

# Classifying Hate Speech Tweets

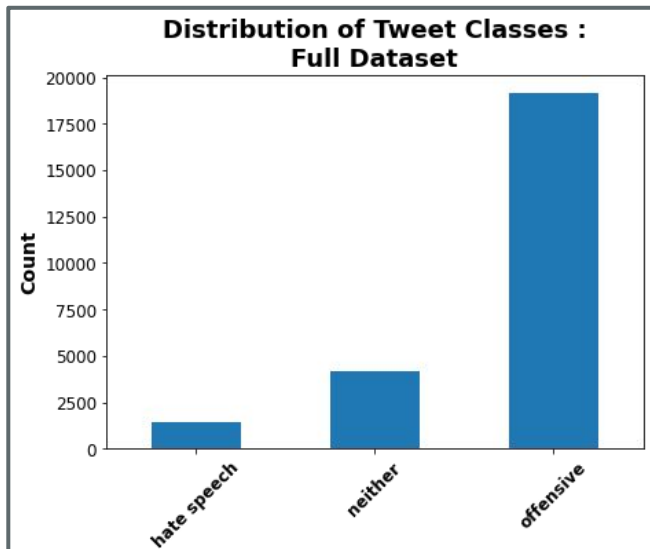
Max Steele

# Purpose

- Accurate classification
    - Hate Speech
    - Offensive Language
    - Neither
  - Maximize overall accuracy
    - Optimize catch rate for hate speech
  - Understand how to best use the model
    - Limit hateful conduct, not free expression
-

# Obtain, Scrub, & Explore

- Dataset from Davidson *et al.* (2017) - "Automated Hate Speech Detection and the Problem of Offensive Language"
  - 24,783 tweets
  - Classified by majority vote



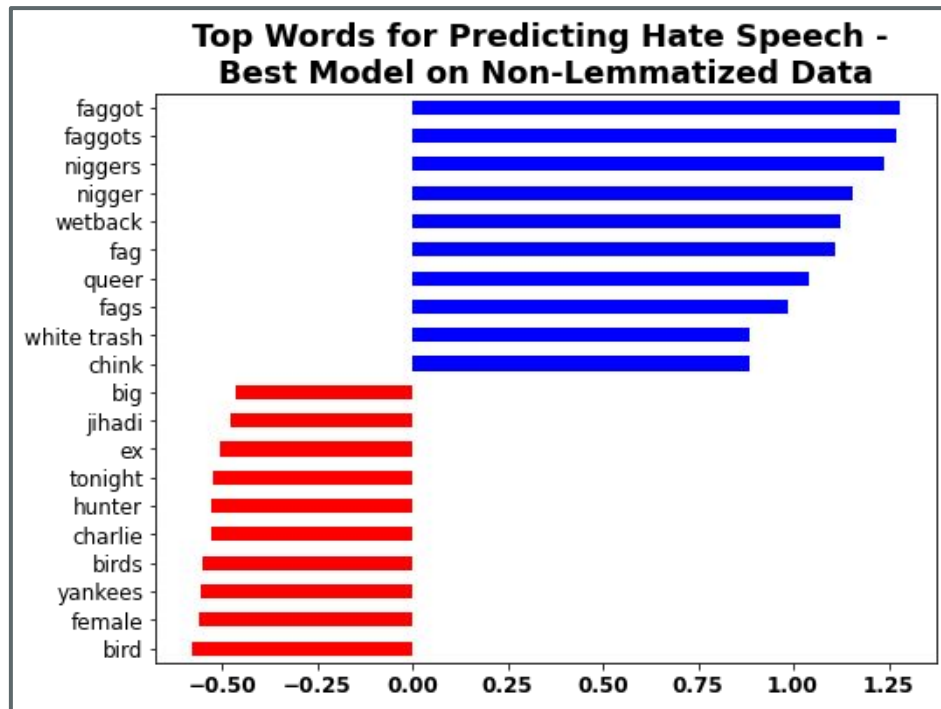
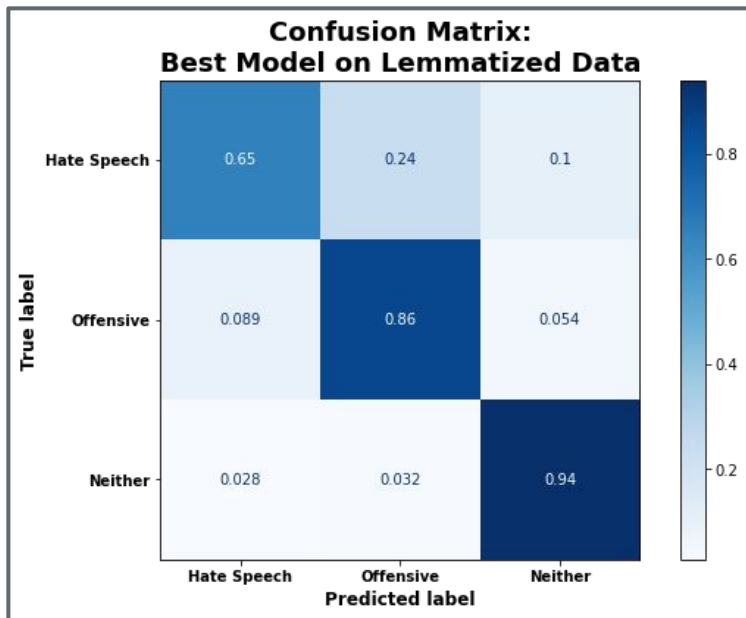
## Most Common Words for Corpus



- Undersampled “offensive language”
  - Final dataset 10,000 tweets
- Processed tweets by removing
  - URLs
  - Mentions
  - Emojis and symbols
  - Stop words

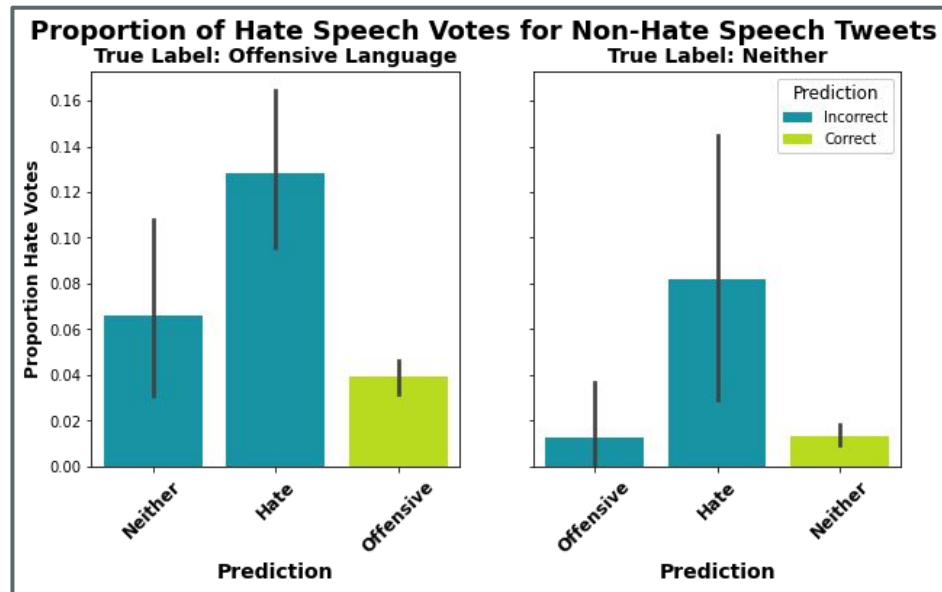
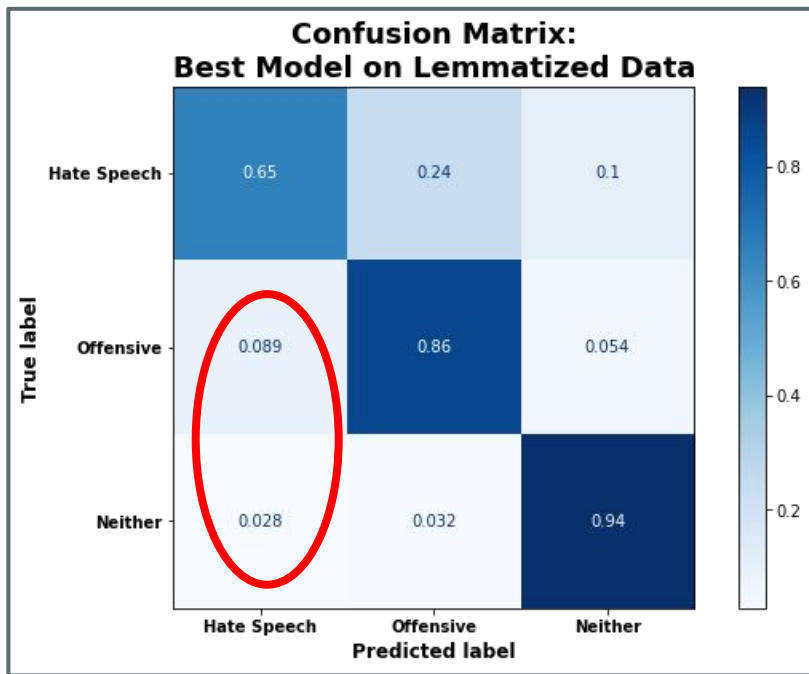
# Modeling

- Compared multiple types of classifier models
- Top model
  - Overall accuracy - 86%
  - Hate speech catch rate - 63%
  - Retrain on lemmatized data



# Recommendations

- Generally do NOT use the model to trigger automatic consequences

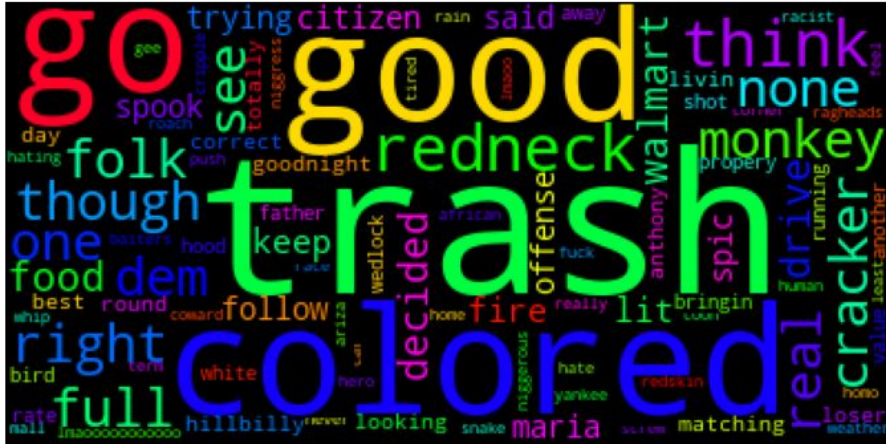


- Maintain a database of tweets investigated under hate policy
  - Decision
  - Reason

# Recommendations

- Actively assess and update the model
  - Investigate patterns of bias
  - Evolution of language

### Most Common Words for Hate Speech Misclassified as Neither



## Most Common Words for Offensive Misclassified as Hate Speech



# Summary

- Current model accurately classifies 86% of tweets
    - Catches 65% of hate speech
  - Useful as a tool to help uphold hateful conduct policy
  - Continue to rely on humans to investigate and report hate speech
    - No consequences triggered automatically
    - Database of investigated tweets
    - Dynamically update algorithm
-

# Future Work

- Add more tweets to dataset
  - Especially hate speech
- Engineer features
  - Mentions
- Deep NLP

---



# Thank you!

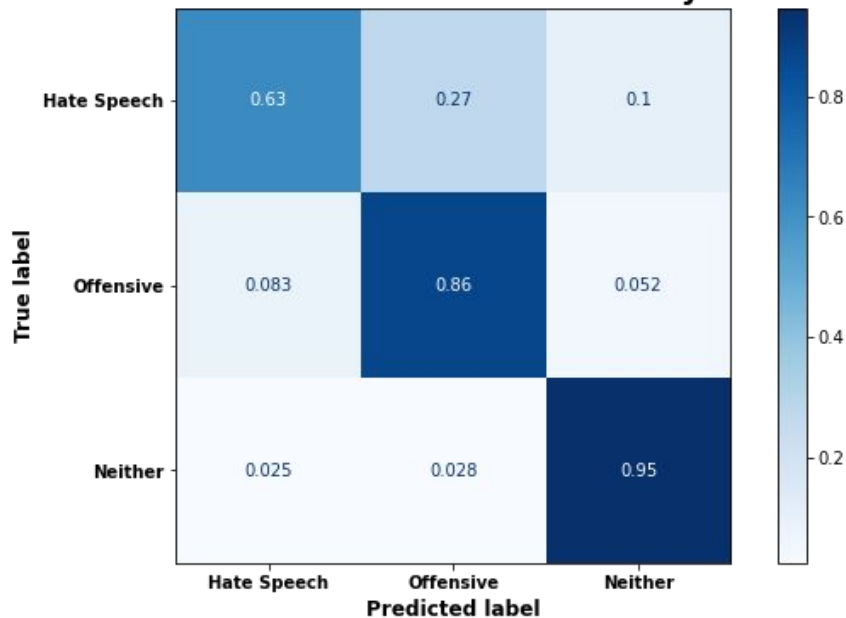
## Questions?

Max Steele

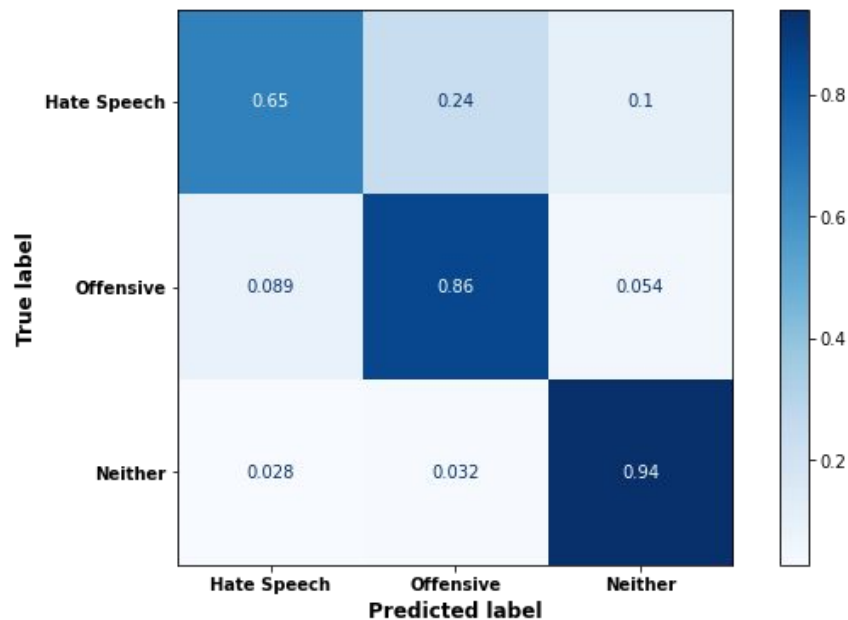
<https://github.com/zero731>

# Appendix I

**Confusion Matrix:  
Best Model - Count Vector LinearSVC  
Tuned for Balanced Accuracy**

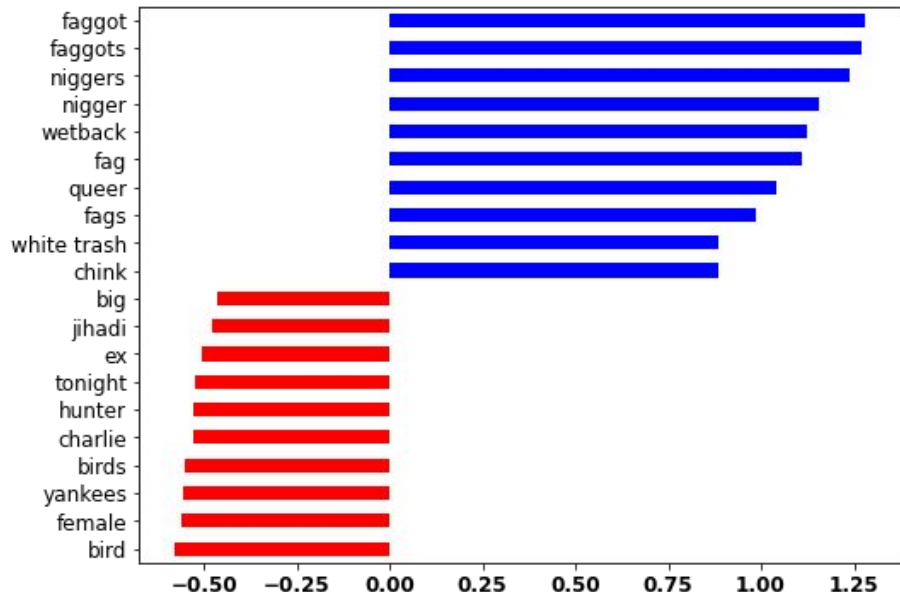


**Confusion Matrix:  
Best Model on Lemmatized Data**

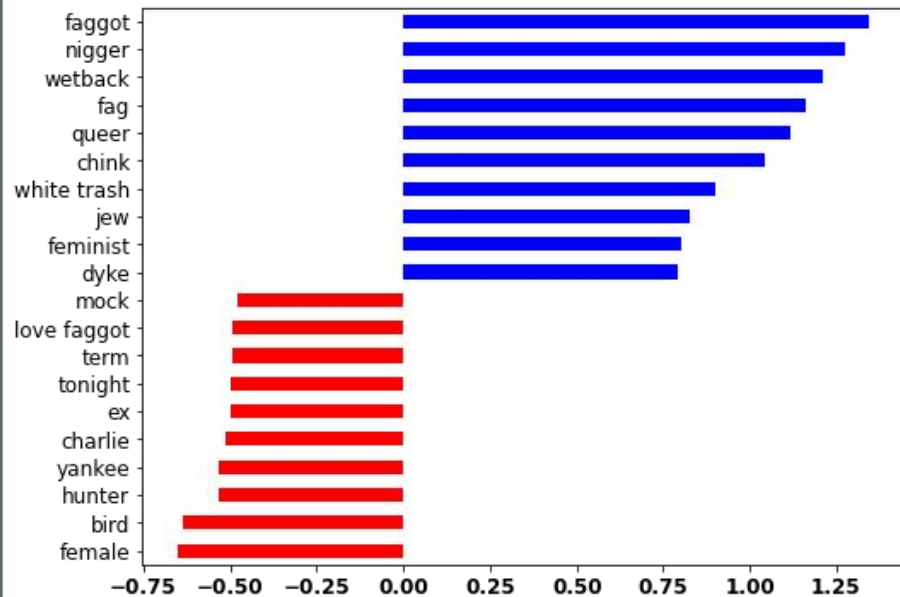


# Appendix II

**Top Words for Predicting Hate Speech -  
Best Model on Non-Lemmatized Data**

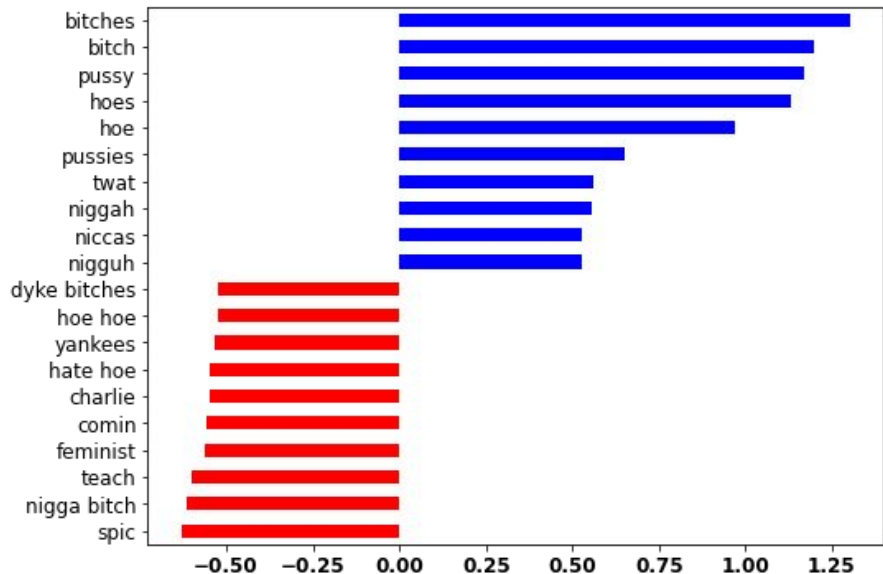


**Top Words for Predicting Hate Speech -  
Best Model on Lemmatized Data**

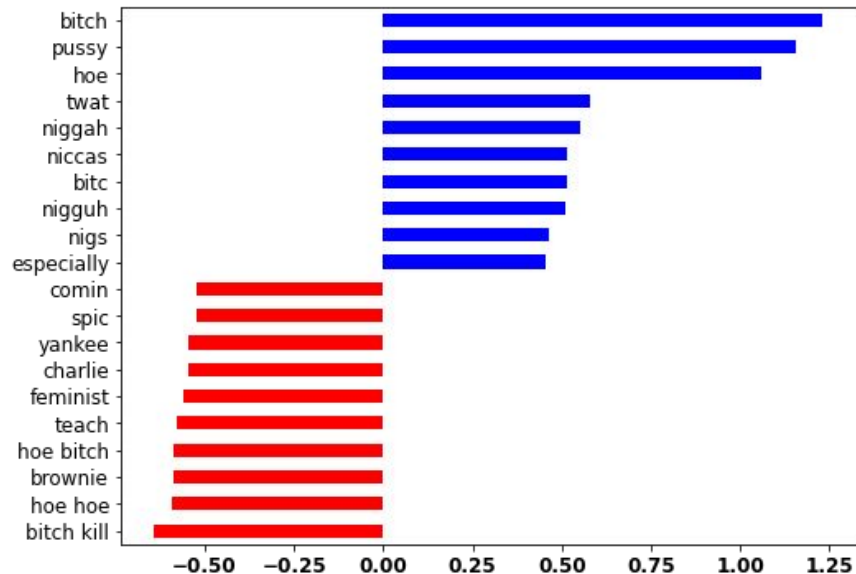


# Appendix III

**Top Words for Predicting Offensive Language -  
Best Model on Non-Lemmatized Data**

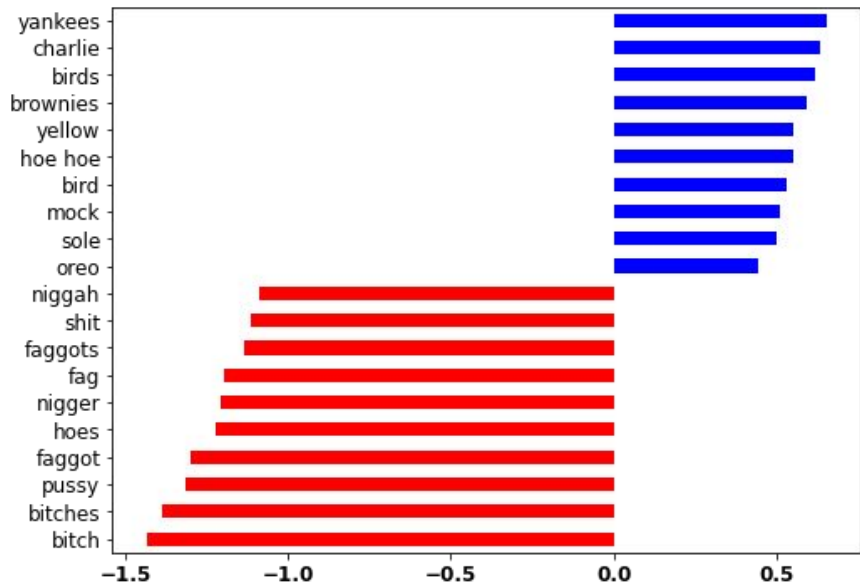


**Top Words for Predicting Offensive Language -  
Best Model on Lemmatized Data**

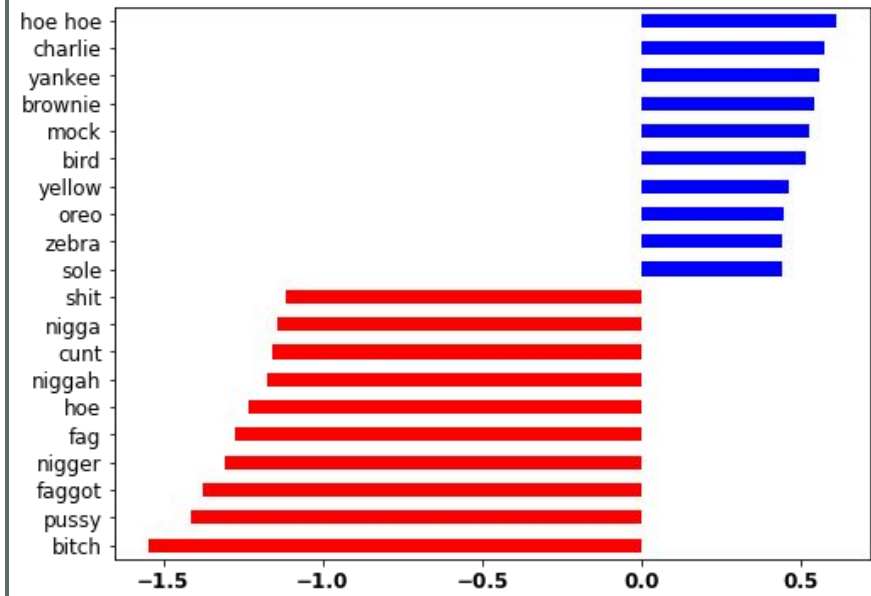


# Appendix IV

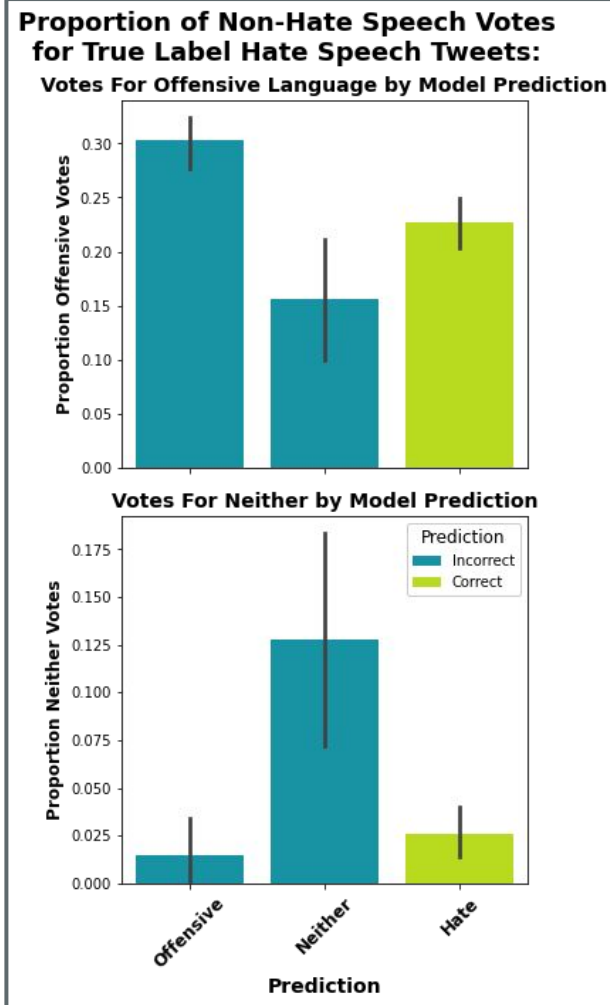
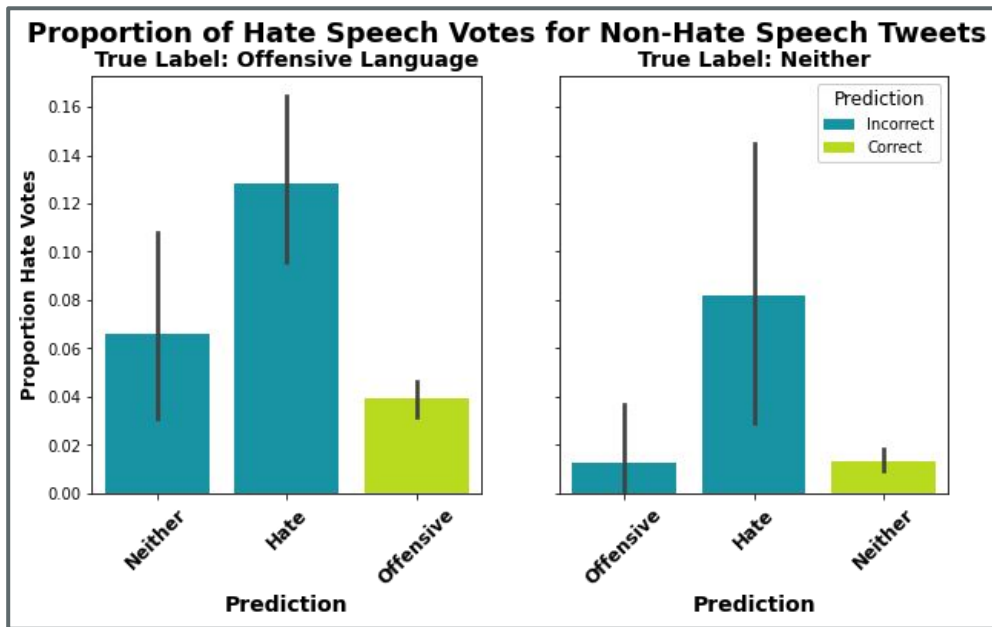
**Top Words for Predicting Neither -  
Best Model on Non-Lemmatized Data**



**Top Words for Predicting Neither -  
Best Model on Lemmatized Data**



# Appendix V



# Appendix VI

## Most Common Words for Hate Speech



# Appendix VII

## Most Common Words for Offensive Language





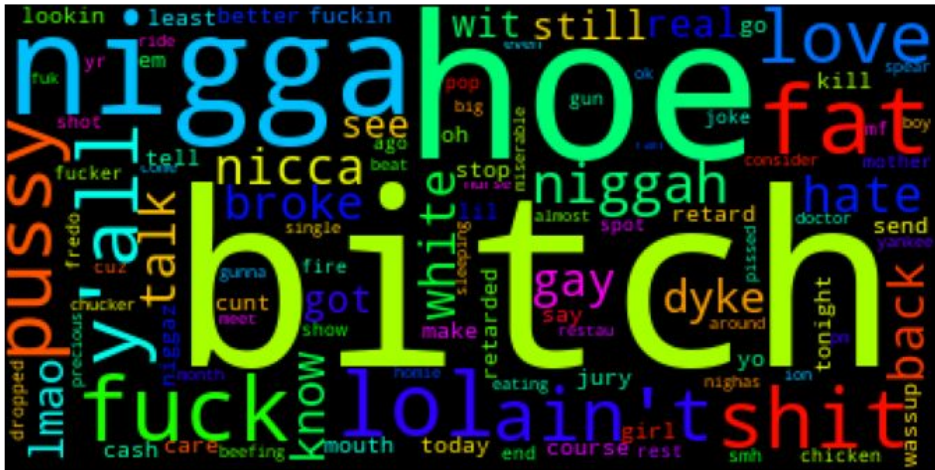
# Appendix VIII

## Most Common Words for Neither

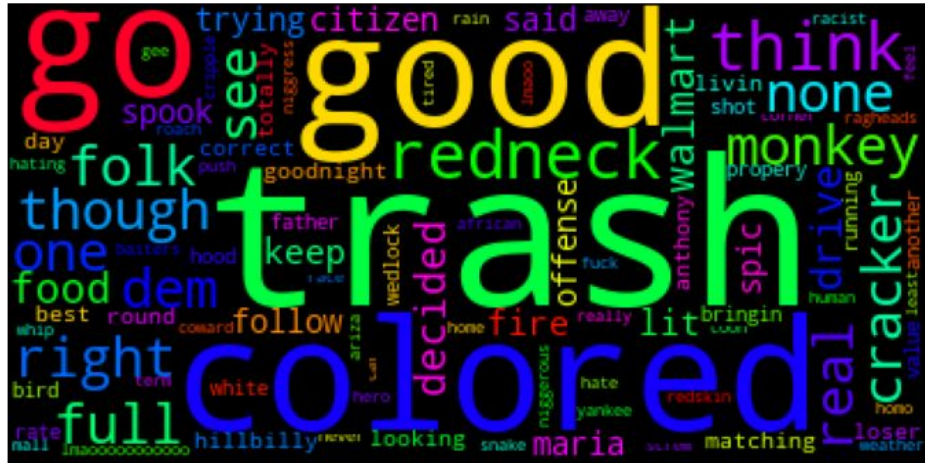


# Appendix IX

## Most Common Words for Hate Speech Misclassified as Offensive



### Most Common Words for Hate Speech Misclassified as Neither

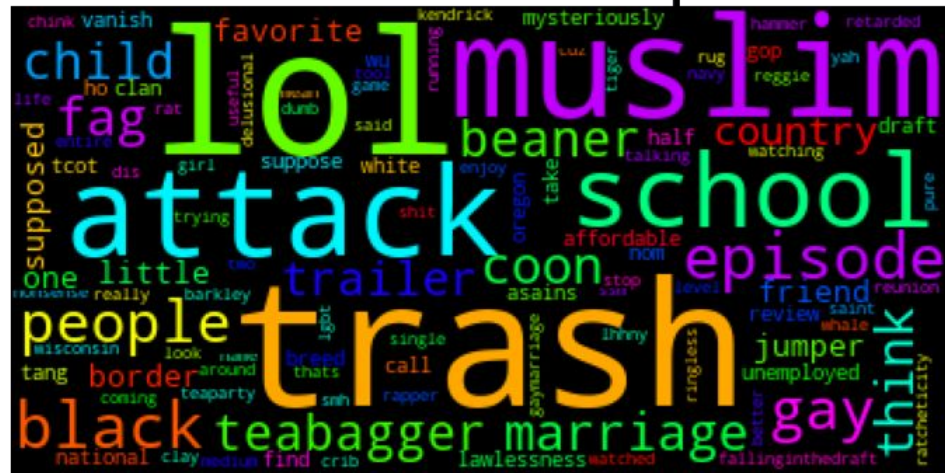


# Appendix X

## Most Common Words for Offensive Misclassified as Hate Speech



### Most Common Words for Neither Misclassified as Hate Speech



# Appendix XI

