**TITLE: FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING**

**PHASE – 4**

**SUBMITTED BY : V S SANGEETHA**

**MAIL ID: sangeethas.devi2004@gmail.com**

**ABOUT THIS PHASE**

In this phase we need to do performing different activities like feature engineering, model training, evaluation as per the instructions in the project.

**DATA COLLECTION**

Gather a dataset of news articles, both real and fake. Ensure that the dataset is well-labeled.

**INSEARTING A COLUMN CLASS AS TARGET FEATURES**

The target column in the training data contains the historical values used to train the model. The target column in the test data contains the historical values to which the predictions are compared. The act of scoring produces a prediction for the target.

**MERGING TRUE AND FAKE DATA FRAMES**

The merge() operation is a method used to combine two dataframes based on one or more common columns, also called keys. The resulting data frame contains only the rows from both dataframes with matching keys. The merge() function is similar to the SQL JOIN operation**.**

**REMOVING COLUMS WHICH ARE NOT REQUIRED**

If your query has columns you don't need, you can remove them. You can select one or more columns, and then either remove the selected ones, or remove the unselected ones, that is the other columns. Consider the difference between removing a column and removing other columns**.**

**RANDOM SHUFFLING THE DATAFRAMES**

One of the easiest ways to shuffle or permute a DataFrame in Pandas is by using the sample() method. The sample() method randomly samples rows from the DataFrame, and you can specify the number of rows to sample using the n parameter.

**CREATING A FUNCTION TO PROCESS THE TEXT**

The text processing of a regular expression is a virtual editing machine, having a primitive programming language that has named registers (identifiers), and named positions in the sequence of characters comprising the text. Using these, the "text processor" can, for example, mark a region of text, and then move it.

**DEFINING DEPENDNT AND INDEPENDENT VARIABLES**

The independent variable is the cause. Its value is independent of other variables in your study. The dependent variable is theeffect. Its value depends on changes in the independent variable.

**CONVERT TEXT TO VECTOR**

Converting words to vectors, or word vectorization, is a natural language processing (NLP) process. The process uses language models to map words into vector space. A vector space represents each word by a vector of real numbers. It also allows words with similar meanings have similar representations.

**MODEL TESTING**

Model-based testing (MBT) is an approach to software testing that requires developers to create a second, lightweight implementation of a software build called a model. Typically, a model consists of business logic and is just a few lines of code.

Importing Dataset

import pandas as pd import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split from sklearn.metrics import accuracy\_score

from sklearn.metrics import classification\_report import re

import string

df\_fake = pd.read\_csv("/content/Fake.csv.zip") df\_true = pd.read\_csv("/content/True.csv.zip")

df\_fake.head()

**title text subject date**

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

df\_true.head(5)

**title text subject date**

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

Inserting a column "class" as target feature

df\_fake["class"] = 0

df\_true["class"] = 1

df\_fake.shape, df\_true.shape ((23481, 5), (21417, 5))

# Removing last 10 rows for manual testing df\_fake\_manual\_testing = df\_fake.tail(10) for i in range(23480,23470,-1):

df\_fake.drop([i], axis = 0, inplace = True)

df\_true\_manual\_testing = df\_true.tail(10) for i in range(21416,21406,-1):

df\_true.drop([i], axis = 0, inplace = True)

df\_fake.shape, df\_true.shape ((23471, 5), (21407, 5))

df\_fake\_manual\_testing["class"] = 0

df\_true\_manual\_testing["class"] = 1

<ipython-input-10-3aaf8ec2aad1>:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-c](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)o df\_fake\_manual\_testing["class"] = 0

<ipython-input-10-3aaf8ec2aad1>:2: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame. Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-c](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)o df\_true\_manual\_testing["class"] = 1

df\_fake\_manual\_testing.head(10)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **title** | **text** | **subject** | **date** | **class** |
| **23471** | Seven Iranians freed in the prisoner swap have... | 21st Century Wire says This week, the historic... | Middle-east | January 20, 2016 | 0 |
| **23472** | #Hashtag Hell & The Fake Left | By Dady Chery and Gilbert MercierAll writers ... | Middle-east | January 19, 2016 | 0 |
| **23473** | Astroturfing: Journalist Reveals Brainwashing ... | Vic Bishop Waking TimesOur reality is carefull... | Middle-east | January 19, 2016 | 0 |
| **23474** | The New American Century: An Era of Fraud | Paul Craig RobertsIn the last years of the 20t... | Middle-east | January 19, 2016 | 0 |
| **23475** | Hillary Clinton: ‘Israel First’ (and no peace ... | Robert Fantina CounterpunchAlthough the United... | Middle-east | January 18, 2016 | 0 |
| **23476** | McPain: John McCain Furious That Iran Treated ... | 21st Century Wire says As 21WIRE reported earl... | Middle-east | January 16, 2016 | 0 |
| **23477** | JUSTICE? Yahoo Settles E-mail Privacy Class-ac... | 21st Century Wire says It s a familiar theme. ... | Middle-east | January 16, 2016 | 0 |
| **23478** | Sunnistan: US and Allied ‘Safe Zone’ Plan to T... | Patrick Henningsen 21st Century WireRemember ... | Middle-east | January 15, 2016 | 0 |
| **23479** | How to Blow $700 Million: Al Jazeera America F... | 21st Century Wire says Al Jazeera America will... | Middle-east | January 14, 2016 | 0 |
| **23480** | 10 U.S. Navy Sailors Held by Iranian Military ... | 21st Century Wire says As 21WIRE predicted in ... | Middle-east | January 12, 2016 | 0 |

df\_true\_manual\_testing.head(10)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **title** | **text** | **subject** | **date** | **class** |
| **21407** | Mata Pires, owner of embattled Brazil builder ... | SAO PAULO (Reuters) - Cesar Mata Pires, the ow... | worldnews | August 22, 2017 | 1 |
| **21408** | U.S., North Korea clash at U.N. forum over nuc... | GENEVA (Reuters) - North Korea and the United ... | worldnews | August 22, 2017 | 1 |
| **21409** | U.S., North Korea clash at U.N. arms forum on ... | GENEVA (Reuters) - North Korea and the United ... | worldnews | August 22, 2017 | 1 |
| **21410** | Headless torso could belong to submarine journ... | COPENHAGEN (Reuters) - Danish police said on T... | worldnews | August 22, 2017 | 1 |
| **21411** | North Korea shipments to Syria chemical arms a... | UNITED NATIONS (Reuters) - Two North Korean sh... | worldnews | August 21, 2017 | 1 |
| **21412** | 'Fully committed' NATO backs new U.S. approach... | BRUSSELS (Reuters) - NATO allies on Tuesday we... | worldnews | August 22, 2017 | 1 |
| **21413** | LexisNexis withdrew two products from Chinese ... | LONDON (Reuters) - LexisNexis, a provider of l... | worldnews | August 22, 2017 | 1 |
| **21414** | Minsk cultural hub becomes haven from authorities | MINSK (Reuters) - In the shadow of disused Sov... | worldnews | August 22, 2017 | 1 |
| **21415** | Vatican upbeat on possibility of Pope Francis ... | MOSCOW (Reuters) - Vatican Secretary of State ... | worldnews | August 22, 2017 | 1 |
| **21416** | Indonesia to buy $1.14 billion worth of Russia... | JAKARTA (Reuters) - Indonesia will buy 11 Sukh... | worldnews | August 22, 2017 | 1 |

df\_manual\_testing = pd.concat([df\_fake\_manual\_testing,df\_true\_manual\_testing], axis = 0) df\_manual\_testing.to\_csv("manual\_testing.csv")

Merging True and Fake Dataframes

df\_merge = pd.concat([df\_fake, df\_true], axis =0 ) df\_merge.head(10)

**title text subject date class**

1. Donald Trump Sends Out Embarrassing New Year’... Donald Trump just couldn t wish all Americans ... News December 31, 2017 0
2. Drunk Bragging Trump Staffer Started Russian ... House Intelligence Committee Chairman Devin Nu... News December 31, 2017 0
3. Sheriff David Clarke Becomes An Internet Joke... On Friday, it was revealed that former Milwauk... News December 30, 2017 0

df\_merge.columns

**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

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

**5** Racist Alabama Cops Brutalize Black Boy While... The number of cases of cops brutalizing and ki... News December 25, 2017 0

Removing columns which are not required

**6** Fresh Off The Golf Course, Trump Lashes Out A... Donald Trump spent a good portion of his day a... News December 23, 2017 0

**7** Trump Said Some INSANELY Racist Stuff Inside ... In the wake of yet another court decision that... News December 23, 2017 0

df = df\_merge.drop(["title", "subject","date"], axis = 1)

**8** Former CIA Director Slams Trump Over UN Bully... Many people have raised the alarm regarding th... News December 22, 2017 0

df.isnull().sum()

**9** WATCH: Brand-New Pro-Trump Ad Features So Muc... Just when you might have thought we d get a br... News December 21, 2017 0

text 0

class 0

dtype: int64

Random Shuffling the dataframe

df = df.sample(frac = 1)

df.head()

|  |  |  |
| --- | --- | --- |
|  | **text** | **class** |
| **7701** | Bobby Jindal abandoned his home state of Louis... | 0 |
| **16528** | Newsflash Hillary WAR is not aesthetically ple... | 0 |
| **9527** | WASHINGTON (Reuters) - The White House confirm... | 1 |
| **18516** | Texas Democrat Rep. Al Green announced on Wedn... | 0 |
| **993** | WASHINGTON (Reuters) - The U.S. State Departme... | 1 |

df.reset\_index(inplace = True)

df.drop(["index"], axis = 1, inplace = True)

df.columns

Index(['text', 'class'], dtype='object')

df.head()

**text class**

1. Bobby Jindal abandoned his home state of Louis... 0
2. Newsflash Hillary WAR is not aesthetically ple... 0
3. WASHINGTON (Reuters) - The White House confirm... 1
4. Texas Democrat Rep. Al Green announced on Wedn... 0
5. WASHINGTON (Reuters) - The U.S. State Departme... 1

Creating a function to process the texts

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

df["text"] = df["text"].apply(wordopt)

Defining dependent and independent variables

x = df["text"]

y = df["class"]

Splitting Training and Testing

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.25)

Convert text to vectors

from sklearn.feature\_extraction.text import TfidfVectorizer vectorization = TfidfVectorizer()

xv\_train = vectorization.fit\_transform(x\_train) xv\_test = vectorization.transform(x\_test)

Logistic Regression

from sklearn.linear\_model import LogisticRegression

LR = LogisticRegression() LR.fit(xv\_train,y\_train)



▾ LogisticRegression

LogisticRegression()

pred\_lr=LR.predict(xv\_test)

LR.score(xv\_test, y\_test) 0.9868983957219252

print(classification\_report(y\_test, pred\_lr))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 0.99 | 0.99 | 5950 |
| 1 | 0.99 | 0.99 | 0.99 | 5270 |
| accuracy |  |  | 0.99 | 11220 |
| macro avg | 0.99 | 0.99 | 0.99 | 11220 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11220 |

Decision Tree Classification

from sklearn.tree import DecisionTreeClassifier

DT = DecisionTreeClassifier() DT.fit(xv\_train, y\_train)





▾ DecisionTreeClassifier

DecisionTreeClassifier()

pred\_dt = DT.predict(xv\_test)

DT.score(xv\_test, y\_test)

0.9952762923351158

print(classification\_report(y\_test, pred\_dt))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 1.00 | 1.00 | 5950 |
| 1 | 1.00 | 0.99 | 0.99 | 5270 |
| accuracy |  |  | 1.00 | 11220 |
| macro avg | 1.00 | 1.00 | 1.00 | 11220 |
| weighted avg | 1.00 | 1.00 | 1.00 | 11220 |

Gradient Boosting Classifier

from sklearn.ensemble import GradientBoostingClassifier

GBC = GradientBoostingClassifier(random\_state=0) GBC.fit(xv\_train, y\_train)

▾

GradientBoostingClassifier

GradientBoostingClassifier(random\_state=0)

pred\_gbc = GBC.predict(xv\_test)

GBC.score(xv\_test, y\_test) 0.9955436720142602

print(classification\_report(y\_test, pred\_gbc))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 1.00 | 0.99 | 1.00 | 5950 |
| 1 | 0.99 | 1.00 | 1.00 | 5270 |
| accuracy |  |  | 1.00 | 11220 |
| macro avg | 1.00 | 1.00 | 1.00 | 11220 |
| weighted avg | 1.00 | 1.00 | 1.00 | 11220 |

Random Forest Classifier

from sklearn.ensemble import RandomForestClassifier

RFC = RandomForestClassifier(random\_state=0) RFC.fit(xv\_train, y\_train)



▾

RandomForestClassifier

RandomForestClassifier(random\_state=0)

pred\_rfc = RFC.predict(xv\_test)

RFC.score(xv\_test, y\_test) 0.9881461675579323

print(classification\_report(y\_test, pred\_rfc))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.99 | 0.99 | 0.99 | 5950 |
| 1 | 0.99 | 0.99 | 0.99 | 5270 |
| accuracy |  |  | 0.99 | 11220 |
| macro avg | 0.99 | 0.99 | 0.99 | 11220 |
| weighted avg | 0.99 | 0.99 | 0.99 | 11220 |

# Model Testing

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 = vectorization.transform(new\_x\_test) pred\_LR = LR.predict(new\_xv\_test)

pred\_DT = DT.predict(new\_xv\_test)

pred\_GBC = GBC.predict(new\_xv\_test) pred\_RFC = RFC.predict(new\_xv\_test)

return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC Prediction: {} \nRFC Prediction: {}".format(output\_lable(pred\_LR[0]),

output\_lable(pred\_GBC[0]), output\_lable(pred\_RFC[0])))

news = str(input()) manual\_testing(news)

news = str(input()) manual\_testing(news)

}news = str(input()) manual\_testing(news)