IoC Skills Bootcamp in Data Science and Cloud Systems

Final Project

Detecting Fake News with Python and Machine Learning

BY : Gitta Ataeimanesh

Table of Content

- Introduction
- Motivation
- Implementation
- Results
- Future work

Introduction

Fake news is misleading information deliberately spread as "news" typically spread online through social media or digital platforms. This is often done to further or impose certain ideas and is often achieved with political agendas. It could negatively influence peoples lives as create confusion and misunderstanding about important social and political topics and Trusting these false information could lead to make wrong and bad decisions that may be harmful to their health and well-being. detecting whether a news is real or fake information could prevent people from making incorrect decisions.

I have mostly use Python programming combined with :

Numpy	Python Library for multidimensional arrays and matrices, functions to perform advanced mathematical and statistical operations, provides vectorization of mathematical operations on arrays and matrices, many other python libraries are built on NumPy
Pandas	• Python Library for Data Science ,used to analyze data .It has been designed to work with DataFrames, adds data structures , tools for reading and loading data, provides tools for data manipulation and allows cleaning, handling missing & plotting data .
SciKit-Learn	• Python Library which provides algorithms for supervised and unsupervised machine learning models such as Classification , Regression and clustering and model validation technics.
Matplotlib, Seaborn	• plotting libraries in python that serves as a visualization utility and high-level interface for drawing statical graphics
PICKLE, Joblib	• set of tools to provide lightweight pipelining in Python • Use the Model for Inferencing ,Deploy the model and use it for new data • useful to save trained model, for use later to predict labels for new data
Flask, Django	• lightweight web applications . Flask is a Web Server Gateway Interface, use to deploy the Classification model and detect any news on the web application as a Fake or Real. • Django is a back-end server- side web framework

Motivation

Research suggest that the rapid spread of Fake News and misinformation online can have profound consequences. Such as:

- Undermining the democratic process. Based on Statista, 83% of people believe fake news negatively affects their country's politics 80% of people are convinced that fake news has a negative effect on other countries' politics, too.

- Platforms for harmful conspiracy theories and hate speech

- Create bias to achieve specific political agenda. Trusting these false information could lead people to make incorrect decisions.

- Spread of false or discredited science such as anti-vax movement, Vaccine hesitancy which could lead to making decisions that may be harmful for health.

- Generating problem for people to distinguish between fake and real news which will create confusion and misunderstanding about important social and political issues. Fake News Makes It Harder For People To See the Truth. 56% of Facebook users can't recognize fake news that aligns with their beliefs, social media statistics show(SSRN)

- Distrust in the media and negative impact on organization's reputation.

  Media trust worldwide has dropped by 8% between 2020 and 2021

  52% of Americans say they regularly encounter fake news online.

  67% of US adults say they've come across false information on social media.

  fake news statistics show Social media is the least trusted news source worldwide

Solution to this predicament:

To deploy a Fake News Detection application based on Machine Learning to predict whether a news is a Real journalism or fake information. This improve lives and wellbeing as a result of better decision making.
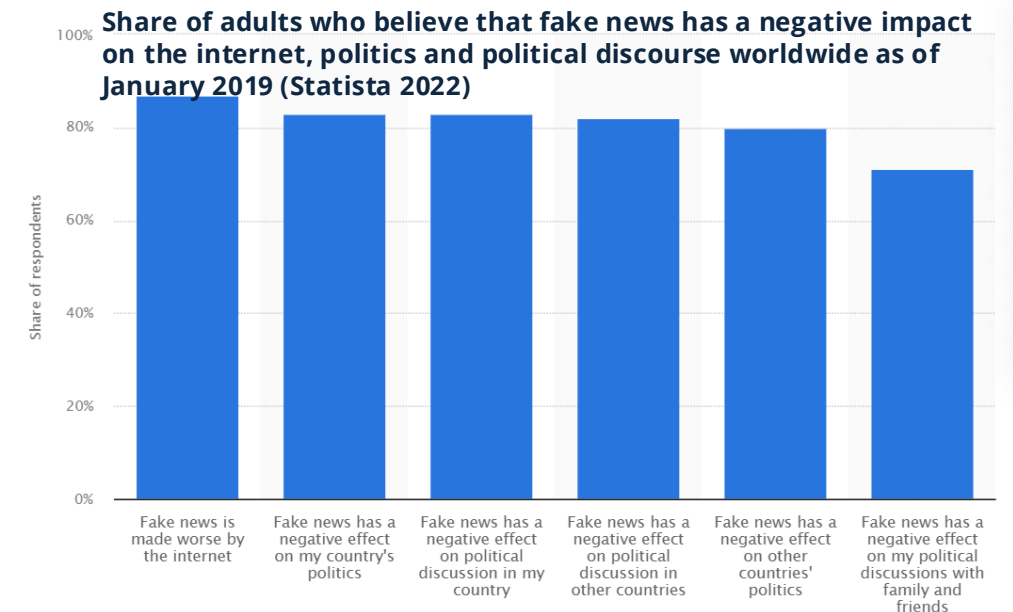
Concern about what is real or fake on the internet

This has risen **12 percentage points** in the last year in the UK

55% (average) agree they are worried
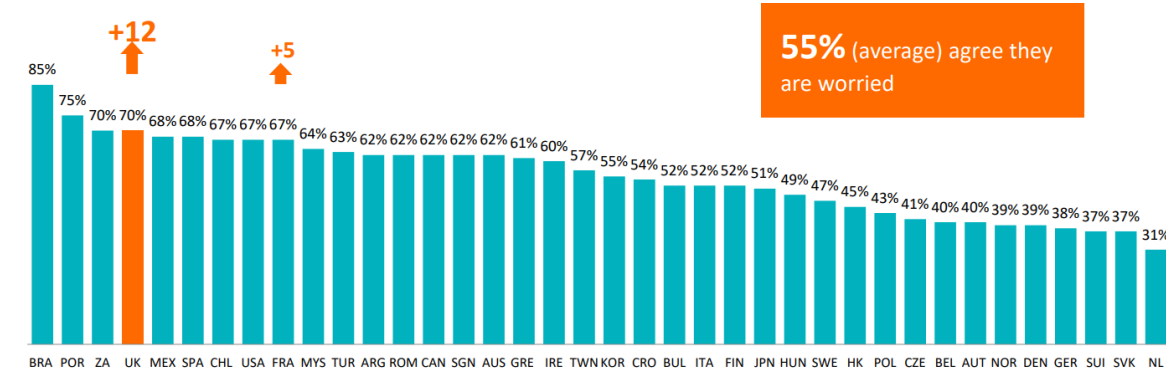
Q. "Thinking about online news, I am concerned about what is real and what is fake on the internet." Base: All markets 2019

www.ofcom.org.uk

Share of adults who believe that fake news has a negative impact on the internet, politics and political discourse worldwide as of January 2019 (Statista 2022)

Implementation

Technic :
Passive Aggressive Classifier, A Classification algorithm falls under the category of online learning in machine learning,will classify a piece of news in Two Categories : REAL or FAKE
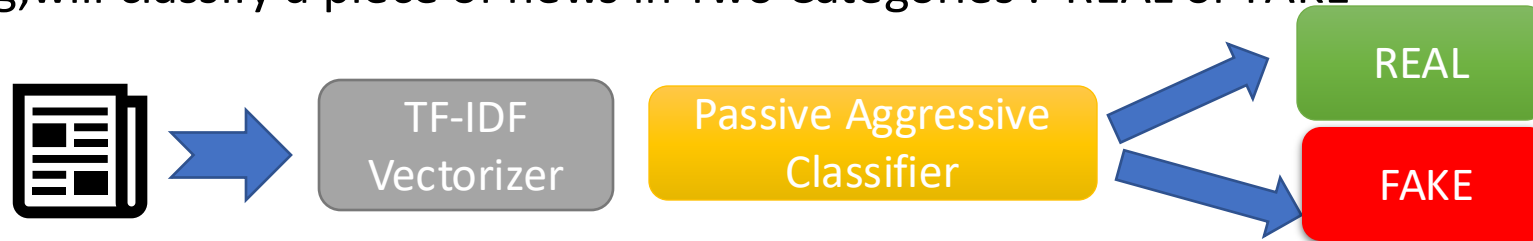
1- The TF-IDF Vectorizer transforms a collection of raw documents into a matrix of TF-IDF features
> will transform the text into meaningful representation of integers or numbers which is used to fit machine learning algorithm for predictions
> TF (Term Frequency):The number of times a word appears in a document is its Term Frequency
> IDF (Inverse Document Frequency):Words that occur many times a document, but also occur many times in many others, may be irrelevant. IDF is a measure of how significant a term is in the entire corpus.

2- Implement a 'Passive Aggressive Classifier' to fit the model
> an online learning algorithm where a system will be trained incrementally by feeding it instances sequentially, individually or in small groups called mini-batches.
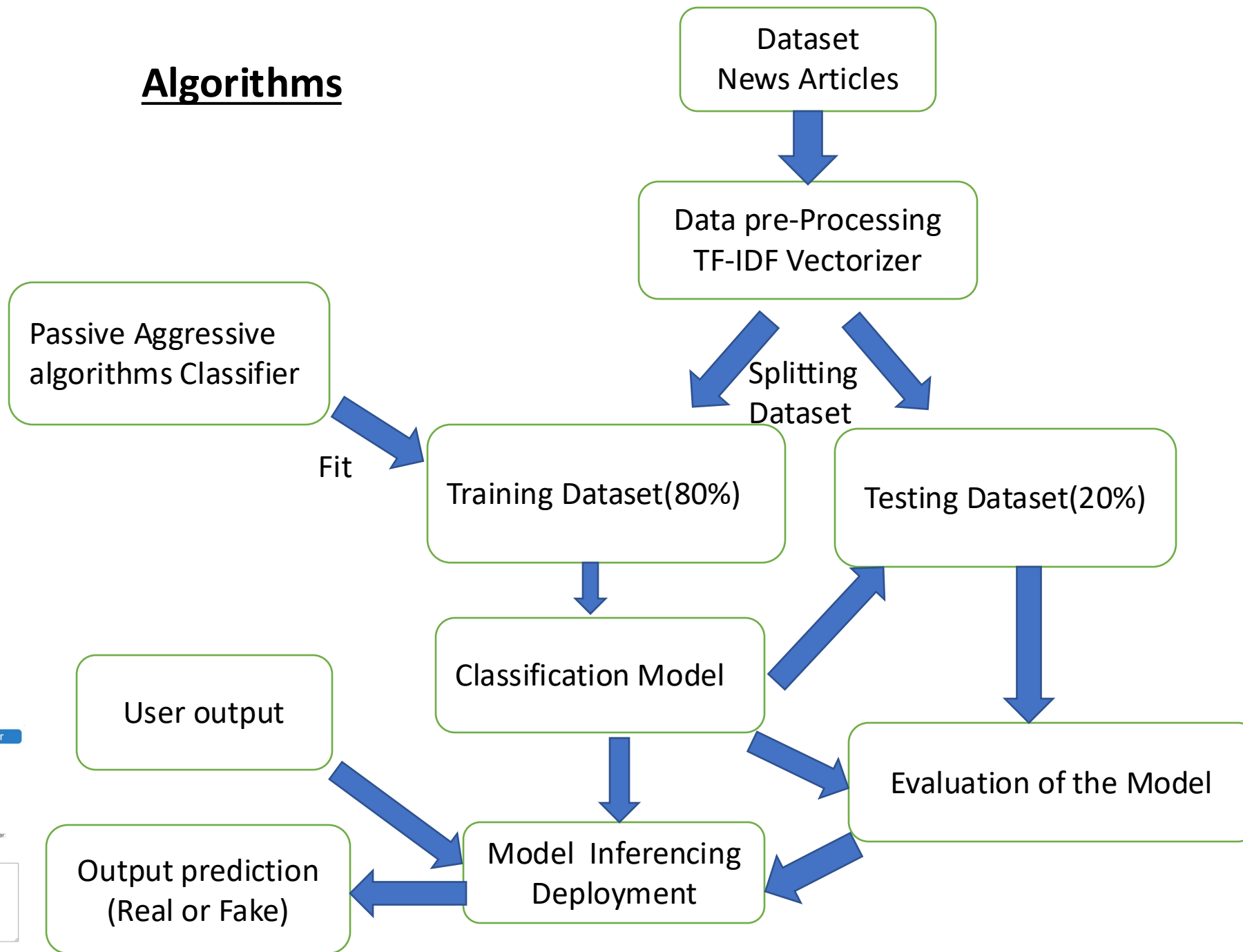> In online learning, a machine learning model is trained and deployed in production in a way that continues to learn as new data sets arrive. It remains passive for correct predictions and responds aggressively to incorrect predictions.

Passive: If the prediction is correct, keep the model and do not make any changes. i.e., the data in the example is not enough to cause any changes in the model.
Aggressive: If the prediction is incorrect, make changes to the model. i.e., some change to the model may correct it.

3- The accuracy Score and Confusion Matrix will evaluate the model

Algorithms

Dataset
News Articles

Data pre-Processing
TF-IDF Vectorizer

Passive Aggressive
algorithms Classifier

Splitting
Dataset

Fit

Training Dataset(80%)

Testing Dataset(20%)

Classification Model

User output

Evaluation of the Model

Output prediction
(Real or Fake)

Model Inferencing
Deployment

Fake news classifier

Real or fake?

You are reading the news, and might wonder:
Is this fake news? Or a trustful article?

Enter the article:

Insert here

Algorithms

1- Importing Libraries

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib
import seaborn as sns
import itertools
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.metrics import accuracy_score, confusion_matrix
```

2- Loading the Dataset

```python
df = pd.read_csv("/Users/gatae/OneDrive/Desktop/news.csv")
```

3- Preparing data (Cleaning Data, Checking for Duplications and missing Data), Text preprocessing using TF-IDF Vectorizer will Convert Text to Vectors

```python
print(df)
df.head()
lb=df.label
lb.head()
df.drop_duplicates(inplace= True)
df.shape
df.isnull().sum()
```

4- Splitting Dataset to Training Sets (80%)
and Test Sets (20%)
Fitting training model With Passive
Aggressive Classifier and predicting
on Test Set

```python
x_train,x_test,y_train,y_test=train_test_split(df['text'], lb, test_size=0.2, random_state=7)

tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
tfidf_train=tfidf_vectorizer.fit_transform(x_train)
tfidf_test=tfidf_vectorizer.transform(x_test)

pac=PassiveAggressiveClassifier(max_iter=50)
pac.fit(tfidf_train,y_train)

y_pred=pac.predict(tfidf_test)
score=accuracy_score(y_test,y_pred)

print("Accuracy: ",round(score*100,2),"%")
```

5- Evaluating the Model,
Measuring performance with
confusion Matrix

```python
pac_con=confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
print("Confusion Matrix for PassiveAgressiveClassifier",end="\n")
print(pac_con)
```

```
[6335 rows x 4 columns]
Accuracy:  92.98 %
Confusion Matrix for PassiveAgressiveClassifier
[[590  48]
 [ 41 588]]
```

Model Accuracy is : 92.98%

```python
ax = sns.heatmap(confusion_matrix(y_test,y_pred), annot=True, fmt="d")
ax.set(xlabel='Prediction', ylabel='Actual')
plt.show()
```

```python
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

conf_matrix = confusion_matrix(y_test, y_pred, labels=['FAKE', 'REAL'])

conf_matrix_table = classification_report(y_test, y_pred, labels=['FAKE', 'REAL'])

print(conf_matrix_table)
```

```
              precision    recall  f1-score   support

        FAKE       0.93      0.93      0.93       638
        REAL       0.93      0.93      0.93       629

    accuracy                           0.93      1267
   macro avg       0.93      0.93      0.93      1267
weighted avg       0.93      0.93      0.93      1267
```
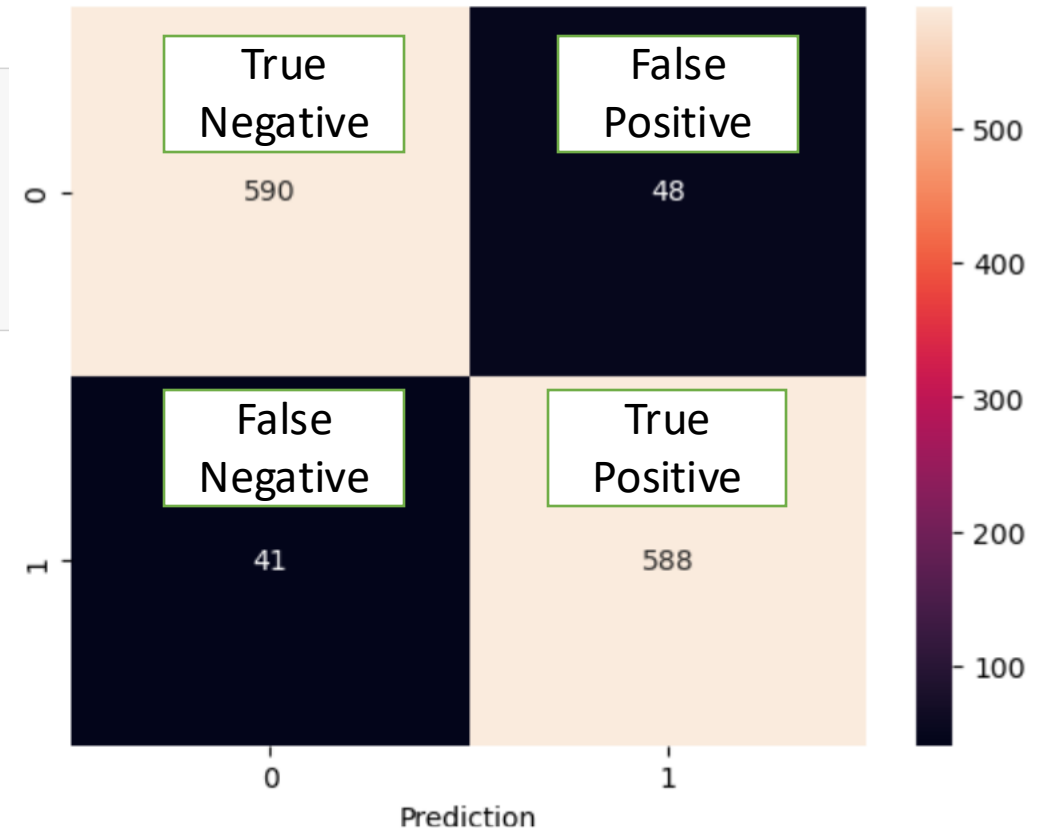

6- Pickle(save) and loading and inferencing the model

Results

a reasonably useful trained model will be saved for use later to predict labels for new data Via inferencing and deployment . A Web Application for End user using Flask or Django can be created

A fake news prediction web application using machine learning Algorithm developed using Django and Heroku

Enter your Text to try it

Write your Text here…..

Predict

Results

User will enter the News in the web Application input Box and will receive a result Based on Machin learning Algorithm which shows whether It is Real or Fake news

Future work

- Detecting and analysing the quality of original sources of information for instance , if news originated from not trustworthy sources that have been know for spreading false or misleading information, it will show it.

- Detecting Types and aspects of Fake News via analysing it :

❖ **Fake/Hoax News** - news that is fabricated with the intention of misleading or confusing readers

❖ **Misleading News** - news stories that report quotes, images, statistics out of context, some of these stories can be old stories that are re-reported with a new misleading headline

❖ **Mimic Websites** - fake news websites that mimic the look of trusted news sources in order to fool readers into thinking a story is real

❖ **Satire** - fake and ironic news stories that are intended to be funny or entertaining. Examples: The Onion, The Borowitz Report

❖ **Clickbait** - news that is promoted with dramatic or misleading headlines that do not reflect the content of the actual story

❖ **Alternative/Alternate Facts** - A different interpretation of facts, usually derived from a misinterpretation of reports or by focusing only on a subset of the available information

- Add more functionality such as Image processing and visual analytics ,Network and Metadata fingerprints analysis and journalist news poster analysis to this application