

Also, I like to add that DataFlair has published a ***series of machine learning Projects*** where you will get interesting and open-source advanced ml projects. Do check, and then share your experience through comments.

Here is the list of top Python projects:

1. Fake News Detection Python Project
2. [Parkinson's Disease Detection Python Project](#)
3. [Color Detection Python Project](#)
4. [Speech Emotion Recognition Python Project](#)
5. [Breast Cancer Classification Python Project](#)
6. [Age and Gender Detection Python Project](#)
7. [Handwritten Digit Recognition Python Project](#)
8. [Chatbot Python Project](#)
9. [Driver Drowsiness Detection Python Project](#)
10. [Traffic Signs Recognition Python Project](#)

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

11. [Image Caption Generator Python 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.

## 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

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

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 install Jupyter Lab to run your code. Get to your command prompt and run the following command:

```
1. C:\Users\DataFlair>jupyter lab
```

You'll see a new browser window open up; create a new console and use it to run your code. To run multiple lines of code at once, press Shift+Enter.

## Steps for detecting fake news with Python

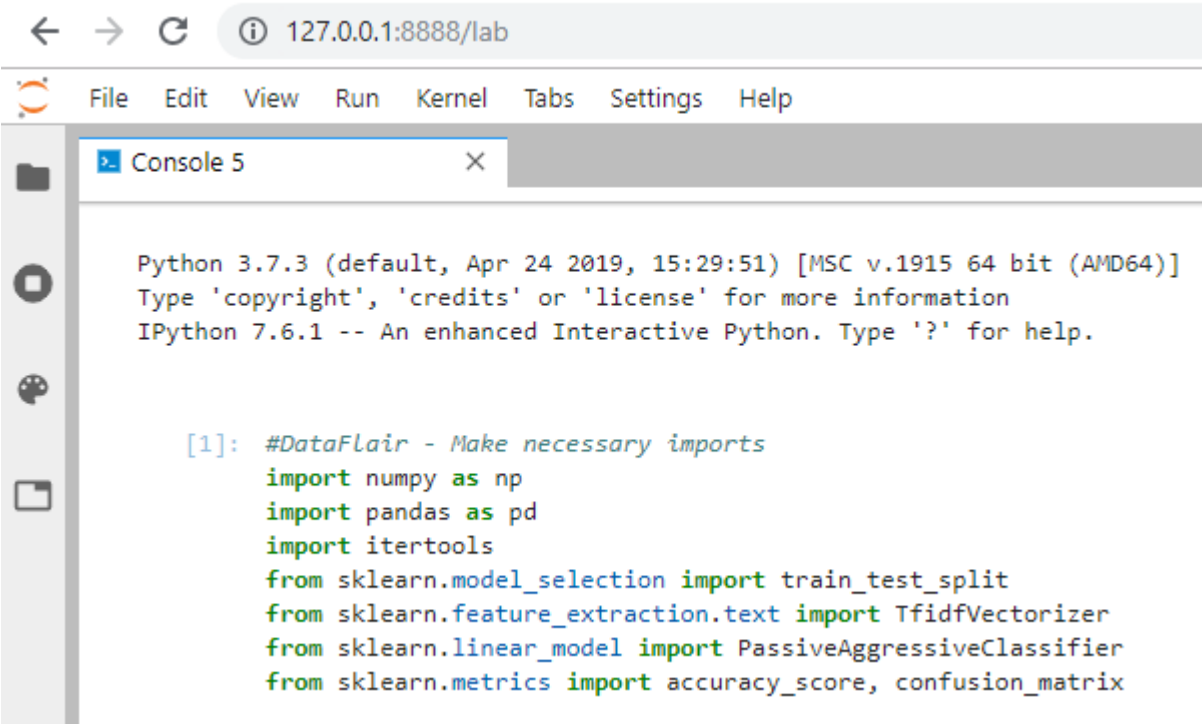
Follow the below steps for detecting fake news and complete your first advanced Python Project –

1. Make necessary imports:

```
1. import numpy as np
2. import pandas as pd
3. import itertools
4. from sklearn.model_selection import train_test_split
5. from sklearn.feature_extraction.text import TfidfVectorizer
6. from sklearn.linear_model import PassiveAggressiveClassifier
7. from sklearn.metrics import accuracy_score, confusion_matrix
```

**Screenshot:**

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+



2. Now, let’s read the data into a DataFrame, and get the shape of the data and the first 5 records.

```

1. #Read the data
2. df=pd.read_csv('D:\\DataFlair\\news.csv')
3.
4. #Get shape and head
5. df.shape
6. df.head()
  
```

### Output Screenshot:

```

[2]: #Read the data
df=pd.read_csv('D:\\DataFlair\\news.csv')

#Get shape and head
df.shape
df.head()
  
```

```
[2]:
```

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 DataFrame.

```

1. #DataFlair - Get the labels
2. labels=df.label
  
```

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

```
3. labels.head()
```

Output Screenshot:

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

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

4. Split the dataset into training and testing sets.

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

Screenshot:

```
[4]: #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.

```
1. #DataFlair - Initialize a TfidfVectorizer
2. tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
3.
4. #DataFlair - Fit and transform train set, transform test set
5. tfidf_train=tfidf_vectorizer.fit_transform(x_train)
6. tfidf_test=tfidf_vectorizer.transform(x_test)
```

Screenshot:

```
[5]: #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)
```

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

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.

```
1. #DataFlair - Initialize a PassiveAggressiveClassifier
2. pac=PassiveAggressiveClassifier(max_iter=50)
3. pac.fit(tfidf_train,y_train)
4.
5. #DataFlair - Predict on the test set and calculate accuracy
6. y_pred=pac.predict(tfidf_test)
7. score=accuracy_score(y_test,y_pred)
8. print(f'Accuracy: {round(score*100,2)}%')
```

Output Screenshot:

```
[6]: #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: 92.82%
```

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.

```
1. #DataFlair - Build confusion matrix
2. confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
```

Output Screenshot:

```
[7]: #DataFlair - Build confusion matrix
confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])

[7]: array([[589, 49],
          [ 42, 587]], dtype=int64)

[ ]:
```

So with this model, we have 589 true positives, 587 true negatives, 42 false positives, and 49 false negatives.

# Summary

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

Today, 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.

Hope you enjoyed the fake news detection python project. Keep visiting DataFlair for more interesting python, data science, and machine learning projects.

Did you like our efforts? If Yes, please give DataFlair 5 Stars on [Google](#) | [Facebook](#)

Tags: [Advanced python project](#) [Python data science projects](#) [python machine learning projects](#) [python mini projects](#) [Top Python Projects](#)

171 RESPONSES

 Comments 17  Pingbacks 0

SAYEED  [November 12, 2021 at 5:25 pm](#)

how to create dataset?

[Reply](#)

Shraddha  [December 2, 2021 at 11:41 am](#)

I am getting following error, Kindly help.

```
ValueError Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_388\4183239203.py in
3
4 # Fit and transform the vectorizer on the train set, and transform the vectorizer on the test set
--> 5 tfidf_train = tfidf_vectorizer.fit_transform(x_train)
6 tfidf_test = tfidf_vectorizer.transform(x_test)

~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in fit_transform(self, raw_documents, y)
1844 """
1845 self._check_params()
-> 1846 X = super().fit_transform(raw_documents)
1847 self._tfidf.fit(X)
1848 # X is already a transformed view of raw_documents so

~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in fit_transform(self, raw_documents, y)
1200 max_features = self.max_features
1201
-> 1202 vocabulary, X = self._count_vocab(raw_documents,
```

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

```
1203 self.fixed_vocabulary_)
1204
~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in _count_vocab(self, raw_documents, fixed_vocab)
1104 vocabulary.default_factory = vocabulary.__len__
1105
-> 1106 analyze = self.build_analyzer()
1107 j_indices = []
1108 indptr = []

~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in build_analyzer(self)
425
426 elif self.analyzer == 'word':
-> 427 stop_words = self.get_stop_words()
428 tokenize = self.build_tokenizer()
429 self._check_stop_words_consistency(stop_words, preprocess,

~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in get_stop_words(self)
358 A list of stop words.
359 """
-> 360 return _check_stop_list(self.stop_words)
361
362 def _check_stop_words_consistency(self, stop_words, preprocess, tokenize):

~\anaconda3\lib\site-packages\sklearn\feature_extraction\text.py in _check_stop_list(stop)
177 return ENGLISH_STOP_WORDS
178 elif isinstance(stop, str):
-> 179 raise ValueError("not a built-in stop list: %s" % stop)
180 elif stop is None:
181 return None

ValueError: not a built-in stop list: English
```

[Reply](#)

**wesam**  [December 30, 2021 at 6:27 pm](#)

Do I get the source code please?

[Reply](#)

**vishal**  [January 11, 2022 at 9:23 am](#)

i didnt get any source code

[Reply](#)

**DataFlair Team**  [January 11, 2022 at 12:06 pm](#)

Source code of this fake news project is published in the article, please copy and use

[Reply](#)



Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

**Vishwas Kshirsagar**  [January 12, 2022 at 12:48 pm](#)

# So with this model, we have 589 true positives, 587 true negatives, 42 false positives, and 49 false negatives.

this is your statement. I believe you have misinterpreted, you have given (y\_test, pred\_y) in the confusion matrix which justifies 589 is True Negative, 587 True Positive, 49 False Positive and 42 False Negatives.

Please reconfirm.

Thank you for the article it is worth of million.

[Reply](#)

**Rumman**  [February 2, 2022 at 4:34 pm](#)

I also have same doubt

[Reply](#)

**kamlesh zore**  [January 30, 2022 at 1:10 pm](#)

Hi , can you please help me with below error

ValueError: Found input variables with inconsistent numbers of samples: [5068, 1267]

[Reply](#)

**387420489**  [March 19, 2022 at 6:39 pm](#)

How can I use this on other news articles? I mean, if I find an article online and I want to see what percentage chance is there that it’s fake news?

[Reply](#)

**Wangari Thuo**  [March 30, 2022 at 10:28 pm](#)

i am getting this error. help

```
-----
ModuleNotFoundError Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_9696\3872586496.py in
--> 1 import intertools
2 from sklearn.model_selection import train_test_split
3 from sklearn.feature_extraction.text import TfidfVectorizer
4 from sklearn.linear_model import PassiveAggressiveClassifier
5 from sklearn.metrics import accuracy_score, confusion_matrix
```

ModuleNotFoundError: No module named ‘intertools’

pleasee

[Reply](#)

**Maiko**  [April 9, 2022 at 3:41 pm](#)

you misspelled itertools as intertools.

[Reply](#)

**alicia**  [April 20, 2022 at 12:38 am](#)

Python Projects	+
Python Django Projects	+
Machine Learning Proje...	+
Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+

tfidf\_train=tfidf\_vectorizer.fit\_transform(x\_train)

tfidf\_test=tfidf\_vectorizer.transform(x\_test)

I kept receiving an error message when running this

[Reply](#)

**harikrishnaraj**  [May 5, 2022 at 1:54 pm](#)

package install is not check the installation once

[Reply](#)

**Hassan**  [June 19, 2022 at 12:43 am](#)

How to give input to model for prediction

[Reply](#)

**Pierre**  [July 13, 2022 at 5:40 am](#)

Hello,

If I understood correctly, we have a dataset with a text and a label to know if the text is truthful or sensationalist. One of the methods to find the truthfulness of a text is to calculate the frequency of words in the text.

What I don't understand is what we are supposed to do with it. And how the PassiveAggressiveClassifier impacts this.

[Reply](#)

**Mohammad Salameh**  [August 27, 2022 at 11:13 pm](#)

I have this as project for a data mining course, I hate my professor, but thank you all for the article it really helped a lot <3

[Reply](#)

**ae**  [December 8, 2022 at 7:26 pm](#)

satara is the biggest city

[Reply](#)

[« Older Comments](#)

LEAVE A REPLY

Comment \*

Name \*

Email \*

Post Comment



Deep Learning Projects	+
AI Projects	+
Python Interview Questi...	+
Python Quiz	+