Detecting Fake News with Python and Machine Learning

18CSE479T- Statistical Machine Learning

Submitted by

Varsha.S (RA2011026010286)

Submitted to

Ms. Akshya J

Assistant Professor, Department of Computing Intelligence

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE ENGINEERING WITH SPECIALIZATION IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

SCHOOL OF COMPUTING

DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

COLLEGE OF ENGINEERING AND TECHNOLOGY SRM

INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR- 603 203

NOV 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this mini project report "Detecting Fake News with Python and Machine Learning" is the bonafide work of Varsha.S (RA2011026010286) who carried out the project work under my supervision.

SIGNATURE

Ms. Akshya J
Assistant Professor
Department of Computational Intelligence
SRM Institute of Science and Technology

Detecting Fake News with Python and Machine Learning project

What is Fake News?

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims, and may end up being viralized by algorithms, and users may end up in a filter bubble.

Abstract :

Today, the increased amount of information sources on internet creates the problem of information overflow. Filtering the relevant and genuine information is another challenge social media facing now. Mobile phones and other electronic gadgets became quite common through which people get up-to-date information. Verifying the authenticity of news needs to have prime importance though a difficult task. This paper outlines a new approach for finding the genuineness of news content. This helps to eliminate the rumors from spreading through social platforms. By using the web scraping method, we assemble the news content related to the news posted for checking. The news prediction is done by implementing techniques like TF-IDF, Bag of words and Natural language processing. The experimental results specify that the system shows an accuracy of 90% when tested against a test set.

Algorithm :

a Fake News Prediction System using Machine Learning with Python. We will be using Logistic Regression model for prediction.

Dataset:

train.csv: A full training dataset with the following attributes:

- **id**: unique id for a news article
- **title**: the title of a news article
- **author**: author of the news article
- **text**: the text of the article; could be incomplete
- **label**: a label that marks the article as potentially unreliable
    - 1: unreliable
    - 0: reliable

test.csv: A testing training dataset with all the same attributes at train.csv without the label.
submit.csv: A sample submission that you can

the Dataset:

1. id: unique id for a news article
2. title: the title of a news article
3. author: author of the news article
4. text: the text of the article; could be incomplete
5. label: a label that marks whether the news article is real or fake:

```
1: Fake news
0: real News
```

Steps for detecting fake news with Python:

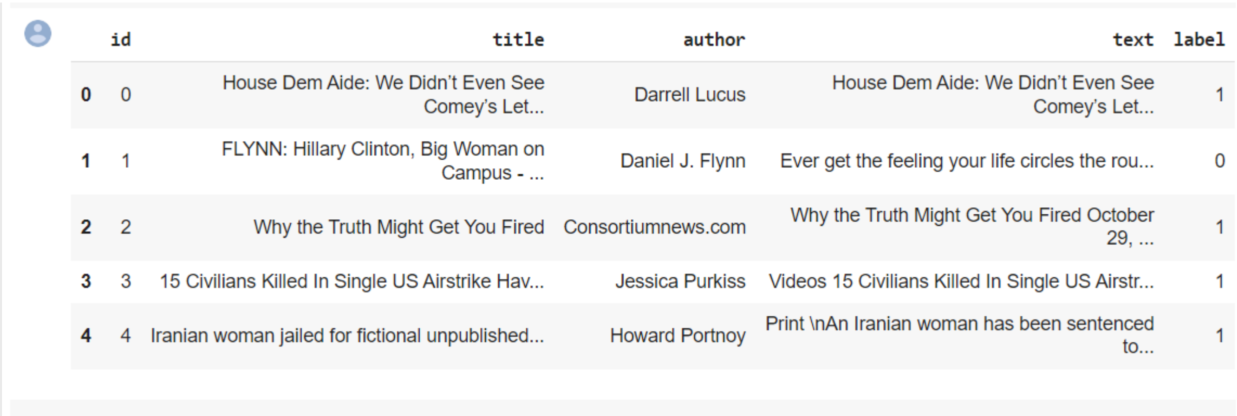
1)Importing the Dependencies

```
import numpy as np
import pandas as pd
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

import nltk
nltk.download('stopwords')

# printing the stopwords in English
print(stopwords.words('english'))
out:
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours',
'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they',
'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is',
'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but',
'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during',
'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then',
'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should',
"should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn',
"doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn',
"mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won',
"won't", 'wouldn', "wouldn't"
```

2) Data Pre-processing

```
 # loading the dataset to a pandas DataFrame

news_dataset = pd.read_csv('/content/train.csv')

news_dataset.shape

# print the first 5 rows of the dataframe
news_dataset.head()
```



	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

```
# counting the number of missing values in the dataset
news_dataset.isnull().sum()
id 0
title 558
author 1957
text 39
label 0
dtype: int64

# replacing the null values with empty string
news_dataset = news_dataset.fillna('')

# merging the author name and news title
news_dataset['content'] = news_dataset['author']+' '+news_dataset['title']

print(news_dataset['content'])

# separating the data & label
X = news_dataset.drop(columns='label', axis=1)
Y = news_dataset['label']
```

3) Stemming:

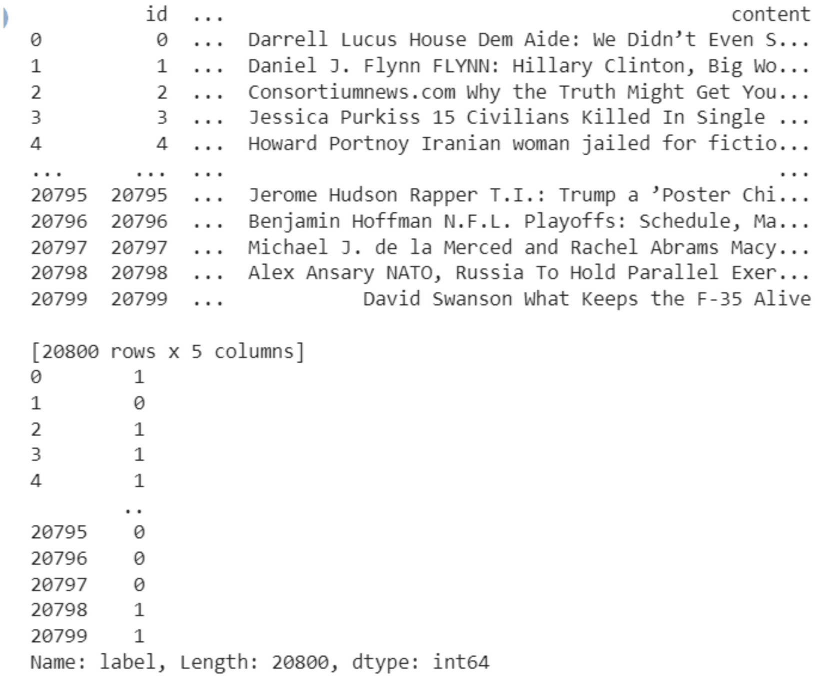
Stemming is the process of reducing a word to its Root word

example: actor, actress, acting --> act

```
def stemming(content):
    stemmed_content = re.sub('[^a-zA-Z]',' ',content)
    stemmed_content = stemmed_content.lower()
    stemmed_content = stemmed_content.split()
    stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
    stemmed_content = ' '.join(stemmed_content)
    return stemmed_content


news_dataset['content'] = news_dataset['content'].apply(stemming)
```

```
print(news_dataset['content'])
0       Darrell Lucus House Dem Aide: We Didn't Even S...
1       Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2       Consortiumnews.com Why the Truth Might Get You...
3       Jessica Purkiss 15 Civilians Killed In Single ...
4       Howard Portnoy Iranian woman jailed for fictio...
                        ...
20795   Jerome Hudson Rapper T.I.: Trump a 'Poster Chi...
20796   Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797   Michael J. de la Merced and Rachel Abrams Macy...
20798   Alex Ansary NATO, Russia To Hold Parallel Exer...
20799           David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

```
#separating the data and label
X = news_dataset['content'].values
Y = news_dataset['label'].values

print(X)

print(Y)
```



```
          id  ...                                            content
0          0  ...  Darrell Lucus House Dem Aide: We Didn't Even S...
1          1  ...  Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2          2  ...  Consortiumnews.com Why the Truth Might Get You...
3          3  ...  Jessica Purkiss 15 Civilians Killed In Single ...
4          4  ...  Howard Portnoy Iranian woman jailed for fictio...
...      ...  ...                                                ...
20795  20795  ...  Jerome Hudson Rapper T.I.: Trump a 'Poster Chi...
20796  20796  ...  Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797  20797  ...  Michael J. de la Merced and Rachel Abrams Macy...
20798  20798  ...  Alex Ansary NATO, Russia To Hold Parallel Exer...
20799  20799  ...            David Swanson What Keeps the F-35 Alive

[20800 rows x 5 columns]
0        1
1        0
2        1
3        1
4        1
        ..
20795    0
20796    0
20797    0
20798    1
20799    1
Name: label, Length: 20800, dtype: int64
```

4) Splitting the dataset to training & test data

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.2, st
ratify=Y, random_state=2)
```

5) Training the Model: Logistic Regression

```
model = LogisticRegression()

model.fit(X_train, Y_train)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
```

6) Evaluation

---

accuracy score

```
accuracy score on the training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)


# accuracy score on the test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)


print('Accuracy score of the test data : ', test_data_accuracy)
```

7) Making a Predictive System

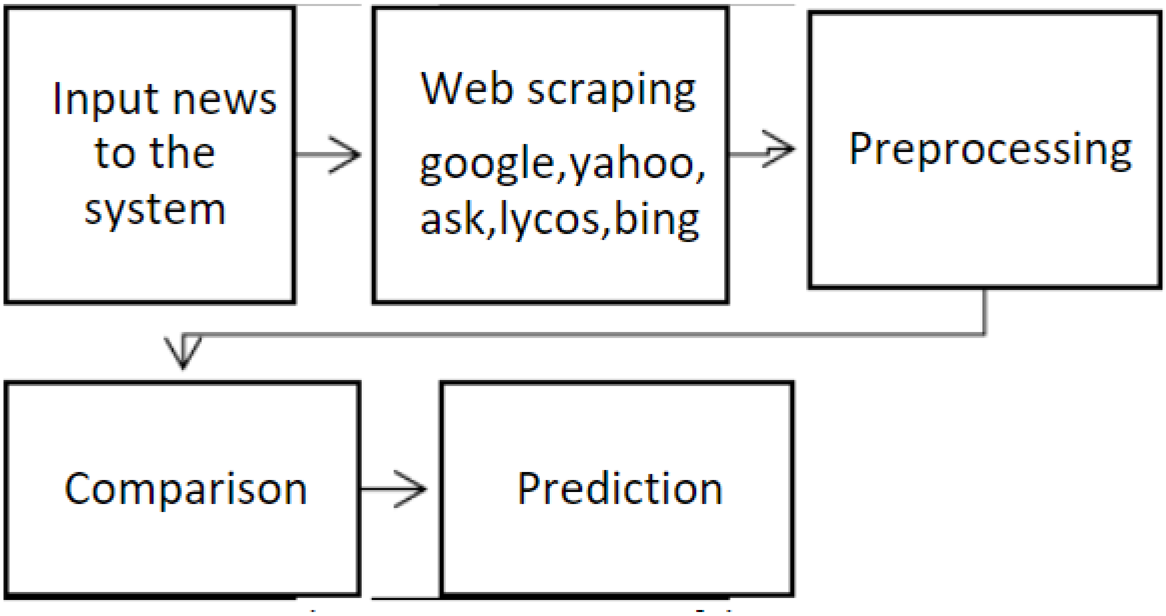
```
X_new = X_test[3]

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

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

out:

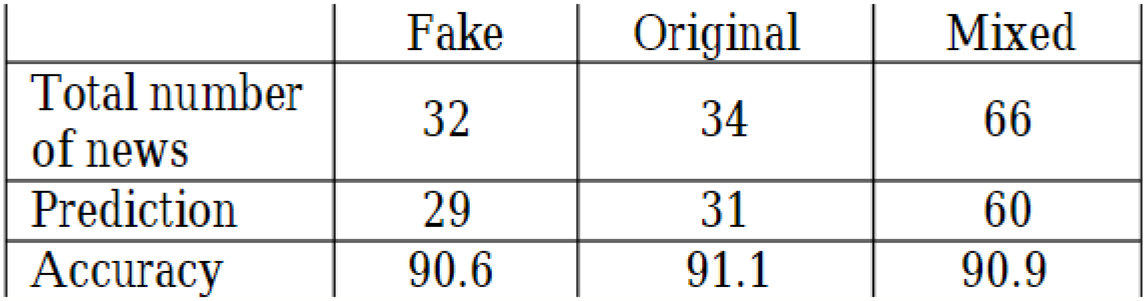
```
[0]
The news is Real
```

```
print(Y_test[3])
```

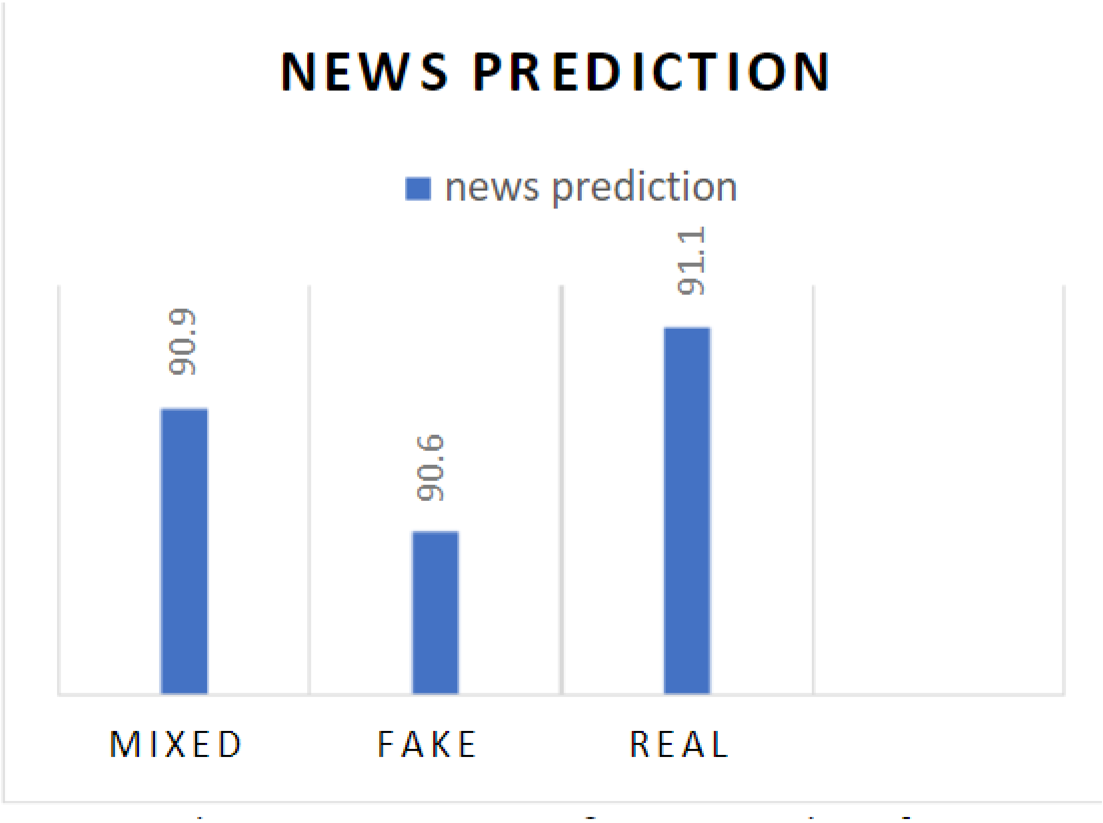
```
out : 0
```



System Architecture



	Fake	Original	Mixed
Total number of news	32	34	66
Prediction	29	31	60
Accuracy	90.6	91.1	90.9



RESULT :

Detecting fake news dataset has been thoroughly read and python code has been implemented.