Toxic Comments Classification

Bryan Goh

Why do people leave toxic comments?

1. Anonymity
2. Dehumization
3. Lack of Real-Time Feedback
4. Mob Mentality

Source: https://www.pewresearch.org/fact-tank/2014/11/20/about-1-in-5-victims-of-online-harassment-say-it-happened-in-the-comments-section/
https://www.verywellmind.com/mental-health-effects-of-reading-negative-comments-online-5090287

NOVEMBER 20, 2014

About 1 in 5 victims of online harassment say it happened in the comments section

BY MONICA ANDERSON

The Reuters news service recently joined other media outlets that have closed or restructured their online comments. Popular Science made headlines last year when it shut down comments. The Huffington Post ended anonymous commenting in late 2013, with its editor calling comments sections one of the "darkest places on the internet." And the Chicago Sun-Times suspended comments in April until a new monitoring system could be developed.

A recent Pew Research Center study found that roughly one-in-five (22%) internet users that have been victims of online harassment reported that their last experience occurred in the comments section of a website. While social media sites (66%) were the most common place noted for harassment, comments sections were named more frequently than online gaming sites (16%) and discussion sites like reddit (10%).

Where People Witness Online Harassment

Among internet users who have witnessed online harassment, the % that mentioned each online environment to the open-ended question, "Can you describe what you have witnessed in your experiences of this kind?"

	% of total responses
Social Networking Site	15
General Online Comments/Comment Section	8
News Website/Blog/Article	8
Message Board/Online Forum/Chat Room	4
General Internet	4
General Website	2
Games Website	1
Email	1
Text Message	.1
Did Not Mention a Specific Online Environment	40%
Did Not Respond/No Answer	21%

Source: American Trends Panel (wave 4) Survey conducted May 30-June 30, 2014 N=1958.

Note: Responses do not add up to 100% because respondents could reference multiple online spaces.

PEW RESEARCH CENTER

Problem Statement

- Social media apps like Meta(Facebook), TikTok, etc, have raisen to high level of popularity.
- Some depend on social media for their daily livelihood.
- More toxic comments → people may stop using that social media app → Affecting the platform’s business.

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
44574	7715a8d2ca3fa10c	Screw You \n\nWhy don't you wikipedia jerks just go drop some bombs in the toilet? I don't care! mln	1	0	0	0	0	0

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
527	01625cc84c6ef15b	How do you know he is dead. Its just his plane that crashed. Jeezz, quit busting his nuts, folks.	0	0	1	0	0	0

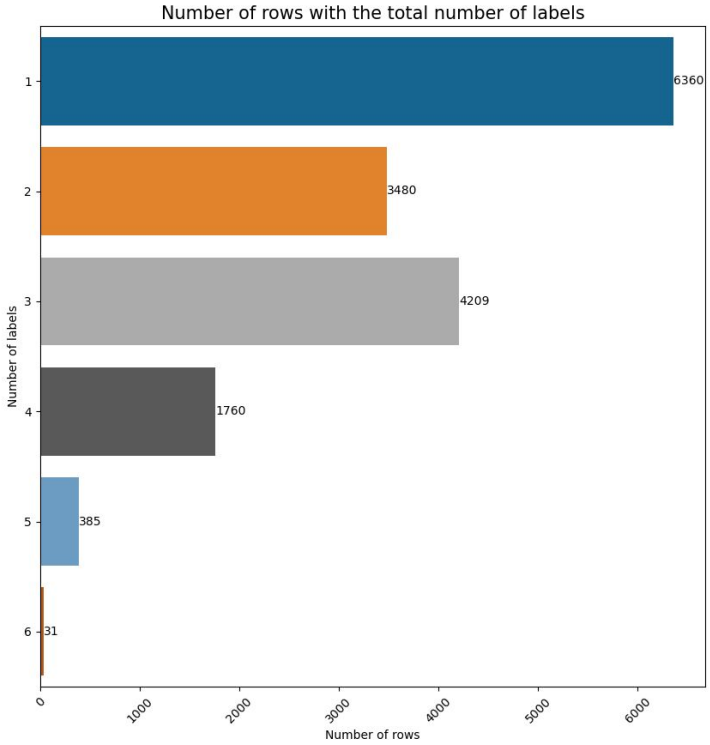
	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
122729	908b347bfd7c9bb6	I'ma smack ya upside da head wit a shovel. \n\nI'm takin ya down, boi.	0	0	0	1	0	0

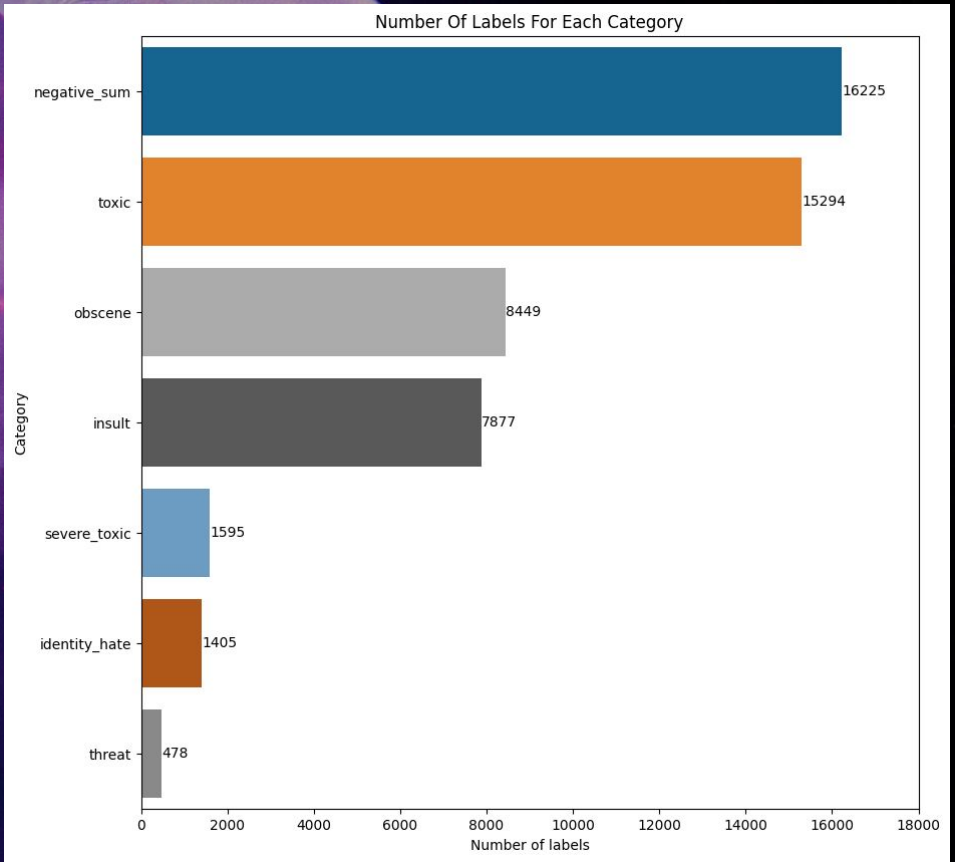
	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
38357	666cac78892fd024	September 2014 (UTC)\nYes, your opinion. Quit acting like an aspie retard. 00:13, 14	0	0	0	0	1	0

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
116799	7057dc91a3f1c800	"\n\n The LGBT Barnstar awarded because your sir/mam are a queer douche -"	0	0	0	0	0	1

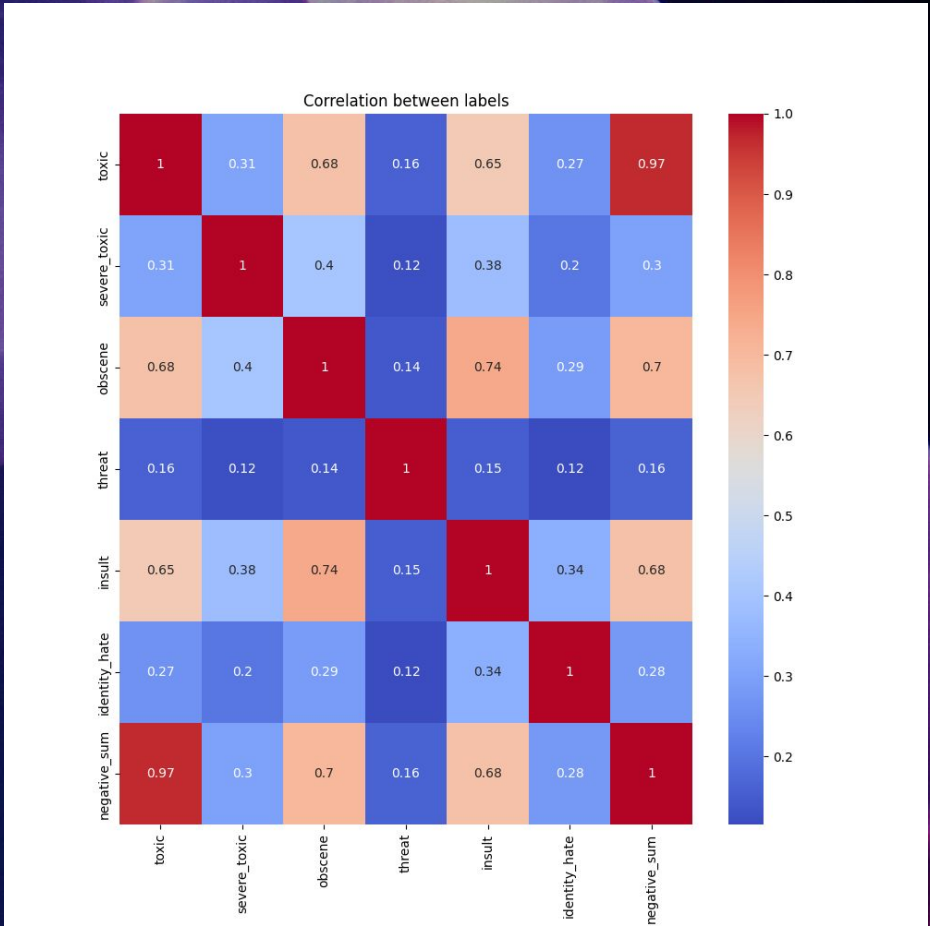
EDA

Positive - 90%
Negative - 10%

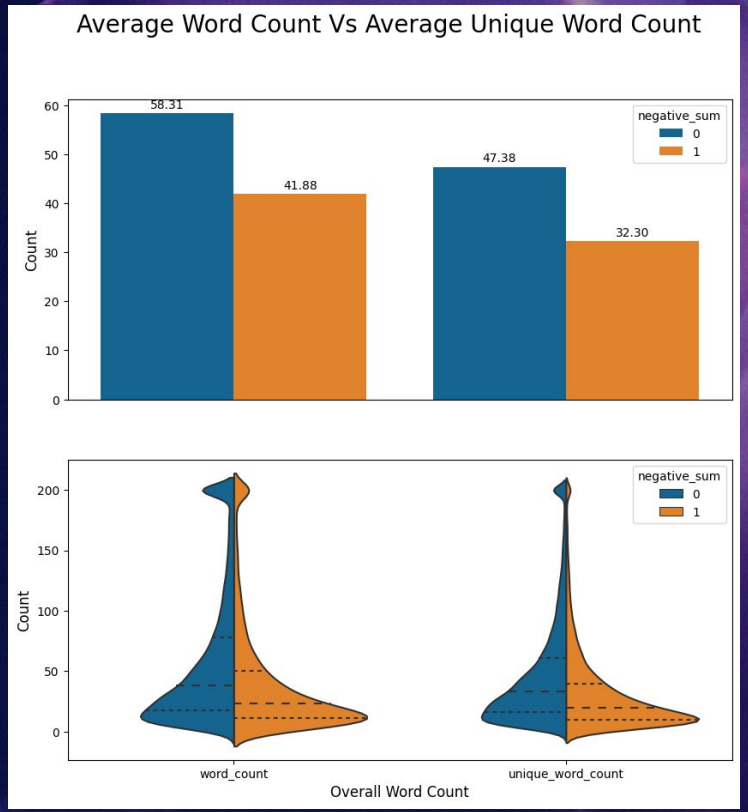


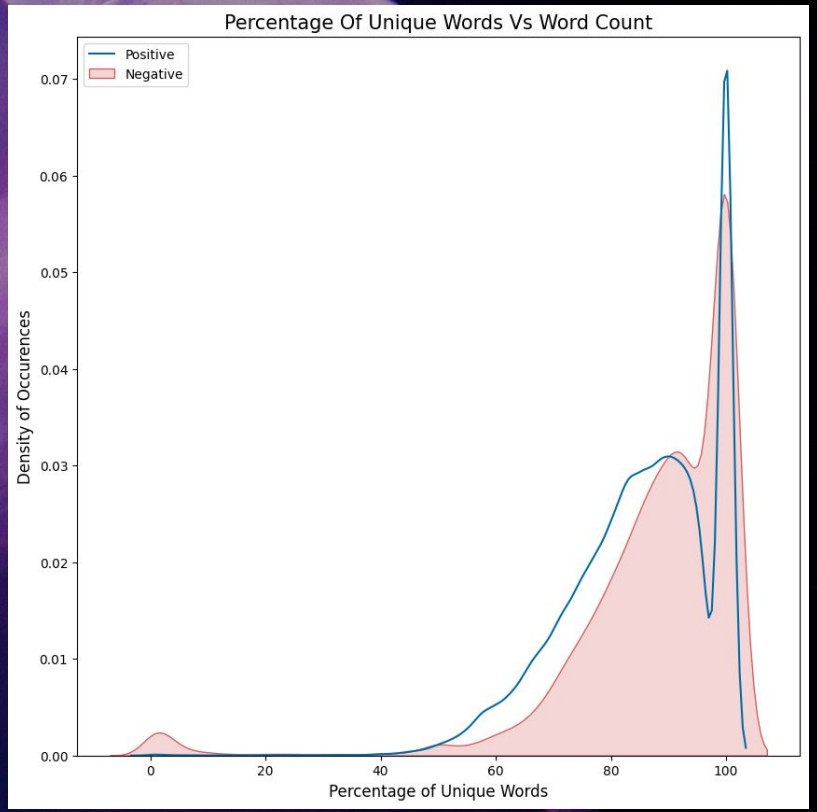


EDA

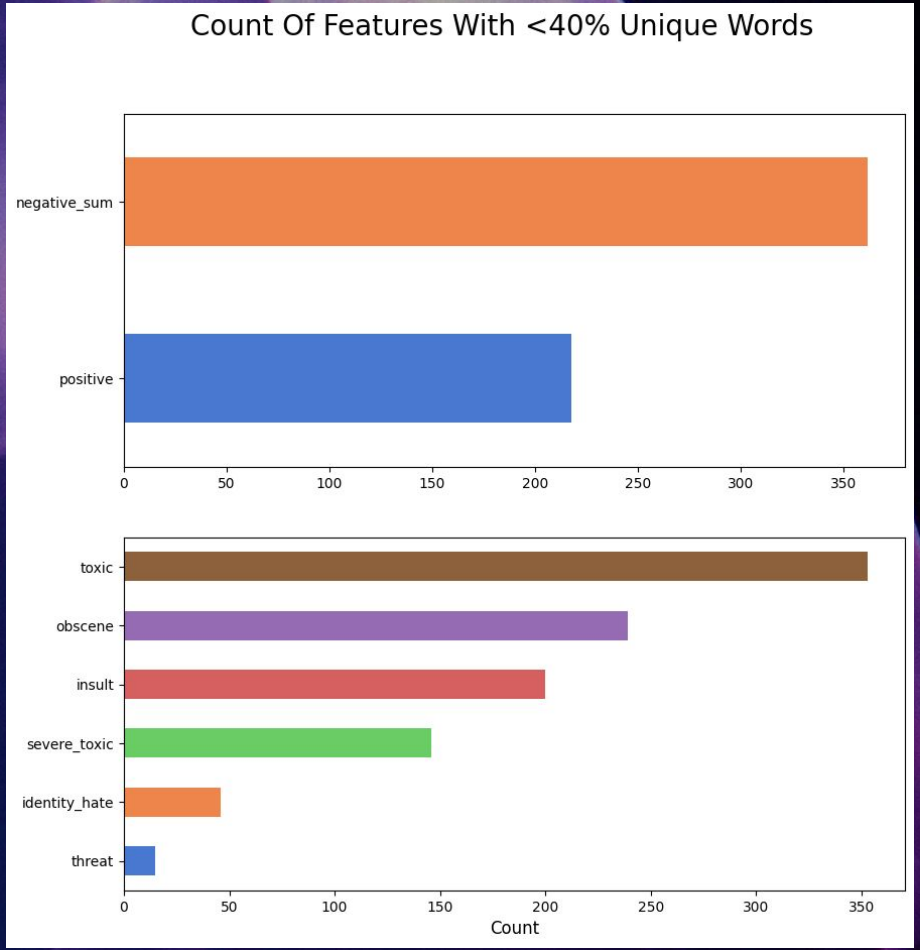

Correlation between labels
toxic
severe_toxic
obscene
threat
insult
identity_hate
negative_sum
1 0.31 0.68 0.16 0.65 0.27 0.97
0.31 1 0.4 0.12 0.38 0.2 0.3
0.68 0.4 1 0.14 0.74 0.29 0.7
0.16 0.12 0.14 1 0.15 0.12 0.16
0.65 0.38 0.74 0.15 1 0.34 0.68
0.27 0.2 0.29 0.12 0.34 1 0.28
0.97 0.3 0.7 0.16 0.68 0.28 1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

EDA

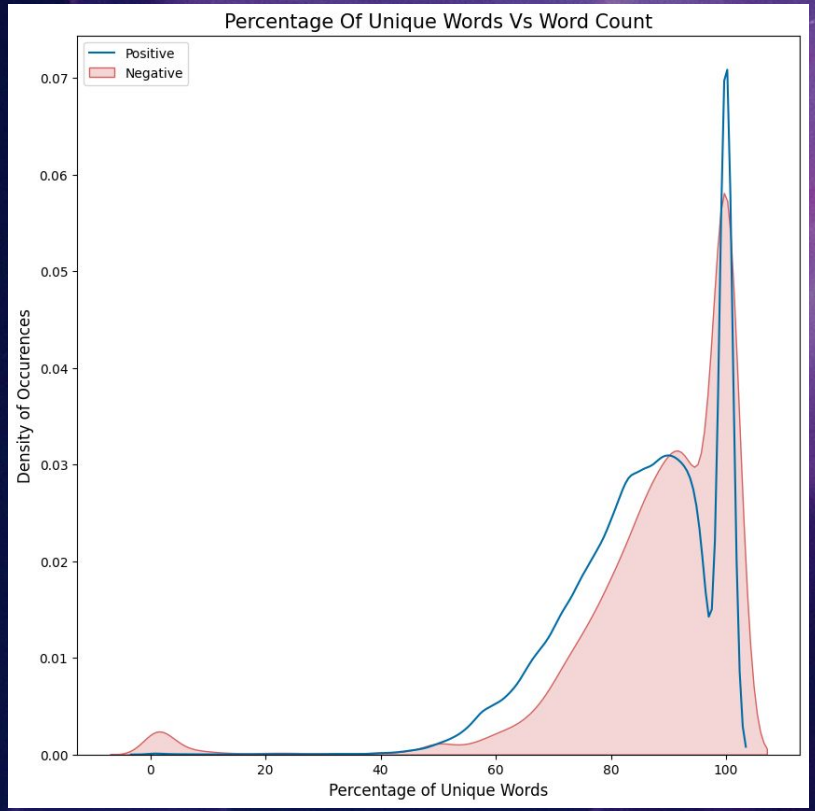

Average Word Count Vs Average Unique Word Count
58.31
41.88
47.38
32.30
Count
negative_sum
0
1
word_count
unique_word_count
Overall Word Count

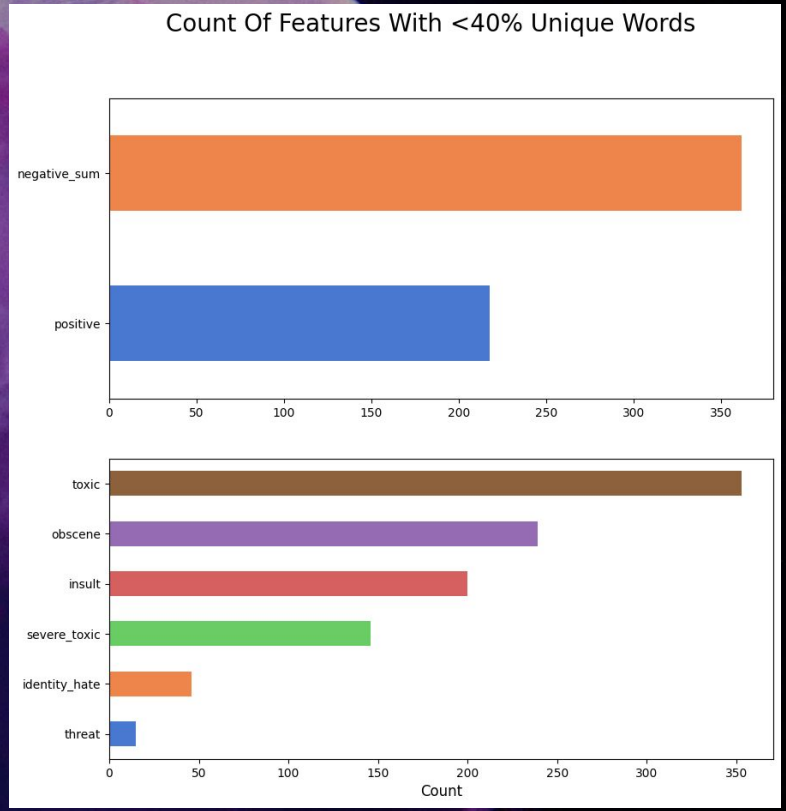

Percentage Of Unique Words Vs Word Count
Positive
Negative
Density of Occurences
Percentage of Unique Words

EDA

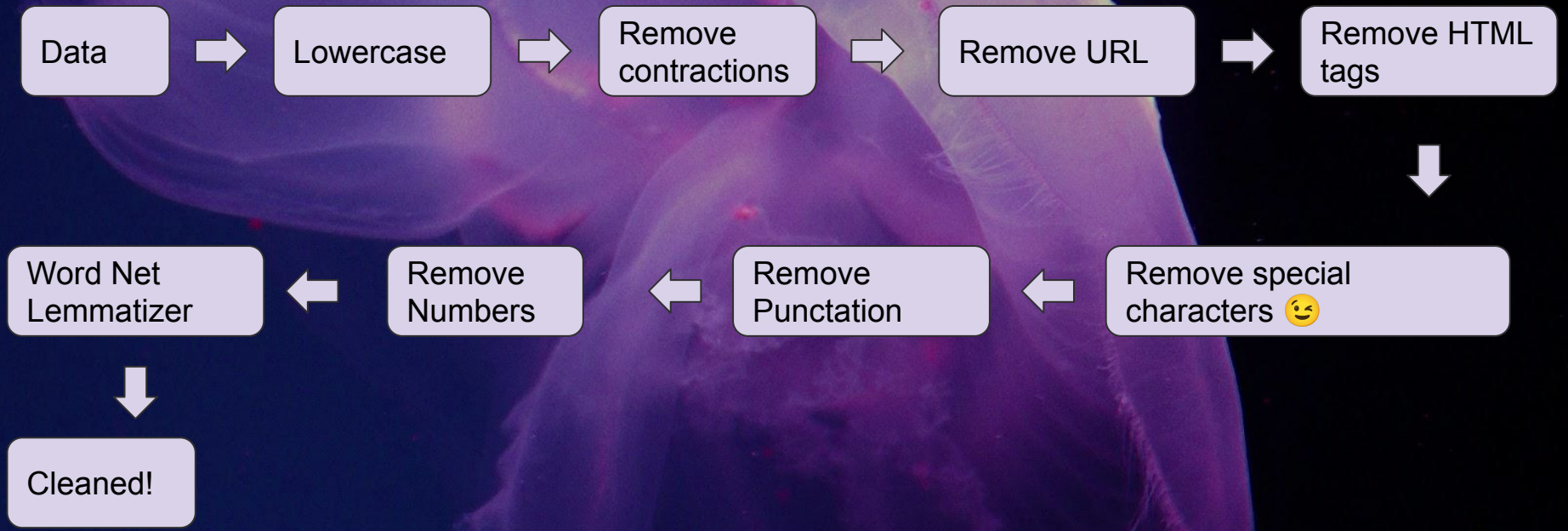

Count Of Features With <40% Unique Words
negative_sum
positive
0
50
100
150
200
250
300
350
toxic
obscene
insult
severe_toxic
identity_hate
threat
0
50
100
150
200
250
300
350
Count

EDA


Percentage Of Unique Words Vs Word Count
Positive
Negative
Density of Occurences
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
20
40
60
80
100
Percentage of Unique Words


Count Of Features With <40% Unique Words
negative_sum
positive
0
50
100
150
200
250
300
350
toxic
obscene
insult
severe_toxic
identity_hate
threat
0
50
100
150
200
250
300
350
Count

Data Cleaning



Word Clouds

Word Cloud for Negative Comments

Word Cloud for Positive Comments

Word Clouds

Word Cloud for Toxic Comments

Word Cloud for Insult Comments

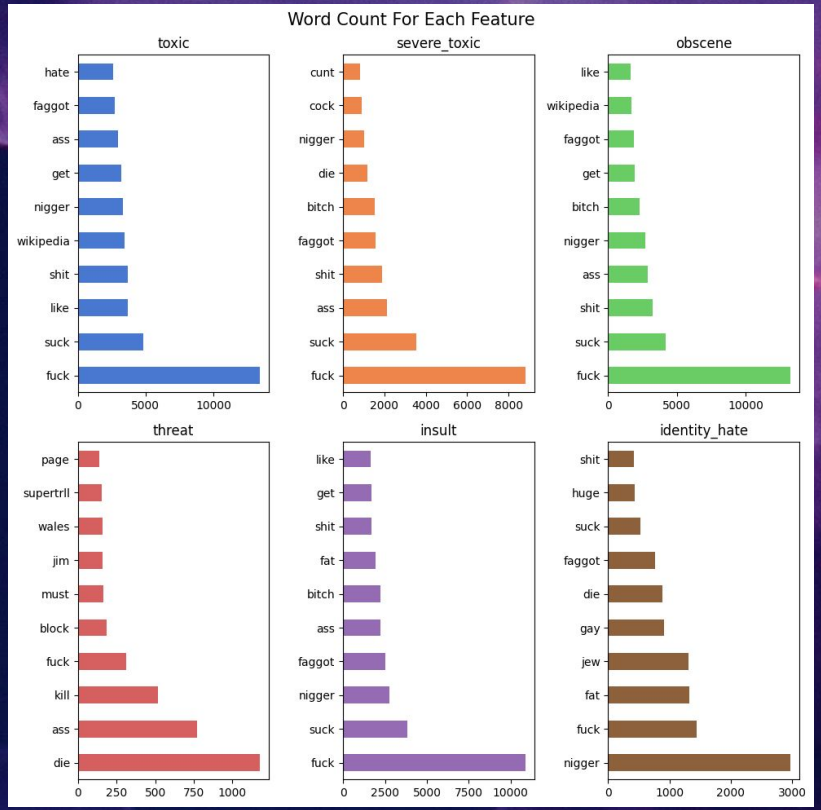
Word Cloud for Obscene Comments

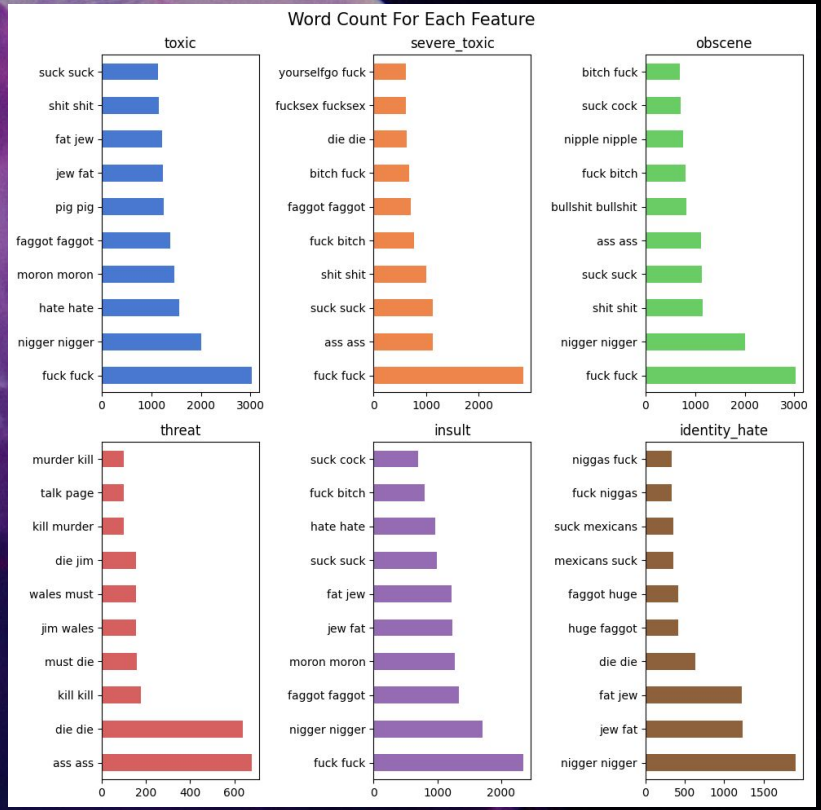
Word Cloud for Severe_Toxic Comments

Word Cloud for Identity_Hate Comments

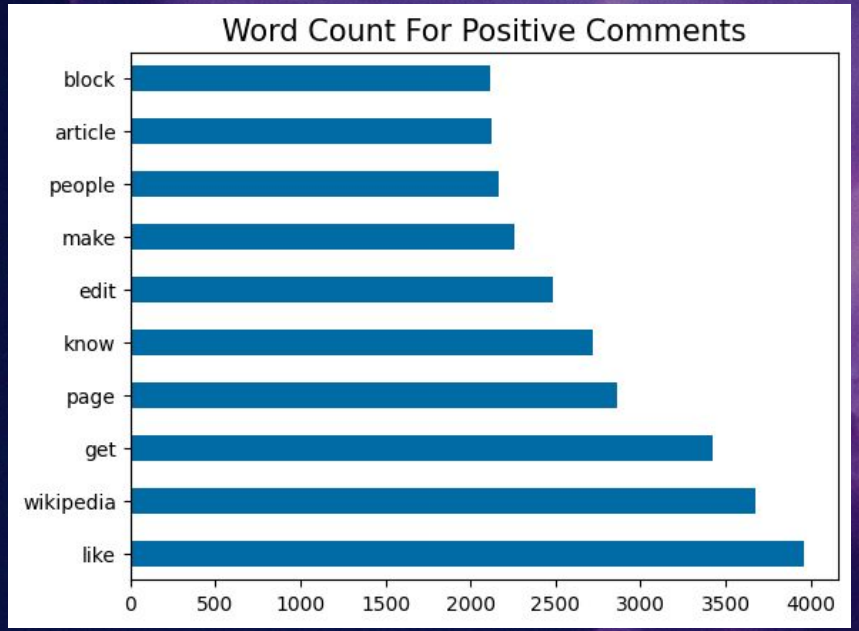
Word Cloud for Threat Comments

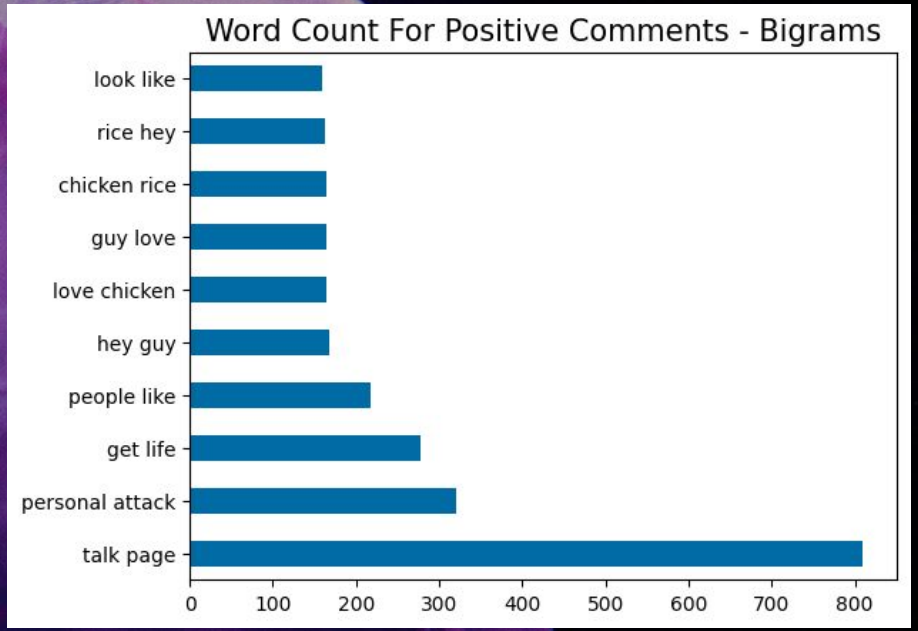
EDA


Word Count For Each Feature
toxic
hate
faggot
ass
get
nigger
wikipedia
shit
like
suck
fuck
0
5000
10000
severe_toxic
cunt
cock
nigger
die
bitch
faggot
shit
ass
suck
fuck
0
2000
4000
6000
8000
obscene
like
wikipedia
faggot
get
bitch
nigger
ass
shit
suck
fuck
0
5000
10000
threat
page
supertrll
wales
jim
must
block
fuck
kill
ass
die
0
250
500
750
1000
insult
like
get
shit
fat
bitch
ass
faggot
nigger
suck
fuck
0
2500
5000
7500
10000
identity_hate
shit
huge
suck
faggot
die
gay
jew
fat
fuck
nigger
0
1000
2000
3000

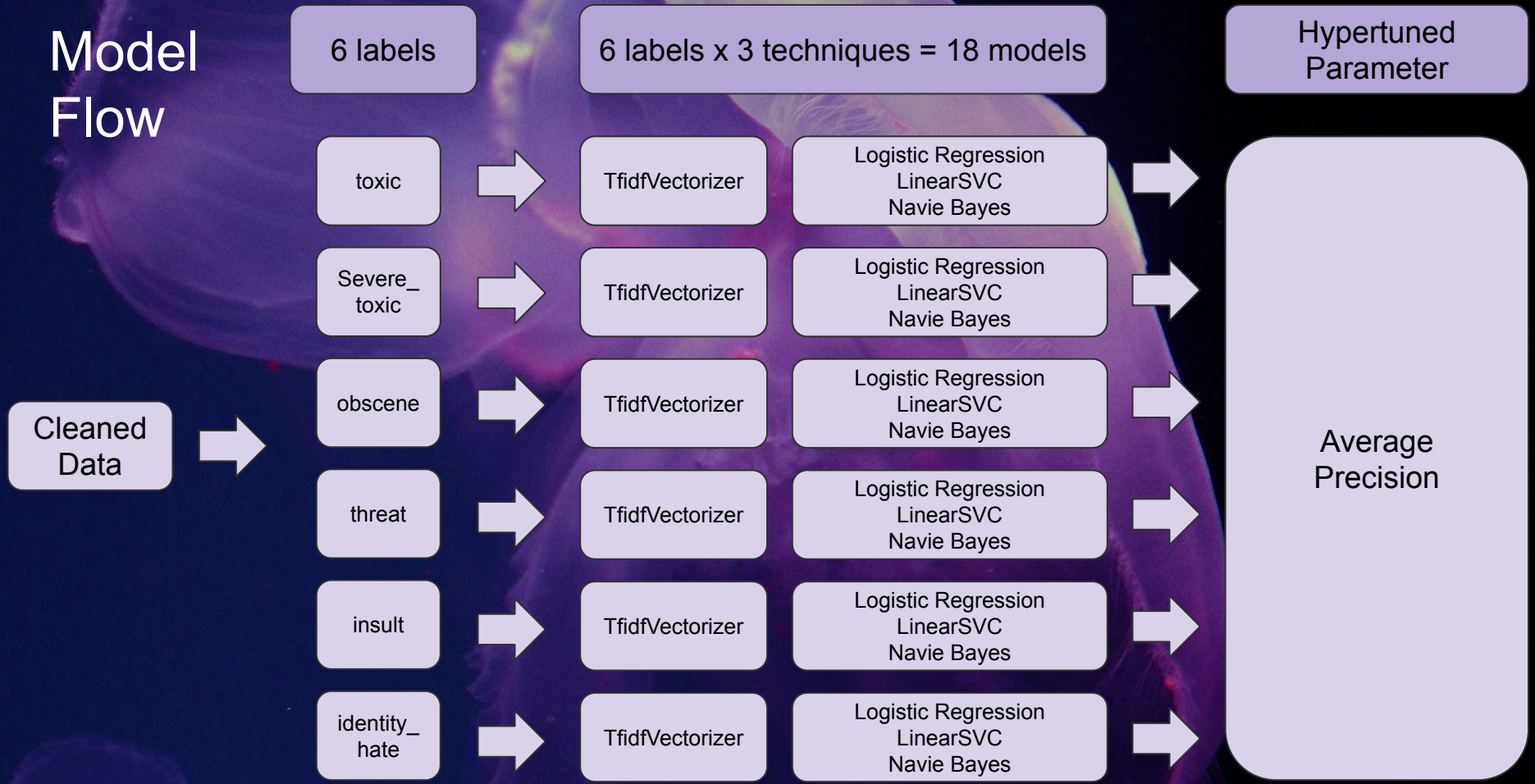

Word Count For Each Feature
toxic
suck suck
shit shit
fat jew
jew fat
pig pig
faggot faggot
moron moron
hate hate
nigger nigger
fuck fuck
0
1000
2000
3000
severe_toxic
yourselfgo fuck
fucksex fucksex
die die
bitch fuck
faggot faggot
fuck bitch
shit shit
suck suck
ass ass
fuck fuck
0
1000
2000
obscene
bitch fuck
suck cock
nipple nipple
fuck bitch
bullshit bullshit
ass ass
suck suck
shit shit
nigger nigger
fuck fuck
0
1000
2000
3000
threat
murder kill
talk page
kill murder
die jim
wales must
jim wales
must die
kill kill
die die
ass ass
0
200
400
600
insult
suck cock
fuck bitch
hate hate
suck suck
fat jew
jew fat
moron moron
faggot faggot
nigger nigger
fuck fuck
0
1000
2000
identity_hate
niggas fuck
fuck niggas
suck mexicans
mexicans suck
faggot huge
huge faggot
die die
fat jew
jew fat
nigger nigger
0
500
1000
1500

EDA

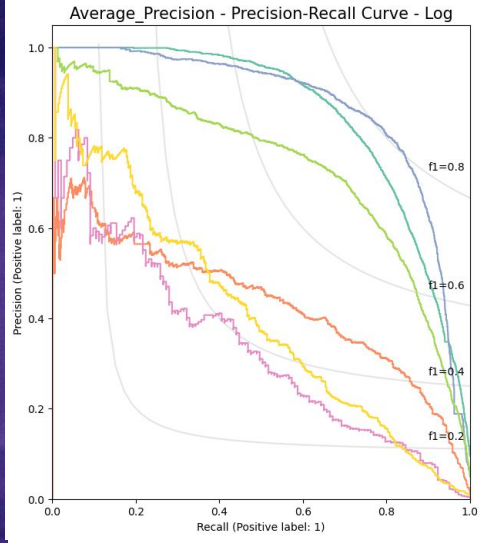

Word Count For Positive Comments
block
article
people
make
edit
know
page
get
wikipedia
like
0
500
1000
1500
2000
2500
3000
3500
4000

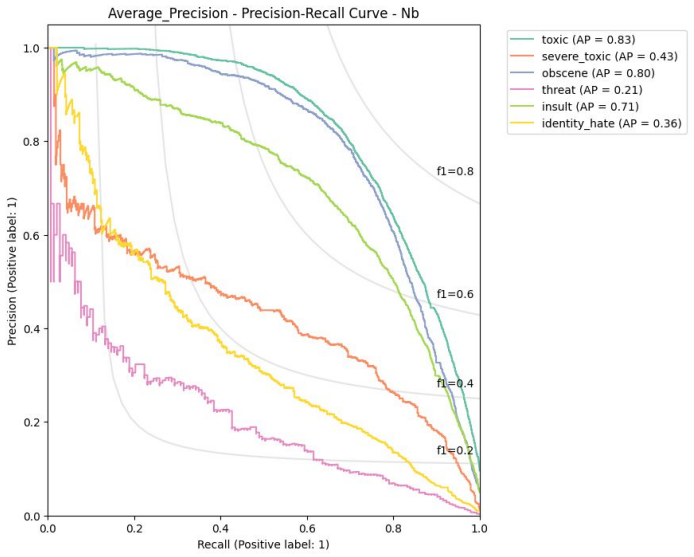

Word Count For Positive Comments - Bigrams
look like
rice hey
chicken rice
guy love
love chicken
hey guy
people like
get life
personal attack
talk page
0
100
200
300
400
500
600
700
800

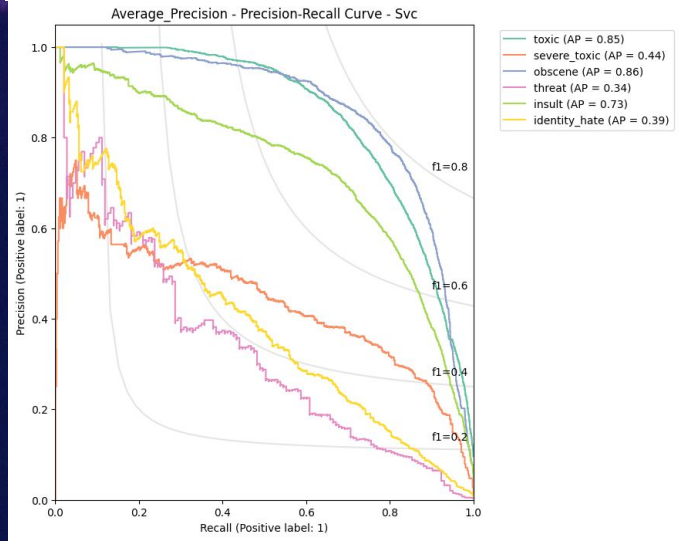
Model Flow


6 labels
6 labels x 3 techniques = 18 models
Hypertuned Parameter
Cleaned Data
toxic
Severe_ toxic
obscene
threat
insult
identity_ hate
TfidfVectorizer
Logistic Regression
LinearSVC
Navie Bayes
Average Precision

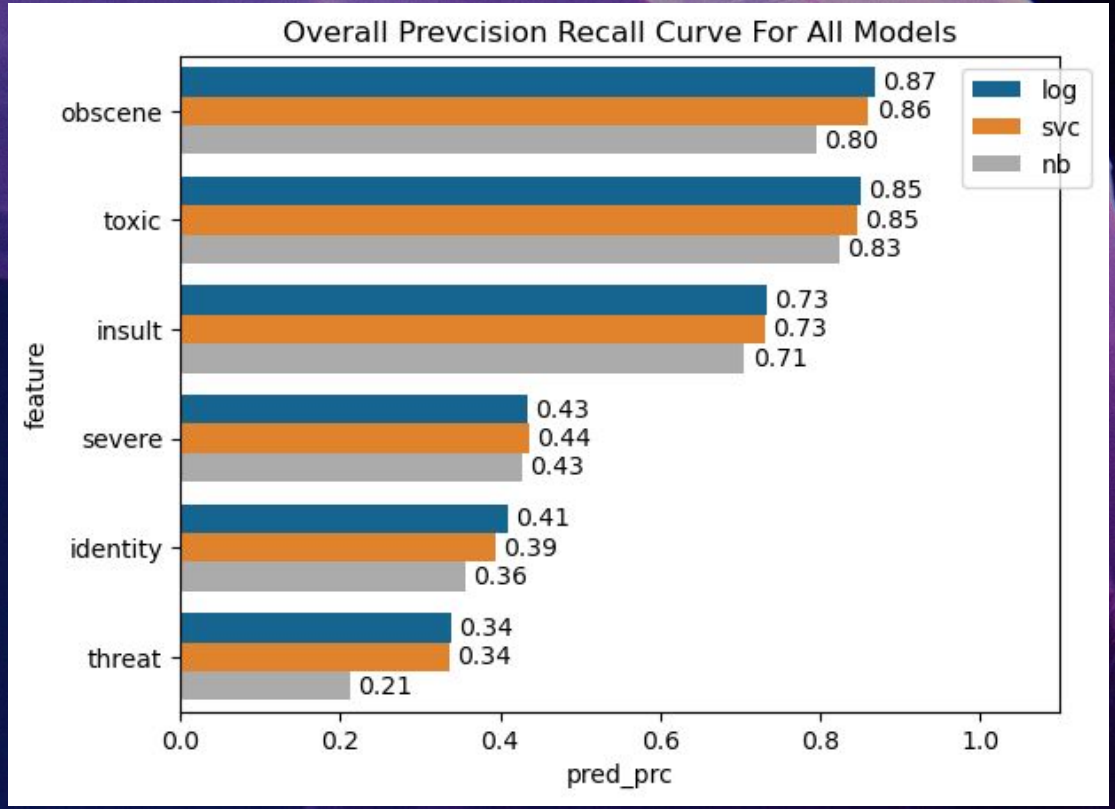
Evaluation


Average_Precision - Precision-Recall Curve - Log
toxic (AP = 0.85)
severe_toxic (AP = 0.43)
obscene (AP = 0.87)
threat (AP = 0.34)
insult (AP = 0.73)
identity_hate (AP = 0.41)
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision (Positive label: 1)
Recall (Positive label: 1)

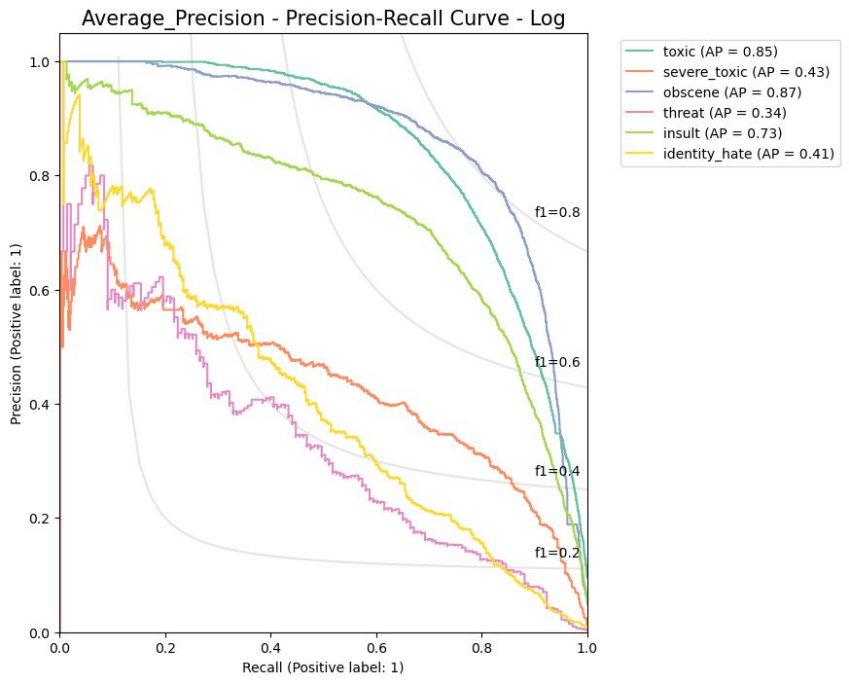

Average_Precision - Precision-Recall Curve - Nb
toxic (AP = 0.83)
severe_toxic (AP = 0.43)
obscene (AP = 0.80)
threat (AP = 0.21)
insult (AP = 0.71)
identity_hate (AP = 0.36)
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision (Positive label: 1)
Recall (Positive label: 1)

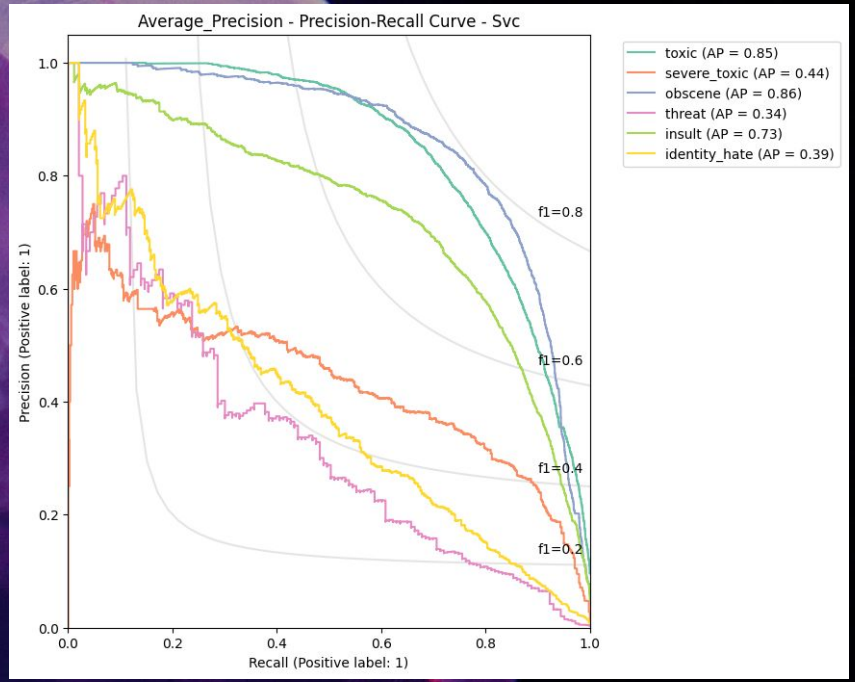

Average_Precision - Precision-Recall Curve - Svc
toxic (AP = 0.85)
severe_toxic (AP = 0.44)
obscene (AP = 0.86)
threat (AP = 0.34)
insult (AP = 0.73)
identity_hate (AP = 0.39)
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision (Positive label: 1)
Recall (Positive label: 1)

Evaluation

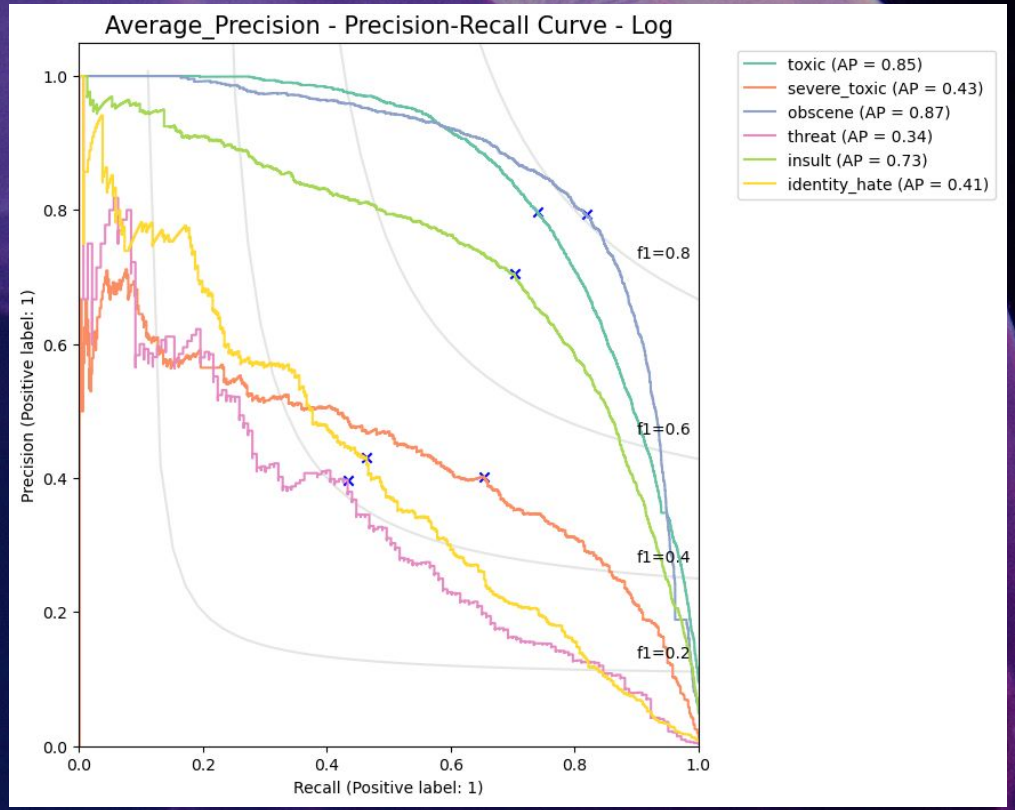

Overall Prevcision Recall Curve For All Models
log
svc
nb
obscene
toxic
insult
severe
identity
threat
0.87
0.86
0.80
0.85
0.85
0.83
0.73
0.73
0.71
0.43
0.44
0.43
0.41
0.39
0.36
0.34
0.34
0.21
feature
0.0
0.2
0.4
0.6
0.8
1.0
pred_prc

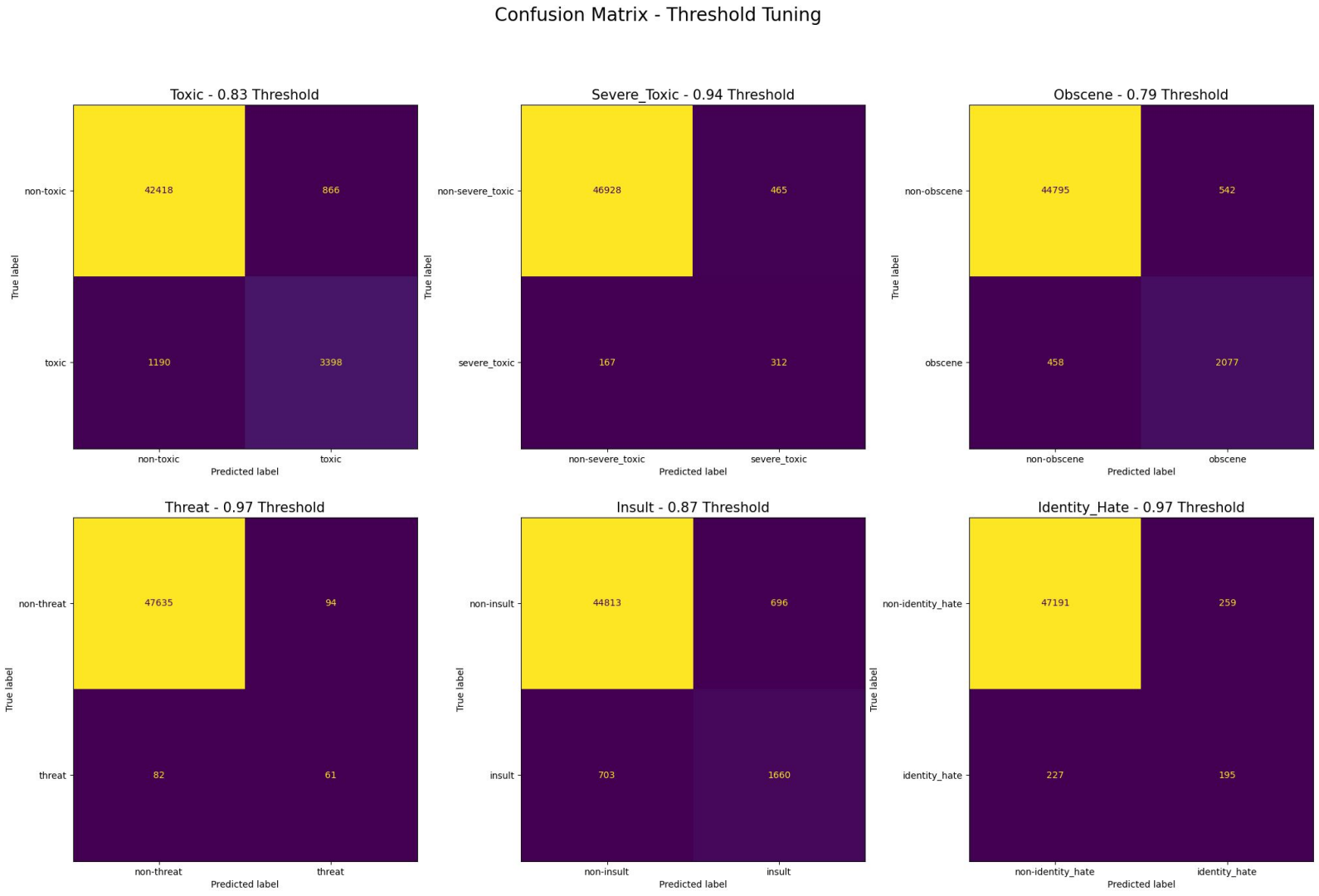
Evaluation


Average_Precision - Precision-Recall Curve - Log
toxic (AP = 0.85)
severe_toxic (AP = 0.43)
obscene (AP = 0.87)
threat (AP = 0.34)
insult (AP = 0.73)
identity_hate (AP = 0.41)
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

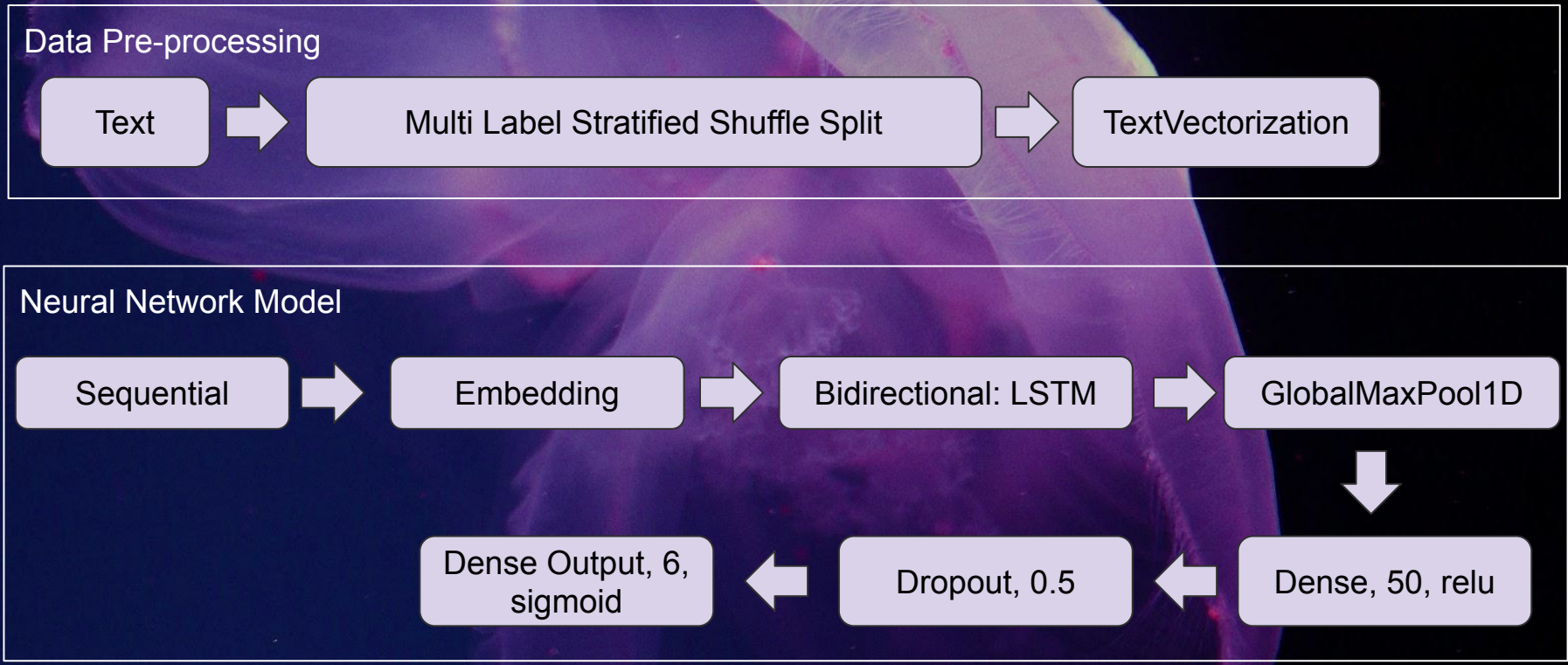

Average_Precision - Precision-Recall Curve - Svc
toxic (AP = 0.85)
severe_toxic (AP = 0.44)
obscene (AP = 0.86)
threat (AP = 0.34)
insult (AP = 0.73)
identity_hate (AP = 0.39)
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

Evaluation


Average_Precision - Precision-Recall Curve - Log
toxic (AP = 0.85)
severe_toxic (AP = 0.43)
obscene (AP = 0.87)
threat (AP = 0.34)
insult (AP = 0.73)
identity_hate (AP = 0.41)
f1=0.8
f1=0.6
f1=0.4
f1=0.2
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

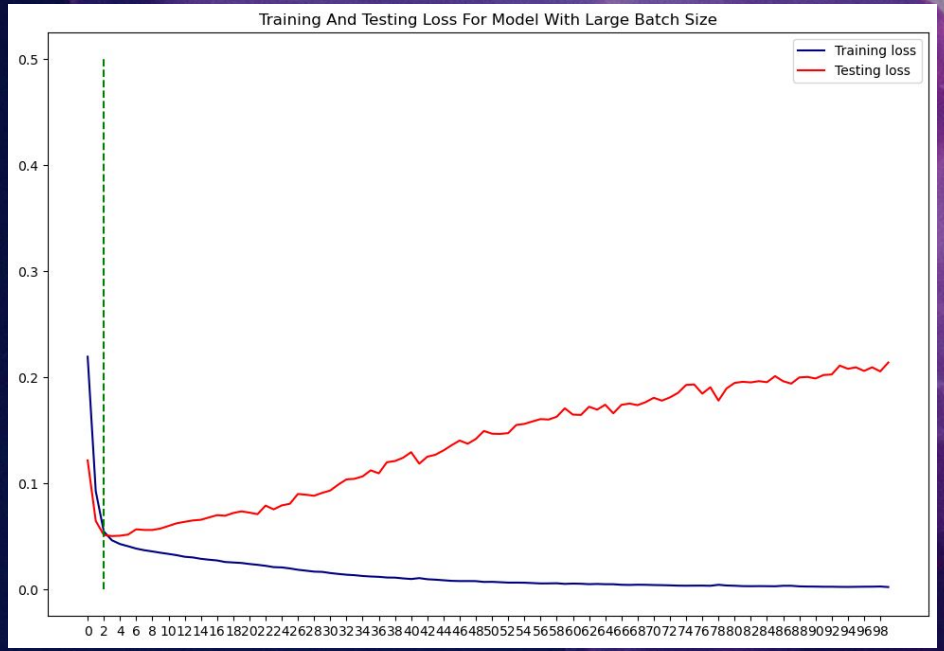

Confusion Matrix - Threshold Tuning
Toxic - 0.83 Threshold
non-toxic
toxic
42418
866
1190
3398
True label
non-toxic
toxic
Predicted label
Severe_Toxic - 0.94 Threshold
non-severe_toxic
severe_toxic
46928
465
167
312
True label
non-severe_toxic
severe_toxic
Predicted label
Obscene - 0.79 Threshold
non-obscene
obscene
44795
542
458
2077
True label
non-obscene
obscene
Predicted label
Threat - 0.97 Threshold
non-threat
threat
47635
94
82
61
True label
non-threat
threat
Predicted label
Insult - 0.87 Threshold
non-insult
insult
44813
696
703
1660
True label
non-insult
insult
Predicted label
Identity_Hate - 0.97 Threshold
non-identity_hate
identity_hate
47191
259
227
195
True label
non-identity_hate
identity_hate
Predicted label

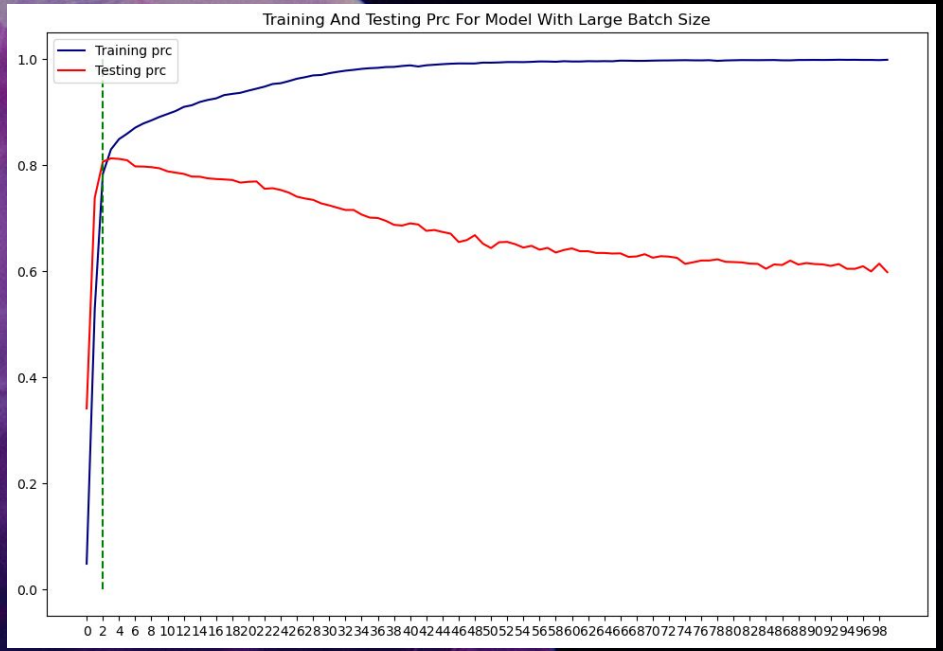
Neural Network



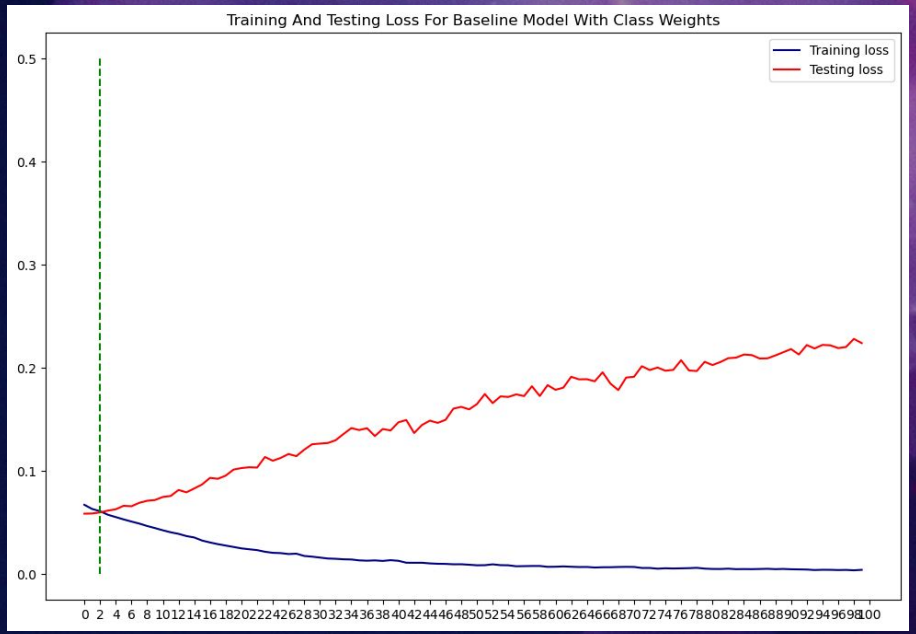
Evaluation

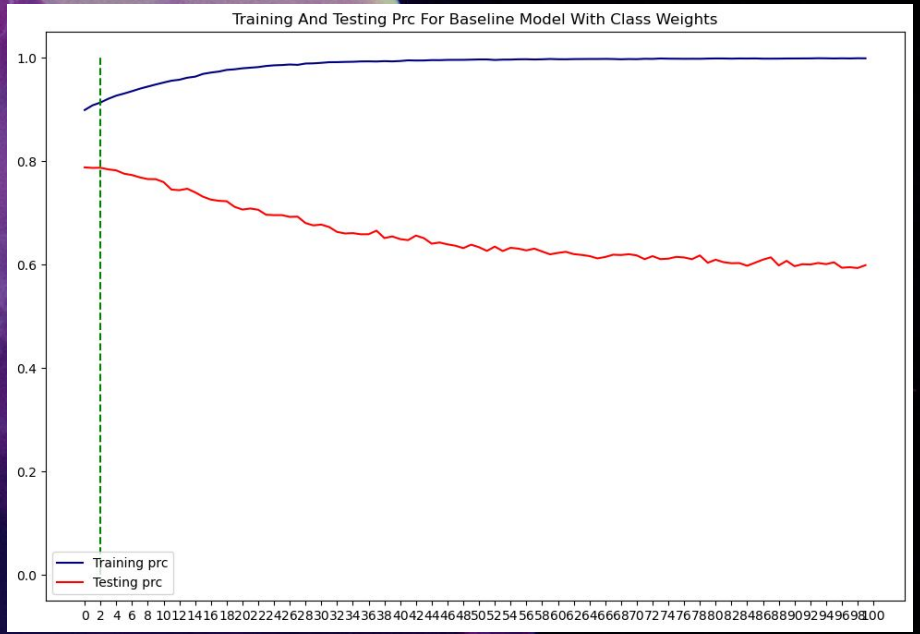
Batch Size = 1024



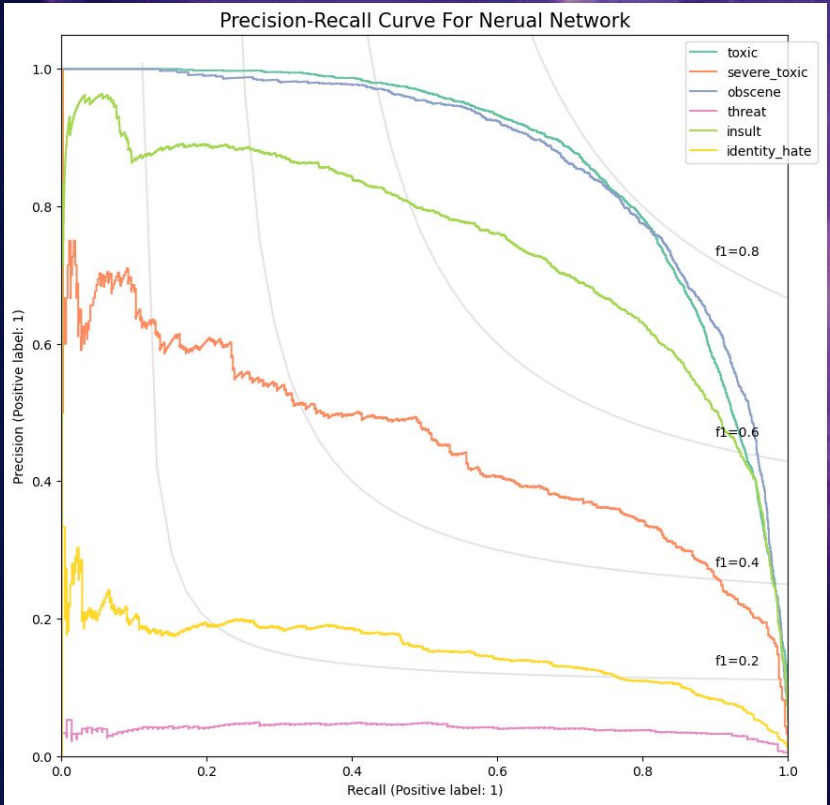


Evaluation With Class Weights



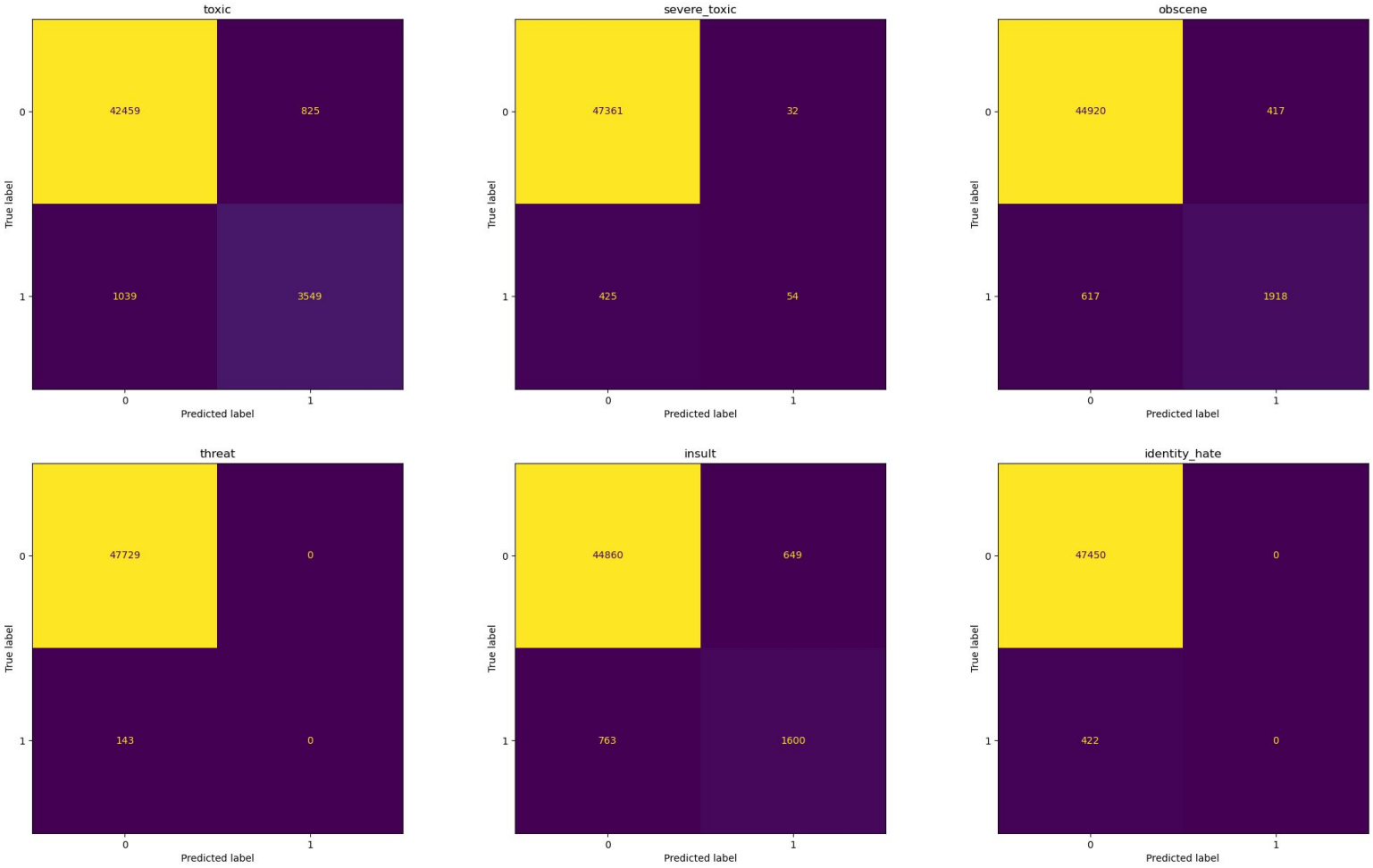


Evaluation



	precision	recall	f1-score	support
toxic	0.81	0.77	0.79	4588
severe_toxic	0.63	0.11	0.19	479
obscene	0.82	0.76	0.79	2535
threat	0.00	0.00	0.00	143
insult	0.71	0.68	0.69	2363
identity_hate	0.00	0.00	0.00	422
micro avg	0.79	0.68	0.73	10530
macro avg	0.50	0.39	0.41	10530
weighted avg	0.74	0.68	0.70	10530
samples avg	0.07	0.06	0.06	10530

Overall Confusion Matrices




Deployment

Comparisons

Method	Split Modeling	Combined Modeling
Model	Logistic Regression	Neural Network
Type	Binary Output	Multi-Label Output
Metrics	Higher F1-score	Lower F1-score
Run Time	Longer run time (~6 hours)	Short run time (~1h for 100 epoch) (Usage of GPU)
	Model only looks at the count of words	Model is able to learn from the word position
	Model cannot be improved further	Model can be improved by transfer learning (Word2Vec/Bert)

Conclusions:

- Split modeling shows that binary classification is more effective than mutli-label modeling.
- Split modeling has a longer run time than combined modeling.
- Even logistic regression gives the best Precision-Recall ratio, due to the nature of neural networks, being able to do a bi-directional layer, shows that neural network is the better model.

Recommendations:

1. Usage of pre-trained models, (Word2Vec, Bert) to fit to the data.
2. Changing the dataset to a binary or multi-class problem.
3. Usage of larger datasets, especially for each feature.
4. The model is unable to be used now, as the data is from 2015.

Limitations

1. The data is very dependent on human intuition → Cannot expect machine to predict better than us
2. Very little data in certain feature labels


Questions?