**FAKE NEWS DETECTION USING NATURAL LANGUAGE PROCESSING (NLP)**



**Name: Prarthana B R**

**Reg.No: 513521104037**

**Department: CSE**

**Year: III**

**NM ID: au513521104037**

**E-Mail: prarthuravi13@gmail.com**

**Phase 3 (Development Part -1)**

**INTRODUCTION:**

The spreading of fake news causes many problems in the society. It easily deceives people and leads to confusion among them. It has the ability to cause a lot of social and national damage with destructive impacts. Sometimes it gets very difficult to know if the news is genuine or fake. Therefore it is very necessary to detect if the news is fake or not.

**Natural Language Processing (NLP)** and deep learning techniques are powerful tools for fake news detection. We tend to use **LSTM approach** for this project.

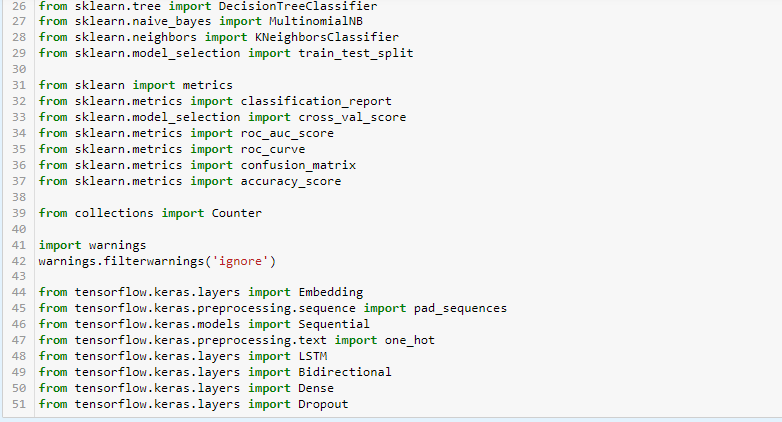
In the continuation of the previous documentation AI\_Phase2, we will be looking forward on the development part of the project. In this phase, we will start building the model by **loading the dataset and processing the data.** It will also give a graphical representation of the amount of fake and true news obtained.

The following steps are followed in this development part.

1. **Importing Libraries:**

In this step we will import all the required libraries and packages from the pre-installed modules.

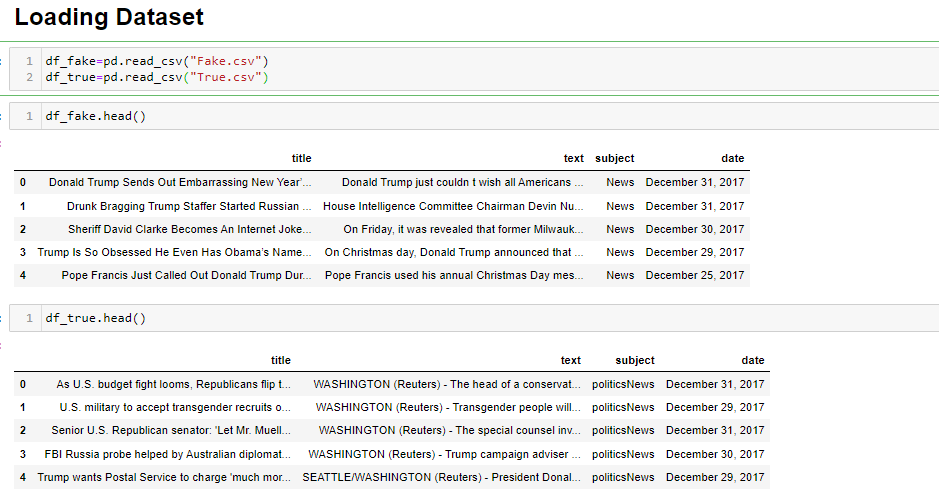




1. **Loading the datasets:**

In this step we will be loading the Fake and True datasets acquired from Kaggle. (<https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset>)

* pd.read\_csv(<filename>) is used to load the datasets.
* df\_<filename>.head() is used to view the datasets.
* The above function will display the 1st 5 entries from the respective datasets.
* Similarly, by passing the number of entries to be viewed as the function argument, we can view the respective number of entries.



1. **Data Cleaning:**

* In this step we will be assigning the labels for the fake and true datasets to specify their authenticity.
* We label fake data to be “F” and true data to be “T”.



* After labelling, we will merge those datasets to form a single dataset.
* As we have used labels for fake and true data, we can distinguish them

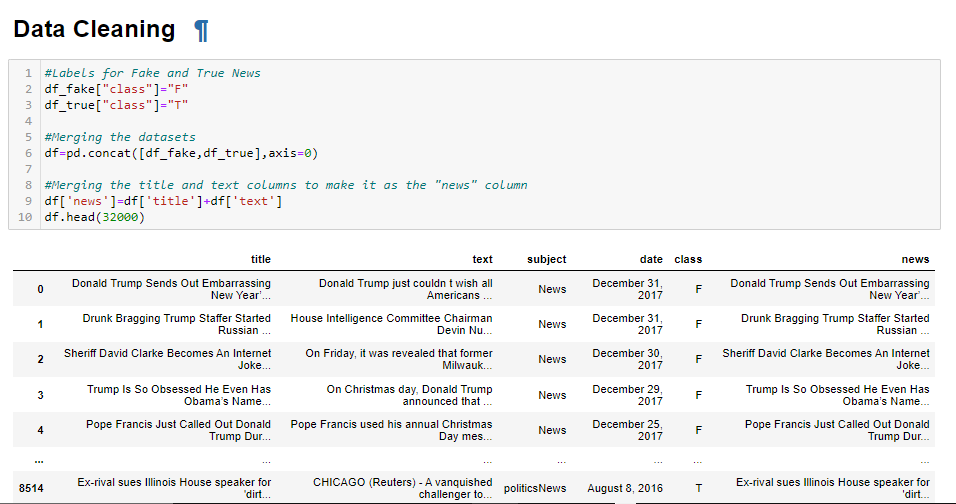
easily.



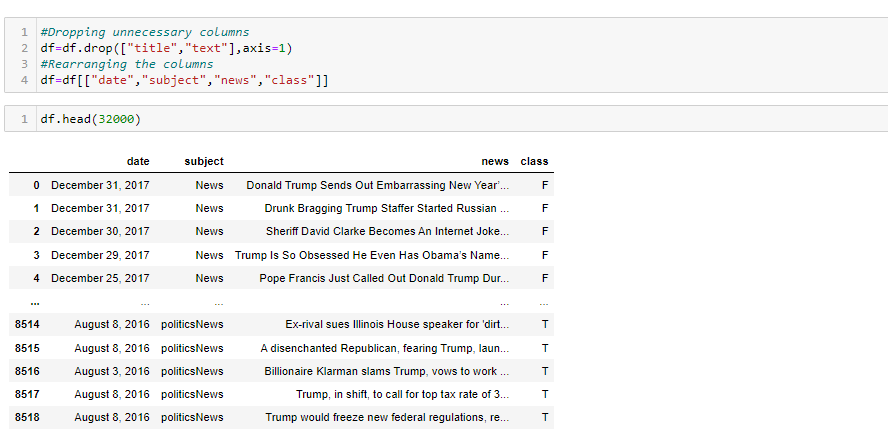
* Now, we will merge the “title” and “text” columns and make it as “news” column. This is to ease the process of detection.



* On performing these, we will get output as



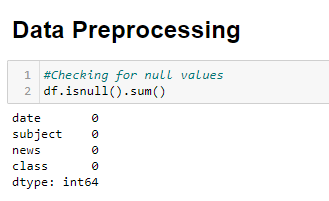
* After merging the datasets and columns we will drop the unnecessary columns from the merged datasets and rearrange them in an order.



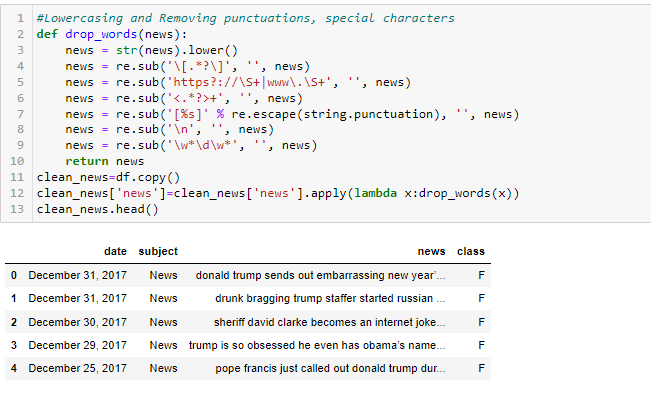
* df.head(32000) is used to view 32000 entries from the dataset with their respective labels.

1. **Data Pre-processing:**

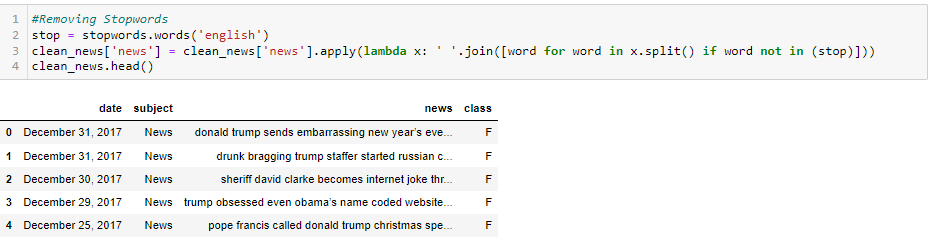
* We will be pre-processing the data for the further development of model
* Firstly we will be checking for the presence of any null values in dataset, as they may interrupt in the evaluation of model.



* Since there are no null values we can proceed with the next step.
* If any null values are found we have to replace them by using fillna().
* Secondly, we will lower case all the fonts in the datasets. We also remove the special characters and punctuations from the dataset as they may be a hindrance in feeding the model.

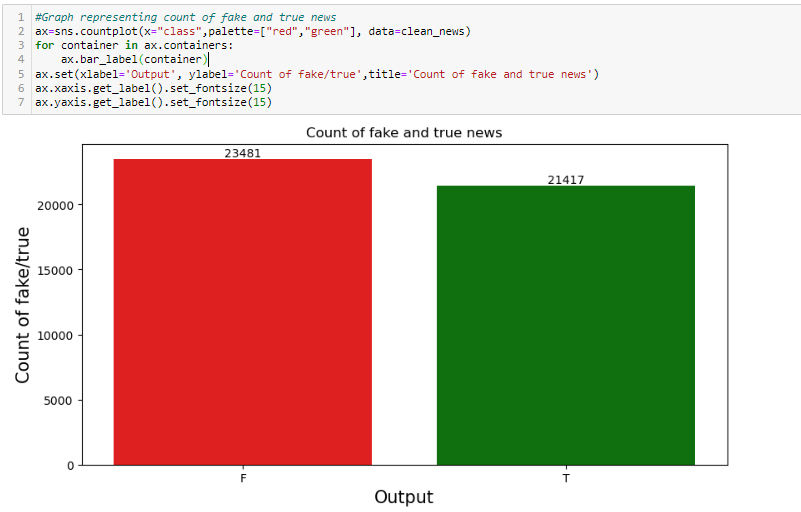


* From above, we can infer that all the letters in the dataset have been lowercased, special characters and punctuations have been removed.
* Finally, we will remove the stop words from the dataset.
* A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that can be ignored when retrieving the data. We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them, by storing a list of words that you consider to stop words.



1. **Dataset Analysis:**

* This is an additional step to have an idea about the news on the datasets.
* Firstly, let’s see the count of fake and true news from the dataset.



* We have a pretty much balanced data. But the count of fake news is higher than the true news but not on a greater extent.
* Next let’s see the count of news based on the subject of the news.

