Classifying Hate Speech Tweets

Max Steele

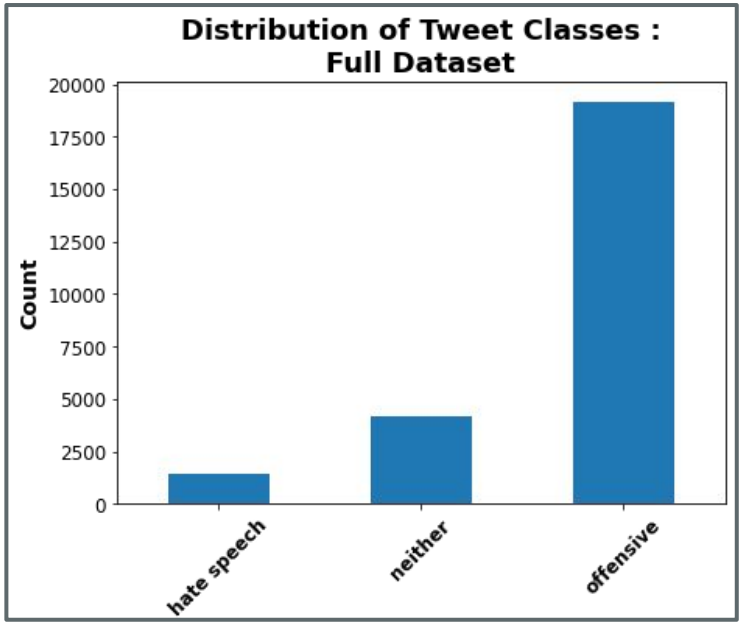
Purpose

- 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
  - Each tweet given a single "true" label based on majority vote

- Undersampled "offensive language"
  - Final dataset 10,000 tweets



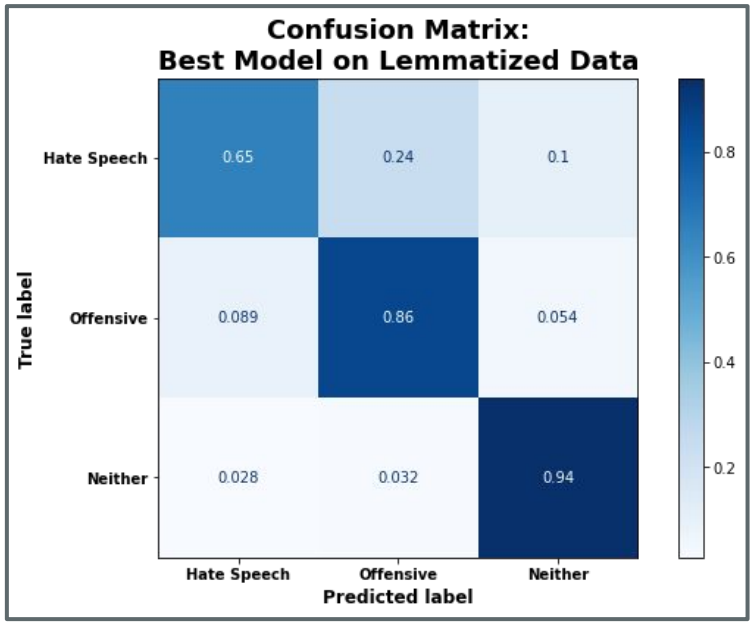
Obtain, Scrub, & Explore

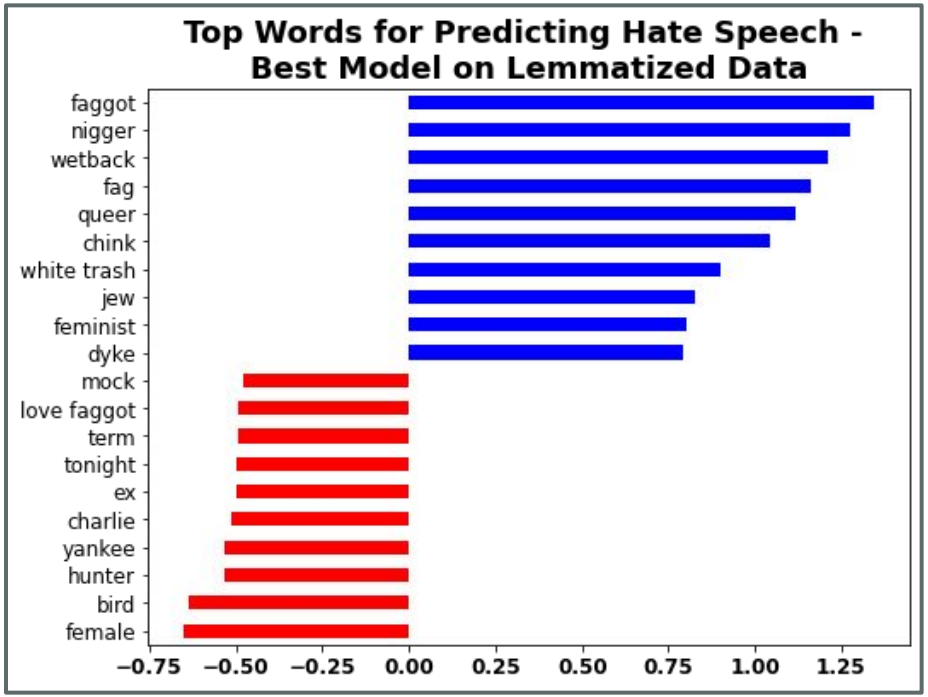
- Processed tweets by removing
  - URLs
  - Mentions
  - Emojis and symbols
  - Stop words

Most Common Words for Corpus

Modeling

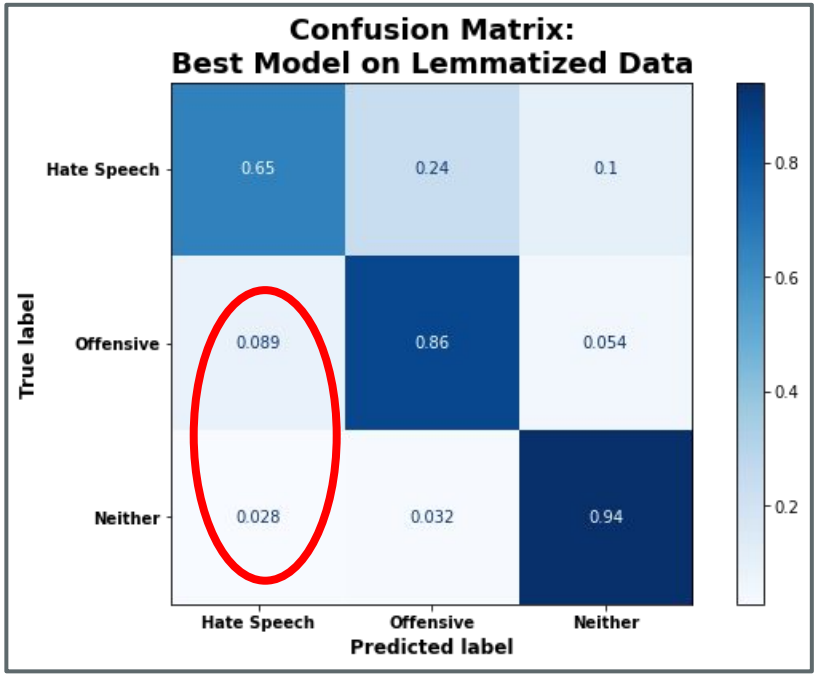
- Compared multiple types of classifier models
- Top model - LinearSVC
  - Overall accuracy - 86%
  - Hate speech catch rate - 63%





Recommendations

- Do NOT use the model to trigger ultimate consequences
  - Use to flag potential hate speech for follow-up investigation

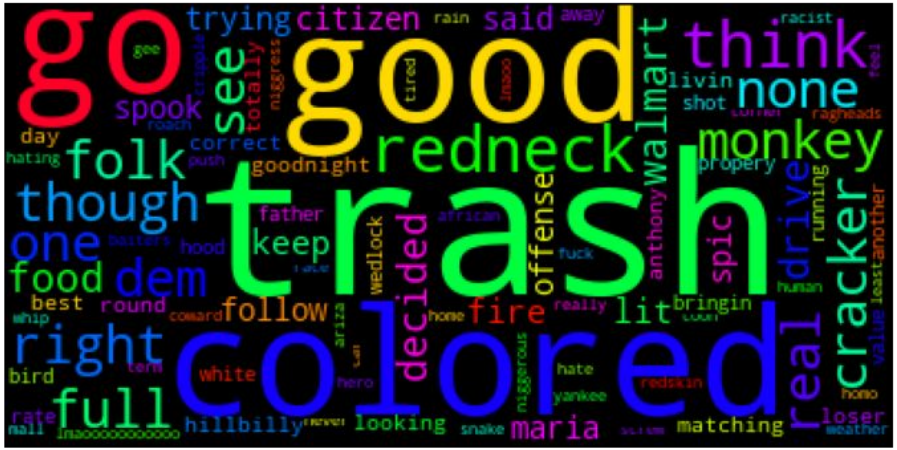


Recommendations

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

Most Common Words for Offensive Misclassified as Hate Speech

Most Common Words for Hate Speech Misclassified as Neither



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

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

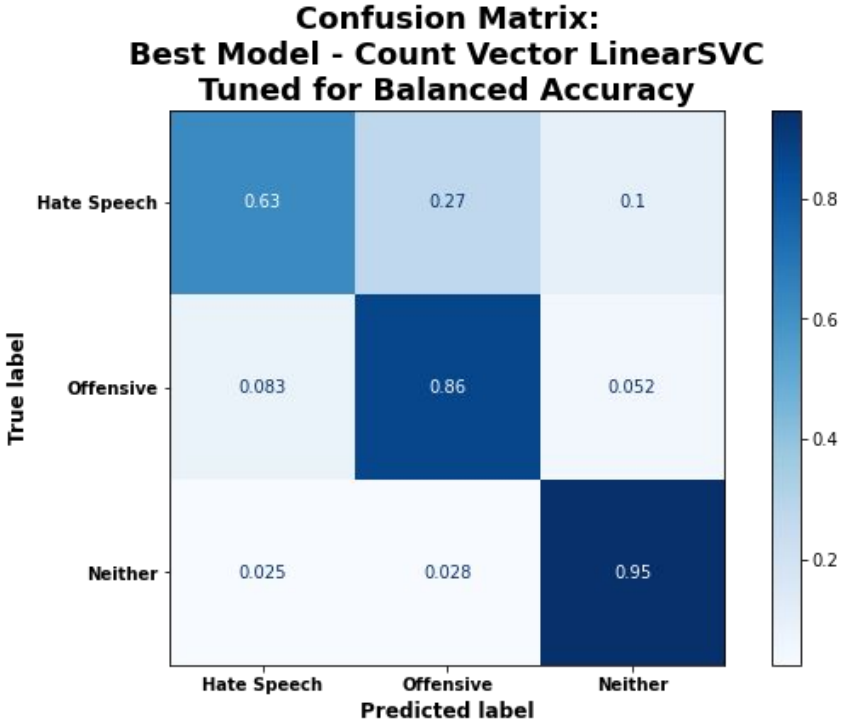
Thank you!

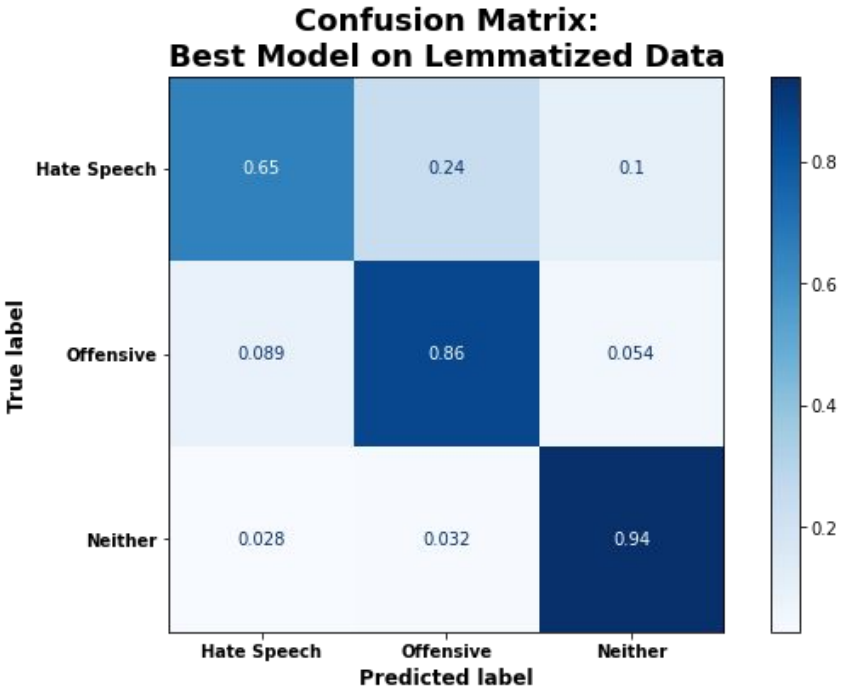
Questions?

Max Steele

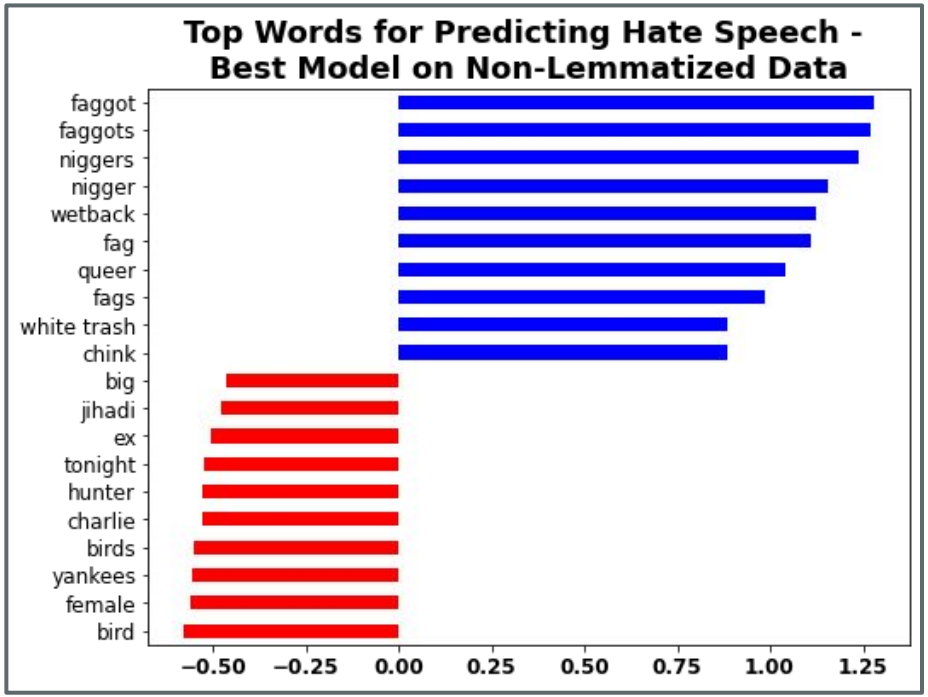
https://github.com/zero731

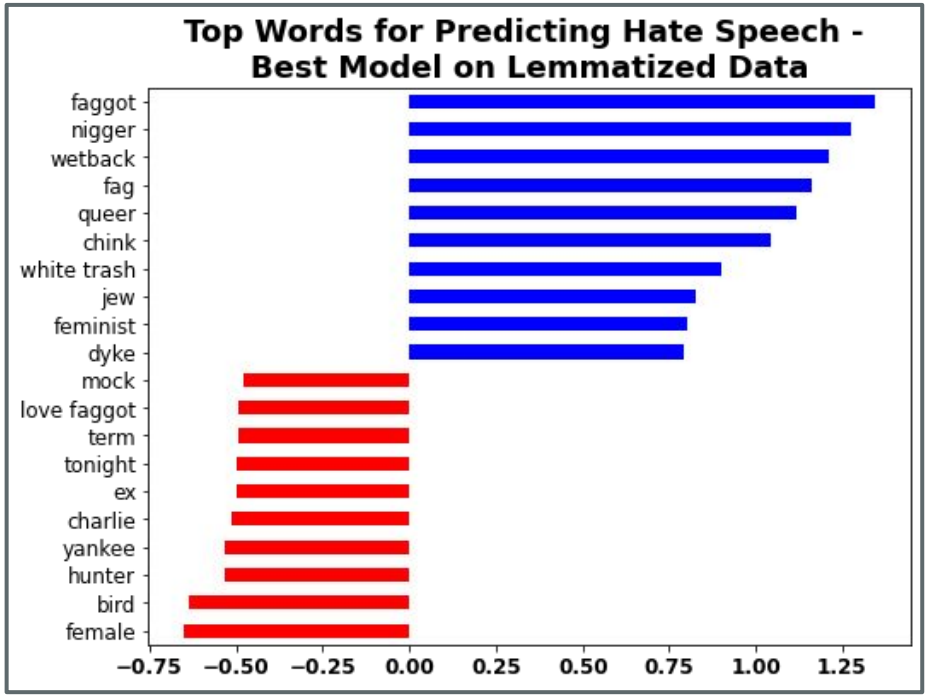
Appendix I


Confusion Matrix:
Best Model - Count Vector LinearSVC
Tuned for Balanced Accuracy
Hate Speech
Offensive
Neither
True label
0.63
0.27
0.1
0.083
0.86
0.052
0.025
0.028
0.95
0.8
0.6
0.4
0.2
Hate Speech
Offensive
Neither
Predicted label

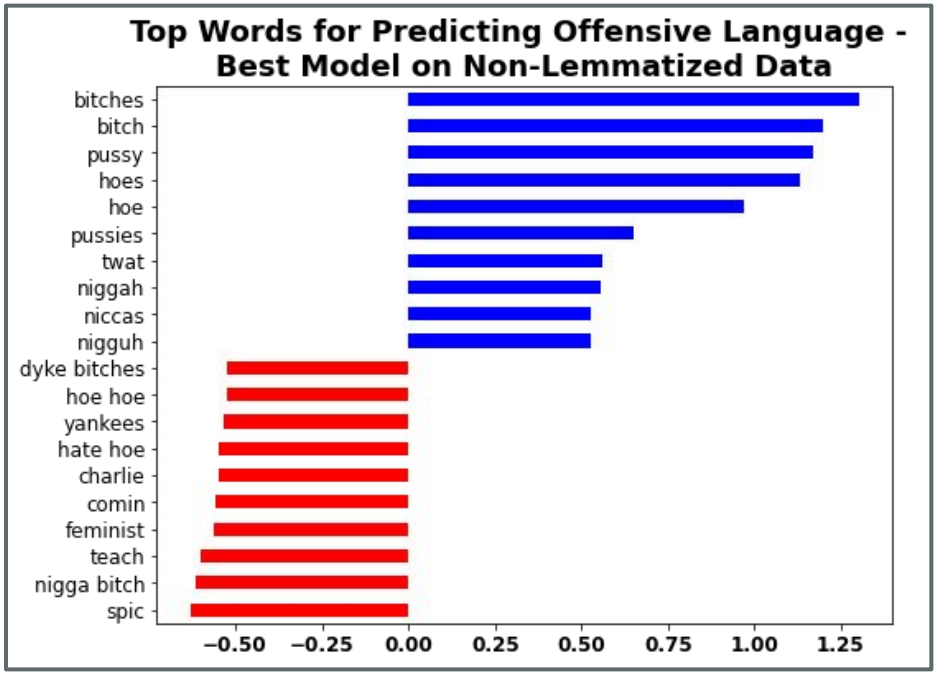

Confusion Matrix:
Best Model on Lemmatized Data
Hate Speech
Offensive
Neither
True label
0.65
0.24
0.1
0.089
0.86
0.054
0.028
0.032
0.94
0.8
0.6
0.4
0.2
Hate Speech
Offensive
Neither
Predicted label

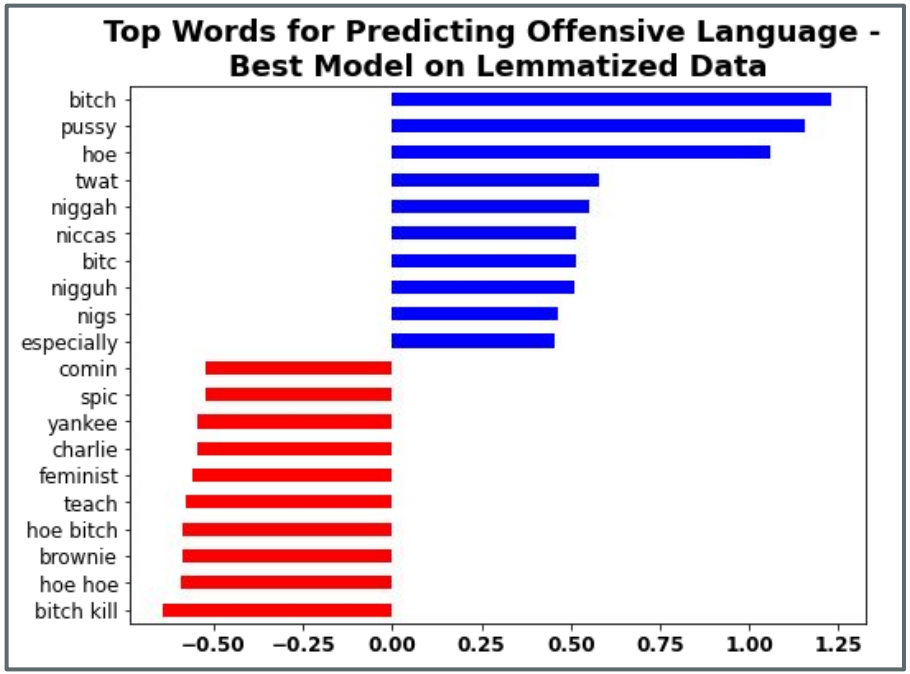
Appendix II


Top Words for Predicting Hate Speech -
Best Model on Non-Lemmatized Data
faggot
faggots
niggers
nigger
wetback
fag
queer
fags
white trash
chink
big
jihadi
ex
tonight
hunter
charlie
birds
yankees
female
bird
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

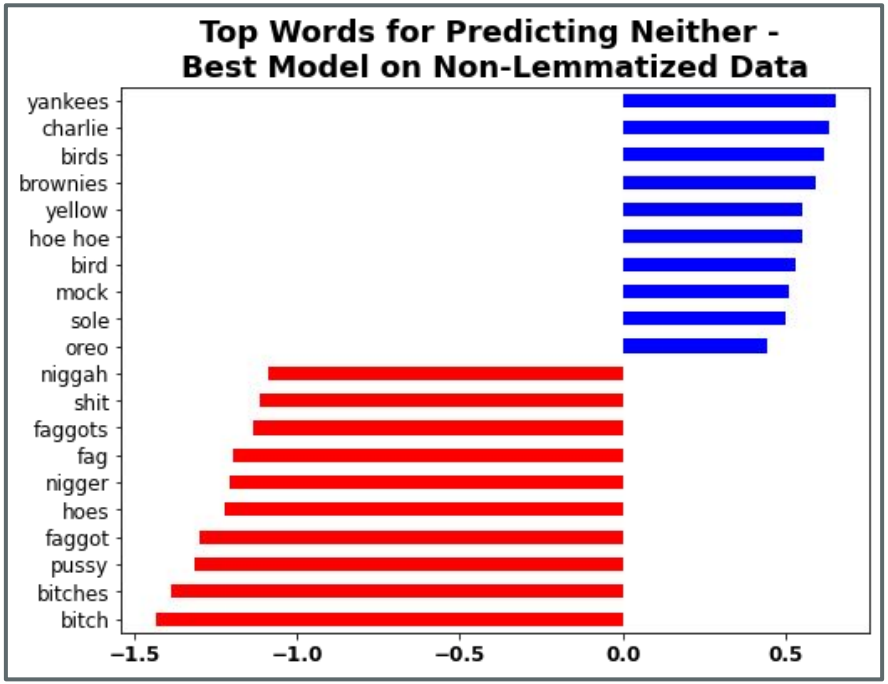

Top Words for Predicting Hate Speech -
Best Model on Lemmatized Data
faggot
nigger
wetback
fag
queer
chink
white trash
jew
feminist
dyke
mock
love faggot
term
tonight
ex
charlie
yankee
hunter
bird
female
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

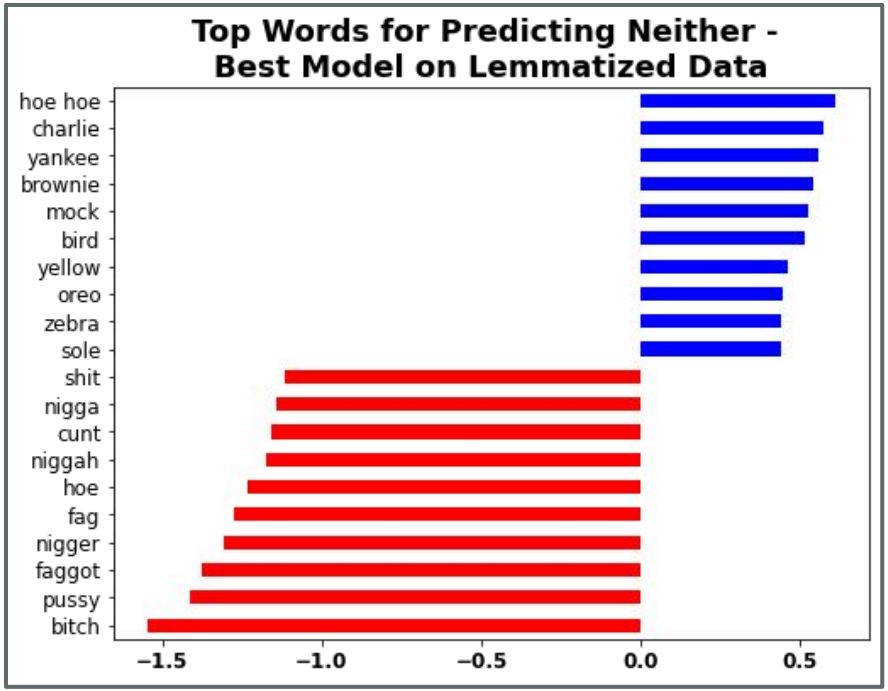
Appendix III



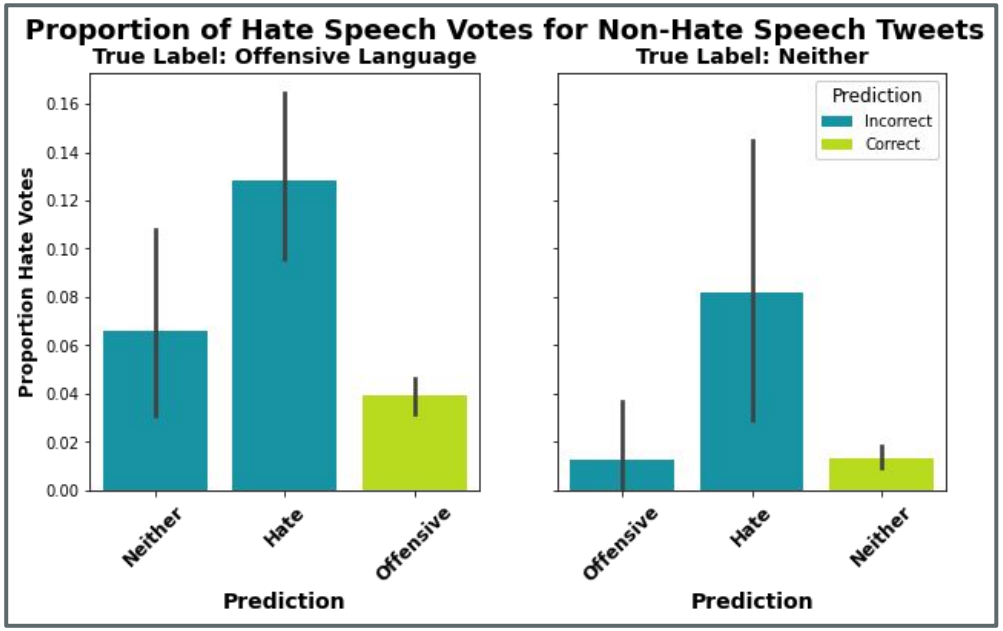


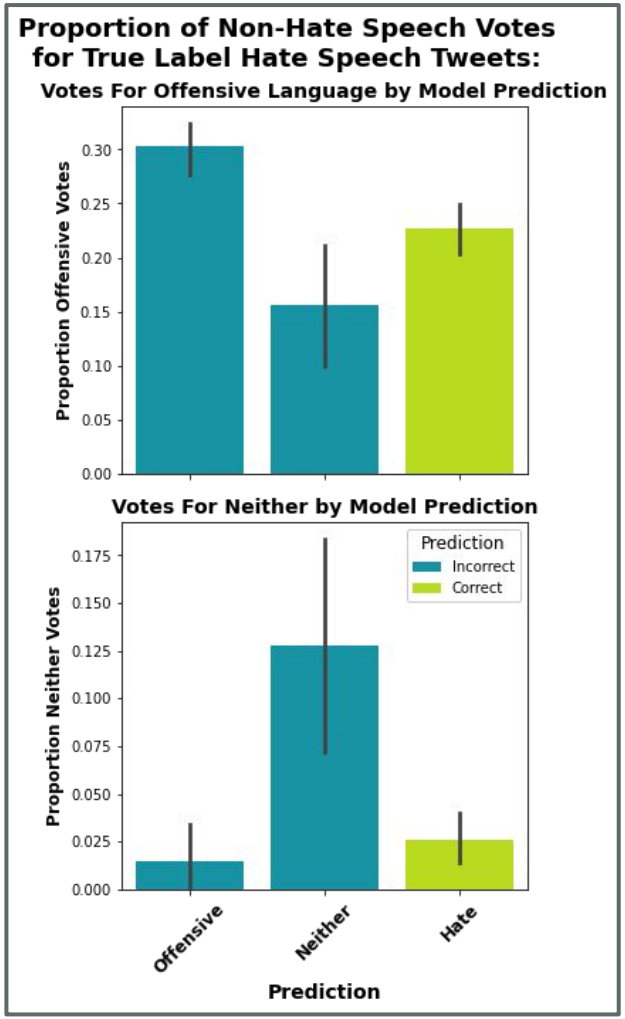
Appendix IV


Top Words for Predicting Neither -
Best Model on Non-Lemmatized Data
yankees
charlie
birds
brownies
yellow
hoe hoe
bird
mock
sole
oreo
niggah
shit
faggots
fag
nigger
hoes
faggot
pussy
bitches
bitch
−1.5
−1.0
−0.5
0.0
0.5


Top Words for Predicting Neither -
Best Model on Lemmatized Data
hoe hoe
charlie
yankee
brownie
mock
bird
yellow
oreo
zebra
sole
shit
nigga
cunt
niggah
hoe
fag
nigger
faggot
pussy
bitch
−1.5
−1.0
−0.5
0.0
0.5

Appendix V


Proportion of Hate Speech Votes for Non-Hate Speech Tweets
True Label: Offensive Language
True Label: Neither
Proportion Hate Votes
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Neither
Hate
Offensive
Offensive
Hate
Neither
Prediction
Prediction
Prediction
Incorrect
Correct


Proportion of Non-Hate Speech Votes for True Label Hate Speech Tweets:
Votes For Offensive Language by Model Prediction
Proportion Offensive Votes
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Votes For Neither by Model Prediction
Proportion Neither Votes
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Offensive
Neither
Hate
Prediction
Prediction
Incorrect
Correct

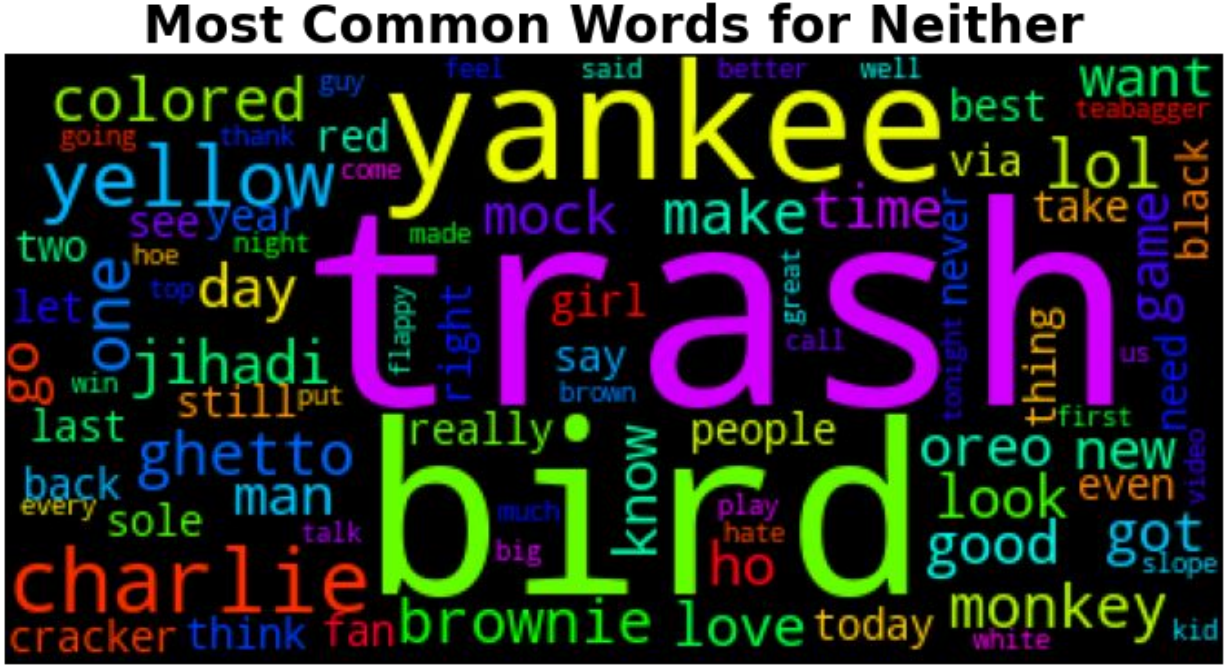
Appendix VI


Most Common Words for Hate Speech
bitch
faggot
nigga
fuck
trash
nigger
white
fucking
hate
bitches
ass
hoe
shit
fag
people
love
queer
retarded
got
coon
go
pussy
stupid
know
think
see
gay
cracker
retard
racist
black
lol
one
kill
want
look
say
stop
new
time
damn
nicca
chink
dyke
colored
right
talking
come
bad
fat
beaner
another
girl
well
said
ur
day
call
dont
really
niggah
make
spic
need
mean
little
tell
oh
even
smh
school
dick
called
ghetto
show
lmao
still
gonna
shut
real
us
boy
ain't
jew
happy
liwe
yo
good
let
cunt
guy
always
back
face
full
america
monkey
stru
y'all
ugly
manfuckin
dumb

Appendix VII

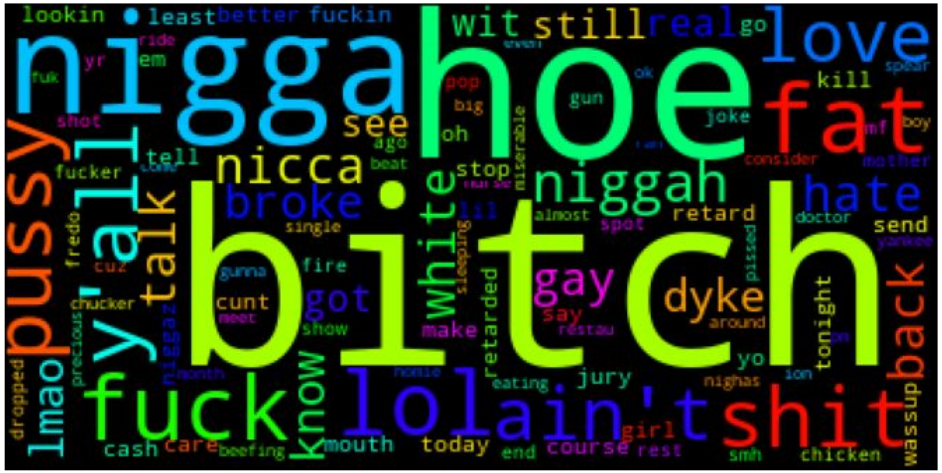
Most Common Words for Offensive Language

Appendix VIII

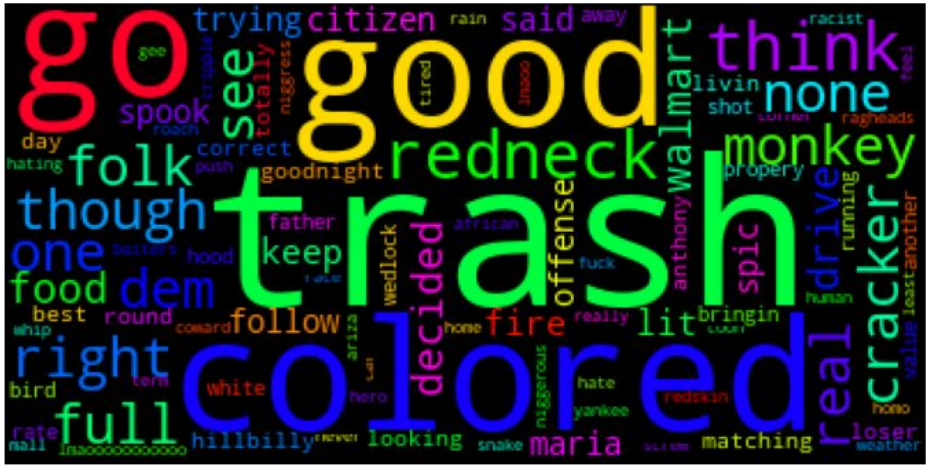


Appendix IX

Most Common Words for Hate Speech Misclassified as Offensive



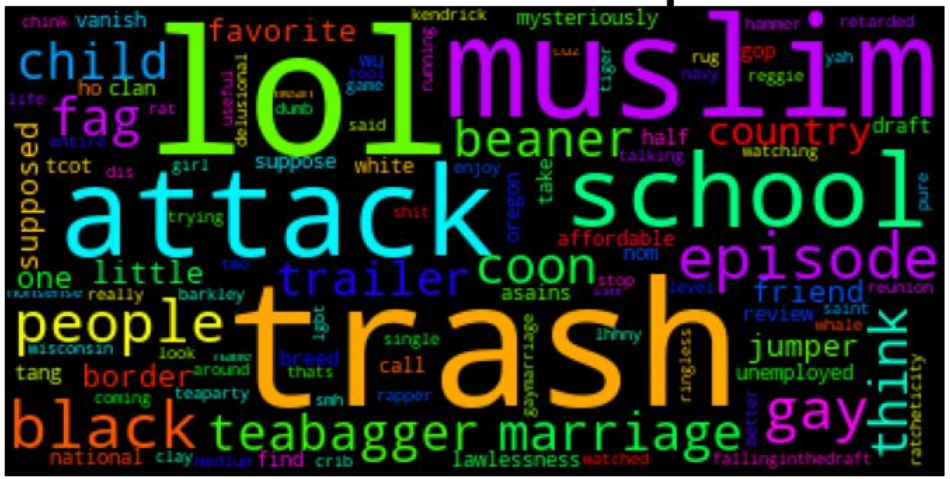
Most Common Words for Hate Speech Misclassified as Neither



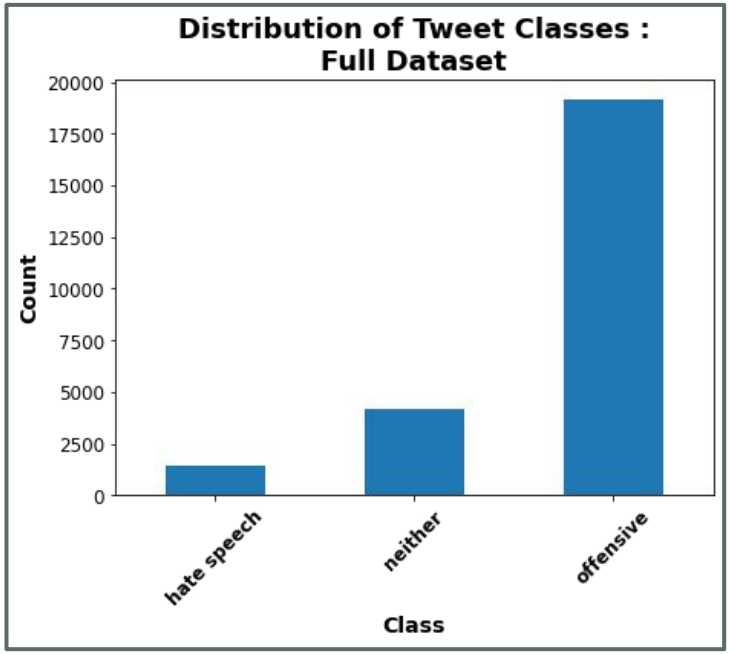
Appendix X

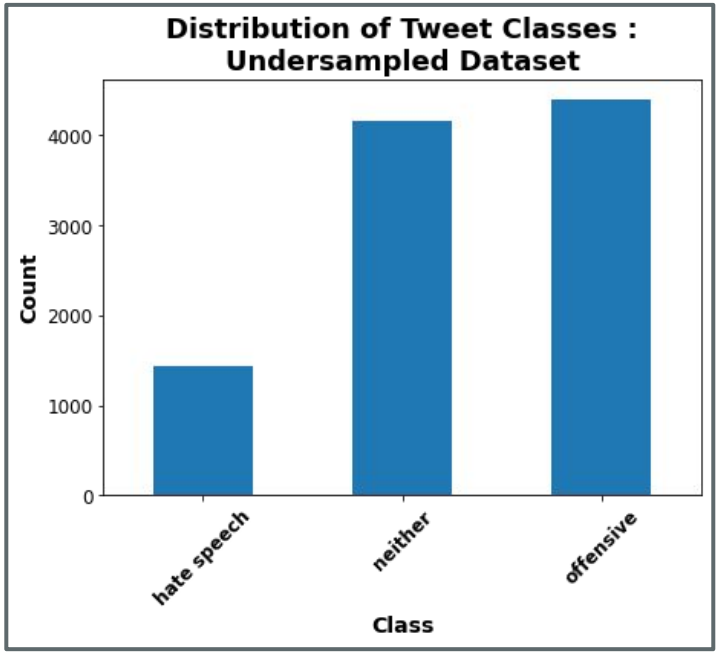
Most Common Words for Offensive Misclassified as Hate Speech

Most Common Words for Neither Misclassified as Hate Speech



Appendix XI


Distribution of Tweet Classes :
Full Dataset
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
hate speech
neither
offensive
Class


Distribution of Tweet Classes :
Undersampled Dataset
Count
4000
3000
2000
1000
0
hate speech
neither
offensive
Class