NEWS GROUP

CLASSIFICATION

NLP

احمد محمد عبدالمجيد محمد	2021170045 CS
عمرو علي سيد بغدادي	2021170376 CS
حسن محمد حسن عبد العظيم	2021170165 CS
زياد عبدالحميد ابو الخير	2021170212 CS
باسيلي اشرف باسيلي	2021170114 CS
حسين اسامه عبدالعظيم	2021170169 CS

Under supervision of

Dr. Sally Saad

DEALING WITH DATA steps:

- After downloading dataset and store it on the hard drive
- We had to open the folder and iterate on the classes folder and for each folder iterate on the files and SAVE the text file in a NEWS list.
- And save the class folder name in another list as it's our target column CATEGORY.

PREPROCESSING steps:

For each text file:

- Removing stop words
- Lower case folding
- Tokenize the text.
- Lemmatize each token.
- Using Regular Expression removes any non-alpha numeric word or white space.

Feature Extraction steps:

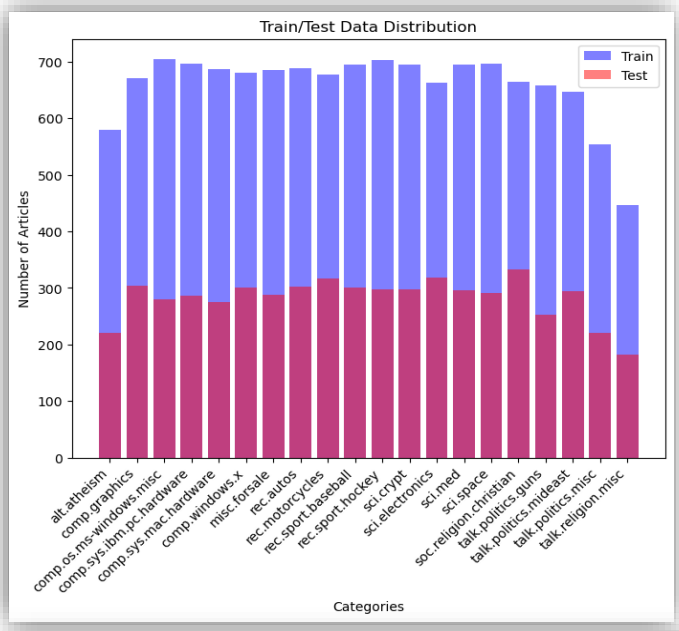
- Using TF-IDF vectorizer we fit the News column which has all the text files for all classes.

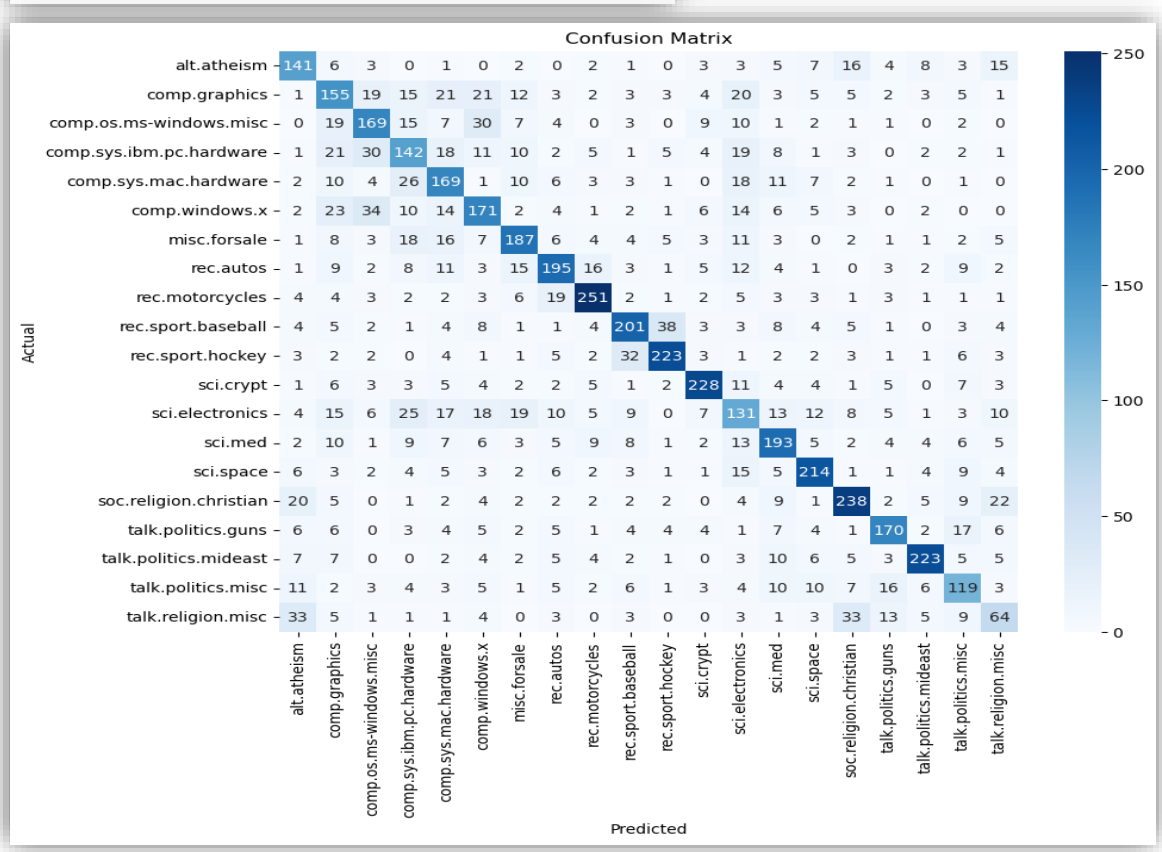
MODELS results:

After splitting data into 0.7 train / 0.3 test we trained these models on the train data and use test data to make predictions

Decision tree:

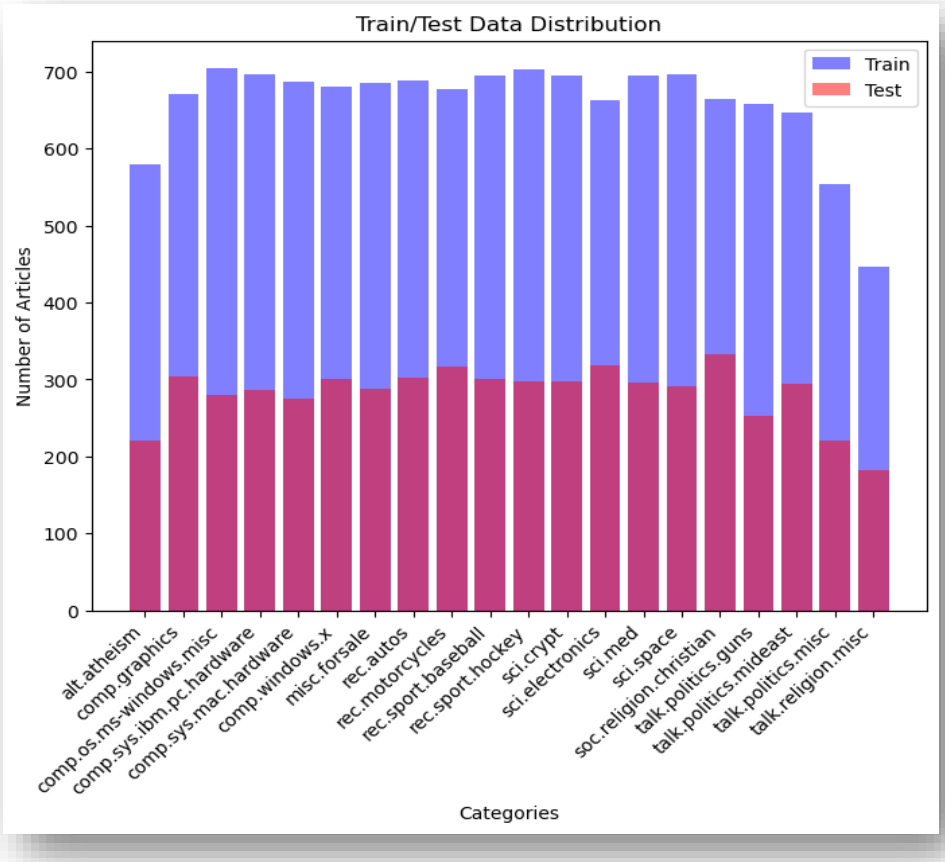
```
Decision tree model Accuracy :  0.6344485749690211
```

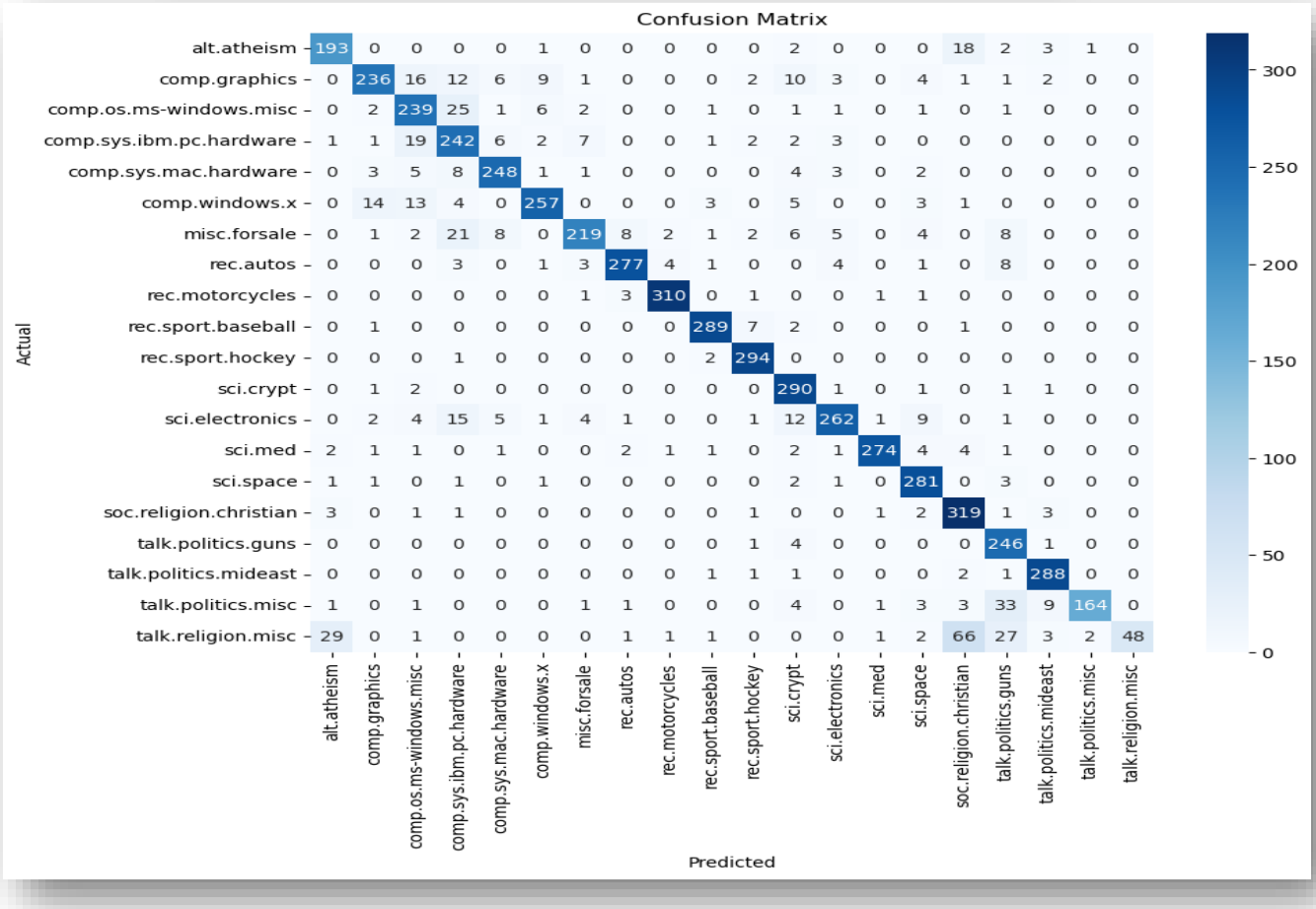




Naive Bayes:

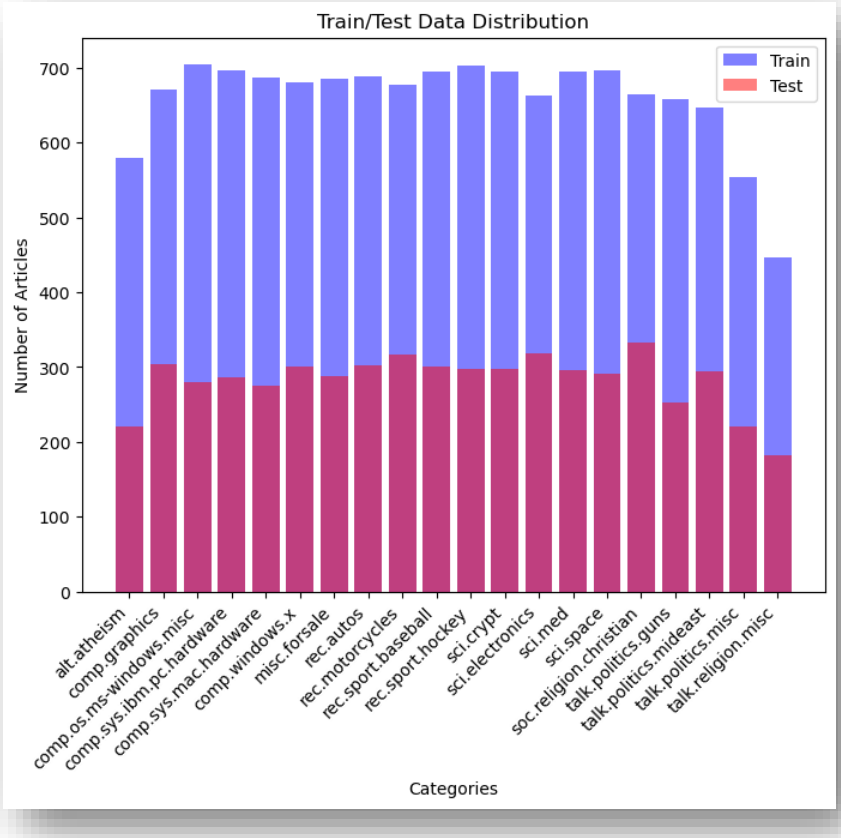
Naive Bayes model Accuracy : 0.8808638697114534

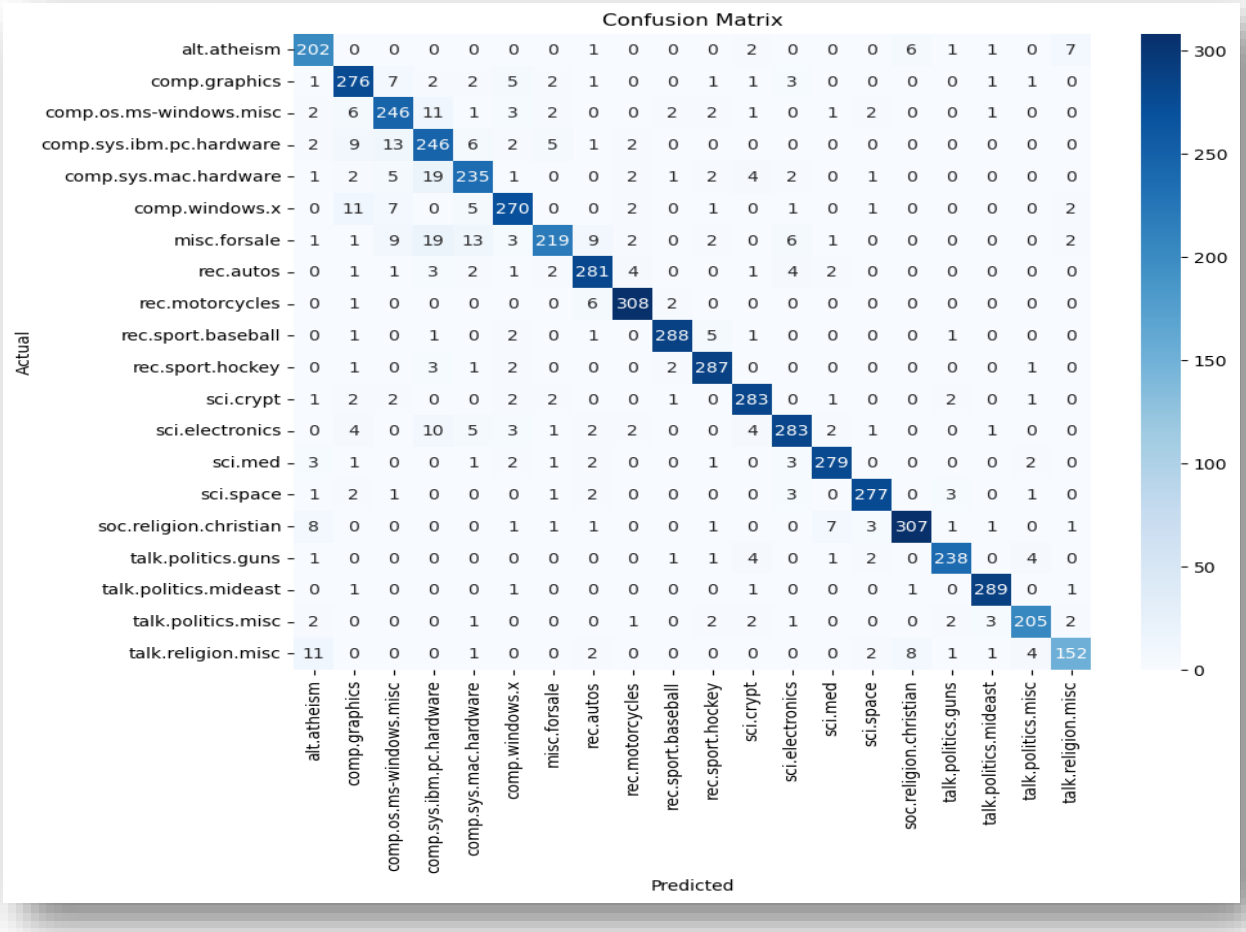




KNN:

KNN model Accuracy : 0.9153832536732165





Now with Our Hero **SVM**

```
SVM model Accuracy :  0.9920339883165162
```