Toxic Language Classification:

Cleaning Up Wikipedia

Brian McMahon

12 March 2018

Disclaimer

This dataset contains text that may be considered profane, vulgar, or offensive.

Why is this project important?

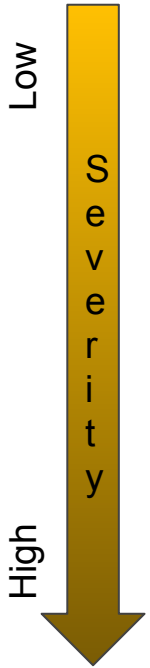
Dataset of wikipedia talk page comments from Alphabet-sponsored Kaggle Competition to develop tools to “help improve online conversation”

Competition Objective to Identify and classify comments labeled as belonging to one or more of six different “toxic language” categories

In **this presentation** we will conduct preliminary topic analysis on the dataset to draw distinctions amongst the different categories of toxic language

The Six Category “Tags”

Dataset contains multi-tagged comments; “rating” refers to the number tags a comment has received



Toxic. General taboo, rude language

Obscene. Profanity

Insult. Rude, degrading

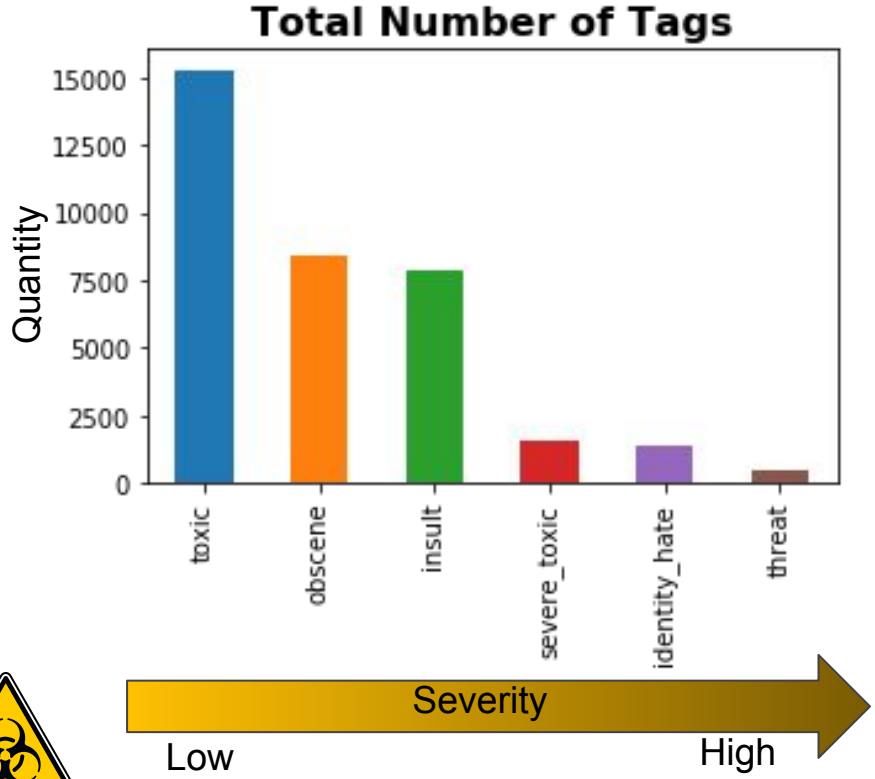
Severe Toxic. Extreme language, statements not classified as Threat or Identity Hate

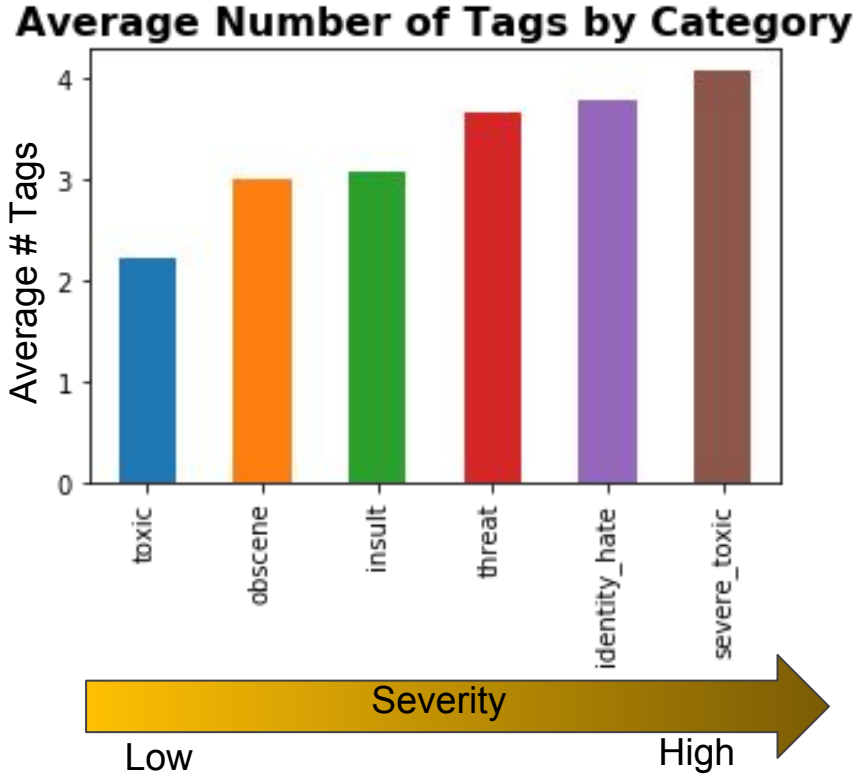
Identity Hate. Attacking race, religion, orientation

Threat. Death wishes, threats

Toxic Comment Mix

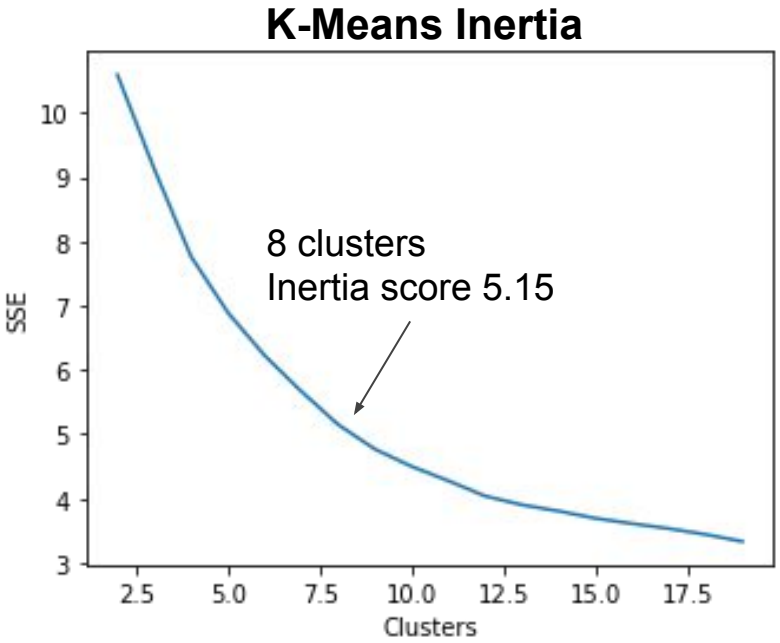
Dataset contains multi-tagged comments; each comment can have from 0 (“clean”) to 6 tags
“Toxic” data filtered to comments with at least 1 tag

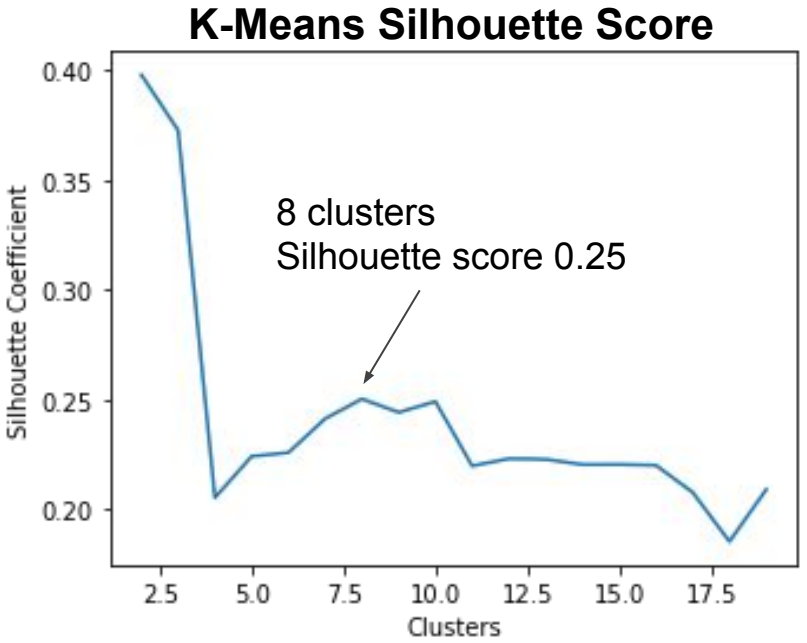




K Means Clustering: Toxic

We can see some of the classifications within these clusters





K Means Clustering: Toxic

We can see some of the classifications within these clusters

Cluster 0

Severe Toxic

Cluster 1

Toxic

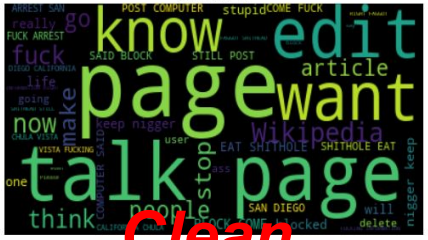
Cluster 2

Toxic

Cluster 3

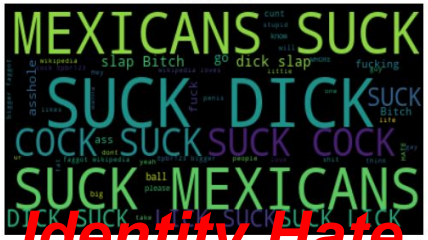
Identity Hate

Cluster 4



Clean

Cluster 5



Identity Hate

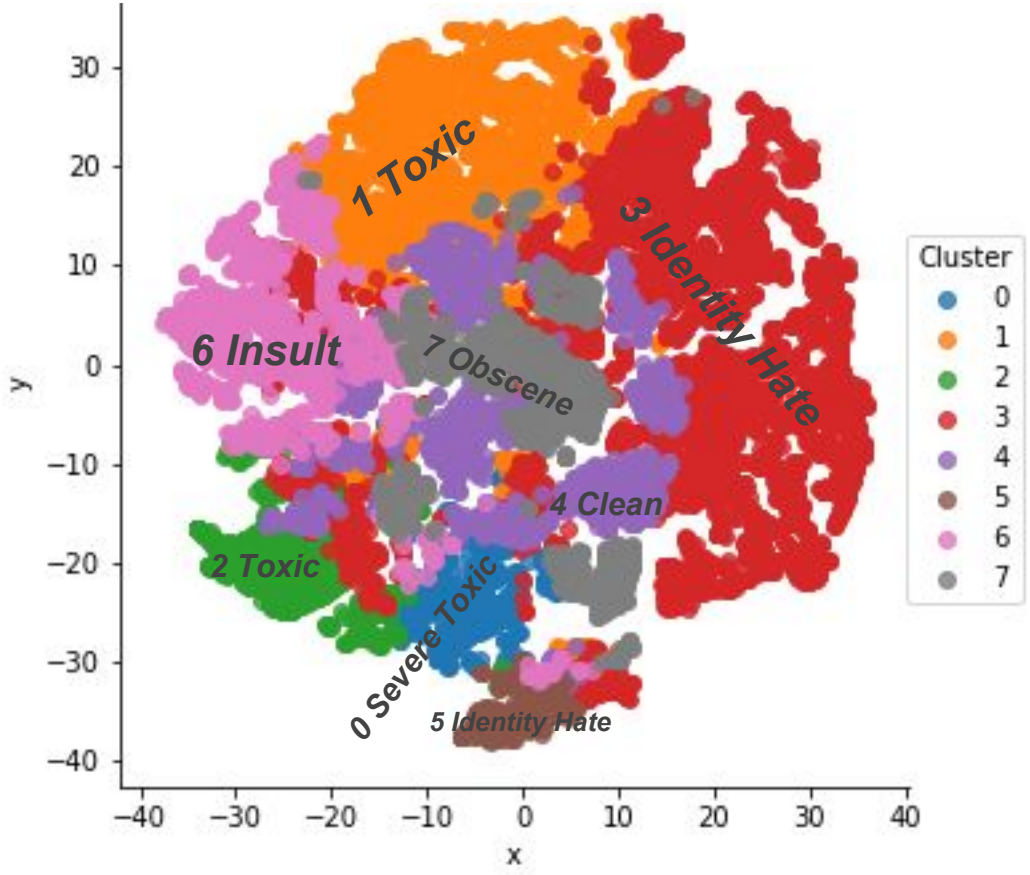
Cluster 6

Insult

Cluster 7

Obscene

T-SNE: Toxic


1 Toxic
3 Identity Hate
6 Insult
7 Obscene
4 Clean
2 Toxic
0 Severe Toxic
5 Identity Hate
Cluster
0
1
2
3
4
5
6
7
x
y
30
20
10
0
−10
−20
−30
−40
−40
−30
−20
−10
0
10
20
30
40

Key Takeaways

Different platforms will have different thresholds for “toxic comments”

Approach toxic language from multiple angles:

- General inappropriate language (i.e., “bad words”) may be acceptable on some forums
- Language which threatens or attacks a person’s identity (“identity hate”) does not always use “bad words” and must be recognized in context

While models today do a reasonable job of identifying these issues, the ultimate decision of appropriateness remains subjective and is a challenge for machines to fully grasp

Questions?

APPENDIX

Techniques Utilized

TF-IDF. Used to gauge the word importance within a collection or corpus, increases proportionally to number of times a word appears in the document and is offset by frequency of word in corpus

NMF. Matrix factorized into multiple matrices with no negative elements

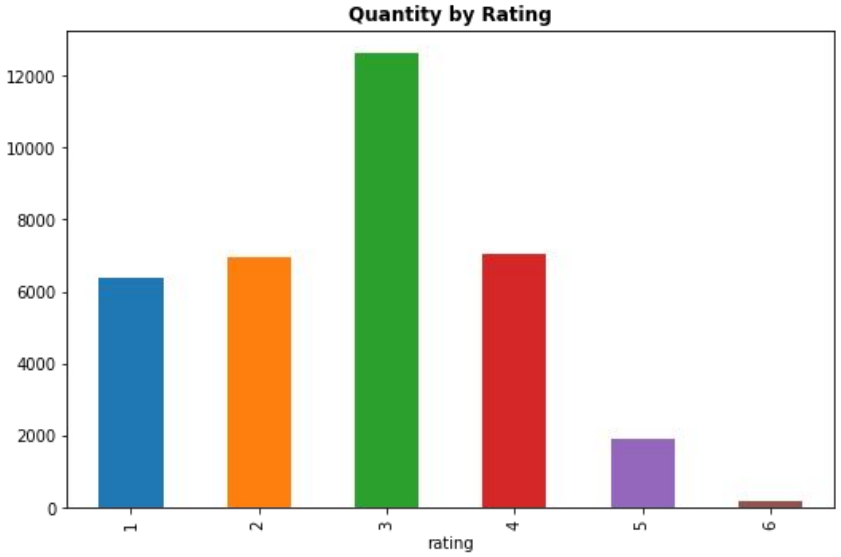
K-Means Clustering. Determine optimized number of clusters by silhouette coefficient

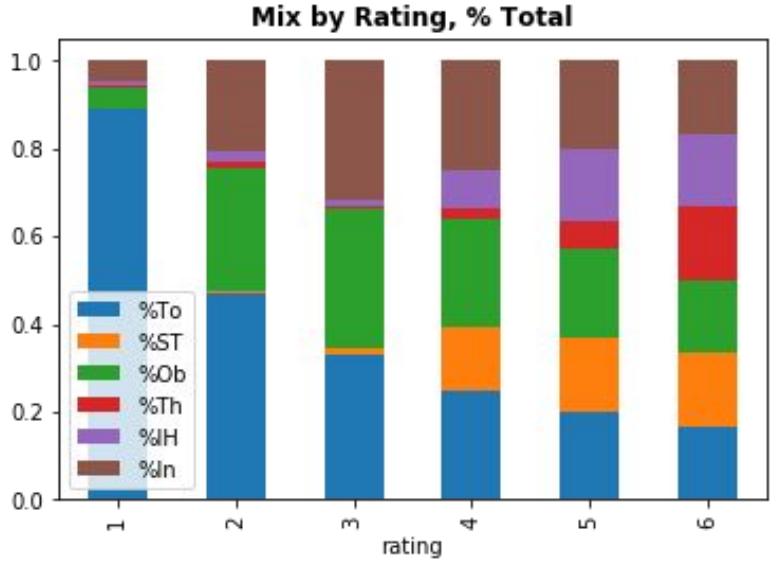
- *Silhouette Score*. A succinct graphical representation of how well each object lies within its cluster
- *Inertia Score*. Choose centroids that minimise the inertia, or within-cluster sum of squared criterion

T-SNE. Embeds high-dimensional data into low-dimensional scatter plot

The Six Categories of Toxic Comments

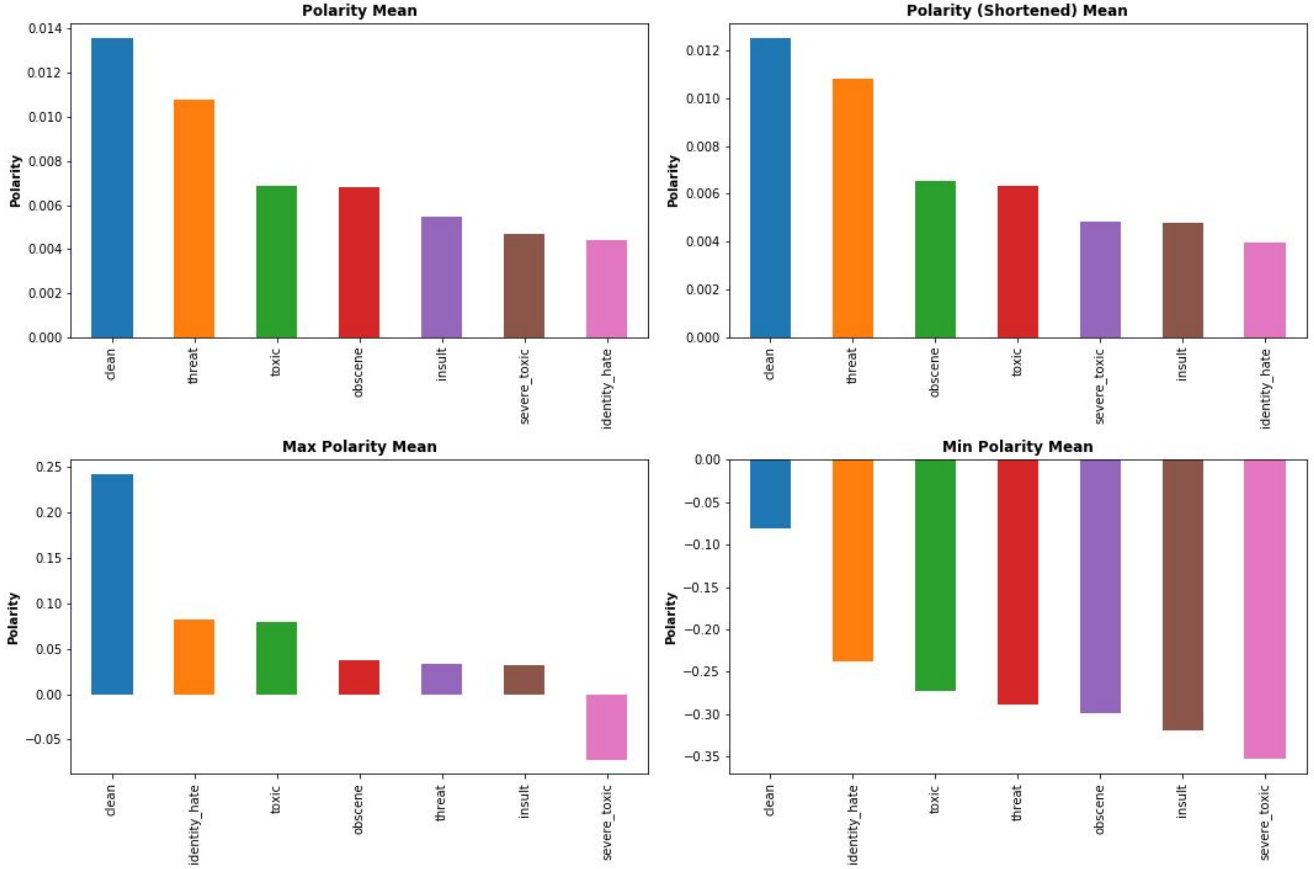
Dataset contains multi-tagged comments; “rating” refers to the number tags a comment has received



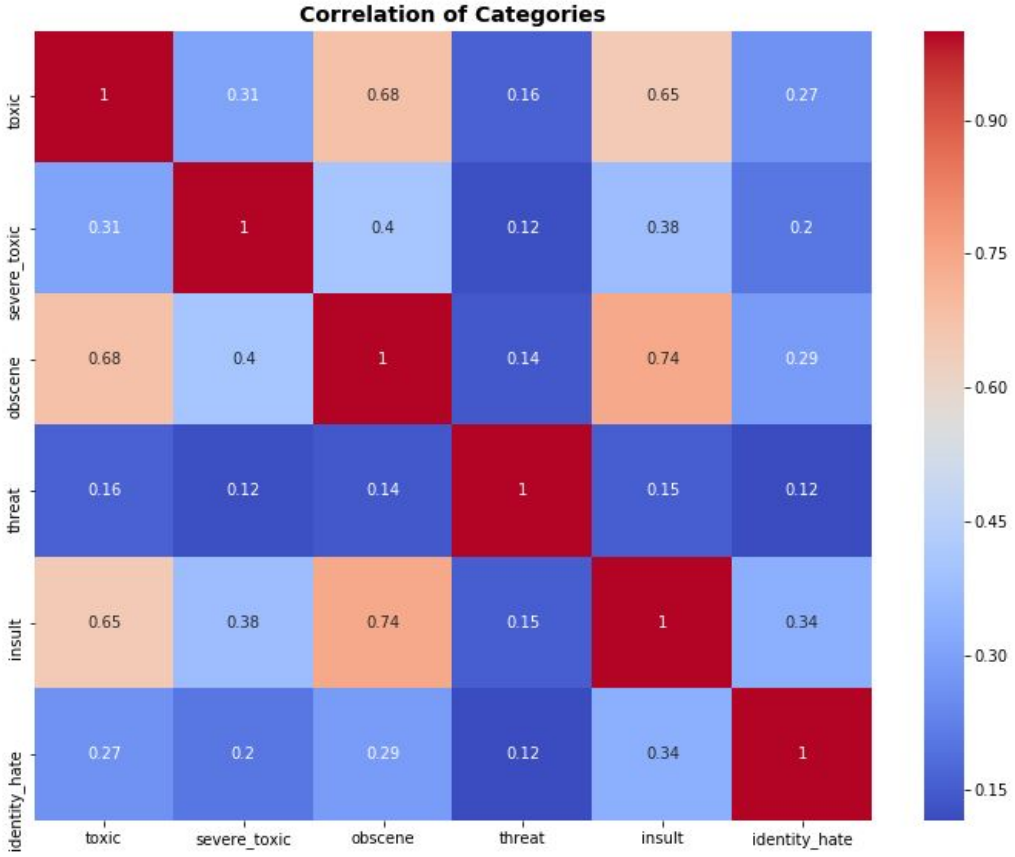


Rating means the number of “tags” per comment

Polarity by Category



Category Correlation



Category Samples

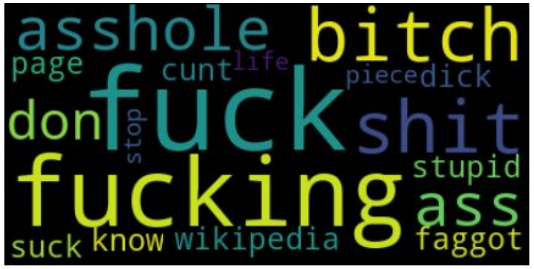
Toxic

Taboo, rude

Obscene

Profanity

Insult



Rude, degrading

Severe Toxic

Extreme language

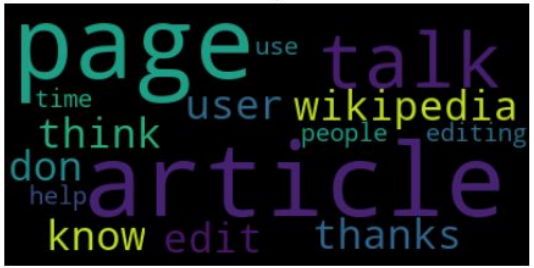
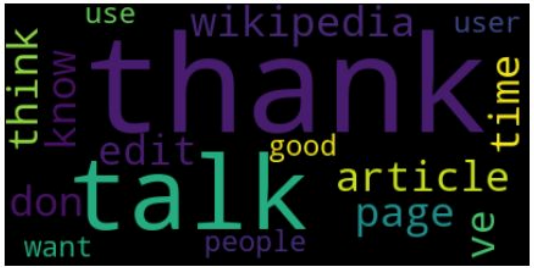
Identity Hate

Attacking race, religion and orientation

Threat

Threats of death

Topic Modeling by Category: TF-IDF/NMF

	TF-IDF/NMF - Frobenius Norm	TF-IDF/NMF - Kullback-Leibler Divergence	LDA
Clean			
	Discussions of article edits		
Toxic			
	Derogatory language		

Topic Modeling by Category: TF-IDF/NMF

	LDA	TF-IDF/NMF - Frobenius Norm	TF-IDF/NMF - Kullback-Leibler Divergence
Threat			

Threats of death

Identity Hate			

Attacking race, religion and orientation

K Means Clustering: All Words (Toxic + Clean)

All_Words Cluster 0

All_Words Cluster 1

All_Words Cluster 2

All_Words Cluster 3

All_Words Cluster 4

All_Words Cluster 5

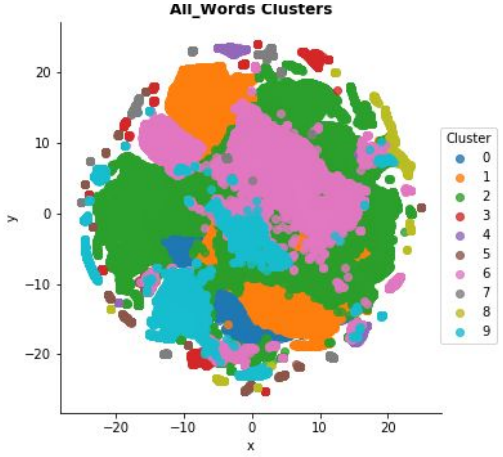
All_Words Cluster 6

All_Words Cluster 7

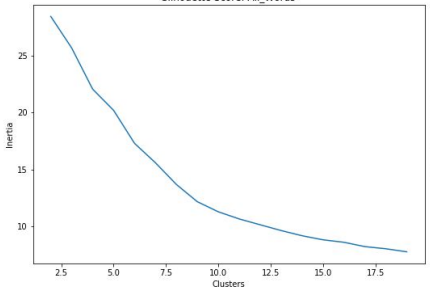
All_Words Cluster 8

All_Words Cluster 9

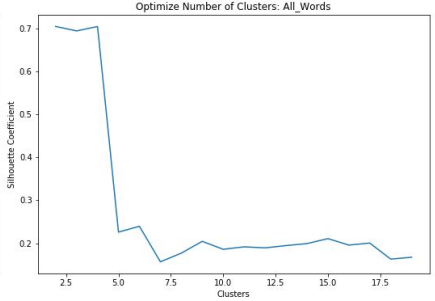
All_Words Clusters



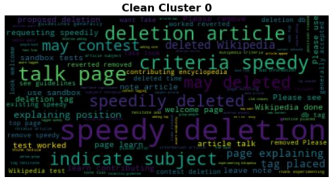
Silhouette Score: All_Words

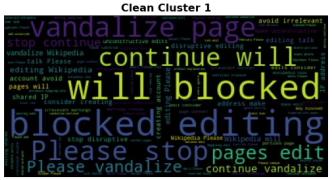


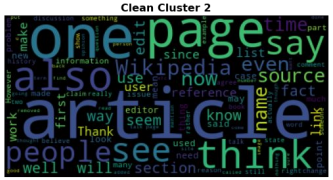
Optimize Number of Clusters: All_Words

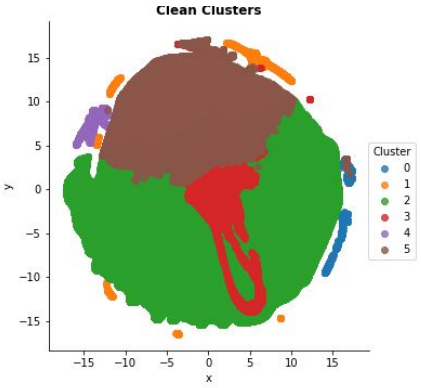


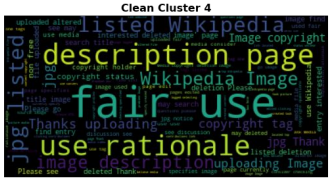
K Means Clustering: Clean Words (All Rated == 0)

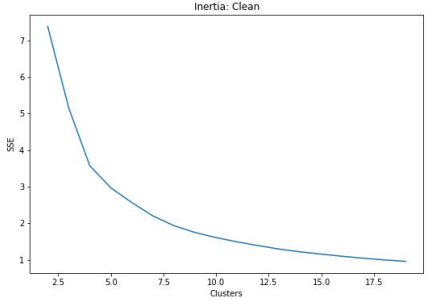


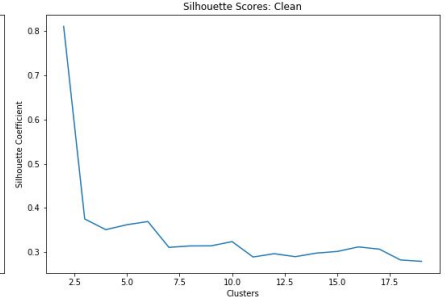






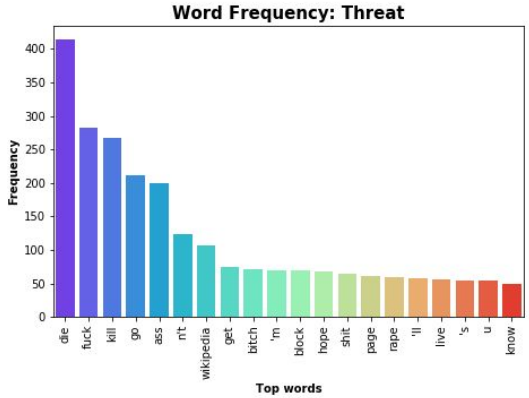


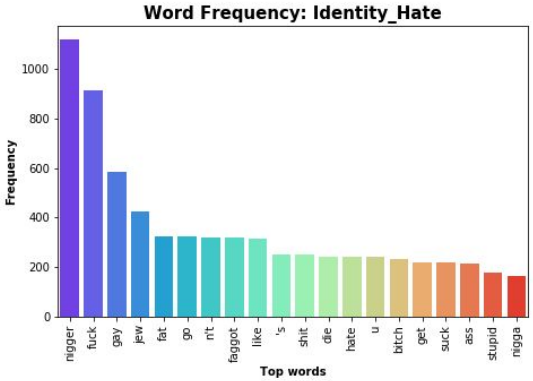


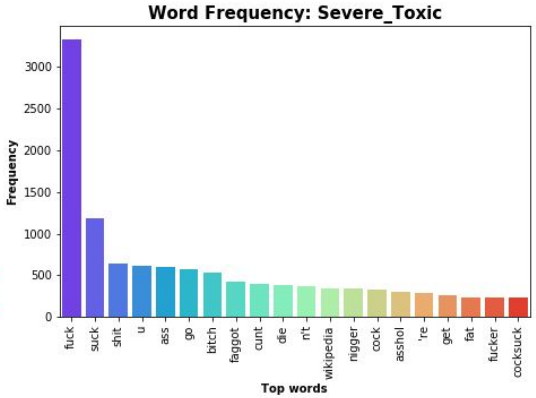


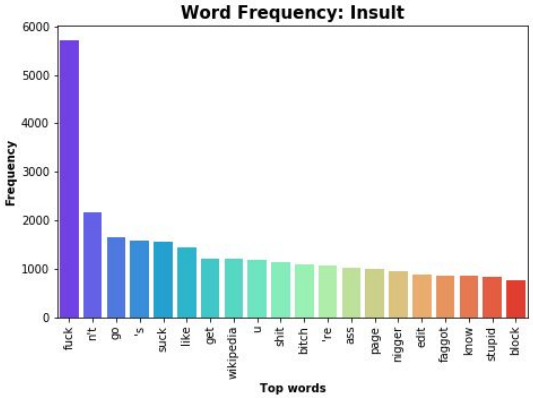
Word Frequency: Singles

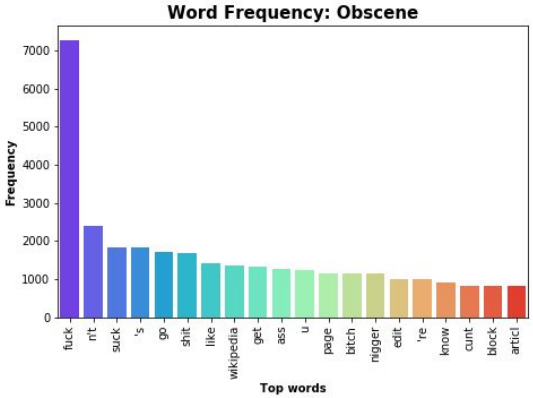
Only stopwords filtered

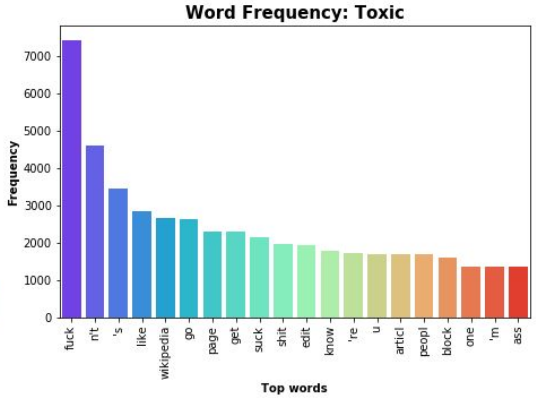




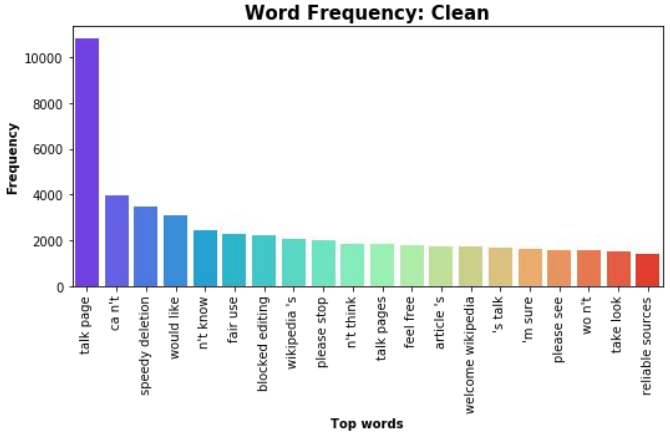


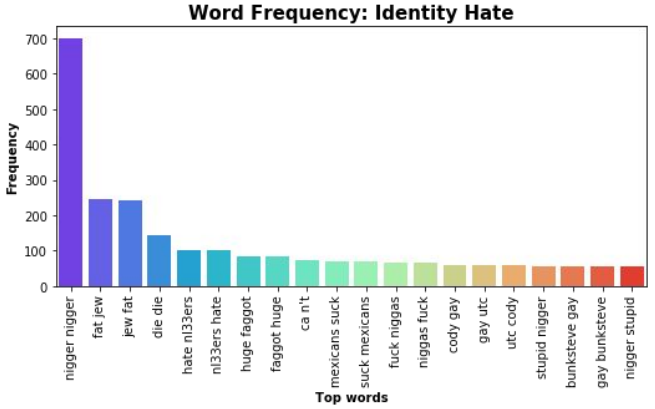


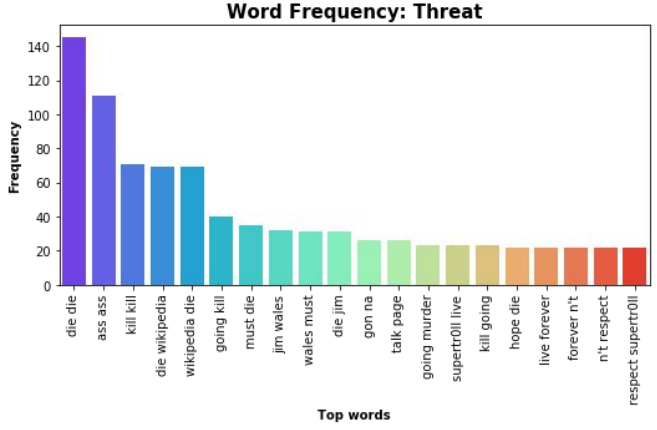


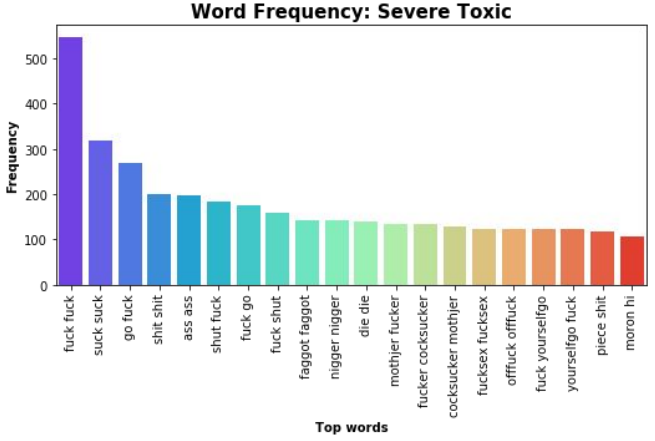


Word Frequency: Bigrams



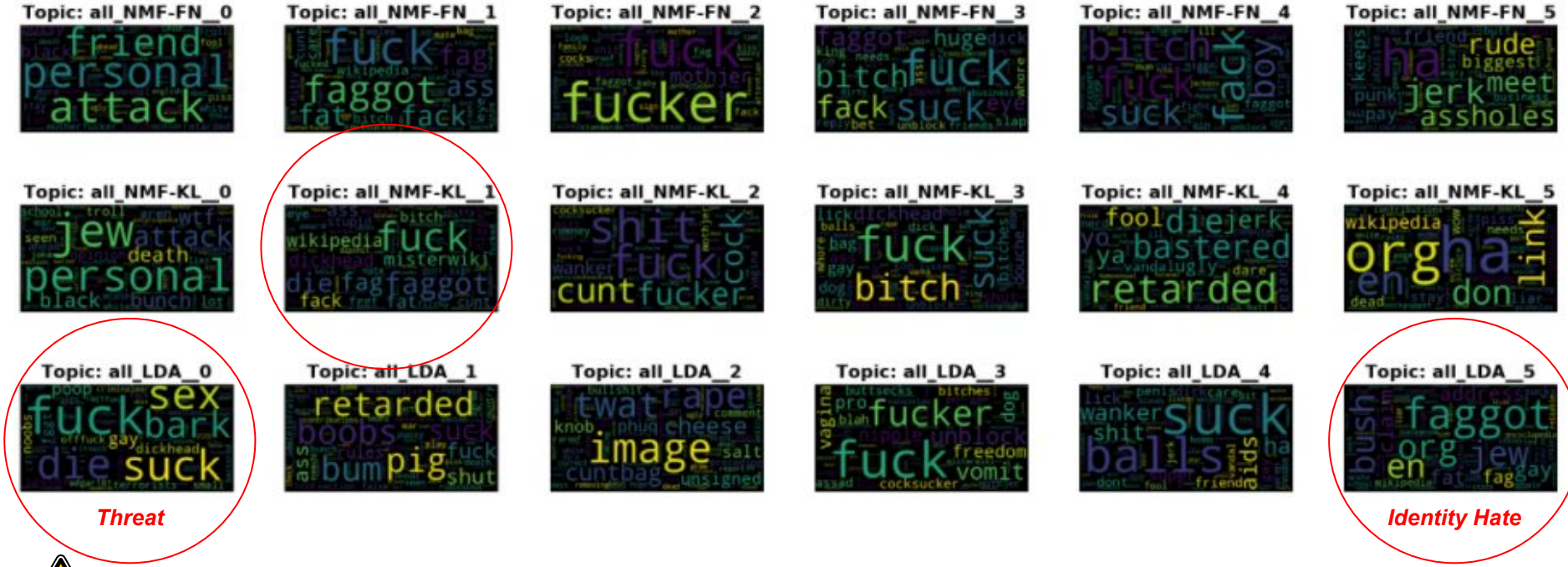






Topic Modeling, 6 Topics from Rating > 0

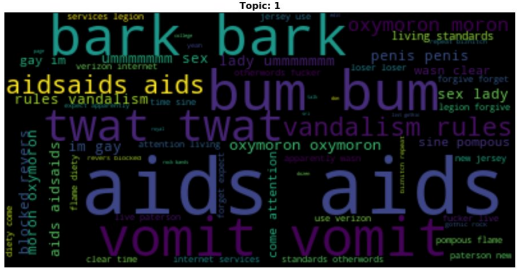
All tagged as at least one form of toxic language
6 categories, topical modeling also roughly ties to these



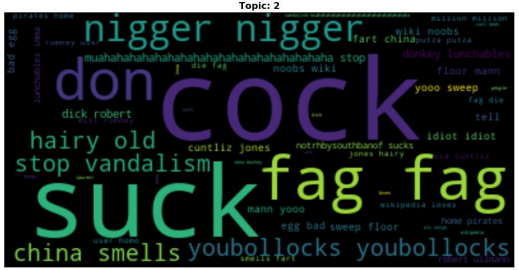
LDA WordClouds By Toxic Bigram Topic

Topic: 0

Topic: 1



Topic: 2



Topic: 3

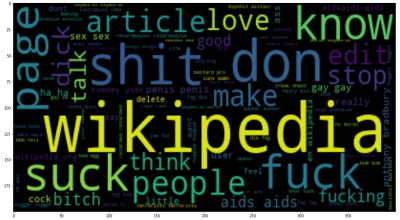
Topic: 4

Topic: 5

LDA WordClouds By Class

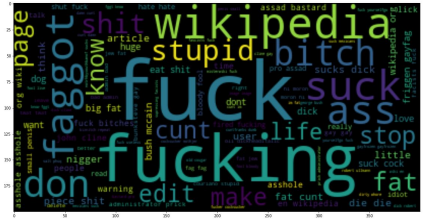
Clean

Toxic



Obscene

Insult



Severe Toxic

Identity Hate

Threat

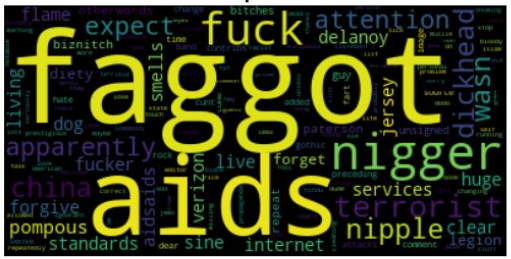
LDA, 6 Topics

Topic: 0

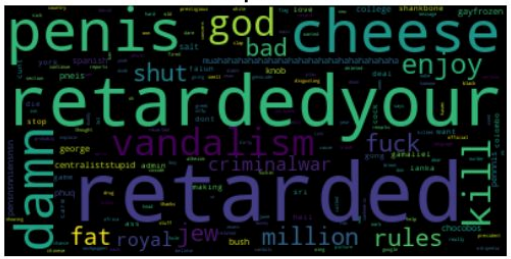
Topic: 1

Topic: 2

Topic: 3



Topic: 4



Topic: 5