Multi Classification Report

Data Cleaning:

1- Removing Duplicates: There were about 4696 duplicates! So I cleaned them not to make redundancy or overfitting
2- Cleaning Text by removing:

- Stop Words
- Punctuations
- Digits
- Special Chars

Classifier:

I used different classifiers and compared between them including:

1- Naive Bayes classifier for multinomial models
2- Linear SVM
3- Logistic Regression
4- Random Forest Classifier

Linear SVM was the best to perform with higher metrics so I used it for classification

Imbalance Problem:

- First: After removing Duplicates It was partially solved
- Second: I used sample weights to take smaller classes into account during modeling

Improving Model:

- First: More data required! Because we lost more than 50% of the data after removing duplicates
- Second: We can use ensemble models to improve performance and use Grid search for hyperparameter tuning
- Third: We can use oversampling to solve imbalance problem

Model Evaluation:

We have a multiclassification problem with imbalance so the accuracy, or precision/recall will not provide the complete picture of the performance of our classifier and accuracy can be very misleading because it does not take class imbalance into account So I recommend using **Cohen's kappa score.** It is a very good measure that can handle very well both multi-class and imbalanced class problems.

Limitations:

- Model is not high efficient about 82% cohen_kappa score and 88% accuracy
- This can be overcome by increasing data or oversampling or using ensemble methods