Fake Data Analysis with Detection

Submitted By

Anamul Haque Shanto
ID: 181472611

A project report submitted in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

SHARMIN AKTER
Lecturer
Department of CSE
City University

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CITY UNIVERSITY
DHAKA, BANGLADESH

March 2022

DECLARATION

We declare that this project report titled “**Fake Data Analysis with Detection**” is the result of our own work except incident news & cited in the references. This project is the partial fulfillment of requirement for the award of degree of Bachelor of Computer Science and Engineering during the period of **2022** in **City University, Dhaka.** The project report has been carried out under by guidance and is a record of work carried out successfully during December-2021 to March-2022. To the best of my knowledge this project has not performed anywhere for a degree.

Submitted by

…………………………………………..
Anamul Haque Shanto
ID: 181472611
Department of Computer Science & Engineering
City University, Bangladesh.

Submitted to

…………………………………………..
Sharmin Akter (Supervisor)
Lecturer
Department of Computer Science & Engineering
City University, Bangladesh.

ACKNOWLEDGEMENT

First of all, we are thankful to the Almighty for his Blessings. Then We would like to express our Heartiest gratitude towards our Supervisor and Lecturer, **Sharmin Akter** Department of Computer Science and Engineering, City University**.** Without her guidance, this project would have never been completed. Her scholarly guidance and valuable advices made it possible to complete this project.

We would like to give special thanks and all credits of the project to our honorable Sir **Md. Safaet Hossain**, Associate professor and Head of the department, Department of Computer Science & Engineering, for his help in various ways from CSE department during his busy schedule. We are also very thankful to the Honorable Dean of Department of Science Faculty, **Prof. Dr. Engr. Md. Humaun Kabir**, for his endless support.

Finally, and this goes without saying, we want to say a special thank you to **City University** for enabling us to conduct this project and providing us with the right support.

Name of Student

………………………………………..
Anamul Haque Shanto
ID: 181472611
Department of Computer Science & Engineering
City University, Bangladesh.

March 2022

DEDICATION

This project is dedicated to all the amazing programmers of all over the world whose are putting their effort to make our life easy and stress-free.

ABSTRACT

Nowadays internet has become an indispensable part of our lives especially on social media platforms such as Facebook, Twitter, and Instagram in which, everyone relies on various online resources for news and study purposes. So, Fake data detection is the most significant issue to be tended to the recent years, there is part of research happening in this field. News spread rapidly among millions of users within a very short-time via online platforms. Alongside, the spread of fake news has far reached with the current usage of social media. Consumers are creating and sharing more information than earlier, some of which are misleading and have no relevance with reality. Especially, automated classification of a text article as disinformation is a challenging task where even experts face difficulties before giving a verdict on the truthfulness of an article. In this project, we propose to perform binary classification of various news articles that are available online with help of concepts of machine learning, natural language processing. We aim to provide the user with the ability to classify whether the news is fake or real.

Table of Contents

LIST OF FIGURES

Chapter 1

Introduction

Information or data is the most important resource in this century. The main issue to be settled is to assess whether the information is significant or insignificant. Counterfeit information tremendously affects part of individuals and associations that might even prompt the finish of the association or frenzy individuals [1].

With the fast improvement of data innovation in the beyond twenty years. PC networks are generally utilized by industry, business and different fields of the human existence. Subsequently, building dependable organizations is a vital assignment for IT executives. Then again, the quick advancement of data innovation delivered a few difficulties to fabricate solid organizations which is an extremely challenging assignment. There are many kinds of assaults undermining the accessibility, uprightness and classification of PC organizations [3].

Machine learning scientists accept that this issue can be addressed utilizing the AI calculations and there is parcel of on-going examination in this field which lead to the new branch [1].

This order isn't that straightforward there are parcel of difficulties to go through to succeed. How about we start with not many of them AI works with the information on the off chance that you are having colossal and clean information, there was an extraordinary possibility making incredible classifier. To make a constant application, the calculation ought to be taken care of with the latest information. Information is of various sizes so that ought to be appropriately cleaned to obtain better outcomes [2].

Social media sites began to show up around 2005 and a large number of them have drawn in hundreds of millions of clients. The quantity of particular profiles at Facebook surpass one billion. Since online media destinations need to draw in however many clients as would be prudent, solid validation of client's character isn't needed by them when another client joins the website [1].

1.1 Motivation and Problem Statement

1.1.1 Problem Statement

Fake Data Analysis is the most important problem to be addressed in the recent years, there is lot of research going on in this field. Because of its serious impacts on the readers. researchers, government and private agencies working together to solve the issue. This Project have been focused on classifying online reviews and publicly available social media posts. Fake News stories usually spread through social media sites like facebook, Instagram, Twitter etc [1].

1.1.2 Motivation

- Fake news is differentiated by the content that mimics news media in form but not in editorial processes.
- People can avoid getting cheated on by using Fake Data Detection.
- People are profiting by clickbait's and publishing fake news on online.
- With Fake Data Analysis less people will be influenced by fake news.
- It helps us to find out real data as valid information from online platform.

1.2 Project Goals

The main objective behind the development and upgradation of existing projects are the following smart approaches:

- ➢ To find out the real news of this recent world.
- ➢ To aware people on fake information.
- ➢ To alert of such article while forwarding to others
- ➢ To develop the news articles as either fake or real.
- ➢ To reduce the number of increasing fake news in social media.
- ➢ To be informative.

- To safely share data for testing the scalability of algorithms and the performance of new software.

1.3 Project Objectives

The objective of developing a web app for fake news detection system includes:

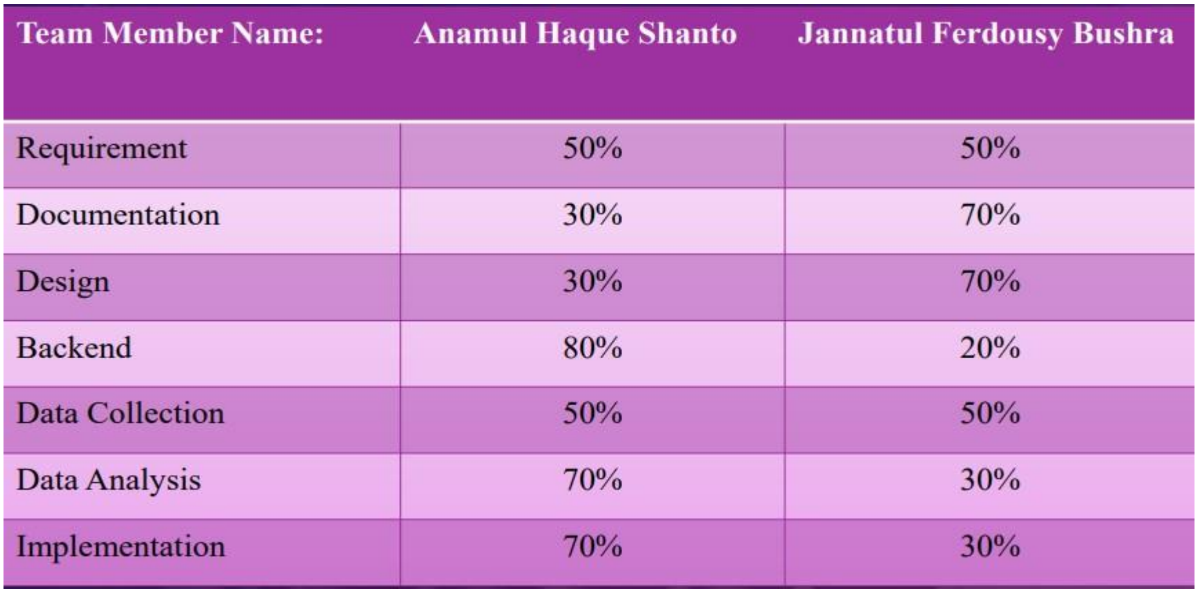
- The application will take input from and will classify if news is real or fake.
- The user will have an option for testing his input data on the chosen classification algorithms.
- The user will also have an option for testing his input csv file based on three classification algorithms. The classification result will download as a csv file.

1.4 Why do we choose this project?

The Fake Data Analysis is used to explore and visualize the data to identify patterns and insights from fake and real news. Fake data news are mandatory, and it has turned into an investigation challenge to reliably take a look at the information, content, and appropriation to name it as right or wrong. Numerous specialists have been attempting to deal with this issue, and they have additionally some way or another been effective. Some have investigated the field of AI, and some have investigated profound learning. All things considered, nobody has at any point created research in the field of opinion investigation or feeling data. So, we would also like to contribute in this path to make our social media experience more secure.

1.5 Team Member Work Distribution

Table 1.1: Team Work Distribution



Team Member Name:	Anamul Haque Shanto	Jannatul Ferdousy Bushra
Requirement	50%	50%
Documentation	30%	70%
Design	30%	70%
Backend	80%	20%
Data Collection	50%	50%
Data Analysis	70%	30%
Implementation	70%	30%

1.6 Time Distribution in Different Tasks

Table 1.2: Estimated day

Tasks	Duration
Task 1 – Analysis, Discussion	10
Task 2 - Planning	20
Task 3 - Design	30
Task 4 - Coding + Testing	30
Task 5 – Document, Report Paper	20
Task 6 – Final Testing, Presentation	10

1.7 Gantt Chart

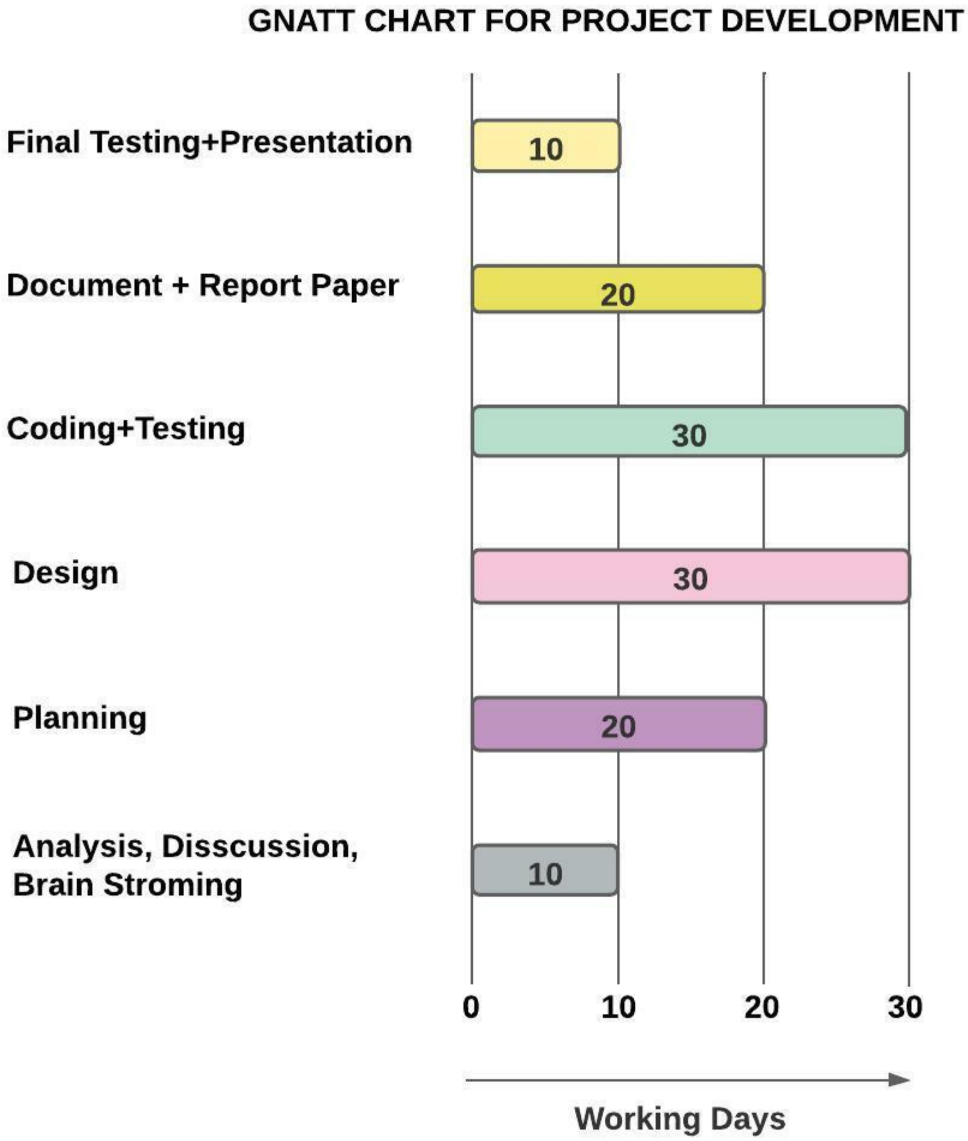


Figure 1:Gantt Chart

Chapter 2

Background Study

When people don’t use the internet, they acquired their news from the radio, TV, and papers. With the internet, the news moved on the online, and shortly, anybody could post data on sites like Facebook, Twitter, Instagram. The spread of phony news has likewise expanded with web-based media. It has become one of the main issues of this century. Individuals utilize the technique for counterfeit news to contaminate the standing of an all-around presumed association for their advantage. The main justification behind such a venture is to outline a gadget to look at the language plans that depict counterfeit and right news through AI. This paper proposes models of AI that can effectively recognize counterfeit news. These models recognize which news is genuine or counterfeit and determine the precision of said news, even in an intricate climate [8].

Fake news is something that everyone is very fond of and needs no introduction. We have seen that internet use has taken off dramatically in recent years, as social media platforms such as Facebook, Twitter, WhatsApp, etc., have evolved. We also should not forget to mention YouTube, one of the biggest culprits in spreading fake news among the population. These applications have many benefits, such as sharing something useful for the betterment of the population [2].

2.1 Features

- Comparative Analysis for different algorithm.
- Real time data
- Confusion matrix visualization on chart
- Fake News detection on web interface
- Data accuracy rate.

2.2 Existing System

2.2.1 Existing Work-1

Title: Fake Data Analysis and Detection Using Ensemble Hybrid Algorithm

A Support Vector Machine(SVM) is a discriminative named classifier where segregation is accomplished by making an isolating hyperplane [1].

It works on the bag of words features where the data of different articles collected is converted into encoded format by using various vectorization techniques based on the requirement some of them are count vectorizer(CV), term frequency and inverse document frequency vectorizer. In this project the following accuracy rate is given on confusion matrix—

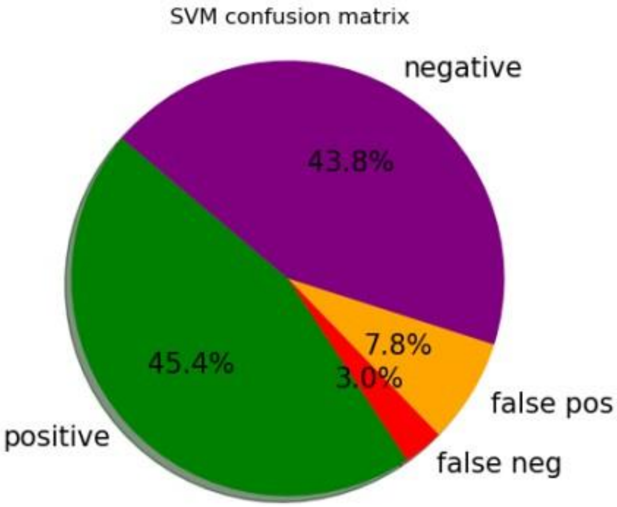


Figure 2:Pie chart representation of SVM

Confusion matrix visualization: Confusion matrix is a great way of analyzing a machine learning model. This has the data of True positive, False positive, False negative, True Negative. The confusion matrix of SVM is as follows [575, 38, 99, 555]

Key Features:

- The ensemble methods used to detect fake news.
- Extracting linguistic features from textual articles and training multiple ML models including K-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), linear support vector machine (LSVM), decision tree (DT), and stochastic gradient descent (SGD),
- Achieving the highest accuracy 82% with SVM and logistic regression.

Weakness of Ensemble Hybrid Algorithm:

- This ensemble process is not better to creat a new observation.
- Ensemble cannot help unknown differences between sample and populations.
- Ensemble can be more difficult to interact.

2.2.2 Existing Work-2

Title: Fake News Detection using NLP

The proposed project uses NLP techniques for detecting the 'fake news', that is, misleading news stories which come from the non-reputable sources. By building a model based on a K-Means clustering algorithm, the fake news can be detected. The data science community has responded by taking actions against the problem. It is impossible to determine a news as real or fake accurately. So the proposed project uses the datasets that are trained using count Vectorizer method for the detection of fake news and its accuracy will be tested using machine learning algorithms.

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. Deep learning is a class of machine learning algorithms that utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning [4].

Figure 3:Fake News Detection using NLP

Key Features:

- visualization for both file
- Data frame for output
- dataset for true and fake file.
- Authentication for news

Weaknesses of Fake News Detection using NLP:

- Complex query language.
- This project is built for a single and specific task only.
- Web less project.

2.3 Critical Remarks of Previous Works:

Table 1.3: Comparison of the previous solutions and the proposed method

Features	Ensemble Hybrid Algorithm	Fake News Detection using NLP	Proposed Method
Machine Learning Algorithm	✓	✓	✓
Natural Language Processing	✓	✓	
Analysis for different algorithm			✓
Real data implementation.		✓	
Confusion matrix visualization on chart	✓		✓
Data accuracy rate.	✓	✓	✓
Fake News detection on web interface			✓

Chapter 3

System Methodology

3.1 Data Selection

This stage is the first strategy of this project. For purpose of this, we searched Bangla dataset but No datasets were found in Bangla language as well as it was hard to gather data of Bangla news. Additionally, Bangla keyword and regular expression wasn't available to the google. Consequently, we have used English dataset that has more than 20,000 data with 4 data columns and collect from Kaggle. Around 2500 articles were collected for our dataset as all these are public articles and newspaper. The dataset consists of fake news and real news. Each file of the dataset consists of more than twenty thousand examples of fake news and real news. The dataset considers the title, text, author, and level and the dataset comprises information used from the fake and real news to datasets.

3.2 Phased development

Our approach evaluates the performance of models by following given scenario

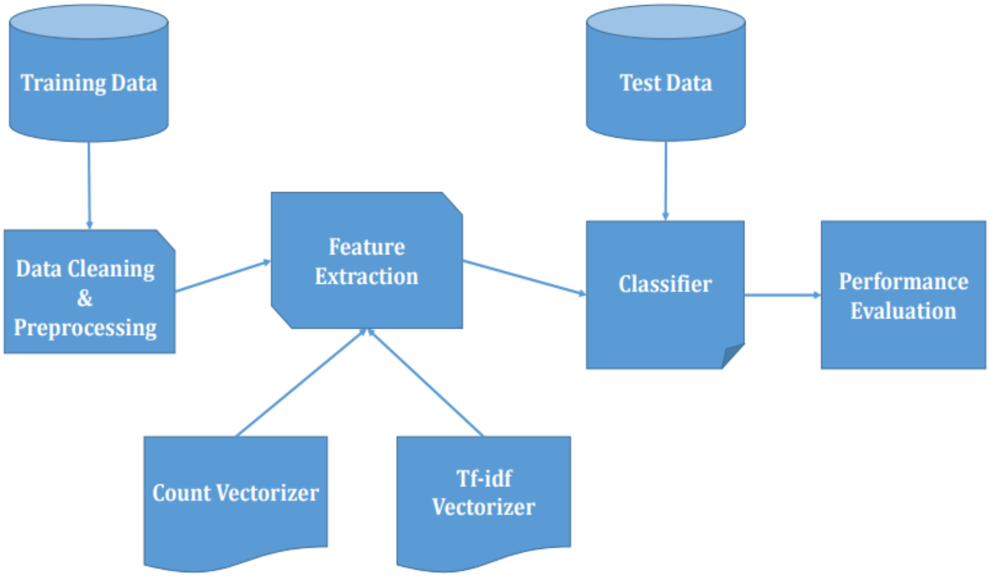


Figure 4:Phased development

3.3 Data Pre-process

Preprocessing plays a significant role on raw text data before feeding data to the classifier. It is very important to apply some preprocessing on the raw text data. A raw text might contain unnecessary symbols or other things that are not important for our classification. We also removed various special characters e.g. @, #,! etc. from our text. Data must be preprocessed to help improve the data quality and its results [5].

There are 4 Steps in Data preprocessing:

Data cleaning: It is first process of data preprocessing to clean the unnecessary data by filling missing values, smoothing noisy data, resolving the inconsistency, and removing outliers. It is used to fill all voids and nulls, eliminate noise discard inconsistent data, and erase isolated data.

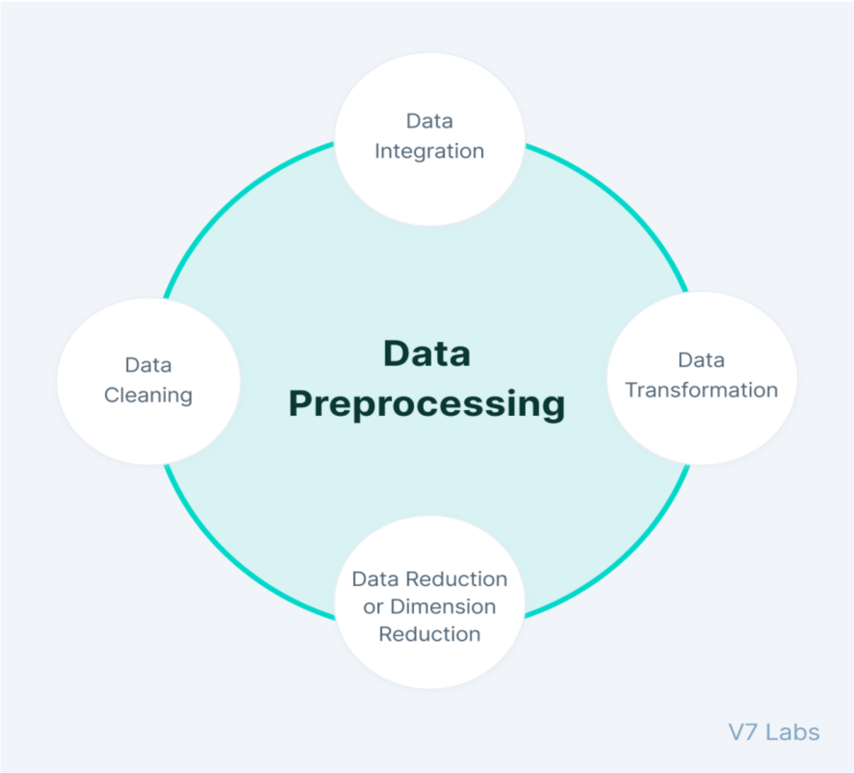


Figure 5:Data Preprocessing

Data integration: It is used to merge the multiple sources data into a single larger data. Data integration brings together data from different sources into a data warehouse which can be larger. The data sources could be ordinary files, data cubes, or several databases. It is reacquired when we are aiming to solve a real world scenario like detecting the presence data from image.

Data reduction: It is used to minimize the size of data through clustering, aggregating, or removal of redundant features. Additionally, it is handled by data analysis and data mining algorithm. This should be done with the integrity of the data still in place.

Data transformations: Data transformation is applied to data that is required to fall within a smaller range like 0.0 to 1.0. It is actually modifying the data that process based on turning the data into the proper format. This helps to enhance the efficiency and accuracy of algorithms with distant measurements. In our project Data transformation happens in Feature Selection that decide which variables like features, characteristics, categories etc. are most important to our analysis. Not only this but also these features will be used to train ML models.

3.4 Feature Extraction Techniques

Feature selection is an important aspect of this project which involves using data mining techniques to select the best set of features which can prove useful in building a predictive model, it is a technique that involves selecting relevant features and removing redundant features.

Feature selection picks a subset of the most important features that apply to a dataset. Few features tend to allow machine learning algorithms to run more effectively and more efficiently with complexity in time with lesser space. Some inputted features that are irrelevant can mislead some machine learning algorithms which can worsen their predictive performance thereby affecting the overall result. To reduce the much data, feature extraction work to be processed by transformed. The input data process of transforming into the set of feature is called feature extraction. For this Technique mainly two approach is considered such as ----

Principle component analysis (PCA): It is used for unsupervised data compression.

Linear discriminant analysis (LDA): Ii is used for supervised dimensionality reduction technique for maximizing class reparability [3].

3.5 TF-IDF Vectorizer

We know that in machine learning there are 2 different ways of converting Text to Numbers for Analysis. Such as Count Vectorizers and TF-IDF verctorizer. Amount them count vectorizer works for convert a given set of strings into a frequency representation. On the other hand, TF-IDF works for providing a numerical representation of a string and statistical analysis. Not only focuses on the frequency of words but also provides the importance of words [6].

Now TF-IDF briefly explained below:

TF stands for Term frequency that works by looking at frequency measurement of a particular term that are connected to the document. IDF stands for Inverse Document Frequency. It looks at how common or uncommon word is amongst the corpus. They measure by following several multiple measures or way of defining frequency [3]. Such as

- By counting the number of words appears in a document.
- By adjusting the length of document in which raw count of occurrences divided by number of words in the document.
- Following Logarithmically scaled frequency.
- By maintaining Boolean frequency.

Measurement process of frequency for TF:

$$tf_{t,d} = \frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document}$$

Here, in the numerator, n is the number of times the term “t” appears in the document “d”. Thus, each document and term would have its own TF value [4].

Measurement process of frequency for IDF:

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

3.6 Classification by Machine Learning

Machine learning tracks how the performance of computers can be improved with regards to the data made available. The ability to automatically learn to make an intelligent decision with regards to the provided data and to identify complex becomes a major research project for computer programmers. In this stage of the process, machine learning classifier algorithms are used to train and test the models to see how they perform in predicting legitimate and fraudulent credit card transactions. Machine learning is a subpart of the Artificial Intelligence framework which allows us to build machine learning models using historic datasets to imitate human thoughts and later to use the trained models to perform tasks automatically without being explicitly programmed. [2].

There are six types of machine learning which evolved from supervised learning which is classified into different methods, each of which has its different algorithm. Algorithms such as multinomial Naive Bayes, passive Aggressive Classifier, Logistic Regression Algorithm, random forest classifier, Decision Tree Classifier, and K Nearest Neighbor (KNN) are common machine learning algorithms used in Fake News Detection Project. The system architecture below shows the proposed techniques used to detect Fake news [10].

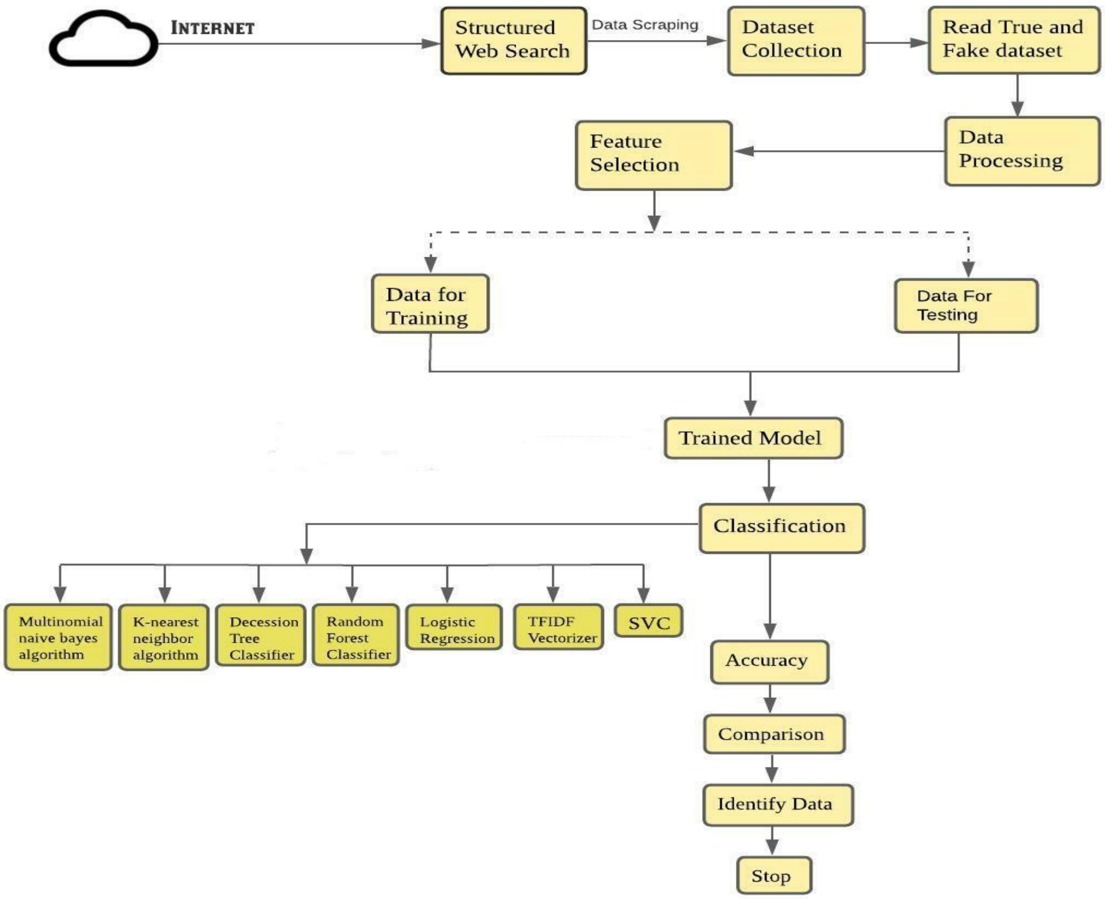


Figure 6:Classification by Machine Learning

3.7 Performance Evaluation

This portion is evaluated the experimental setup that is related to the results. All performance was implemented on Goggle Colab, which provided a cloud environment. The libraries of Python that we used for training and testing. These are Numpy, Pandas, Sklearn, Natural language Tool kit (NLTK) and Matplotlib. We separated our dataset into the testing and training dataset. The ratio of them are 20:80.

The results of our project were tested through a confusion matrix and a sklearn library. First, the TF-IDF vectorizer was evaluated on the test dataset. Then all algorithms were performed to test and found almost more than 90 % accuracy rate except K-NN algorithm with confusion matrix [5].

Chapter 4

Work Procedure

This chapter gives an overview of the project Outline of methodology, Data Collection, Data Analysis & Pre-Processing, Build the Designing Model and Design and implementation of the system.

4.1 Outline of methodology

1. Analysis
2. Planning
3. Design
4. Development & Testing
5. Deployment.

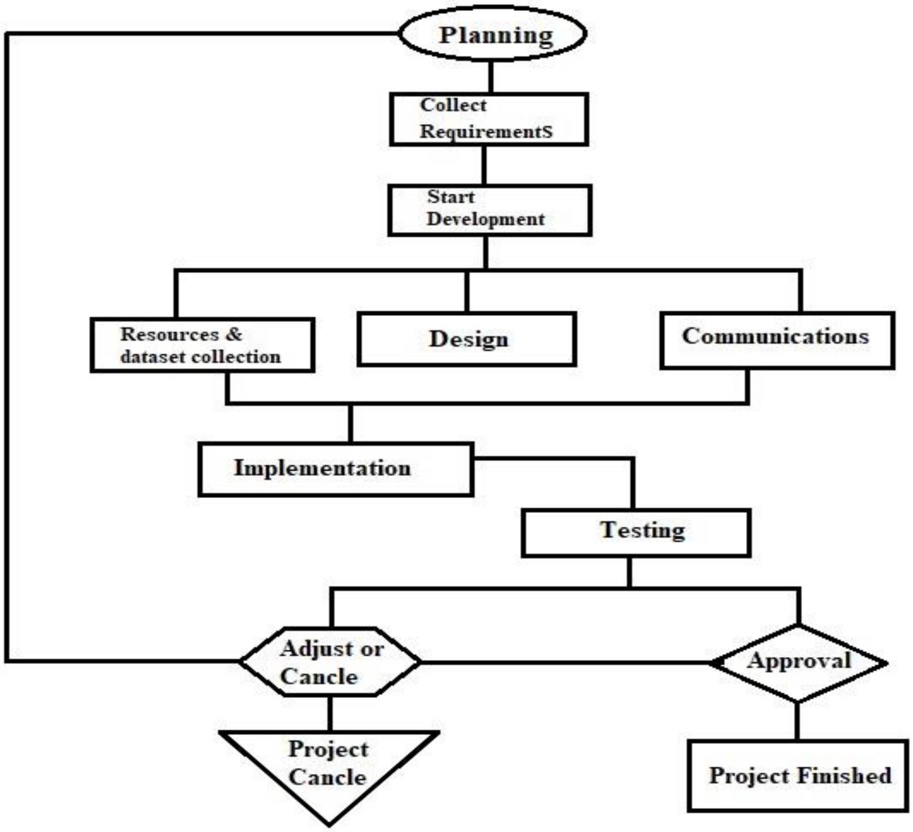


Figure 7:Development Workflow for Fake Data Analysis

4.2 Planning

At the planning phase, the project is initialized by determining the problem statement and the motivation to start the project. After that, the scope and objectives of the project will be set and documented. All system requirements will also be identified and collected. Then build the development part.

After that, the resources searching and design will be done by group members communications. This process is to determine the previous arts or current solution towards the problem. The strength and weakness of any previous work will be identified and will be used to compare to our proposed solutions in order to produce a better application. Other than that, all the requirements will be also be filtered and justified before proceed to the next phase. Upon completion of the planning and analysis phases, we will start to detect the system.

During the development process of the system of a different version, we will verify and analysis the project tasks for the system of current version. After that, we will be proceeding to the design phase. During the design phase, we will be designing the system of the current version by drawing the architecture pattern, activity diagram and class diagram to stimulate the real system application process. The system then will be developed and tested in the implementation phase. The development process of the system will be repeated for every version until the system is completely built.

Chapter 5

Design Specification

5.1 System Architecture

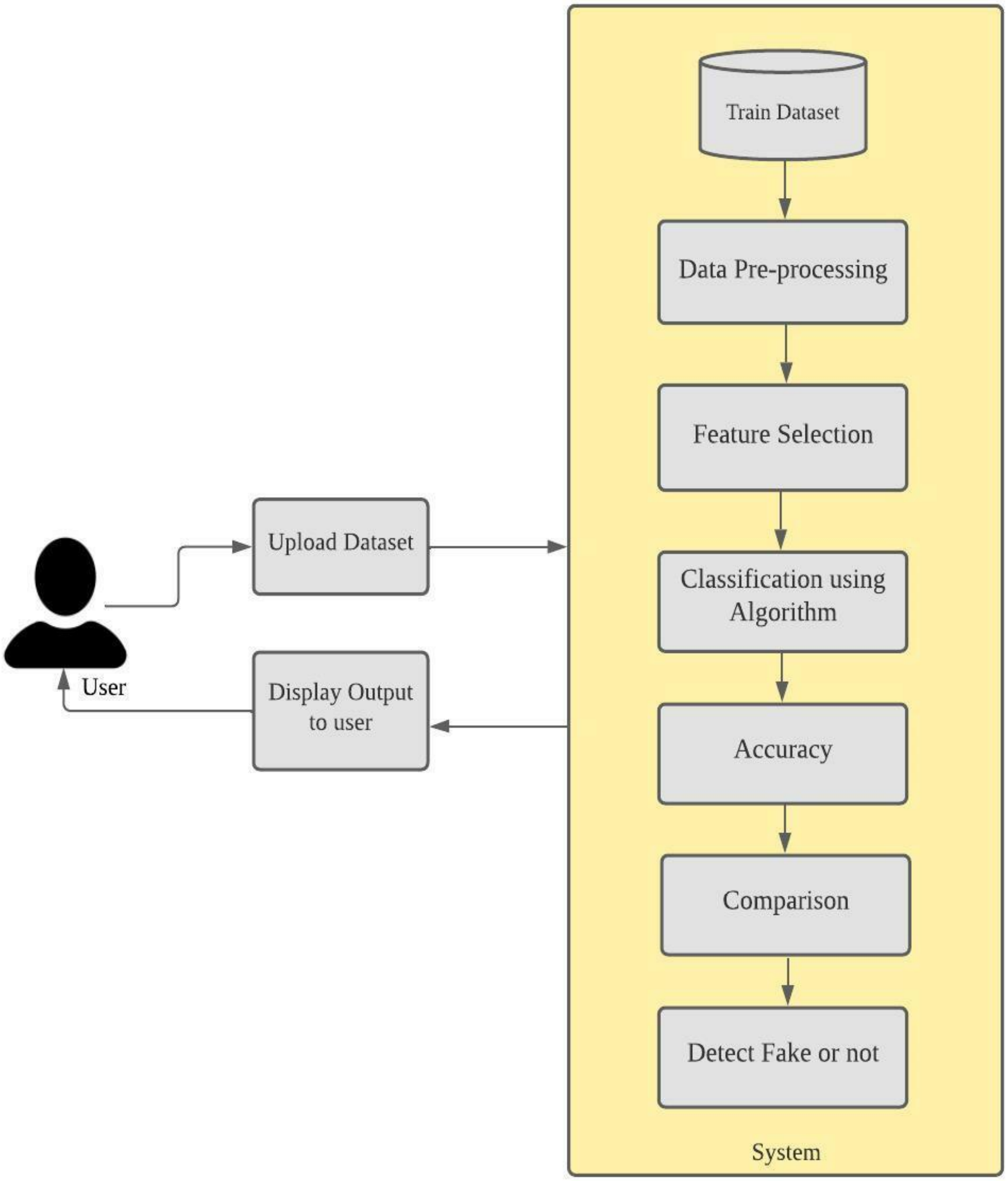


Figure 8:Architecture Diagram

5.2 Use Case Diagram

Here, Use case diagram is a dynamic diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use case diagram is used to show which operations are performed by the user and which operation are performed by the system. Use case diagrams are valuable for visualizing the functional requirements of a system that will translate into design choices and development priorities [5]. The following image shows the Use Case Diagram of the Fake Data Analysis.

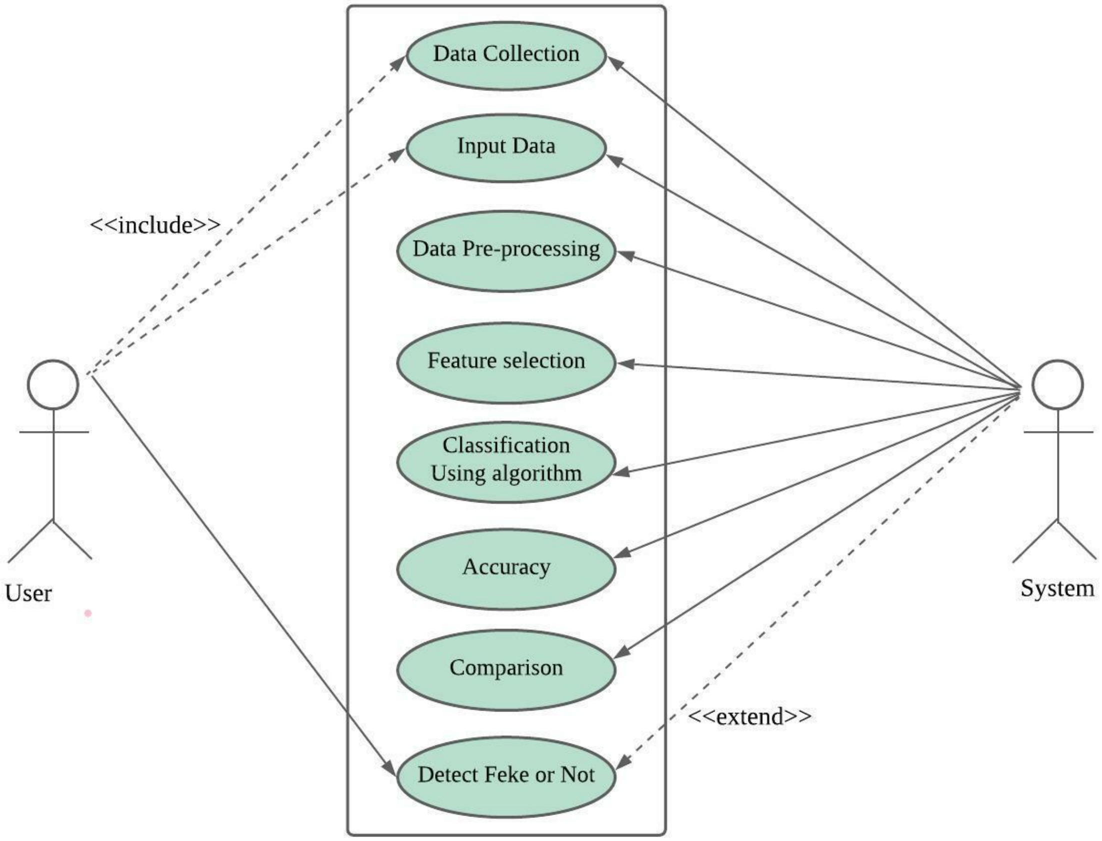


Figure 9:Architecture Diagram

5.3 Sequence Diagram

In sequence diagram bit by bit arrangement of steps is shown. In above chart first preprocess all train information and test information. Then, at that point, by applying the train information Train the machine and construct the module and at the last apply AI calculation on it.

For testing reason apply the test information on module and see the arrangement either phony or genuine [5]. The following image shows the Sequence Diagram of the Fake Data Analysis.

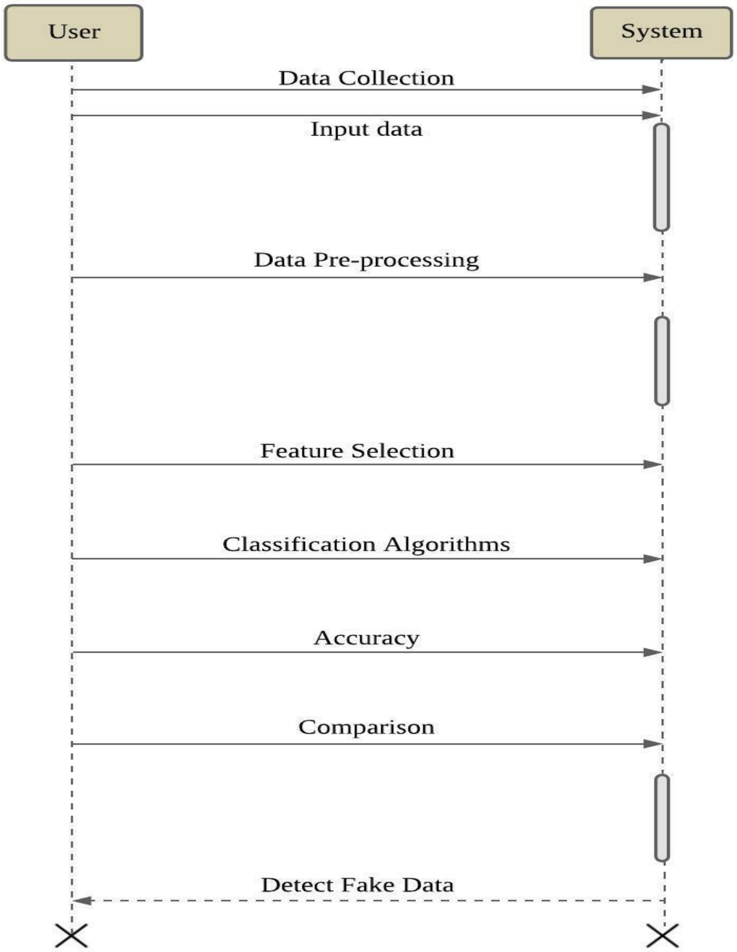


Figure 10:Sequence Diagram

5.4 Flow Chart

A flowchart is a diagram that depicts a process, system or computer algorithm. They are widely used in multiple fields to document, study, plan, improve and communicate often complex processes in clear, easy-to-understand diagrams [5]. The following image shows the Flow Chart of the Fake Data Analysis.

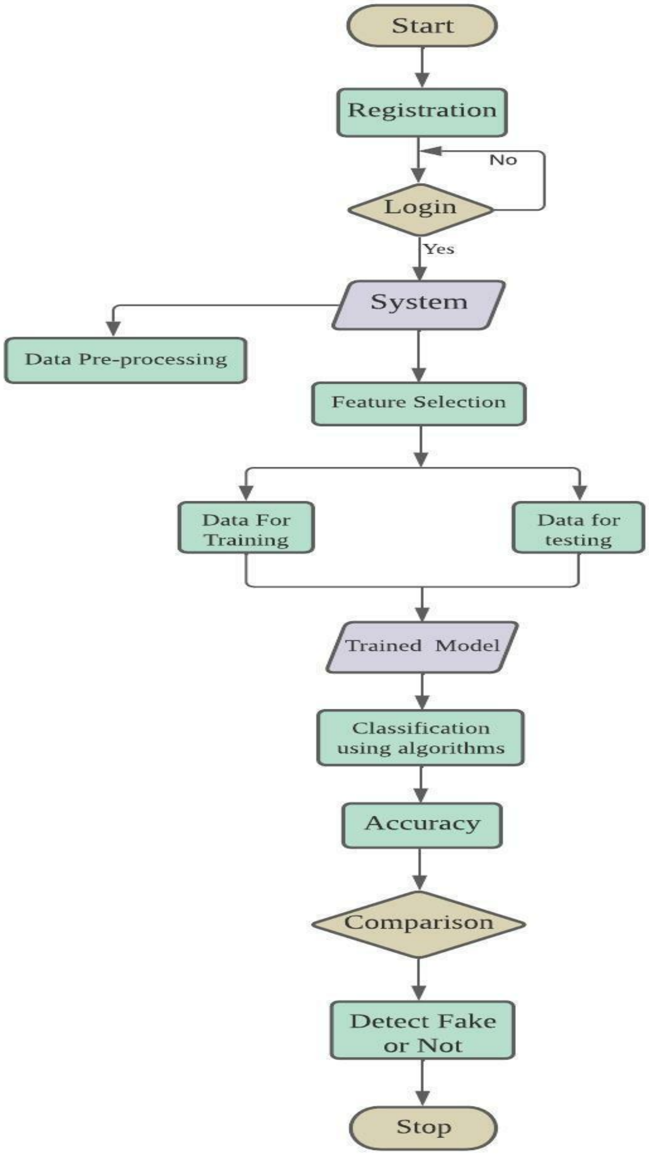


Figure 11:Activity Diagram

Chapter 6

Analysis of Algorithms

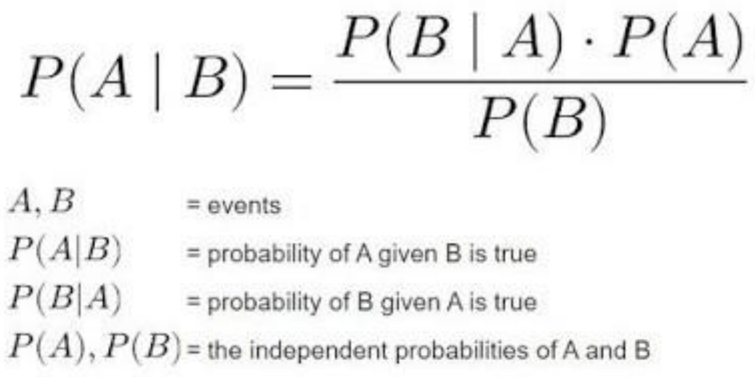
6.1 Data Pre-processing

The data needs to be pre-processed before the training, testing, and modeling phases. Before moving to these phases, the real news and fake news are concatenated.

After collecting the data from Dataset repository we performed a data preprocessing step. In this step we ran our code through all the data files present in the folders and collected the data relevant. With this we filtered out the data for this research only.

6.2 Multinomial Naive Bayes Algorithm

Naive Bayes: The Naive Bayes Classifier method depends on the Bayesian theorem and is especially fit when then, at that point, high layered information [6].



$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

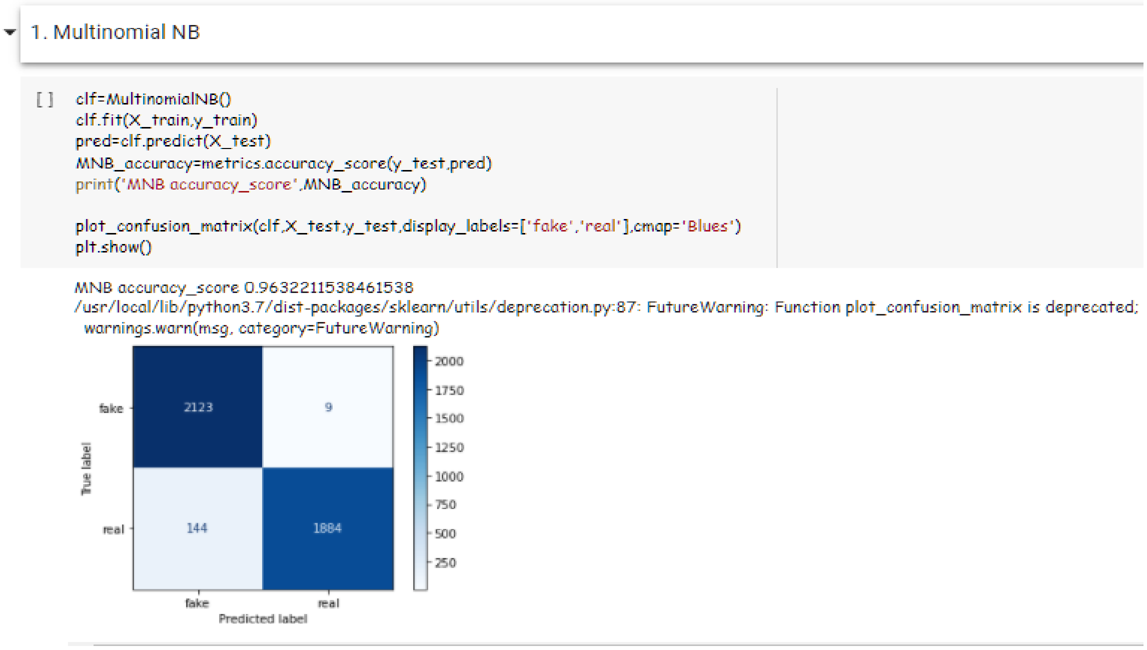
A, B	= events
$P(A\|B)$	= probability of A given B is true
$P(B\|A)$	= probability of B given A is true
$P(A), P(B)$	= the independent probabilities of A and B

It is used for classification when is in discrete structure. It is exceptionally helpful in text handling. Every text will be changed over to a vector of word count. It can't manage negative numbers.

It is predefined in Scikit Learn Library. So we can import that class in our undertaking then we make an object of Multinomial Naive Bayes Class.

1. Fit the classifier on our vectorized train information

2. Whenever the classifier fitted effectively on the preparation set then we can utilize the foresee technique to anticipate the outcome on the test set [6].



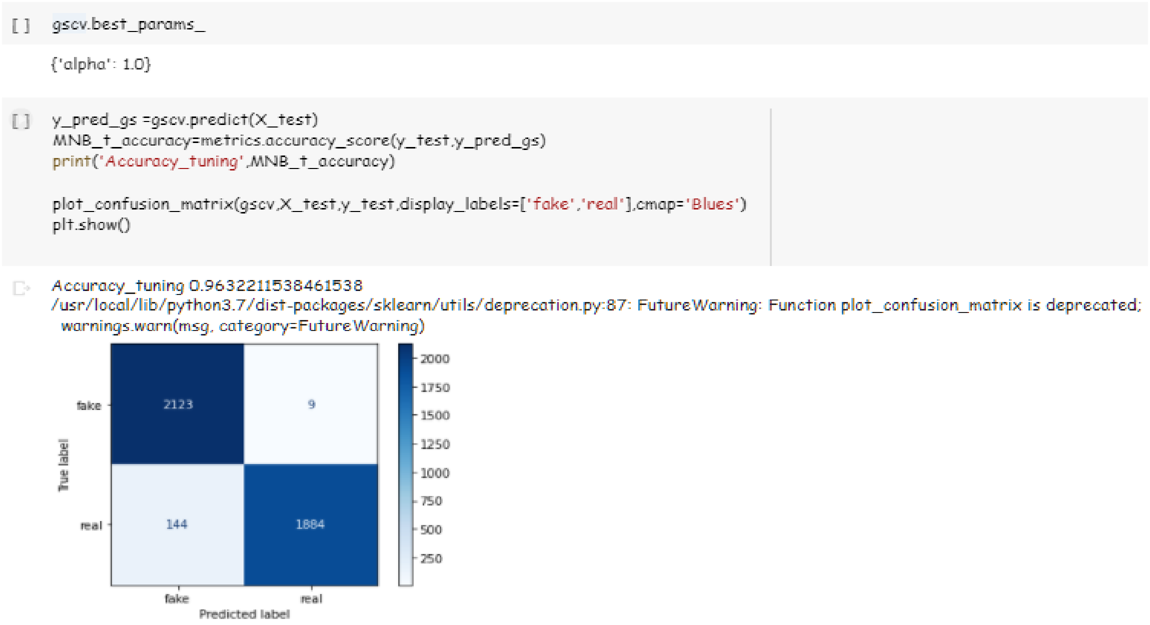
1. Multinomial NB

```
[ ] clf=MultinomialNB()
    clf.fit(X_train,y_train)
    pred=clf.predict(X_test)
    MNB_accuracy=metrics.accuracy_score(y_test,pred)
    print('MNB accuracy_score',MNB_accuracy)

    plot_confusion_matrix(clf,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
    plt.show()
```

```
MNB accuracy_score 0.9632211538461538
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
  warnings.warn(msg, category=FutureWarning)
```

Figure 12:Multinomial Naive Bayes Algorithm

6.2.1 Multinomial Naive Bayes Algorithm with Hyper Parameter



```
[ ] gscv.best_params_

    {'alpha': 1.0}
```

```
[ ] y_pred_gs =gscv.predict(X_test)
    MNB_t_accuracy=metrics.accuracy_score(y_test,y_pred_gs)
    print('Accuracy_tuning',MNB_t_accuracy)

    plot_confusion_matrix(gscv,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
    plt.show()
```

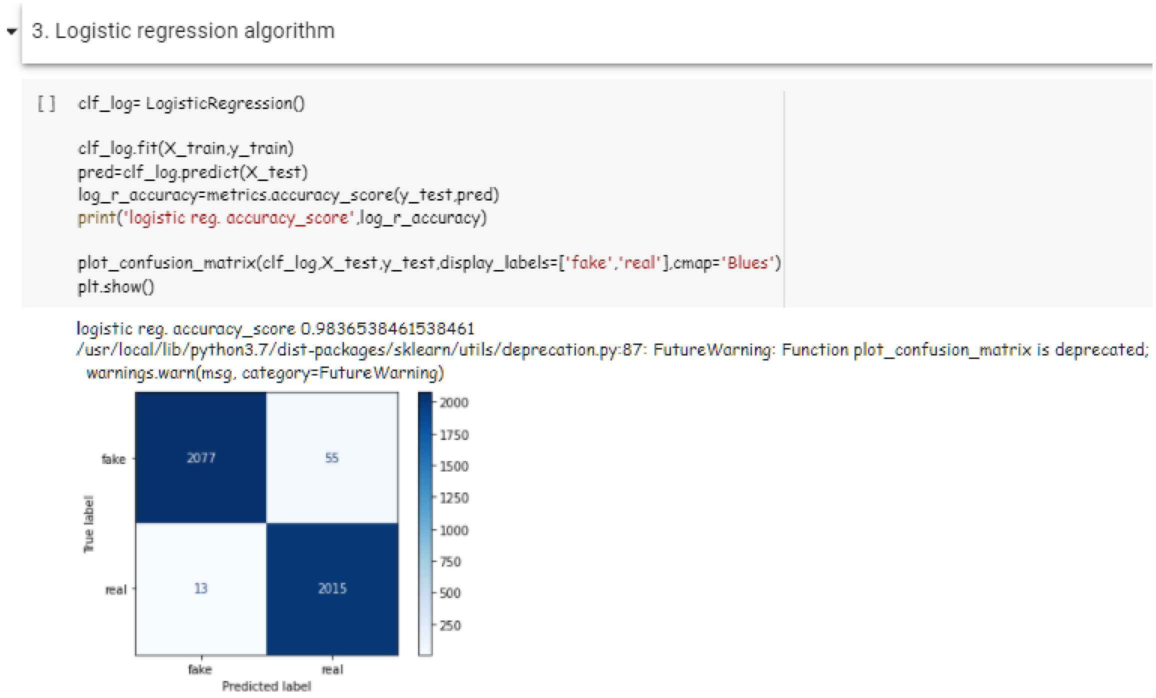
```
Accuracy_tuning 0.9632211538461538
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
  warnings.warn(msg, category=FutureWarning)
```

Figure 13:Hyper Parameter

6.3 Logistic Regression Algorithm

The third procedure that we are utilizing to make this model work accurately is the Logistic regression algorithm. Logistic regression in machine learning directs that calculated relapse can find an association among the features and probability of a particular outcome. A logistic regression classifier is utilized while the anticipating esteem is clear cut. For example, while anticipating the worth, it will give either a valid or misleading reaction. Logistic regression can find an association among the features and probability of a particular outcome. The Logistic regression model can be imported from the SKLEARN linear model [7].

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$



3. Logistic regression algorithm

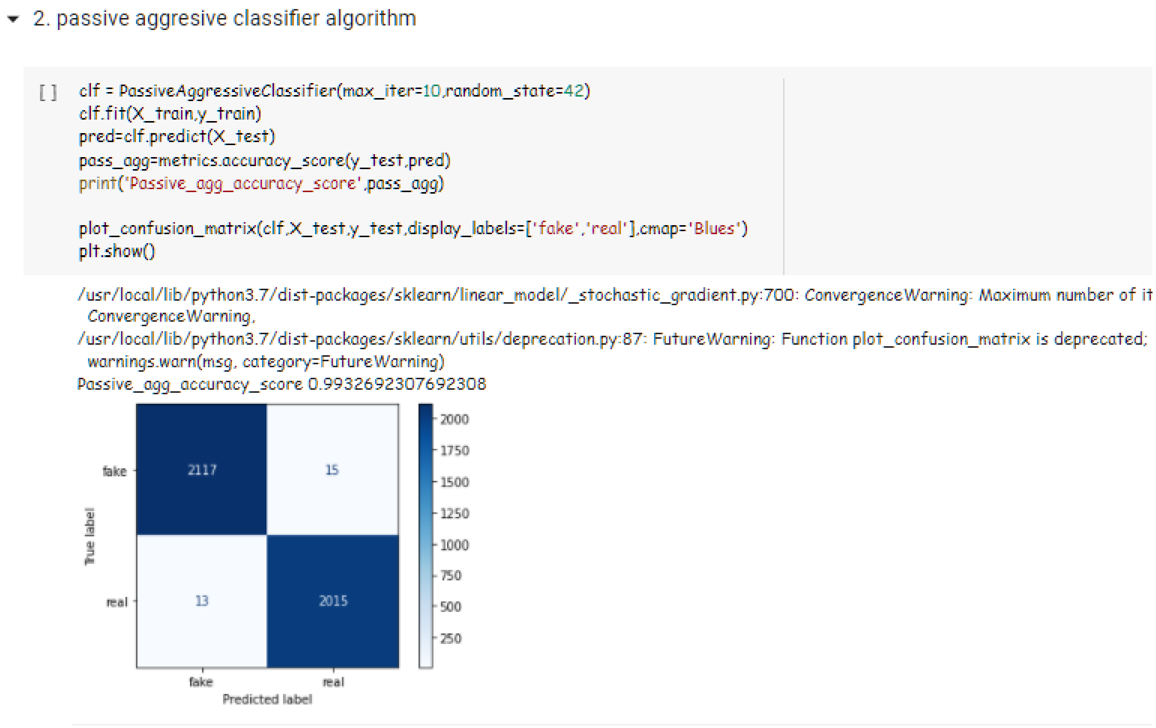
```
[ ]  clf_log= LogisticRegression()

     clf_log.fit(X_train,y_train)
     pred=clf_log.predict(X_test)
     log_r_accuracy=metrics.accuracy_score(y_test,pred)
     print('logistic reg. accuracy_score',log_r_accuracy)

     plot_confusion_matrix(clf_log,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
     plt.show()
```

```
logistic reg. accuracy_score 0.9836538461538461
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
  warnings.warn(msg, category=FutureWarning)
```

Figure 14:Logistic Regression Algorithm

6.4 Passive Aggressive Classifier Algorithm

Passive Aggressive calculations are internet learning calculations. Such an algorithm remains passive for a right order result, and turns forceful in case of an error, refreshing and changing. Not at all like most different calculations, it doesn't meet. Its motivation is to make refreshes that right the misfortune, causing next to no adjustment of the standard of the weight vector [7].

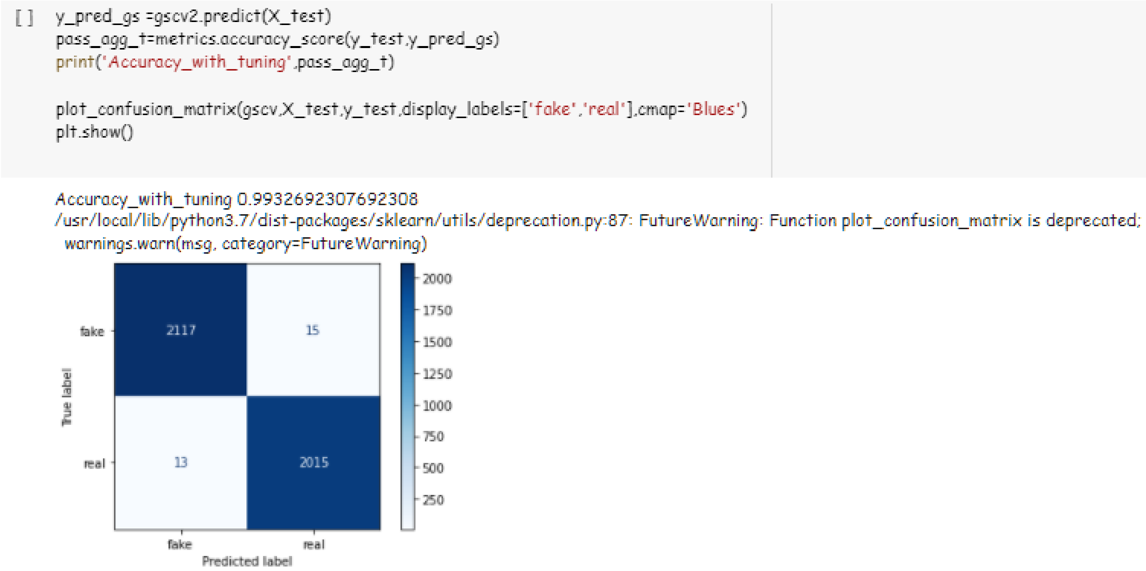


```
clf = PassiveAggressiveClassifier(max_iter=10,random_state=42)
clf.fit(X_train,y_train)
pred=clf.predict(X_test)
pass_agg=metrics.accuracy_score(y_test,pred)
print('Passive_agg_accuracy_score',pass_agg)

plot_confusion_matrix(clf,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_stochastic_gradient.py:700: ConvergenceWarning: Maximum number of it
  ConvergenceWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
  warnings.warn(msg, category=FutureWarning)
Passive_agg_accuracy_score 0.9932692307692308
```

Figure 15:Passive Aggressive Classifier Algorithm

6.4.1 Passive Aggressive Classifier Algorithm with tuning



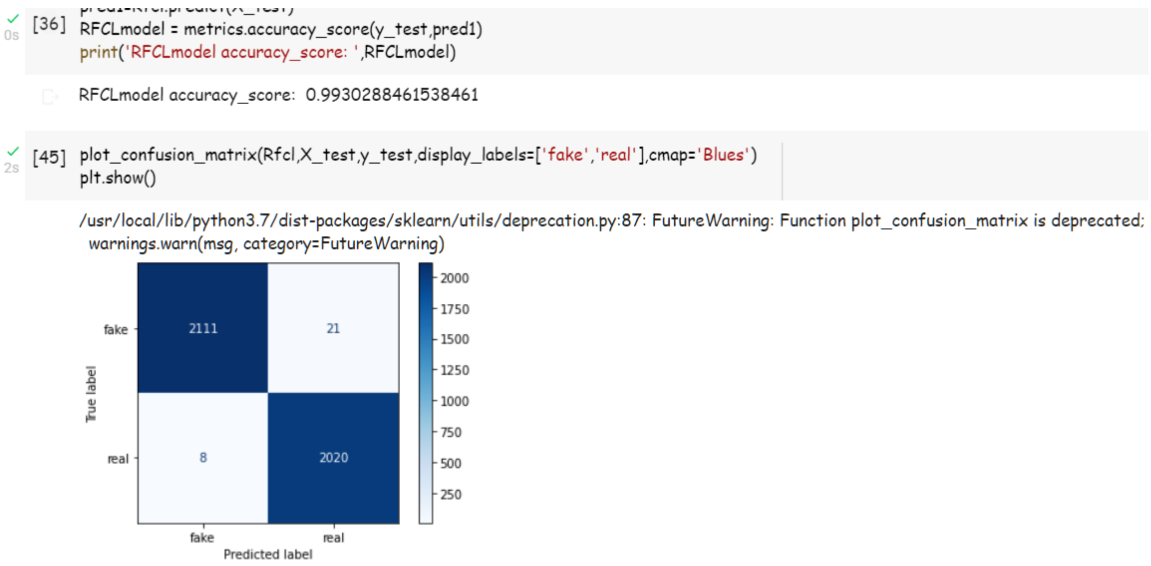
```
[ ]  y_pred_gs =gscv2.predict(X_test)
     pass_agg_t=metrics.accuracy_score(y_test,y_pred_gs)
     print('Accuracy_with_tuning',pass_agg_t)

     plot_confusion_matrix(gscv,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
     plt.show()
```

```
Accuracy_with_tuning 0.9932692307692308
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
  warnings.warn(msg, category=FutureWarning)
```

Figure 16:Passive Aggressive Classifier Algorithm with tuning

6.5 Random Forest Classifier

The random forest has almost the same hyper parameters as a decision tree or a sacking classifier. This technique adds more arbitrariness to the model while developing the trees. First of all, a random forest classifier is a technique that makes different choice trees and consolidates them to produce a more exact and stable prediction. The random forest has hyper parameters that are almost the same as a decision tree or a sacking classifier. This technique adds more arbitrariness to the model while developing the trees [1].

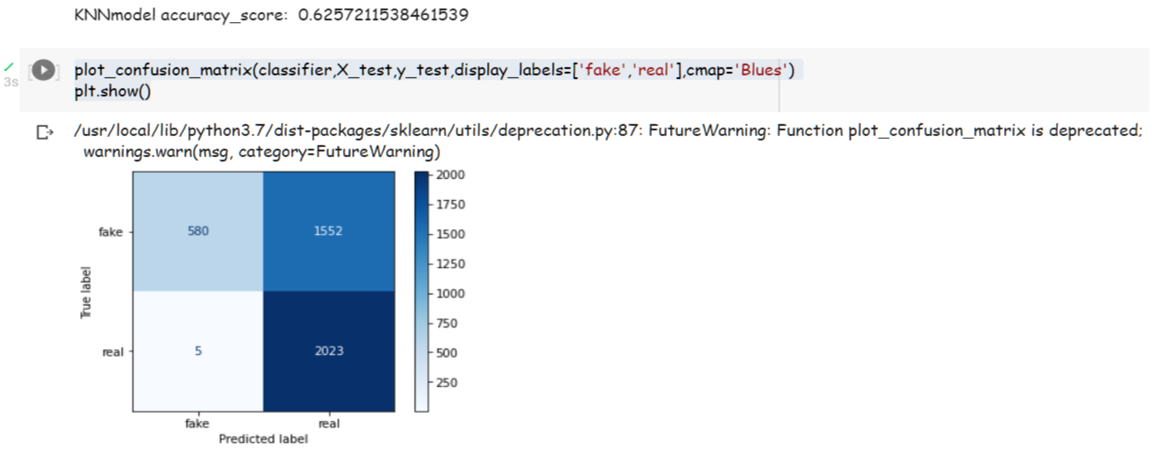


```
pred1=Rfcl.predict(X_test)
[36] RFCLmodel = metrics.accuracy_score(y_test,pred1)
     print('RFCLmodel accuracy_score: ',RFCLmodel)

     RFCLmodel accuracy_score:  0.9930288461538461

[45] plot_confusion_matrix(Rfcl,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
     plt.show()

     /usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
       warnings.warn(msg, category=FutureWarning)
```

Figure 17:Random Forest Classifier

6.5 K Nearest Neighbor Algorithm

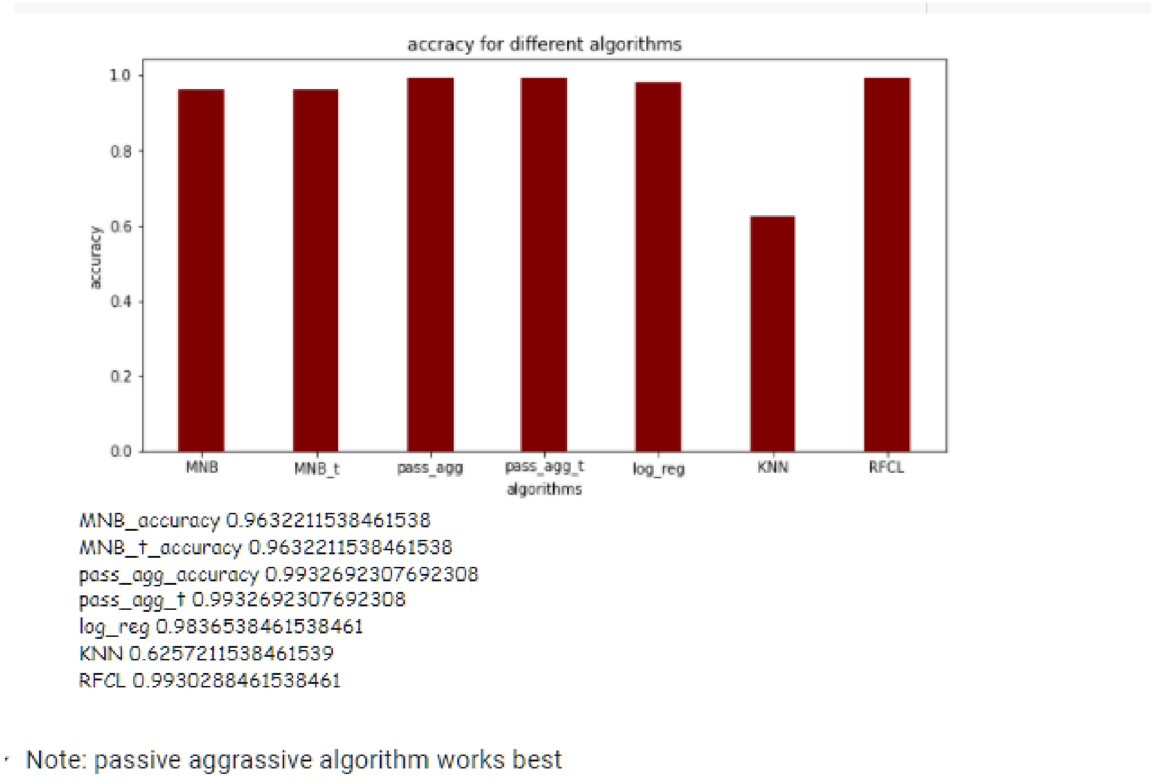
The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows [3].



```
KNNmodel accuracy_score:  0.6257211538461539

plot_confusion_matrix(classifier,X_test,y_test,display_labels=['fake','real'],cmap='Blues')
plt.show()

/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function plot_confusion_matrix is deprecated;
  warnings.warn(msg, category=FutureWarning)
```

Figure 18:K Nearest Neighbor Algorithm

6.6 Comparison

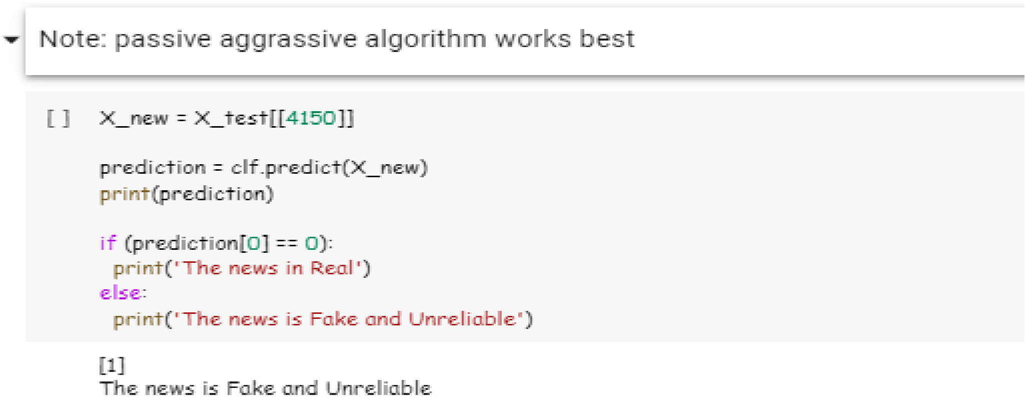


```
MNB_accuracy 0.9632211538461538
MNB_t_accuracy 0.9632211538461538
pass_agg_accuracy 0.9932692307692308
pass_agg_t 0.9932692307692308
log_reg 0.9836538461538461
KNN 0.6257211538461539
RFCL 0.9930288461538461
```

Note: passive aggrassive algorithm works best

Figure 19:Comparison

As we got the highest accuracy from Passive Aggressive Algorithm. Using this algorithm, we detect the data whether it is fake or real. The following figure is given below:



Note: passive aggrassive algorithm works best

```
[ ]  X_new = X_test[[4150]]

     prediction = clf.predict(X_new)
     print(prediction)

     if (prediction[0] == 0):
       print('The news in Real')
     else:
       print('The news is Fake and Unreliable')

     [1]
     The news is Fake and Unreliable
```

Chapter 7

Implementation

7.1 Python

Python is a high-level, object-oriented programming language that is easy to use, simple, and can be easily interpreted. Python Syntax which illustrates clarity and readability aims to assist programmers to write logical and clear code for large- and small-scale projects and also minimizes the maintenance cost of the software. Python provides packages and modules which enhance the reuse of codes. Python is however perfect for Artificial Intelligence Projects and Machine learning due to its consistency, simplicity, and its access to outstanding libraries and frameworks for Artificial Intelligence and Machine Learning [9].

7.2 Libraries

The flexibility of Python has motivated a lot of developers to develop new libraries, which has made Python quite popular amongst most Machine learning professionals. Some of the Machine learning libraries used in the implementation of this model include Sklearn, Matplotlib, pandas, NumPy, NLTK libraries [10].

SKlearn: Also known as Scikit-learn is an open-source python machine learning which is built on Matplotlib, NumPy, and Scipy by introducing algorithms for classification, clustering, and regression. Sklearn has a lot of tools that can be used for data analysis.

Matplotlib: Matplotlib is an interactive and very common python library that can be used across different platforms. It can be used for two-dimensional plotting and visualization. It helps with the visualization of patterns in a dataset, it offers various forms of visualization in form of plots and graphs such as bar charts, histograms, error charts, pie charts, etc [10].

Pandas: Pandas is an easy-to-use high-performance open-source python library with a variety of data analysis, data structure, and data manipulation tools for python language. Pandas can be used to data from different sources such as SQL Database, CSV, Excel, and JSON files.

NumPy: Also known as Numerical Python is one of the best in terms of science and mathematical programming library. It uses matrix processing and multi-dimensional array with high-level mathematical models. It is majorly used for computational analysis which makes it one of the most used Machines learning Python packages [9].

NLTK: Natural Language Toolkit (NLTK) NLTK is an essential library supports tasks such as classification, stemming, tagging, parsing, semantic reasoning, and tokenization in Python. It's basically your main tool for natural language processing and machine learning [10].

7.3 Flask

Flask is a micro web structure written in Python. It is named a micro framework on the grounds that it doesn't need specific instruments or libraries. It has no data set reflection layer, structure approval, or whatever other parts where previous outsider libraries give normal capacities. However, Flask upholds augmentations that can add application highlights as though they were executed in Flask itself. Expansions exist for object-social mappers, structure approval, transfer dealing with, different open confirmation innovations and a few normal systems related instruments [9].

7.4 Jupyter Notebook

Jupyter Notebook is an open-source web application that enables the sharing and creating of documents that contain live code, visualizations, equations, and narrative text. It is used to transform and clean data, statistical modeling, numerical simulation, and machine learning. It supports over 40 programming languages such as Python. Scala, R, and Julia [9].

Chapter 8

Screenshots of the project

In our Website at first we can see a Home page. In this page four NAV link content available. If we press any NAV link, we will go to the particular section. Also, we can see a prediction button in which by clicking we will move to the prediction page and be able to identify the data weather it is real or fake.

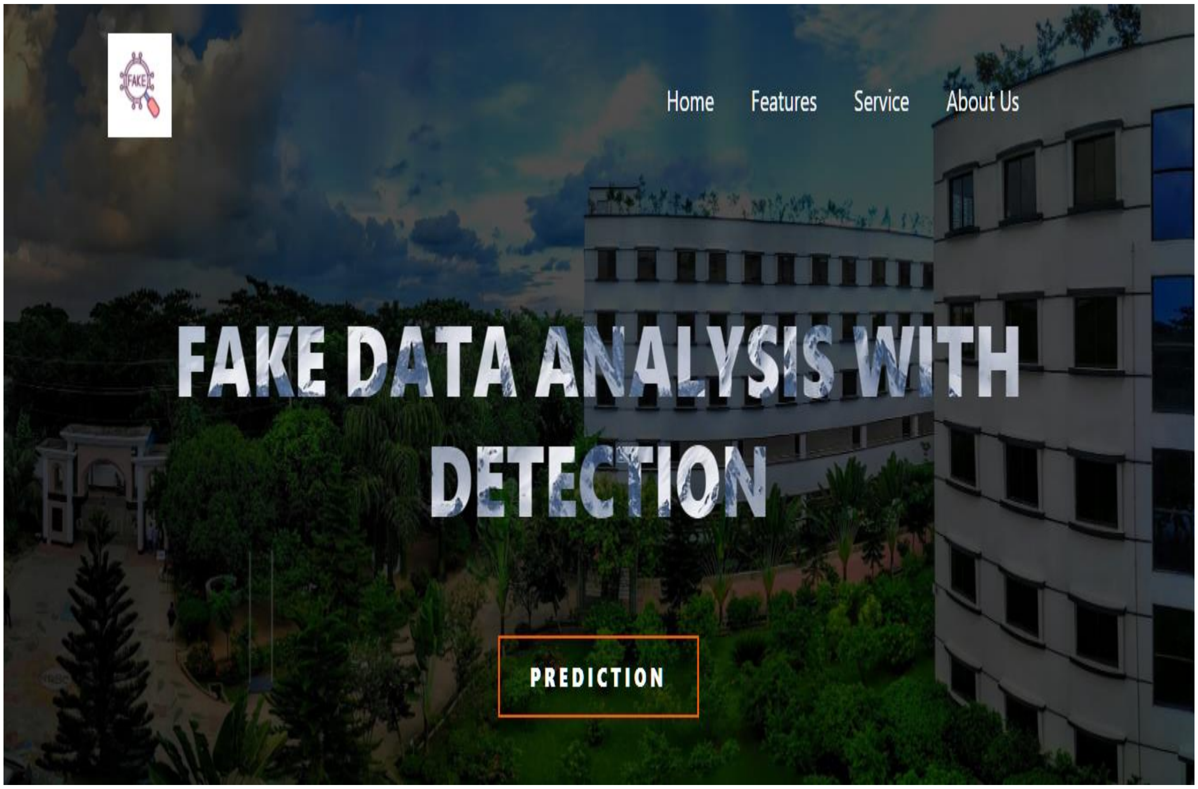
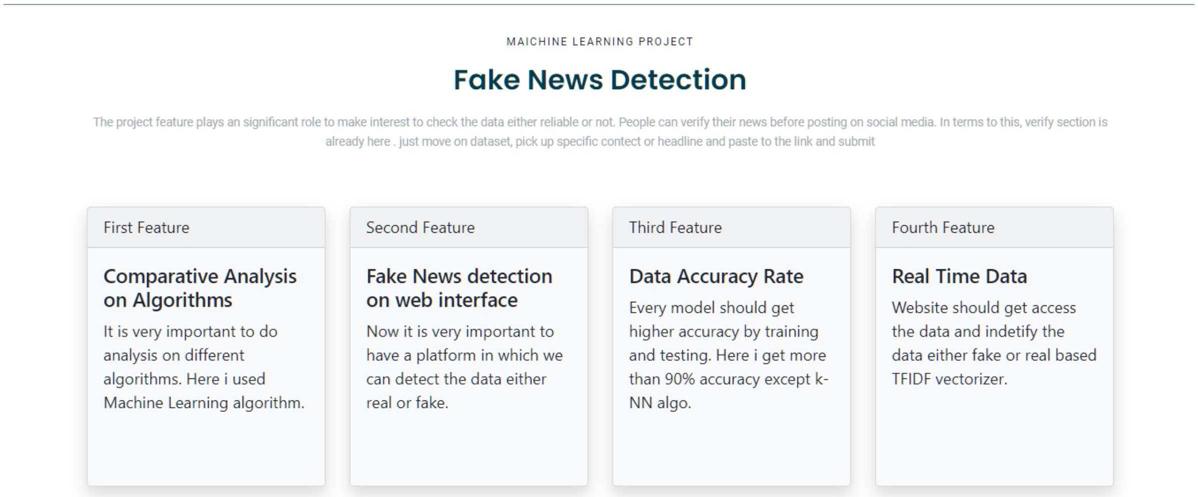


Figure 20:Home Page in Fake Data Analysis Web app

Here this part we put four features with shot description.

- **Comparative Analysis on Algorithms**
  Comparative analysis refers to the correlation of at least two cycles, archives, informational collections or different items. Design investigation, sifting and choice tree examination are types of similar examination.
- **News Detection on web interface**
  The spread of phony news is increasing day by day. On Twitter, Facebook, Reddit, individuals exploit counterfeit word to get out reports, win political advantages and snap rates.

  Identifying counterfeit news is basic for a sound society, and there are various ways to deal with distinguish counterfeit news. From an AI angle, counterfeit news identification is a parallel arrangement issue. For perspective of this angle we predict the news either real or fake on our website.
- **Data Accuracy Rate**
  The results were evaluated through a confusion matrix and a Scikit library classification. Firstly, the dataset dived into two portion one is training and another testing then TF-IDF vectorizer was evaluated on the test dataset. Secondly, different algorithm was evaluated based on the test dataset. We used five algorithms. such as
  - Multinomial Naive Bayes: The model was performed with 96.32% accuracy. The model was able to determine a total of 2123 fake news instances and 1884 real news instances.
  - Passive Aggressive Classifier: The model was performed with 99.32% accuracy. The model was able to determine a total of 2117 fake news instances and 2015 real news instances.
  - Logistic Regression: The model was performed with 98.36% accuracy. The model was able to determine a total of 2077 fake news instances and 2015 real news instances.
  - Random Forest Classifier: The model was performed with 99.30% accuracy. The model was able to determine a total of 2111 fake news instances and 2020real news instances.
  - K Nearest Neighbor: The model was performed with 62.52% accuracy. The model was able to determine a total of 580 fake news instances and 2023 real news instances. Among them 1052 data were miss calculated.
- **Real time data**
  For now, all data that is consist on our data set only these are predicted by our model. Whenever we take any real time data on social media, email, newspaper etc. and predict them in our UI then we get fake news because we put this portion as our future goal.



MAICHINE LEARNING PROJECT

Fake News Detection

The project feature plays an significant role to make interest to check the data either reliable or not. People can verify their news before posting on social media. In terms to this, verify section is already here . just move on dataset, pick up specific contect or headline and paste to the link and submit

First Feature

Comparative Analysis on Algorithms

It is very important to do analysis on different algorithms. Here i used Machine Learning algorithm.

Second Feature

Fake News detection on web interface

Now it is very important to have a platform in which we can detect the data either real or fake.

Third Feature

Data Accuracy Rate

Every model should get higher accuracy by training and testing. Here i get more than 90% accuracy except k-NN algo.

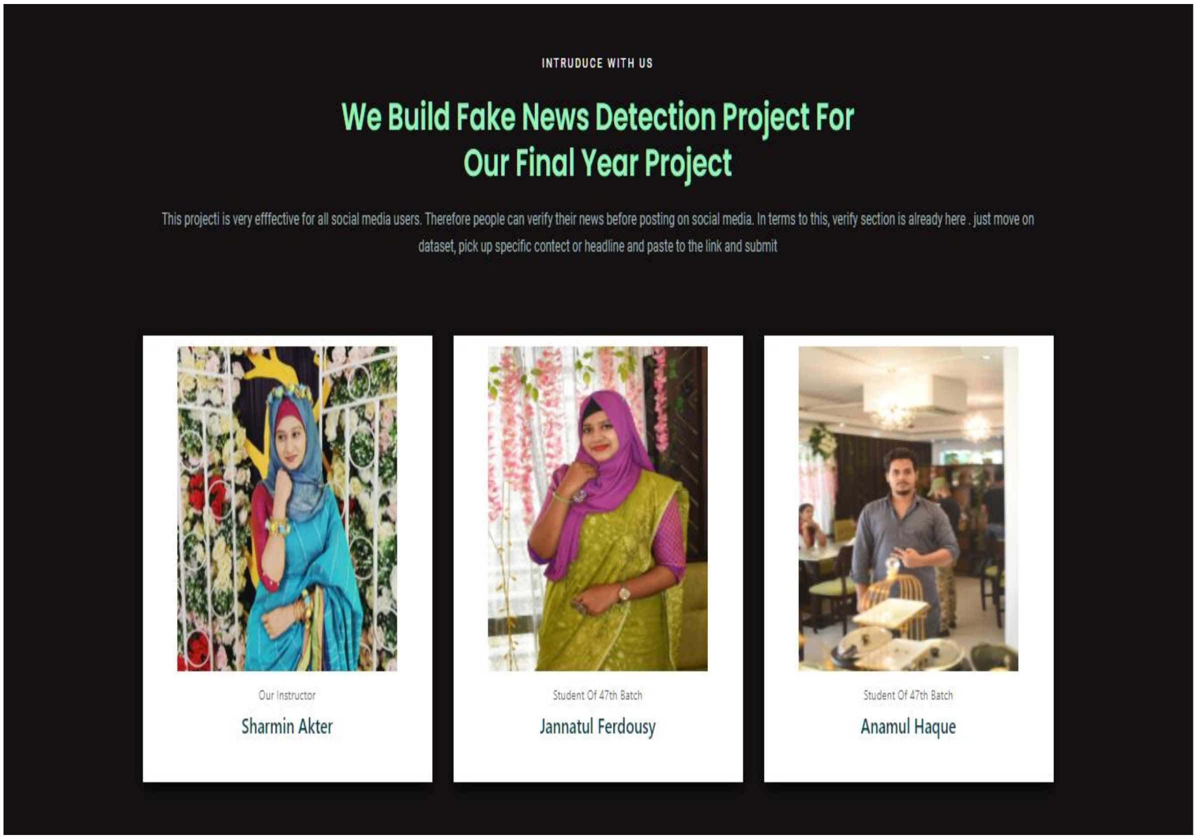
Fourth Feature

Real Time Data

Website should get access the data and indetify the data either fake or real based TFIDF vectorizer.

Figure 21:Features

Here our team with honorable supervisor **"Sharmin Akter"**



INTRUDUCE WITH US

We Build Fake News Detection Project For Our Final Year Project

This projecti is very efffective for all social media users. Therefore people can verify their news before posting on social media. In terms to this, verify section is already here . just move on dataset, pick up specific contect or headline and paste to the link and submit

Our Instructor

Sharmin Akter

Student Of 47th Batch

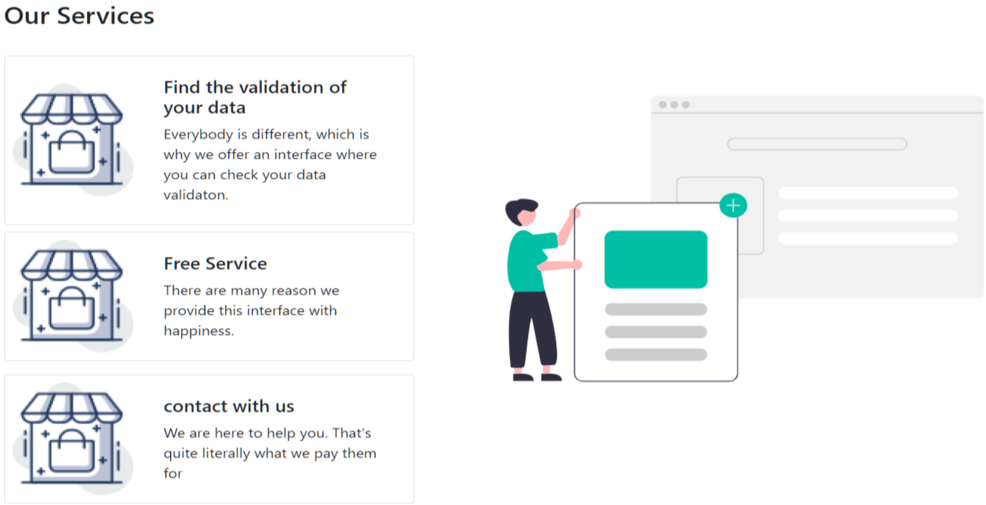
Jannatul Ferdousy

Student Of 47th Batch

Anamul Haque

Figure 22:Project Team

In this section Users can get touch on our services. Each services will be provided very firstly.



Our Services

Find the validation of your data

Everybody is different, which is why we offer an interface where you can check your data validaton.

Free Service

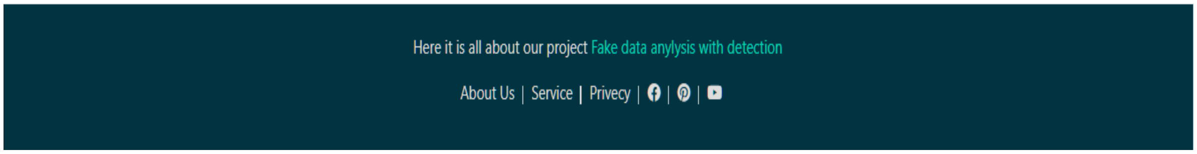
There are many reason we provide this interface with happiness.

contact with us

We are here to help you. That's quite literally what we pay them for

Figure 23: Services

This is our web page Footer part.



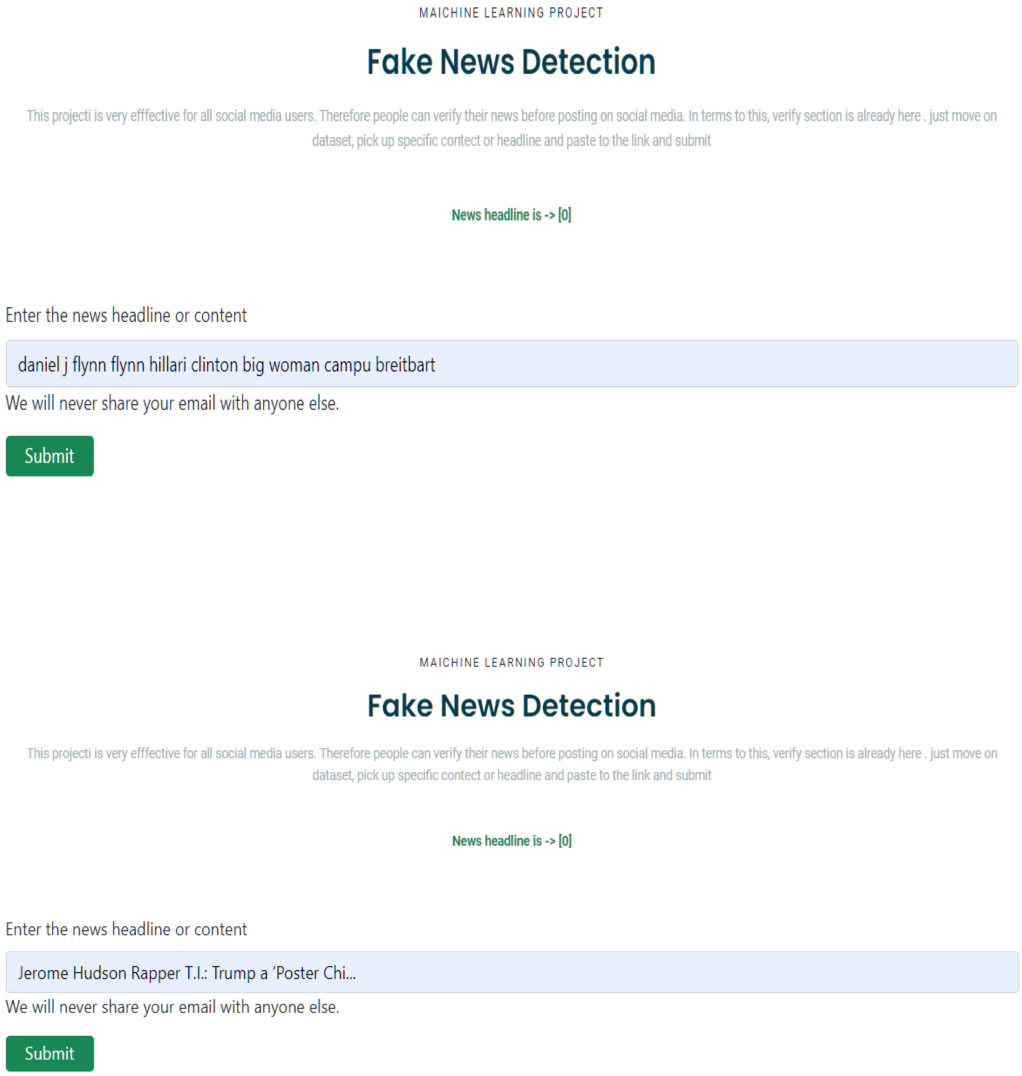
Here it is all about our project Fake data anylysis with detection

About Us | Service | Privecy | | |

Figure 24:Footer Part

This portion provide for detect the news title, social media headline weather real or fake. If the news headline shows zero (0) that means the data is real, otherwise its fake when it shows one (1).

Note: All data must be contained in data set.



MAICHINE LEARNING PROJECT

Fake News Detection

This projecti is very efffective for all social media users. Therefore people can verify their news before posting on social media. In terms to this, verify section is already here . just move on dataset, pick up specific contect or headline and paste to the link and submit

News headline is -> [0]

Enter the news headline or content

daniel j flynn flynn hillari clinton big woman campu breitbart

We will never share your email with anyone else.

MAICHINE LEARNING PROJECT

Fake News Detection

This projecti is very efffective for all social media users. Therefore people can verify their news before posting on social media. In terms to this, verify section is already here . just move on dataset, pick up specific contect or headline and paste to the link and submit

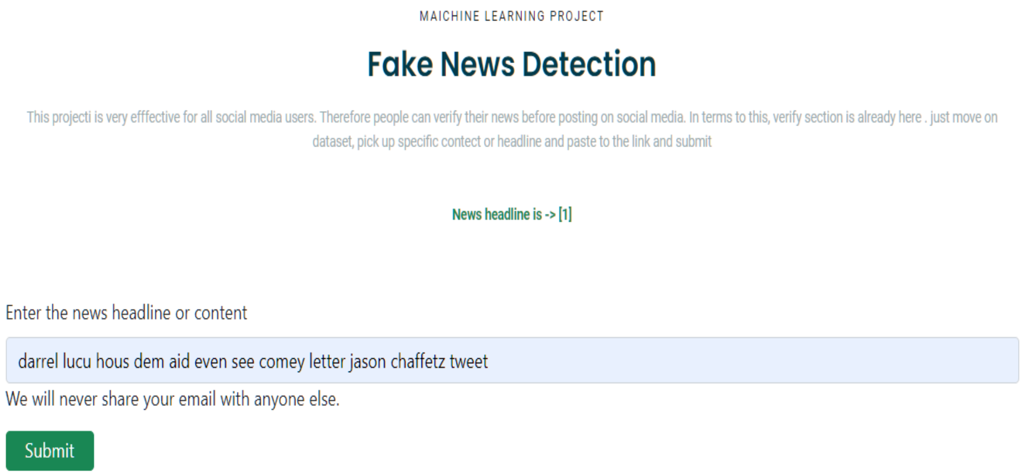
News headline is -> [0]

Enter the news headline or content

Jerome Hudson Rapper T.I.: Trump a 'Poster Chi...

We will never share your email with anyone else.

Figure 25:Real News Detected



MAICHINE LEARNING PROJECT

Fake News Detection

This projecti is very efffective for all social media users. Therefore people can verify their news before posting on social media. In terms to this, verify section is already here . just move on dataset, pick up specific contect or headline and paste to the link and submit

News headline is -> [1]

Enter the news headline or content

darrel lucu hous dem aid even see comey letter jason chaffetz tweet

We will never share your email with anyone else.

Figure 26:Fake News Detected

8.1 Target user

Our targeted user are all kinds of Social Media users in the world & any Online News Papers readers can use it and will get benefited. The content writer and the vlogger's also use this site for the real information.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

In this project, we have discussed about classifying fake news articles with the help of Machine Learning and natural language processing. We also have proposed prediction model which gives the accuracy rate more than 90% and it cover all important news. In this paper, several experiments will perform and tested to evaluate the efficiency and the performance by following machine learning classifiers. Several performance metrics are computed (accuracy rate, precision, false negative, false positive, true negative and true positive). The methodology of this project is incremental development method. Specifically, the phased development will be used to develop the Fake Data Analysis. All the collaborating functions will be implemented in the Fake Data Analysis with the use of Machine Learning, Natural Language Processing so many thinks.

9.2 Future Work

For future work, more in-depth research that focuses on genetic machine learning algorithms with more complex feature selection, which would give a better result, even with the work on Natural Language Processing will be get the best result. To get better performance we will move to the deep learning. we will must ensure that real time data will be checked definitely. Not only this features we will bring but also authenticity will be added for security purposes [9].

REFERENCES

1. P. B. P. Reddy, M. P. K. Reddy, G. V. M. Reddy and K. M. Mehata, "Fake Data Analysis and Detection Using Ensembled Hybrid Algorithm," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 890-897, doi: 10.1109/ICCMC.2019.8819741.
2. S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," 2017 IEEE 15th Student Conference on Research and Development (SCOReD), 2017, pp. 110-115, doi: 10.1109/SCORED.2017.8305411.
3. M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), 2017, pp. 000277-000282, doi: 10.1109/SISY.2017.8080566.
4. A. Uppal, V. Sachdeva and S. Sharma, "Fake news detection using discourse segment structure analysis," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 751-756, doi: 10.1109/Confluence47617.2020.9058106.
5. “IJASRET”, [Online] Available: http://www.ijasret.com/ [Last Accessed on 16-02-2022 at 9.30 pm]
6. “Medium”, [Online] Available: https://medium.com/ [Last Accessed on 21-02-2022 at 10.00 pm]
7. “Towards Data Science”, [Online] Available: https://towardsdatascience.com/ [Last Accessed on 22-02-2022 at 9.00 pm]
8. “Journal of Machine Learning Research”, [Online] Available: https://www.jmlr.org/papers/volume21/17-360/17-360.pdf [Last Accessed on 22-03-2022 at 9.00 am]
9. “Developers”, [Online] Available: https://www.toptal.com/python/python-machine-learning-flask-example [Last Accessed on 25-03-2022 at 5.00 pm]
10. “Medium One”, [Online] Available: https://medium.com/@Med1um1/using-matplotlib-in-jupyter-notebooks-comparing-methods-and-some-tips-python-c38e85b40ba1 [Last Accessed on 25-03-2022 at 5.00 pm]