




Toxic Comment Classification

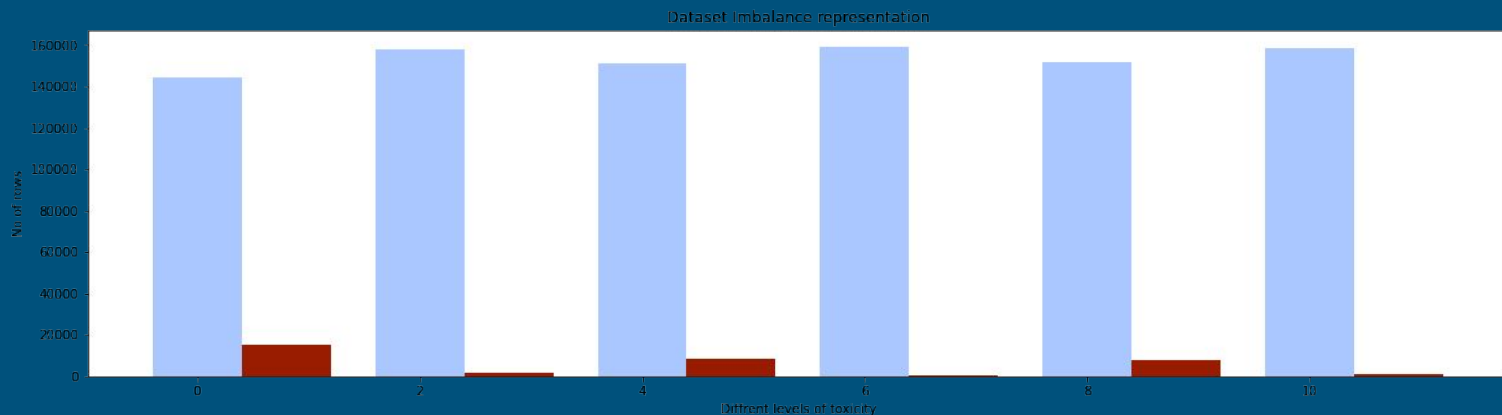


Nirav Patel (187242)
Saurav Shah (187157)
Srikar Reddy (187263)
Srujan Kumar (187247)



Dataset

- Comments and 6 levels of toxicity : toxic, severe toxic, obscene, threat, insult, identity hate
- 2335 unique characters and 1542 alphabets, so comments in other languages also (107 to be exact!)
- 53229 unique words before any pre processing!



Data Preprocessing

- Case of the alphabets doesn't help in extracting information, so used lower case
- The multiple spaces were converted to single space
- Characters other than the alphabets were removed
- Stop words removed
- Removed links
- Used regular expression for filtering out

This IS A aPPLe => this is a apple

this is => this is a apple

a apple

I, am, this, that

Data Representation

- Bag of Words
- Document Matrix
- Term Frequency Inverse Document Frequency (tf-idf)
- Binary Document Matrix
- Skelarn's TfidfVectorizer, CountVectorizer, set then CountVectorizer

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

Logistic Regression

- Document is transform into Term Frequency/ Inverse document matrix
- Every unique words is a feature
- Different model for every toxicity level
- Best value of Regularization constant was determined using cross-validation set
- Best value of Regularization constant was determined to be $C=100$
- Used sklearn's LogisticRegression

	Accuracy					
	Toxic	Severe Toxic	Obscene	Threat	Insult	Identity hate
Train	99.73	99.86	99.85	99.99	99.70	99.96
Test	99.72	99.90	99.87	99.70	99.70	99.97

Naive Bayes

- Document is transform into Document Vector
- Every unique word is a feature
- Different Models for each toxicity level
- Multinomial naive bayes with value of $\alpha=0.01$
- It is a generative model
- Used sklearn's MultinomialNB

	Accuracy					
	Toxic	Severe Toxic	Obscene	Threat	Insult	Identity hate
Train	97.74	99.45	98.74	99.81	98.42	99.60
Test	97.73	99.46	98.76	99.80	98.47	99.60

Support Vector Machine

- Comment into Tf-Idf
- Linear Kernel, as NLP problem, many (223530) unique words so already in high dimension, linearly separable data, is faster
- Best value of regularization parameter was determined to be 500
- Used sklearn's LinearSVC

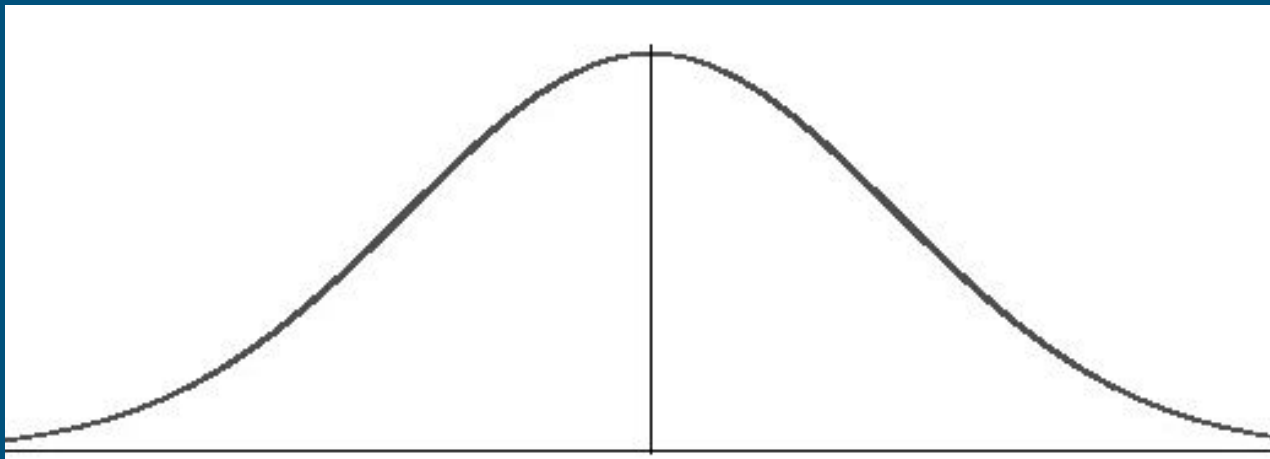
Accuracy						
	Toxic	Severe Toxic	Obscene	Threat	Insult	Identity hate
Train	99.91	99.78	99.90	99.99	99.84	99.98
Test	99.91	99.83	99.90	99.99	99.85	99.99

Final Aggregated Results

	Logistic Regression	Naive Bayes	Support Vector Machine
Train	99.84%	98.96%	99.9%
Test	99.81%	98.97%	99.91%
Train Time (Order)	20-30 minutes	1-2 minutes	10-15 minutes
Hyper Parameter (best)	C=100	alpha=0.05	C=500

Further Exploration

- What if the dataset is not labelled?
- Can consider toxic comments as anomaly
- Train a model to fit a gaussian curve (learn sigma, mu)



References & links

Jupyter Notebook

https://colab.research.google.com/drive/1iw7JbSa_PhWz3dY6JKudbOv1bZFE_4eG?usp=sharing

Readings

<https://web.stanford.edu/~jurafsky/slp3/4.pdf>

<https://towardsdatascience.com/introduction-to-anomaly-detection-c651f38ccc32>

SKLearn's official documentation