Arabic Article Classification

A machine learning report

Amr Mahmoud 19200010

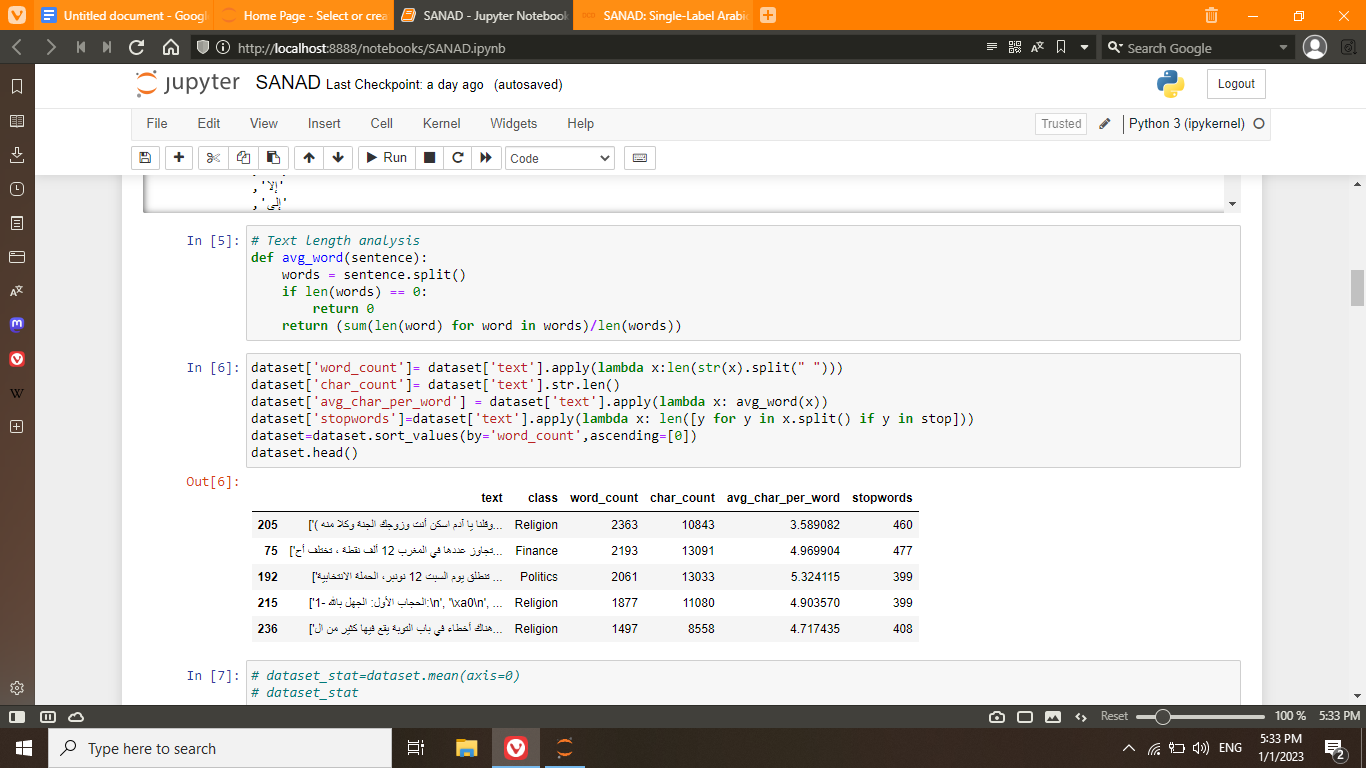
Muhammad Helmy 1610114

In this project, we implement arabic text classification.

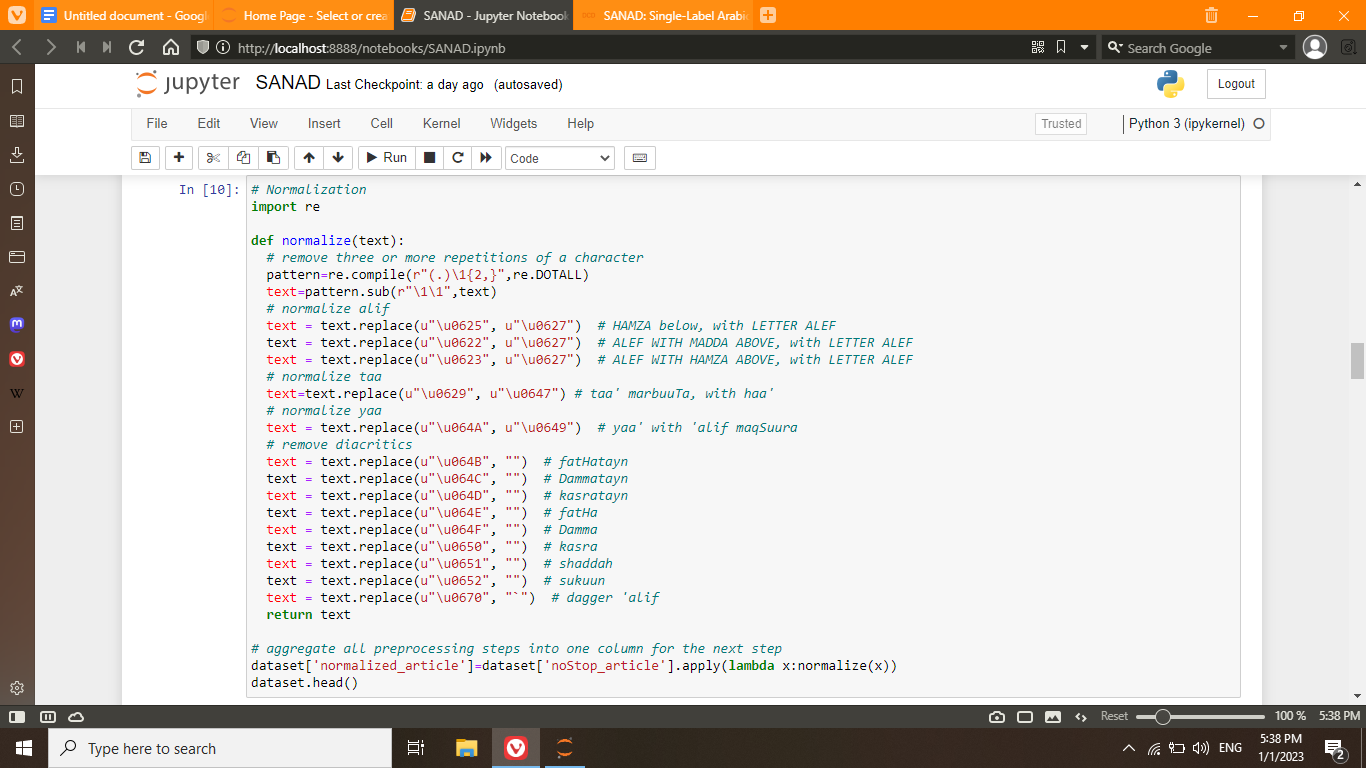
We trained the model using the following dataset: <https://data.mendeley.com/datasets/57zpx667y9/2>

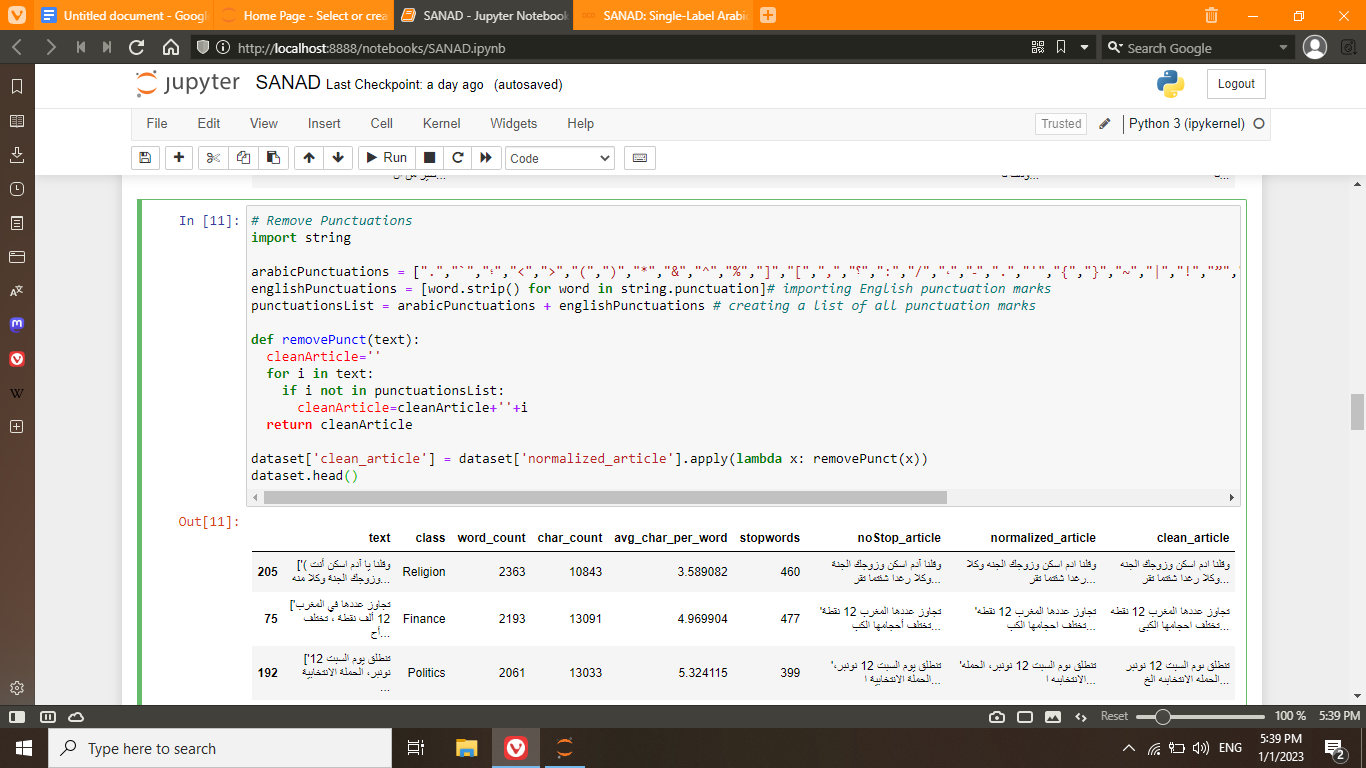
The dataset consists of articles from several news sources, separated into folders based on their topic. We only used data from one of the newspapers for convenience and due the large amount of data it was unnecessary to use the rest.

We then performed exploratory data analysis on the data to explore information such as average article word count, character count, average number of characters per word and the number of stop words.

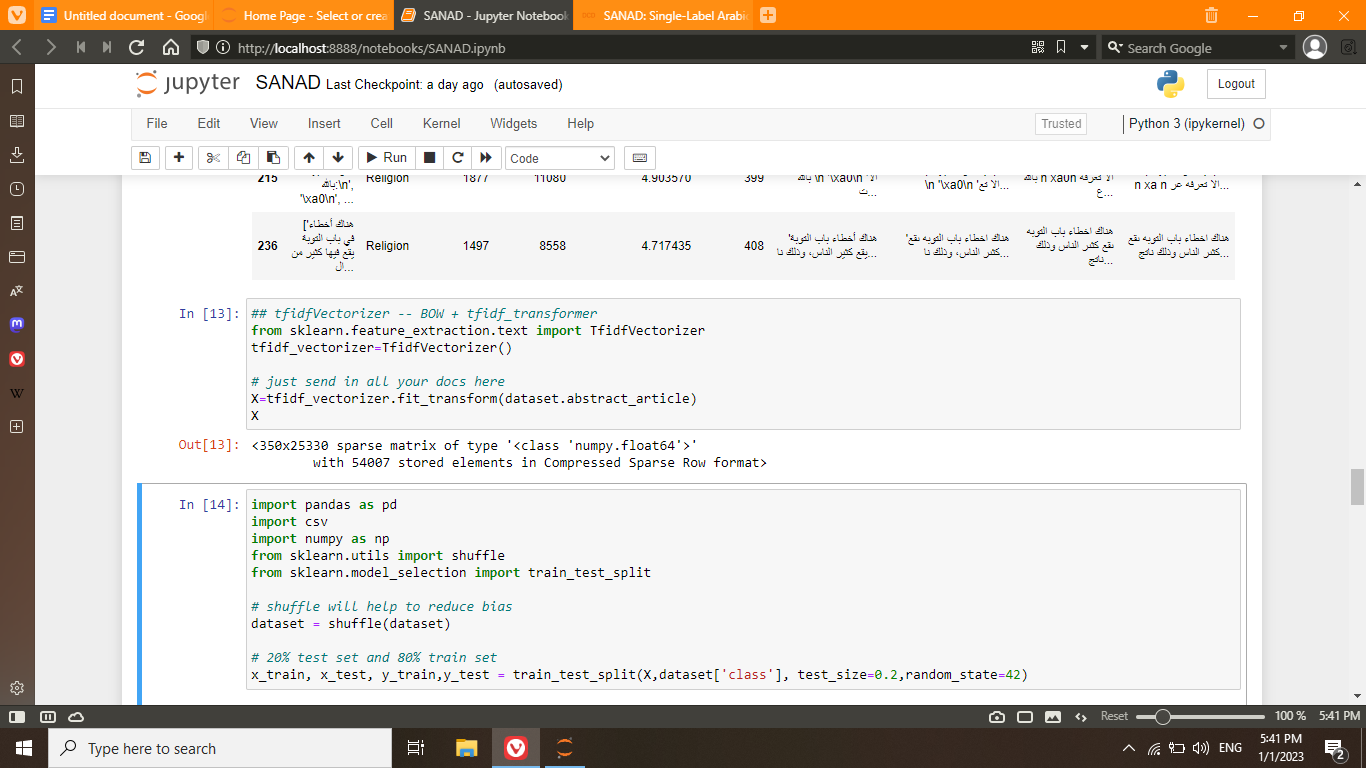


Then in the text preprocessing step we removed all stopwords and any other unnecessary characters such as spaces, numbers, punctuations or any characters that were repeated in excess within a single word.

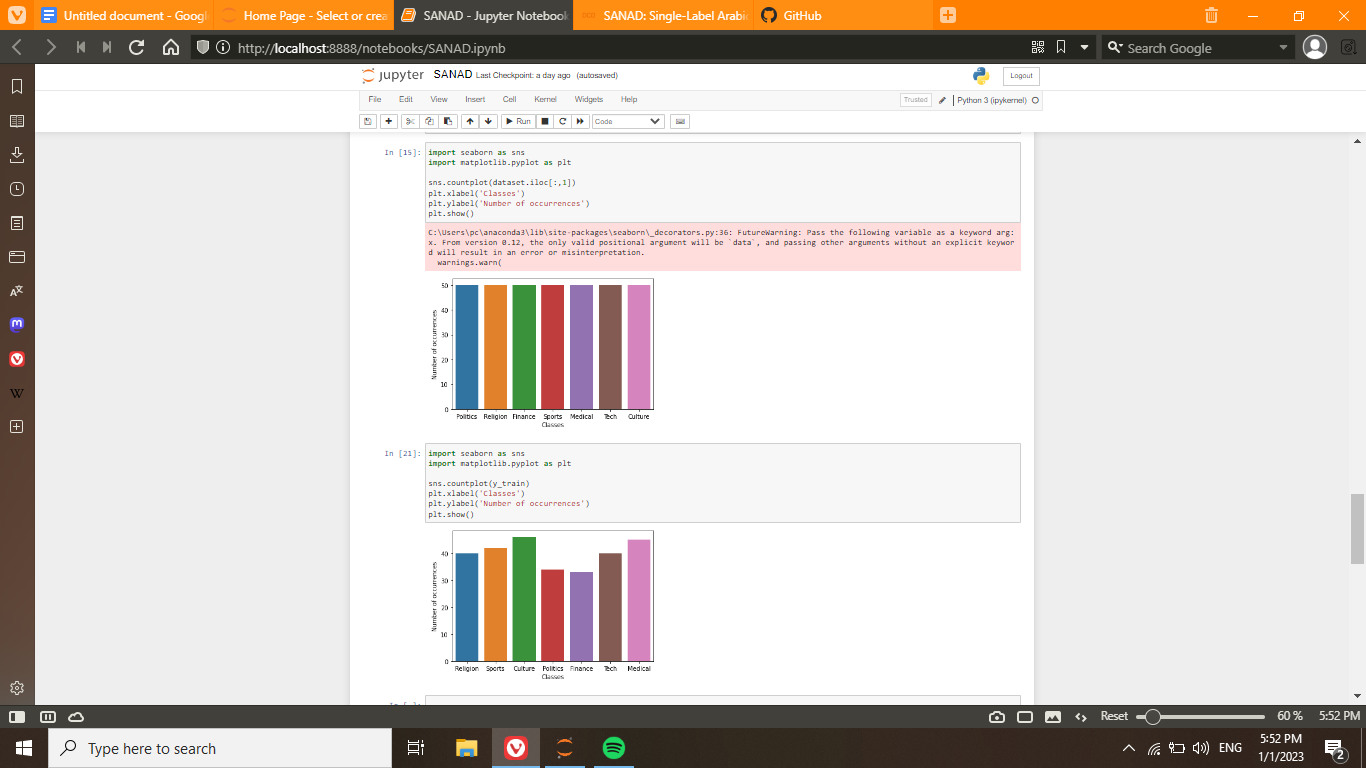




The text data was then passed through a TFIDF Vectorizer, and that became our X. Then the data was split into training and testing data.



We then plot the number of occurrences of every class in the training set to make sure the distribution matched the original distribution and there was no bias.



We then used a random forest classifier with the data. The accuracy was very poor (15%) and after anaysing the data, the accuracy for the training set was very high so the conclusion was that this low accuracy was due to overfitting.

In an attempt to reduce this overfitting, we used a cross validation grid search to be able to find the best parameters for the Random Forest Classifier. The resulting parameters led to similar results as the initial. It is also worth noting that we had to reduce the data size as performing cross validation on 14000 entries took a very long time.

We experimented with several other models. One which yielded great success was logistic regression. It gave an accuracy of over 90%.