**A PROJECT REPORT**

**ON**

**FAKE NEWS DETECTOR**

**USING MACHINE LEARNING ALGORITHM**

**PASSIVE AGGRESSIVE CLASSIFIER**

Submitted as a part of

TRAINING

IN

Data Analytics, Machine Learning using Python



Efforts By :

**Nidhi Gupta** ( GLA UNIVERSITY , MATHURA )

Under the guidance and teachings of

**Mr. Bipul Shahi**

***ACKNOWLEDGEMENT***

**P**resenting the ascribed project paper in this very simple and official form , I would like to place my deep gratitude and thanks to Mr. Bipul Shahi ,our coordinate trainer and developer. He has been teaching us the concept of Machine Learning in a very efficient way giving us real life examples and applications and more over teaching things graphically with all this experience and patience.

Without his help, I wouldn’t have been able to complete this project. In this quantarine period, this has been my best learning and all the credit goes to him for arranging all the lectures, proper content ,recordings ,and clearing our every doubt that occurred. And at last but not the least I would like to thank my dear parents for helping me to grab this opportunity to get trained with Dignique Labs.

Thanking You.

***Introduction to The Problem Statement: Fake News Detector***

**What is fake news?**

Fake news has been a society problem, used to propagate false or rumour information in order to change people’s behaviour. It has been there with its hoaxes since before the advent of the Internet.

The widely accepted definition of fake news is :”fictious articles deliberately fabricated to deceive the readers”. Social media and news outlets publish fake news to increase the readership or as a part of psychological warfare. Clickbaits lure users with flashy headlines.

**Why Detection?**

Fake news can create chaos among the real world as it can change the mindset of the people and harm the people in ways that is beyond psychological imagination.

So the purpose of this project paper could be used to come up with a solution so as to detect and filter out sites containing false and misleading information.

**TECHNOLOGY AND CONCEPTS USED**

**MACHINE LEARNING**

Learning algorithms are widely used in computer vision applications. Before considering the model related tasks, we are going to have a brief look at basics of machine learning.

Machine learning has emerged as a useful tool for modelling problems that are otherwise difficult to formulate exactly. Classical computer programs are explicitly programmed by hand to perform a task. With machine learning, some portion of the human contribution is replaced by a learning algorithm. As availability of computational capacity and data has increased, machine learning has become more and more practical over the years, to the point of being almost ubiquitous.

It can be used in two ways:

* *Supervised Learning*
* *Unsupervised Learning*

So machine learning is basically an art of providing artificial intelligence to the computer system by means of some features that can be observed**. Features** are the data that is collected that help to solve the problem (input to the model).

And what we get is the answer generally called the **Label** (output /Observation from the model)

**. Supervised Learning:**

Goes under the supervision making the machine learn on the existing data (Feature and the Labels).Then we make predictions on the trained data and hence check its accuracy. Hence forth we make predictions on the testing data and check the accuracy of the model. And optimize the error every time to reach the desired accuracy.

**. Unsupervised Learning:**

This goes the same method as for supervised learning except for the fact that instead of comparing the label with the predictions here the features are being grouped according to the behaviour and clusters of the same features are being created.

There are various Machine Learning Algorithms we can use based upon the problems and accuracy we received from the models created and as per now for the fake news detector ,the algorithm I am using is the Passive Aggressive Classifier and the Sklearn library.

***Feature Extraction***

Now that the fake news has been defined and the target has been set as the data is being exported from [*www.kaggle.com*](http://www.kaggle.com) ,it is needed to analyse what features can be used in order to classify fake news. We have four raw components:-

1. **Source**: Where does the news come from, who wrote it, is this source reliable or not
2. **Headline**: Short summary of the news content that tries to attract the reader.
3. **Body Text**: the actual context of the text.
4. **Images/Vedios**: if present.

***Algorithm Used***

*Passive Aggressive Classifier:*

Passive Aggressive Algorithms are online learning algorithms. Such an algorithm remains passive for correct Classification outcome, and turns aggressive in the event of miscalculation, updating and adjusting. Unlike most other algorithm 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.

***TF( TERM FREQUENCY ) VECTORIZER:***

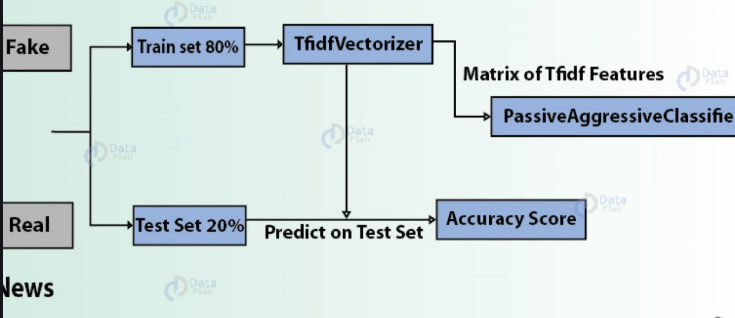
The number of times the word appear in the document is the Term Frequency .A higher value means a term appear more often than others, and so the document is the good match when the term is the part of search frequencies.

***IDF (INVERSE DOCUMENT FREQUENCY ) VECTORIZER:***

Words that occur many times a document, but may occur many time in many others, may be irrelevant. IDF is the measure of how significant a term is in the entire corpus.

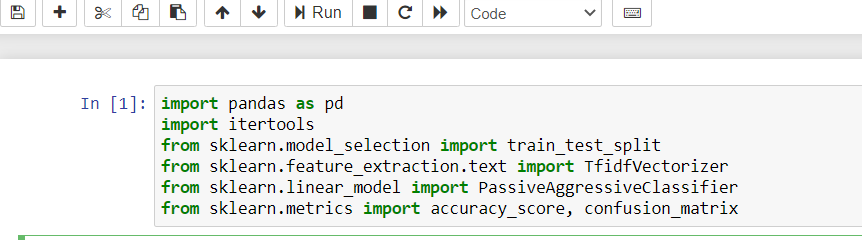
The TFIDF Vectorizer converts a collection of raw documents into the matrix of TD-IDF features.

*Below lies an image ,which clearly explain the process.*



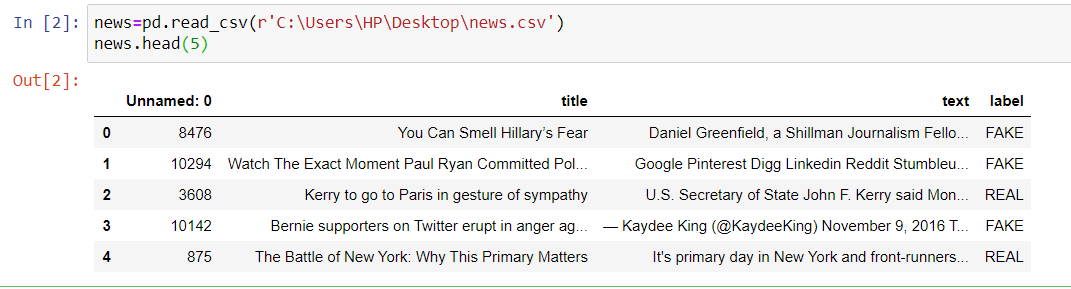
***Steps for preparing the model:***

***Step 1:*** Firstly we will import all the python libraries we’ll be using in this model.



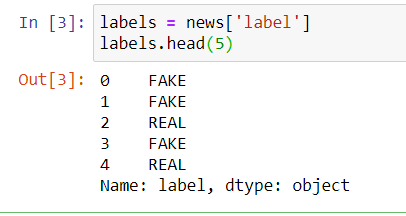
**Step 2:**

Now we’ll be reading the dataset which contains a list of features and the label and display the first five lines to give an idea of how it looks.



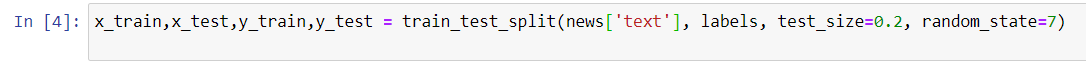
**Step 3:**

Now we’ll get the labels from the dataset

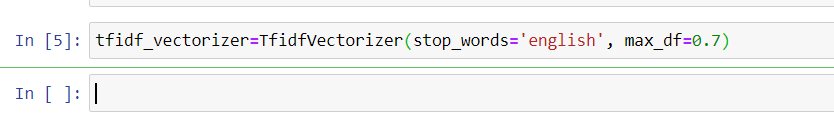


**Step 4:**

From our dataset , we ‘ll divide our data into two –the training part and the testing part .first we’ll train our data from the training part and then test.

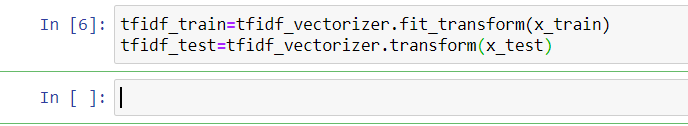
**Step 5:**

We are initializing TFIDF Vectorizer.  Stop words are the most common words in a language that are to be filtered out before processing the natural language data. And all the words with document frequency greater than 0.7 will be discarded(considered fake).



***Step 6:***

We fit the vectorizer on the train set and transform it on the training set.

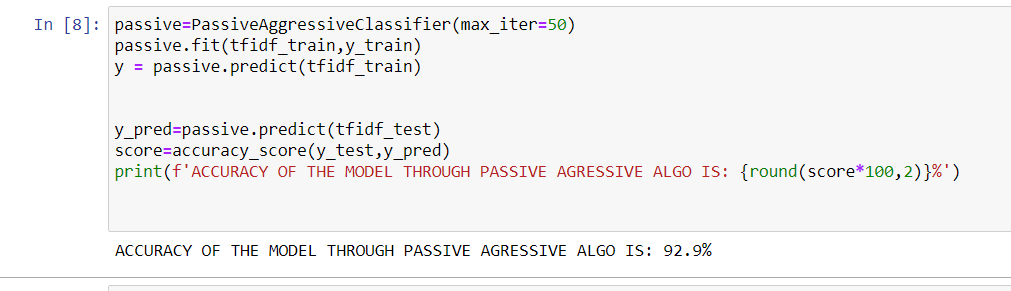


**Step 7:**

We’ll initialize a passive Agressive Classifier .we’ll predict them on the test set and calculate accuracy.

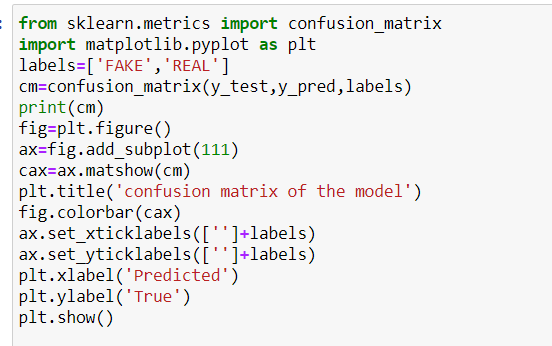
And the accuracy we so obtained from this model is 93% approximately.

To get a more clearer picture of the result we’ll show a confusion matrix.

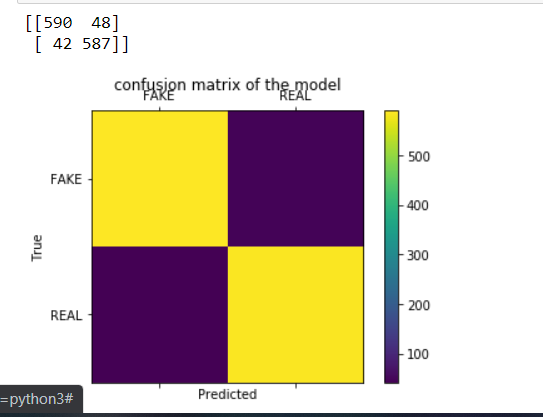


**Step 8:**

Confusion Matrix



Output:



***Conclusion***

This project paper has given an algorithm showing its implementation to detect a fake news and a real one. Not only this single algorithm but we can use many other machine learning algorithm to implement the same model.

All they differ is in accuracy.in this model I have taken 80% of the data as training data for the machine and 20% data for testing of the model and got 93% accuracy.

Machine learning has its various applications as well in diferent fields:

1. Machine Learning in Education
2. Machine Learning in Search Engine
3. In Digital Marketing
4. In health Care
5. Data Mining and automation of tasks

So in today’s world, Machine Learning is an application of Artificial Intelligence.

Google says, ”Machine Learning is the future” and it’s really gonna be very bright. As humans are getting more and more dependent and addicted to machines, we are to witness a new revolution which will take over the world

And become its future…

***References:-***

**1.***www.google.com/fake\_news\_article*

*This is article about what fake news actually is.*

*2.www.kaggle.com*

*Source of the dataset I have worked upon*

*3.www.github.com/nidhi2611*

*This is the link to the official repository of my project where the code has been wriiten.*