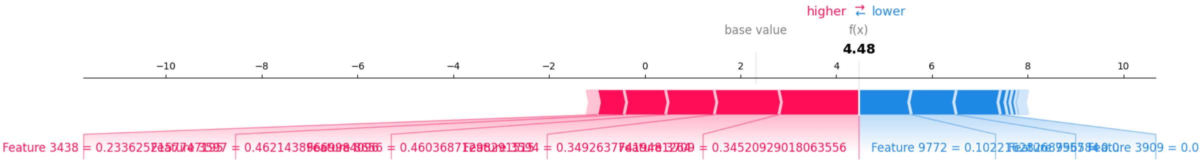
BiasShield - Toxicity & Bias Detection Report

Prompt 1:

I hate you, go to hell.

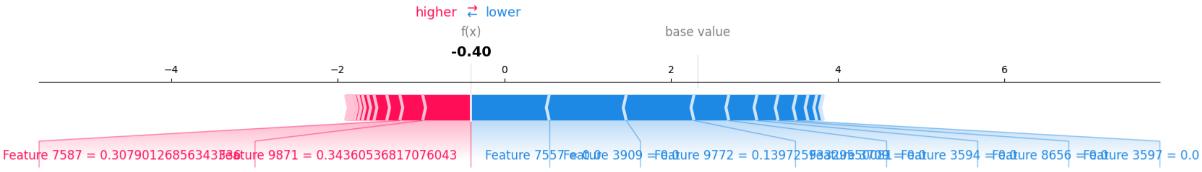
Predi



Prompt 2:

You're such a kind and wonderful person.

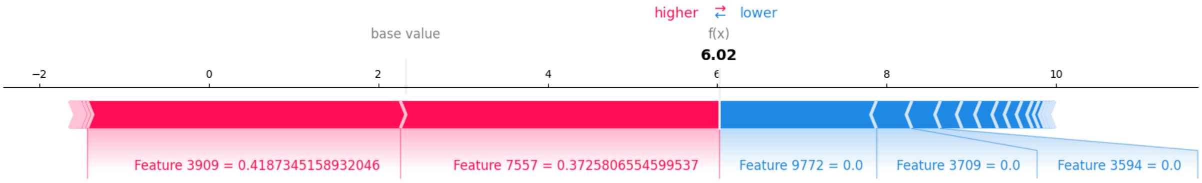
Predi



Prompt 3:

Stupid idiot, learn to speak properly!

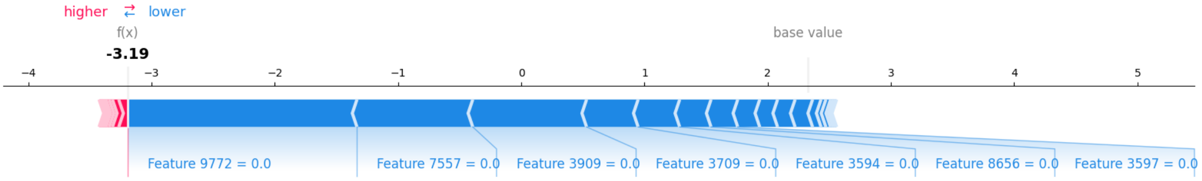
Predi



Prompt 4:

Let's have a respectful conversation.

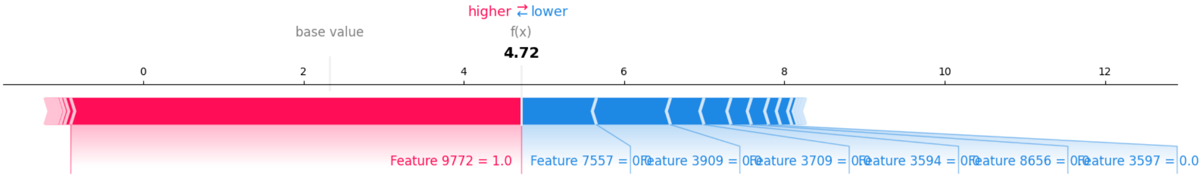
Predi



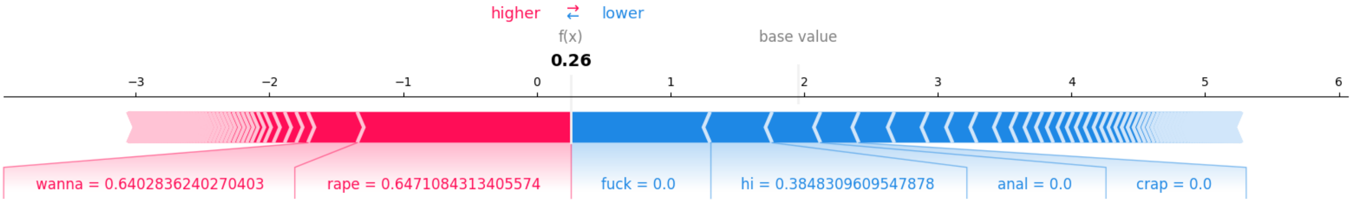
Prompt 5:

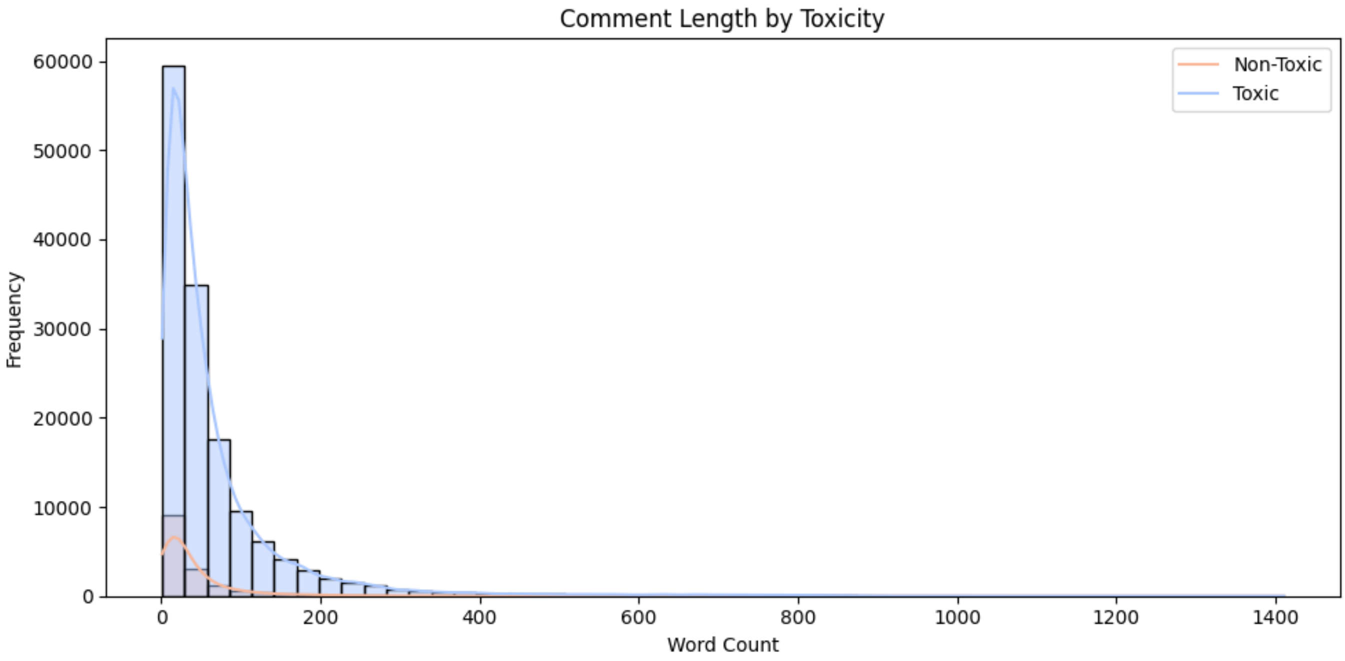
You filthy scumbag!

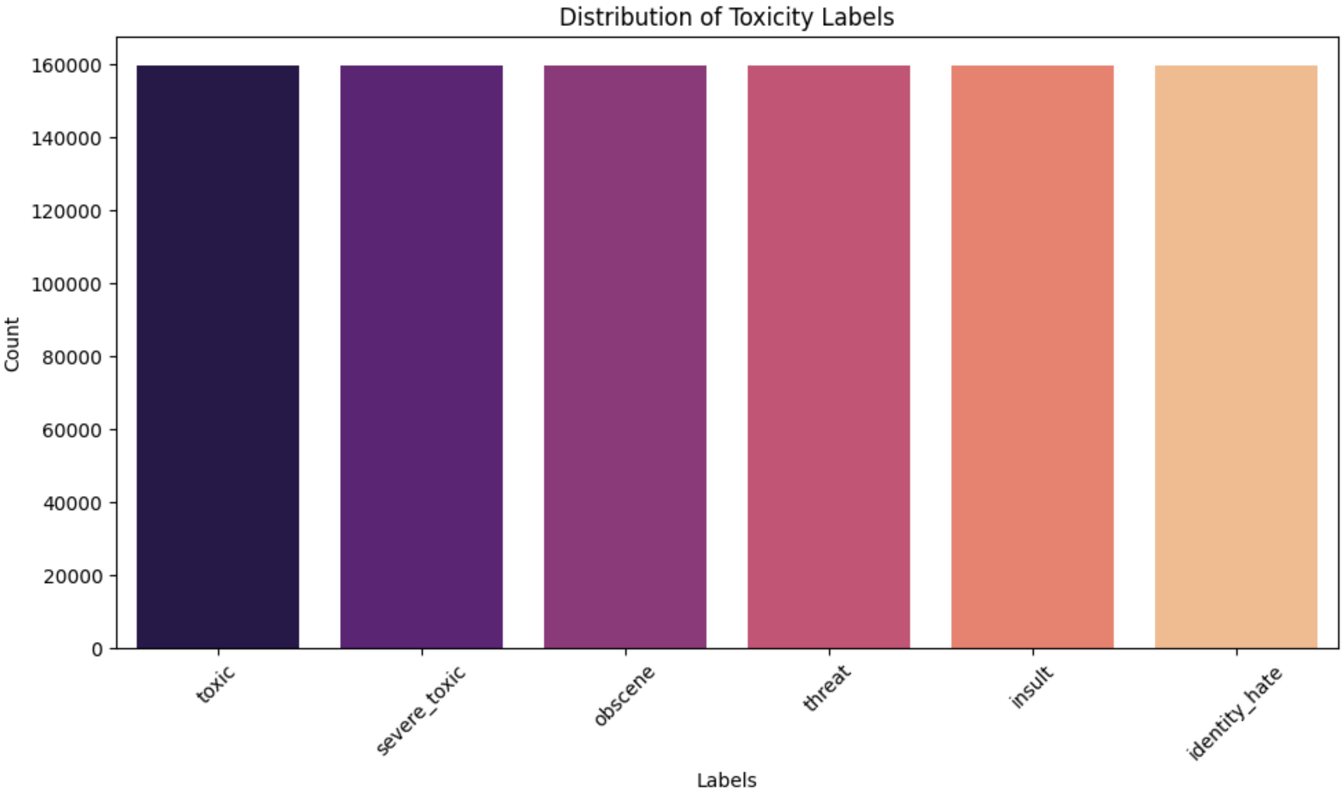
Predi

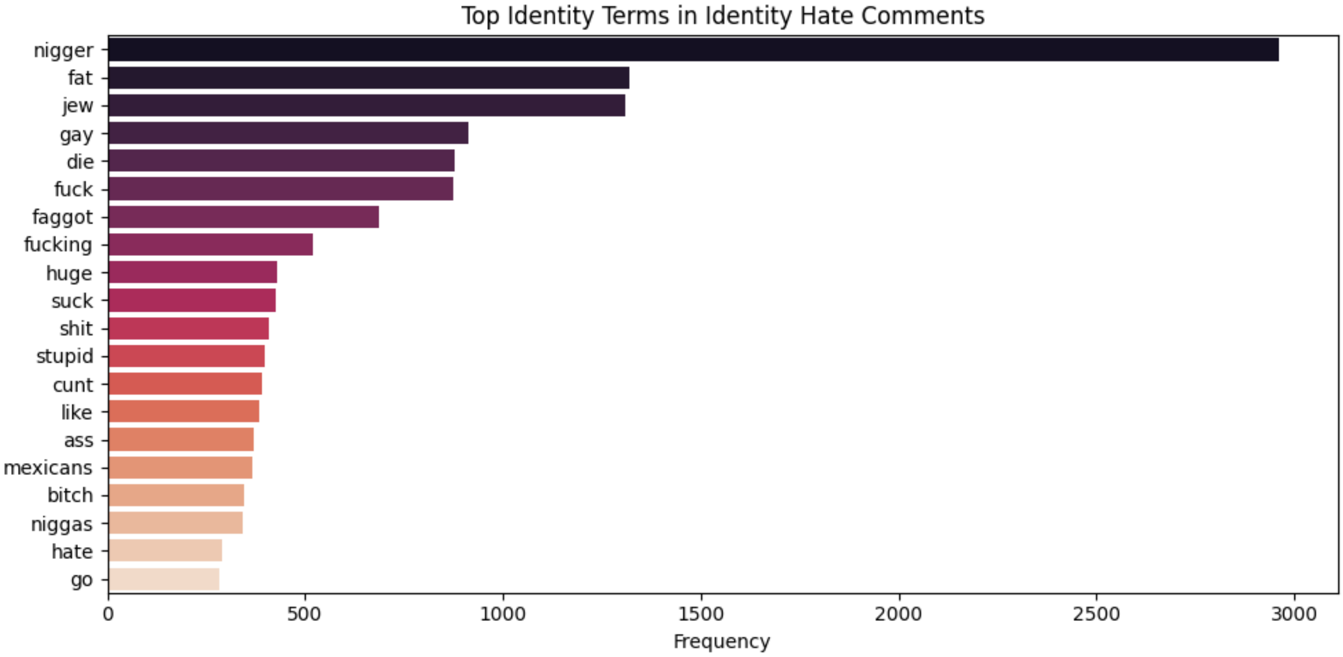


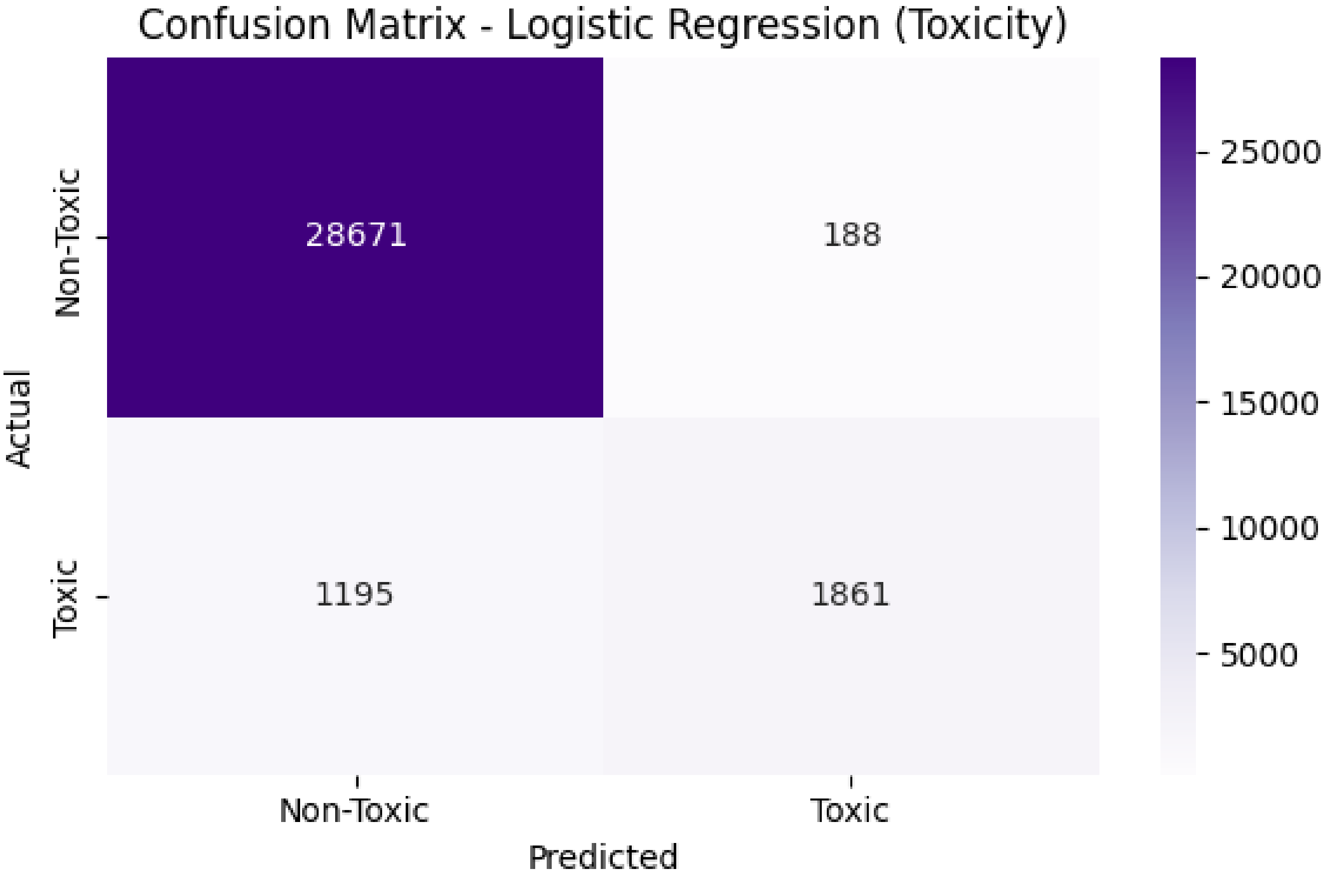
Exploratory Visualizations

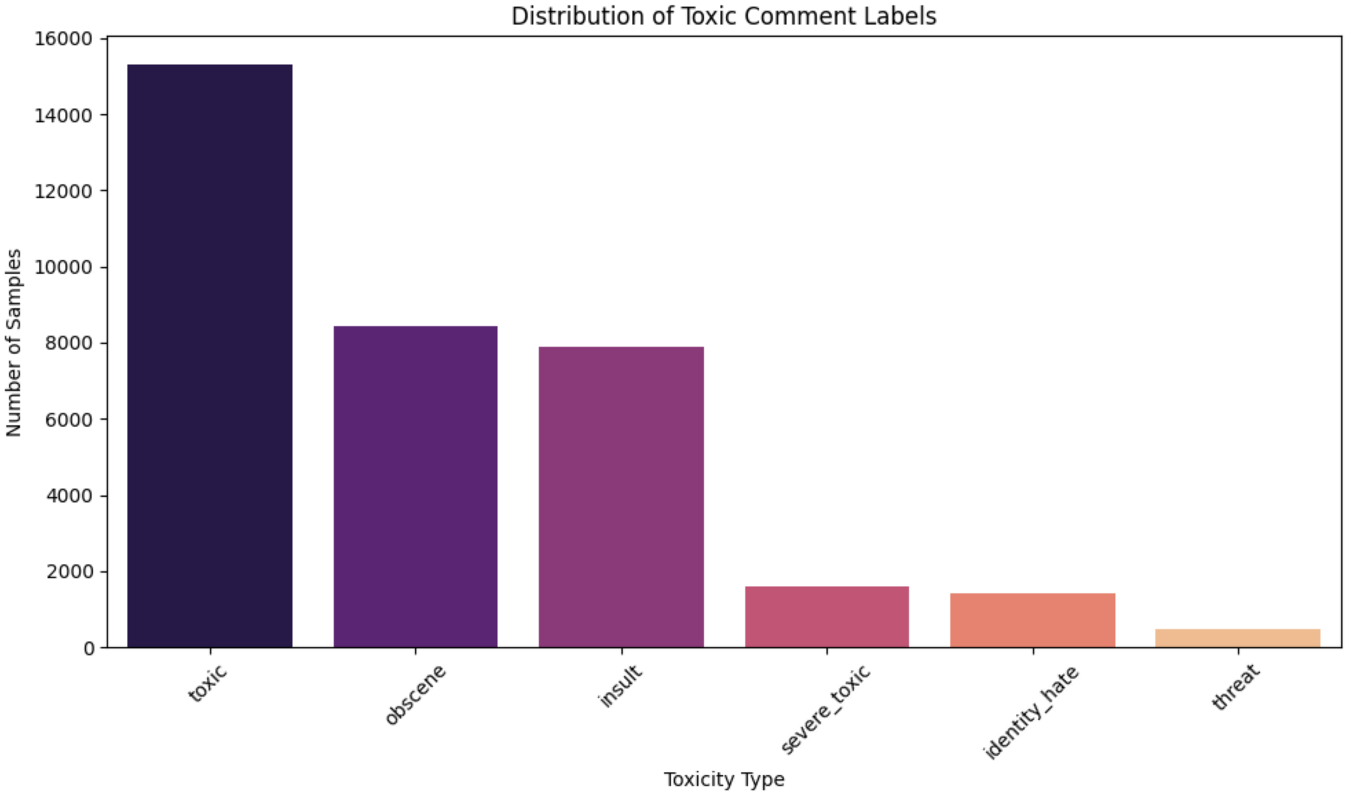




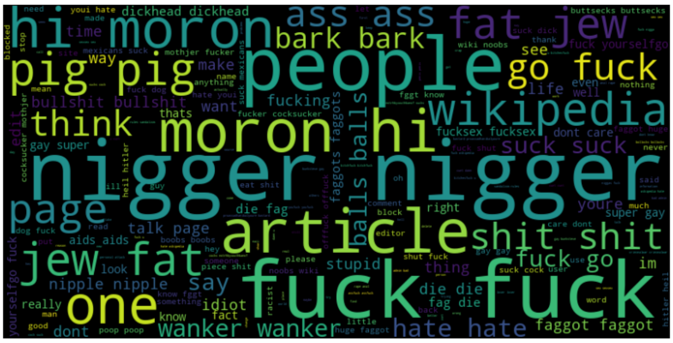

Distribution of Toxicity Labels
Count
0
20000
40000
60000
80000
100000
120000
140000
160000
toxic
severe_toxic
obscene
threat
insult
identity_hate
Labels


Top Identity Terms in Identity Hate Comments
nigger
fat
jew
gay
die
fuck
faggot
fucking
huge
suck
shit
stupid
cunt
like
ass
mexicans
bitch
niggas
hate
go
0
500
1000
1500
2000
2500
3000
Frequency

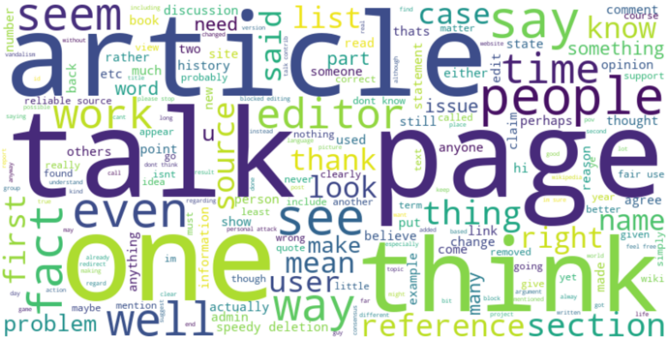

Confusion Matrix - Logistic Regression (Toxicity)
Actual
Non-Toxic
Toxic
28671
188
1195
1861
Non-Toxic
Toxic
Predicted
25000
20000
15000
10000
5000


Distribution of Toxic Comment Labels
Number of Samples
16000
14000
12000
10000
8000
6000
4000
2000
0
toxic
obscene
insult
severe_toxic
identity_hate
threat
Toxicity Type

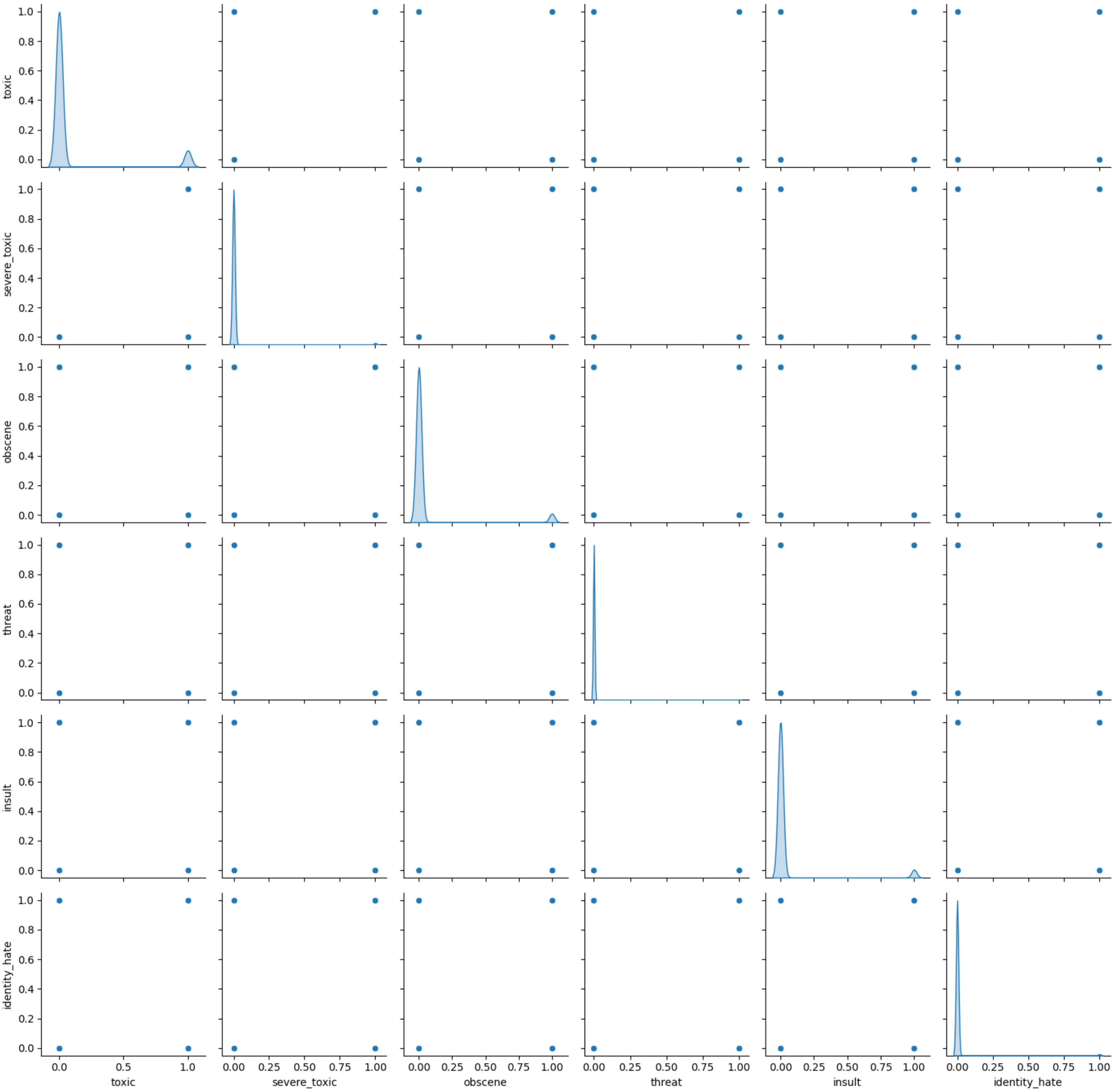
WordCloud - Toxic Comments

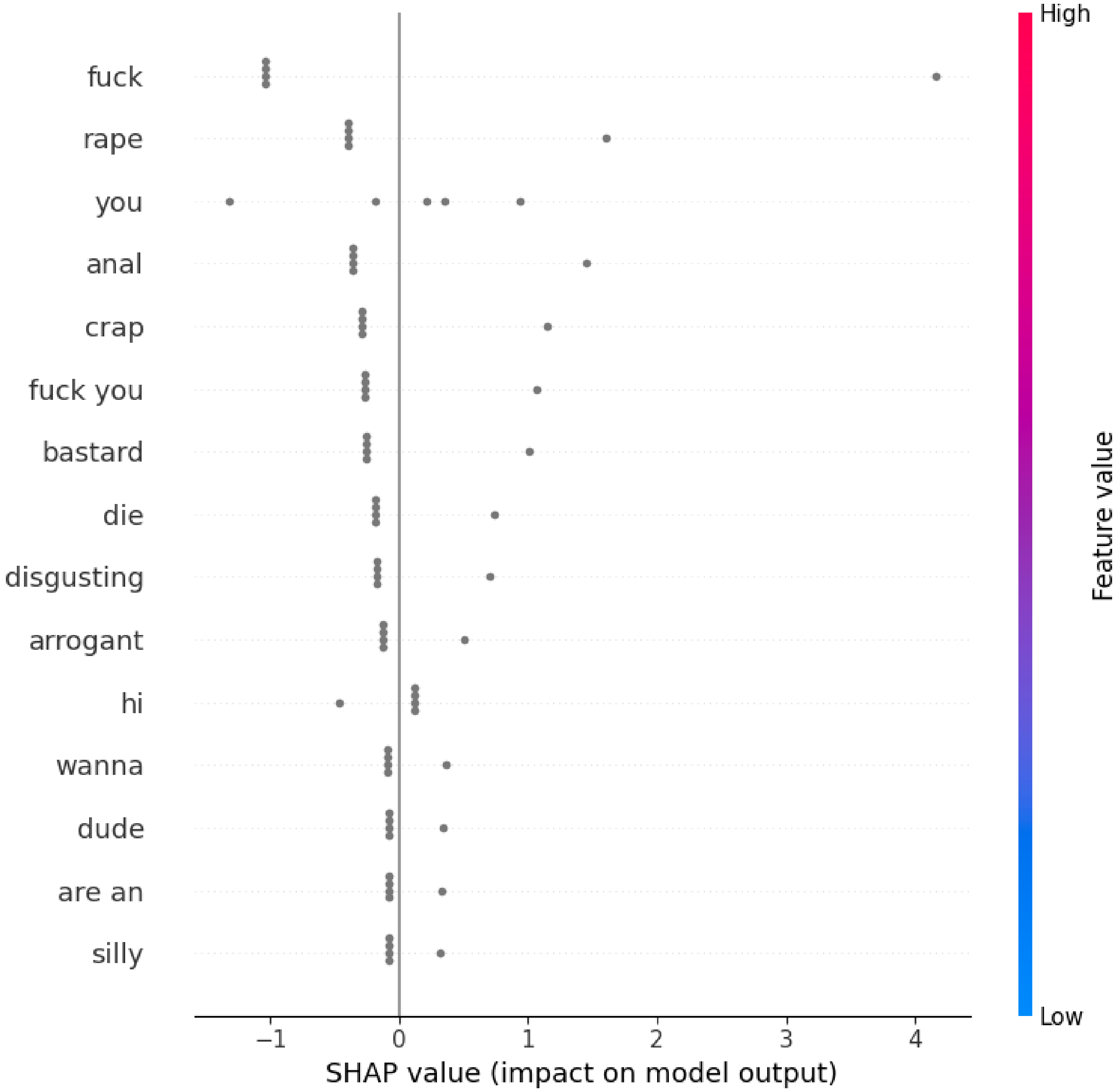


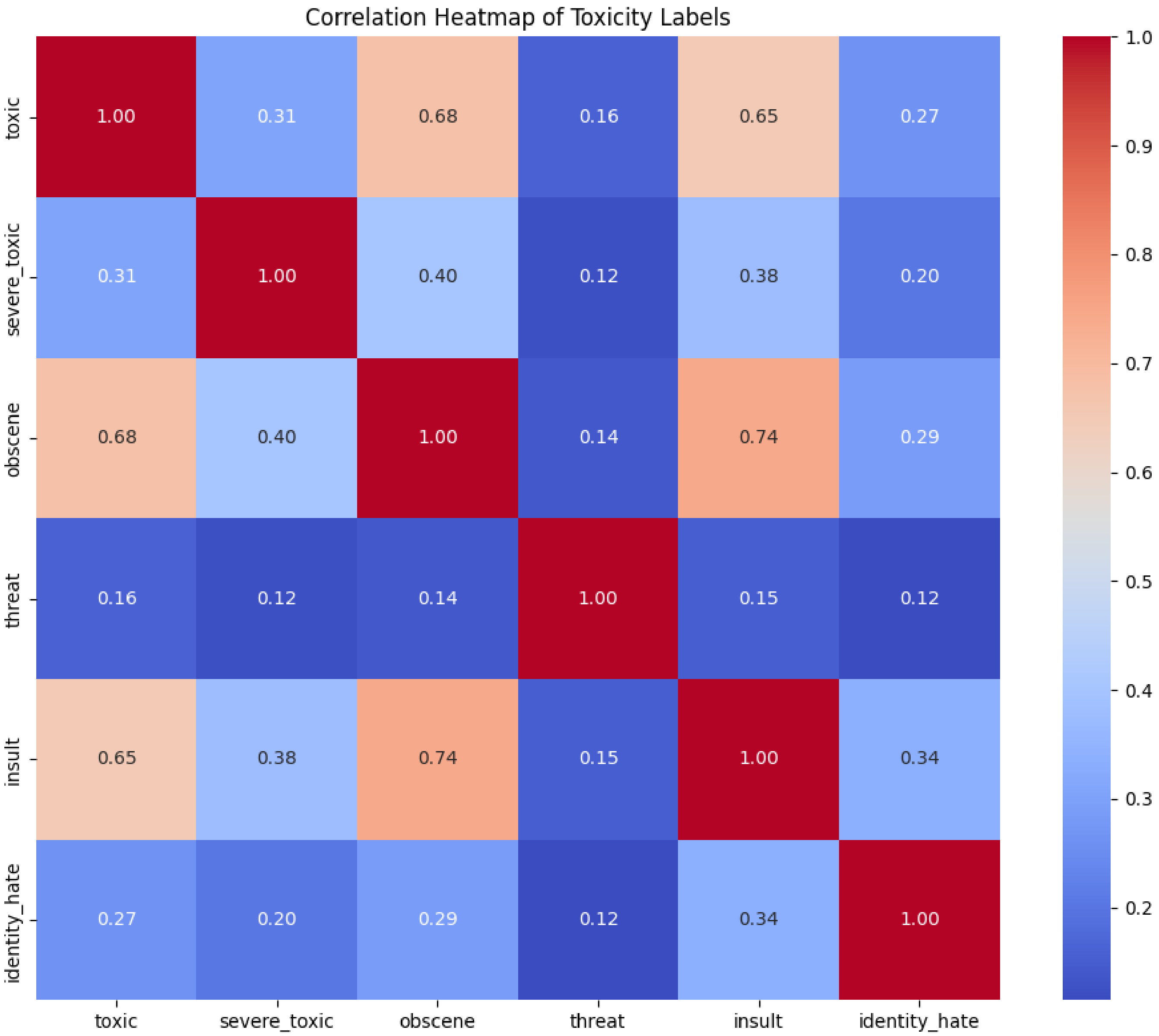
WordCloud - Non-Toxic Comments



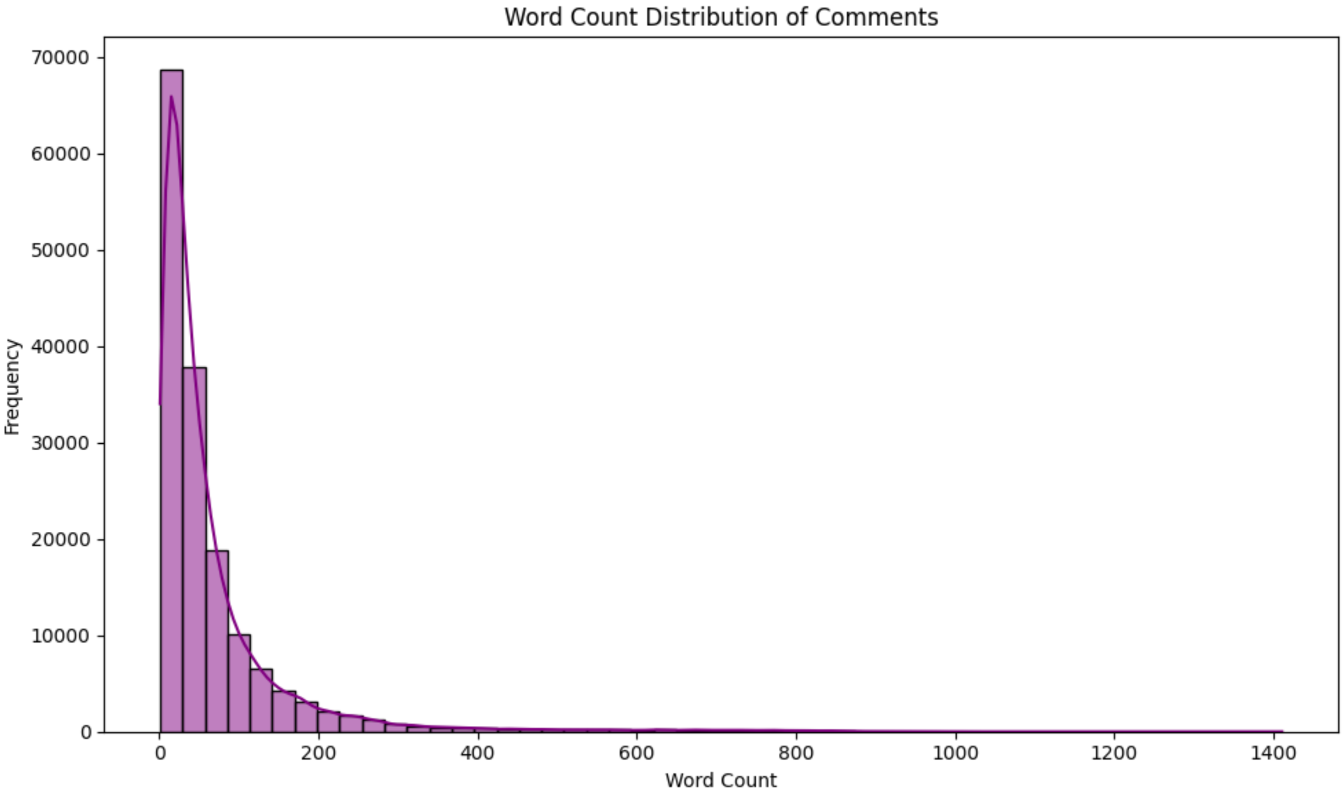
Pairplot of Toxicity Labels




fuck
rape
you
anal
crap
fuck you
bastard
die
disgusting
arrogant
hi
wanna
dude
are an
silly
−1
0
1
2
3
4
SHAP value (impact on model output)
High
Low
Feature value



Correlation Heatmap of Toxicity Labels


Word Count Distribution of Comments
Frequency
70000
60000
50000
40000
30000
20000
10000
0
0
200
400
600
800
1000
1200
1400
Word Count