Fake News & its Virality

The influence and simplicity of use of social networks has transformed the creation and dissemination of knowledge in today's society. However, easier access to information does not imply an increase in public awareness. In addition, unlike traditional media outlets, social networks enable the spread of misinformation to be more rapid and widespread. The viral spread of incorrect information has major consequences for the public's behaviours, attitudes, and beliefs, and can eventually jeopardise democratic processes. The key difficulty that researchers face today is limiting the detrimental impact of misleading information by early detection and control of its transmission.

Aim:

This project aims to build a model to accurately classify a piece of news as REAL or FAKE.

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 in 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.

Process :
Using sklearn, we build a TfidfVectorizer on our dataset. Then, we initialise a PassiveAggressive Classifier and fit the model. In the end, the accuracy score and the confusion matrix tell us how well our model fares.

Fake News Data Set : 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.