Data Glacier

Your Deep Learning Partner

Final Presentation

Advance NLP : Hate Speech detection using Transformers

Hiten Chadha
Team Name- Arjohi
31.08.2022

Agenda

Problem Description

EDA Presentation

Proposed Modeling Technique

Tested Models

Chosen Model and Final Recommendatinon

Data Glacier

Your Deep Learning Partner

Please Note

- For week 13, I lost communication with my group without any explanation and also was quite a bit occupied myself with work at my home university with my thesis finalization. So the work done for the last week was completed by just me and also I could not devote as much time as I wanted to for applying further advanced Deep learning models like BERT without my team. I have tried classifiers that are known well to me and worked with good accuracy for this dataset. Hope it will be graded accordingly

Checking the Shape of Training and Test Data

```
print("Training Set:"% training_data.columns, training_data.shape)
print("Test Set:"% testing_data.columns, testing_data.shape)
```

```
Training Set: (31962, 3)
Test Set: (17197, 2)
```

We have 31962 and 17197 tweets in the training and test data set respectively.

Null Data

```
print('Train_Set -----')
print(training_data.isnull().sum())
print('Test_set -----')
print(testing_data.isnull().sum())
training_data.head()
```

```
Train_Set -----
id       0
label    0
tweet    0
dtype: int64
Test_set -----
id       0
tweet    0
dtype: int64
```

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

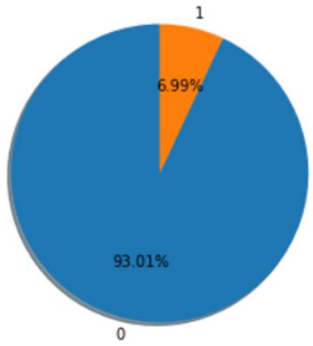
There are no null data in the datasets.

Positive and Negative Tweets

```
training_data['label'].value_counts() #counting no of positives and negatives
```

```
0    29720
1     2242
Name: label, dtype: int64
```

There are 2242 hate speech tweets (represented in yellow color in the given pie chart) in the training data and the rest contains no hate speech.

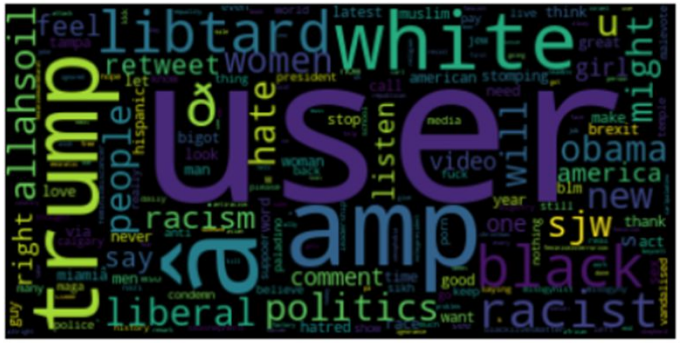



```
0    23783
1     1786
Name: label, dtype: int64
```

Most Frequent Hate Words

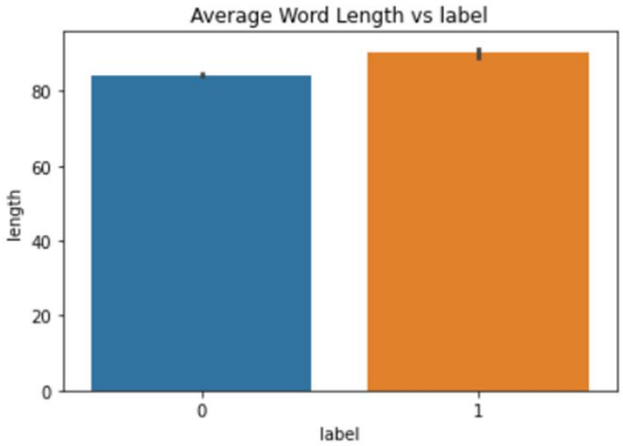
Non-Hate Comments

Hate Comments



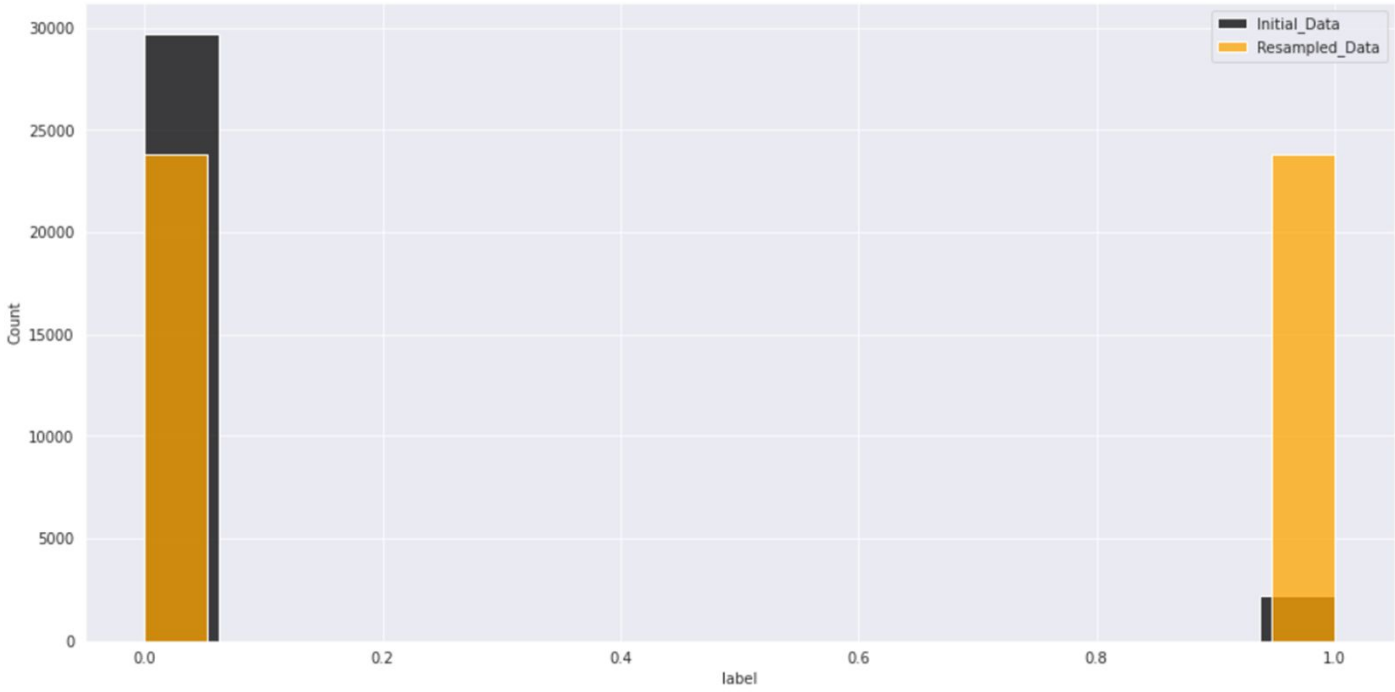
Visual representation of most frequent hate words

Average Word Lengths



Average word lengths for hate speech (orange) and non hate speech (blue) tweets.

Undersampling and Overssampling results



Initial data(black) and after sampling data(orange) for hate and non hate words

Recommended Models

- XGBClassifier
- LogisticRegression
- MultinomialNB
- SGDClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- KNeighborsClassifier
- LinearSVC
- SVC
- BERT
- RoBERTa

Tested Models

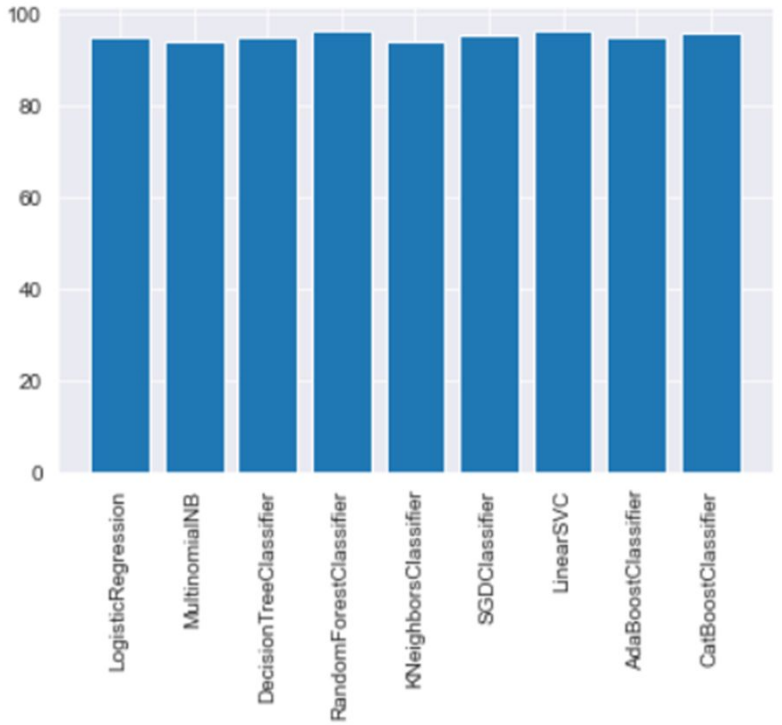
- CatBoost Classifier
- LogisticRegression
- MultinomialNB
- SGDClassifier
- KNeighborsClassifier
- LinearSVC
- DecisionTree Classifier
- RandomForest Classifier
- Adaboost Classifier
- BERT(Failed to execute completely)

Tested Model Accuracies

RandomForestClassifier Accuracy Score : 96.03%
AdaBoostClassifier Accuracy Score : 94.73%
KNeighborsClassifier Accuracy Score : 93.87%
LogisticRegression Accuracy Score : 94.82%

CatBoostClassifier Accuracy Score : 95.54%
DecisionTreeClassifier Accuracy Score : 94.78%
MultinomialNB Accuracy Score : 94.99%

SGDClassifier Accuracy Score : 95.28%
LinearSVC Accuracy Score : 96.39%



Final chosen model

- Linear SVC

Chosen Model metrics

LinearSVC Accuracy Score : 96.39%

	precision	recall	f1-score	support
0	0.99	0.97	0.98	6080
1	0.59	0.86	0.70	313
accuracy			0.96	6393
macro avg	0.79	0.91	0.84	6393
weighted avg	0.97	0.96	0.97	6393

Final Recommendation

- For the given problem, the recommended model used should be Linear SVC. However, Randomforest can be a good choice as well and might work better with other test data. BERT model might work even better but unfortunately I was unable to implement it in full, so the future work might be to test the data with the Huggingface library.

Thank You