Detecting Fake News on Covid-19 using Machine Learning Algorithms

by

Farhan Hossain
17101484
Md Zahid Hasan
17201060
Sourov Hasan
17301091
Mobassherul Alam
17301135
Afrin Shahana
19201120

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
May 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Farhan Hossain

Farhan Hossain
17101484

Zahid Hasan

Md Zahid Hasan
17201060

Sourov Hasan

Sourov Hasan
17301091

Mobassherul Alam

Mobassherul Alam
17301135

Afrin Shahana

Afrin Shahana
19201120

Approval

The thesis/project titled " Detecting Fake News on Covid-19 using Machine Learning Algorithms "submitted by

1. Farhan Hossain (17101484)
2. Md Zahid Hasan (17201060)
3. Sourov Hasan (17301091)
4. Mobassherul Alam (17301135)
5. Afrin Shahana (19201120)

Of Spring, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on May 29, 2022.

Examining Committee:

Supervisor:
(Member)

Ms. Sifat E Jahan
Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

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

Abstract

The expansion of the Internet and swift adoption of social media platforms such as Facebook, Twitter, Instagram, Reddit, etc., has seen news and information publicized in such a way that has never been perceived in human history before. This easy access to information has resulted in an exponential increase in the misleading and falsification of news. News articles with no valid source get circulated within a society causing chaos and confusion. This work examines existing techniques and technologies used to detect fake news and demonstrates a model that sees fake news using machine learning algorithms and evaluates its performance on real-world datasets

Keywords: Covid-19, Fake Covid-19 News, Machine Learning, Classifiers, Datasets

Acknowledgement

To begin, we would want to express our gratitude to the Almighty for keeping us well and safe.

Secondly, we want to express our gratitude to our supervisor, Ms. Sifat E Jahan, and co-supervisor, Mr. Annajiat Alim Rasel, for their guidance and unwavering support. We would not have been able to do this without their assistance.

Finally, we want to thank our parents for their unquestioning support throughout our lives.

Last but not least, we owe a debt of gratitude to the COVID-19 frontline workers and COVID fighters for giving us hope and inspiring us to pursue this research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In December 2019, a unique virus known as COVID-19 was discovered in China, and It has subsequently spread to other regions of the globe, killing many people. COVID 19 is a disease that can cause severe problems in the human respiratory system. And every day, the people are getting an update about the rating point of death and what remedies are being worked on. Someone may spread fake news in today's internet age. Fake news is disseminated to ruin a person's, organization's, or institution's reputation.

As a consequence, false information has infiltrated practically every aspect of human life, with the dissemination of false information during the 2019 COVID-19 epidemic outbreak posing the most significant threat in recent memory. It might be misinformation directed towards a political party, an organization, or an institution. False news may now be spread on various websites, ranging from print media to internet sites, blog posts, social networking sites, and other online content mediums. Users may now access the most up-to-date information with a single click. These networking sites such as Facebook, Twitter, WhatsApp, and Reddit are robust and vital in their present status. They only allow individuals to talk about and share thoughts about schooling, government, politics, and healthcare. However, these sites are also utilized to make money badly. The spread of false news has accelerated in recent years, as seen by the creation of vaccination news and the epidemic symptoms on a human beings. This proliferation of non-factual material on the Internet has resulted in many issues extending beyond politics to include schooling, athletics, healthcare, and even research.

Several investigators are attempting to identify bogus information. This is when machine learning becomes very helpful. Researchers use many different algorithms to detect fake news. Identifying inaccurate information, according to scholar Wang 2017, is a difficult task. To detect bogus data, they employed machine learning. According to their studies, misinformation has grown in popularity over time, which is why you must be able to spot phony news. This is why models are developed. Following the development of supervised learning, misleading information will be detected easily. The importance of machine learning in detecting fake news is demonstrated. It is critical to comprehend the concept of misleading information. It will be necessary to look into how ml algorithms may assist us in recognizing misleading stories in the future.

1.1 Research Problem

Every year, many people suffer from fake news, which occurs genetically by giving wrong information. In one of the world statistics of 2019, the percentage of shared fake news on TV is 51%. We can't deny that many people rely on this platform. But this also includes social media and many other online media. This news detection can prevent violent crime, property crime, organized crime, victimless crime, cybercrime, political power abuse, insecurity of life, proper knowledge of how to prepare to prevent the present pandemic, and consuming wrong information.[3]

We consume news from various platforms such as trusted sources like famous websites, journalists, and outlets, but we are not getting the truth news all the time. Usually, false newsmakers use the same type of trusted website name and structure to attract people to read their written information to boost their site. Mainly they create this to influence people wrongly or as their point of view. Furthermore, this news is made for business purposes where they don't care about ethics or provide people with correct information. Sometimes we see a lot of false information about popular and respected humans. It's a trap to make a wrong impression on their character, which causes terrible effects on victims' careers and also their personal life [4].some victims suicide because of this inaccurate viral information. Because it's a big shame for victims, and often people don't believe them. We have also seen some journalists and politicians who were victims of fake news and destroyed their whole ca- reers. We can also see structural differences in false news or misleading news on popular sites like Facebook, Twitter, and website information. Clickbait is a trap where only they promote advertising but show attractive headlines to grab people's attention. Ordinary people face a significant disturbance due to this reason. Misleading Headings is another kind where fake publishers, such as news that is not entirely false, add some extra incorrect information to make it more eye-catching, which will be shown on the audience's newsfeed. Moreover, publishing false news is a dangerous factor for today's world as it can cause death too.

In the medical sector, we can see our country is not performing as well as other countries. In our medical industry, we have a shortage of doctors, but if we have seen different pages, it shows a different ratio of failings of doctors. This misguidance is not on a topic only. Fake news is a subject that hasn't been very much in the news but is increasingly getting attention. That is the fact that fake news also exists in medicine. Unfortunately, it seems like in all spheres of life. This fake medical news appears to be increasing. This is the consequence of growing sources on the Internet. Some are unsourced, which means presenting opinion pieces without any evidence. Most newspapers had highly specialized reporters who did not research appropriately on medical and science but wrote reports.

Some people publish news, but they are usually not employed by papers. We are also facing a publishing crisis in the medical literature. Some investigators discovered conflict of interest or sometimes manipulated data. Fake data was also published because of economic relationships. Not only in medical literature but also in dermatology in other areas get corrupted in the journal publishing process. And some people are promoting fake medicines and fake treatments by publishing false news

on their sites. As we can see, some people selling the covid-19 drugs on their online platforms ensure that it's working 99%, but they are not verified as trustworthy platforms. But many people fall into their trap. But it is also true that we can learn much vital information on health, but we also get fake and false information. Some examples can be publishing alternative medicine for several medical treatment cases, but they are not scientifically proven to cure cancer or other diseases. We often see a conflict between two online platform informants regarding medicine and treatment. The same case, but other illogical advice is given. Overall many unqualified doctors and sellers provide wrong information regarding treatment, medication, and current health situation.[12]

Some fake stories are not entirely fake but destroy actual events. Some publishers took old days' news, copied them, and made new misleading headlines and publication dates. These are copyright infringements. Satire is a thing where information is labeled commonly; sometimes, it's funny. But when we watched it, we realized it wasn't the news. In this kind, these advertisements are designed in such a way that encourages to click, which initiates monetization for the creator and adds reviews. Simulated news results in violence like bullying. Here the primary victims are innocent and also poor people. We can see the impact of fake news in the democratic section through elections. These issues create massive conflict from nation to nation and worldwide.[6]

1.2 Research Objectives

We all know that spreading false information is not a new phenomenon. It can stretch to more individuals rapidly with the help of social networks and other online platforms.The provider is well-known, reputable, and has proven trustworthy in the prior. The title doesn't often capture the complete picture, even in genuine media articles. However, misinformation, specifically sarcastic attempts, might contain various telling signals in the content. The tagline is also another obvious indicator of a counterfeit profile if it has one.
There are a few instances where it is seen that some writers aren't even genuine. Furthermore, legitimate or authoritative-sounding sources will be used by misinformation or reports. However, a closer examination reveals that these sources do not support the allegation. So, we can find a secure structural configuration to find out the techniques to understand fake news and conveniently detect them.

Usually, in this process, the system takes data. It matches it with its dataset, and with the help of Machine Learning (ML) tools, it returns feedback, whether the taken information is correct or not in a boolean value. the main research objectives are given below

1. To understand how the process works
2. To understand how this detector captures fake news
3. To develop a model of news detection
4. To offer recommendations to improve the mode
5 . To establish a standard structure of news detection
6. To increase the performance of this system.

Chapter 2

Related Work

After a thorough review of some research papers and journals, we understood various algorithms.Some notable testing models were used to detect false news in these research publications.

Nowadays, fake news is the only common thing we face every day as there are many articles published regularly. Social media is one of the primary sources for searching datasets. They used machine learning algorithms such as ensemble decision trees, random forests, and different tree classifiers to train the model. The suggested model's performance is assessed using a variety of evaluation indicators. On Twitter and Weibo, the experiment was conducted. The suggested model had a 90% perfection ratio within 5 minutes of the news breaking. But, the instigators used a minute dataset, limiting the applicability of this type of research in the real world. Using a deep learning method provides another framework for detecting misleading information. To classify tweets as f, the authors developed a Bidirectional LSTM with CNN.

DCDistance and Word2Vec are two document portrayal methods for text classification mentioned in[5].In brief, it works like this: It is given the original input matrix, which is said to be rather large. The addition of the feature vectors of objects of the exact origin is done. That number of features will represent each object.

On the other hand, each syllable was plotted to a vector of a specific size using Word2Vec. That is, Word2Vec's purpose was to convert every syllable into a quantitative vector. Beyond using natural language models, they also devised algorithms with specific characteristics whose output is to be derived from unprocessed news texts. These custom features are Exclamation points, quotation marks, exclamation marks throughout the text, various unique words, phrases, and symbols, and a high proportion of capital letters sentence by sentence, Descriptive words, adverbs, and nouns in proportion to the message's tone, as well as any spelling mistakes. They ran four sets of tests on each dataset: one utilizing their unique features, two employing Word2Vec and DCDistance were the computational models that produced the highest performance. Other classification techniques, such as inaccurate NOT and F1-Score measures, were surpassed by these two algorithms. The bag-of-words procedure produced the best results among the feature set. They came to the conclusion that the best strategy to detect false news is to use solely text characteristics that are as language-independent as feasible.

Following a famous article published by[2] , they did a great job gathering important ways of an algorithm. Many instant ways have been the only way to find out fake news around us, Including supervised and unsupervised algorithms. Using more and more software to reduce errors is also an excellent finding. So the main framework they proposed is Linguistic Inquiry and Word Count(LIWC), which describes the professional approach as a framework. By searching the dataset, maintaining the category of a public figure, sports, entertainment, and some website that contains fake news in the meantime. The following articles are ISOT fake news dataset and KAGGLE. So when we try to speak about LWIC, we can see that no encoding is needed for the categorical values.

There are two parts shown: training and testing, etc.; using the best learning algorithms helps confirm the accuracy of the outcomes will be verified quickly. When it comes to Algorithms, they used logistic regression, vector machine, multilayer perceptron, and KNN. So these algorithms follow specific rules to make output. They also observed a benchmark algorithm that included CNN, SVM, and Bi-LSTM. Datasets used in this article are found from open sources with both valid and invalid data. And also, precision and accuracy methods are used there. So there are some valuable findings as follows. Individual learners scored 47.75 percent correct, while entity learners scored 81.5 percent correct. Individual learners have an accuracy rate of 80%, while entity learners have a rate of 93.5 percentage for DS3. Perez-LSVM, unlike DS2, is the best-working strategy on DS3, with a 96 percent efficiency. Individual learners averaged 85 percent accuracy, whereas entity learners averaged 88.16 percent accuracy. Wang-Bi-LSTM was the poorest performing algorithm, with a 62 percent accuracy. The random forest method with Perez-LSVM obtained a maximum accuracy of 99 percent on DS1. Linguistic Inquiry and Word Count (LIWC) feature sets and various Ensemble approaches in .

A learning algorithm was utilized in tandem with the suggested technique to assess the effectiveness of false news detection classifiers.Random forest outperformed other methods on all samples. The overall accuracy of boosting classifiers (XGBoost) was higher (Table 3). The enhancing classifier (XGBoost) has an accuracy score of 95.25 percent across all data sets. The accuracy of the random forest (RF) was 96.3 percent. Enhancing classifier (XGBoost) received a score of 96.3 percent. XGBoost surpassed other learning approaches on all evaluation criteria. The main aspect contributing to XGBoost's remarkable performance is its operational concept, which successfully discovers and lowers defects in each phase. The key premise behind XGBoost's functioning was to use numerous categorization and regression trees (CART), which integrate many weak classifiers to give incorrectly classified data pieces larger weights. As a result, after each loop, the model can accurately detect the incorrectly classified points, while normalization settings limit the overfitting issue. Finally, while logistic regression was a very simple strategy, it attained an overall accuracy of over 90%.

Assessment of the fake news detection, categorized into two parts, was provided in@article. Machine learning techniques are commonly used in the Language Method and Networking Method to train classifiers to fit the assessment[1].The dataset re-

quires a" Bag of Words" representation in the first approach. Each word is a significant unit, and this method keeps track of the word occurrence of the content. When such specifications are matched with location-based terms or emotive aspects, probability ranges can provide misleading language clues.

Furthermore, deep linguistic components (sentence construction) have indeed been studied in order to anticipate failures, which is done using Probability Context-Free Grammars (PCFG). With 85-91 percent accuracy, this approach discriminates rule types for ambiguous identification[11]. The second technique, which relies on network features and behavior to forecast deceit, supports material tactics that depend on misleading speech and contamination indications. Knowledge Networks might be a big step toward scalable computational fact-checking algorithms. Due to the recursive interaction between subject and predicate via other nodes, contextual proximity is attributed to searches depending on retrieved notion claims. If the nodes are near enough, a subject-predicate-object sentence is much more likely to work. Finally, within confined areas, a language and p2p strategy has demonstrated good efficiency percentages in categorization tests.

There are so many works of literature review that should add the most valuable review on counterfeit news detection. As everyone uses different algorithms, a group from west university, Timisoara, showed the importance of artificial intelligence algorithms. The Bag-of-words model, N-gram model, term frequency-inverse document frequency model, and others are examples of models that neural networks may use to detect real and fake news. They experimented with the publication date, title, author, etc., link, etc., to match whether it's real or fake. They gathered a considerable amount of articles to find the news type. By which they applied metrics that provided curves based on the papers.

Chapter 3

Methodology

We used the supervised learning algorithm in our model to detect fake news about Covid-19. To do so, we implemented ML algorithms and NLP for the data cleaning. The figure below shows our proposed model.

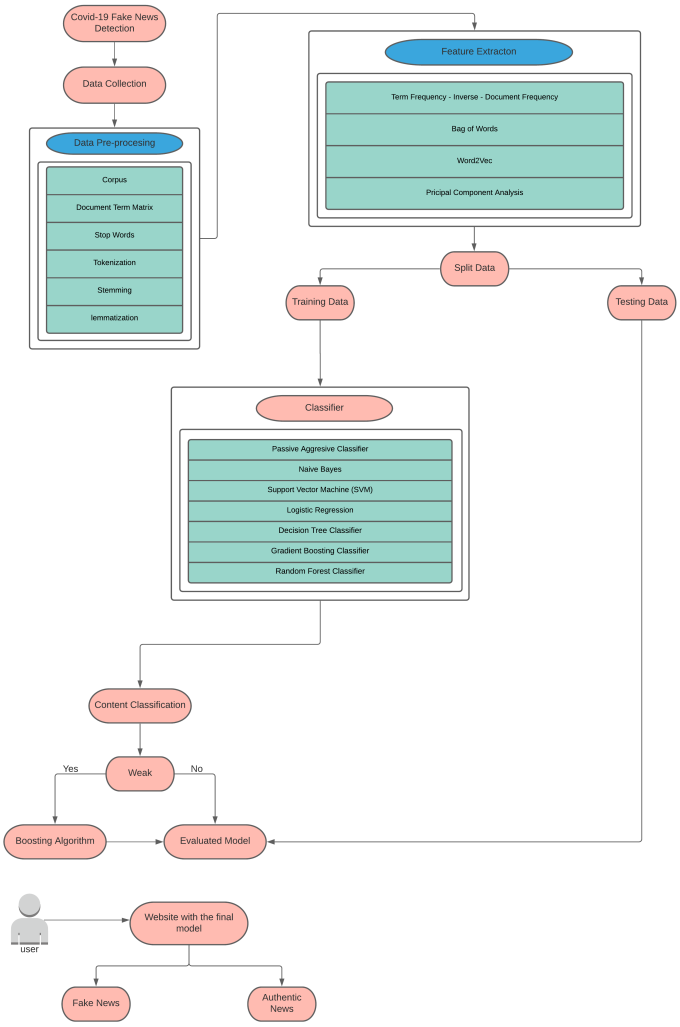


Figure 3.1: Methodology

3.1 Input Data

Inputs that are often recognized as attributes are unrestrained variables. Inputs are also known to be predictors.A data set is identical to one or more database tables when associated with tabular data. A distinct register of the data set in query corresponds to each table row, and each table column represents a particular variable. A datum is a term that refers to any value. A dataset can also, be made up of a group of papers or files. Since covid-19 is a recent occurrence, collecting data has been a significant challenge. We did find some datasets con- training about 400k real news and 100k fake news. These were available in different individual .csv files. We had to merge all the .csv files into one that contained all the values. After that, we trimmed it down to good real and fake news of about 220k and 40k data, respectively. These data included information from various websites, articles, journals, news from social media, and news from individual replies on social media. Since the social media replies only had the ID no, we had to trim them all, leaving us with about 50000 real news and 8000 fake news. The labeled dataset RealNews.csv contains 220k data and 50000 raw, real news, whereas the dataset FakeNews.csv contains 40k data in whole and 8000 essential fake information

Dataset	Real News	Fake News
Data before trimming	222000	40000
Data after trimming	50000	8000
Data for training	30000	4800
Data for testing	20000	3200

Table 3.1: Data Distribution

features in brief are:

* **ArticleID:** Contains the serial number of the news that is used.

* **Source.Name:** Contains the .csv filename.

* **Column1:** Contains the serial number of the total amount of news in the dataset.

* **fact_check_url:** Contains the information of who confirmed the news.

* **news_url:** Contains the website address from where the news has been extracted.

* **Title:** Contains the body of the news.

* **Label:** Contains whether the news is real or not. The graph from the figure shows

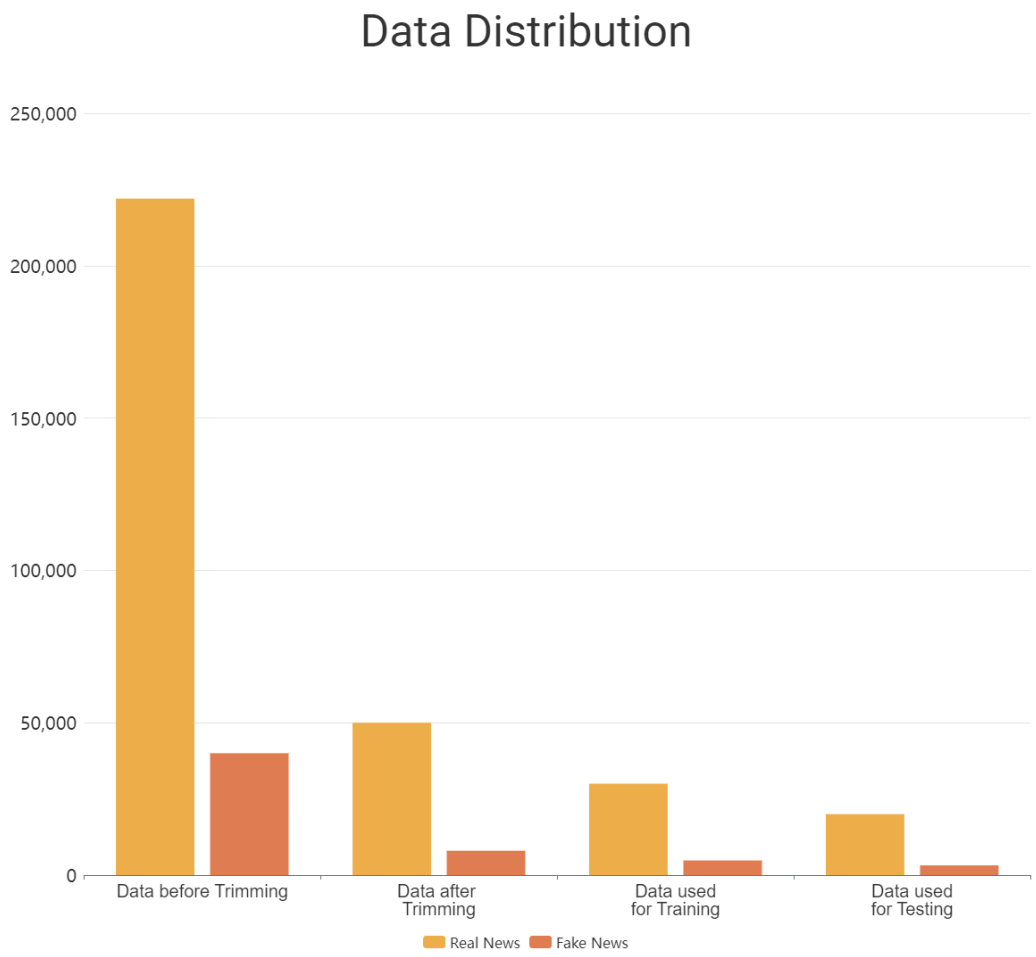


Figure 3.2: Data Distribution

The graph from the figure shows the distribution among the dataset. Around 220k news falls under the actual category, and nearly 40k information falls under the fake type. The blue bar represents the real news, whereas the red bar represents the fake news.

3.2 Data pre-processing

To maintain a well-formed dataset from an un-organized dataset, data pre-processing uses a strategy. Most of the datasets do not always follow the basic format of data. For that, It has a significant dependency on the success of project work. Data pre-processing always state that about the accuracy of the data and make it complete by correcting all the error found in the dataset following some programming. Here we want to keep the data without punctuation like commas and quotes. And we

	title	class
0	"Spraying chlorine or alcohol on the skin kills viruses in the body"	0
1	"Only older adults and young people are at risk"	0
2	"Children cannot get COVID-19"	0
3	"COVID-19 is just like the flu"	0
4	"Everyone with COVID-19 dies"	0
...	...	...
449	How can we protect people seeking tuberculosis care during the COVID-19 pandemic?	1
450	Can tuberculosis and COVID-19 be tested on the same type of specimen?	1
451	Should all people being evaluated for tuberculosis also be tested for COVID-19 and vice-versa?	1
452	Is tuberculosis treatment different in people who have both TB and COVID-19?	1
453	Does the BCG vaccine protect people from COVID-19?	1

Figure 3.3: Dropping unnecessary coloumns

also keep Those by decreasing the size together. All punctuations, for example (!;"), need to be removed during data pre-processing. During frequency generations, all punctuations must be removed. As a result, we devised specific methods to use. After applying, all the punctuations have been taken out from our dataset. Stop words are

	title
0	spraying chlorine or alcohol on the skin kills viruses in the body
1	only older adults and young people are at risk
2	children cannot get
3	is just like the flu
4	everyone with dies
...	...
449	how can we protect people seeking tuberculosis care during the pandemic
450	can tuberculosis and be tested on the same type of specimen
451	should all people being evaluated for tuberculosis also be tested for and viceversa
452	is tuberculosis treatment different in people who have both tb and
453	does the bcg vaccine protect people from

Figure 3.4: Removing punctuations

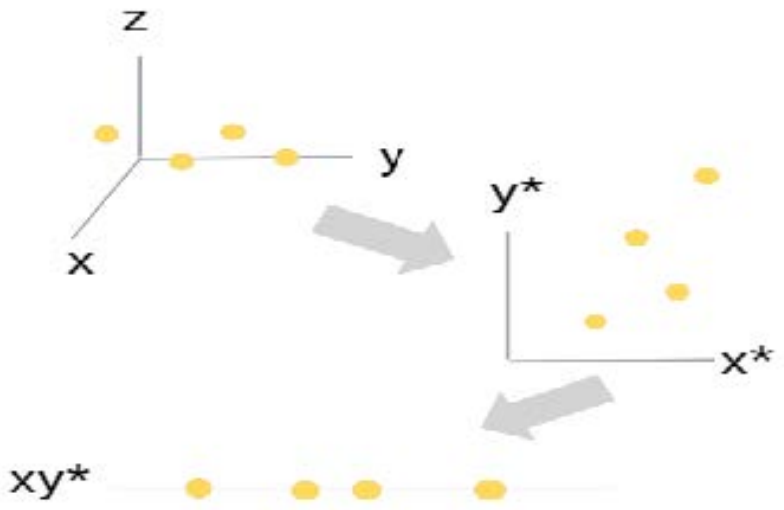
the words when we use those before processing regular languages. These are known as articles, prepositions, conjunctions, etc. When we write some sentences, these words don't contain information like the, an, or so. Because there is less number of the token, stop words help reduce the size of the database and save the training time following it. A piece of news: The death rate of COVID -19 in Bangladesh was 17.93%. Text removal after stop words: death rate COVID19 17.93% A piece of news: The death rate of COVID -19 in Bangladesh was 17.93%.Text removal after stop words: death rate COVID19 17.93%

	title
0	spraying chlorine alcohol skin kills viruses body
1	older adults young people risk
2	children
3	flu
4	dies
...	...
449	protect people seeking tuberculosis care pandemic
450	tuberculosis tested type specimen
451	should people evaluated tuberculosis tested viceversa
452	tuberculosis treatment people tb
453	bcg vaccine protect people

Figure 3.5: Removing StopWords

3.3 Feature Extraction

Machine Learning Feature Extraction once the preliminary information is extraordinarily different, which we can't utilize for Machine Learning modeling, we use feature extraction. In Feature extraction technique for extracting the only necessary data from information into a smaller set of choices.If we've got a bent to ponder, c is taken out of the equation because of its tiny value. By doing so, we are doing Feature Extraction.[14]



Feature extraction is the method of remodeling or jutting an area composed of many dimensions into an area of fewer dimensions. Representing information in multiple

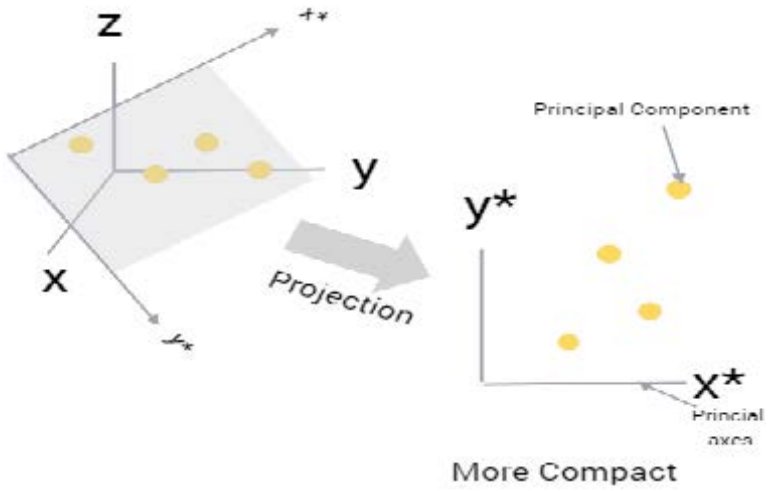


Figure 3.6: Feature Extraction

dimensions to ones that square measure less. As projection from a three-dimensional plane to a 2-dimensional plane. When the projection of parts happens, a new access square measure is created to explain the connection, known as the principal axes the new information is named principal components.[13]

3.3.1 Term Frequency-Inverse Document Frequency(TF-IDF)

As a queuing constraint for alternatives, TF-IDF is used in machine learning and data mining. The gist is that the load will increase because the word frequency during a document will increase. It means the weight will increase the longer a term within the document[18]. However, its offset helps eliminate the importance of common words like the, or a word seems within the entire information set or corpus in the document. The equation of TF-IDF examination TF-IDF code that is given below:

TF-IDF score for term i in document j=TF(i,j)*IDF(j)

Here, IDF= Inverse Document Frequency TF=Term Frequency t=Term

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ document\ j}{Total\ words\ in\ document\ j}$$

$$IDF(i,j) = log2\frac{Total\ documents}{documents\ with\ term\ i}$$

In order to utilize TF-IDF, we've integrated the TF-IDF Vectorizer from the sci-kit-learn package. We generated a TF-IDF Vectorizer class after developing the module. Anytime we've gotten passed our corpus, the entity we regard as match transformation (corpus) executes in an assortment. Once printed, we can see the results of TF-IDF.

3.3.2 Word2vec

Word2vec is a technique in technology that permits Computer science to try and do arithmetic with the word. For example, if we tend to provide an equation:

King - men + ladies

Here the computer can tell the solution is queen. The computer does not perceive text; they perceive numbers. If there's how to represent the word king as a variety, it will accurately represent the word in number. The range of words cannot be one number. It desires a set of ranges and a mathematical set of numbers known as a vector[20]. For example, if we tend to represent operating into a vector that is simply a bunch of numbers, it will represent the which means of any word accurately.Word2Vec is an algorithmic rule that distinguishes words as vectors that are in a position to search out similarities in words. This algorithmic rule uses a neural network model to be told word associations from an outsized text corpus. In this model, every word is described as a vector of many dimensions rather than one range. Word2Vec will observe similar words, and also the distinction between words is additionally preserved. This technique has the potential to save precious linguistic data.This figure shows that when victimizing the Word2Vec rule, the machine can

Figure 3.7: Word2Vec

recognize its similar values. The unlabeled raw corpus is transformed into tagged knowledge (by mapping the target word to its context word) the illustration of words is learned in an intensive classification task.Word2Vec is simply straightforward to implement. It needs to import the Word2Vec library from genesis. The model then places the corpus that the pre-processed text knowledge into the Word2Vec perform, also by providing the min count, if the value is minor, then twice as high as it is now, it'll merely ignore that word. One of Word2Vec's many capabilities is finding the most related terms. It is simple for a novice to grasp the theory and put it into practice.

3.3.3 Principal Component Analysis(PCA)

The main purpose of PCA is to discover trends in an information set and then distill the variables down to their most important features or alternatives, simplifying the data without sacrificing key characteristics. PCA asks if all the scales of the information set spark joy and provides the user with the choice to eliminate ones that don't. Principal part Analysis can be a technique to reduce distinct options

from a dataset.[19] Generally, several independent features result in overfitting the result. To avoid overfitting, PCA is a solution technique. This can not be precisely a feature extraction methodology. If truth be told, it's the contrary. In real scenarios, we work with thousands of words and features. For that reason, our information may get overfitted. As a result, PCA may be an excellent option for avoiding this. As a result, PCA can be an excellent option for avoiding this problem. PCA can work within the figure below: Principal element analysis, or PCA, is a technique that permits the methodology for summarizing relevant information in large data fields

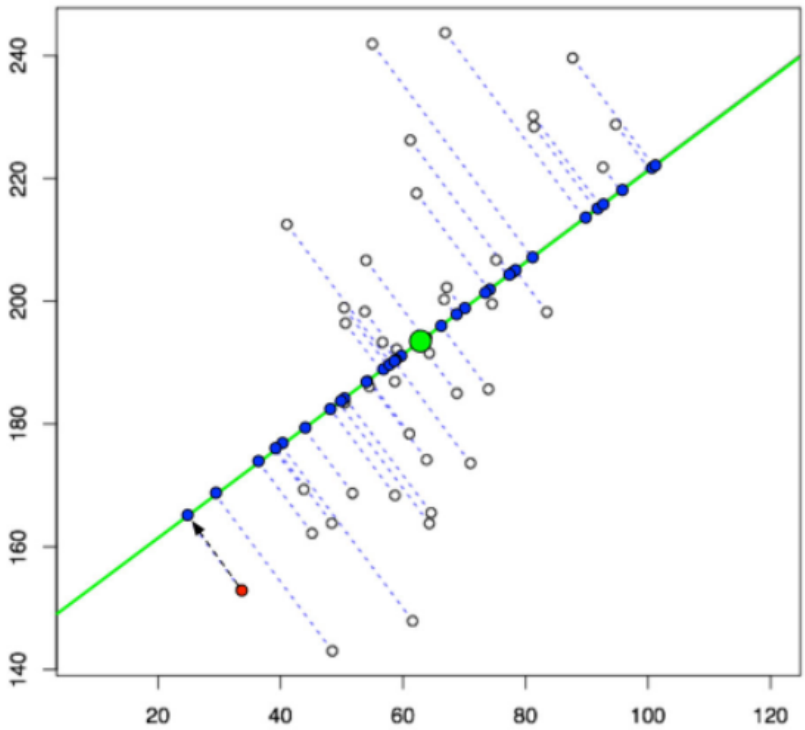


Figure 3.8: Princial Component Analysis Workflow

by proposing a reduced collection of "abstract indexes" that may be seen and studied is PCA (principal component analysis). First and foremost, the information must be scaled before using the PCA approach. Once scaling the info, we need to import PCA from sci-kit learn. The number of components required to lower the dimension in the object parameter is fixed after the PCA object is produced. Notwithstanding thousands of options in the data frame, this method can scale back the independent part to one hundred. When there are many options, a few algorithms don't work well.PCA helps to bring the most effective suited options for the algorithm and helps to Improve Algorithm Performance, Stop overfitting, Improves Visualization, and make Independent variables less interpretable.

3.4 Classifier

A classifier is said to be a machine learning method that dynamically organizes or characterizes data within one or more "classes." Because there is a lot of unstructured text on the web, such as emails, social chats, webpages, and social media, extracting value from this data is complex.[15] For instance, consider an email classifier that examines emails and determines whether or not they are spam. Text classifiers may organize, arrange, and categorize any text found in papers, files, or on the Internet. The classifiers that we used in our datasets are as follows:

3.4.1 Logistic Regression

The independent variable predicts the dependent variable using logistic regression, a form of regression.[16] It is an effective method for binary classification problems as our purpose is to detect whether the news is true or false, so this classifier is an excellent choice.

$$f(x) = \left(\frac{1}{1+e^{-x}}\right)$$

Here, y is anticipated to be 1 whenever the curve approaches positive infinite, while y is anticipated to be 0 whenever the curve approaches negative infinity. If the sigmoid function's result is greater over 0.5, the result is usually categorized as 1 or YES, and if it is below 0.5, it is usually classified as 0 or NO. Here is the code below how it works.

3.4.2 Support Vector Machine

This is a regression and classification approach using supervised machine learning. SVM finds a hyperplane that separates the different types of data. It develops the finest hyperplane frequently, which is then used to minimize an inaccuracy.

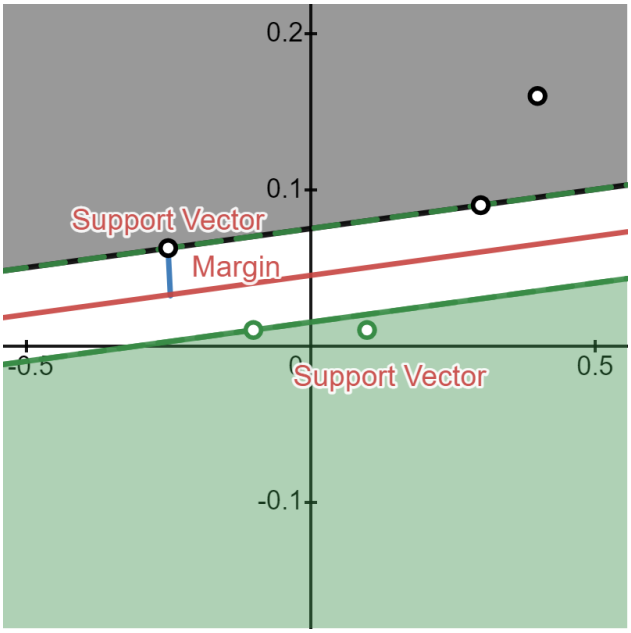


Figure 3.9: SVM

In the graph above, support vectors are the data points closest to the hyperplane. These locations will further define the dividing line by computing margin. Margin distance between two lines at the nearest layer point.A larger distance between the groups is considered beneficial; a smaller profit is found to be negative[17]. So, SVM creates a decision boundary between fake and real news data and selects extreme cases. The severe instance of fake and real news will be seen. It will be classified as real news based on the support vectors.

3.4.3 Naïve Bayes

Naive Bayes estimates that each given data point might belong to one or more of a number of topics (or not). This is used to arrange opinions, web pages, mails, and other types of content into categories, topics, or "tags" based on specified criterion. It's based on the Bayes theorem, with forecast independence assumptions. This algorithm accepts that such value of predictor (B) for a given class (A) is unrelated to the values of other forecasters. The equation is -

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

From the above equation, If B is true, the chances of A is equal to the chances of B being true multiplied by the chances of A being true divided by the chances of B being actual. So, If B is true, the probability of A is equal to the chance of B being true multiplied by the probability of A being true divided by the probability of B being actual. So, the dataset is separated into training and test sets for this classifier, and the due process has been done, respectively.

3.5 Boosting Algorithms

[10]Boosting is a collection of algorithms whose primary goal is to turn weak students into strong students. They are familiar with machine learning because they have worked with it for years, and they have become mainstream in the Data Science sector. Boosting is a machine learning strategy for reducing errors in predictive data analysis. Machine learning software, also known as machine learning models, is trained on labeled data by data scientists to generate educated estimates about unlabeled data. Depending on the accuracy of the training data set, a single machine learning model could generate prediction errors. Boosting algorithms are unique algorithms used to supplement the data model's existing results and assist in correcting faults. They employ the notion of timid learners and strong learners' discussion, balanced median scores, and greater support ratings to make predictions. These systems employ choice labels and margin-maximizing categorization for operation.' Boosting' refers to a range of techniques for helping weak learners become strong. The figure shows a general overview of the Boosting algorithm. The Boosting algorithms work in a different way than other classifiers. They find the weak rules in the classifier by applying machine learning algorithms with different distributions. To select the best distribution, the steps should be followed:

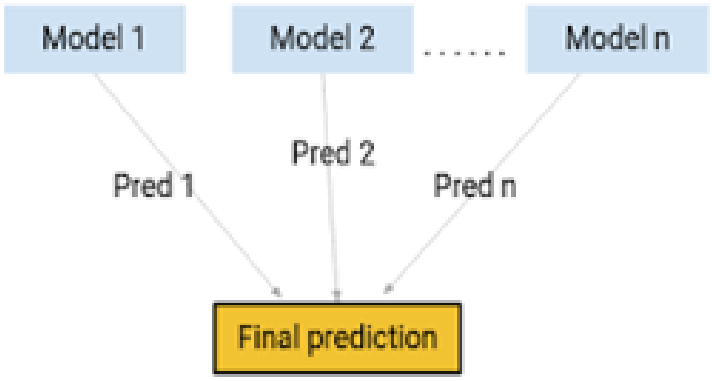


Figure 3.10: Boosting Algo

Step 1: The timid learner examines all the probabilities and gives every insight with identical values.
Step 2: If a projection creates an inaccuracy as a result of the first flawed learning approach,the observation's estimation inaccuracy is given greater weight. The following algorithm is used for weak learning.
Step 3: Iterate through the 2nd step until the base learning algorithm reaches its limit or achieves desirable accuracy. Finally, it combines the weak and robust learners to generate a strong learner who gains accuracy with time, improving the model's prediction capability. Boosting focuses more on examples that have been misclassified and have more errors as a result of preceding weak rules. The boosting algorithm we are using:
1) AdaBoost (Adaptive Boosting)
2) Gradient boosting
3) XGBoost

3.6 AdaBoost (Adaptive Boosting)

The AdaBoost algorithm is the short form of Adaptive Boosting. Adaptive Boosting technique in Machine Learning used as an Ensemble Method. Adaboost starts by creating basic assumptions on the given dataset, then gives every event identical values. If the learner's original assessment is wrong, the incorrectly predicted phrase takes precedence, resulting in a recursive cycle. It keeps adding new apprentices until the technique's maximum is reached. In Adaptive Boosting, reassigning all values to each occurrence matches the series of feeble learners at various rates, with bigger weights given to erroneously classified models. Noisy data and outliers make AdaBoost vulnerable. The figure shows the algorithm process diagram.[8] Adaptive Boosting operates similarly to the previous examples. By uniting a group of inferior learners depending on their weight and age, it develops a robust learner. It gives each data set equal weight in the first iteration and then begins forecasting that data set. Then it is put all together to make a firm prediction. [7]

3.6.1 Gradient Boosting

The machine learning approach that may be used to define and reduce loss functions is gradient boosting. Gradient boosting combines the predictions from numerous decision trees to get the final prediction. Gradient boosting is a technique for sequentially training several models. Each latest design decreases the damage value of the whole structure using the Gradient Descent approach. The learning process fitted new models in a sequential manner to get a more precise approximation of the output parameter. In gradient boosting machines, all weak learners are decision trees. The goal is to collect diverse signals/information from data, which is accomplished by selecting the optimal split using a different subset of attributes. This implies that the trees aren't all the same, and thus they can catch distinct signals from the data. The figure shows gradient boosting workflow. This boosting approach successively trains numerous models. Gradient Boosting is used to solve problems of classification using prediction models. GBRT, commonly known as Gradient Tree Boosting, is used in the Python Sklearn module. It's a variant of boosting that applies to any differentiable loss function. It applies to both regression and classification problems.[7]

3.6.2 XGBoost

Extensive gradient boosting is abbreviated as XGBoost.Because of the sequential model training, gradient boosting machines are often slow to implement. XGBoost is mainly used to solve classification difficulties, but it can also solve regression issues. It's a variant of a gradient boosted decision tree meant to help you make better decisions. Features of XGBoost are:

- **Regularized Learning:** By softening the resulting learning values, it assists in minimizing over-fitting. Models that employ fundamental, analytical variables will benefit from the systematized goal.

- **Gradient Tree Boosting:** Conventional Euclidean spatial minimization approaches can be utilized to enhance the tree ensemble framework.

- **Shrinkage and Column Subsampling:** In contrast to the formalized objective, two extra measures are used to mitigate overfitting. The first strategy, shrinkage, was presented by Friedman. Shrinkage reduces freshly inserted values by a fraction after every stage of tree enhancing. Shrinkage, like a training ratio in variable optimization, reduces the impact of every tree while allowing subsequent trees to improve the system.

The XGBoost can execute parallel processing on a given computer and contains simultaneously a tree training process and a linear design training technique.It's ten times quicker than every existing gradient boosting method. The gradient descent design is employed by XGBoost and GBMs in all three techniques.XGBoost distinguishes itself from competing GBMs in the field of networkEfficiency and algorithmic improvements.[9]

Chapter 4

Result

We used the TF-IDF Vectorizer after pre-processing the title. We discovered connections between the title and the label. With TF-IDF Vectorizer extraction, imported from sklearn. Feature extraction. Text, the tags with their labels (1 or 0) are represented as an array. Next, the data is dissected into two categories: training and testing. The testing data size is 40%, with the remaining 60% being utilized for training. To fit the model we acquired from the training dataset, we used Multinomial Na¨ıve Bayes, Passive Aggressive, Decision Tree, Random Forest, Suppor Vector Machine, and Logistic Regression.

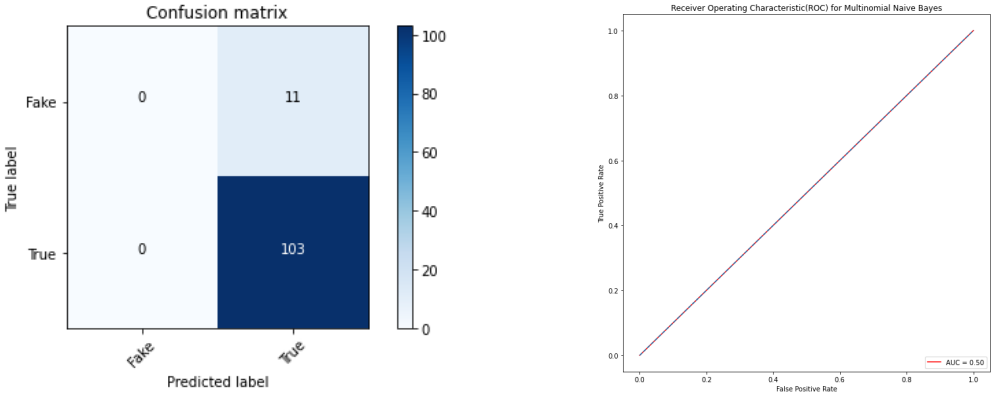


Figure 4.1: Confusion Matrix and ROC Curve for Naive Bayes

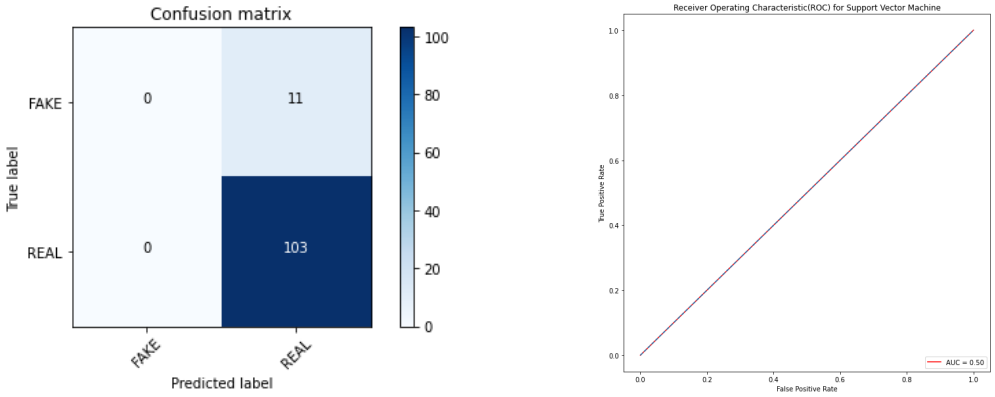


Figure 4.2: Confusion Matrix and ROC Curve for Support Vector Machine

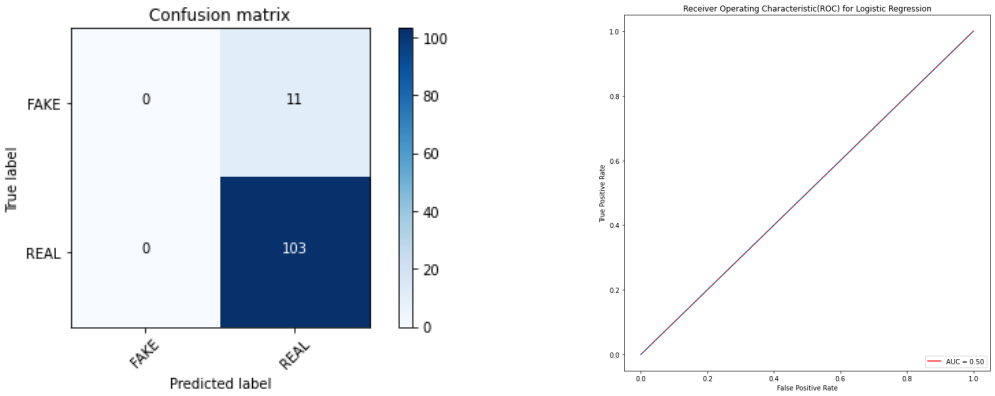


Figure 4.3: Confusion Matrix and ROC Curve for Logistic Regression

In summary, Naive Bayes, Logistic Regression, and Support Vector Machine have given the most accuracy, precision, recall, and f1-score (0.90), and Passive-Aggressive Classifier has given the least accuracy, precision, recall, and f1-score (0.87). After implementing the mentioned classifiers on our datasets, the results that we acquired are given below.

Classifier	Accuracy	Precision	Recall	f1-score
Passive Aggressive	0.88	0.91	0.96	0.93
Naive Bayes	0.91	0.91	1.00	0.95
SVM	0.90	0.90	1.00	0.95
Logistic Regression	0.90	0.90	1.00	0.95
Decision Tree	0.89	0.91	0.98	0.94
Random Forest	0.90	0.90	1.00	0.94

Table 4.1: comparison of classifiers

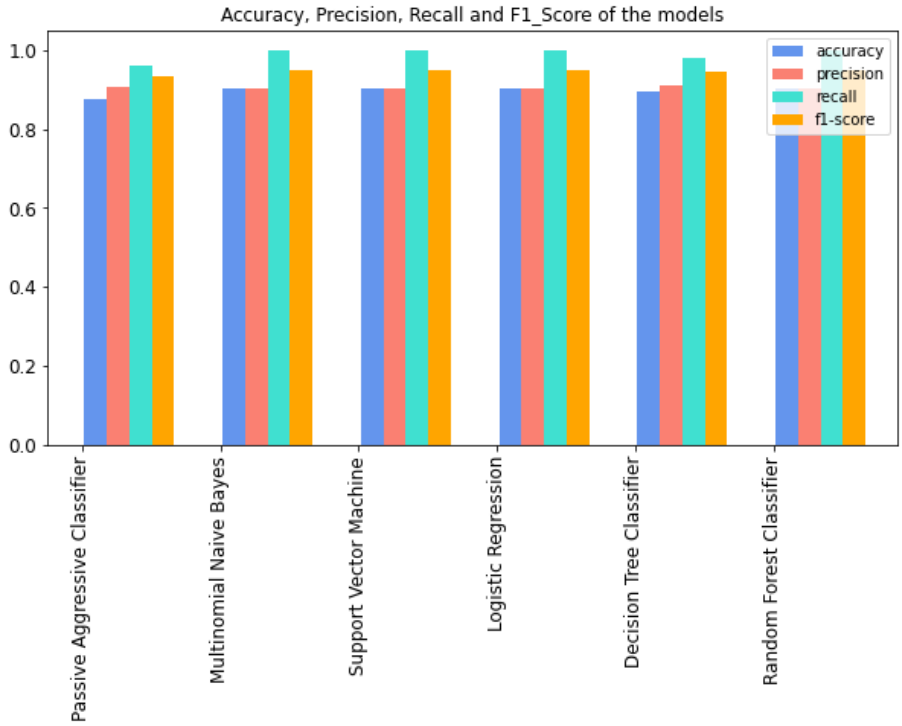


Figure 4.4: Graphical Representation of Results

Chapter 5

Conclusion

Now in this digital world, it is straightforward to spread fake news. Many people are always connected to the internet and social media platforms. People are using Facebook, Twitter, Instagram, etc. These platforms are used to spread fake news. Fake news starts proliferating against personal, political parties, institutions, or institutions when some bad people take benefit of these tenets. It can tear down a reputation or affect daily life or career. Through fake news, citizens' opinions towards a political party can also be changed. It is necessary to find a strategy to spot misleading information. Different machine learning methods are used, unlike intention, which can also be used to detect fake news. The study's main objective is to find out misinformation from general news. The classifiers that are first trained with a dataset is called training dataset. Then these classifiers can automatically detect fake news. The detection of false information has many problems that need investigators' observation. For example, identifying the critical factors involved in spreading disinfo is essential to reduce the spread of false information. To point out the primary origins of spreading fake news, various ML techniques can be used. After collecting it, we pre-processed our dataset using two distinct forms: Corpus

and Document-Term Matrix. The Multinomial Nave Bayes classifier was used to arrive at our conclusion. The word" title" has an accuracy score of 90%. We need to include other false news in our dataset to enhance accuracy. To create a better Document-Term Matrix, we must also tokenize and lemmatize our corpus. We'll also need to expand our collection of "Stop Words." To improve accuracy, we'll use boosting algorithms like XGBoost, AdaBoost (Adaptive Boosting), and Gradient Boosting. Even though our dataset has 40k records, we first chose approximately 8000 false news stories. For more accurate findings in the future, we will employ some more classifiers.

Bibliography

[1] N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the association for information science and technology*, vol. 52, no. 1, pp. 1–4, 2015.

[2] B. Al Asaad and M. Erascu, "A tool for fake news detection," in *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, IEEE, 2018, pp. 379–386.

[3] S. Shabani and M. Sokhn, "Hybrid machine-crowd approach for fake news detection," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, IEEE, 2018, pp. 299–306.

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

[5] P. H. A. Faustini and T. F. Covoes, "Fake news detection in multiple platforms and languages," *Expert Systems with Applications*, vol. 158, p. 113 503, 2020.

[6] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021.

[7] *A gentle introduction to the gradient boosting algorithm for machine learning*, https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/, (Accessed on 05/24/2022).

[8] *Adaboost algorithm: Boosting algorithm in machine learning*, https://www.mygreatlearning.com/blog/adaboost-algorithm/, (Accessed on 05/24/2022).

[9] *An intro to xgboost for machine learning*, https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/, (Accessed on 06/03/2022).

[10] *Best boosting algorithm in machine learning*, https://www.analyticsvidhya.com/blog/2021/04/best-boosting-algorithm-in-machine-learning-in-2021/, (Accessed on 05/24/2022).

[11] *Explained: What is fake news? | social media and filter bubbles*, https://www.webwise.ie/teachers/what-is-fake-news/, (Accessed on 05/24/2022).

[12] *Fake medical news and other costly misinformation on cancer | uicc*, https://www.uicc.org/news/fake-medical-news-and-other-costly-misinformation-cancer, (Accessed on 05/24/2022).

[13] *Fake news worldwide - statistics & facts | statista*, https://www.statista.com/topics/6341/fake-news-worldwide/, (Accessed on 05/24/2022).

[14] *Feature extraction - wikipedia*, https://en.wikipedia.org/wiki/Feature{_}extraction, (Accessed on 05/24/2022).

[15] *Machine learning classifiers - the algorithms & how they work*, https://monkeylearn.com/blog/what-is-a-classifier/, (Accessed on 05/24/2022).

[16] *Python logistic regression tutorial with sklearn & scikit | datacamp*, https://www.datacamp.com/tutorial/understanding-logistic-regression-python, (Accessed on 05/24/2022).

[17] *Scikit-learn svm tutorial with python (support vector machines) | datacamp*, https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python, (Accessed on 05/24/2022).

[18] *Understanding tf-id: A simple introduction*, https://monkeylearn.com/blog/what-is-tf-idf/, (Accessed on 05/24/2022).

[19] *What is principal component analysis (pca) and how it is used?* https://www.sartorius.com/en/knowledge/science-snippets/what-is-principal-component-analysis-pca-and-how-it-is-used-507186, (Accessed on 05/24/2022).

[20] *Word2vec | tensorflow core*, https://www.tensorflow.org/tutorials/text/word2vec, (Accessed on 05/24/2022).