

Motivation

A central challenge for automatic hate speech detection on social media is the separation of hate speech from other instances of offensive language. There are important qualitative differences between different types of potentially abusive language that need to be considered. For example a tweet quoting rap lyrics that contain potentially racist or sexist terms should not be regarded in the same way as a tweet that directs racist slurs at another user. Existing work often does not make this distinction and many types of abusive language are often conflated.

Data

We collected tweets that contained terms in the Hatebase.org lexicon and labeled a sample of 25k of these into three categories to distinguish between hate speech (language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group), offensive language that does not meet this definition, and non-offensive language. We use tweets where two or more human coders agreed.

Model

We use a multi-class classifier with 5-fold cross-validation to predict each of the three classes. We first use logistic regression with an L1 penalty to select the best features then use L2 regularization to predict the class (Linear SVC led to similar results). Overall the model gets an F1 score of 0.91, although performance is worst for the hate class.

Features

- Bigram, unigram, and trigram features, each weighted by its TF-IDF.
- Syntactic structure: sequences of one, two, and three adjacent Penn Part-of-Speech (POS) tags.
- Tweet quality: Flesch-Kincaid Reading Ease and Fleisch Readability scores.
- Tweet sentiment: amount of positive, negative, neutral, and overall sentiment of each tweet using VADER sentiment.
- Binary and count indicators for hashtags, mentions, and URLs.
- Number of characters, words, and syllables.
- We also tried character n-grams and a custom word embedding but they did not improve our model.

Conclusions

Future work should consider more fine-grained definitions to avoid misclassifying offensive language as hate speech. Much of what has been considered hate speech is likely much more innocuous. Human coders appear to be biased towards recognizing certain types of hate speech and not others.

Automated Hate Speech Detection and the Problem of Offensive Language

Thomas Davidson¹, Dana Warmley², Michael Macy^{1,3}, Ingmar Weber⁴
 {¹Department of Sociology, ²Center for Applied Mathematics, ³Department of Information Science}, Cornell University, Ithaca, NY, USA, ⁴Qatar Computing Research Institute, Doha, Qatar.

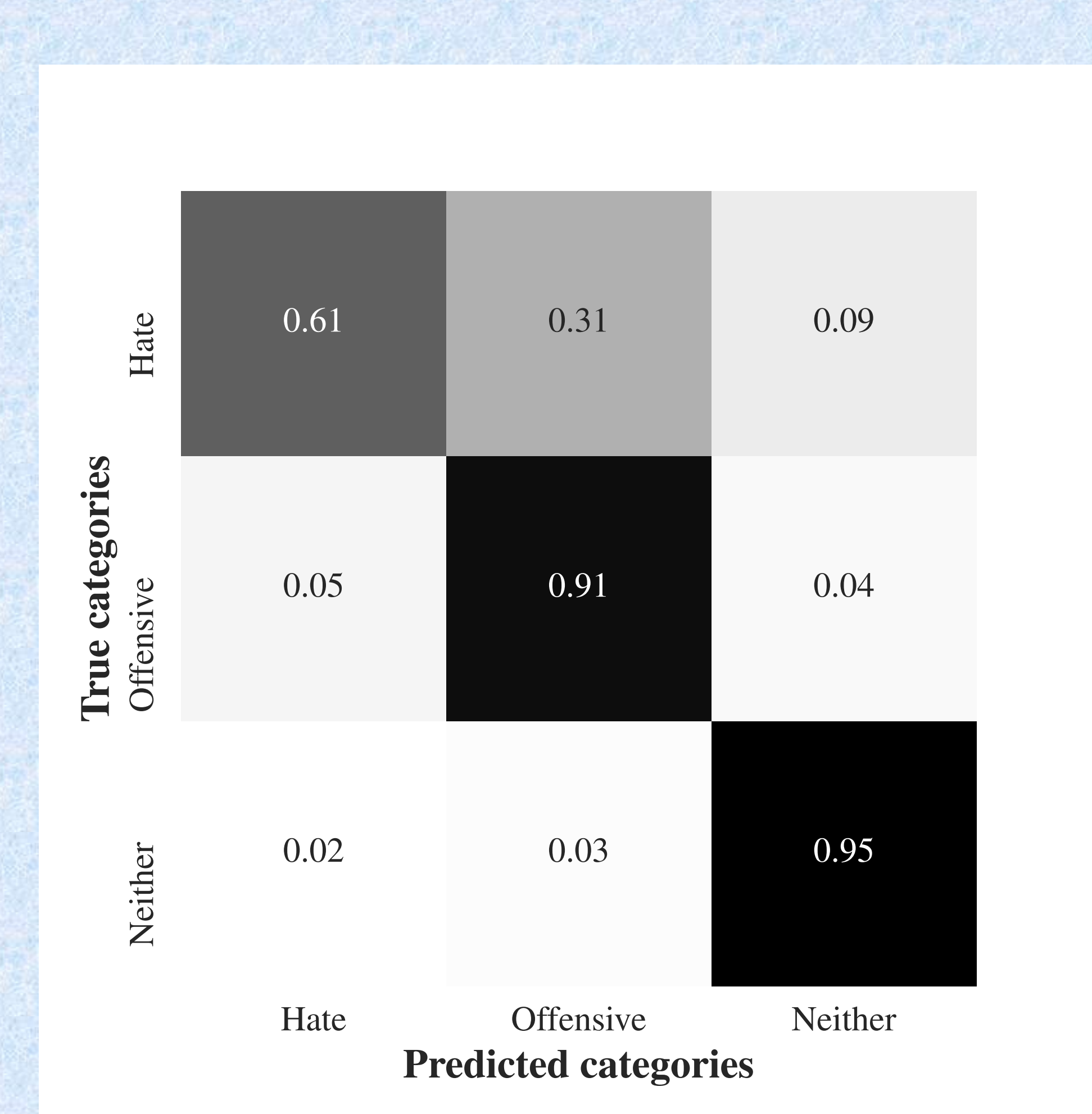


Figure 1: True versus predicted category

Hate speech

- Precision and recall scores for the hate class are 0.44 and 0.61, respectively (see Figure 1).
- Hate speech misclassified as offensive language appears to be less hateful and perhaps mislabeled by coders, although there are some true misclassifications.
- Hate speech classified as neither tends to lack profanity and slurs.
- Classifier is biased towards detecting the more prevalent forms of hate speech – particularly anti-black racism and homophobia

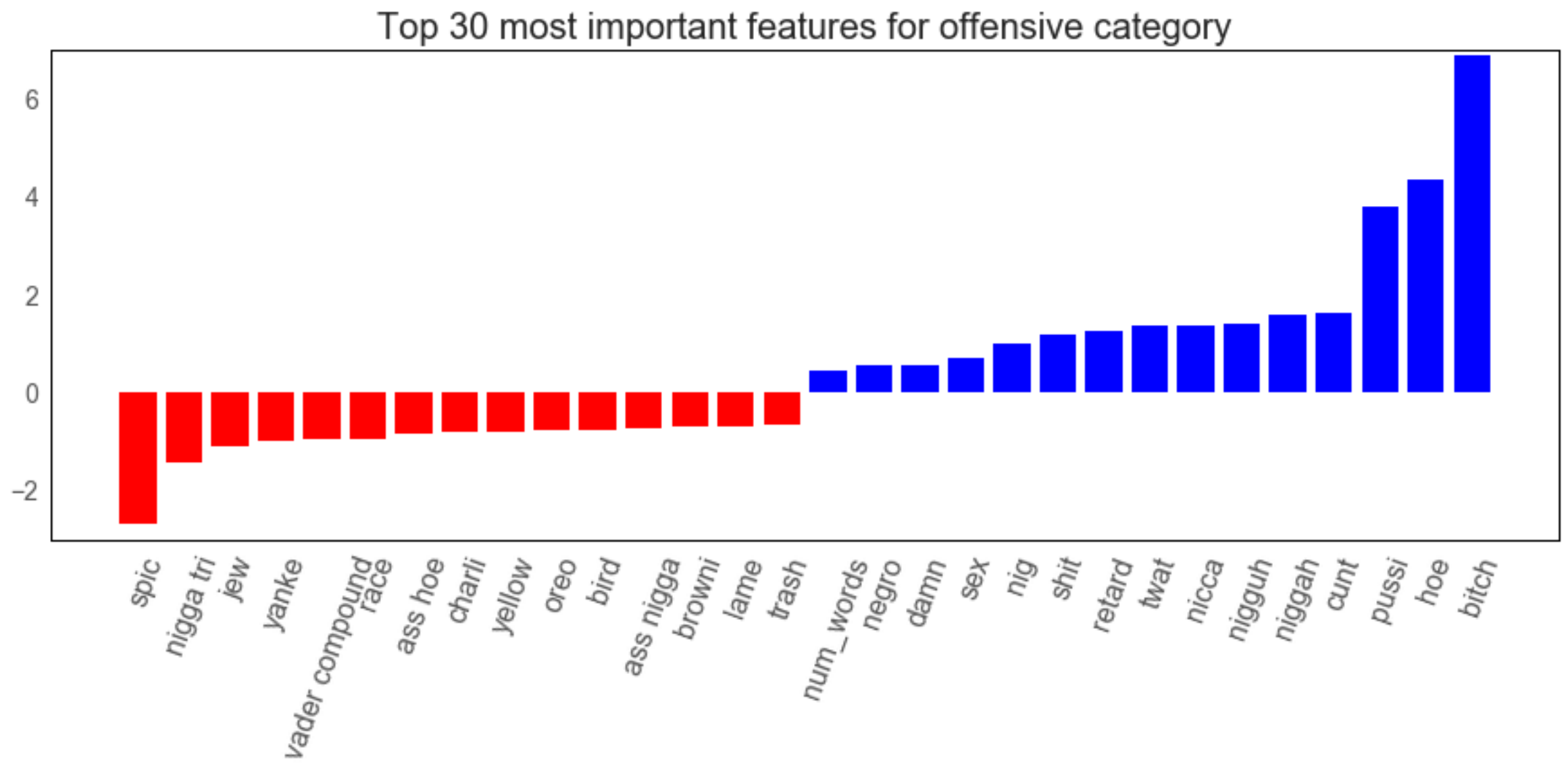
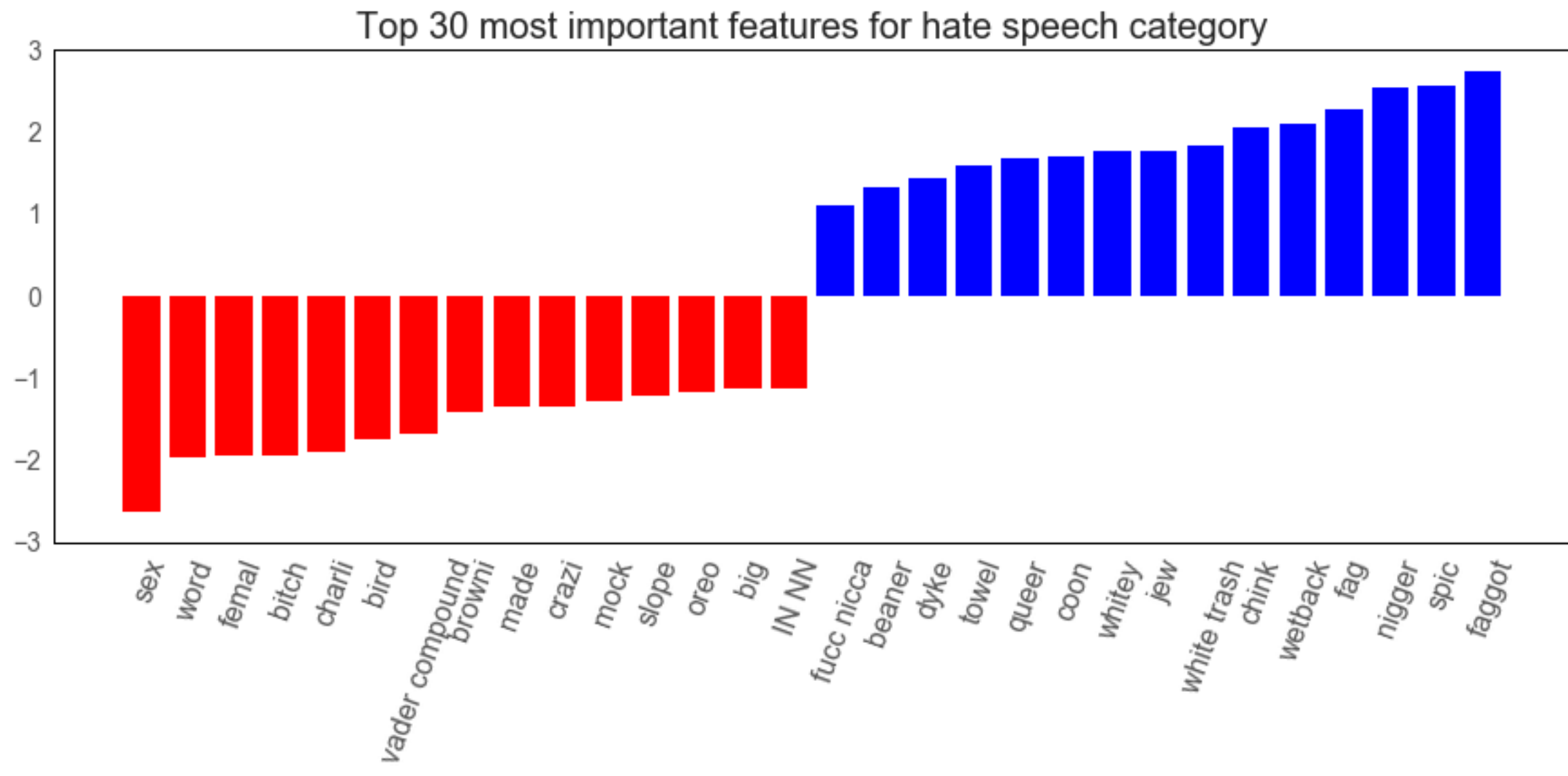


Figure 2: Example tweets from each section



Offensive language

- Human coders appear to consider racist or homophobic terms to be hateful but consider words that are sexist and derogatory towards women to be only offensive.
- While our model still misclassifies some offensive language as hate speech we are able to avoid the vast majority of these errors by differentiating between the two.
- Offensive classified as neither tends to contain profanities but not any racial or sexist slurs.

Neither

- Tweets with overall positive sentiment and higher readability scores are more likely to belong to the neither class. Presence of any profanity or slurs is a strong negative predictor.
- Most misclassifications appear to contain potentially offensive language.



Check out our Github page for more information and to access our data:
<https://github.com/t-davidson/hate-speech-and-offensive-language>

