Fake News Detection

13

Fake News Detection

Introduction:

Misinformation presents a huge challenge in online society. As a result, there have been many attempts to identify and classify misinformation. Specifically, in social networking sites, blogs, as well as online newspapers.

Since this is a very broad research area, in this documentation we'll look at:

Natural language processing to classify the news if it is fake or real based on the title and content of the news

Methodology:

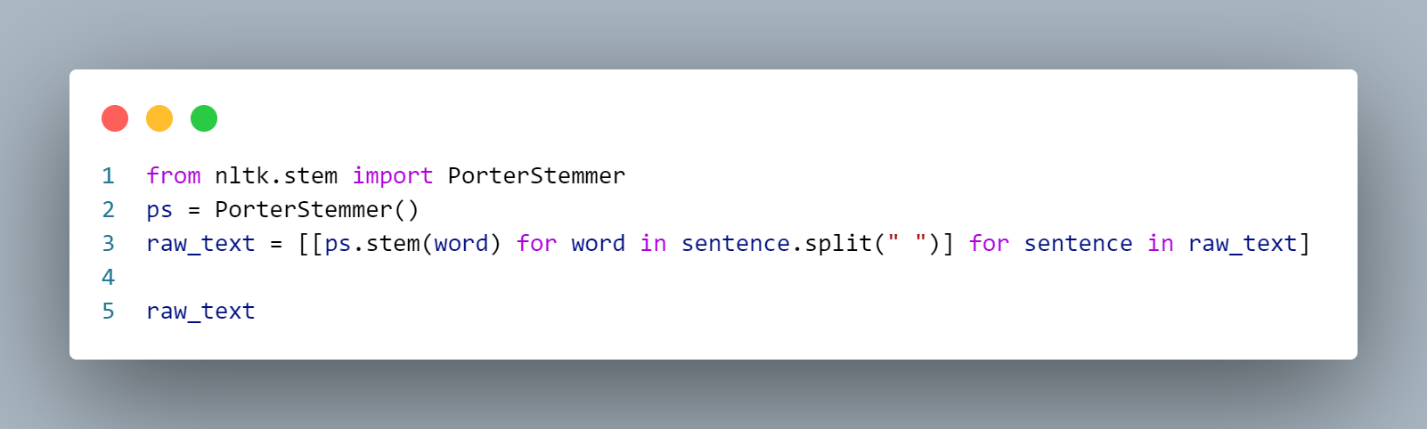
***Preprocessing:***

***1-Preparing data:***



we found out that there is a column doesn’t affect the classification, so we dropped it and encoded the values (“ fake ”,” real ”) to 0s and 1s to be easily classified, and we concatenated the text and title column to be one feature and use feature extraction on it.

***2-Stemming:***



We used the porter stemmer to stem the data and reduce the number of words to deal with in our corpus.

***3- Tokenization and Stop words and punctuation removal:***

******

Tokenize the combined columns title and text and remove the stop words and punctuation to be ready for feature extraction

***Features Extraction:***

******

Used word2vector to extract features by giving every word a vector represents it

***Hyperparameter Tuning:***

******

Using grid search to find the best hyperparameters for the Logistic Regression model to obtain the highest accuracy

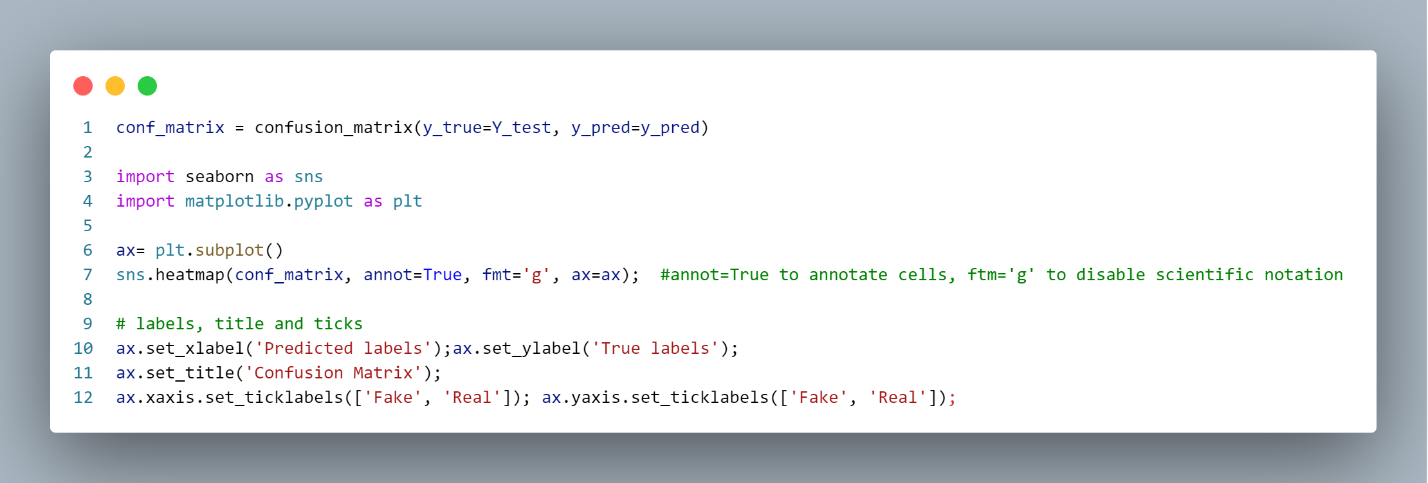
***Model Training:***



Splitting the dataset to test and training then train the logistic regression model using the train dataset then testing the accuracy

***Visualization: Graphical user interface, text

Description automatically generated***



Data Set Summary:

1-What is the data set used?

Used the dataset provided with the project which is “news”

2- What is the summary of the dataset columns?

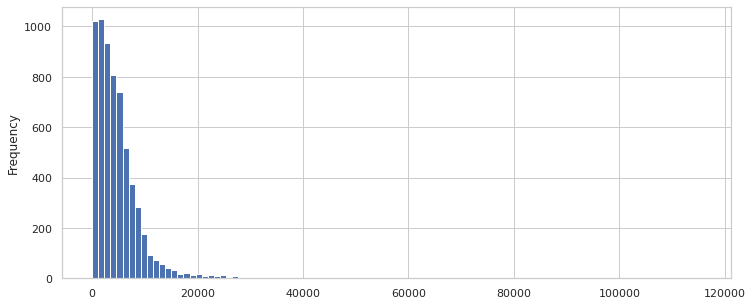
First column dropped because it was not related to data .

Title and text Columns were merged to be used in the preprocessing process .

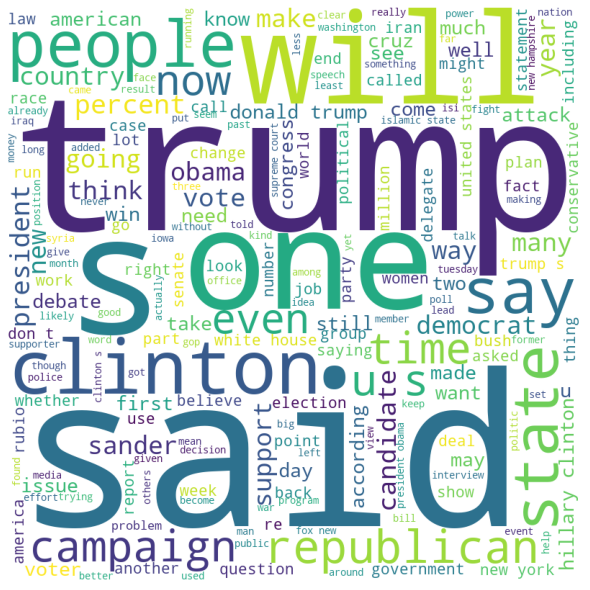
Label column was encoded to 0s for fake and 1s for real.

3- Visualize the dataset statistics\*/

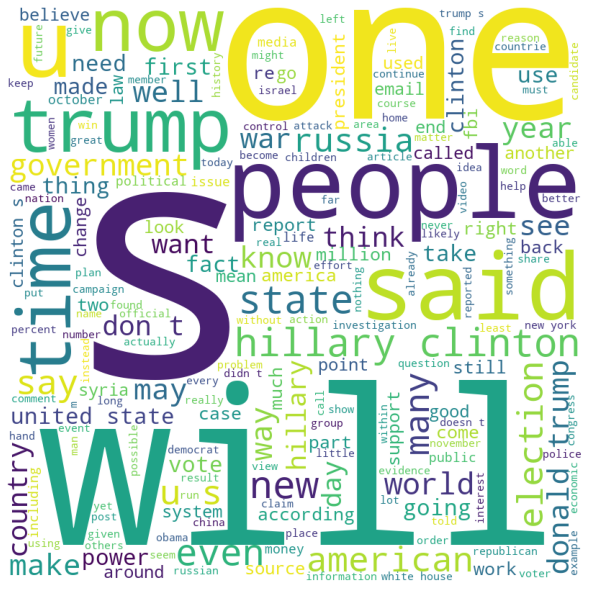
-Frequency Statistic for the text



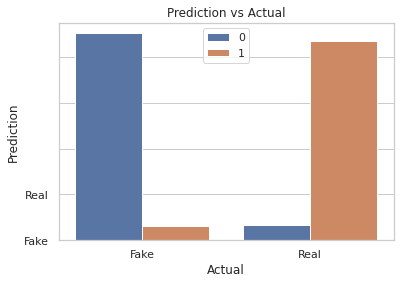
Word cloud for real News :

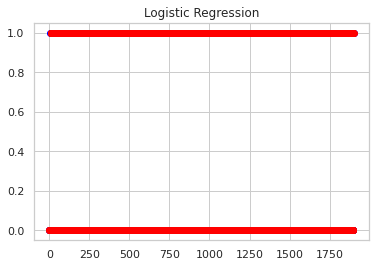


Word cloud for real News :



Results:





Table

Description automatically generated

