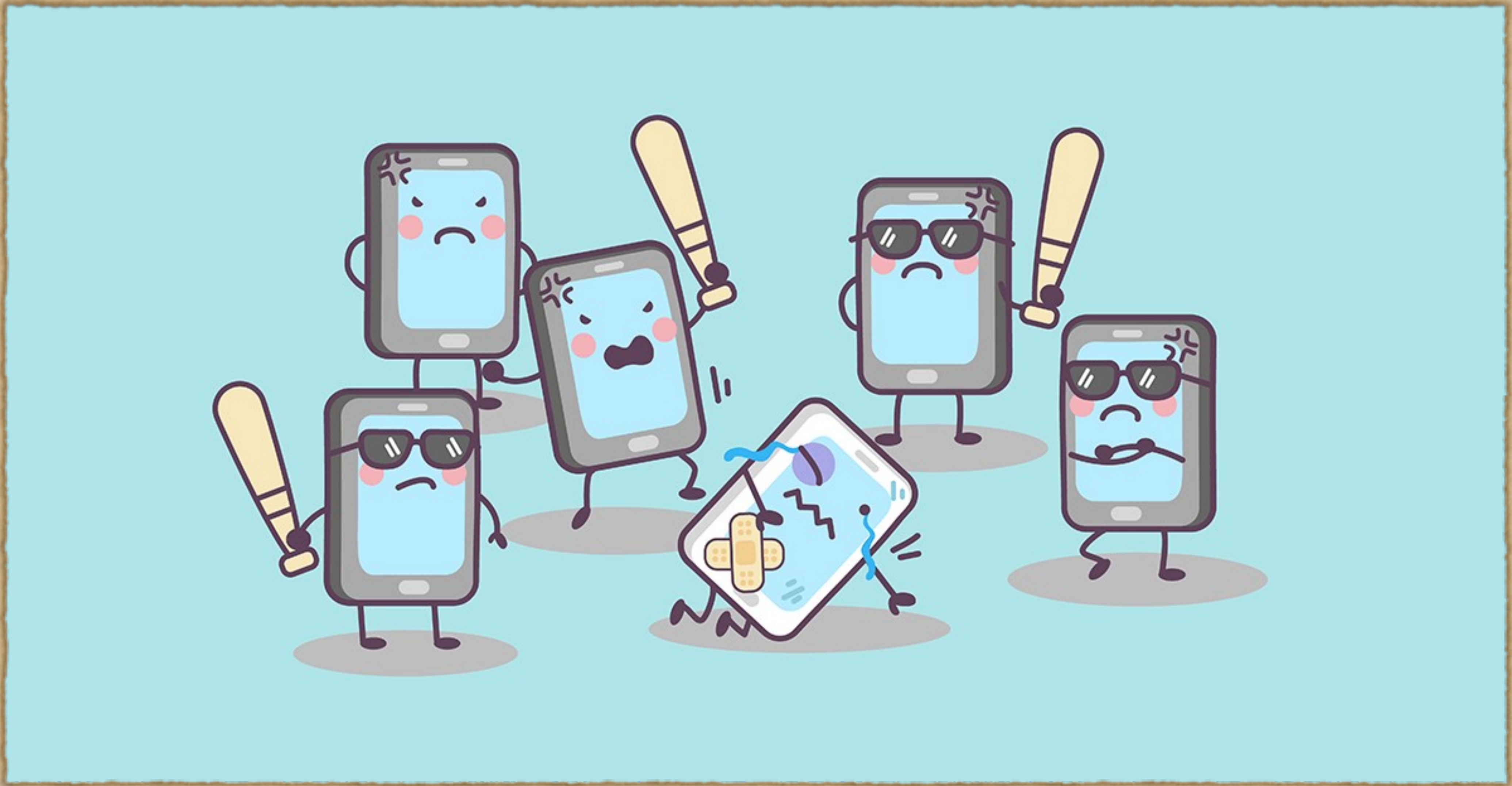


Toxic Comment Classification

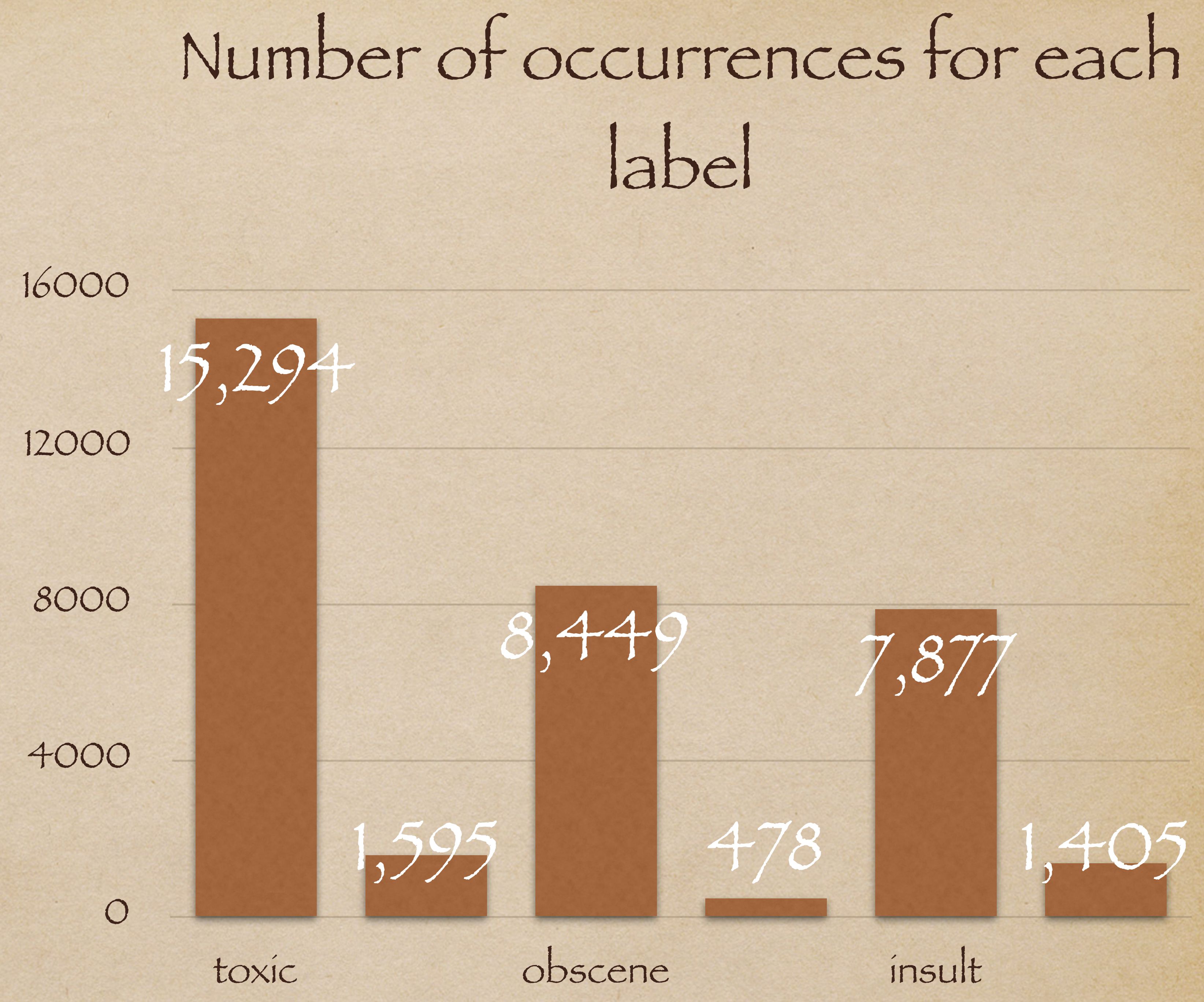
By Minh Ngoc Pham (Jean)

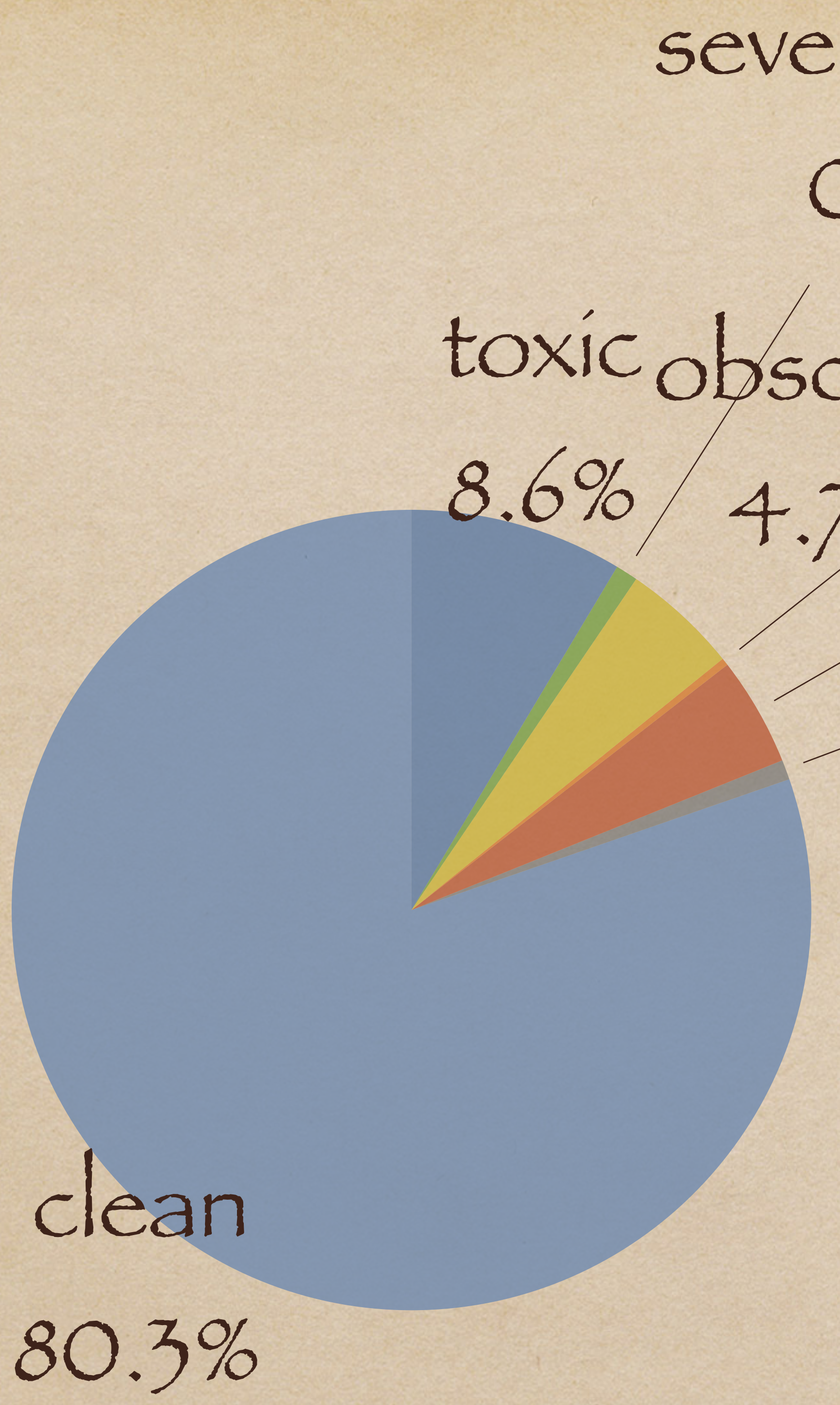
Project Introduction

kaggle

- Toxic Comment Classification Challenge
- Published in 2018
- 159,571 comments in training set
- 63,978 comments in test set

- Highest count in toxic comments
- There are a lot of clean comments





- 80.3% of the comments in the training test are clean
- Class imbalance issue

- High number of comment with NO label
- There are only 31 labels classified as all of the 6 labels

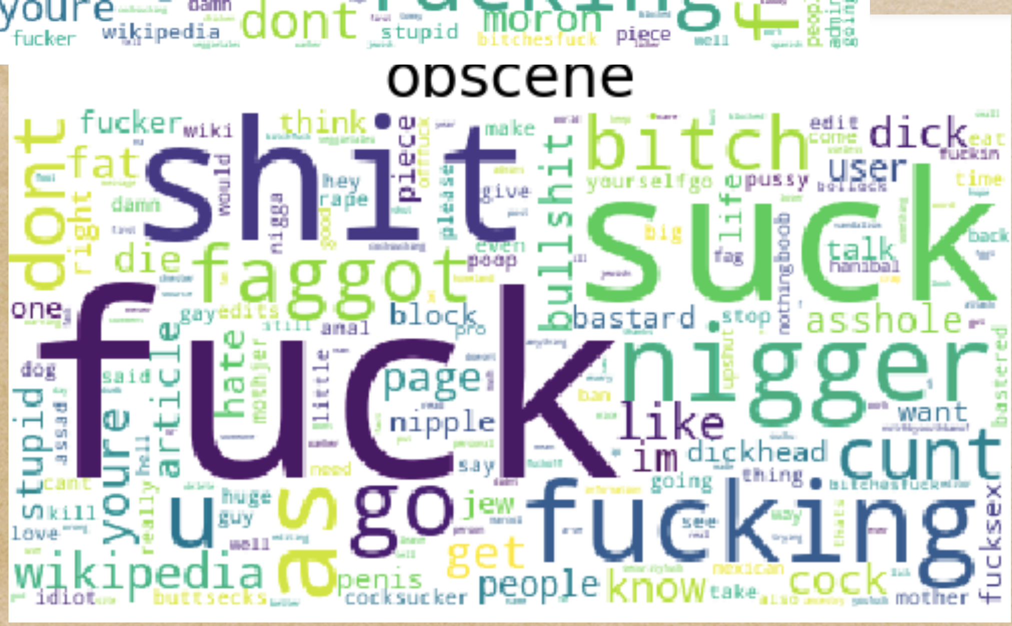
Word Cloud


toxic
fuck
dont
like
shit
nigger
fucking
wikipedia
faggot
page
gay
hate
bitch
sex
die
go
youre
moron


severe_toxic
suck
shit
fuck
as
bitch
cock
nigger
faggot
fucking
cunt
go
die

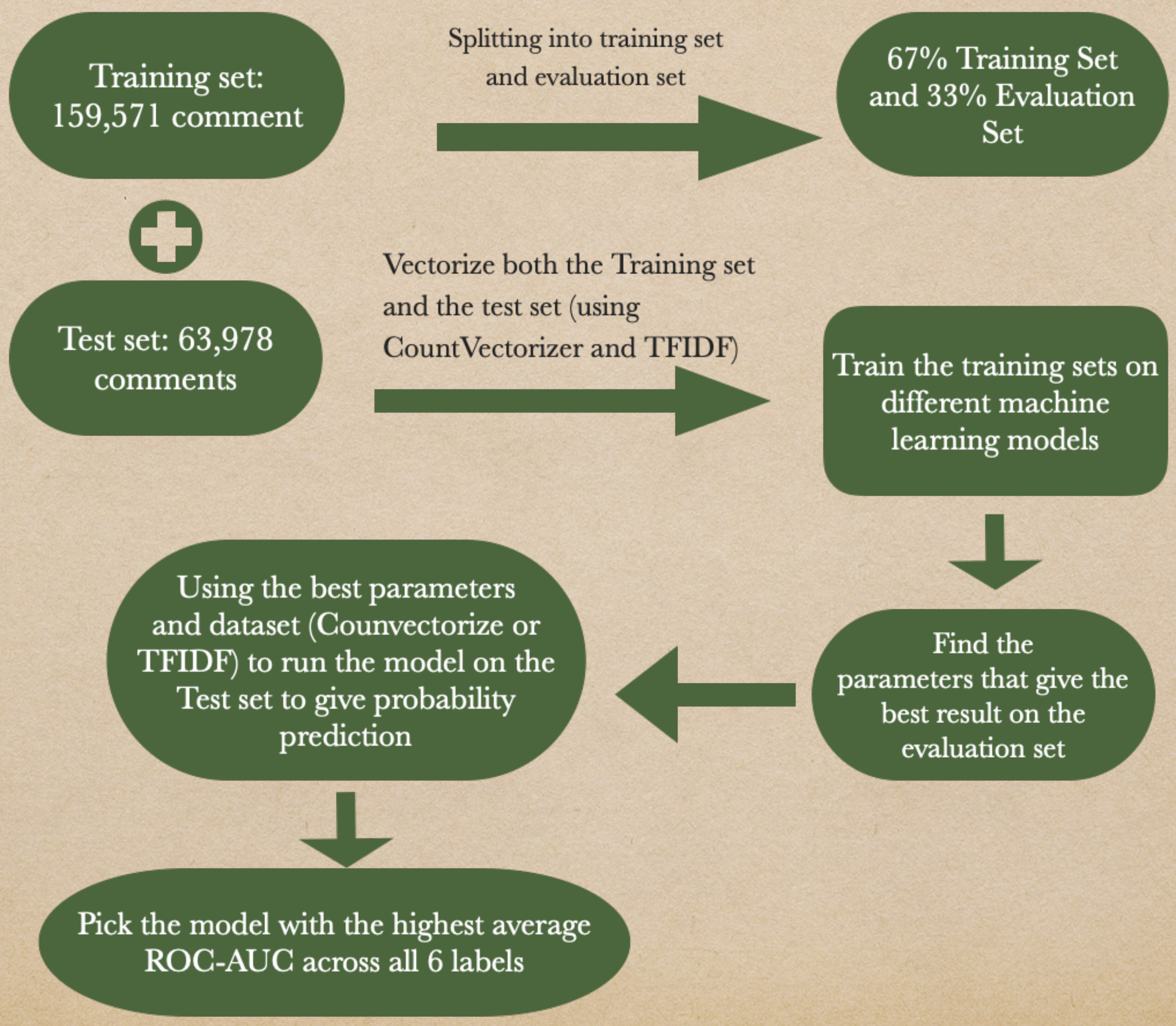

insult
fuck
suck
nigger
fucking
faggot
shit
as
go
bitch
fat
dont
youre
hate
cunt


threat
die
as
kill
going
must
block
wale
page
wikipedia
fucking
fuck


opscene
fuck
shit
suck
nigger
fucking
faggot
bitch
cunt
go
as
dont
like
asshole
wikipedia


identity_hate
nigger
jew
fat
gay
faggot
fuck
die
fucking
suck
nigga
huge
shit
likestupid
mexican
hate
bitch

Method used


Training set: 159,571 comment
Test set: 63,978 comments
Splitting into training set and evaluation set
67% Training Set and 33% Evaluation Set
Vectorize both the Training set and the test set (using CountVectorizer and TFIDF)
Train the training sets on different machine learning models
Find the parameters that give the best result on the evaluation set
Using the best parameters and dataset (Counvectorize or TFIDF) to run the model on the Test set to give probability prediction
Pick the model with the highest average ROC-AUC across all 6 labels

Logistic Regression

	OneVsRestClassifier (with C=1, estimator penalty =l1)- TFIDF	ClassifierChain ((with C=1, estimator penalty =l1) - TFIDF	OneVsRestClassifier (with C=1, estimator penalty =l2) - TFIDF	OneVsRestClassifier (with C=1, estimator penalty =l2) - CountVectorizer
Toxic	0.968258	0.968258	-	0.799235
Severe Toxic	0.984863	0.980751	-	0.765841
Obscene	0.981258	0.970786	-	0.778556
Threat	0.979104	0.965633	-	0.619170
Insult	0.971562	0.947265	-	0.774192
Identity	0.970317	0.955580	-	0.671779
Average	0.975894	0.964712	0.977498	0.734796

	TFIDF Vectorizer data for only 'words'	TFIDF Vectorizer data for both 'word' and 'char'
Result on evaluation dataset	0.9759	0.9831
Result on Test dataset	0.9729	0.979

Naive Bayes

	MultinomialNB() with alpha = 0.1	MultinomialNB() with alpha = 1	MultinomialNB() with alpha = 5	MultinomialNB() with alpha = 10	MultinomialNB() with alpha = 50
Average ROC-AUC	0.943539	0.853696	0.781917	0.762324	0.737935

	Toxic	Severe Toxic	Obscene	Threat	Insult	Identity Hate	Average
With TFIDF dataset	0.954660	0.973322	0.958055	0.912945	0.957876	0.933435	**0.948382**
With CV dataset	0.917226	0.929333	0.919417	0.848131	0.916570	0.864621	**0.899216**

Predictive Power on Test set: 0.9364

Decision Tree Classifier

	max_depth=10, criterion = 'gini' - TFIDF dataset	max_depth=10, criterion='entropy' - TFIDF dataset	OneVsRestClassifier, max_depth=10,criterion='entropy' - CountVectorize dataset	ClassifierChain, max_depth=10,criterion='entropy' - CountVectorize dataset	OneVsRestClassifier, max_depth=15,criterion='entropy' - CountVectorize dataset
Toxic	0.740011	0.841011	0.788924	0.788485	0.827153
Severe Toxic	0.805550	0.736647	0.800206	0.823473	0.684435
Obscene	0.841200	0.858569	0.862610	0.830148	0.858615
Threat	0.614732	0.703729	0.719132	0.558249	0.680063
Insult	0.766586	0.843947	0.836273	0.836210	0.822210
Identity Hate	0.747754	0.769022	0.780570	0.659727	0.777247
Average	**0.752639**	**0.792154**	**0.780570**	**0.749382**	**0.774954**

Predictive Power on Test set: 0.7941

Random Forest Classifier

	OneVsRestClassifier, n_estimators = 100, max_depth=15 - TFIDF dataset	ClassifierChain, n_estimators = 100, max_depth=15 - TFIDF dataset	OneVsRestClassifier, n_estimators = 100, max_depth=15 - CV dataset	OneVsRestClassifier, n_estimators = 1000, max_depth=15 - TFIDF dataset	OneVsRestClassifier, n_estimators = 1000, max_depth=10 - TFIDF dataset
Toxic	0.933500	0.933731	0.920494	0.936943	0.925093
Severe Toxic	0.979556	0.977144	0.973829	0.982700	0.981119
Obscene	0.976782	0.974873	0.963179	0.979460	0.974446
Threat	0.921531	0.928160	0.928774	0.951529	0.950142
Insult	0.960064	0.959790	0.943864	0.963969	0.957683
Identity Hate	0.946986	0.948915	0.934240	0.961600	0.956791
Average	**0.953070**	**0.953769**	**0.944063**	**0.962700**	**0.957546**

Predictive Power on Test set: 0.9668

Conclusion

All four models perform relatively well in predicting the labels for the toxic comments.

	Logistic Regression	**Naive Bayes**	**Decision Tree Classifier**	**Random Forest Classifier**
Average ROC-AUC	0.979	0.948382	0.7941	0.9668

Online community is no longer what it used to be and toxic comments disrupt the healthy discussion that can exist and drive serious users away. Social media platforms such as instagram and facebook and among others can certainly make use of a more effective way of filtering toxic comments to ensure a better online experience for all.
While Instagram has setting for users to manually filter a set of keywords, it should not be up to the users to have to ensure they are not exposed to a toxic online environment.