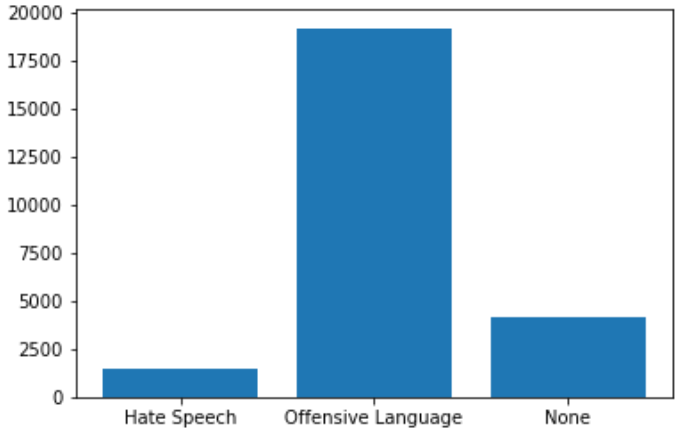
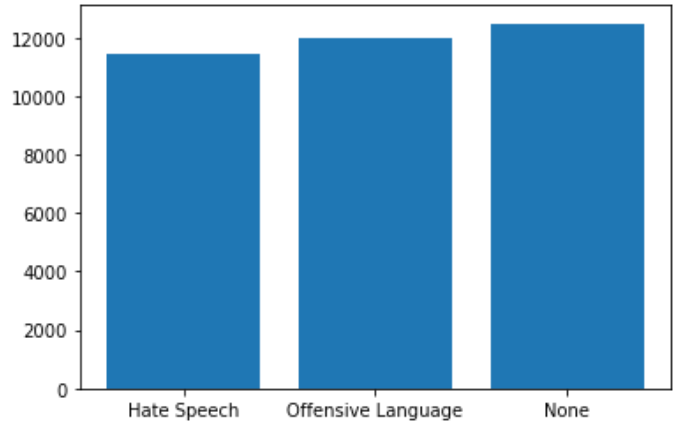
TWITTER HATE SPEECH CLASSIFICATION

- T DAVIDSON DATASET



ORIGINAL DATASET



BALANCED DATASET

The dataset was balanced by repeating the Hate Text and None Text by appropriate factor.
Hate Speech: 11440
Offensive Language: 12000
None: 12489

PREPROSSING STEPS TAKEN IN TWEETS

1. Removed the @Usernames
2. Removed the URLS in the tweets
3. Removed all numbers and Special Characters
4. Removed the stop words
5. Replaced some Slangs in tweets
6. Lemmatization of text

FEATURES OF TEXT DATA

1. Total unique words in the corpus: 14,146 (50% of them had frequency >=3)
2. Average Length of tweets: 7
3. Max length: 28

TF-IDF Vectors

1. They were formed on the top 8000 occurring words
2. Shape of TF-IDF Training Vector: (35929,8000)

Classes = ['Hate Speech', 'Offensive Language', 'None']

ML MODELS TRAINED ON TF-IDF

1. **SVM**

   model = linearSVC(class_weight='balanced',multi_class='crammer_singer', max_iter = -1)

   Evaluated Metrics:

	precision	recall	f1-score	support
0	0.88	0.96	0.92	562
1	0.95	0.85	0.89	604
2	0.96	0.98	0.97	631
accuracy			0.93	1797
macro avg	0.93	0.93	0.93	1797
weighted avg	0.93	0.93	0.93	1797

2. **SVM – 2**

   model_2 = LinearSVC(class_weight='balanced',C=1, penalty='l2', max_iter = 1500, loss='squared_hinge',multi_class='ovr')

   Evaluated Metrics:

	precision	recall	f1-score	support
0	0.89	0.96	0.92	562
1	0.95	0.85	0.90	604
2	0.95	0.98	0.97	631
accuracy			0.93	1797
macro avg	0.93	0.93	0.93	1797
weighted avg	0.93	0.93	0.93	1797

Result was same as before

3. **SGD Classifier**

   model_3 = SGDClassifier(n_jobs=-1, class_weight =' balanced',penalty =' l2')

   Evaluated Metrics:

	precision	recall	f1-score	support
0	0.81	0.81	0.81	562
1	0.87	0.81	0.84	604
2	0.90	0.95	0.92	631
accuracy			0.86	1797
macro avg	0.86	0.86	0.86	1797
weighted avg	0.86	0.86	0.86	1797

4. **Logistic Regression**

   model = LogisticRegression(n_jobs = -1, penalty='l2', multi_class='multinomial', class_weight = 'balanced',verbose=1)

   Evaluated Metrics:

	precision	recall	f1-score	support
0	0.87	0.93	0.90	562
1	0.92	0.84	0.88	604
2	0.95	0.97	0.96	631
accuracy			0.91	1797
macro avg	0.91	0.91	0.91	1797
weighted avg	0.92	0.91	0.91	1797

DEEP LEARNING MODEL TRAINED ON EMBEDDINGS

Number of Words in Vocabulary: 8000
Embedding Dimension: 32
Padded Length of each Tweet: 24

```
Model: "Twitter Hate Text Classification"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 24, 32)            256000
_________________________________________________________________
simple_rnn_1 (SimpleRNN)     (None, 24, 8)             328
_________________________________________________________________
global_max_pooling1d_1 (Glob (None, 8)                 0
_________________________________________________________________
dense_1 (Dense)              (None, 20)                180
_________________________________________________________________
dropout_1 (Dropout)          (None, 20)                0
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 63
=================================================================
Total params: 256,571
Trainable params: 256,571
Non-trainable params: 0
```

EVALUATED METRICS:

Training Accuracy: 98.39 %
Validation Accuracy: 96.25%
Test Accuracy: 95.99%

	precision	recall	f1-score	support
0	0.94	0.99	0.96	572
1	0.97	0.90	0.94	586
2	0.96	0.99	0.98	639
accuracy			0.96	1797
macro avg	0.96	0.96	0.96	1797
weighted avg	0.96	0.96	0.96	1797