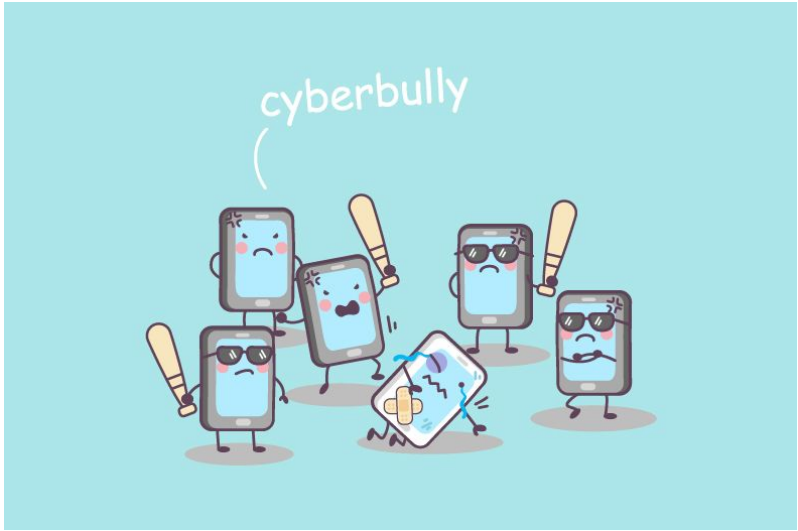
Toxic Comments Filter

Shanka Attanayake
Andres Lojan Yepez
Leandra Lai
Princy Chahal
Reza

Table Content

Intro

- A lot of toxic comment exist on the internet, especially in social media
- It destroy the way we communicate



Problem Statement

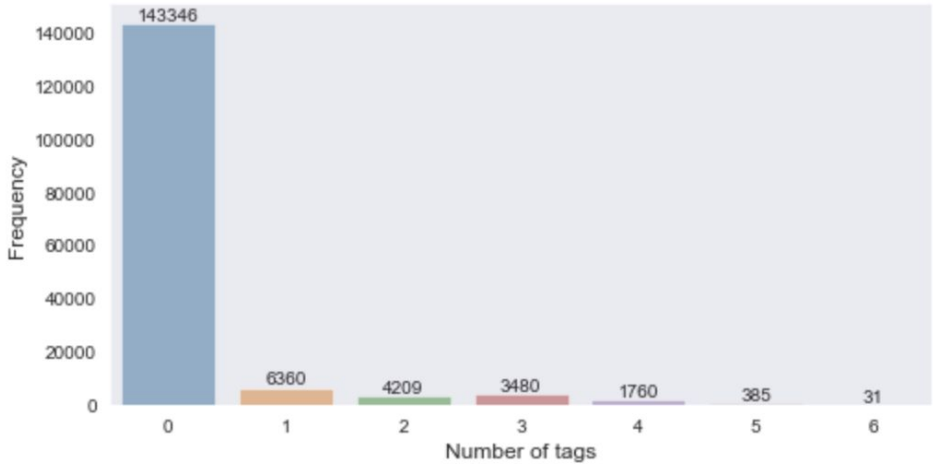
We want to identify if a comment is toxic or not on social media to enable a better and healthier communication environment

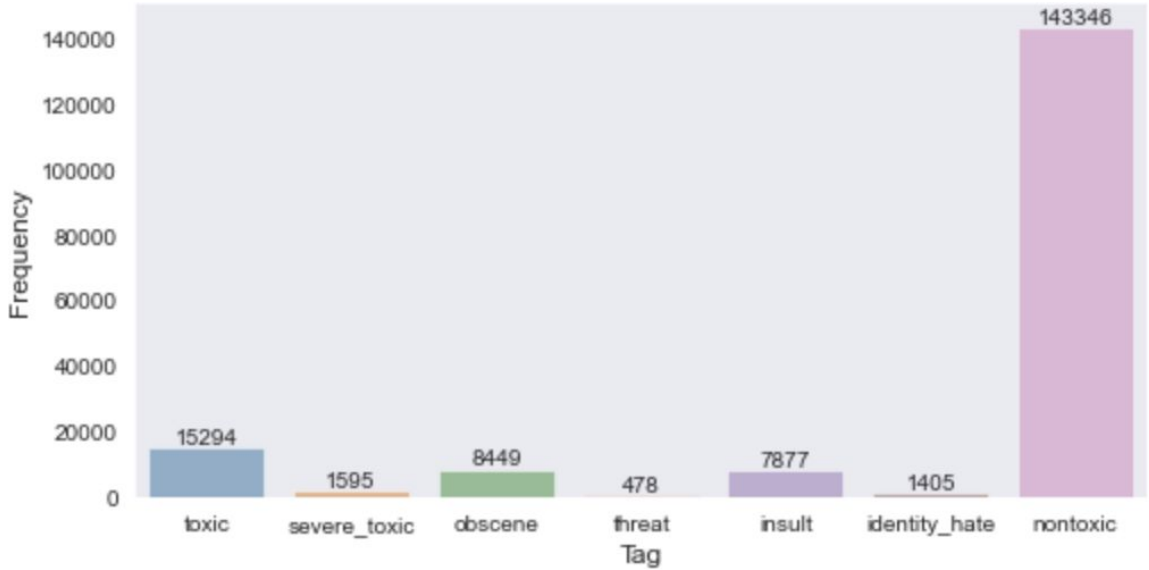
Dataset

- From Kaggle
- Consist comments from Wikipedia's talk page edits
- Separate test and train dataset
- Around 10% = toxic

	: train	test
# of rows	: 159571	153164
percentage	: 51	49

Total # of comments = 159571
Total # of nontoxic comments = 143346
Total # of tags = 35098





Note that the toxicity is not evenly spread out. Therefore, we must be careful about imbalance issues.

Data Cleanup

```
"Rex Mundi \n\nI've created a stub on Rex Mundi at Rex Mundi High School.  Only thing I know about it is that both my Aunt Donn
a and Bob Griese went there.  Please add anything you might know about it.\n\nBTW, my dad was a Panther; I live in Princeton my
self."
```

```
'rex mundi I have create stub rex mundi rex mundi high school thing know aunt donna bob griese go please add anything might kno
w it btw dad panther live princeton '
```

Model Pipeline

Machine Learning Without Cleaning

Countvectorizer & TFIDF

ML Classifier	Label	Precision	Recall	F1-Score	Support	Accuracy
Random Forest	Non-Toxic	0.94	0.96	0.95	57888	0.91
	Toxic	0.55	0.46	0.50	6090	
Logistic Regression	Non-Toxic	0.98	0.93	0.95	57888	0.92
	Toxic	0.55	0.79	0.65	6090	
Decision Tree	Non-Toxic	0.97	0.89	0.93	57888	0.88
	Toxic	0.42	0.76	0.54	6090	
Multinomial NB	Non-Toxic	0.92	1.00	0.96	57888	0.92
	Toxic	0.81	0.18	0.29	6090	

Machine Learning With Cleaning

Countvectorizer & TFIDF

ML Classifier	Label	Precision	Recall	F1-Score	Support	Accuracy
Random Forest	Non-Toxic	0.96	0.97	0.97	57888	0.94
	Toxic	0.70	0.63	0.66	6090	
Logistic Regression	Non-Toxic	0.98	0.93	0.95	57888	0.92
	Toxic	0.54	0.78	0.64	6090	
Decision Tree	Non-Toxic	0.98	0.91	0.94	57888	0.90
	Toxic	0.48	0.80	0.60	6090	
Multinomial NB	Non-Toxic	0.92	1.00	0.96	57888	0.92
	Toxic	0.89	0.19	0.32	6090	

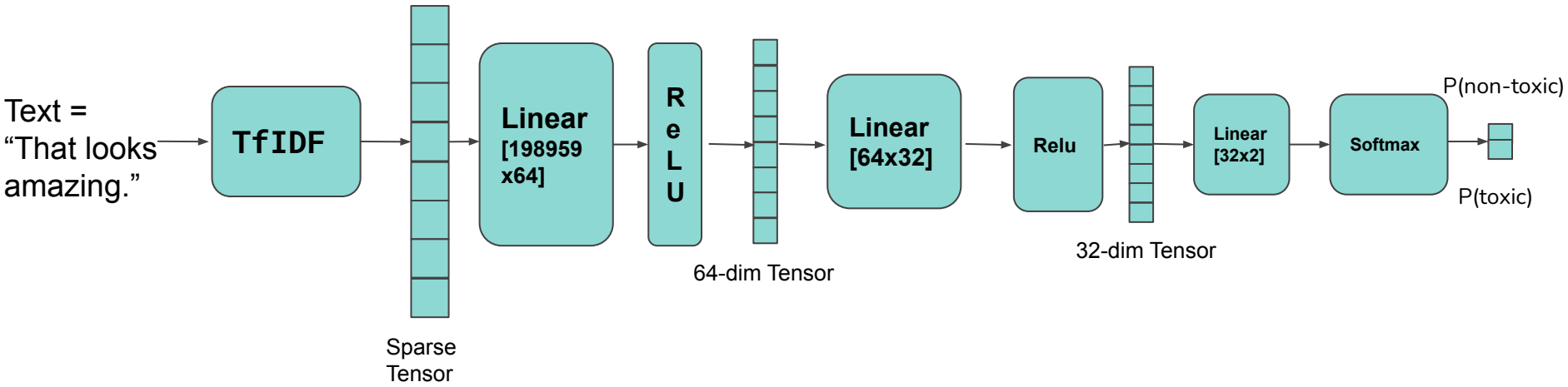
Best Machine Learning Model

Random Forest WIth Countvectorizer and Tf-idf Transformer

```
steps = [("vect", CountVectorizer()),
         ("tfidf", TfidfTransformer()),
        ("clf", RandomForestClassifier())]
```

	precision	recall	f1-score	support
0	0.96	0.97	0.97	57888
1	0.70	0.63	0.66	6090
accuracy			0.94	63978
macro avg	0.83	0.80	0.81	63978
weighted avg	0.94	0.94	0.94	63978

Neural Network for Toxic Comment Classification

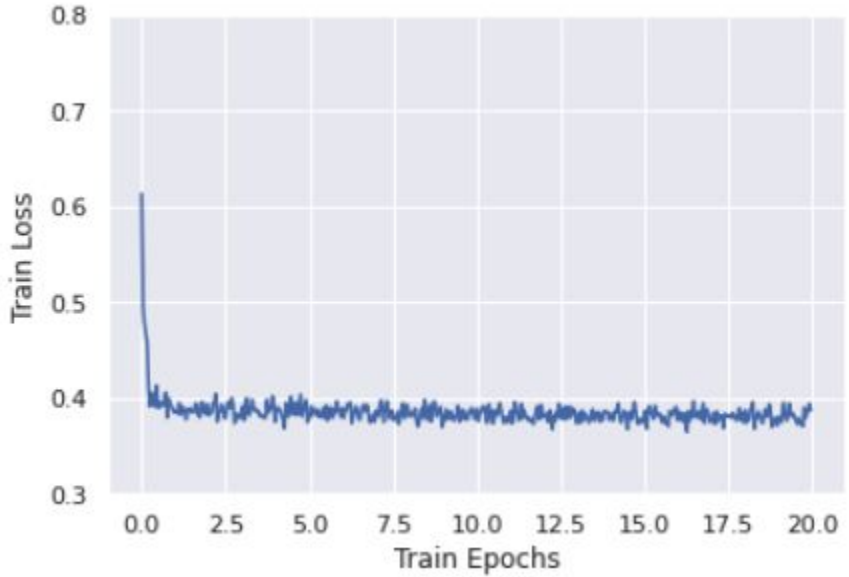


Neural Network Loss Weights

Cross Entropy loss = $-w_y \log p(y)$

We set $w_0 = 1$ and $w_1 = 4$ to account for imbalance.

Train epochs = 20

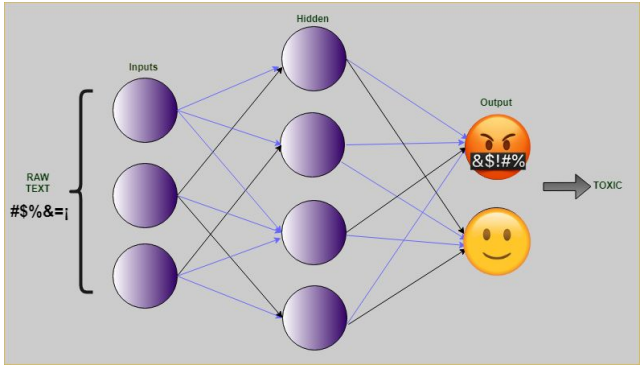


Neural Network Results

ML Classifier	Label	Precision	Recall	F1-Score	Support	Accuracy
Neural Network (MLP)	Non-Toxic	0.97	0.95	0.96	57888	0.94
	Toxic	0.60	0.72	0.65	6090	

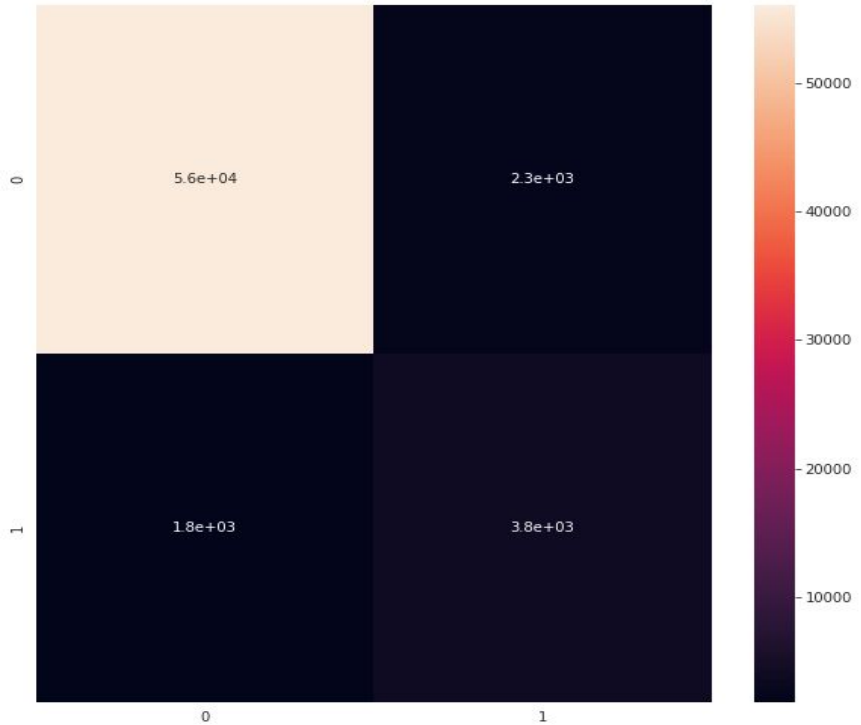
Demo

- “ I hope you die in a house fire.”
- “ I wish your dog dies so it doesn’t have to wake up everyday to the sight of your face”



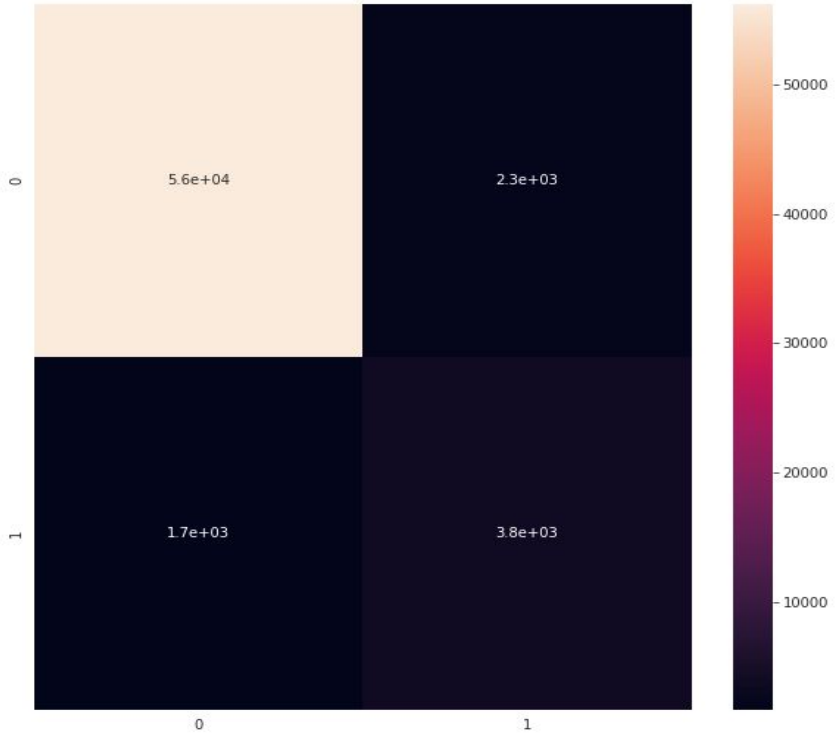
Confusion Matrix:
Random Forest with CountVectorizer

	precision	recall	f1-score	support
0	0.96	0.97	0.97	57888
1	0.68	0.63	0.65	6090
accuracy			0.94	63978
macro avg	0.82	0.80	0.81	63978
weighted avg	0.93	0.94	0.94	63978



Confusion Matrix: Random Forest with CountVectorizer and Tf-idf Transformer

	precision	recall	f1-score	support
0	0.96	0.97	0.97	57888
1	0.70	0.63	0.66	6090
accuracy			0.94	63978
macro avg	0.83	0.80	0.81	63978
weighted avg	0.94	0.94	0.94	63978



Conclusion

- Clean data VS Raw data makes the difference in terms of accuracy
- Based on the simplicity of a binary classifier task, Classic Algorithms turn out a better option rather than NN
- We are looking forward to seeing new results by trying with some other variations of this methodology

Thank You! Questions?