Capstone Project

Content Moderation (Trust & Safety)

NLP Multi-Label Classification

Lo Kok Fu
DSIF 2

INTRODUCTION

- On the online space, there are multiple platforms where users are able to generate and post text contents as they deem fit. With these ease of access, content violations like violence, explicit, cyber bullying or racial hate contents are constantly on the rise.

- *Natural Language Processing*
- *Multi-Label Classification*

***DISCLAIMER**: THIS PROJECT CONTAINS TEXTS WITH EXPLICIT OR COARSE LANGUAGE / CONTENT. VIEW AT YOUR OWN DISCRETION*

CONTENTS

Problem Statement

- Toxic contents may be also communicated to vulnerable groups such as the minors or racially sensitive groups. In which, would incite violent , hate behaviors or suicide tendencies.

- Therefore there is a need to protect vulnerable groups from such toxic comments through filtering or surfacing at large scale for safe content viewership.

Image ref: https://www.socialsamosa.com/2019/07/instagram-account-disabling-content-violation-polices/

Solution

- Develop an online content NLP Machine Learning Algorithm to detect user generated toxic words and classify them for surfacing to platform censorship processes with accordance to community guidelines violations policies for online user text content enabled generation platforms.

Data Sets

- Extracted from Kaggle
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
- 159,571 rows
- 6 Target columns
- ‘Toxic’, ‘Severe Toxic’, ‘Obscene’, ‘Threat’, ‘Insult’, ‘Identity Hate’

Methodology

1. Data Cleaning

Clean Function

Removed regular expressions

Lower cased

0
Duplicated Data

0
Missing Data

Created additional Column

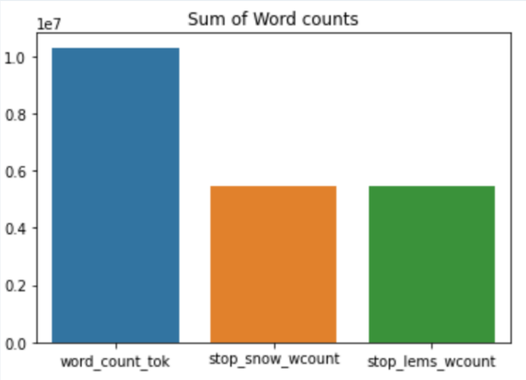
“Safe” column = 1 (Positive) where all target value = 0 (Negative)

Comment Text rows	Target Columns
159, 571	2 unique labels

2. Text Preprocessing

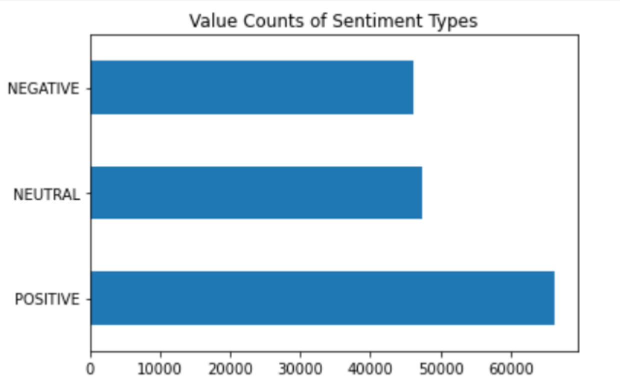
Cleaned Text › Tokenize › Snowball Stemming / Lematization

Column	Word count
word_count_tok	10299557
stop_snow_wcount	5469019
stop_lems_wcount	5460040



2. Text Pre-processing (continued)

Sentiment Analysis (Vader) › Compound Score › Compound Type › Positive >= 0.25
Neutral > -0.25 and <0.25
Negative <= -0.25



comment_text	senti_scores	compound	sentiment_type	tokens
why the edits made under my username hardcore ...	{'neg': 0.0, 'neu': 0.892, 'pos': 0.108, 'comp...	0.5574	POSITIVE	[why, the, edits, made, under, my, username, h...
daww he matches this background colour im seem...	{'neg': 0.118, 'neu': 0.71, 'pos': 0.172, 'com...	0.2263	NEUTRAL	[daww, he, matches, this, background, colour, ...
hey man im really not trying to edit war its j...	{'neg': 0.083, 'neu': 0.849, 'pos': 0.068, 'co...	-0.1779	NEUTRAL	[hey, man, im, really, not, trying, to, edit, ...
i cant make any real suggestions on improvemen...	{'neg': 0.044, 'neu': 0.893, 'pos': 0.063, 'co...	0.2500	POSITIVE	[i, cant, make, any, real, suggestions, on, im...
you sir are my hero any chance you remember wh...	{'neg': 0.0, 'neu': 0.663, 'pos': 0.337, 'comp...	0.6808	POSITIVE	[you, sir, are, my, hero, any, chance, you, re...

2. Text Pre-processing (continued)

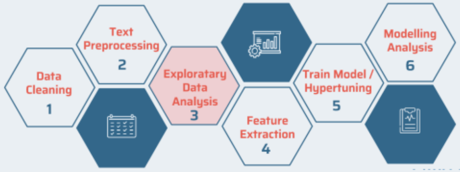
20	000b08c464718505	regarding your recent edits once again please read wpfilmplot before editing any more film articles your edits are simply not good with entirely too many unnecessary details and very bad writing please stop before you do further damage the	{'neg': 0.236, 'neu': 0.671, 'pos': 0.093, 'compound': -0.7905}	-0.7905	NEGATIVE	[regarding, your, recent, edits, once, again, please, read, wpfilmplot, before, editing, any, more, film, articles, your, edits, are, simply, not, good, with, entirely, too, many, unnecessary, details, and, very, bad, writing, please, stop, before, you, do, further, damage, the]

12	0005c987bdfc9d4b	hey what is it talk what is it an exclusive group of some wp talibanswho are good at destroying selfappointed purist who gang up any one who asks them questions abt their antisocial and destructive noncontribution at wpask sityush to clean up his behavior than issue me nonsensical warnings	{'neg': 0.161, 'neu': 0.69, 'pos': 0.149, 'compound': -0.4019}	-0.4019	NEGATIVE	[hey, what, is, it, talk, what, is, it, an, exclusive, group, of, some, wp, talibanswho, are, good, at, destroying, selfappointed, purist, who, gang, up, any, one, who, asks, them, questions, abt, their, antisocial, and, destructive, noncontribution, at, wpask, sityush, to, clean, up, his, behavior, than, issue, me, nonsensical, warnings]

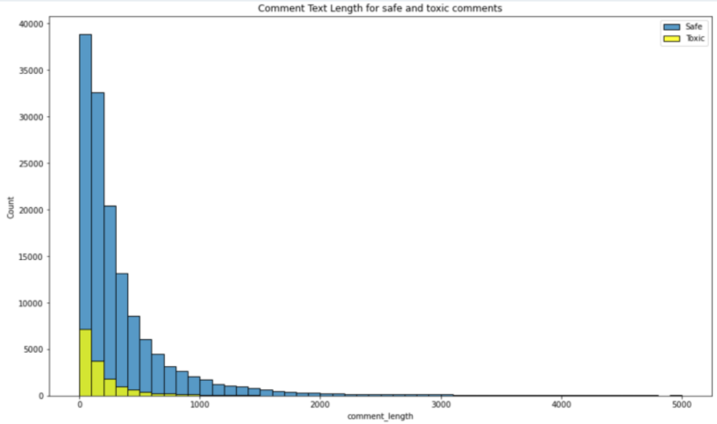
Comments Safe or not safe?

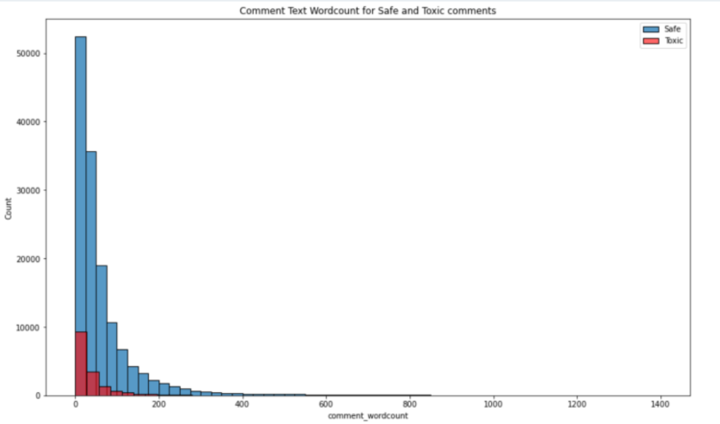
330	00d429d337eaa672	god is deadi dont mean to startle anyone but god is dead we should not worry about him anymore just thought i would let everyone know well goodbye and good luck with your newfound crisis of faith	{'neg': 0.191, 'neu': 0.429, 'pos': 0.379, 'compound': 0.8121}	0.8121	POSITIVE

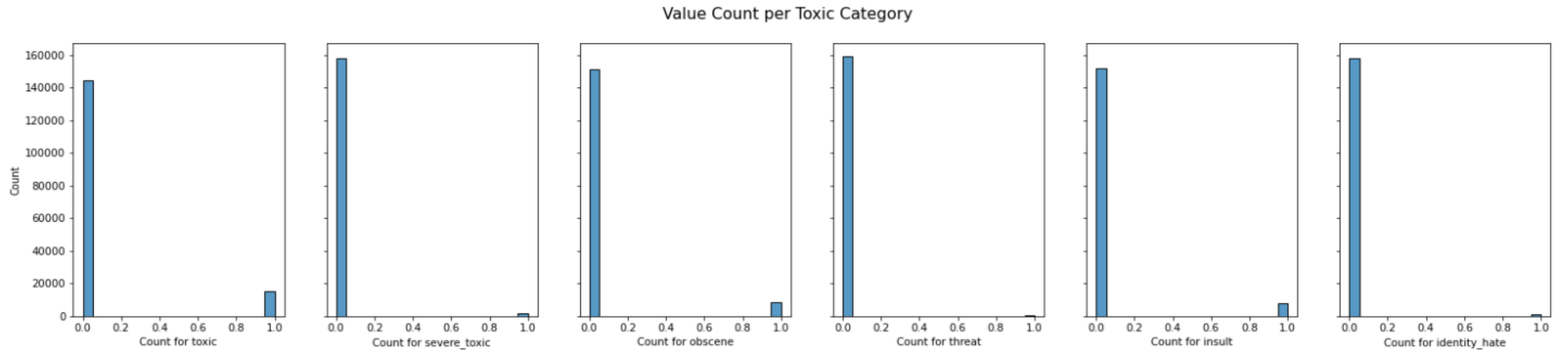
5	00025465d4725e87	congratulations from me as well use the tools well · talk	{'neg': 0.0, 'neu': 0.464, 'pos': 0.536, 'compound': 0.7964}	0.7964	POSITIVE



3. Exploratory Data Analysis

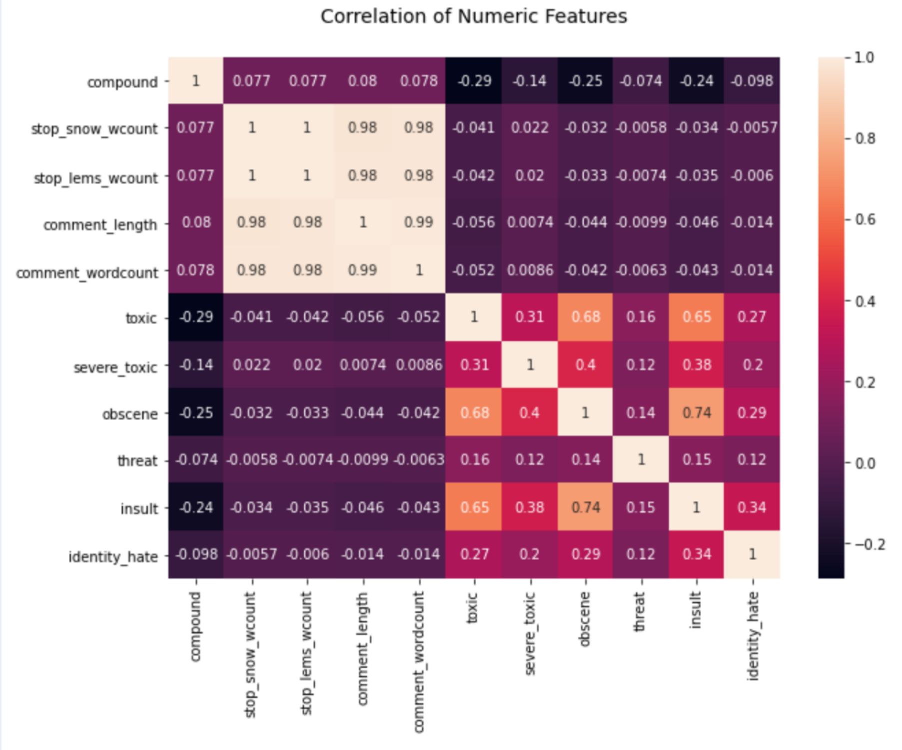




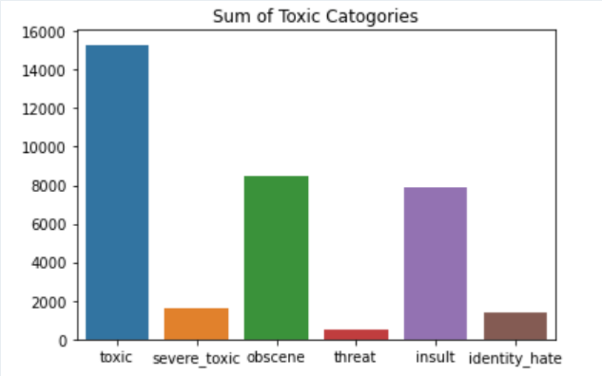


- Text Length and counts skewed towards to the left
- Data imbalanced

3. Exploratory Data Analysis



- Toxic comments are moderately co-related to obscene and insult



3. Exploratory Data Analysis

4. Feature Extraction

TF-IDF Vectoriser :

- Focuses on the frequency of words present in the corpus but also provides the importance of the words.
- Remove the words that are less important

N grams

Unigrams and Bigrams

Max Features

8,000

5. Train Model / Hypertuning

Data Imbalance

Treat with class weight = 'balance'

```
Test Accuracy Score of Logistic Reg.: 0.9154989597172436
Precision : 0.9463262238090764
Recall    : 0.9018602672875019
F1-score   : 0.9235583370585606
```

Logistic Regression Scores without treating data imbalance

One Vs Rest Classifier Logistic Regression

uses the binary relevance method to perform multilabel classification, which involves training one binary classifier independently for each label

Scores

Model	Test Accuracy	Precision	Recall	F1	Train Score	Test Score
Logistic Regression	0.854310	0.819183	0.907188	0.860943	0.856314	0.854310
Random Forrest Classifier	0.898253	0.898253	0.802176	0.847500	0.898344	0.898253
Decision Tree Classifier	0.763969	0.637818	0.783797	0.703312	0.985352	0.763969

GridsearchCV

Random Forrest Classifier

Logistic Regression

6. Model Analysis

0.866315
Train: 0.856314
Test: 0.854310

Logistic Regression

0.847500
Train: 0.898344
Test: 0.898253

Random Forrest Classifer

**Scoring based on F1 Score*

Conclusion

Regular Expressions

Deep Learning

Deployment

Sentiment Analysis

Thank You