CodeFest: Linguipedia

Objective : **to detect hate speech in tweets (racist or sexist sentiment).**

Data : labelled dataset of 31,962 tweets (training data)

Evaluation Metrics : F1 score

PREPROCESSING :

- Tokenization(break into words and tag Part of Speech)
- Conversion of whole text to lower case
- Elimination of extra white space
- Removal of common English stopwords
- Removal of punctuation Marks
- Lemmatisation (Grouping together inflected form of words so that they are all treated as single word)

CONVERSION TO TF-IDF VECTOR

OVERSAMPLING:

The data is highly imbalanced which means ratio of non-hate spech to hate speech is very high. In such cases model becomes biased i.e, model tends to predict majority class in most cases. To avoid this training data has benn oversampled. Oversampling has been done simply appending the data of minority class. This leads to relatively less imbalanced training data.

MODEL:

Linear SVM is fitted on the given training dataset. Optimum value of parameter C is found through grid search and the value has C=0.7 which was found using grid search is directly used in the given code. Many researchs show that most text categorizations are linearly seperable and also the problem has high dimensional input space . Thats why linear SVM is preferred in text classification.

Programminng Language Used : Python

Toolkit used : NLTK , Scikit