Detecting Toxic contents in online conversations

Amal Sony, Mohd Sharique Khan, Natansh Negi,
Siddu Madhure Jayanna

Team P05

Introduction and Related Work

- Problem Statement:
  - How to facilitate online conversation ?
  - How to Handle toxic and divisive content ?
  - How to handle these online bullies ?

Solution: Classify their comments and suspend their accounts.

Introduction and Related Work

- Technology Stack:

Python + the following libraries:

- Nltk
- Wordcloud
- Sklearn
- Gensim
- Keras-Tensorflow

Introduction and Related Work

Related Work

[1] John Pavlopoulos & Prodromos Malakasiotis & Ion Androutsopoulos (2017) Deeper attention to abusive user content moderation. In EMNLP.

[2] Betty van Aken & Julian Risch & Ralf Krestel & Alexander Loser Challenges for Toxic Comment Classification: An In-Depth Error Analysis.

[3] Ellery Wulczyn & , Nithum Thain & , Lucas Dixon & Ex Machina: Personal Attacks Seen at Scale

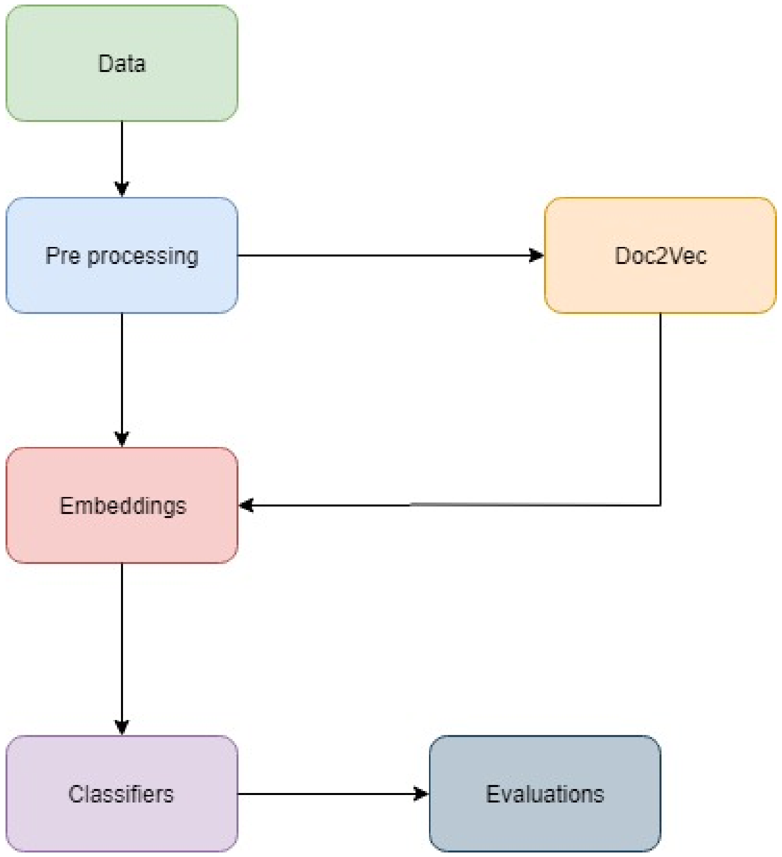
Introduction and Related work

Hypotheses

- Simple ensemble classifiers (like voting, average ,weighted average) and complex ensemble classifiers (like AdaBoost and Random Forest) would outperform shallow classifiers.
- Deep Neural Network classifiers > ensemble classifiers (simple & complex).

Method

Flow Chart



Natansh Negi (P-05)

Method

Data

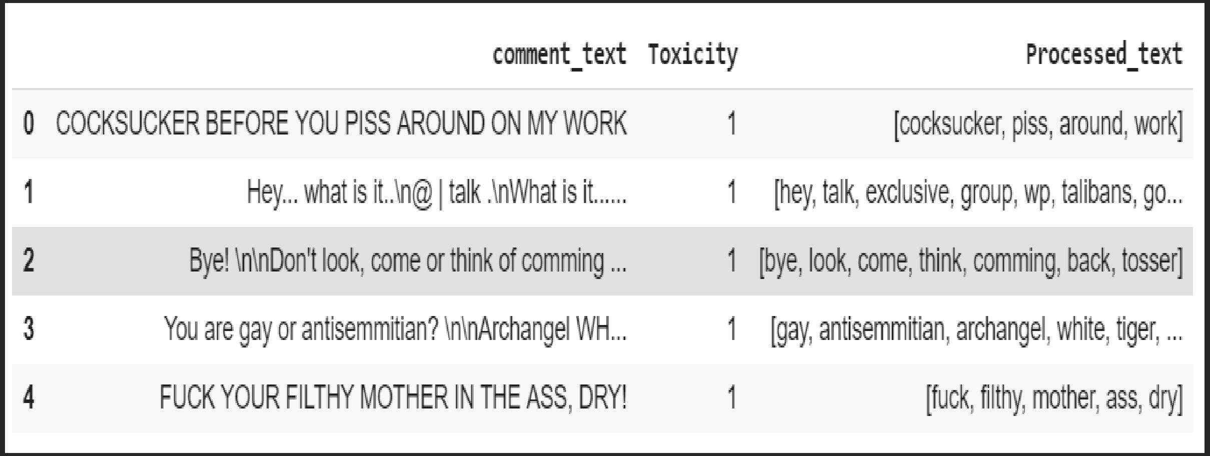
- Considered dataset of Quora and Jigsaw.
- Comprises mainly of 2 columns (**comment_text, Toxicity**)

Data Preprocessing

- Removed null data.
- Eliminated stop words.
- Tokenization & stemming
- Case, negation and punctuation handling
- Transformation
- Exploration (word Cloud)
- Sampling (stratified)

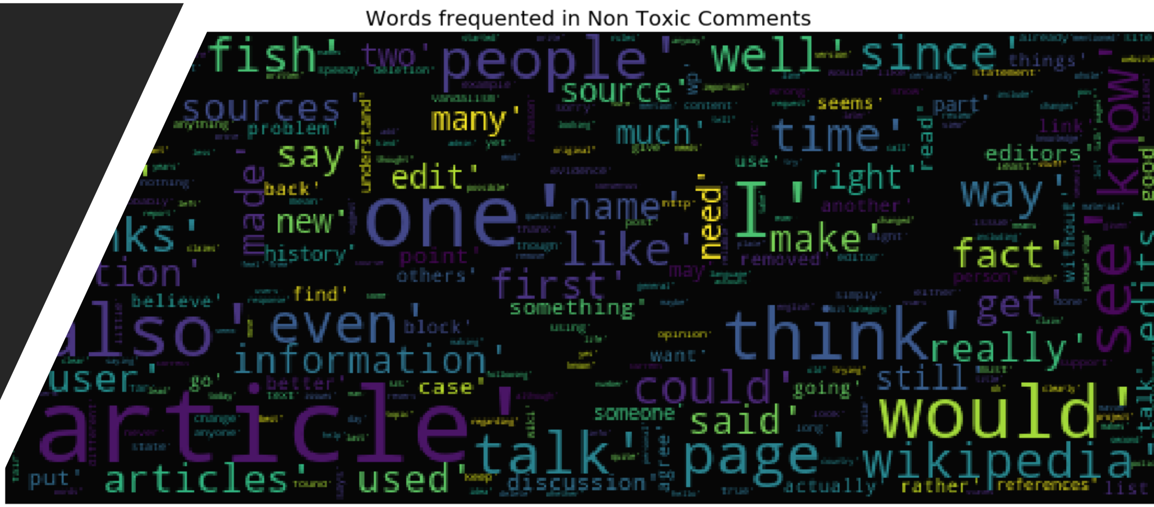
Method

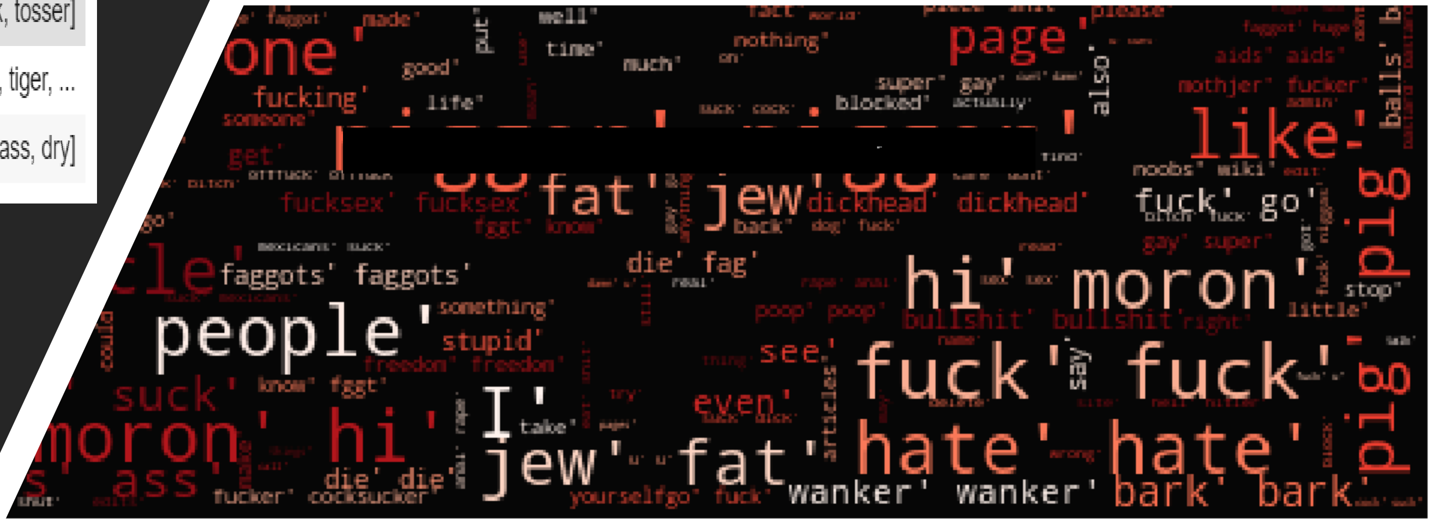
- Processed Data (left)
- Word Cloud (right)



	comment_text	Toxicity	Processed_text
0	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	[cocksucker, piss, around, work]
1	Hey... what is it..\n@ | talk .\nWhat is it......	1	[hey, talk, exclusive, group, wp, talibans, go...
2	Bye! \n\nDon't look, come or think of comming ...	1	[bye, look, come, think, comming, back, tosser]
3	You are gay or antisemmitian? \n\nArchangel WH...	1	[gay, antisemmitian, archangel, white, tiger, ...
4	FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!	1	[fuck, filthy, mother, ass, dry]

Words frequented in Non Toxic Comments



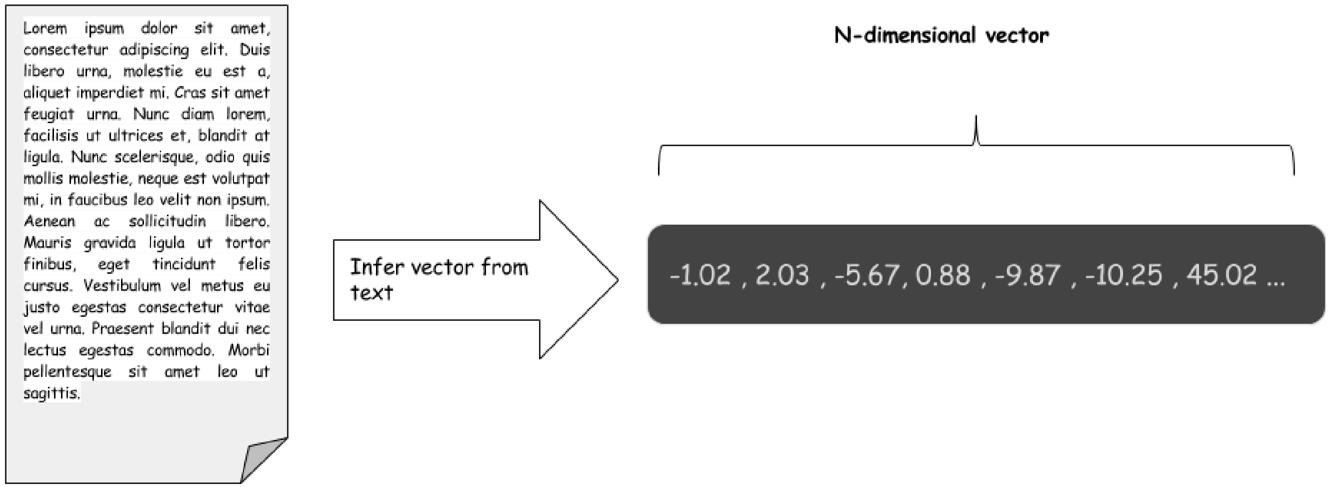


Mohd Sharique Khan (P-05)

Method

Doc2Vec

- Used to create a numeric representation of a document.
- They inherit the semantics of word vectors
- They take word order into account



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis libero urna, molestie eu est a, aliquet imperdiet mi. Cras sit amet feugiat urna. Nunc diam lorem, facilisis ut ultrices et, blandit at ligula. Nunc scelerisque, odio quis mollis molestie, neque est volutpat mi, in faucibus leo velit non ipsum. Aenean ac sollicitudin libero. Mauris gravida ligula ut tortor finibus, eget tincidunt felis cursus. Vestibulum vel metus eu justo egestas consectetur vitae vel urna. Praesent blandit dui nec lectus egestas commodo. Morbi pellentesque sit amet leo ut sagittis.

Infer vector from text

N-dimensional vector

-1.02 , 2.03 , -5.67, 0.88 , -9.87 , -10.25 , 45.02 ...

Method: Classifiers

Logistic Regression

Naïve Bayes

Decision Tree

Deep Neural Network

Ensemble simple (Logistic regression + Decision Tree)

Ensemble complex (Random Forest +Adaboost)

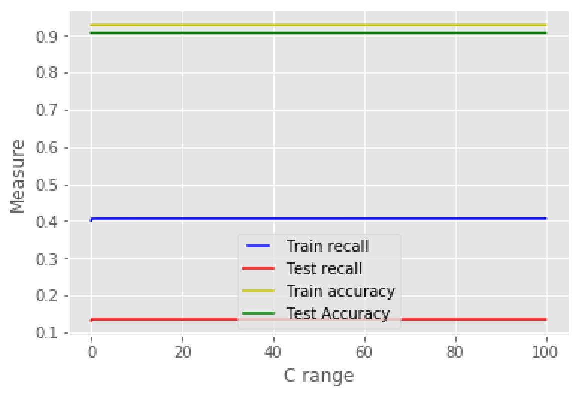
Method

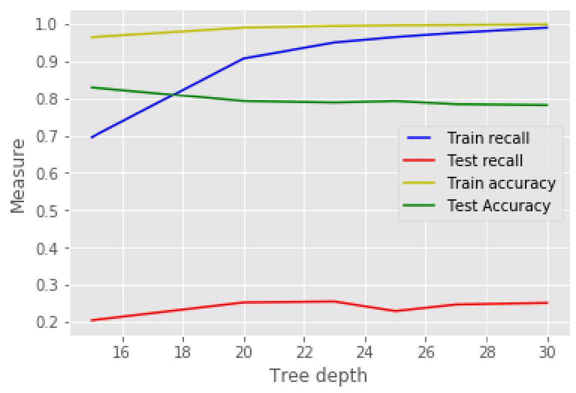
Hyperparameter approaches:

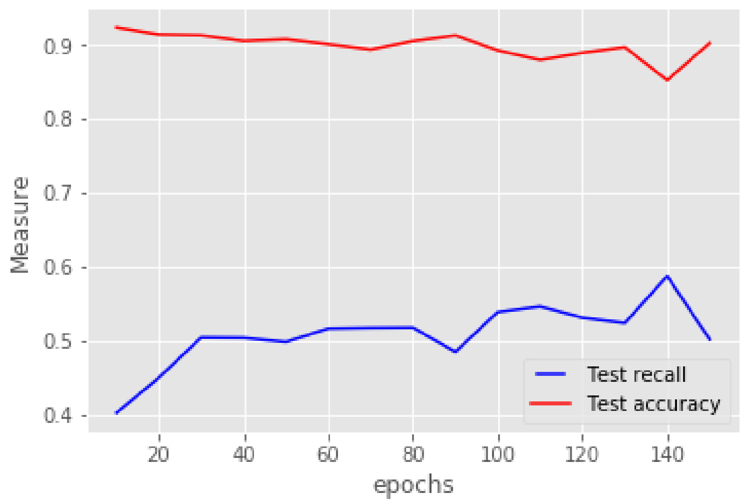
- Grid search
- Random Search

Hyperparameters tuned:

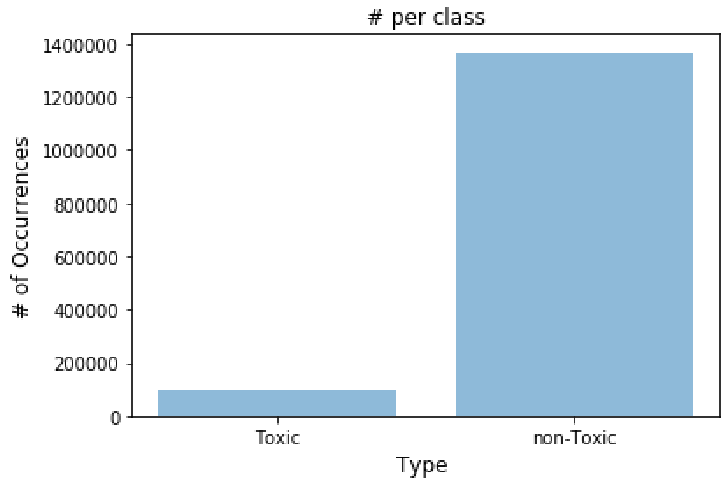
- Epochs (Deep Neural Network, Word2Vec)
- Max depth (Decision tree)
- C parameter (Logistic Regression)
- Vec_size, alpha (Word2Vec)







Results



Experiment 1:

When we used all the data from both the datasets for training and testing.

Accuracy ~ 94%

Is this good enough ?

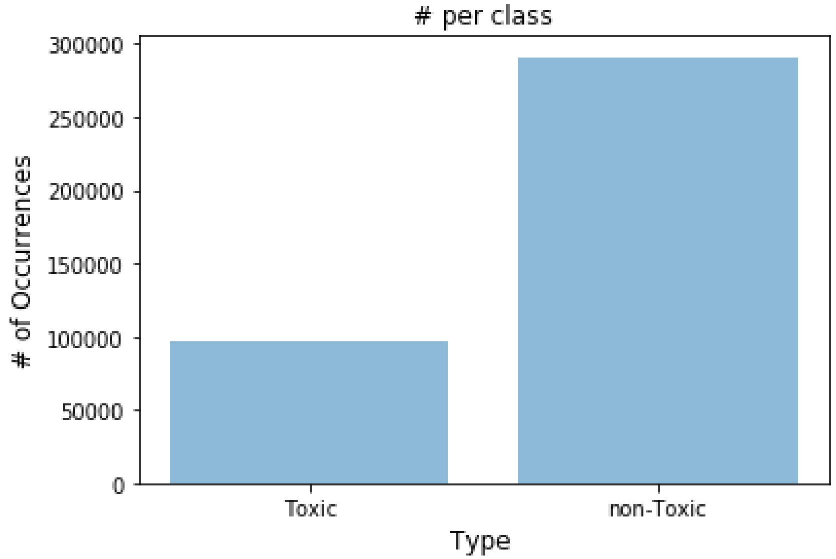
Results

Classifier	Accuracy	Precision	Recall	F-measure
Logistic	93.3%	75%	0.2%	0.4%
Naïve Bayes	92.9%	6.4%	0.4%	0.8%
Decision Tree	93.0%	NaN	0%	NaN
Ada Boost	91.32%	26%	1.7%	3.19%
Deep Neural Network	95.20%	4.2%	17.53%	6.7%
Ensemble (simple)	93.77%	67.13%	1.2%	2.35%
Ensemble (complex)	93.95%	31.65%	2.2%	4.1%

Results

- Experiment 2:

  Took 1:3 dataset for training our models



Classifier	Accuracy	Precision	Recall	F-measure
Logistic	65.18%	48.5%	71.96%	57.94%
Naïve Bayes	34.58%	33.7%	99.6%	50.3%
Decision Tree	66.66%	NaN	0%	NaN
Ada Boost	62.8%	31.11%	17.20%	22.15%
Deep Neural Network	78%	42.12%	47.33%	44.57%
Ensemble (Simple)	66.66%	NaN	0%	NaN
Ensemble (complex)	63.9%	NaN	0%	NaN

Results

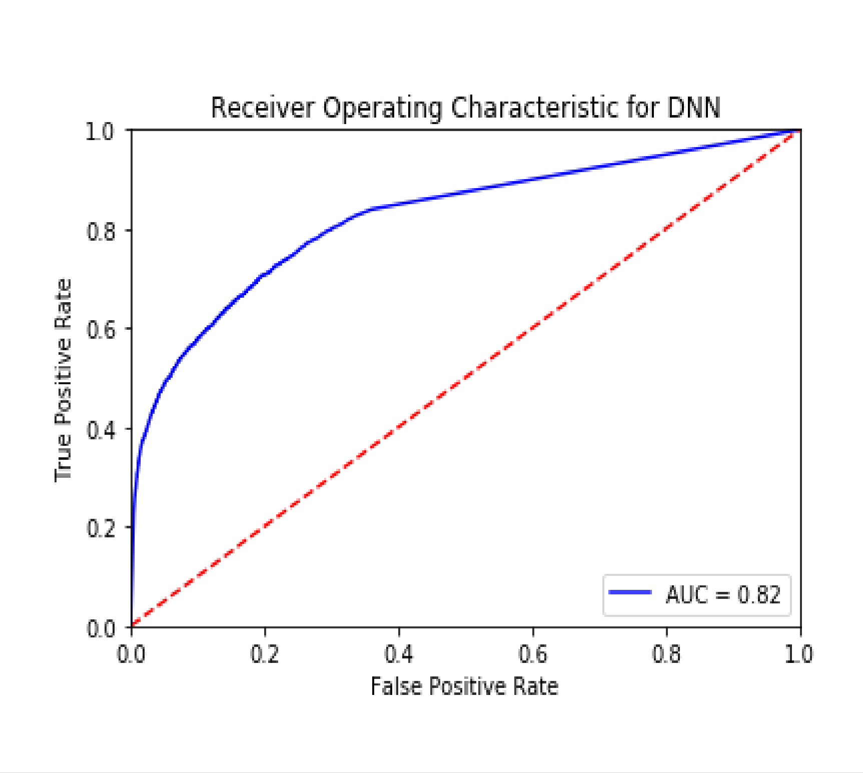
. . . . Experiment N

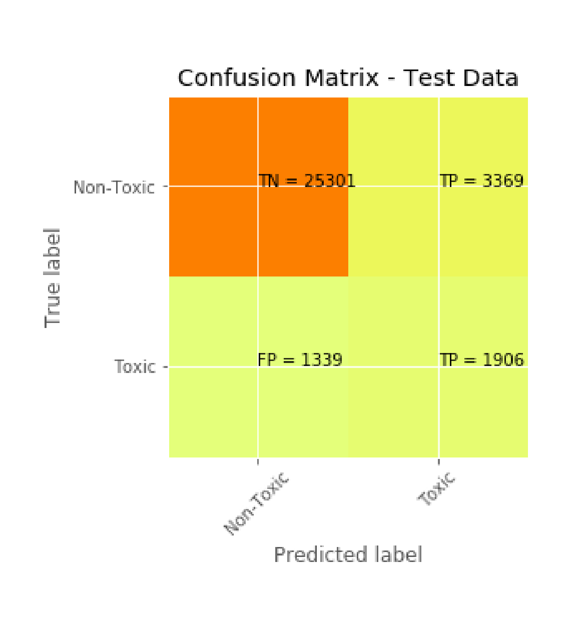
Results

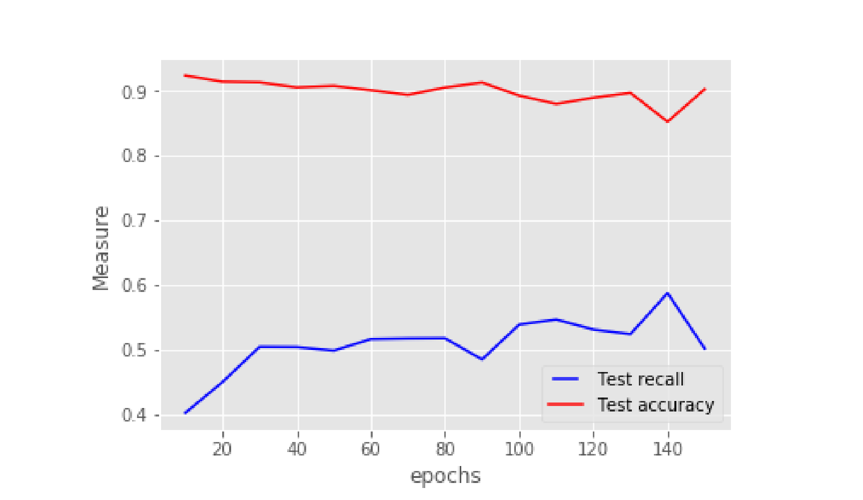
Results from our Best model (DNN)

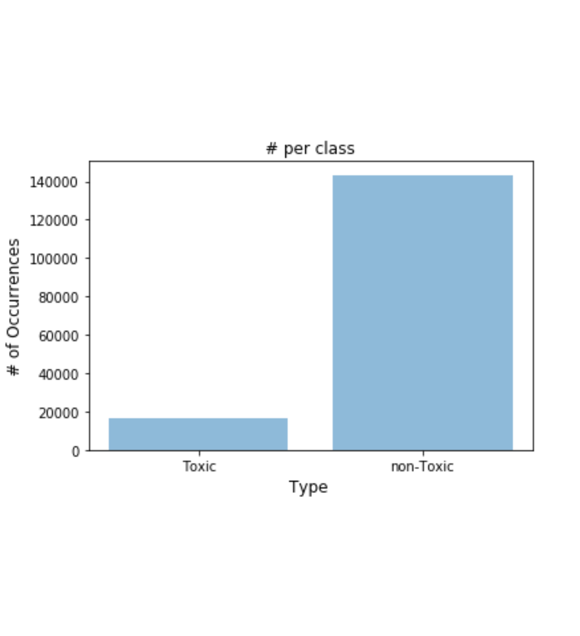
- Hyperparameters
- Confusion Matrix
- ROC and AUC

Accuracy	85.25%
Precision	36.13%
Recall	58.74%
F- Measure	44.74%
AUC	0.82









Key Learning

- Inconsistency of data plays a big role in model generation.
- Large quantities of data cannot not guarantee a good model.
- Selecting the right type of data and hyperparameters is important.
- DNN>Ensemble>shallow classifiers. (Hypotheses proved).

Future Work

We can segregate the toxic comments according to their severity. Eg(Insult , Racism, Aggression etc.)

Compare our results with any unsupervised classifier.