

## Detecting Fake News with Python and Machine Learning

**Aim:** Detecting Fake News with Python and Machine Learning is to develop a robust and accurate machine learning model that can analyze news articles and determine whether they are genuine or fake, thereby helping to combat the spread of misinformation and improve media literacy.

Attributes in dataset:

**Title:** Denotes the title of the news.

**Text:** text information under the title.

**Label:** This represents the FAKE or REAL news.

The dataset is shown below.

	title	text	label
8476	You Can Smell Hillaryâ€™s Fear	Daniel Greenfield, a Shillman Journalism Fellow at the	FAKE
10294	Watch The Exact Moment Paul Ryan Co	Google Pinterest Digg Linkedin Reddit Stumbleupon Print	FAKE
3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Monday that	REAL
10142	Bernie supporters on Twitter erupt in anger	â€” Kaydee King (@KaydeeKing) November 9, 2016 The	FAKE
875	The Battle of New York: Why This Primary	It's primary day in New York and front-runners Hillary	REAL
6903	Tehran, USA		FAKE
7341	Girl Horrified At What She Watches Boy	Share This Baylee Luciani (left), Screenshot of what	FAKE
95	â€” Britainâ€™s Schindlerâ€™ Dies at 10	A Czech stockbroker who saved more than 650 Jewish ch	REAL
4869	Fact check: Trump and Clinton at the 'cc	Hillary Clinton and Donald Trump made some inaccurate	REAL
2909	Iran reportedly makes new push for ura	Iranian negotiators reportedly have made a last-ditch	REAL
1357	With all three Clintons in Iowa, a glimps	CEDAR RAPIDS, Iowa â€” â€œI had one of the most	REAL
988	Donald Trumpâ€™s Shockingly Weak De	Donald Trumpâ€™s organizational problems have gone	REAL
7041	Strong Solar Storm, Tech Risks Today	Click Here To Learn More About Alexandra's	FAKE

Do you trust all the news you hear from social media?

All news are not real, right?

How will you detect fake news?

The answer is Python. By practicing this advanced python project of detecting fake news, you will easily make a difference between real and fake news.

Before moving ahead in this machine learning project, get aware of the terms related to it like fake news, tfidfvectorizer, PassiveAggressive Classifier.

### 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.

# What is a TfidfVectorizer?

**TF (Term Frequency):** The number of times a word appears in a document is its Term Frequency. A higher value means a term appears more often than others, and so, the document is a good match when the term is part of the search terms.

**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. The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF features.

# What is a PassiveAggressiveClassifier?

Passive Aggressive algorithms are online learning algorithms. Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.

# Detecting Fake News with Python

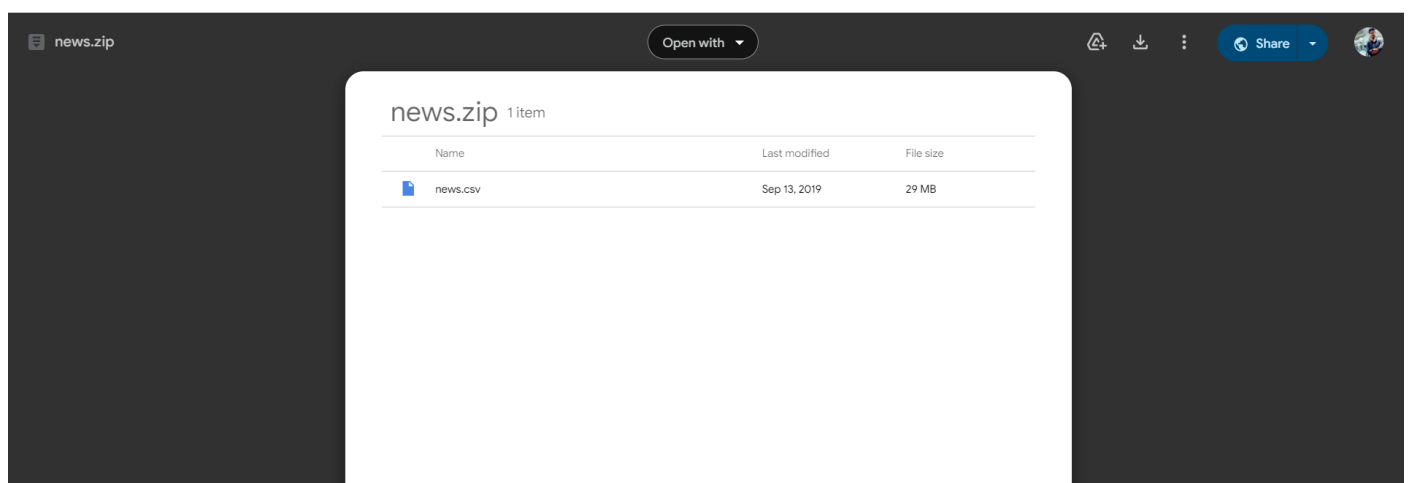
To build a model to accurately classify a piece of news as REAL or FAKE.

# About Detecting Fake News with Python

This advanced python project of detecting fake news deals with fake and real news. Using sklearn, we build a TfidfVectorizer on our dataset. Then, we initialize a PassiveAggressive Classifier and fit the model. In the end, the accuracy score and the confusion matrix tell us how well our model fares.

# The fake news Dataset

The dataset we'll use for this python project- we'll call it news.csv. This dataset has a shape of 7796×4. The first column identifies the news, the second and third are the title and text, and the fourth column has labels denoting whether the news is REAL or FAKE. The dataset takes up 29.2MB of space and you can [download it here](#).



# Project Prerequisites

You'll need to install the following libraries with pip:

```
1. pip install numpy pandas sklearn
```

You'll need to use Google Colab to run your code. Get into files and import the Dataset into it:

## Steps for detecting fake news with Python

Follow the below steps for detecting fake news and complete your first advanced Python Project –

1. Make necessary imports:

```
[ ] import numpy as np
import pandas as pd
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. Now, let's read the data into a Data Frame, and get the shape of the data and the first 5 records.

```
[5] df=pd.read_csv('news.csv')
#Get shape and head
df.shape
df.head()
```

	Unnamed: 0		title	text	label
0	8476		You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...		Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy		U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...		FAKE
4	875	The Battle of New York: Why This Primary Matters		It's primary day in New York and front-runners...	REAL

3. And get the labels from the Data Frame.

```
[6] #DataFlair - Get the labels
labels=df.label
labels.head()
```

```
0    FAKE
1    FAKE
2    REAL
3    FAKE
4    REAL
Name: label, dtype: object
```

4. Split the dataset into training and testing sets.

```
[7] #DataFlair - Split the dataset
x_train,x_test,y_train,y_test=train_test_split(df['text'], labels, test_size=0.2, random_state=7)
```


5. Let's initialize a TfidfVectorizer with stop words from the English language and a maximum document frequency of 0.7 (terms with a higher document frequency will be discarded). Stop words are the most common words in a language that are to be filtered out before processing the natural language data. And a TfidfVectorizer turns a collection of raw documents into a matrix of TF-IDF features.

Now, fit and transform the vectorizer on the train set, and transform the vectorizer on the test set.

```
✓ 6s #DataFlair - Initialize a TfidfVectorizer
tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
#DataFlair - Fit and transform train set, transform test set
tfidf_train=tfidf_vectorizer.fit_transform(x_train)
tfidf_test=tfidf_vectorizer.transform(x_test)
```


6. Next, we'll initialize a PassiveAggressiveClassifier. This is. We'll fit this on `tfidf_train` and `y_train`.

Then, we'll predict on the test set from the TfidfVectorizer and calculate the accuracy with `accuracy_score()` from `sklearn.metrics`.

✓ 0s  #DataFlair - Initialize a PassiveAggressiveClassifier  
pac=PassiveAggressiveClassifier(max\_iter=50)  
pac.fit(tfidf\_train,y\_train)  
#DataFlair - Predict on the test set and calculate accuracy  
y\_pred=pac.predict(tfidf\_test)  
score=accuracy\_score(y\_test,y\_pred)  
print(f'Accuracy: {round(score\*100,2)}%')

Accuracy: 92.98%

7. We got an accuracy of 92.82% with this model. Finally, let's print out a confusion matrix to gain insight into the number of false and true negatives and positives.

✓ 0s  confusion\_matrix(y\_test,y\_pred, labels=['FAKE','REAL'])  
  
array([[591, 47],  
 [ 42, 587]])

## Result

We learned to detect fake news with Python. We took a political dataset, implemented a TfidfVectorizer, initialized a PassiveAggressiveClassifier, and fit our model. We ended up obtaining an accuracy of 92.82% in magnitude.