TASK: HTSPC- Hate Speech Classification

Ananlysis:

- “Hate speech” labels count: 2031 v/s “Not Hate speech” labels count: 3235
- Since classes are not balanced, F1 score would be better metric to evaluate our model.

Preprocessing:

- Convert to lowercase
- Remove URL
- Remove digits
- Remove mentions
- Remove stop words
- Apply `WordNet Lemmatizer`
- Apply Snowball Stemmer
- Remove punctuation

Vectorizer:

- Used TFIDF vectorizer to tokenize data

Model used:

Tested these models using k-fold cross validation:

- Logistic regression
- Naive Baye’s
- Random forest
- Support Vector classifier.

Best F1 score was obtained on Support Vector classifier so used it as final model

Key observations:

- Since labels are not balanced, it maybe a bit misleading if we just look at accuracy for evaluating our model. For e.g: 61.4% samples in training set are lablled as 1(not hate speech). So if we build a dummy classifier which just returns label 1 for every sample, we would still get 61.4% accuracy. So better approach would be to use F1 score and confusion matrix to evaluate our model.

- The model can’t clearly differntiate between the context in which an abuse or bad word is used. In some cases, it maybe hate speech, while in other cases it might just be valid harsh criticism. For e.g: Below tweet was labelled as hate speech although it’s just valid criticism.

```
This truly is a crisis and, worst part is, his delusional admirers will probably never know. They'll just blame ev
eryone else for his stupidity, fear mongering, hate, greed and lies. #FuckTrump

Prediction: hate
Actual: not hate
```