Untitled

November 23, 2022

0.1 Analysis and Classification of Haters Speech and Offensive Language Twitter Dataset

Name

Institution

Professor

Date

0.1.1 Background

Since the dawn of civilization, our technology has come a long way. Now anyone with a smartphone device and an internet connection can connect and communicate freely. This is very positive but having many people accessing the social world leads to mixed opinions and words thrown at each other. It is easier to type something cruel than to say it to someone's face. This leads to people commenting or talking negatively with cruel comments, harsh words, racist comments, insulting expressions, and many more. This is called hate speech. The barrier that can divide people from expressing hatred through comments and texts can be hate speech or offensive language detectors and ban them. This is why we need to come up with hate speech detectors that can easily erase the hate speeches and make the social world a little bit cleaner.

0.1.2 Introduction

Hate speech detection is an important tool for combating hate speech, particularly on social media. Several methods have been developed for the task, including a recent proliferation of machine-learning and deep-learning model-based approaches. Because hate speech data sets are rarely clean, they must be pre-processed before classification algorithms can detect hate speech. Different machine learning models have different strengths, making some better than others at certain tasks like detecting hate speech. Some models are more accurate than others in terms of efficiency. It is critical to employ various models and compare their performance in order to identify the best one for hate speech detection. Pre-training methods have grown in popularity in recent years, and it is important to test whether they work well with hate speech detection algorithms. It is also important to see how hate speech detection models can be used to address domain changes.

0.1.3 Methodology

Dataset Description: The 'Hate speech and offensive language dataset' is used for this project and over6000 rows from the dataset are used to train where 80% rows are labeled as training dataset

and 20% rows are labeled as testing dataset. Each row in the dataset is labeled as hate speech and offensive language. #### Technolgies Used: Following libraries were be used for this project

Pandas - for reading the CSV file

Numpy- for arrays

Matplotlib - for graph plotting

Sklearn - for implementing machine learning models

Seaborn - for making statistical graphics

NLTK - for text data preprocessing

0.1.4 Project Overview

Our project is a combination of several subtasks and processes. Starting with loading the dataset into the pandas dataframe, we first oversampled the dataset to combat discrimination among the different classes. Then we applied text preprocessing to get rid of unhelpful and noisy data. Moving on, we applied feature engineering to vectorize the raw text data. Following this, we split the dataset into train and test sets and fed the train data into our models. Finally, we ran accuracy metrics to measure and compare the performance of all three algorithms.

0.1.5 Models used:

After the completion of count vectorization using bag-of-words, our dataset is finally ready to train. In this project, we have used three classification algorithms to train over our dataset. The algorithms are as follows. 1. Logistic Regression 2. Decision Tree 3. Ensemble Model - Algorithms used are: Xgboost Model, AdaBoost and Gradient Boosting Classifier Mode 4. SVM

Results

0.1.6 Accuracy Matrix

This section elaborates on the performance of the severals models we've used for text classification. From first table it is evident that SVM has the best performance among the other models. The accuracy metrics we've taken are precision score, recall score, and f1 score. The precision score measures how accurately the model predicts the true positives. The recall score measures the model's ability to find the true positives among all the positive correct outputs. F1 score is a harmonic mean of the precision score and recall score. The results for the three algorithms are stored in Tables 1, 2, and 3. From Tables 1, 2, and 3. It is clearly seen that Logistic Regression and Decision Tree performed the worst among all the other three algorithms while SVM performed the best. For example, the precision scores for correctly predicting the true positives for "Hate Speech" for Logistic Regression, SVM, and Decision Tree are respectively 0.81, 0.91, and 0.81. The relatively bad performance of Decison Tree was due to the algorithm's conditional independence of its features. The algorithm starts to show weakness as the number of features increases.

Logistic Regression

	precision	recall	f1-score	support

```
           0       0.61      0.86      0.71      3424
           1       0.86      0.70      0.78      5896
           2       0.95      0.90      0.93      5074

    accuracy                           0.81     14394
    macro avg       0.81      0.82      0.80     14394
    weighted avg       0.84      0.81      0.81     14394
```

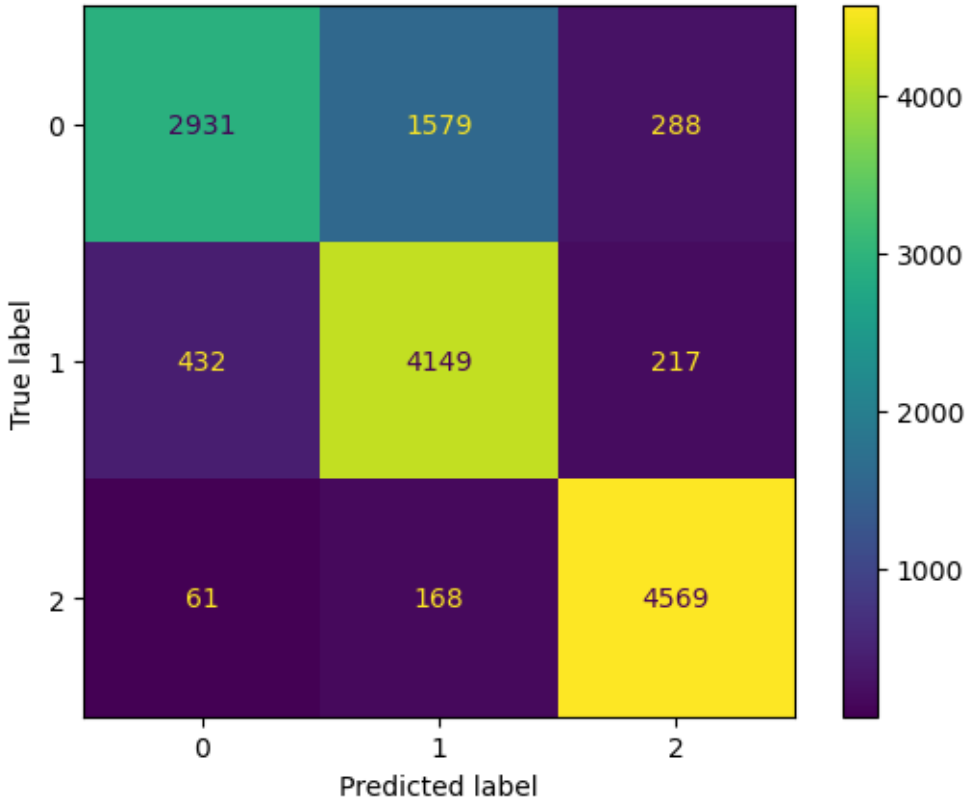
SVM

```
              precision    recall  f1-score   support

           0       0.75      0.98      0.85      3663
           1       0.97      0.80      0.88      5784
           2       0.94      0.91      0.92      4947

    accuracy                           0.88     14394
    macro avg       0.88      0.90      0.88     14394
    weighted avg       0.90      0.88      0.89     14394
```

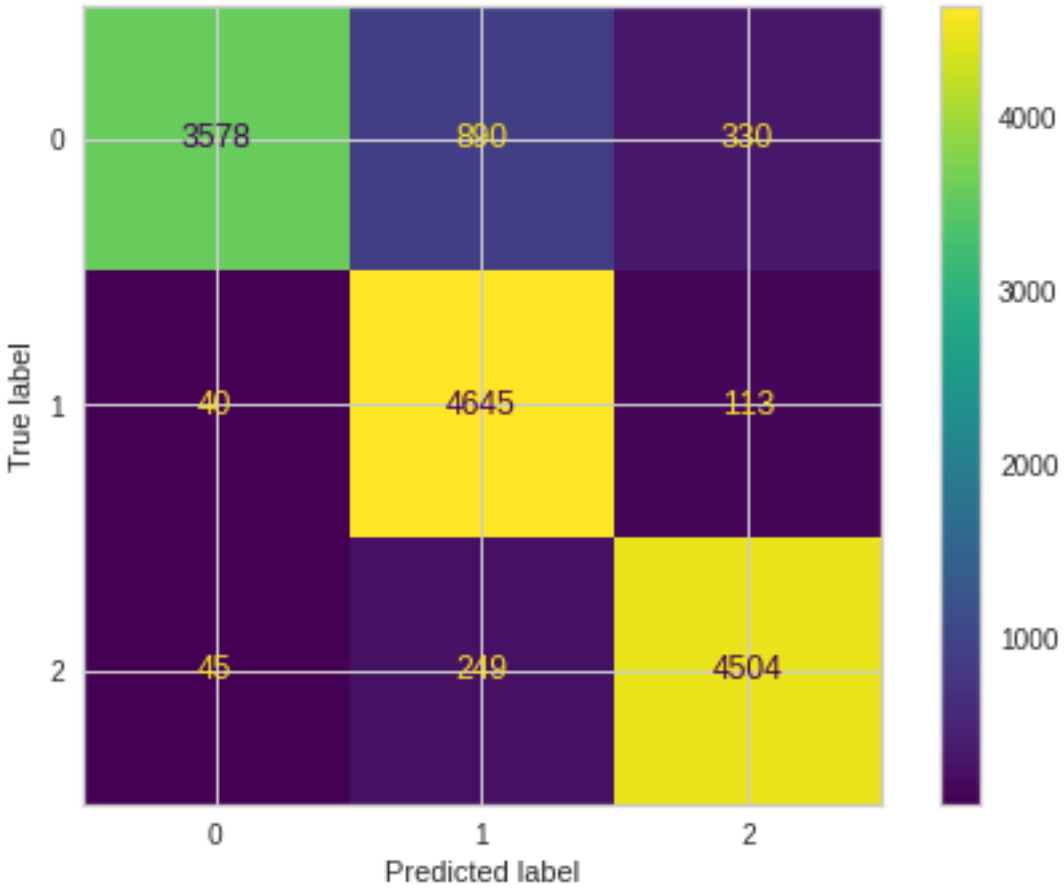
Decison Tree

```
              precision    recall  f1-score   support

           0       0.63      0.87      0.73      3498
           1       0.91      0.69      0.79      6299
           2       0.88      0.92      0.90      4597

    accuracy                           0.81     14394
    macro avg       0.81      0.83      0.81     14394
    weighted avg       0.83      0.81      0.81     14394
```

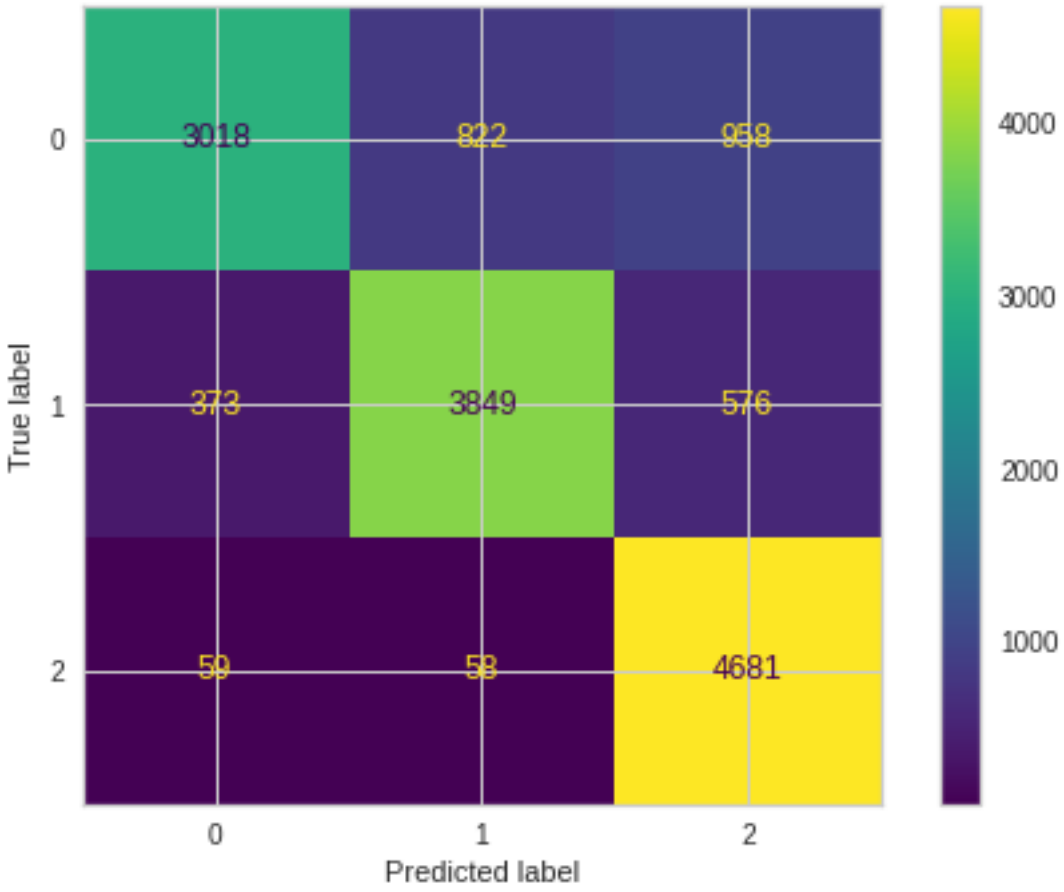
0.1.7 Confusion Matrix



Logistic Regression


3578
890
330
40
4645
113
45
249
4504
0
1
2
True label
0
1
2
Predicted label
4000
3000
2000
1000

SVM



Decision Tree

0.1.8 Conclusion

Throughout this project we automated text classification techniques to classify hate speech and offensive tweets . We have used several classification algorithms to train over our dataset - Logistic Regression, Decision Tree, SVM, and Ensemble model. Among them SVM has the best accuracy while Decision Tree and Logistic Regression has the worst performance results.

[ ]: