SUMMER TRAINING REPORT

ON

FAKE NEWS DETECTION SYSTEM

Submitted to Guru Gobind Singh Indraprastha University, Delhi(India)

In partial fulfilment of the requirement for the reward of the degree of

BTECH

IN

INFORMATION AND TECHNOLOGY

Submitted by:

Manas Lamba

ROLL NO:00515003121

DEPT. OF INFORMATION AND TECHNOLOGY

MAHARAJA SURAJMAL INSTITUTE OF TECHNOLOGY

NEW DELHI-110058

OCTOBER 2023

ACKNOWLEDGEMENT

A research work owes its success from commencement to completion to the people in love with researchers at various stages. Let me in this page express my gratitude to all those who helped us in various stages of this study. First, I would like to express my sincere gratitude indebtedness to **Dr . Tripti Sharma** (HOD, Department of Information Technology, Maharaja Surajmal Institute of Technology ,New Delhi) for allowing me to undergo the summer training of 4 weeks at DCIT.

I am grateful to our guide Ms. Nidhi, for the help provided in completion for the help provided in completion of the project, which was assigned to me. Without his friendly help and guidance it was difficult to develop this project.

Submitted By:

Manas lamba

(Roll No.:00515003121)

CERTIFICATE

This is to verify that Mr. Manas Lamba of Bachelor of Information Technology has completed Summer Training on the topic Machine Learning and Artificial Intelligence from **DICT** as Partial fulfilment of Bachelor of Engineering IT. The summer Training report and presentation by him is genuine work done by him and the same is being submitted for evaluation.

Signature

CANDIDATE'S DECLARATION

I, Abhishek, Roll No. 00515003121, B.Tech(Semester-5th) of the Maharaja Surajmal Institute Of Technology, New Delhi declare that the Training Report entitled "**ML/AI**" is An original work and data provided in the study is authentic to the best of my knowledge. This report has not been submitted to any other Institute for any other degree.

Manas Lamba

(00515003121)

Place:

Date:

ABSTRACT OF PROJECT

Fake News Detection System is a ML model created to classify whether a certain article represents a Fake or Real news as we can see that fake news has become a real nuisance in

The age of Internet so we can use this model to identify whether to identify whether this news article is providing us real or fake news.

ORGANISATION INTRODUCTION

Founded in 2001, DICT today is one of the most reputed institution in the country accredited under NIELIT Scheme and conducting courses in Computer Science from 'O' (Foundation) to 'A'(Advance Diploma.) levels. DICT's social sprit and innovative approach to academics are the foundations upon which the institute has established itself as a premier global resource for IT man power. Through a commitment to its fundamental principles, it will continue to produce technicians who can invent the best IT solutions. DICT is one of the pioneers in IT training and education with about Two decade of commendable presence in the Industry. The organization has been making steady progress in diverse areas of IT Services and Software Industry. From its inception, the organization has been striving to maintain its high standards in all its endeavours.

CONTENT

Chapter 1

Technology Used

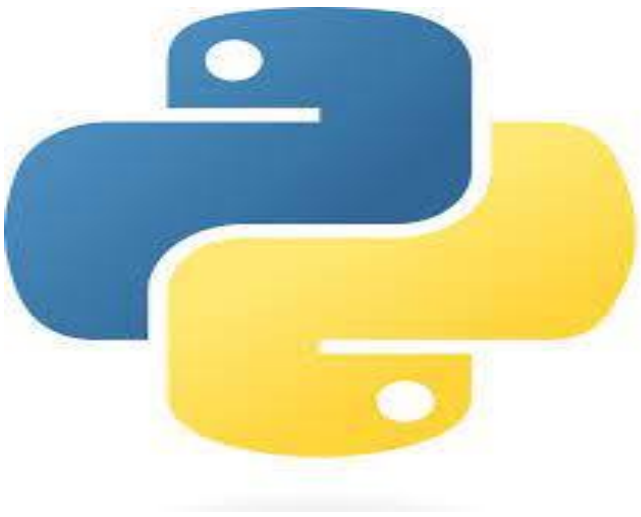
Python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000. Python 3.0, released in 2008, was a major revision not completely backward-compatible with earlier versions. Python 2.7.18, released in 2020, was the last release of Python 2.

Python consistently ranks as one of the most popular programming languages.



Python

Python for ML

Python today is used for various ML project because of it's simple syntax and wide varieties of libraries not only that python's compilation process makes it more suitable for making ML models .

Some of the libraries used for making ML models are:

1.Pandas: For editing datasets.

2.Matplotlib: For Data visualization

3.Ski-learn : Used for pre -processing of data, creating module and properly analyses the efficiency of model.

4.StatsModel.api: Used for creating, understanding statistics and improving the given model

Pandas

Pandas is an open-source library in Python that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library of Python. Pandas is fast and it has high performance & productivity for users.

History of Pandas Library

Pandas were initially developed by Wes McKinney in 2008 while he was working at AQR Capital Management. He convinced the AQR to allow him to open source the Pandas. Another AQR employee, Chang She, joined as the second major contributor to the library in 2012. Over time many versions of pandas have been released. The latest version of the pandas is 1.5.3, released on Jan 18, 2023.

Pandas

Why Use Pandas?

- Fast and efficient for manipulating and analyzing data.
- Data from different file objects can be easily loaded.
- Flexible reshaping and pivoting of data sets
- Provides time-series functionality.

SKLEARN

Scikit-learn is an open-source Python library that implements a range of machine learning, pre-processing, cross-validation, and visualization algorithms using a unified interface.

Important features of scikit-learn:

- Simple and efficient tools for data mining and data analysis. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, etc.
- Accessible to everybody and reusable in various contexts.
- Built on the top of NumPy, SciPy, and matplotlib.
- Open source, commercially usable – BSD license.

Modelling process of Sklearn

Load Dataset

A dataset is nothing but a collection of data. A dataset generally has two main components:

- **Features**: (also known as predictors, inputs, or attributes) they are simply the variables of our data. They can be more than one and hence represented by a **feature matrix** ('X' is a common notation to represent feature matrix). A list of all the feature names is termed **feature names**.
- **Response**: (also known as the target, label, or output) This is the output variable depending on the feature variables. We generally have a single response column and it is represented by a **response vector** ('y' is a common notation to represent response vector). All the possible values taken by a response vector are termed **target names**.

Splitting the dataset

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model.
But this method has several flaws in it, like:

- The goal is to estimate the likely performance of a model on **out-of-sample** data.
- Maximizing training accuracy rewards overly complex models that won't necessarily generalize our model.
- Unnecessarily complex models may over-fit the training data.

A better option is to split our data into two parts: the first one for training our machine learning model, and the second one for testing our model.

Step 3: Training the model

Now, it's time to train some prediction models using our dataset. Scikit-learn provides a wide range of machine learning algorithms that have a unified/consistent interface for fitting, predicting accuracy, etc.

SKLEARN

RE MODULE

A **Regular Expression (RegEx)** is a special sequence of characters that uses a search pattern to find a string or set of strings. It can detect the presence or absence of a text by matching it with a particular pattern and also can split a pattern into one or more sub-patterns. Python provides a **re** module that supports the use of regex in Python. Its primary function is to offer a search, where it takes a regular expression and a string. Here, it either returns the first match or else none.

MATPLOTLIB

Matplotlib is an amazing visualization library in **Python** for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

CHAPTER 2

ABOUT PROJECT

Fake News Detection System

This project main goal was to create a fully functional ML model which can classify whether a given news article is fake or real.

We have created a fully functional model using python libraries which can classify whether news is fake or real without much error by using text of article.

Scope

The main role of model is to classify whether news is fake or real with the help of text written in a news article.

Functions

In this model there several steps which later help us get final result:

1. Dataset is first cleaned as unnecessary variables from datasets are removed. In the given Dataset we only need text of article for classification
2. Later Text of article is pre-processed so that it becomes is for model to classify.
3. After that processed data is converted into vector as ML model can't run using text.
4. Later text converted vector is used for classification either using logistic regression or decision tree.

Feasibility Study:

As we can see what this model is used for we can also think about it's uses in our lives news has become more accessible with coming time it also caused circulation of false news quite easy thus we need systems like these to find out whether news is fake or real. This can help solve various crime in this country as we can weed out false news and those who are spreading it.

Market Analysis

As we can see use of this system we can also face quite a competition in this sector as we have various social media apps using this system in order to make their platform more safe as on the other hand various government agencies have also started applying their own systems as it become very necessary for our country to maintain peace.

Use of Technology in this System

Pandas

This library is used for loading dataset used for training model and cleaning so that it can be properly used for further processes

Re module

This module is used for preprocessing text variable by creating a function in python thus we are able to get a text which is much more readable for system.

Sklearn

Sklearn is one of the more important libraries used in this system as it was for various purposes:

1.Splitting data for training and testing

2.Converting text into vector form

3.Using Logistic regression or decision tree for classification of news.

4.Check efficiency of model by checking it’s F1-score,precision and recall as measuring unit ranging from 0 to 1.

CHAPTER 3

PROJECT SOURCE CODE

SOURCE CODE SCREENSHOT

Import Dataset

```
fake=pd.read_csv("C:\\Users\\manas\\OneDrive\\Desktop\\Fake.csv")
true=pd.read_csv("C:\\Users\\manas\\OneDrive\\Desktop\\True.csv")
```

```
fake.head()
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

```
true.head(5)
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

Insert new column 'type' ¶

```
fake['type']=0
true['type']=1
```

```
fake.shape,true.shape
```

```
((23481, 5), (21417, 5))
```

```
# remove last news from both table for manual test
fake_mt=fake.tail(1)
true_mt=true.tail(1)
fake.drop([23480],axis=0,inplace=True)
true.drop([21416],axis=0,inplace=True)
```

```
fake_mt
```

	title	text	subject	date	type
23480	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

```
true_mt
```

	title	text	subject	date	type
21416	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

```
manual_test=pd.concat([fake_mt,true_mt],axis=0)
```

In this photo we picked two article for manual testing and demonstration.

Merge Two dataframes ¶

```
news=pd.concat([fake,true],axis=0)
news.head()
```

	title	text	subject	date	type
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0

Remove Columns which are not requires

```
news.columns
```

```
Index(['title', 'text', 'subject', 'date', 'type'], dtype='object')
```

```
news=news.drop(['title','subject','date'],axis=1)
```

```
news.isnull().sum()
```

```
text    0
type    0
dtype: int64
```

Shuffle rows ¶

```
news=news.sample(frac=1)
```

```
news.head(20)
```

Process text

```
import re
import string
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
news['text']=news['text'].apply(wordopt)
```

creating dependent and independent variables

```
x=news['text']
y=news['type']
```

Splitting Dataset

```
from sklearn.model_selection import train_test_split
```

```
x_t,x_te,y_t,y_te=train_test_split(x,y,test_size=0.25)
```

Convert text to vector ¶

```
from sklearn.feature_extraction.text import TfidfVectorizer
vect=TfidfVectorizer()
xv_t=vect.fit_transform(x_t)
xv_te=vect.transform(x_te)
```

Logistic Regression

```
from sklearn.linear_model import LogisticRegression
l=LogisticRegression()
l.fit(xv_t,y_t)
```

▾ LogisticRegression

LogisticRegression()

```
pred_l=l.predict(xv_te)
```

```
l.score(xv_te,y_te)
```

0.9854775481111903

```
from sklearn.metrics import classification_report
print(classification_report(y_te,pred_l))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5779
           1       0.99      0.98      0.99      5445

    accuracy                           0.99     11224
   macro avg       0.99      0.99      0.99     11224
weighted avg       0.99      0.99      0.99     11224
```

Decision Tree Classification

```
from sklearn.tree import DecisionTreeClassifier
d=DecisionTreeClassifier()
d.fit(xv_t,y_t)
```

▾ DecisionTreeClassifier

DecisionTreeClassifier()

```
pred_d=d.predict(xv_te)
```

```
d.score(xv_te,y_te)
```

0.994743406985032

```
print(classification_report(y_te,pred_d))
```

```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      5779
           1       1.00      0.99      0.99      5445

    accuracy                           0.99     11224
   macro avg       0.99      0.99      0.99     11224
weighted avg       0.99      0.99      0.99     11224
```

Run Model

```
def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Not A Fake News"

def manual_testing(news):
    testing_news = {"text":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["text"] = new_def_test["text"].apply(wordopt)
    new_x_test = new_def_test["text"]
    new_xv_test = vect.transform(new_x_test)
    pred_LR = l.predict(new_xv_test)
    pred_DT = d.predict(new_xv_test)


    print("LR Prediction:",output_lable(pred_LR[0]),"\nDt Prediction:",output_lable(pred_DT[0]))
```

```
manual_testing(manual_test.text[23480])
```

```
LR Prediction: Fake News
Dt Prediction: Fake News
```

```
manual_testing(manual_test.text[21416])
```

```
LR Prediction: Not A Fake News
Dt Prediction: Not A Fake News
```

Chapter 4

Conclusion

We can tell that by now how this model can use text of news article to detect whether news is fake or real some might ask if we can also use feature ‘title’ but if we think logically we will be able to understand that title are nothing but announcement made it in the article main content that we can verify whether the facts stated by it are correct or not.

Text used for classification have a very good result as you can see in screenshots

BIBLOGRAPHY

PYTHON: Wikipedia

Sklearn: GeeksforGeeks

RE AND PANDAS: GeeksforGeeks

Dataset Used:Kaggle