UNINTENDED TOXIC COMMENT CLASSIFICATION

PARTH PATEL

Who might care?

Government

Social Media Company

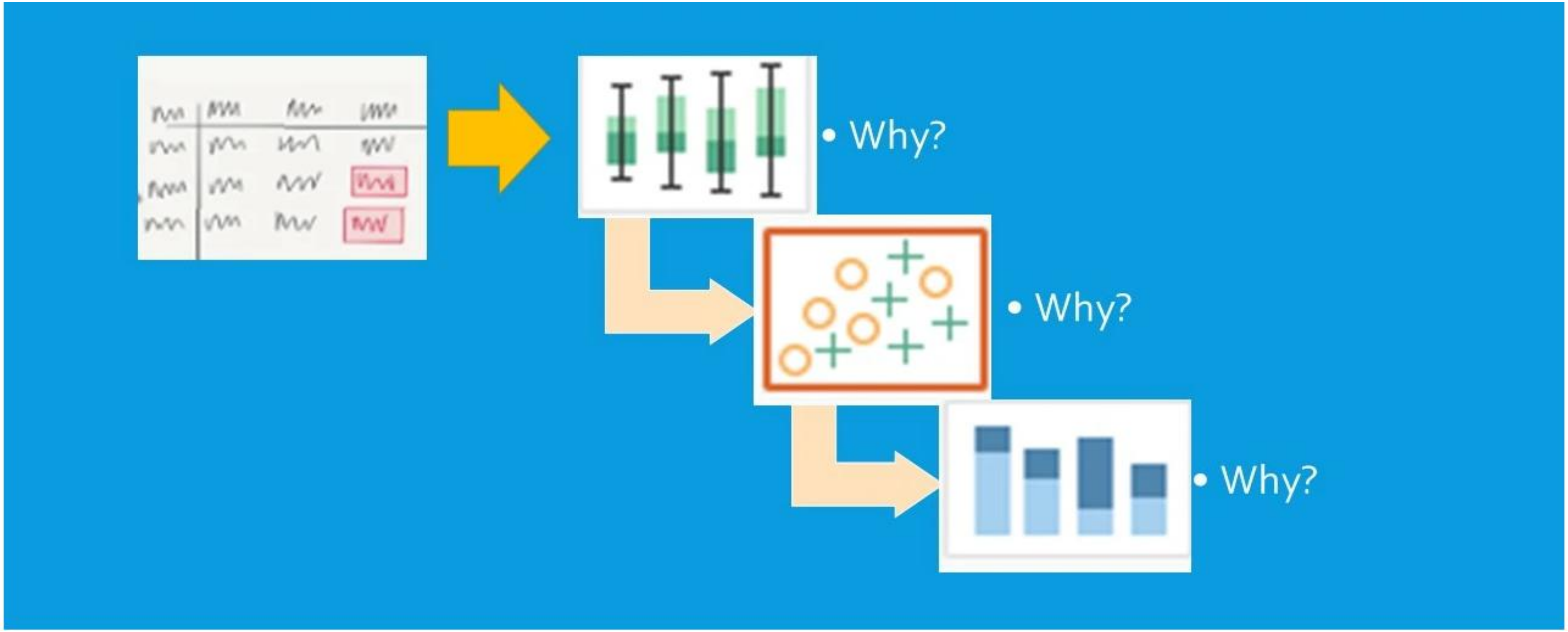
Prediction Problem

Comment	Prediction
Nonsense ? Kiss off, geek. What I said is true. I'll have your account terminated.	Toxic
Ban one side of an argument by a bullshit	Toxic
Nazi admin and you get no discussion	Toxic
You are acting like a gay	Toxic
Why can you put English for example on some players but other people don't like it – why ?	Non-Toxic
I am gay Women	Toxic

Data Overview

- Data set obtained from Kaggle
- Number of rows ~2M
- Column Description
  - Comments
  - 5 toxicity Type
  - 24 Attribute describing ethnicity and religion , Gender, Sexual Orientation and Disability

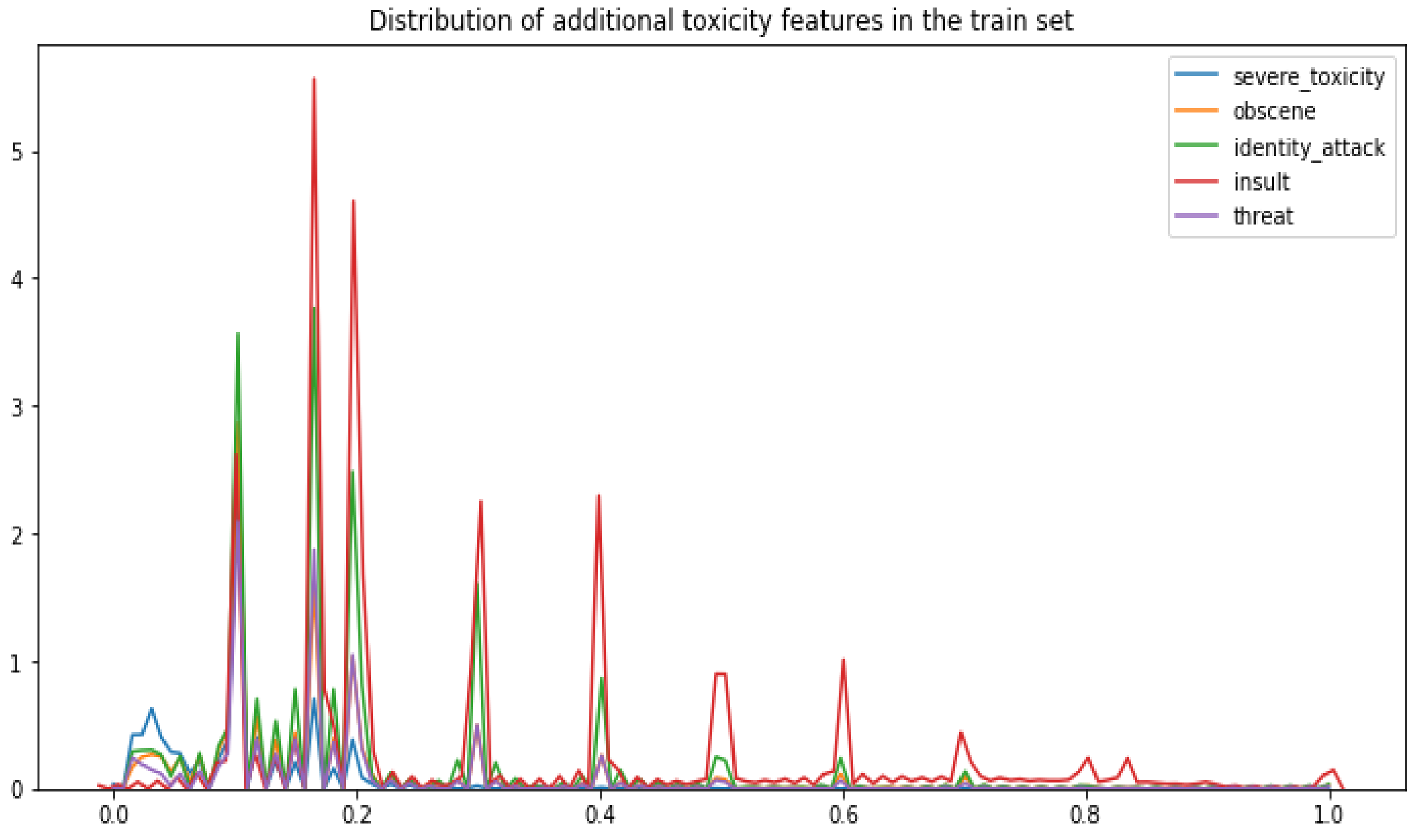
	id	target	comment_text	severe_toxicity	obscene	identity_attack	insult	threat	asian	atheist
0	59848	0.000000	This is so cool. It's like, 'would you want yo...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN
1	59849	0.000000	Thank you!! This would make my life a lot less...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN
2	59852	0.000000	This is such an urgent design problem; kudos t...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN
3	59855	0.000000	Is this something I'll be able to install on m...	0.000000	0.0	0.000000	0.00000	0.0	NaN	NaN



Exploratory Data Analysis

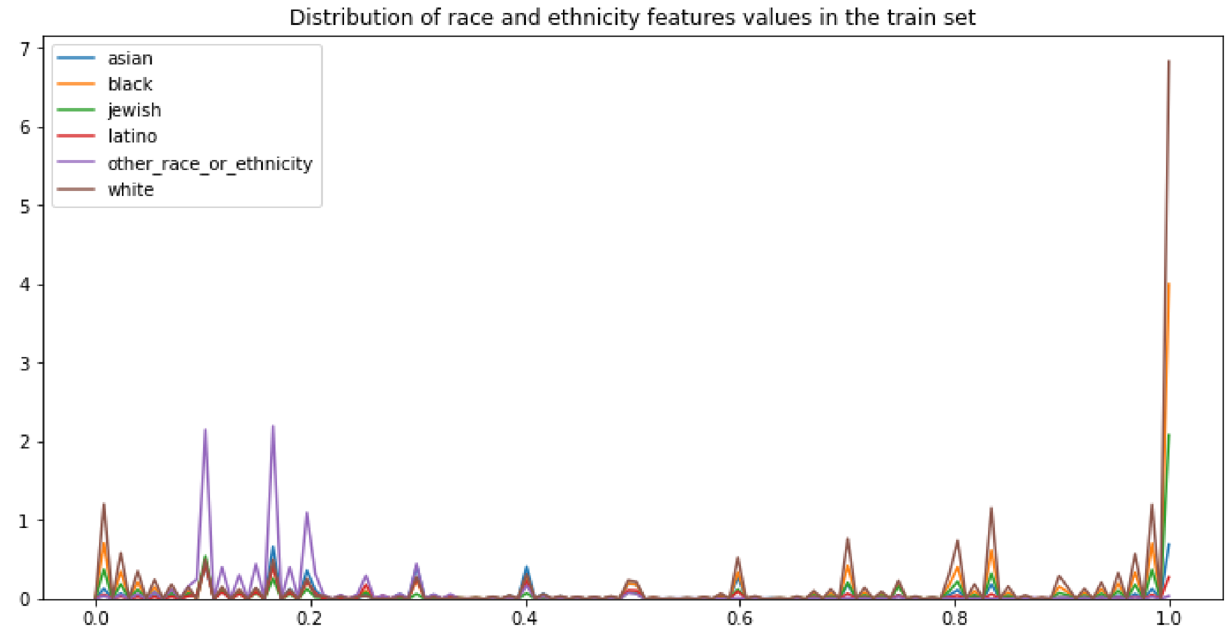
Toxicity Feature Distribution

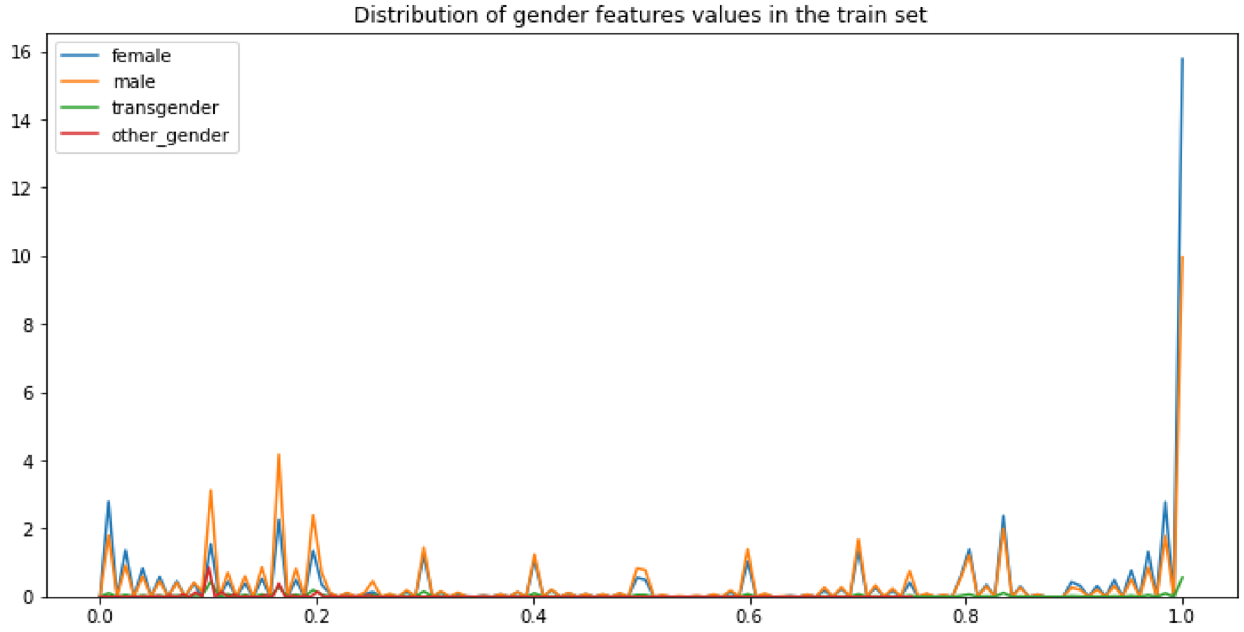
- Type of comment(Toxicity feature) does not have any direct relation with toxicity
- All types do have toxic as well as non-toxic comments
- More insulting comments compared to other toxicity type



Distribution Plot based on and Ethnicity and Gender

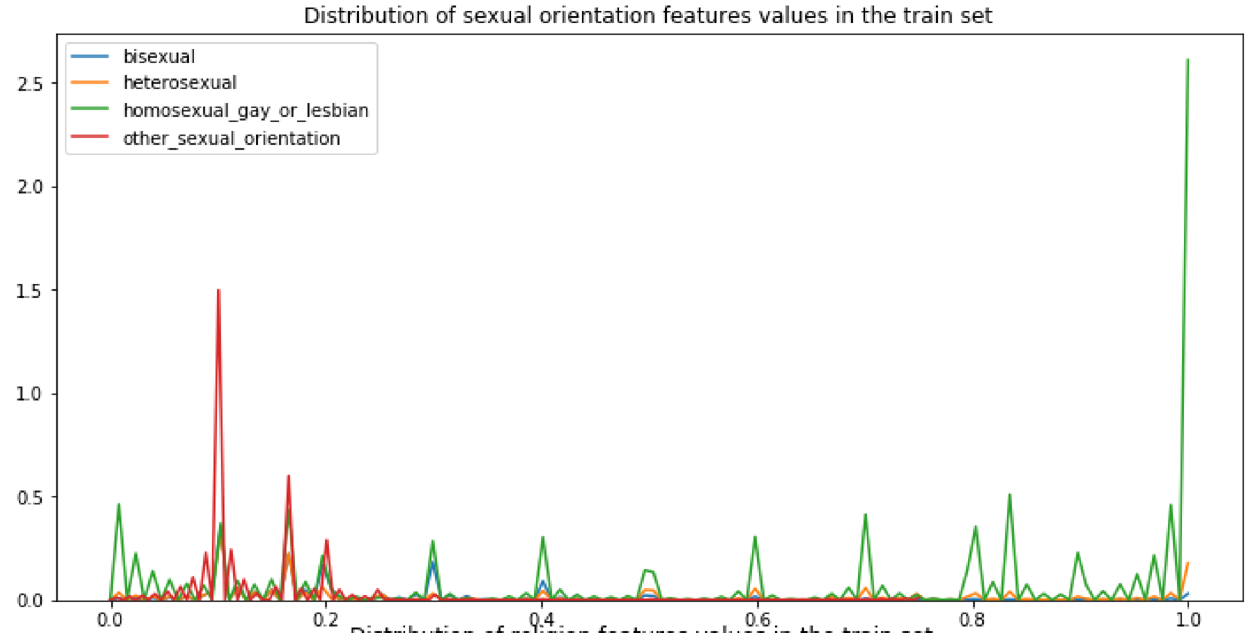
- Comments with other races or ethnicity category were non-toxic
- Comments with White and black race had many toxic comments compared to non- toxic
- comments mentioning female have more probability being toxic compared to non-toxic

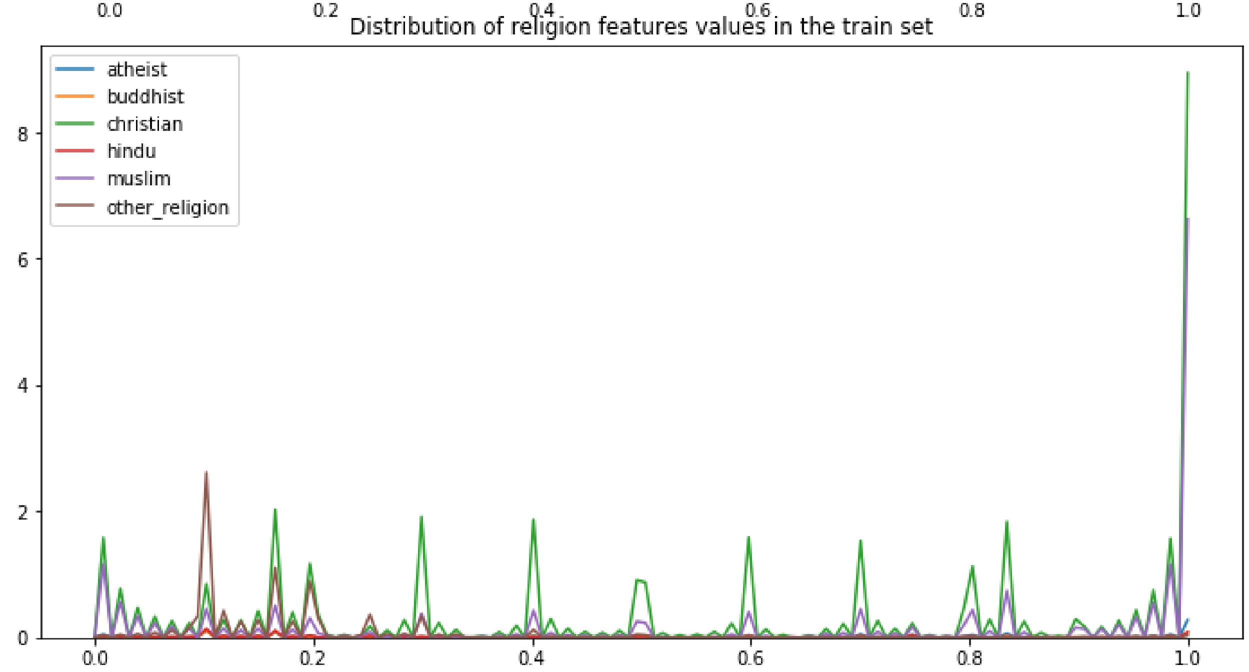




Distribution Plot based on Sexual Orientation and Religion

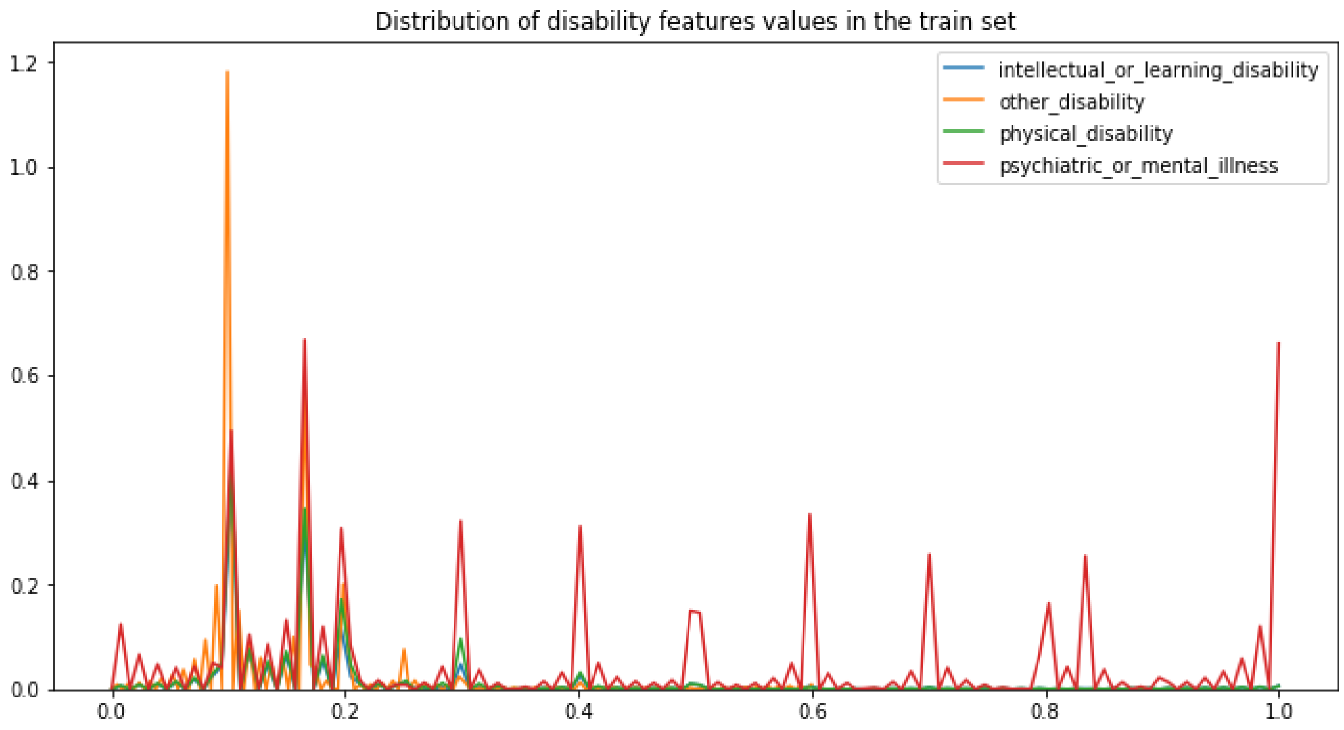
- Comments towards Homosexual people were mostly toxic
- Comments on Christian and Muslim has more probability being toxic
- Comments towards Hindus and Buddhist or Other religion are usually non-toxic





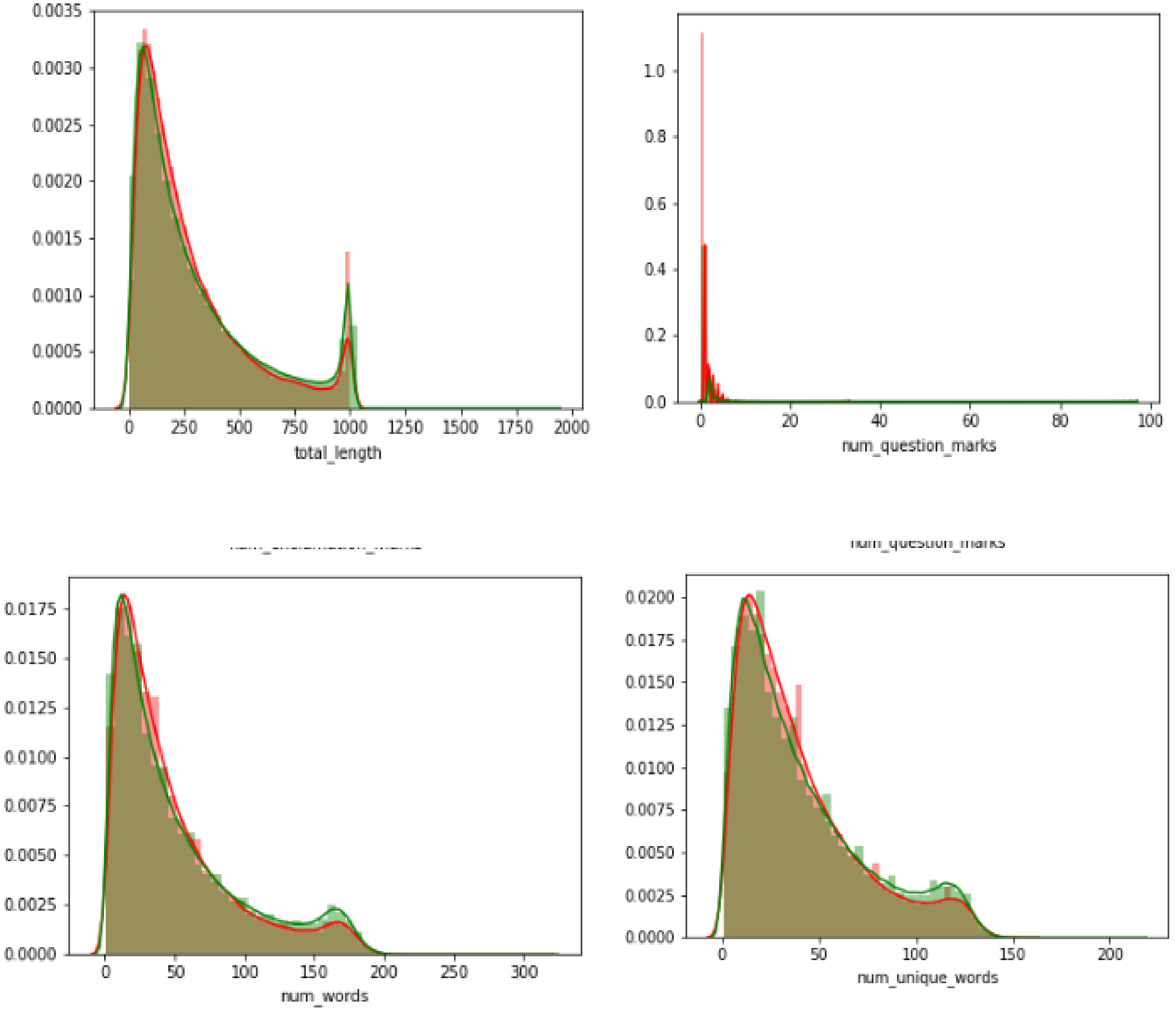
Distribution Plot based on Disability and other general observation

- Comments towards Mentally disabled people were mostly toxic
- Overall we did have some significant patterns which can be used as feature but number of observations with those features were so in-significant that we had to drop the idea of using them as features



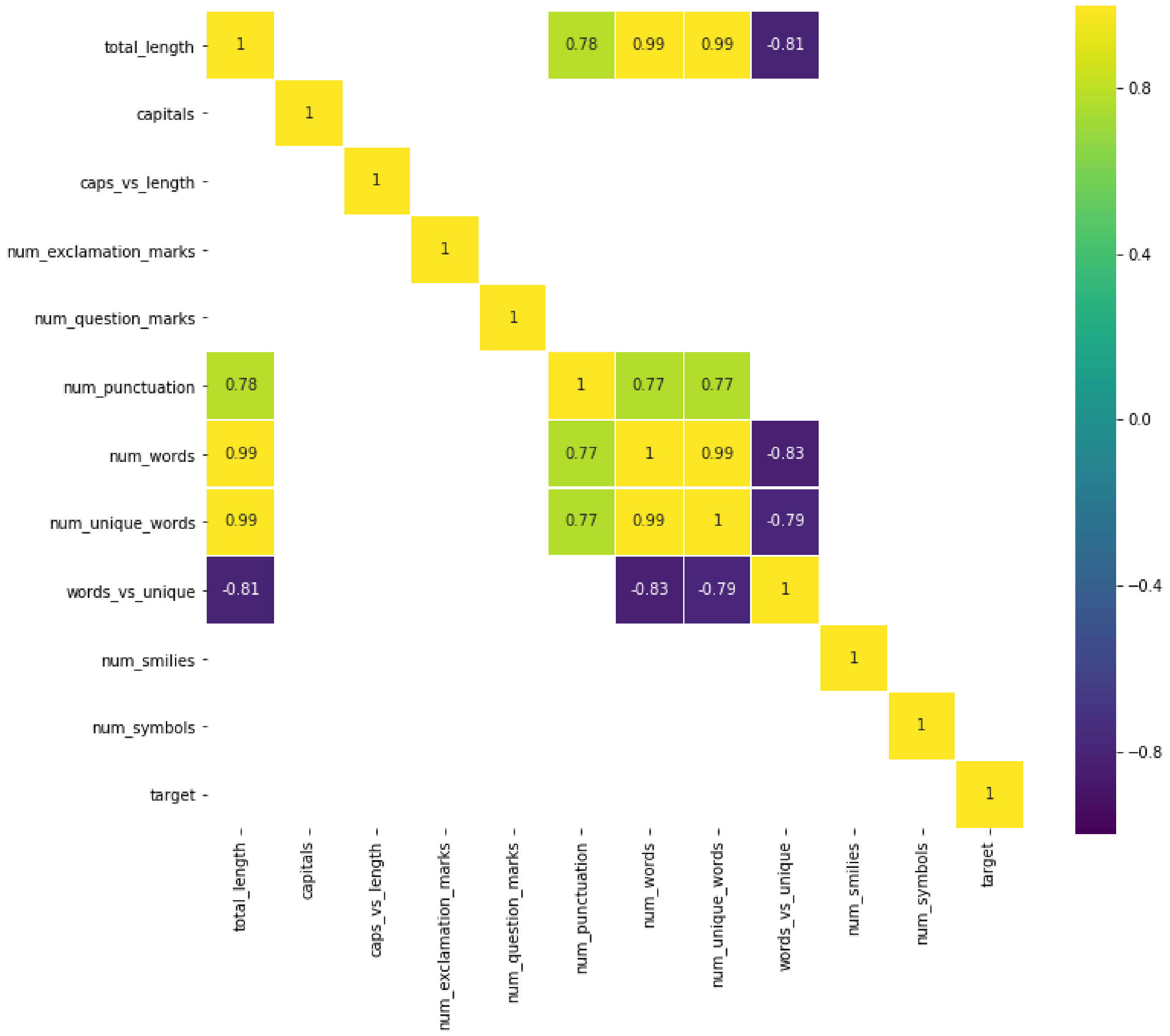
Distribution Plot of New Feature

- Comments with question marks are toxics
- No of Unique Words Almost equal to number of Words



Co Relation between new variable and Target variable

- none of new feature have high co-relation with target
- number of words and number of unique words are very highly corelated
- we do have positive correlation with number of punctuation and number of words

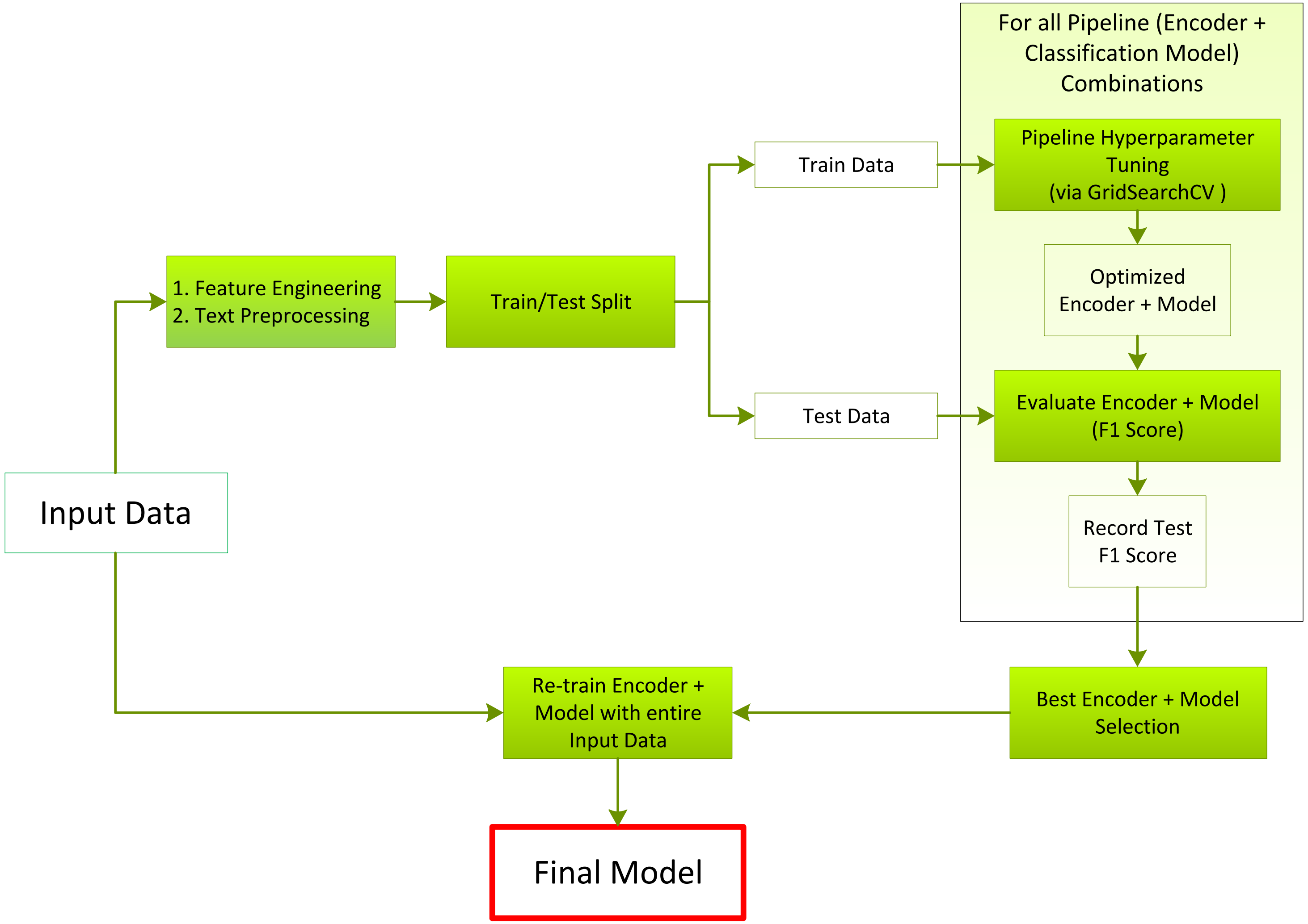


Word Cloud [Toxic–NonToxic]

Modeling

Modeling Overview

- Type: Supervised Learning
- Pipeline consists of
  - Encoder
  - Classification Model



Raw	Lowercased
Canada CanadA CANADA	canada

Lower casing

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose

Lemmatization

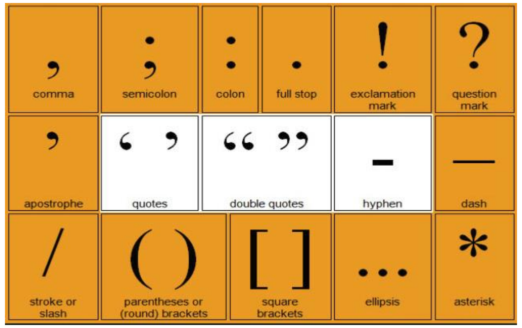
Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Stop words removal

Text Preprocessing

rating_description	no_contract
always something to do here at work just wishe...	[always, something, to, do, here, at, work, ju...
Working here has been a great opportunity for ...	[Working, here, has, been, a, great, opportuni...
Great place to work. Would love to work here a...	[Great, place, to, work., Would, love, to, wor...
Great experience and worked with alot of great...	[Great, experience, and, worked, with, alot, o...
learned a lot also got to know a lot of cool p...	[learned, a, lot, also, got, to, know, a, lot,...
Everyday I get to learn new stuff. I am really...	[Everyday, I, get, to, learn, new, stuff., I, ...
Pick your role carefully so that you don't get...	[Pick, your, role, carefully, so, that, you, d...
As a contractor, I did not have much structure...	[As, a, contractor,, I, did, not, have, much, ...
I've been wanting to work at [illegible]	[I have, been, wanting, to, work, at, google,...
Lot of creative freedom. Always something to w...	[Lot, of, creative, freedom., Always, somethin...

Remove Contraction



Remove Punctuations

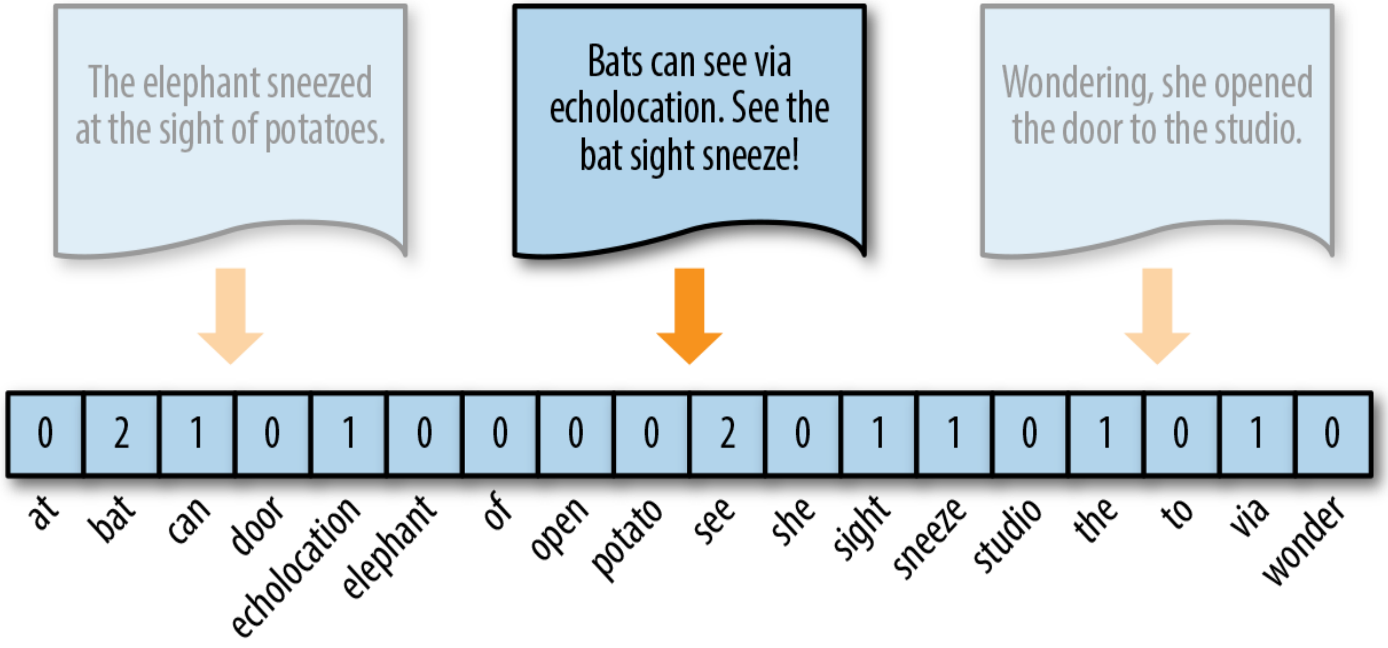
```
def keep_alpha(sentence):
    alpha_sentence = re.sub('[^a-z A-Z]+', ' ', sentence)
    return alpha_sentence
```

Keep only alphabetic strings

Text Preprocessing

Text Encoding Techniques

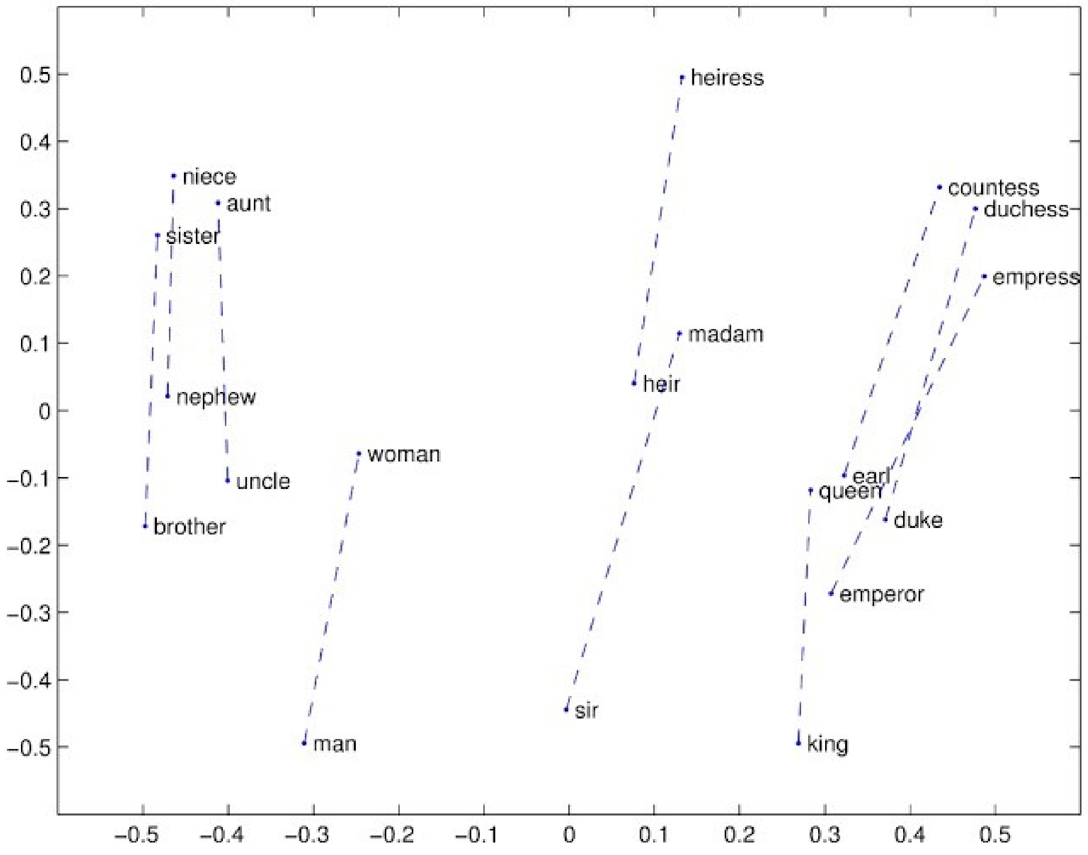
- Bag of Word
- TF IDF



Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0

Text Encoding Techniques

- GloVe
- Fast Text



Count Vectorizer	Ngram = (1, 2)	min_df =0.01	analyzer="word"
Logistic Regression	C=1.4	max_iter= 10	

Countvectorizer	Ngram = (1, 2)	min_df =0.01	analyzer="word"
Random Forest	max_depth=3	min_samples_split =10	min_samples_leaf=7

Count Vectorizer	Ngram = (1, 2)	min_df =0.01	analyzer="word"
Dummy Classifier	strategy="most_frequent"		

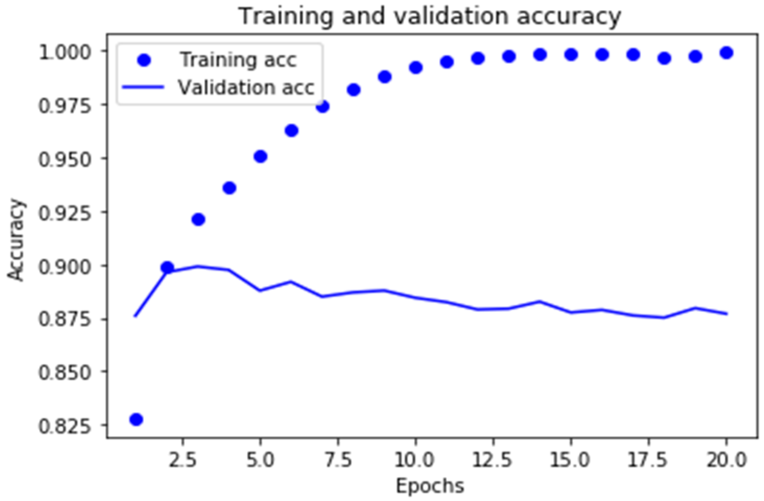
TFIDF	Ngram = (1, 1)	min_df =1	analyzer="word"
SVM Classifier	max_iter': 100	kernel': 'rbf'	C= 1.1

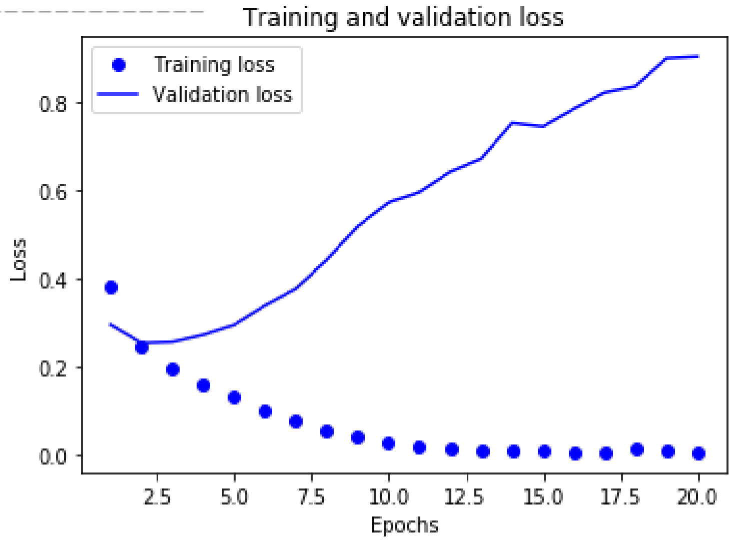
Count Vectorizer	Ngram = (1, 2)	min_df =0.01	analyzer="word"
Bernoulli Naive Bayes	alpha=100		

TF IDFVectorizer	Ngram = (1, 1)
Bernoulli Naive Bayes	alpha= 0.03

Scikit Learn Model Design

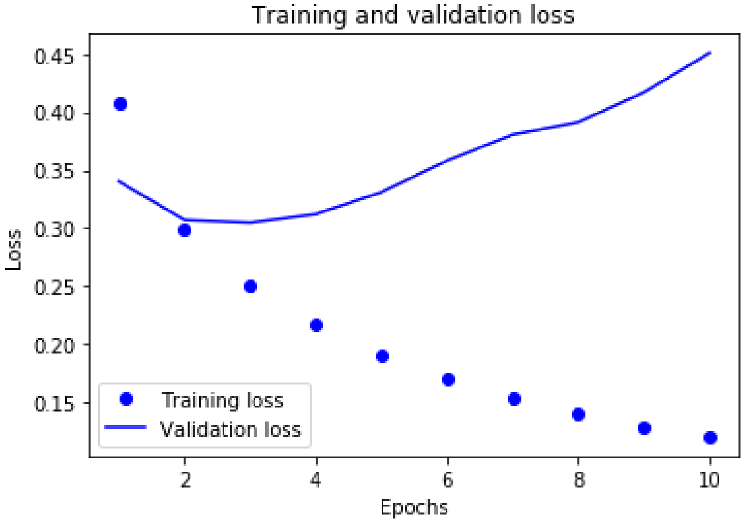
Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 150, 100)	9000000
flatten_1 (Flatten)	(None, 15000)	0
dense_1 (Dense)	(None, 32)	480032
dense_2 (Dense)	(None, 1)	33

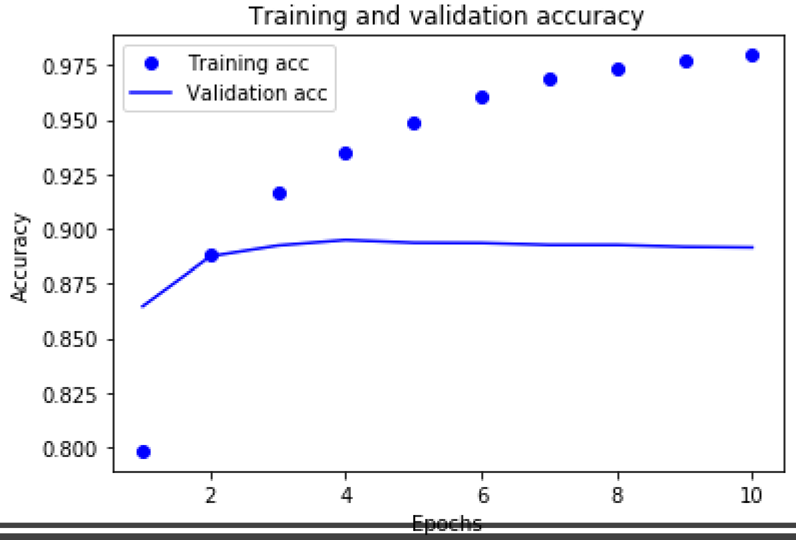




GloVe Model Design

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
embedding_1 (Embedding)      (None, 150, 300)          27000300  
_________________________________________________________________
flatten_1 (Flatten)          (None, 45000)             0         
_________________________________________________________________
dense_1 (Dense)              (None, 32)                1440032   
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 33        
=================================================================
```

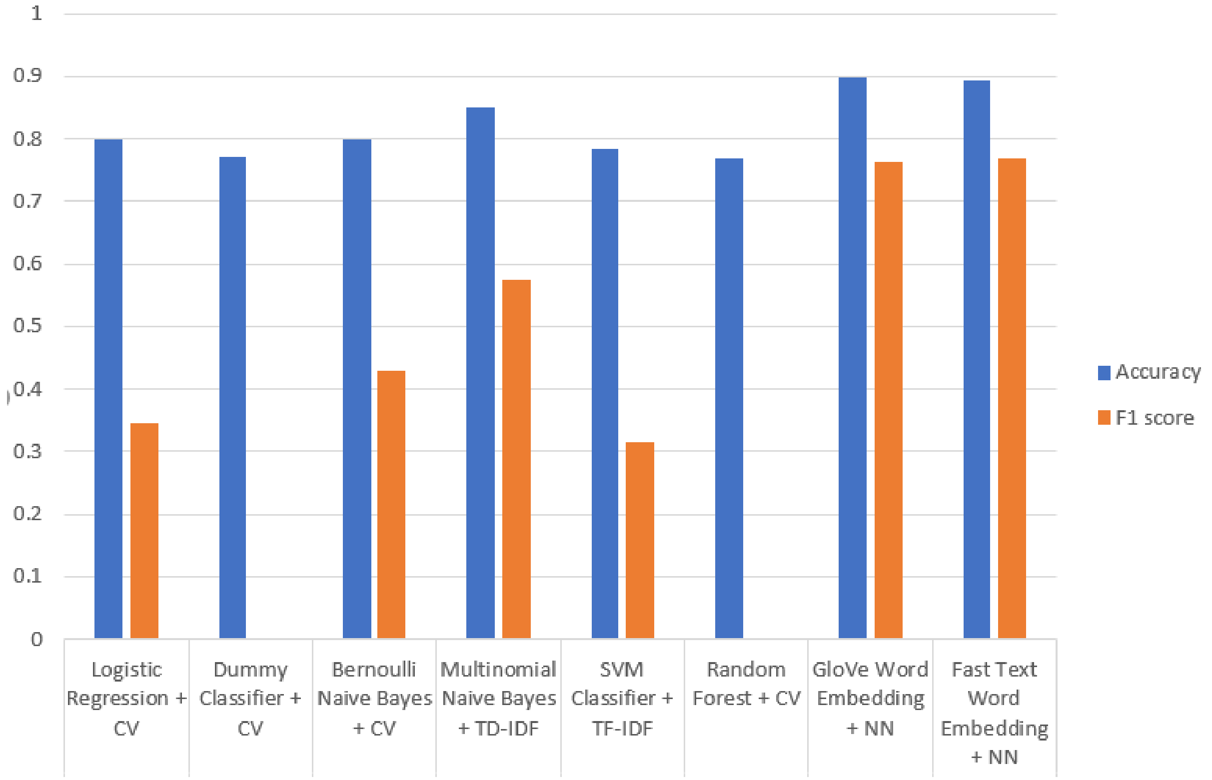




Fast Text Model Design

Model Summary

- Neural network does perform much better than count or TF-IDF vectors
- TF-IDF vectors are better than Count word vectors
- Most algorithm using count vector try to predict same result as non-toxic comment
- Glove Word and Fast Text has much better prediction accuracy with F1-score indicating most toxic comments are identified correctly also keeping non-toxic separate



- We had very imbalance dataset gathering more toxic comments would help improve our model.
- Due to limiting processing power we had removed certain nontoxic comments
- TF-IDF vectors were very large in dimension by performing dimensionality reduction on those vectors will improve our execution process and time of execution.
- Current model with word embedding are built using simple one hidden layer of neural network , this can definitely be improved by adding more layers like Convolutional and pooling which can help to extract more feature
- Web app can be created where users can enter text comments and predicted toxicity can be displayed
- API can be created for user so that toxic comment can be classified and flagged as soon as they are entered which can help websites or application block or restrict unwanted content.

Limitations and Ideas

Thank you