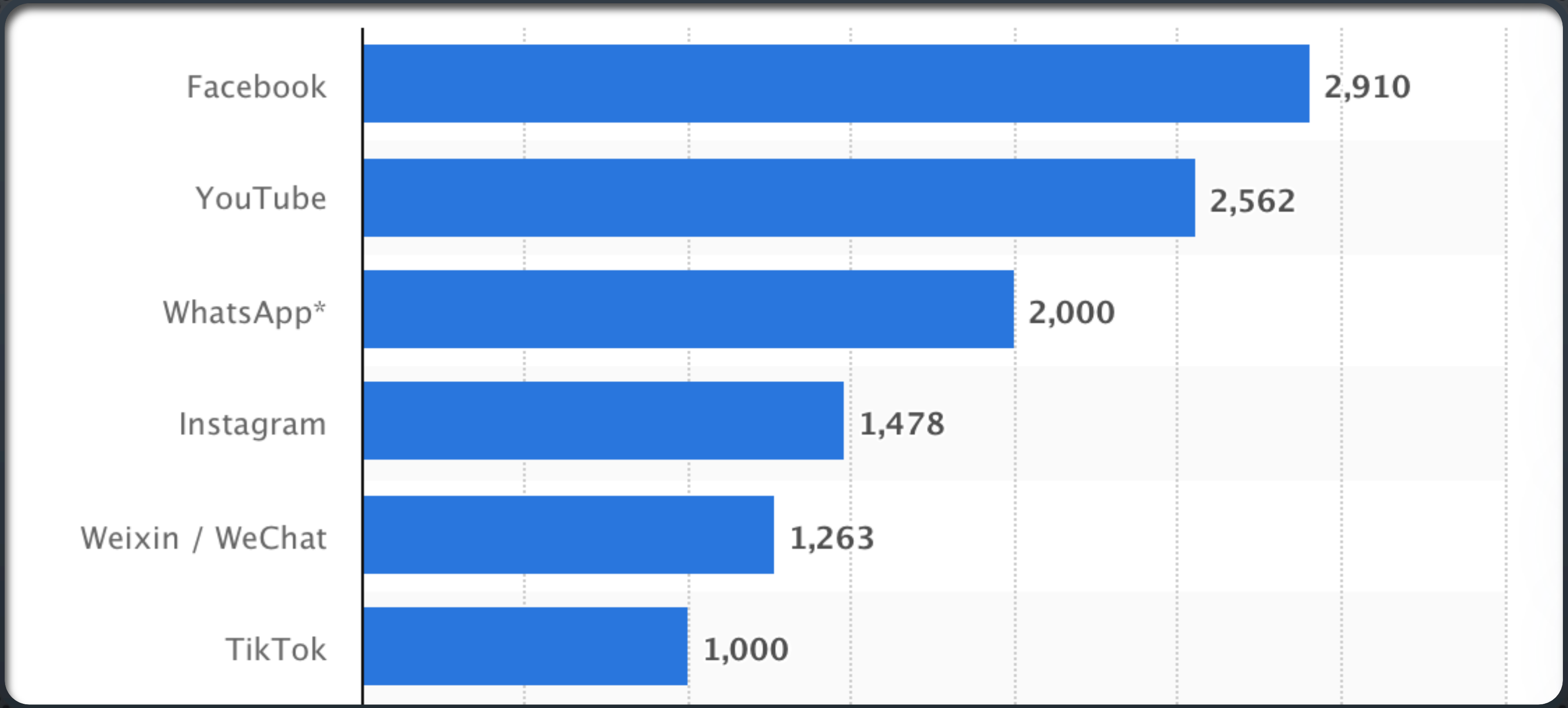
TOXIC COMMENT CLASSIFICATION

Ricardo Cepeda

Problem

- Create a model that can identify threats, obscenity, insults, identity-based hate in comments for future implementation in online social interactions.



MOST POPULAR SOCIAL NETWORKS 2022

IN MILLIONS

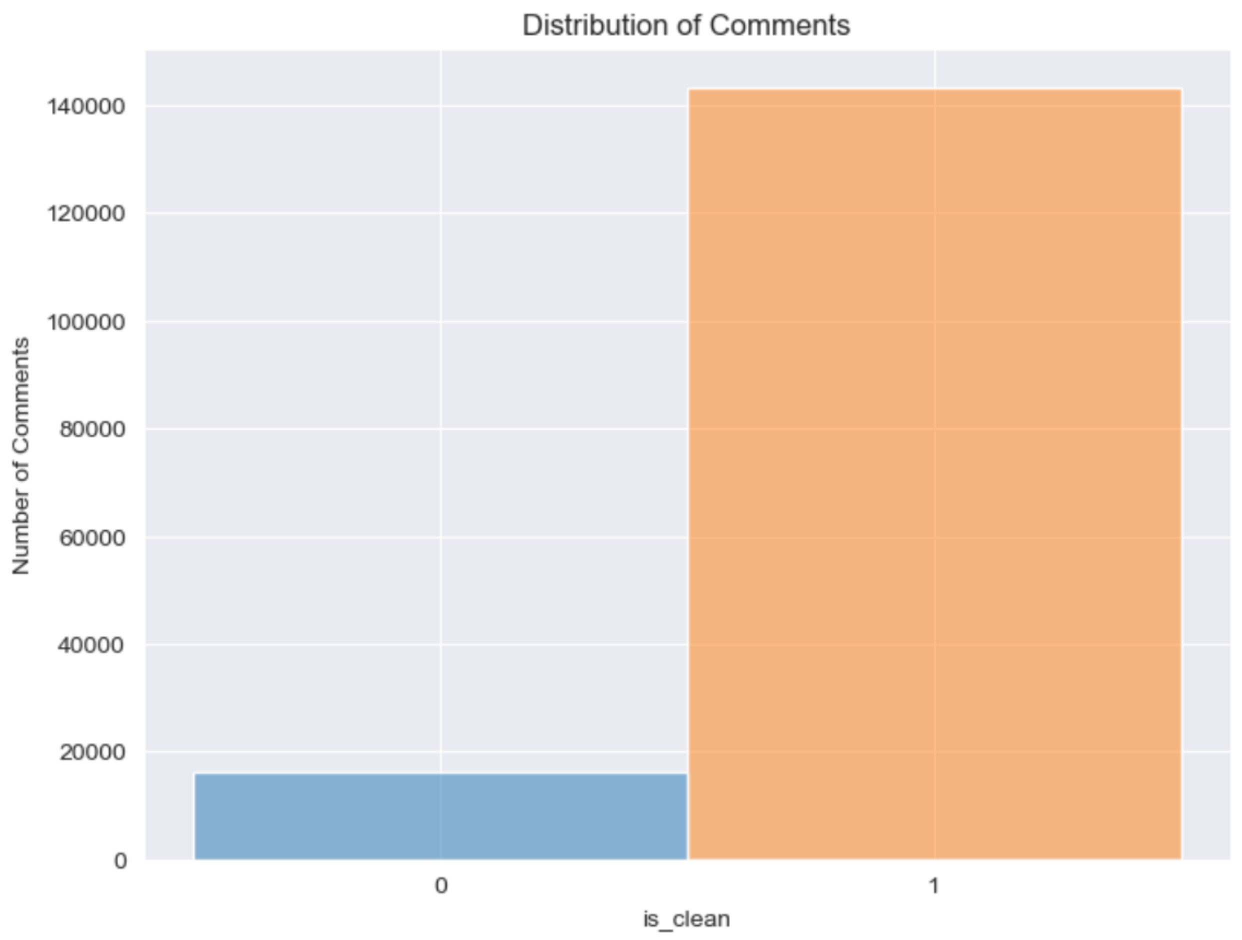
DATA

- Large number of Wikipedia comments which have been labeled by human raters for toxic behaviors.
- Source: Kaggle (159,571 comments)
- Types of toxicity:
  - Toxic
  - Severe_toxic
  - Obscene
  - Threat
  - Insult
  - Identity_hate

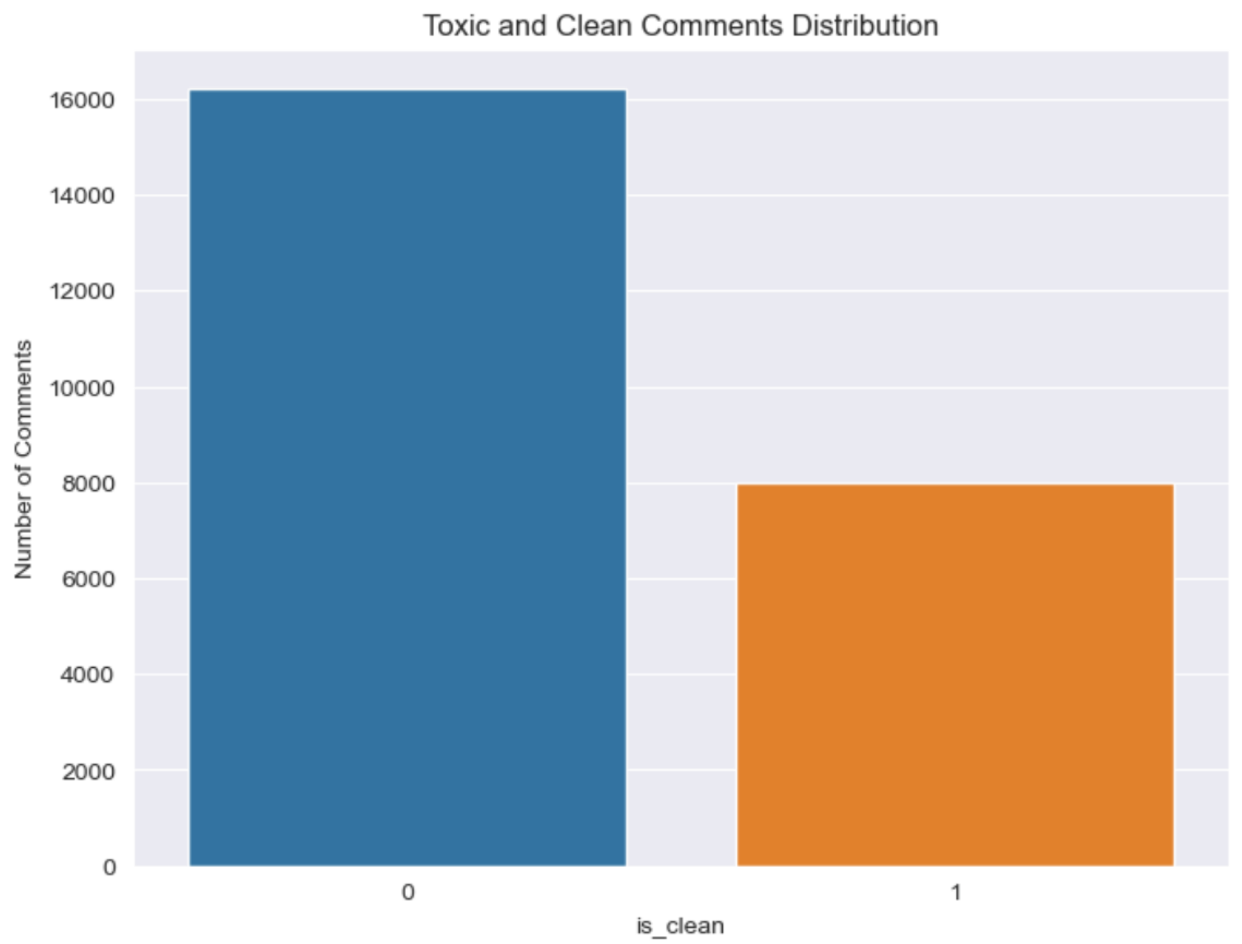
id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
b5db69dcbe54df49	Thanks for taking the time to inform me of the...	0	0	0	0	0	0
701794650c6273aa	Bogaert is used as a reference for different p...	0	0	0	0	0	0
a60b6a4c70b19be9	The Electrolux Timeline at states Kelvinator,...	0	0	0	0	0	0
b889d0235cc46712	The truth is difficult sometimes, best to acce...	0	0	0	0	0	0
2625e751c827b607	Suck my cheesy dick)	1	1	1	0	1	0

DATA SAMPLES

- "Nonsense? Kiss off, geek. What I said is true, I'll have your account terminated."
- "ban one side of an argument by a bullshit Nazi admin and you get no discussion because the Islamic editors feel they 'won'."
- "Why can you put English for example on some players, but other people don't like it – why?"

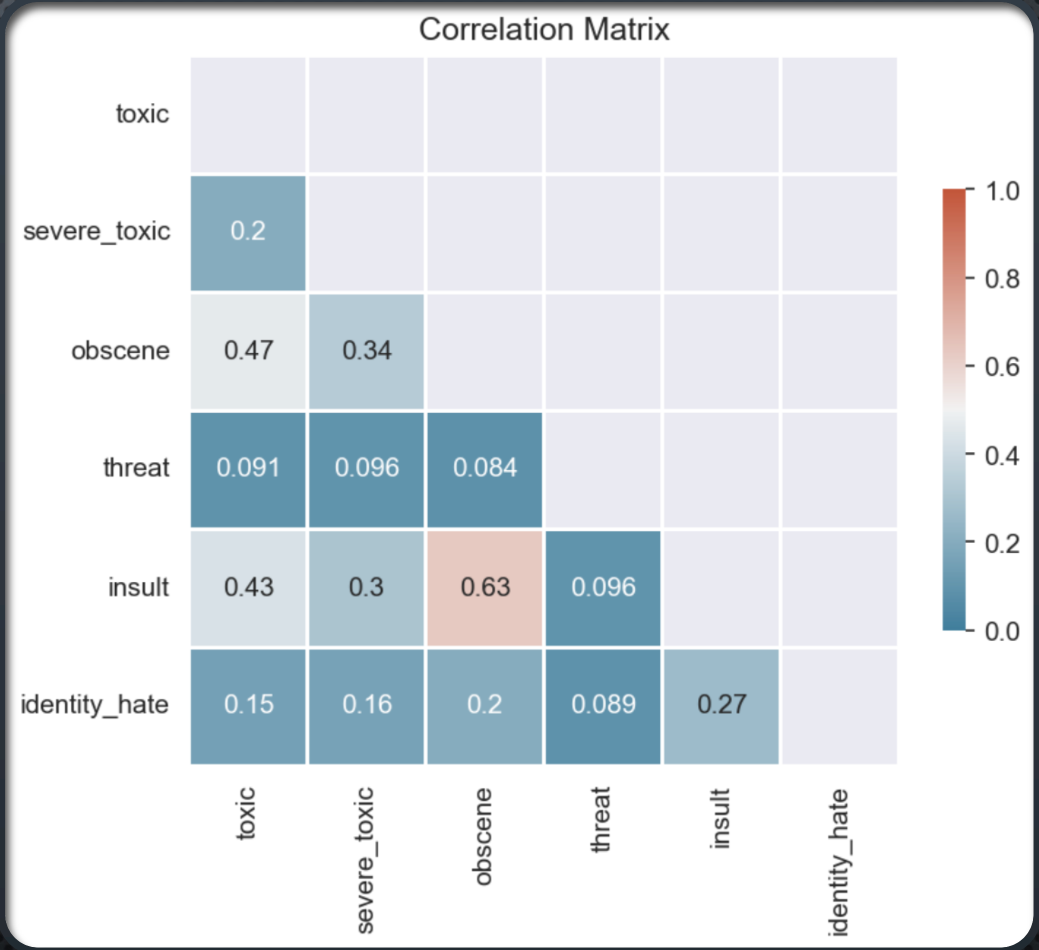


DATASET DISTRIBUTION



REBALANCE THE DATA

RELATION BETWEEN LABELS



WORD CLOUD REPRESENTATION

MODELING

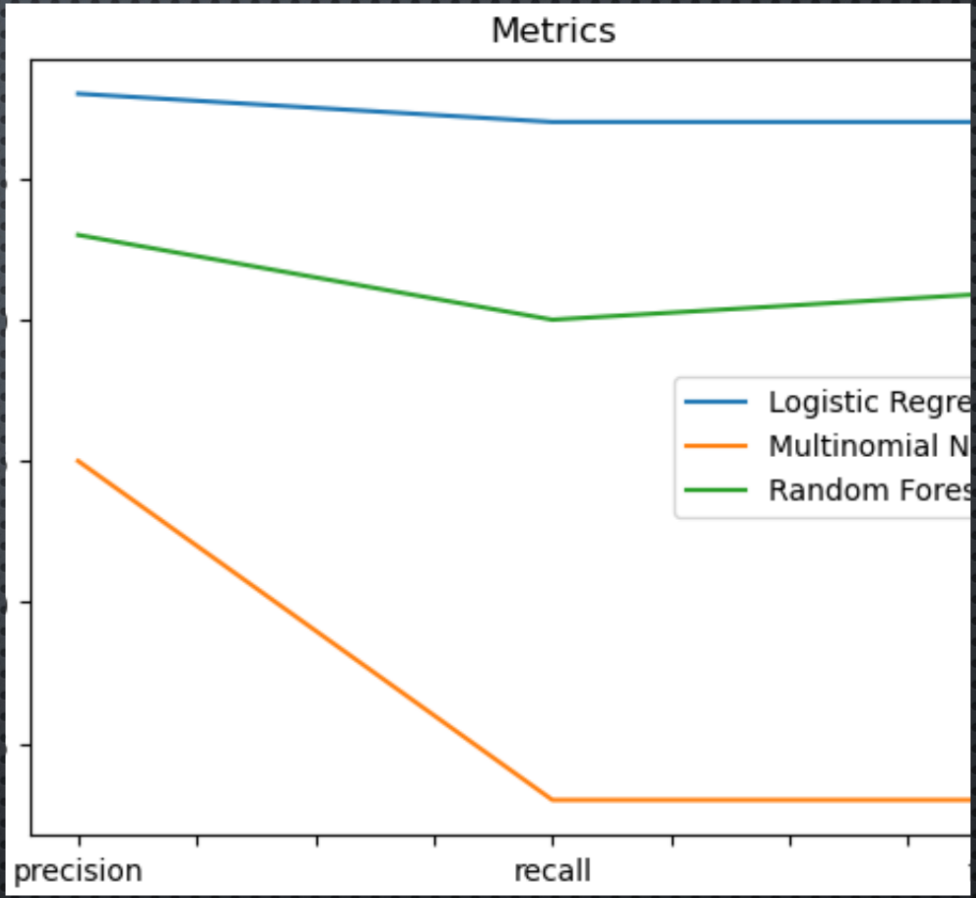
Model	Accuracy Score
OneVsRestClassifier Logistic Regression	**% 96.4**
OneVsRestClassifier MultinomialNB	% 63.3
Random ForestClassifier	% 90.9

	toxic	severe_toxic	obscene	threat	insult	identity_hate	is_clean
Logistic Regression	1.0	0.0	1.0	0.0	1.0	0.0	0.0
Multinomial NB	1.0	0.0	1.0	0.0	1.0	1.0	0.0
Random Forest	1.0	0.0	1.0	0.0	1.0	0.0	0.0
True Values	1.0	0.0	1.0	0.0	1.0	0.0	0.0

PREDICTIONS

"I AM A CUM GUZZLING MOTHERFUCKER THAT LIKES BOYS … "

evere_toxic	obscene	threat
0.0	1.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0
0.0	1.0	0.0



METRICS

Conclusions

Logistic regression the model to go with for this problem

Future improvements with deep learning