**Department of Electronics and Communication**

Machine Learning Mini Project – Report

Course Code- UE17EC337

# Topic- NEWS VALIDATON using Python

Group members

Rohan Ajay – PES1201700454

M. Tejeshwar – PES1201700675

Ajay Raj K T – PES1201700685

**INDEX**

|  |
| --- |
| * Introduction * Objective and Problem Statement * Tools , Modules and Concepts used * Methodology * Implementation * Result and Analysis * Conclusion * Future scope * References * Code and output |

**Introduction**

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 vitally

used by algorithms, and users may end up in a

filter bubble.

By practicing this advanced python project of

detecting fake news, we can easily sort out the real news from the fake ones.

**Problem Statement :**

Build a model to accurately classify a piece of news as REAL or FAKE using Python.

**Objectives :**

* Obtain a valid dataset (.csv file).
* Split the dataset for training and testing of the ML model.
* Train the model using a selected range of the data.
* Using this, predict the outcome of the test dataset.
* Compare the predicted values with test values to obtain an accuracy score.
* Plot confusion matrix for the same.

# Tools used :

* Python 3
* Jupyter Lab

**Modules Used** :

1. **Numpy** - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
2. **Pandas** - Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

.

1. **Sklearn** - Skearn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
2. **Matplotlib** - Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits.
3. **Itertools** - Functions creating iterators for efficient looping.

**Concepts used :**

* **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.
* **Passive Aggressive Classifier**:An algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting.It doesn’t converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.
* **Accuracy Score :** It is a function that compares the predicted values to the actual test values of out dataset and returns the accuracy of our ML model.
* **Confusion Matrix :** - The confusion matrix is a 2x2 matrix which gives us the number of True Positives, False positives, False Negatives and True Negatives of predictions made by an ML model,

**Methodology :**

* This project deals with fake and real news.
* Using sklearn, we build a TfidfVectorizer on our dataset.
* The TfidfVectorizer converts a collection of raw documents into a matrix of TF-IDF features.
* 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 dataset:**

The dataset we have for this python project is named “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.

Link to the dataset :

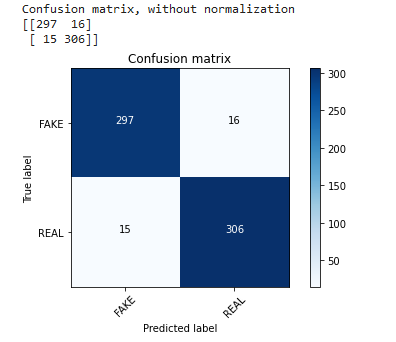
<https://tinyurl.com/NewsDataSet>

**Implementation:**

* Downloaded necessary libraries and modules specified above.
* Implemented the code on Jupyter Notebook.

**Results:**

* Obtained an accuracy score of 95.11% for a train to test ratio of 9:1.
* The confusion matrix obtained was :



**Analysis:**

* Accuracy improved as we decreased test size.
* Accuracy kept on increasing as on multiple executions of the code.

**Conclusions:**

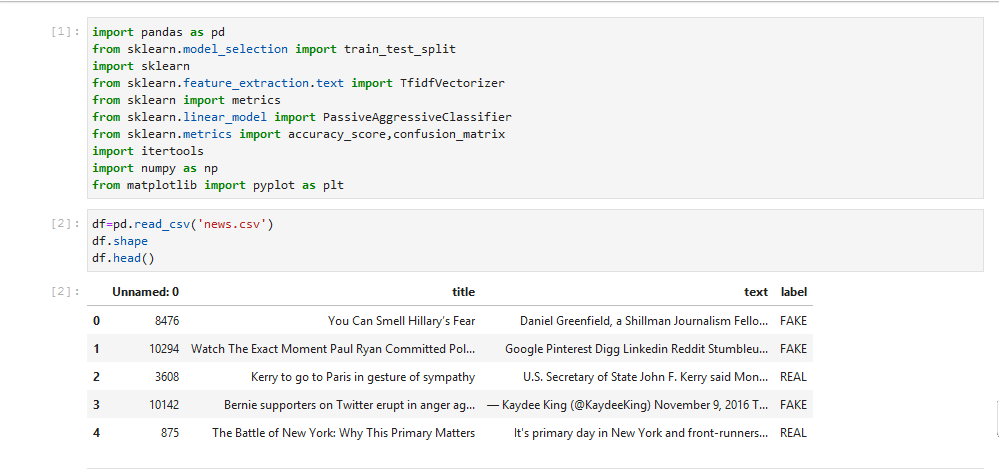
* Built a ML model and obtained a confusion matrix and an accuracy score successfully.
* Accuracy improves by providing a larger training dataset and multiple iterations.

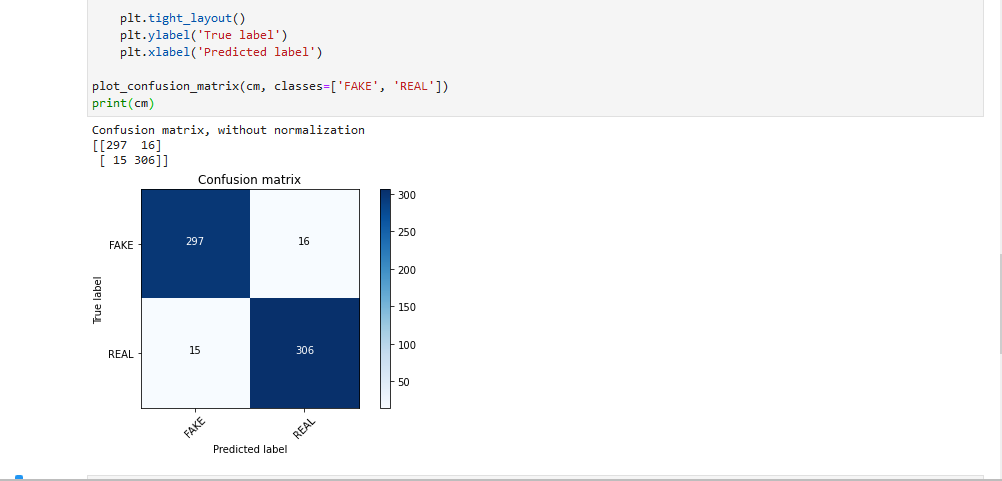
**Future Scope:**

* Can include user entered news by appending it to the CSV file.
* Can predict and display the result of a particular entry of the dataset.
* Can improve accuracy by adding enough entries to the dataset.

**References :**

1. <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=2ahUKEwjo2JLO9u_oAhXdwcQBHRKYBS8QFjACegQIAxAB&url=https%3A%2F%2Fjupyter.readthedocs.io%2Fen%2Flatest%2Finstall.html&usg=AOvVaw0QqiNjt854nhSa_prfu5Qn>
2. <https://docs.python.org/3/installing/index.html>
3. <https://www.youtube.com/watch?v=HW29067qVWk>
4. https://tinyurl.com/NewsDataSet

**Code and outputs :** 



-----X-----X-----x-----