Twitter sentiment analysis

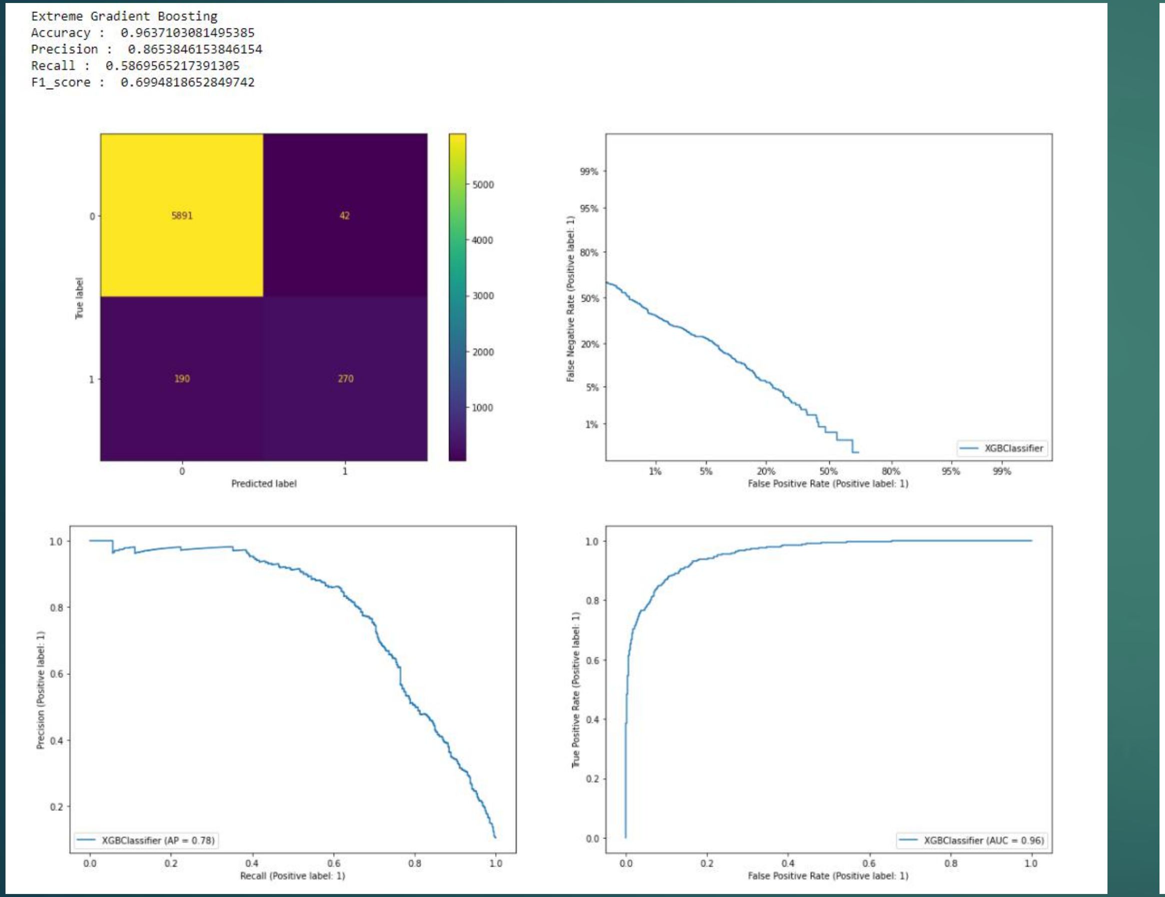
SOUNAK MANDAL

This project attempts at classifying tweets into racist and non-racist categories. The model figures out the connotations behind the word and how likely it is to be linked to racism. The words are encoded by different word embedding techniques. The statistical or frequency based embedding techniques include count vectorization and TF-IDF. The neural word embedding is using word2vec model and is simply averaged to obtain encoding for each tweet.

The models used were standard logistic regression, support vector machines, random forest and xgboost. With all models neural embedding worked best while statistical embedding performed poorly. Finally the xgboost model was fine tuned.

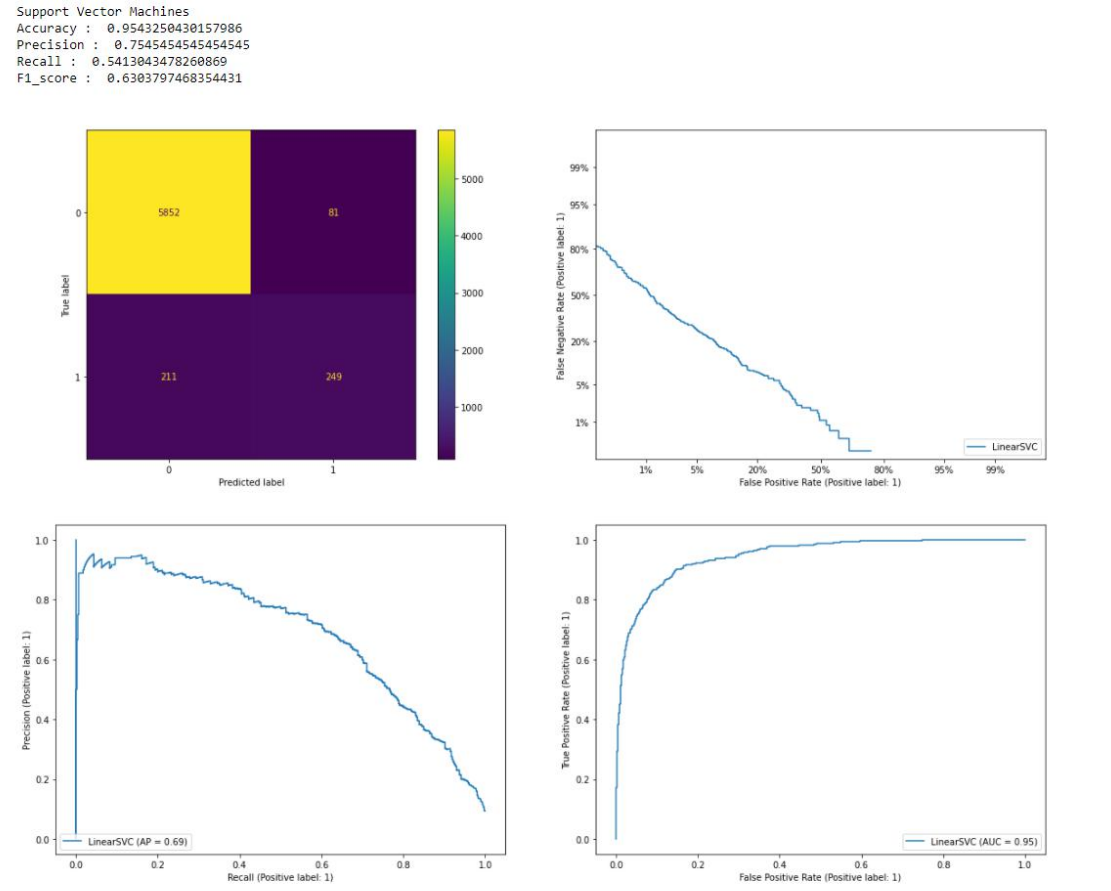
GitHub link : ML-Projects/twitter sentiment analysis at main · SounakMandal/ML-Projects (github.com)

Some of the plots available in the notebook. For each embedding, model combination confusion matrix, det curve, precision recall curve and roc curve were plotted.



```
Extreme Gradient Boosting
Accuracy :  0.9637103081495385
Precision :  0.8653846153846154
Recall :  0.5869565217391305
F1_score :  0.6994818652849742
```

5891
42
190
270
0
1
True label
Predicted label
5000
4000
3000
2000
1000
99%
95%
80%
50%
20%
5%
1%
False Negative Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
XGBClassifier
1.0
0.8
0.6
0.4
0.2
0.0
Precision (Positive label: 1)
Recall (Positive label: 1)
XGBClassifier (AP = 0.78)
True Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.96)



```
Support Vector Machines
Accuracy :  0.9543250430157986
Precision :  0.7545454545454545
Recall :  0.5413043478260869
F1_score :  0.6303797468354431
```

5852
81
211
249
0
1
True label
Predicted label
5000
4000
3000
2000
1000
99%
95%
80%
50%
20%
5%
1%
False Negative Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
LinearSVC
1.0
0.8
0.6
0.4
0.2
0.0
Precision (Positive label: 1)
Recall (Positive label: 1)
LinearSVC (AP = 0.69)
True Positive Rate (Positive label: 1)
LinearSVC (AUC = 0.95)

Thank You