Ethics
& Biases

Exclusive Algorithms

K. KENNETH DAVIS

CAPSTONE PROJECT
DSI 1213

Machine Learning
Engineer

K.Kenneth Davis

The Trans Capitalist Consulting

Thetranscapitalist.com

Problem Statement

Does the Figure Eight Inc hate speech algorithm misclassify hate speech or offensive language tweets, and in doing so, is this harmful to marginalized communities?

Hate Speech Algorithms

Hate Speech	Offensive Language
Neither	Count

Data

01 Used 10k rows of tweets & 'class' is the target y

02 Multi-classification data

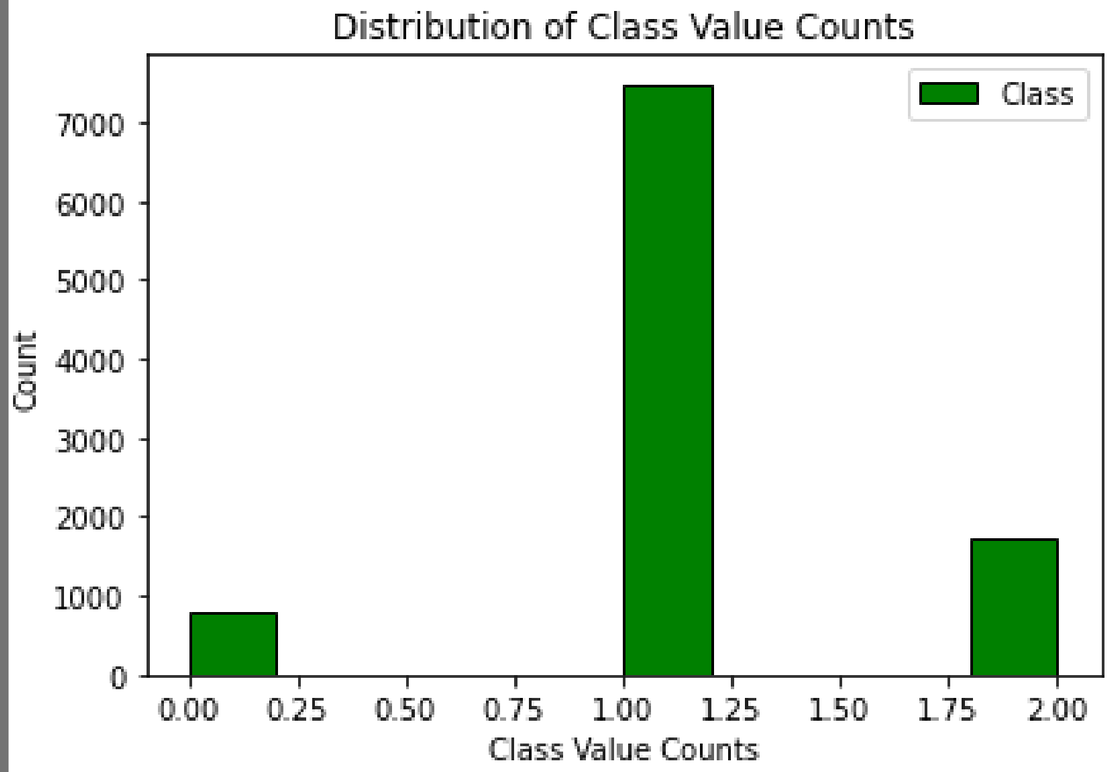
Cleaning

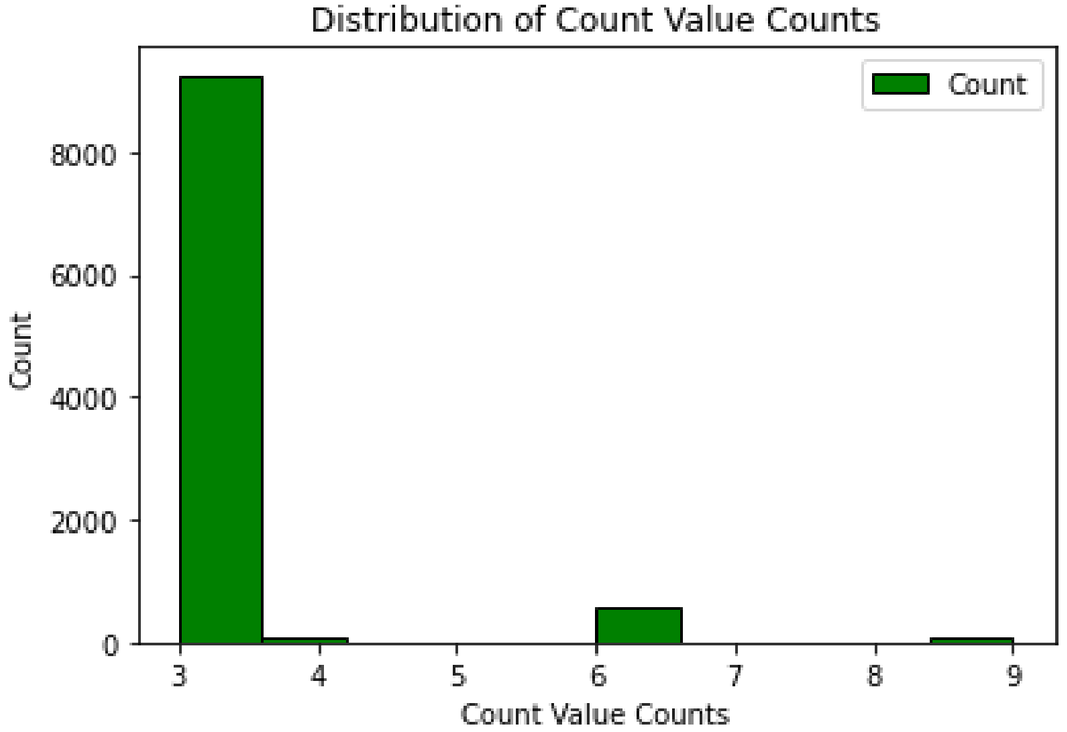
01 Removing punctuation and HTML

02 Creating text files and csv files for different dataframes.

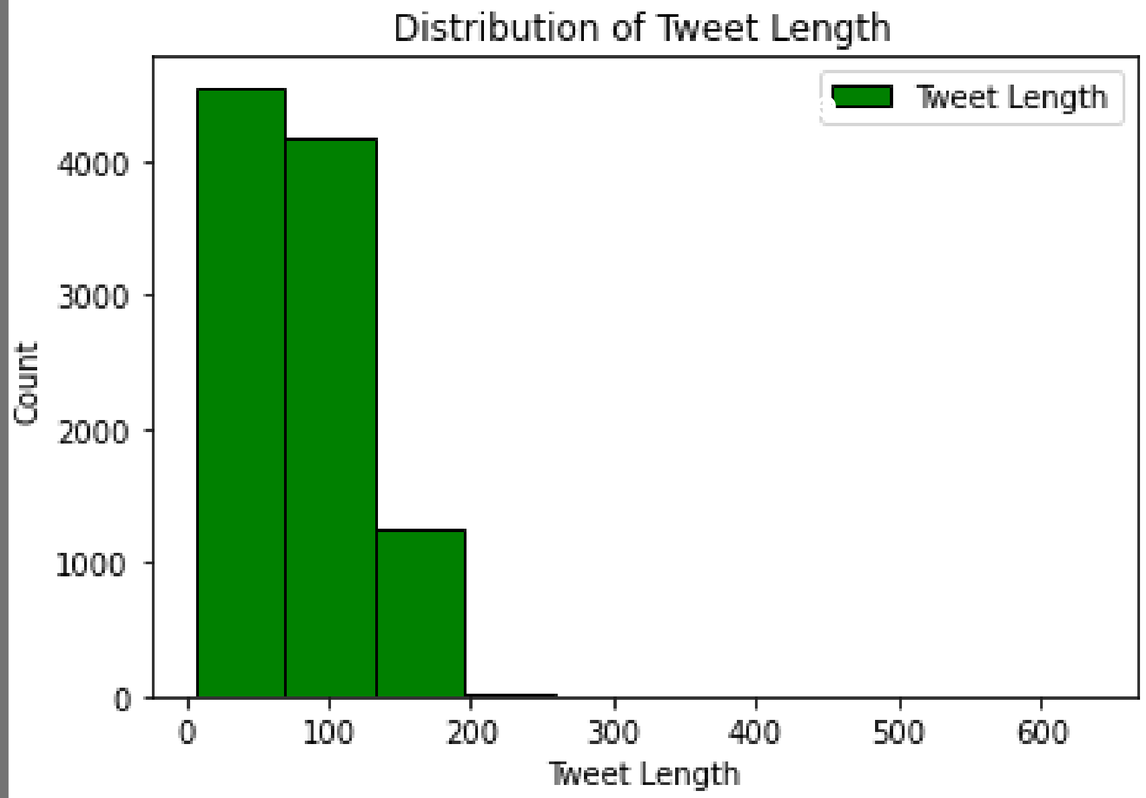
	count	hate_speech	offensive_language	neither	class	tweet
0	3	0	0	3	2	RT mayasolovely As a woman you shouldn't comp...
1	3	0	3	0	1	RT mleew17 boy dats cold...tyga dwn bad for c...
2	3	0	3	0	1	RT UrKindOfBrand Dawg RT 80sbaby4life You eve...
3	3	0	2	1	1	RT C_G_Anderson viva_based she look like a tr...
4	6	0	6	0	1	RT ShenikaRoberts The shit you hear about me ...

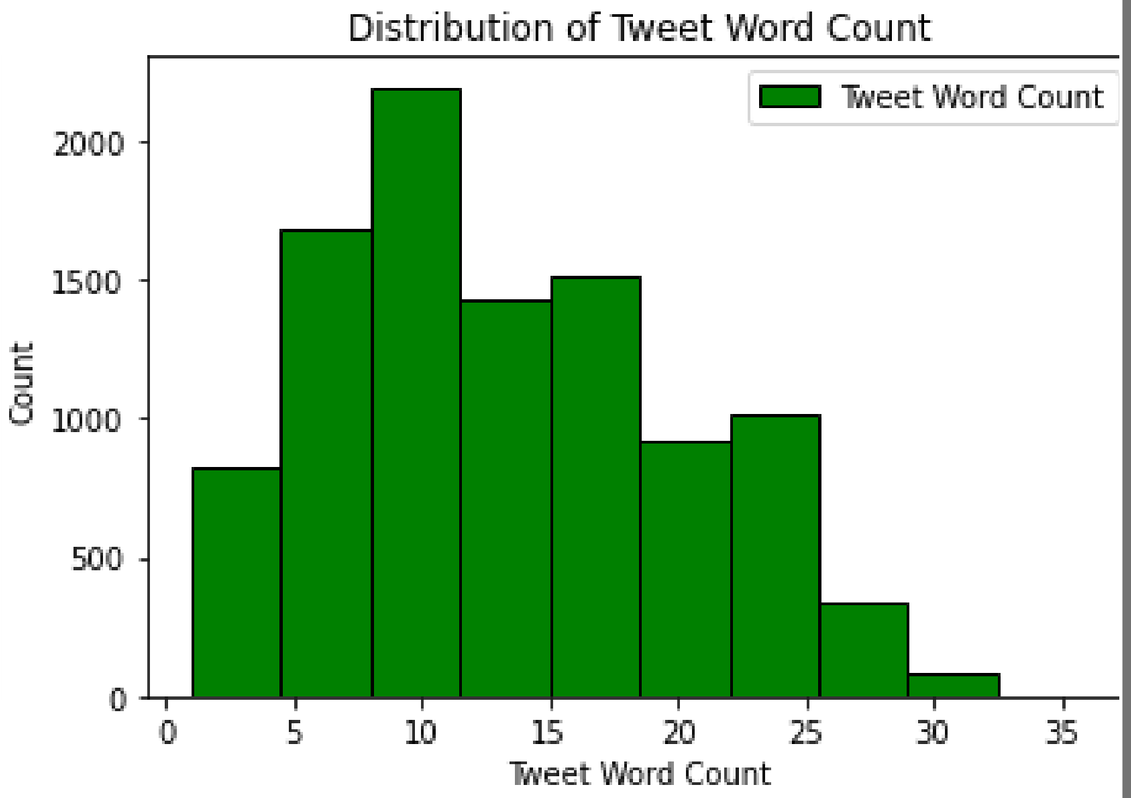
Value Counts Distributions



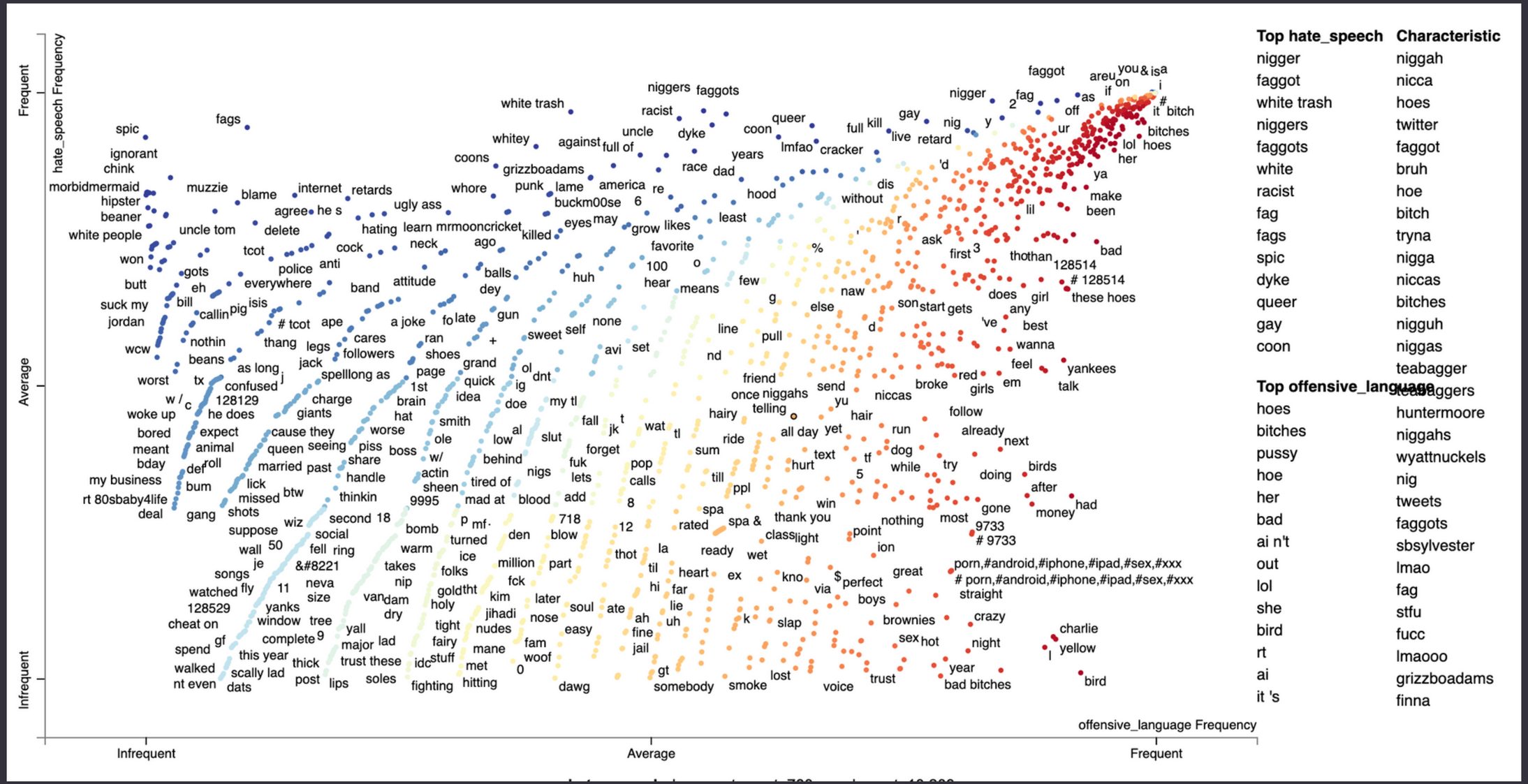


Tweet Length v. Word Count





Scatter Text Chart



Scatter Text Context

hate_speech document count: 780; word count: 10,203
offensive_language document count: 9,220; word count: 119,528

Term: fags

hate_speech frequency:
48 per 25,000 terms
22 per 1,000 docs
Some of the 18 mentions:

"Why people think gay marriage is okay is beyond me. Sorry I don't want my future son seeing 2 **fags** walking down the street holding hands an

😳 RT KingHov1313 They recruited buku **fags** in here😳...a warehouse is no place for them😉

Derek Kirchner7 tori mills02 **fags&**#128567;😁

Im Thirst randies are **fags**

RepulsiveTool And the only band that sucked worse live is those **fags** Guns & Poses.

SophieRo3 if only there was a medal for that "we hate **fags,** but we won't kill them" medal.

Thank u for the privilege to let **fags** live.

YourDudeFerg gay **fags**

Zohdaa **fags**

_Nadiaxo **fags**

caykelly16 MadisonEarhart u guys are **fags.**

davidly62793584 how many fuckin **fags** did a even get? Shouldnt be allowed into my wallet whilst under the influence haha

femalejokerhoee yeah he sucks just like the other **fags** who are... oh nevermind there are NO fags in the #NFL

gellebeaux_ whatever hoe bitch #1 them **"fags"** beat yall sad asses too

genuineMitzi NonVictional **fags**

offensive_language frequency:
2 per 25,000 terms
1 per 1,000 docs
Some of the 9 mentions:

"TheGamerKingKev I wonder if the girl that gave head to 24 boys for a drink in magaluf has twitter" she probly sends nudes for **fags**

HeilSidious dyldobaggenss whatta you **fags** doin?

MIKEB IN THE D rams are for **fags**

most gays are not **fags.**

ZaneIsClasker Photoshop putting my head on **fags** is not proof.

We're **fags.**

josh helton Xbox live **fags**

Facebook is for **fags**

GOOD Music >>>>> Young Money **fags**

Tweet Misclassifications

```
# 3 votes for offensive but not hate speech
# although that is considered hate speech to lgbtq community

hate_10k['tweet'].iloc[86]
```

✓ 0.3s

```
'"BrokenPiecesmsc ItsNotAdam faggot read my tweets after dat k" it wasn\'t even funny lol'
```

```
# Three people voted
# 2 votes for offensive language, 1 vote for neither
#This clearly is hate speech and has been misclassified

hate_10k['tweet'].iloc[3]
```

✓ 0.3s

```
' RT C_G_Anderson viva_based she look like a tranny'
```

Tweet Close Call Classifications

```
# 5 voted hate, # 4 voted neither
# Classified as hate speech
hate_10k['tweet'].iloc[5749]
```

✓ 0.1s

```
'clinchmtn316 sixonesixband AMERICA today, the rule of thumb is when in doubt, blame "whitey"'
```

```
# Classified as hate speech
# 7 voted hate, 1 voted offensive, 1 voted neither
hate_10k['tweet'].iloc[6171]
```

✓ 0.3s

```
'infidelpamelaLC I\'m going to blame the black man, since they always blame "whitey" I\'m an equal opportunity hater.'
```

```
# Classified as hate speech
# 6 hate, 3 offensive
hate_10k['tweet'].iloc[6378]
```

✓ 0.3s

```
"kcSnowWhite7 SamSaunders42 don't forget nappy headed, butt ugly bitch, who's married to a Muslim"
```

SpaCy Matcher

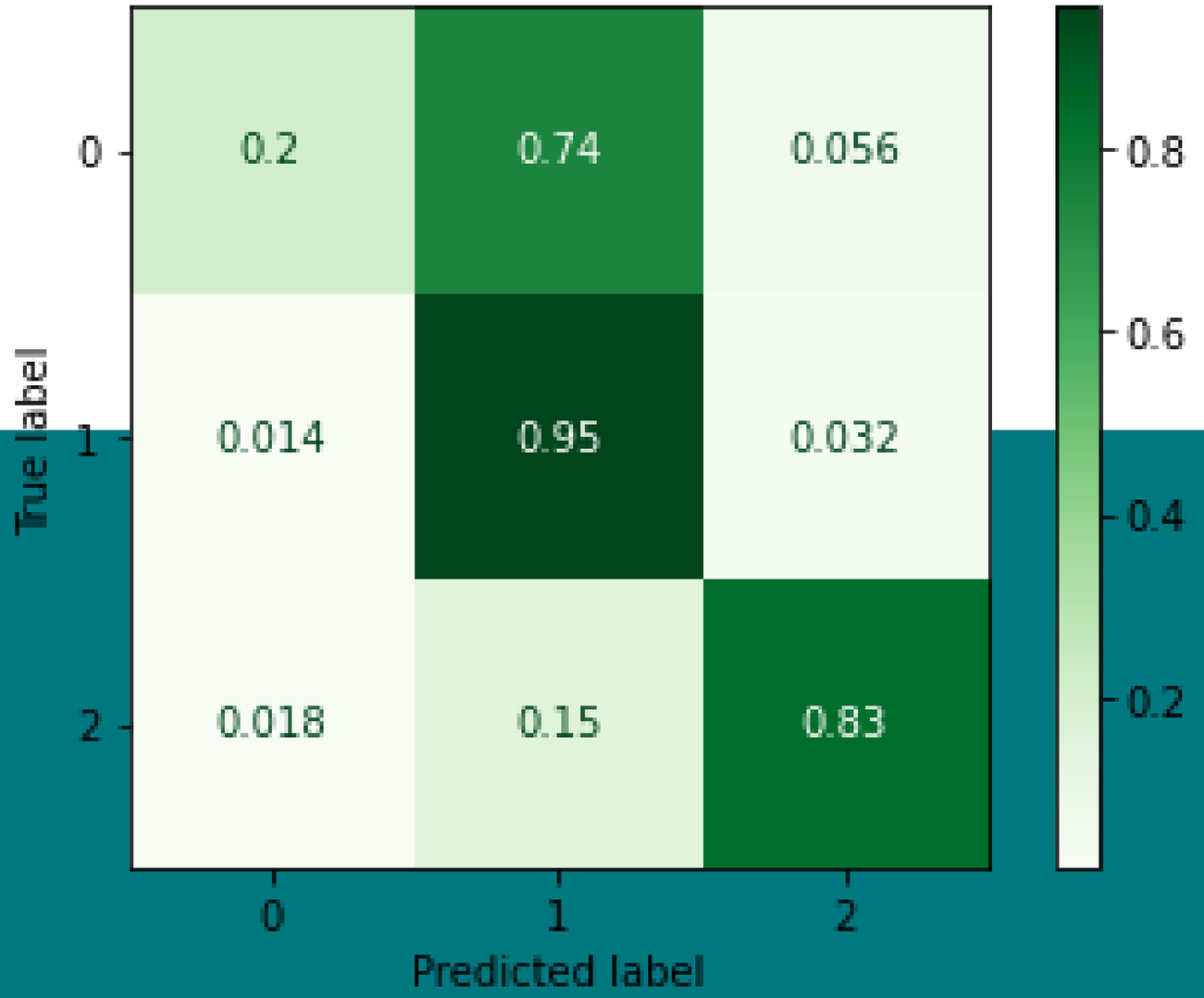
```
606
(96, 250, 260) Murda Gang bitch its Gang Land "', '
(96, 562, 572) Murder Game pussy nigga shut up "', '
(96, 928, 939) Pill Chamberlain these bitches love my music "', '
(96, 944, 960) Teanna Trump probably cleaner than most of these twitter hoes but........."', '
(96, 1282, 1294) emoji doe?"y he say she looked like scream lmao', '
(96, 1407, 1413) Damn Skippy lol', '
(96, 1543, 1551) niggas act like bitches..', '
(96, 3220, 3238) KelsieBelsi You\'re not a man if you refer to every girl as a bitch"', '
(96, 3392, 3407) KissMySmilee Don\'t got time for bitches to be actin iffy."', '
(96, 4006, 4012) south america bitch', '
(96, 4022, 4031) toto santi is like nasty pussy', '
(96, 4629, 4640) dis bitch in da wrong tree bro.', '
(96, 5087, 5101) SmeegBaby Dese hoes be LYIN to all of us nigga"', '
(96, 6004, 6018) marackaf aye b you pullin hoes just like o taught you', '
(96, 6160, 6167) ya homeboy fucc nicca', '
(96, 8834, 8841) JUDYANN\'S SO PRETTY', '
```

CountVect Logistic Regression Modeling

Models	Accuracy	*Bal.Acc	Precision	Recall	f1
Null Model	75%	33%	-	-	-
CtVect w/LogReg	87%	66%	-	-	-
0	-	-	53%	20%	29%
1	-	-	90%	95%	92%
2	-	-	84%	83%	83%

0 = Hate Speech, 1= Offensive Language, 2 = Neither

CountVectorizer with Logistic Regression



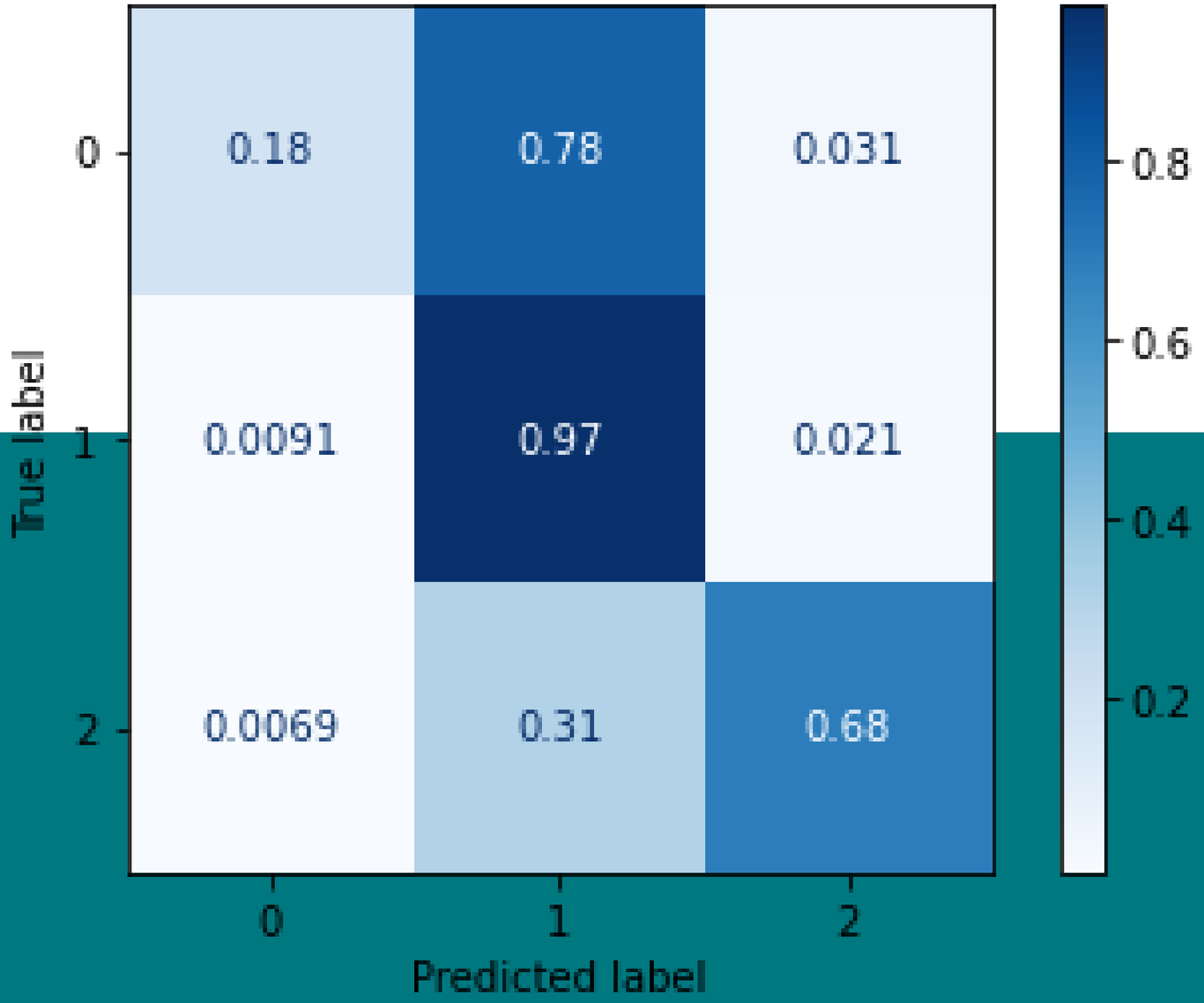
0 = Hate Speech, 1= Offensive Language, 2 = Neither

Tfidf with Logistic Regression

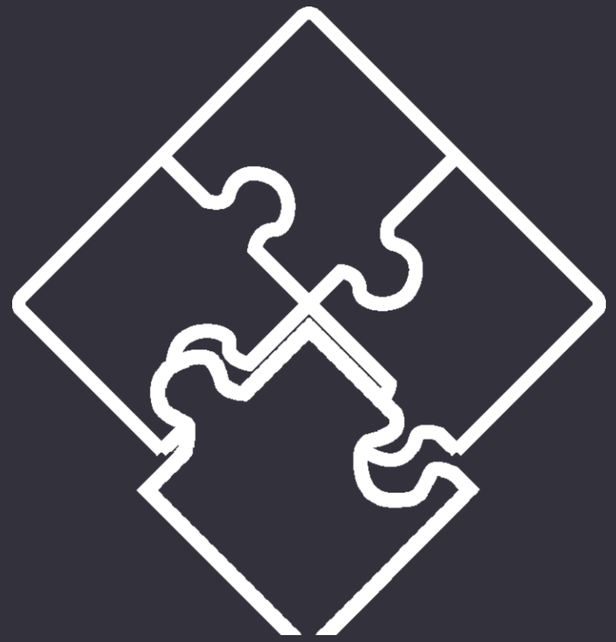
Models	Accuracy	*Bal.Acc	Precision	Recall	f1
Null Model	75%	33%	-	-	-
Tfidf w/LogReg	86%	61%	-	-	-
0	-	-	64%	18%	29%
1	-	-	86%	97%	91%
2	-	-	7%	68%	76%

0 = Hate Speech, 1= Offensive Language, 2 = Neither

Tfidf with Logistic Regression



0 = Hate Speech, 1= Offensive Language, 2 = Neither



Solutions

01 Train models on marginalized communities vernacular & vocabulary

02 Increase diverisity hires and increase D & I seminars for employees

03 Create stricter rules and greater consequences for hate speech.

Key Takeaways

Yes, the algorithm misclassifies, and it is harmful due to internal bias.

Be aware of your own unconscious biases in the code and models you create.

Next Steps

01 Rebalance Data

02 Create a convolutional neural network (CNN) model

03 Compare biases of race & gender

Thank You

Any Questions?