Toxic Comment Classification

Analysis and Model Generation using Python

Lakshveer Singh
Laksh.9258@gmail.com

Contents

1 **Introduction**

1.1 Background

Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) are working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far, they have built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

1.2 Problem Statement

Build a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate better than Perspective's current models, using a dataset of comments from Wikipedia's talk page edits. Improvements to the current model will hopefully help online discussion become more productive and respectful.

Exploring and Analyzing Data

2.1 Data

Top 5 rows of the train dataset

id	comment_text	toxic	severe_toxic	obscene	threat	Insult	identity_hate
0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

2.2 Data Size and Structure

The training set consist on 159571 rows and 8 rows.

Variable Name	Data Type
id	object
comment_text	object
toxic	int64
severe_toxic	int64
obscene	int64
threat	int64
insult	int64
identity_hate	int64

A description of each variable, including description of some key values, is given below.

Variable Name	Description	Key
id	Id of the person who has commented	
comment_text	Comment_text by the person	
toxic	Is the comment toxic?	0 = No, 1 = Yes
severe_toxic	Is the comment severe_toxic?	0 = No, 1 = Yes
obscene	Is the comment obscene?	0 = No, 1 = Yes
threat	Is the comment threat?	0 = No, 1 = Yes
insult	Is the comment insult?	0 = No, 1 = Yes
identity_hate	Is the comment identity_hate?	0 = No, 1 = Yes

Converted toxic, severe_toxic, obscene, threat, insult, identity_hate into factor/category type and comment_text into string data type.

Percentage of train and test data:

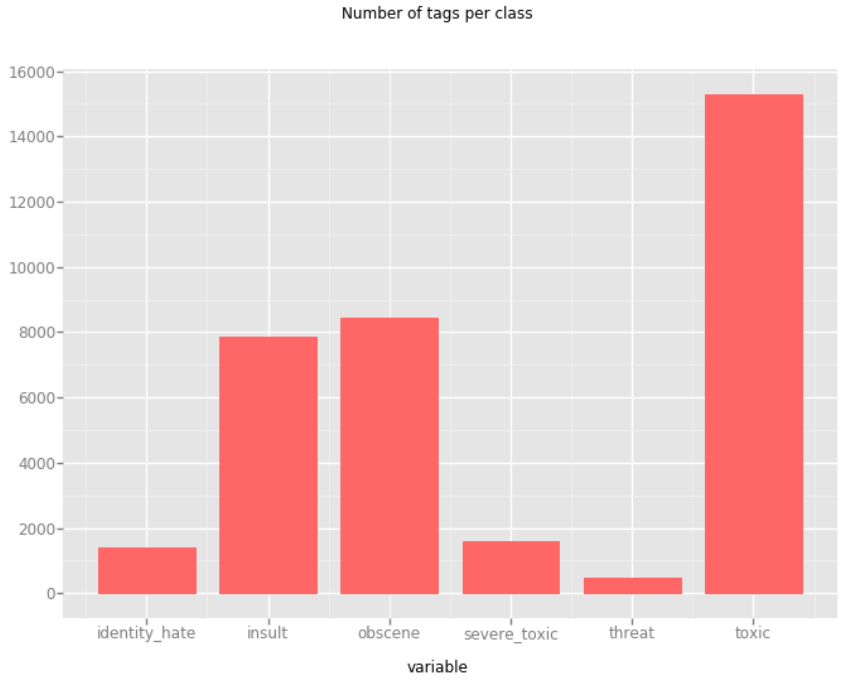
	Train	Train
Rows	159571	153164
Percentage	51	49

2.3 Class Imbalance

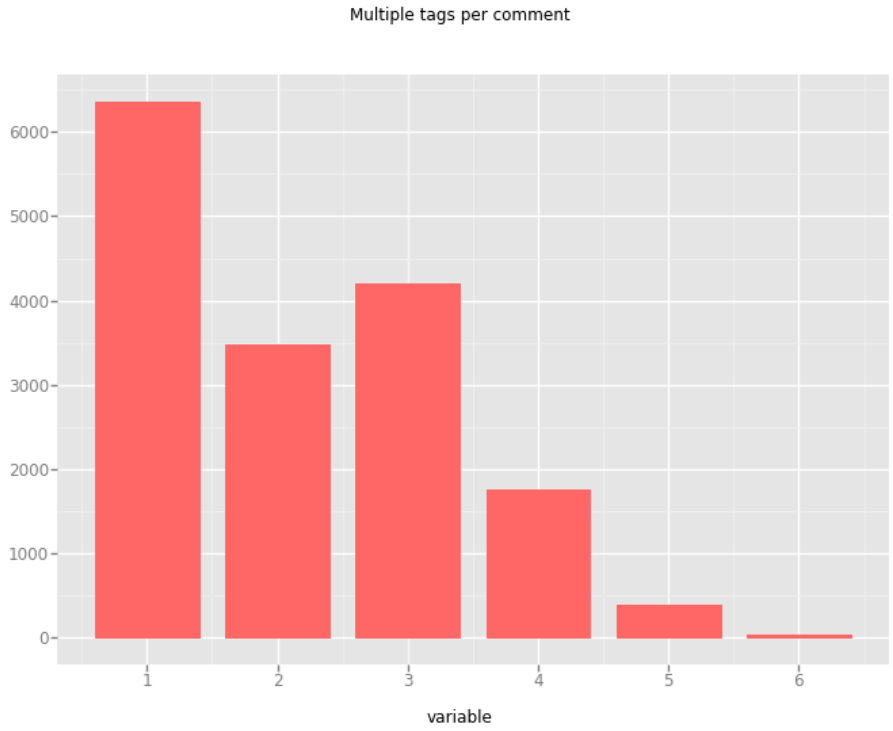
Class Imbalance for train dataset.

Total number of comments	159571
Total number of clean comments	143346
Total number of tag	35098

Total Tag distribution



As we can see no of clean comments plus total number of tags is not equal to the total no of comments which means many comments have been multi tagged. Let's calculate the no of comments in each class which are muti-tagged.



2.4 Completeness of the data

Missing value in train data :

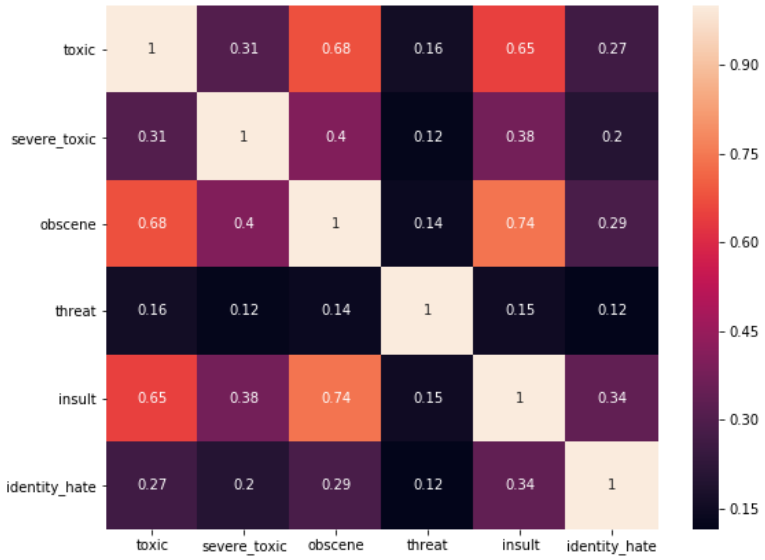
id	0
comment_text	0
toxic	0
severe_toxic	0
obscene	0
threat	0
insult	0
identity_hate	0

Missing value in test data :

id	0
comment_text	0
toxic	0
severe_toxic	0
obscene	0
threat	0
insult	0
identity_hate	0

There seems to be no missing value in the dataset and there are no ‘unknown’ text as well in the comment_text dataset.

2.5 Correlation Analysis



2.6 Cross Tab

Since it will be pretty difficult to visualize crosstab between all the 6 classes, lets take a look at toxic with other tags:

	severe_toxic		obscene		threat		insult		identity_hate	
severe_toxic	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
toxic										
0	144277	0	143754	523	144248	29	143744	533	144174	103
1	13699	1595	7368	7926	14845	449	7950	7344	13992	1302

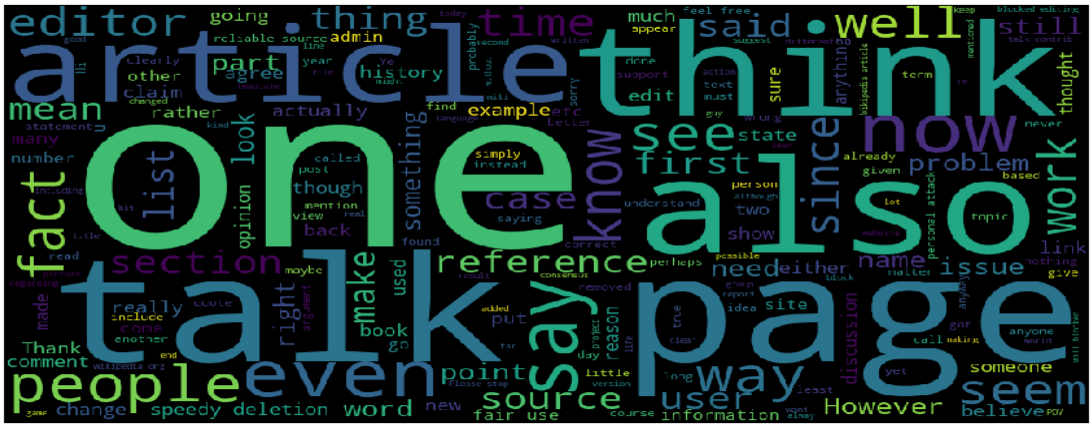
The above table shows that all the severe_toxic comment are tagged as toxic as well which means severe_toxic is the subset of toxic.

Cross tab between insult and other classes:

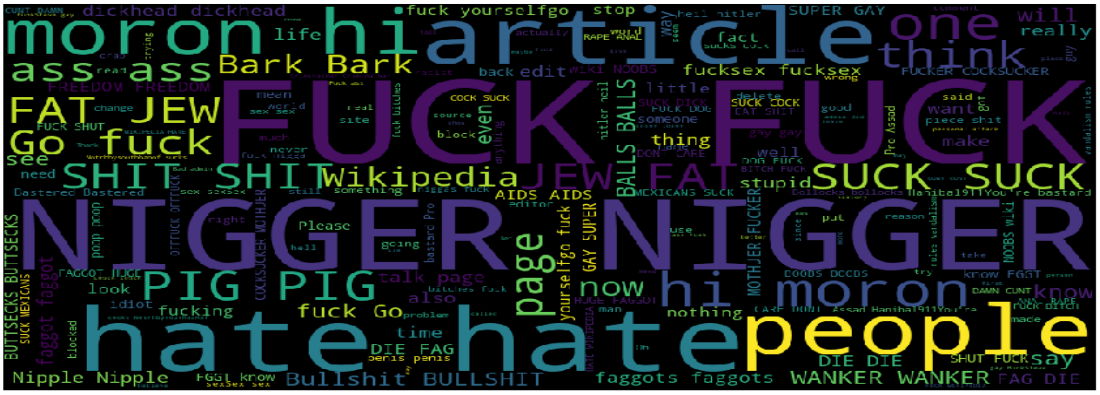
	toxic		severe_toxic		obscene		threat		identity_hate	
toxic	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
insult										
0	143744	7950	151470	224	149400	2294	151523	171	151449	245
1	533	7344	6506	1371	1722	6155	7570	307	6717	1160

2.7 Word Cloud

Wordcloud of Clean Comments



WordCloud of Toxic Comments



WordCloud of Severe Toxic Comments

WordCloud of Threat Comments

WordCloud of Obscene Comments

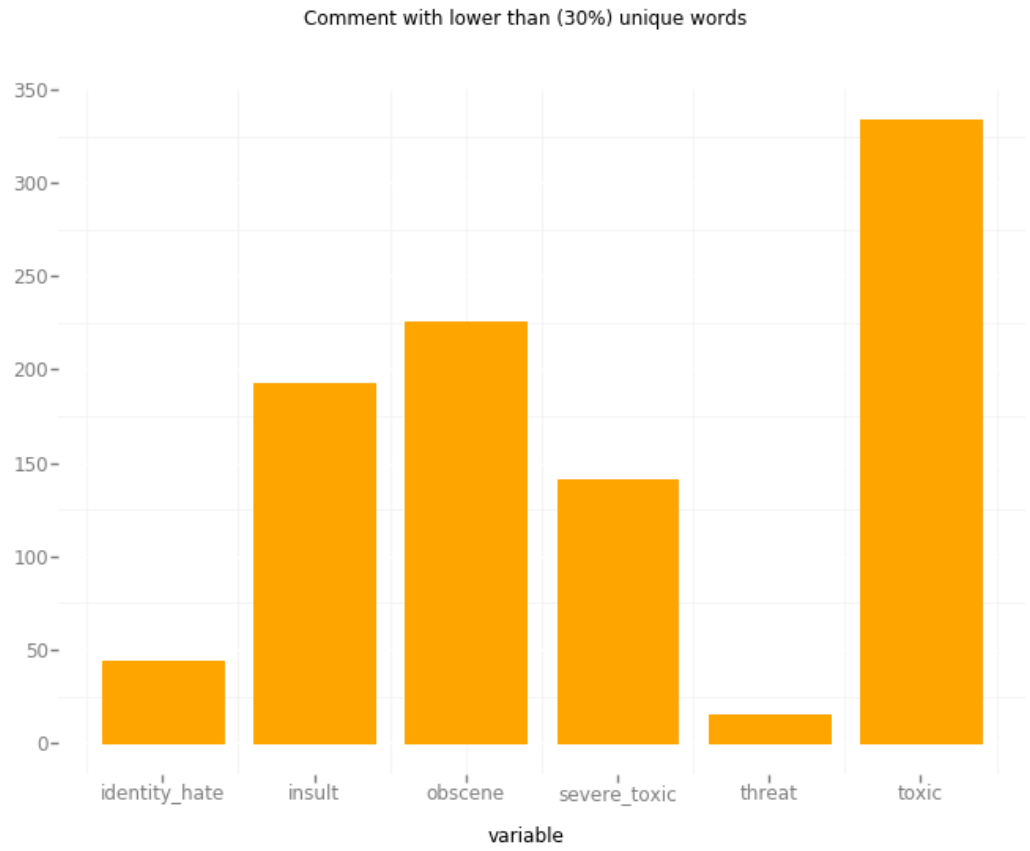
Wordcloud of Insult Comments

WordCloud of Identity Hate Comments

3 **Feature Engineering:**

id	count_sentence	count_words	count_uqwords	count_stopwords	count_letters	count_punctuations	count_uppercase	avglen_words	unqword_percent	punct_percent
0000997932d777bf	2	43	41	18	264	10	2	5.162791	95.348837	23.255814
000103f0d9cfb60f	1	17	17	2	112	12	1	5.588235	100.000000	70.588235
000113f07ec002fd	1	42	39	20	233	6	0	4.571429	92.857143	14.285714
0001b41b1c6bb37e	5	113	82	56	622	21	5	4.486726	72.566372	18.584071

Spammers:



Spammers are toxic to the model

An example of spammer in Identity hate

YOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU
PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF
ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!\nYOU PIECE OF ASIAN
AMERICAN SCUM!\nYOU PIECE OF ASIAN AMERICAN SCUM!'

An example of toxic spammer

YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!
YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!

YOU ARE A MOTHJER FUCKER COCKSUCKER! YOU ARE A MOTHJER FUCKER COCKSUCKER!'

An example of threat spammer

SuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC

FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!\nSuPeRTR0LL WiLL LiVe FoReVeR!\niF You DoN'T ReSPeCT THe SuPeRTR0LL You WiLL Die You PaTHeTiC FooL!"

4 Text Pre-processing

4.1 Corpus Cleaning

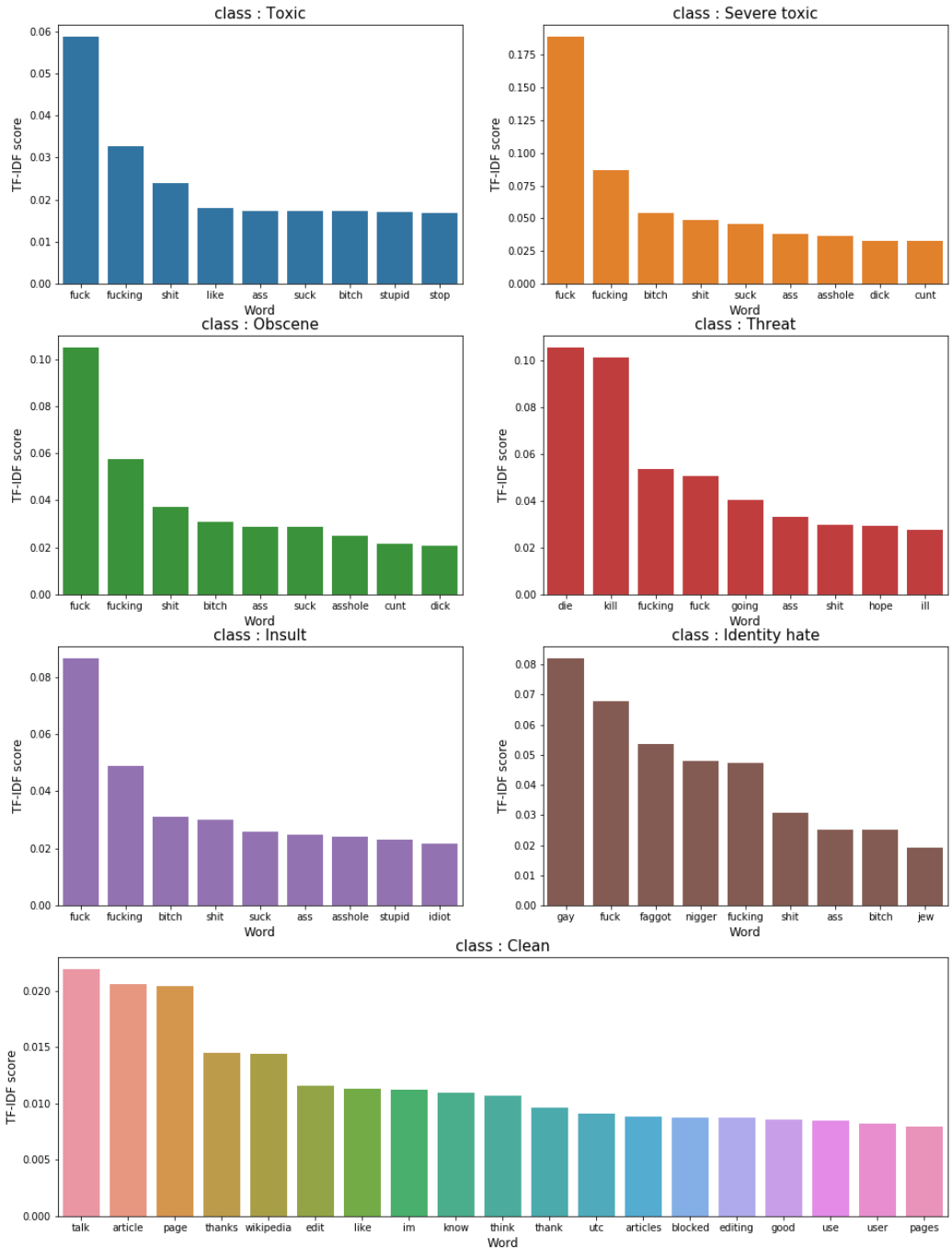
Comment text before cleaning:

'"\n\nSorry I find that there\'s nothing ""honorable"" about samurai ethics, it\'s largely blind loyalty and fanaticism, and lack any of the Judeo-Christian values such as compassion and mercy. Such are the values indoctrinated into Imperial Japanese soldiers when they slaughtered over 300,000 Chinese civillians in the Nanjing Massacre, tortured Allied POWs in prison camps such as Changi, launched suicide Kamikaze attacks against Allied ships etc. Not even the Germans, who had at least at some Judeo-Christian values left even under the Nazi regime, match the level of fanaticism and pure evil of Imperial Japan.60.242.159.224 \n\n"'

Comment text after cleaning:

' sorry find theres nothing honorable samurai ethics largely blind loyalty fanaticism lack judeochristian values compassion mercy values indoctrinated imperial japanese soldiers slaughtered chinese civillians nanjing massacre tortured allied pows prison camps changi launched suicide kamikaze attacks allied ships etc even germans least judeochristian values left even nazi regime match level fanaticism pure evil imperial japan '

TF_IDF Top words per class(unigrams)



5 Prediction

Logistic Regression.

Used Logistic Regression to find the coefficient values of all the word extracted from TfidfVectorizer.

Then fitting the logistic model with the train unigrams and the target classes from the train dataset and then predicting the tags for the test dataset.

Coefficient value of the feature with their corresponding feature

	0	0
0	abide	-0.549982
1	ability	-1.389539
2	able	0.553125
3	abortion	1.759678
4	absence	-0.396224
5	absolute	2.390534
6	absolutely	-0.239823
7	absurd	-0.212042
8	abuse	-1.643592
9	abused	-0.933796
10	abusing	-1.696815
11	abusive	0.572757
12	academic	-1.108304
13	academics	-0.519579
14	academy	-1.012687

Validating the model on train dataset as we have tags for those.

Column: identity_hate

Confusion matrix	
152284	5882
17	1388

	precision	recall	f1-score	support
	1.00	0.96	0.98	158166
	0.19	0.99	0.32	1405
avg / total	0.99	0.96	0.98	159571

Column: toxic

Confusion	matrix
[[142208	2069]
10093	5201]]

	precision	recall	f1-score	support
	0.93	0.99	0.96	144277
	0.72	0.34	0.46	15294
avg / total	0.91	0.92	0.91	159571

Column: severe_toxic

Confusion	matrix
[[151835	6141]
466	1129]]

	precision	recall	f1-score	support
	1.00	0.96	0.98	157976
	0.16	0.71	0.25	1595
avg / total	0.99	0.96	0.97	159571

Column: obscene

Confusion	matrix
[[147671	3451]
[4630	3819]]

	precision	recall	f1-score	support
	0.97	0.98	0.97	151122
	0.53	0.45	0.49	8449
avg / total	0.95	0.95	0.95	159571

Column: threat

Confusion	matrix
[[152082	7011]
[219	259]]

	precision	recall	f1-score	support
	1.00	0.96	0.98	159093
	0.04	0.54	0.07	478
avg / total	1.00	0.95	0.97	159571

Column: insult

Confusion	matrix
[[148184	3510]
[4117	3760]]

	precision	recall	f1-score	support
	0.97	0.98	0.97	151694
	0.52	0.48	0.50	7877
avg / total	0.95	0.95	0.95	159571

Top 10 rows of the predicted dataset:

id	toxic	severe_toxic	obscene	threat	insult	identity_hate
00001cee341fdb12	0.999963	0.967598	0.999993	0.315525	0.998839	0.991273
0000247867823ef7	0.011015	0.006086	0.001791	0.002538	0.025985	0.009702
00013b17ad220c46	0.061991	0.001125	0.046332	0.000154	0.107666	0.041826
00017563c3f7919a	0.009743	0.019207	0.004942	0.000342	0.006412	0.000743
00017695ad8997eb	0.132900	0.017170	0.040452	0.000622	0.057539	0.007543
0001ea8717f6de06	0.059103	0.003909	0.013374	0.000167	0.073134	0.010970
00024115d4cbde0f	0.009373	0.001050	0.013767	0.000210	0.010836	0.003833
000247e83dcc1211	0.956231	0.009964	0.306070	0.042331	0.554623	0.016294
00025358d4737918	0.058902	0.005907	0.025885	0.006005	0.016970	0.004821
00026d1092fe71cc	0.015931	0.002080	0.025197	0.001053	0.017526	0.031499

6 Python Code:

Please refer ToxicityPrediction.html for the complete code. Follow the link or find the attached file with the name ToxicityPrediction.html in the folder.