Natural Language Processing of Hate Speech

Jojeena Jogy Kolath
MSc in Big Data

1. Problem and Aim

Recently, social media has served as a major platform for the extensive propagation of online hate speech, and has proved to further propagandize real-life hate crimes.

Aim – To develop natural language processing algorithms to identify, classify text relating to online hate speech.

2. Approach

Dataset – Dataset was built primarily using:

- Twitter querying
- Datasets from related papers
- Data extracted from manifestos.

Data was split into train data (70%) and test data (30%).

Methods – Classifier models using Python sklearn pipeline architecture were built with steps:

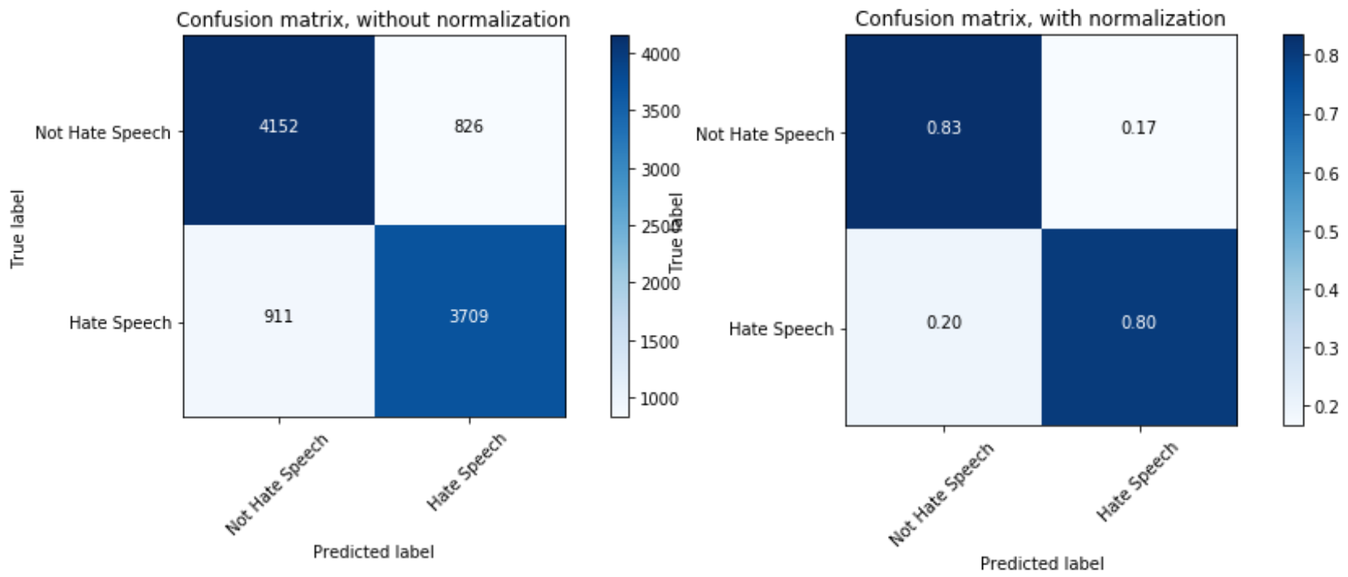
- Vectorizer – Tfidf, Count and custom built keyword vectorizers with stopword exclusion
- Classifier algorithms – RandomForest, SVC, NaiveBayes and Logistic Regression – wrapped in a OneVsRest classifier
- Feature Selection using SelectKBest or n-grams
- Neural networks models using word embeddings - GloVe, ElMo.

Amazon Web Services was used to store datasets (using S3) and to leverage cluster computing (using Elastic Container Service).

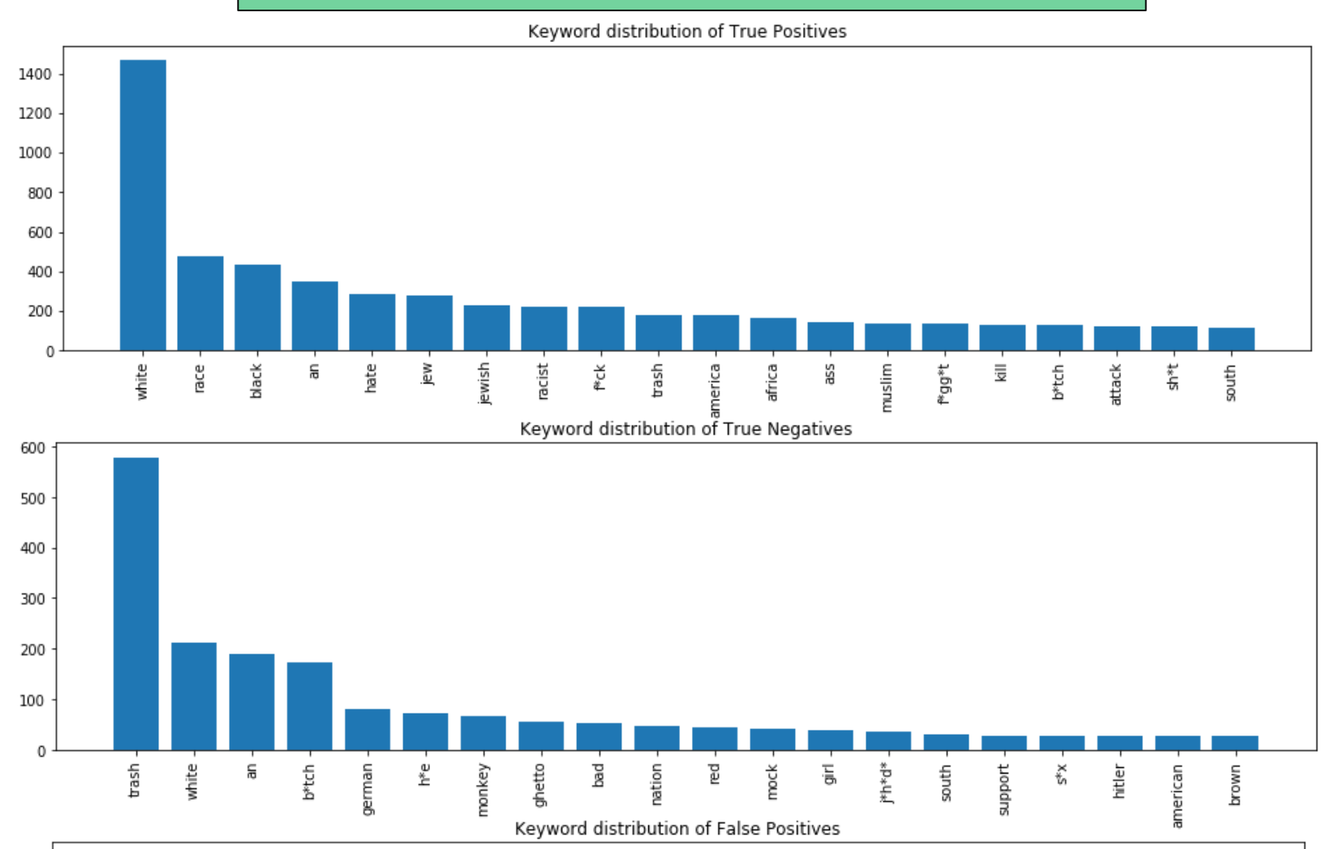
3. Results

5-fold cross-validation was applied and model efficiency was evaluated using:

- Classification report and confusion matrices
- Plots of True Positives, True Negatives, False Positives and False Negatives



Validation Accuracy = 0.82



4. Next Steps

- Further experimenting with CNNs,
- RNN models with Long Short Term Memory
- Sentiment analysis

For further information contact: *Jojeena Jogy Kolath*

Joa00073@students.stir.ac.uk