Minimum Viable Product Proposal - Hate Speech Classifier
Leon Zhou

Domain and Summary

Hate speech and offensive language, while both unsavory, are distinct in legality and content. The intent of this product is to devise a classifier that may accurately identify instances of hate speech, as classified by human readers, from a collection of internet comments consisting of hate speech, offensive but not hateful speech, and benign speech.

To accomplish this, multiple classifiers will need to be created, including a part-of-speech tagger, a parser, and the final hate speech classifier.

Data

Field	Type	Description
count	Numeric	Number of human readers who classified comment
hate_speech	Numeric	Number who classified comment as hate speech
offensive_language	Numeric	Number who classified comment as offensive language
neither	Numeric	Number who classified comment as neither hateful nor offensive
class	Numeric	Classification of comment based upon majority rule of readers
tweet	String	Text of comment to be classified

The data consists of approximately 24,000 individual tweets that were reviewed and classified by human readers. For the purposes of classification model training, only the text of the comment will be analyzed.

Features to train the classifier will exclusively be drawn from the text of the comment itself. The individual words, as well as bigrams and trigrams may be easily extracted from each comment and analyzed. Hashtags and emoji use can also be examined for correlation with hate speech.

Known Unknowns

The majority of natural language processing techniques are relatively new to me, but the task appears to be doable. Documentation for many of the tools needed to complete the product are plentiful online. The biggest hurdle thus far has been my inexperience in memory management and optimization; working with the large data structures that text involves appears to be pushing my current hardware to its limits.