from the data. Under data exploration, we are giving the various visualizations that we obtained while exploring our dataset.

#### 5.1.1 NUMBER OF FAKE AND TRUE ARTICLES IN DATA

Fake news articles in our data set are 23481 and true news articles are 21417 as represented in the below Figure 5.1:

```
target
fake    23481
true    21417
Name: target, dtype: int64
```

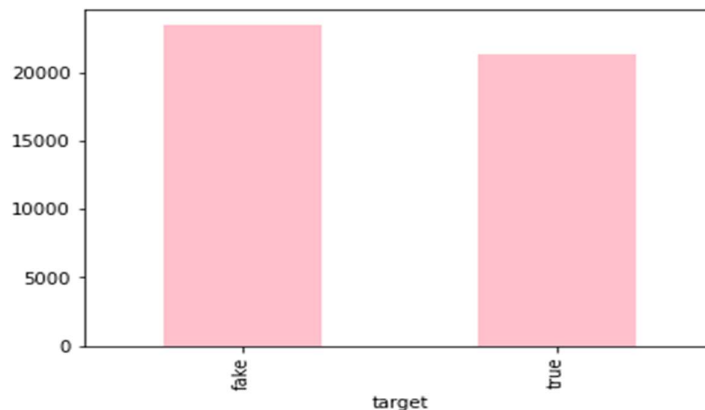


Figure 5.1. Number of Fake and True Articles

### 5.1.2 ARTICLES PER SUBJECT

As the data we are using here is in the text and having the subject as a column which contains various subjects being included in news articles so we tried to get the number of articles each subject is having. The following Figure 5.2 shows the number of news articles per subject where the main subjects are: GovernmentNews, Middle-east, News, US\_News, left-news, politics, politicsNewsz, and wordnews.

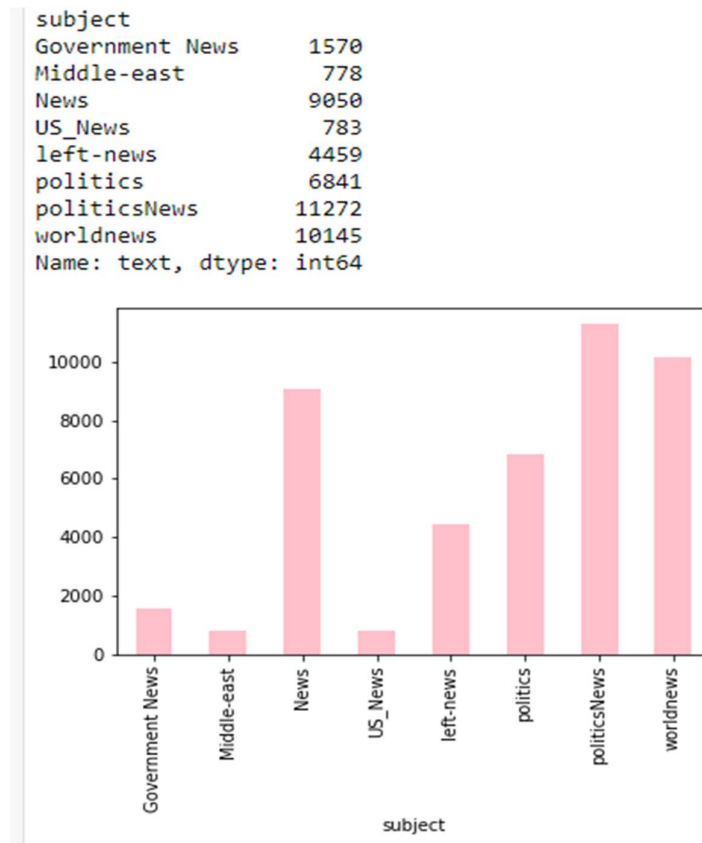


Figure 5.2. Articles per Subject

As we can see in the above Figure 5.2 the subject politicsNews contains the maximum number of articles and worldnews having the minimum number of articles.

### 5.1.3 WORDS APPEARING FREQUENTLY IN FAKE AND TRUE NEWS ARTICLES

As there are two target variables in our dataset i.e. fake and true so we tried analyzing which words are occurring frequently in both the target variables. The count of words

that occurs most frequently in false information articles are given in below Figure 5.3:

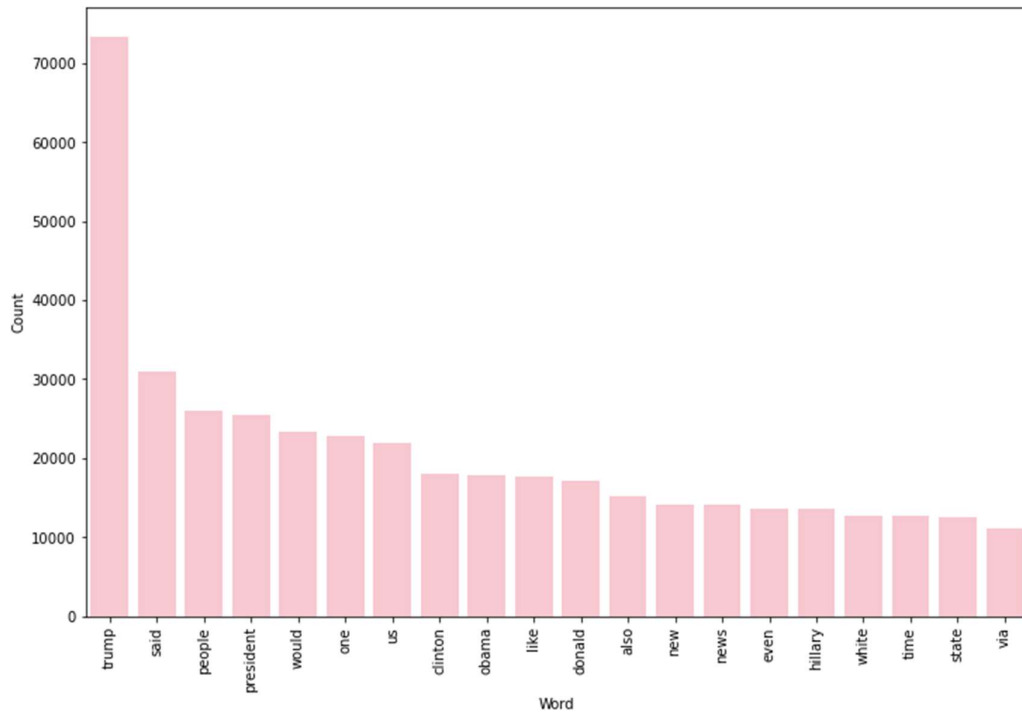


Figure 5.3. Frequent Word Count in Fake News Articles

From the above figure, we can see that the word ‘trump’ appears most frequently and ‘via’ appears least frequently in the fake news articles.

The count of words that occurs most frequently in real news articles are given in below Figure 5.4:

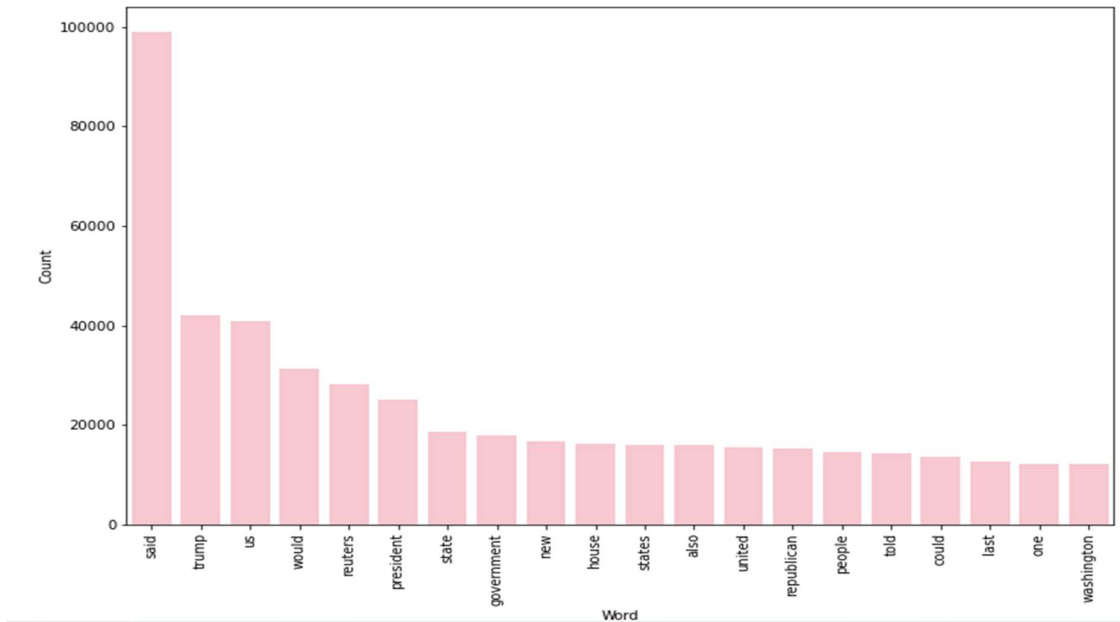


Figure 5.4. Frequent Word Count in True News Articles

From the above figure we can see that the word ‘said’ appears most frequently and ‘washington’ appears least frequently in the real news articles.

### 5.1.4 WORDCLOUD FOR FAKE AND TRUE NEWS

Wordcloud is an effective way to represent and visualise the text processing that depicts the most frequently words appearing in larger and bolder characters and in different colours as well. The word that is the smaller in size would be having the lesser importance. In Figure 5.5 and Figure 5.6 we are highlighting the words that are occurring in fake and true news respectively. These can clearly give an idea just by seeing at the wordcloud which words are appearing frequently in both the articles.

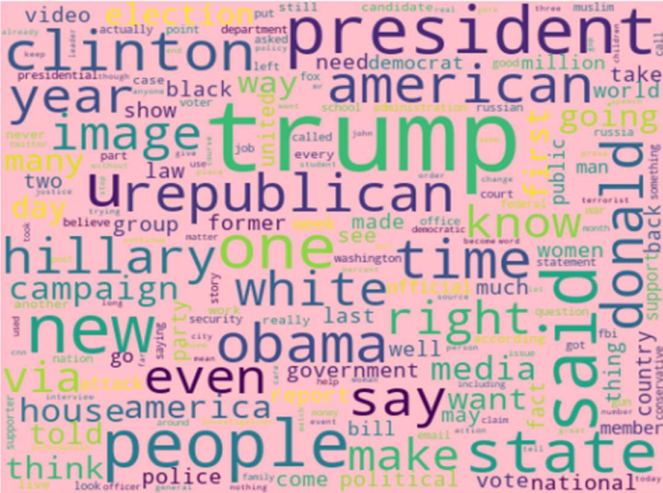


Figure 5.5. WordCloud for Fake News Articles

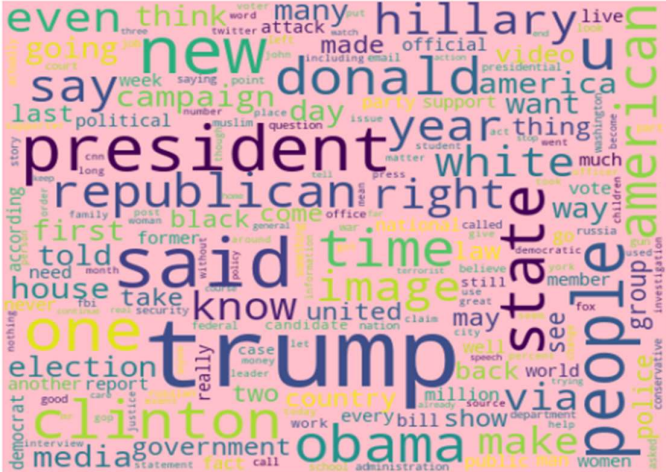


Figure 5.6. WordCloud for True News Articles

## 5.2 CONFUSION MATRIX OBTAINED FROM ML CLASSIFIERS

Here we are giving the various confusion matrices that we obtained after applying five supervised machine learning algorithms i.e. Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, and Random Forest. As discussed in methodology each confusion matrix contains the four values TP, FP, TN and FN. In our experiment, TP represents that the news which was actually fake is also predicted as fake. FP represents that the news which was actually real is predicted as fake. FN represents that the news which was actually fake but is predicted as real. TN represents that the news which was actually real is also predicted as real. The real and fake true label and predicted label are shown in confusion matrix.

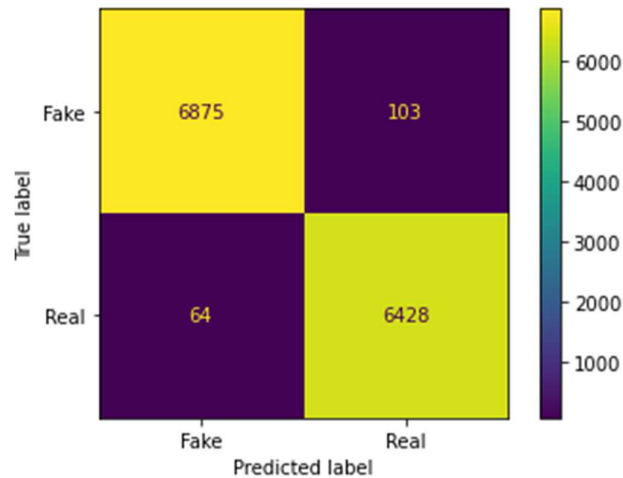


Figure 5.7. CM using Logistic Regression

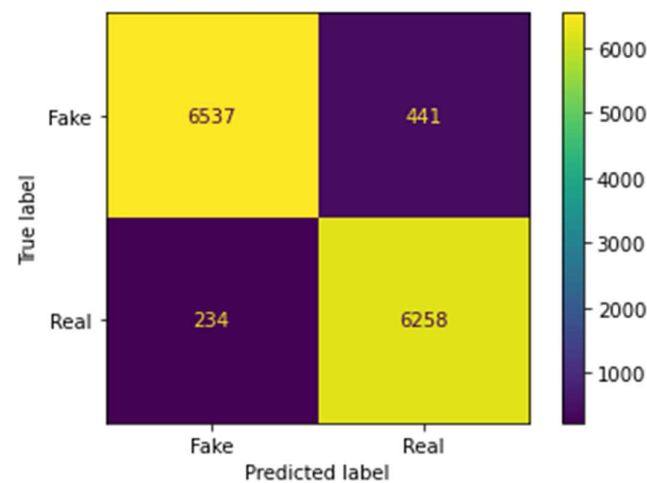


Figure 5.8. CM using Naïve Bayes

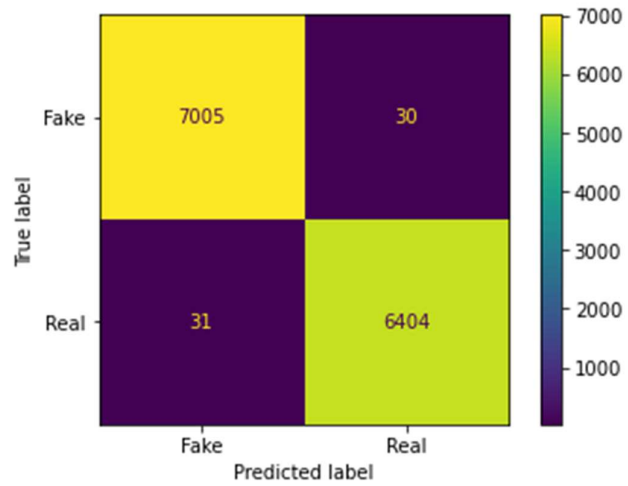


Figure 5.9. CM using Support Vector Machine

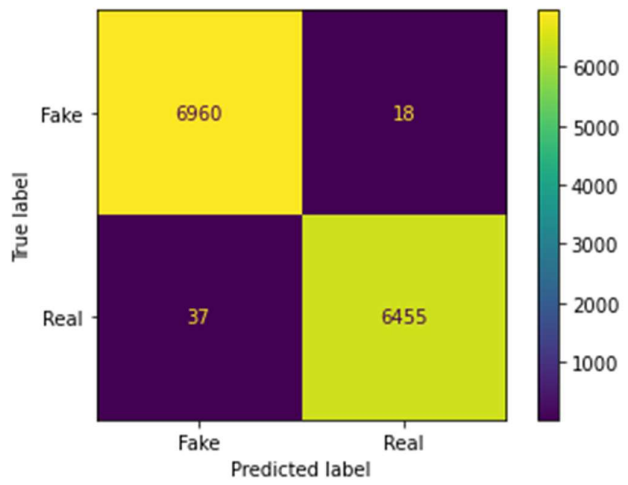


Figure 5.10. CM using Decision Tree

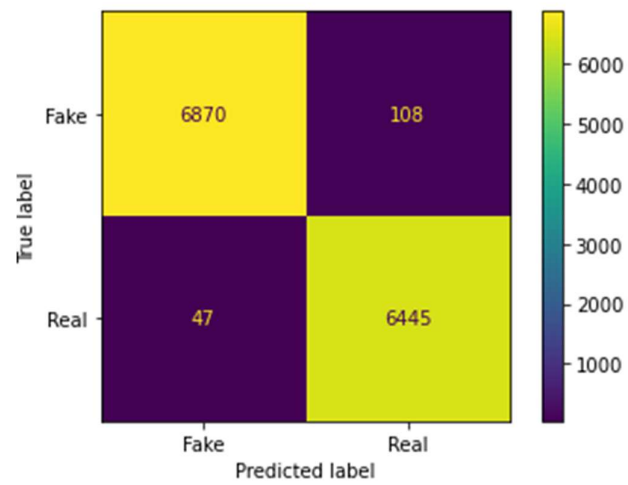


Figure 5.11. CM using Random Forest

The CM in Figure 5.7, Figure 5.8, Figure 5.9, Figure 5.10, Figure 5.11 shows the TP, FP, TN and FN values of each classifiers. Table 5.1 shows all the values obtained from confusion matrix when we applied all ML classifiers.

Table 5.1 TP, TN, FP and FN values for Classifiers

<b>Classifier</b>	<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
LR	6875	6428	64	103
NB	6537	6258	234	441
SVM	7005	6404	31	30
DT	6960	6455	37	18
RF	6870	6445	47	108

### 5.3 ACCURACY AND PERFORMANCE MEASURES

We calculated the accuracy and the various performance measures. The following Table 5.2 represents obtained the accuracy and the performance measures (precision, recall, F1-score) in this experiment:

Table 5.2 Accuracy and Performance Measures

<b>Classifier</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-Score</b>
Logistic Regression	98.76 %	0.99	0.99	0.99
Naïve Bayes(MNB)	94.99 %	0.95	0.95	0.95
Support Vector Machine(LSVM)	99.58 %	1.00	1.00	1.00
Decision Tree	99.59 %	1.00	1.00	1.00
Random Forest	98.85 %	0.99	0.99	0.99

As described earlier we have implemented all the five supervised ML algorithms i.e. Logistic Regression, Naïve Bayes, SVM, Decision Tree, Random Forest. From Table 5.2 we can see that Decision Tree outperforms here which has the accuracy 99.59% and all performance metrics (precision, recall, and F1-score) as 1.00 perform the best. After decision tree SVM performs with a very negligible difference in accuracy when compared to Decision Tree. Naïve bayes has the lowest accuracy i.e. 94.99% and all performance metrics i.e. (recall, F1-score and precision) as 0.95.



Following graph shows the comparison of all the ML models used in this experiment:

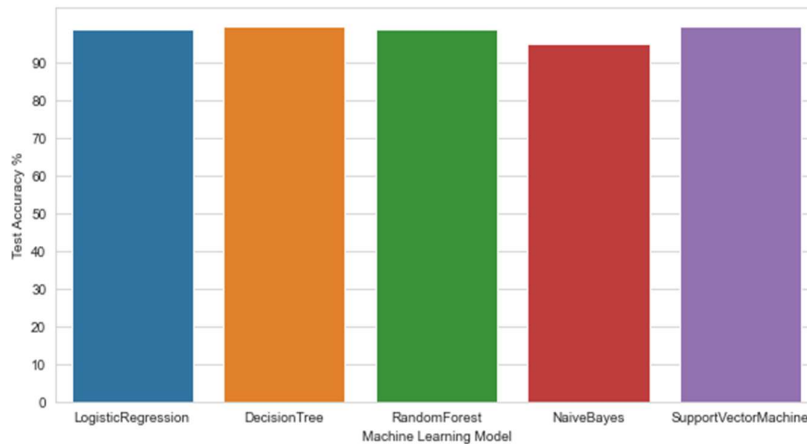


Figure 5.12. Graph for comparison of all the classifiers

The graph in Figure 5.12 which shows that Decision Tree performs the best which has the highest accuracy of 99.59% and the Naïve Bayes classifier has the lowest accuracy of 94.99%. The SVM with 99.58% and very negligible difference in accuracy with DecisionTree i.e. 99.59% is performing at second best.

In our experiment Decision Tree is performing best. As the data is categorical that is the news is either fake or real so in such cases this algorithm performs better as compared to other supervised ML algorithm.

When compared to NB is a generative model while DT is discriminative model. When compared to SVM solves non-linear issues using the kernels method, whereas decision trees handle the problem by deriving hyper-rectangles in input space.

Hence in our study decision tree is performing best with an accuracy of 99.58%

## **CHAPTER 6**

### **CONCLUSION AND FUTURE SCOPE**

Fake news has a huge influence on our social life, as well as in other domains, such as politics and education. Fake news may create significant social and societal harm, as well as have potentially dangerous consequences. It is becoming more difficult for the citizens/consumers to obtain the information that is precise and error free and reliable because of increasing the dimensions of social media. It's critical to discover such false information early on in order to avoid the global harm it can do. As a result, in this paper, we designed a methodology for detecting false news that combines NLP techniques with supervised learning classification algorithm. In this work, we have presented a machine learning approach using various machine learning classifiers to detect fake news. After comparing the performance of each model the conclusion can be drawn that Decision Tree outperforms the other algorithms being used i.e. with the accuracy of 99.59% and secondly Support Vector Machine performs well with the slightest difference in accuracy of 99.58%. This approach would be helpful to identify fake news effectively and with higher accuracy in future.

Future work could include comparing multiple deep learning approaches and new ensemble learning methods to the classification techniques used in this study and determining the best strategy for detecting fake news. Also, we may integrate a larger data set from different sources like various URLs and news publication sites as it would be having bigger journalese and could be used for obtaining better results in a generalized manner.

## REFERENCES

- [1] H. Ahmed, I. Traore, and S. Saad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques," in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, vol. 10618, I. Traore, I. Woungang, and A. Awad, Eds. Cham: Springer International Publishing, 2017, pp. 127–138. doi: 10.1007/978-3-319-69155-8\_9.
- [2] M. Granik and V. Mesyura, "Fake News Detection Using Naive Bayes Classifier," 2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc., pp. 900–903, 2017, doi: 10.1109/UKRCON.2017.8100379.
- [3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective." arXiv, Sep. 02, 2017. Accessed: May 26, 2022. [Online]. Available: <http://arxiv.org/abs/1708.01967>.
- [4] A. Zubiaga et al., "Discourse-aware rumour stance classification in social media using sequential classifiers," *Information Processing & Management*, vol. 54, no. 2, pp. 273–290, Mar. 2018, doi: 10.1016/j.ipm.2017.11.009.
- [5] A. Campan, A. Cuzzocrea, and T. M. Truta, "Fighting fake news spread in online social networks: Actual trends and future research directions," Dec. 2017, pp. 4453–4457. doi: 10.1109/BigData.2017.8258484.
- [6] C. M. M. Kotteti, Xi. Dong, N. Li, and L. Qian, "Fake News Detection Enhancement with Data Imputation," in 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), Athens, Aug. 2018, pp. 187–192. doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00042.
- [7] S. Aphiwongsophon and P. Chongstitvatana, "Detecting Fake News with Machine Learning Method," in 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and

- Information Technology (ECTI-CON), Chiang Rai, Thailand, Jul. 2018, pp. 528–531. doi: 10.1109/ECTICon.2018.8620051.
- [8] A. Jain and A. Kasbe, "Fake News Detection," 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018, pp. 1-5, doi: 10.1109/SCEECS.2018.8546944.
- [9] O. Ajao, D. Bhowmik, and S. Zargari, "Fake News Identification on Twitter with Hybrid CNN and RNN Models," in Proceedings of the 9th International Conference on Social Media and Society, Copenhagen Denmark, Jul. 2018, pp. 226–230. doi: 10.1145/3217804.3217917.
- [10] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised Learning for Fake News Detection," IEEE Intell. Syst., vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/MIS.2019.2899143.
- [11] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc., no. August, pp. 3391–3401, 2018, doi: 10.48550/arXiv.1708.07104.
- [12] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some Like it Hoax: Automated Fake News Detection in Social Networks," arXiv, arXiv:1704.07506, Apr. 2017. doi: 10.48550/arXiv.1704.07506.
- [13] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu, "Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users Based on Weakly Supervised Learning." arXiv, Dec. 13, 2020. Accessed: May 26, 2022. [Online]. Available: <http://arxiv.org/abs/2012.04233>
- [14] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," Physica A: Statistical Mechanics and its Applications, vol. 540, p. 123174, Feb. 2020, doi: 10.1016/j.physa.2019.123174.
- [15] A. Kesarwani, S. S. Chauhan, and A. R. Nair, "Fake News Detection on Social Media using K-Nearest Neighbor Classifier," in 2020 International

- Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, Jun. 2020, pp. 1–4. doi: 10.1109/ICACCE49060.2020.9154997.
- [16] A. Nagaraja, S. K N, A. Sinha, J. V. Rajendra Kumar, and P. Nayak, “Fake News Detection Using Machine Learning Methods,” in International Conference on Data Science, E-learning and Information Systems 2021, Ma’an Jordan, Apr. 2021, pp. 185–192. doi: 10.1145/3460620.3460753.
- [17] S. M. S.-U.-R. Shifath, M. F. Khan, and M. S. Islam, “A transformer based approach for fighting COVID-19 fake news.” arXiv, Jan. 28, 2021. Accessed: May 26, 2022. [Online]. Available: <http://arxiv.org/abs/2101.12027>
- [18] M. Alloghani, D. Al-Jumeily Obe, J. Mustafina, A. Hussain, and A. Aljaaf, “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science,” 2020, pp. 3–21. doi: 10.1007/978-3-030-22475-2\_1.
- [19] I. Stančin and A. Jović, “An overview and comparison of free Python libraries for data mining and big data analysis,” in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2019, pp. 977–982. doi: 10.23919/MIPRO.2019.8757088.
- [20] M. Kumar Jain, D. Gopalani, Y. Kumar Meena, and R. Kumar, “Machine Learning based Fake News Detection using linguistic features and word vector features,” in 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Nov. 2020, pp. 1–6. doi: 10.1109/UPCON50219.2020.9376576.
- [21] S. Vinothkumar, S. Varadhaganapathy, M. Ramalingam, D. Ramkishore, S. Rithik, and K. P. Tharanies, “Fake News Detection Using SVM Algorithm in Machine Learning,” in 2022 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2022, pp. 1–7. doi: 10.1109/ICCCI54379.2022.9740886.

- [22] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake News Detection Using Machine Learning Approaches," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1099, no. 1, p. 012040, Mar. 2021, doi: 10.1088/1757-899X/1099/1/012040.
- [23] V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, "Analysis of Classifiers for Fake News Detection," Procedia Computer Science, vol. 165, pp. 377–383, 2019, doi: 10.1016/j.procs.2020.01.035.
- [24] N. Smitha and R. Bharath, "Performance Comparison of Machine Learning Classifiers for Fake News Detection," in 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, Jul. 2020, pp. 696–700. doi: 10.1109/ICIRCA48905.2020.9183072.
- [25] J. E. T. Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison," International Journal of Computer Trends and Technology (IJCTT), vol. 48, pp. 128–138, Jun. 2017, doi: 10.14445/22312803/IJCTT-V48P126.
- [26] "Fake and real news dataset." <https://www.kaggle.com/clmentbisailon/fake-and-real-news-dataset> (accessed May 26, 2022).
- [27] "Supervised Machine learning - Javatpoint," [www.javatpoint.com](http://www.javatpoint.com). <https://www.javatpoint.com/supervised-machine-learning> (accessed May 26, 2022).
- [28] "Unsupervised Machine learning - Javatpoint," [www.javatpoint.com](http://www.javatpoint.com). <https://www.javatpoint.com/unsupervised-machine-learning> (accessed May 26, 2022).
- [29] "Support Vector Machine (SVM) Algorithm - Javatpoint," [www.javatpoint.com](http://www.javatpoint.com). <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm> (accessed May 26, 2022).
- [30] "Text Preprocessing NLP | Text Preprocessing in NLP with Python codes," <https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/> (accessed May 26, 2022).

- [31] “What is a Random Forest?,” TIBCO Software. <https://www.tibco.com/reference-center/what-is-a-random-forest> (accessed May 30, 2022).
- [32] T. A. Team, “Decision Trees Explained With a Practical Example – Towards AI.” <https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53>, <https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53> (accessed May 30, 2022).
- [33] Shaikh, J., Patil, R.: Fake News Detection using Machine Learning. In: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC). pp. 1–5. IEEE, Gunupur Odisha, India (2020). <https://doi.org/10.1109/iSSSC50941.2020.9358890>.
- [34] Zhang, X., Ghorbani, A.A.: An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*. 57, 102025 (2020). <https://doi.org/10.1016/j.ipm.2019.03.004>.

## **PUBLICATION**

### **1. Paper 1**

**“Fake news detection using supervised machine learning classification algorithm”**

Accepted in 4th International Conference on Inventive Computation and Information Technologies ICICIT 2022.

Venue: Coimbatore, India

Organizer: RVS Technical Campus Coimbatore, India