Project 12 Analysis of Hate Speech 1

This project aims to put forward an efficient approach for identifying hate speech in online forum and free documents.

To start you can study the following tutorial on text classification with various features and classifiers. https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/

First, we consider the Wikipedia Comment Corpus available at https://meta.wikimedia.org/wiki/Research:Detox/Data_Release . Download the datasets corresponding to Personal Attack, Aggression and that of Toxicity.

2. Construct a database of non-aggressive, non-personal and non-toxicity dataset. For this purpose, build a program that generates such dataset using two distinct approaches. The first one creates a negation of the attack / aggression / toxicity sentence. The second one picks up a sentence at random from Wikipedia on a completely unrelated topic. This creates a database of at least twice the size of the downloaded dataset.

3. Construct a classifier that classifies a post into either a personal attack or non-attack (use only attack and neutral dataset). Discuss the various combinations of features and classifiers and report the accuracy rate and F1 score.

4. Repeat question 3) for aggression and for toxicity cases.

5. Now we would like a classifier to learn the three types of hate speech simultaneously. Use the structure of the dataset and design an ensemble of classifiers for classifying a post into neutral, attack, aggression and toxicity with various combination of feature sets in order to test the performance of the various classifiers. Report a comparative analysis with metric F1 score and accuracy of classification.

6. Use the LIWC features alone or combined with other features investigated in 3-4) in order to test the performances of the suggested classifiers. Use the majority voting rule to test the combination of the various classifiers.

7. Design a GUI interface where the user can input the URL of the online forum and output the various categories of the hate speech.