A Data Mining Approach for Classification of Toxic Comments and Empowering Positive Discourse

By,

Deeraj Nair
Malhar Dhopate
Rishit Puri

Motivation & Problem Statement

In an era of ubiquitous online communication, hate speech and toxic remarks have attracted a lot of attention. Using data mining and machine learning algorithms, this project attempts to create a model that is able to recommend to the user to change/update their comments to reduce online negativity and promote a positive discussion, or discourse. The primary focus of this project is,

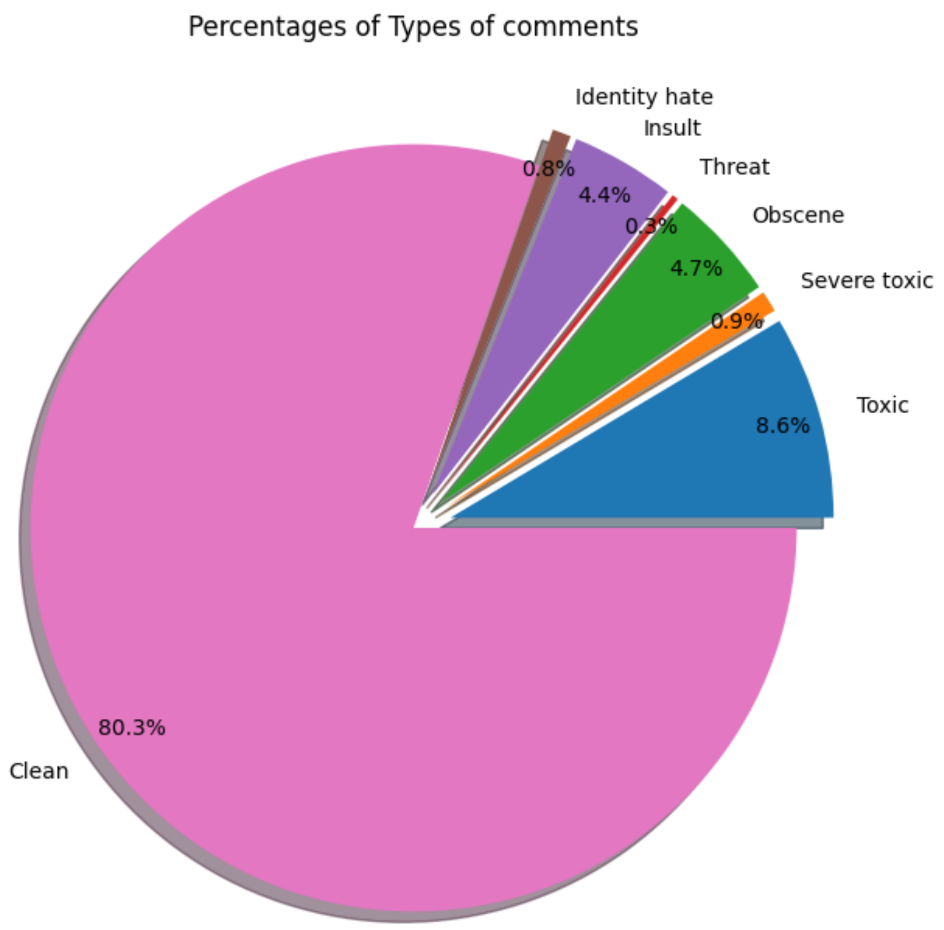
1. To create a model to analyze the sentiment of user comments in an online setting to identify comments for – ‘toxic’, ‘severe_toxic’, ’obscene’, ‘threat’, ‘insult’, and, ’identity_hate’ classes.
2. To help users to choose a positive discourse in an effort to reduce online negativity in form of - bullying, or abusing which can affect the mental and psychological health of the content creators.

Dataset Overview

- The dataset is readily available on Kaggle with -
  - 159571 rows and 8 columns
- The dataset has a variety of comments classified into 6 different levels of toxicities.

EDA

- Data Composition -
- - The number of ‘toxic’ comments is, 10,652
- - The number of ‘severe_toxic’ comments is, 1091
- - The number of ‘obscene’ comments is, 5876
- - The number of ‘threat’ comments is, 338
- - The number of ‘insult’ comments is, 5474
- - The number of ‘identity_hate’ comments is, 950
- - The number of ‘clean’ comments is, 99,384

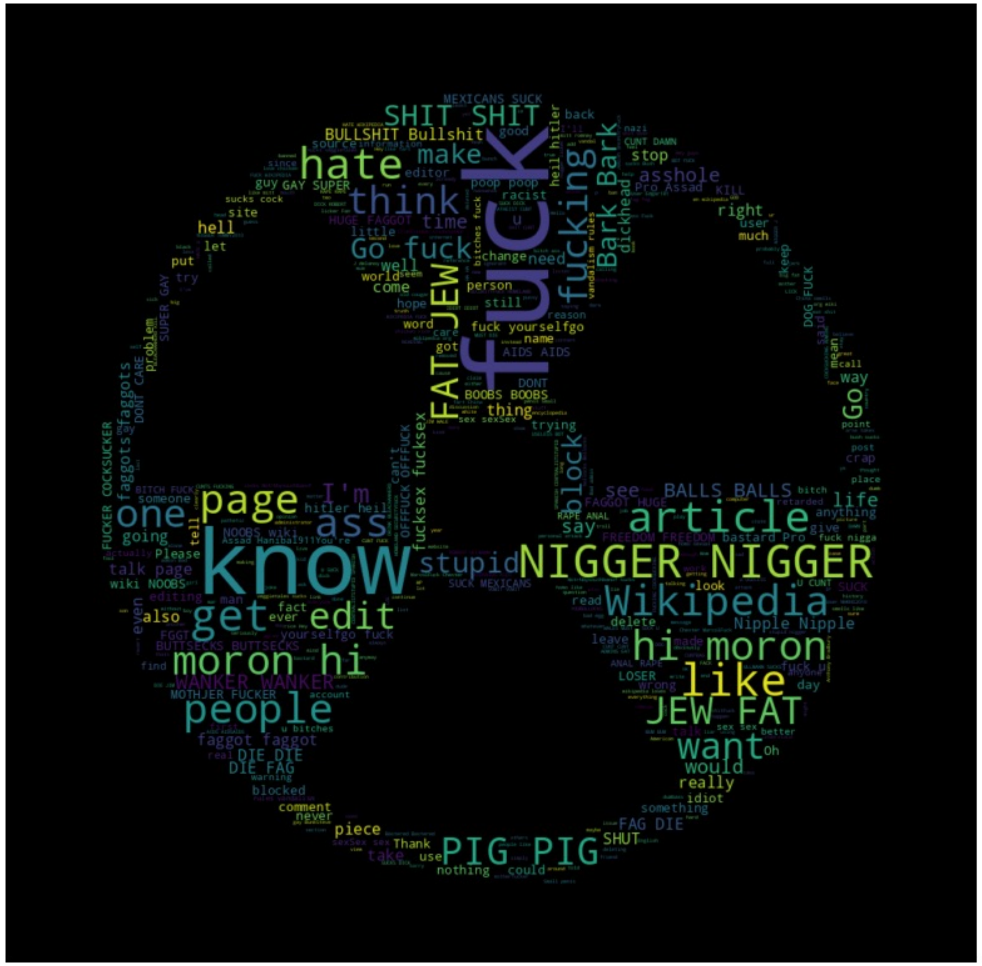


Disclaimer – The following slides may include explicit language, sensitive topics, or triggering themes that may be distressing to some users. Reader discretion is advised.

Wordclouds for 'toxic' and 'severe_toxic' comments

Word Cloud for Severe Toxic Comments

Word Cloud for Toxic Comments



Wordclouds for 'threat' and 'obscene' comments

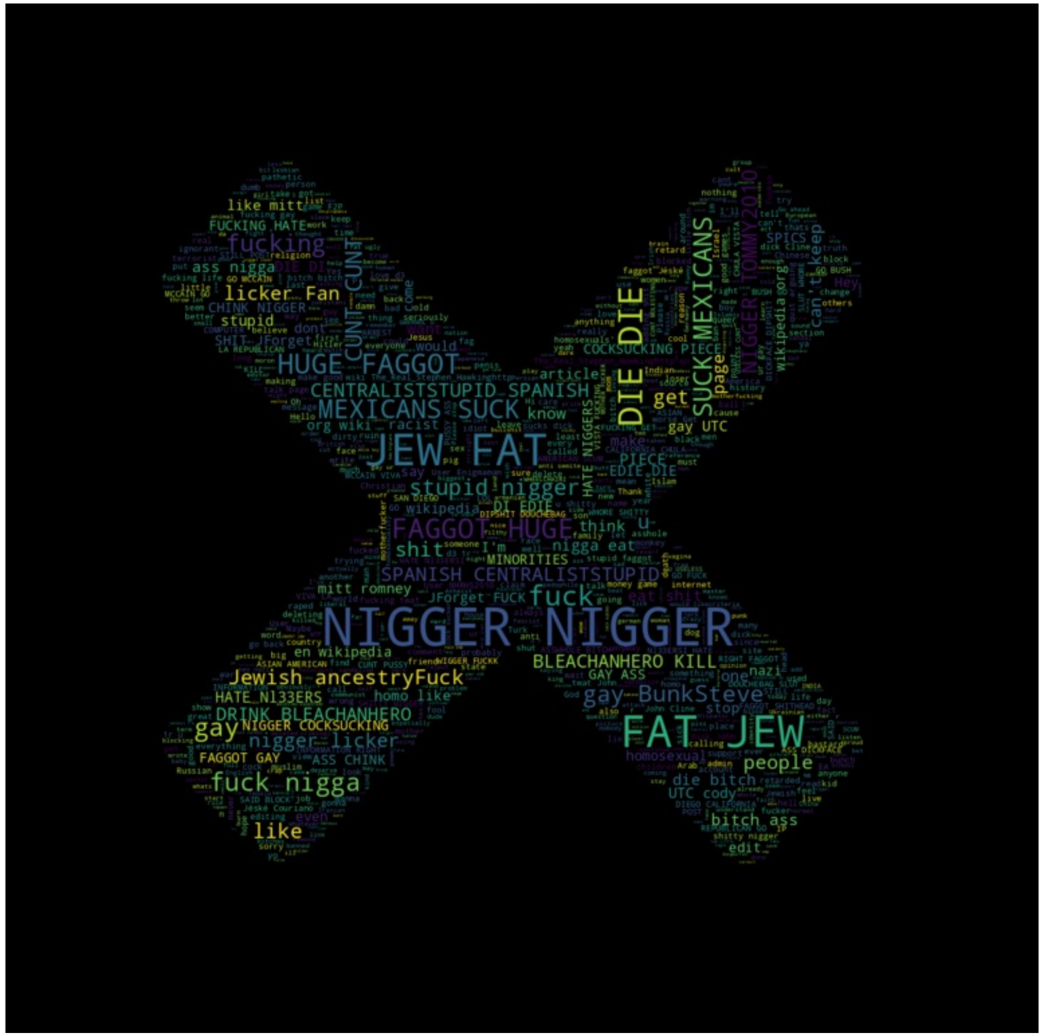
Word Cloud for Threat Comments

Word Cloud for Obscene Comments

Wordclouds for 'insult' and 'identity_hate' comments

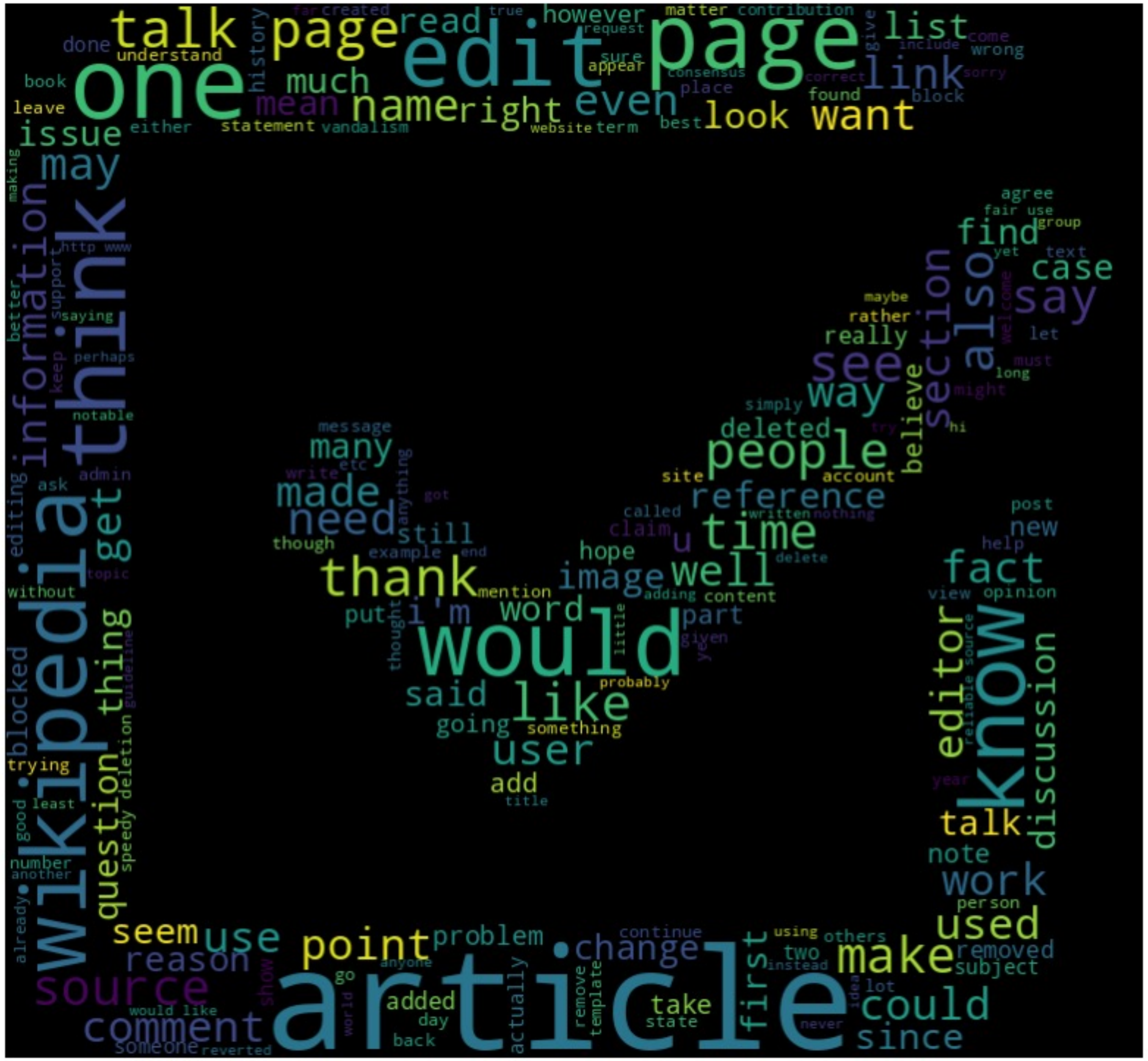
Word Cloud for Insult Comments

Word Cloud for Identity hate Comments



Wordcloud for 'clean' comments

Word Cloud for Clean Comments



Current Models & Observations

	Label	Accuracy	Precision	Recall	F1 Score
0	toxic	0.9569	0.9551	0.9569	0.9532
1	severe_toxic	0.9903	0.9879	0.9903	0.9885
2	obscene	0.9767	0.9756	0.9767	0.9746
3	threat	0.9978	0.9971	0.9978	0.9971
4	insult	0.9689	0.9657	0.9689	0.9653
5	identity_hate	0.9917	0.9897	0.9917	0.9890
6	Combined	0.9183	0.8928	0.9183	0.9022

Logistic Regression – 91.8% accuracy

	Label	Accuracy	Precision	Recall	F1 Score
0	toxic	0.9572	0.9548	0.9572	0.9548
1	severe_toxic	0.9898	0.9849	0.9898	0.9858
2	obscene	0.9778	0.9766	0.9778	0.9767
3	threat	0.9977	0.9969	0.9977	0.9969
4	insult	0.9687	0.9658	0.9687	0.9664
5	identity_hate	0.9915	0.9897	0.9915	0.9885
6	Combined	0.9172	0.8907	0.9172	0.8991

Random Forest – 91.7% accuracy

	Label	Accuracy	Precision	Recall	F1 Score
0	toxic	0.9600	0.9582	0.9600	0.9573
1	severe_toxic	0.9902	0.9879	0.9902	0.9858
2	obscene	0.9792	0.9781	0.9792	0.9779
3	threat	0.9978	0.9971	0.9978	0.9972
4	insult	0.9710	0.9683	0.9710	0.9685
5	identity_hate	0.9919	0.9905	0.9919	0.9891
6	Combined	0.9213	0.8974	0.9213	0.9061

SVM – 92.1% accuracy

```
Layer (type)                Output Shape              Param #
=================================================================
input_3 (InputLayer)        [(None, 100)]             0

embedding_1 (Embedding)     (None, 100, 300)          52248900

spatial_dropout1d_1 (Spati  (None, 100, 300)          0
alDropout1D)

bidirectional (Bidirection  (None, 100, 256)          439296
al)

conv1d (Conv1D)             (None, 100, 64)           16448

max_pooling1d (MaxPooling1  (None, 50, 64)            0
D)

flatten (Flatten)           (None, 3200)              0

dense (Dense)               (None, 128)               409728

dropout (Dropout)           (None, 128)               0

batch_normalization (Batch  (None, 128)               512
Normalization)

dense_1 (Dense)             (None, 6)                 774

=================================================================
Total params: 53115658 (202.62 MB)
Trainable params: 866502 (3.31 MB)
Non-trainable params: 52249156 (199.31 MB)
_________________________________________________________________
```

LSTM Neural Net

(Please note, the project is still being optimized, we are still trying to improve the recall and reduce computational time.)

Expected Outcomes & Future Improvements

- We are still involved in optimizing the models further, to improve the recall for multiple labels.
- Multi language recognition
- Develop real-time monitoring capabilities to identify and address toxic comments as soon as they are posted.
- Integrate user feedback mechanisms to enhance the model's performance on real-world usage.

THANK YOU