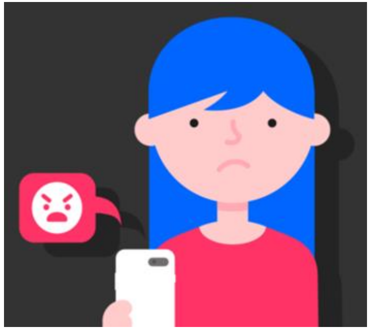
Identifying Toxic Comments on Social-media and Categorizing the level of Toxicity

Motivation

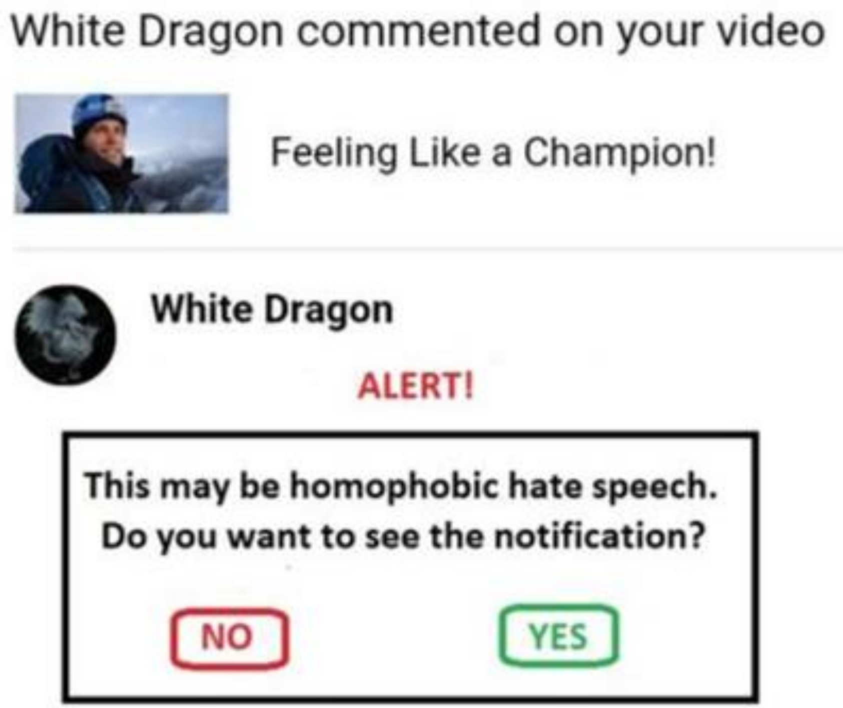
Social media gives people the freedom and platform to express their thoughts publically behind a username. Though this brought like-minded people together, many people feel entitled to use toxic speech towards brand or creators insensitively.

This can directly affect the victim’s self-esteem and has very real impacts on their mental wellbeing with many reporting high rates of **anxiety** and **depression**. Hate speech online has been linked to a global increase in **violence toward minorities**.



Application

Policies used to curb hate speech risk limiting free speech and are inconsistently enforced. Countries such as the United States grant social media companies broad powers in managing their content and enforcing **hate speech rules**. Others, including Germany, can force companies to **remove posts within certain time periods**. Along with law forces, there is a need of new applications or features to **automatically detect, report and block** hate speech.



White Dragon commented on your video

Feeling Like a Champion!

White Dragon

ALERT!

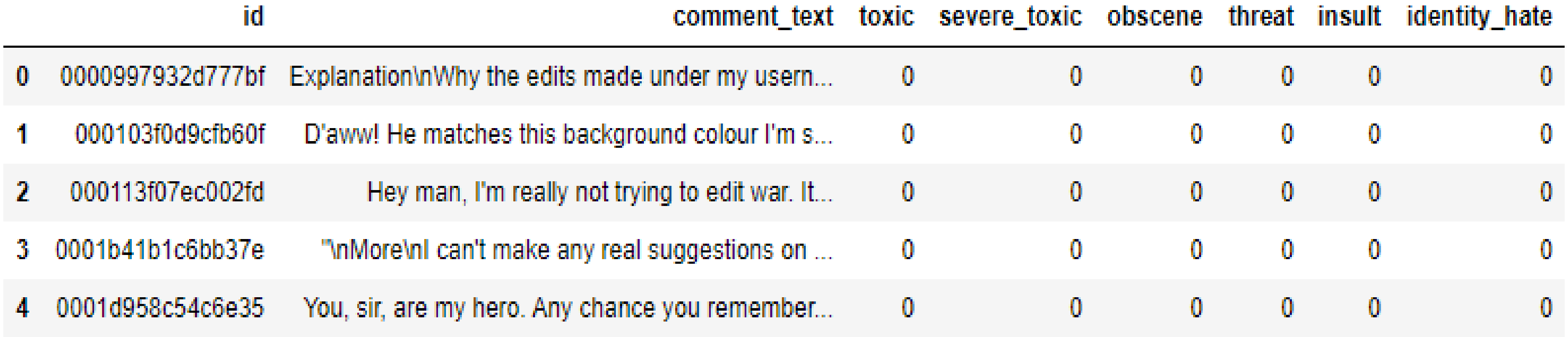
This may be homophobic hate speech.
Do you want to see the notification?

Dataset

Source:- https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

The dataset collected have been labelled by human raters for the toxic behavior.
The toxicity types are labelled as **toxic**, **severe toxic**, **obscene**, **threat**, **insult** and **identity hate.**

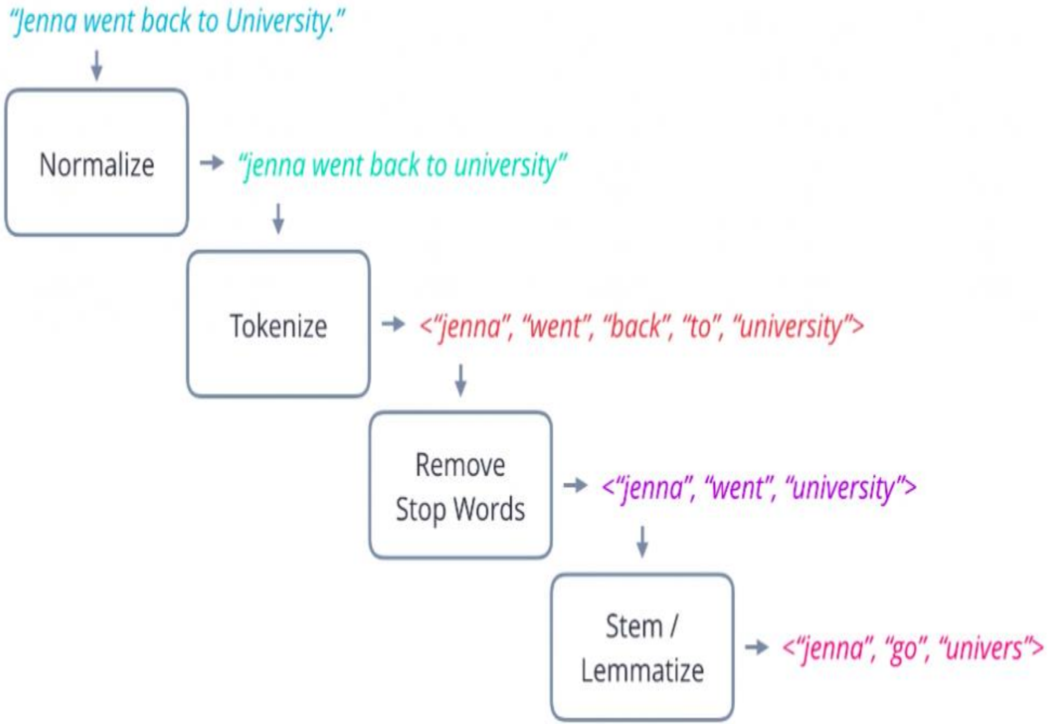
It contains approximately **1,60,000** text comment samples.



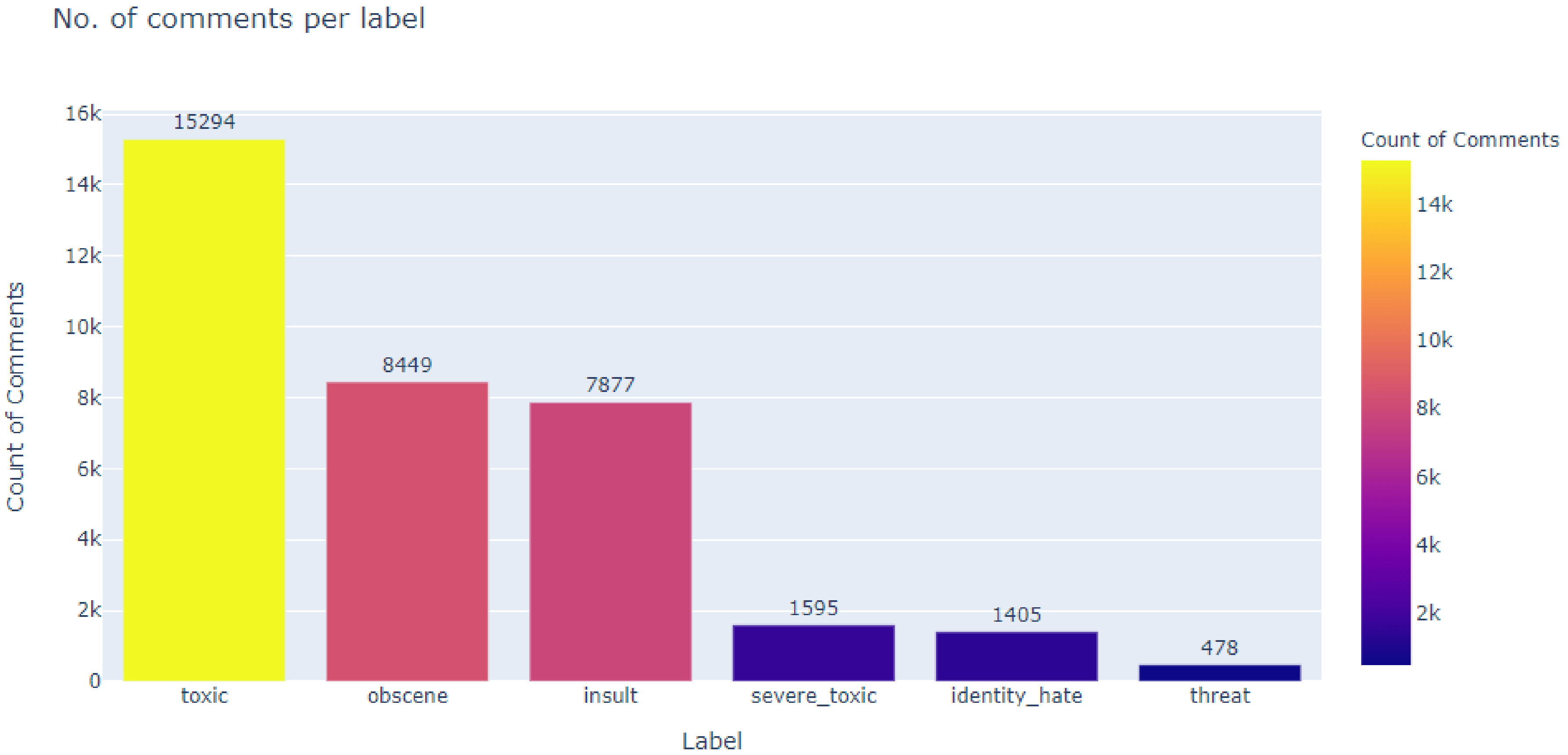
	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

Data Preprocessing

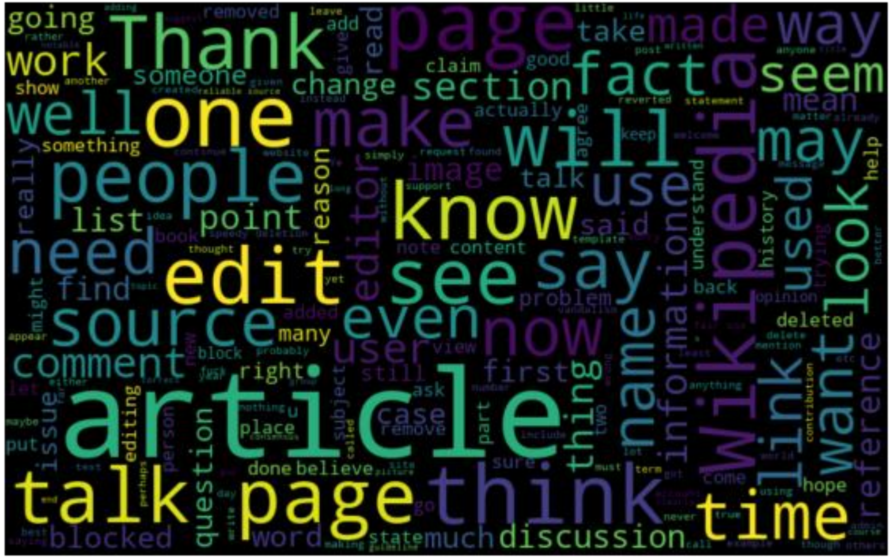
1. Removing extra spaces
2. Removal of IP address and URLs if present
3. Removing Stopwords
4. Tokenization and Lemmatization of words present in the comment


"Jenna went back to University."
Normalize
"jenna went back to university"
Tokenize
<"jenna", "went", "back", "to", "university">
Remove Stop Words
<"jenna", "went", "university">
Stem / Lemmatize
<"jenna", "go", "univers">

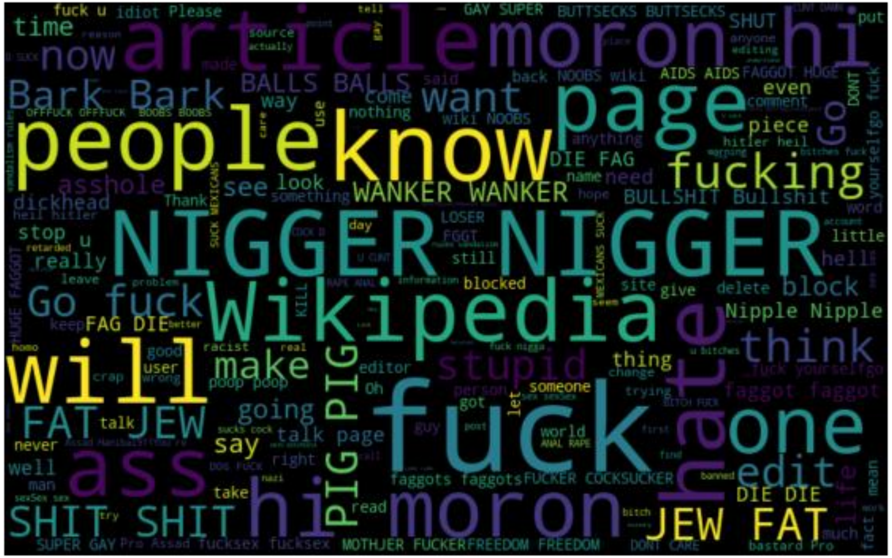
Exploratory Data Analysis



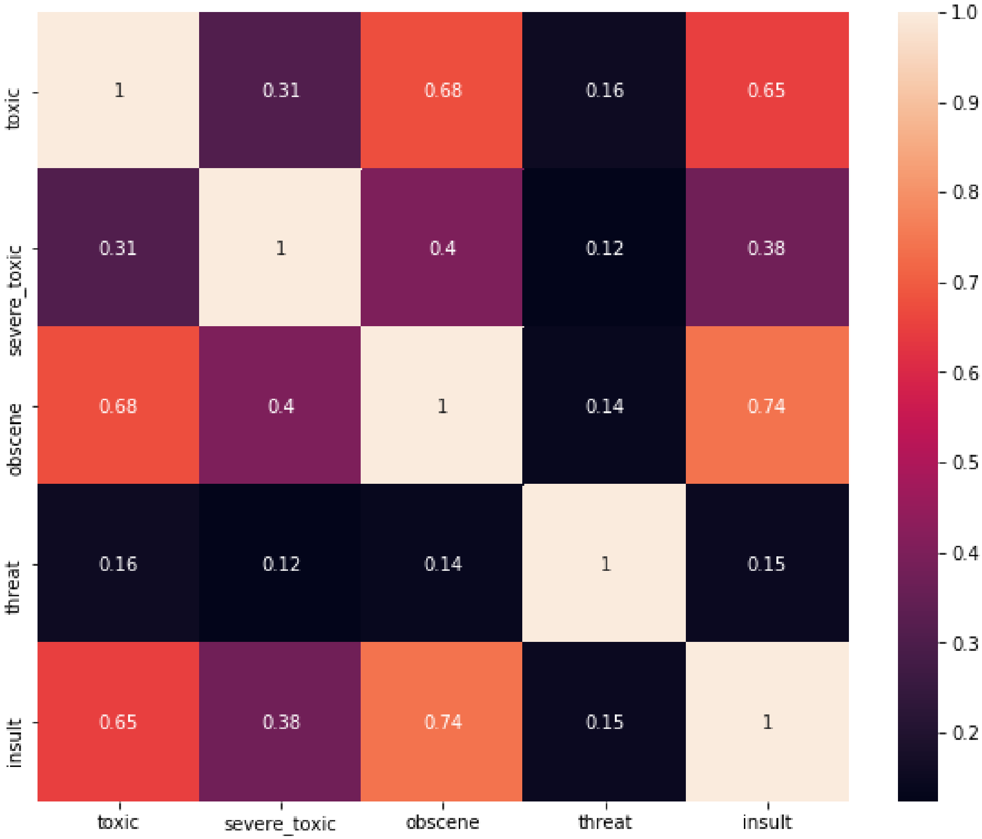
WordCloud



Entire Text in Comments

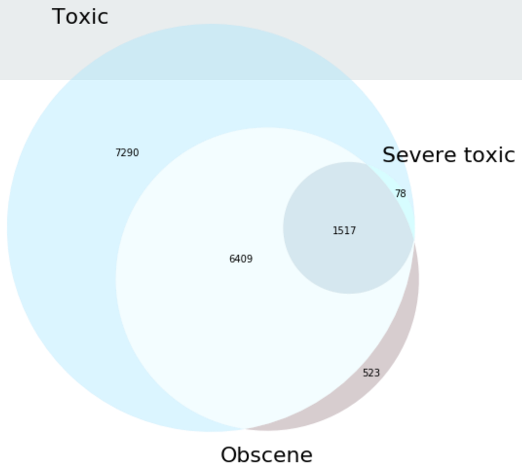


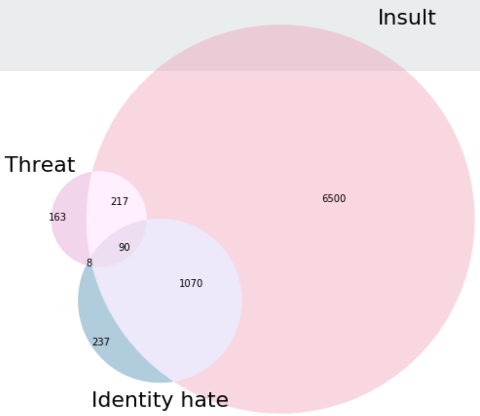
Comments Identified as Toxic



Correlation Matrix for different levels of toxicity

Insult and Obscene, Toxic and obscene exhibit a strong correlation

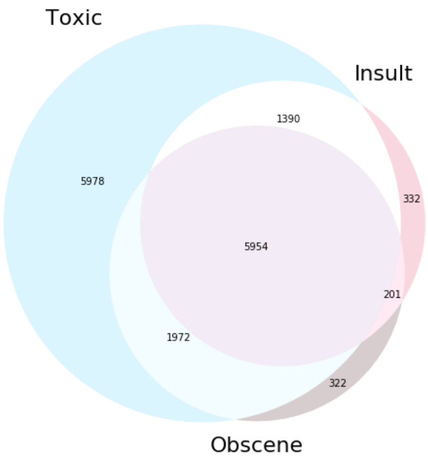


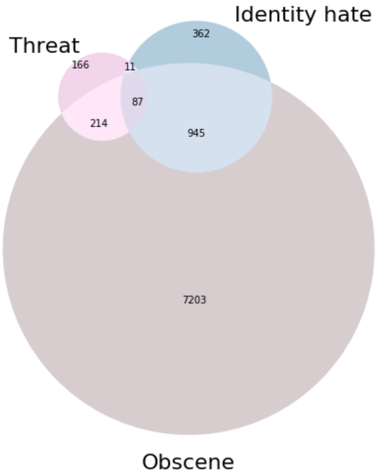


Interrelation between Categories

using Venn Diagram

- ❖ Almost all severe toxic comments are toxic
- ❖ Most severe toxic comments are also obscene in nature
- ❖ toxic, insult and obscene comments share a good amount of overlap

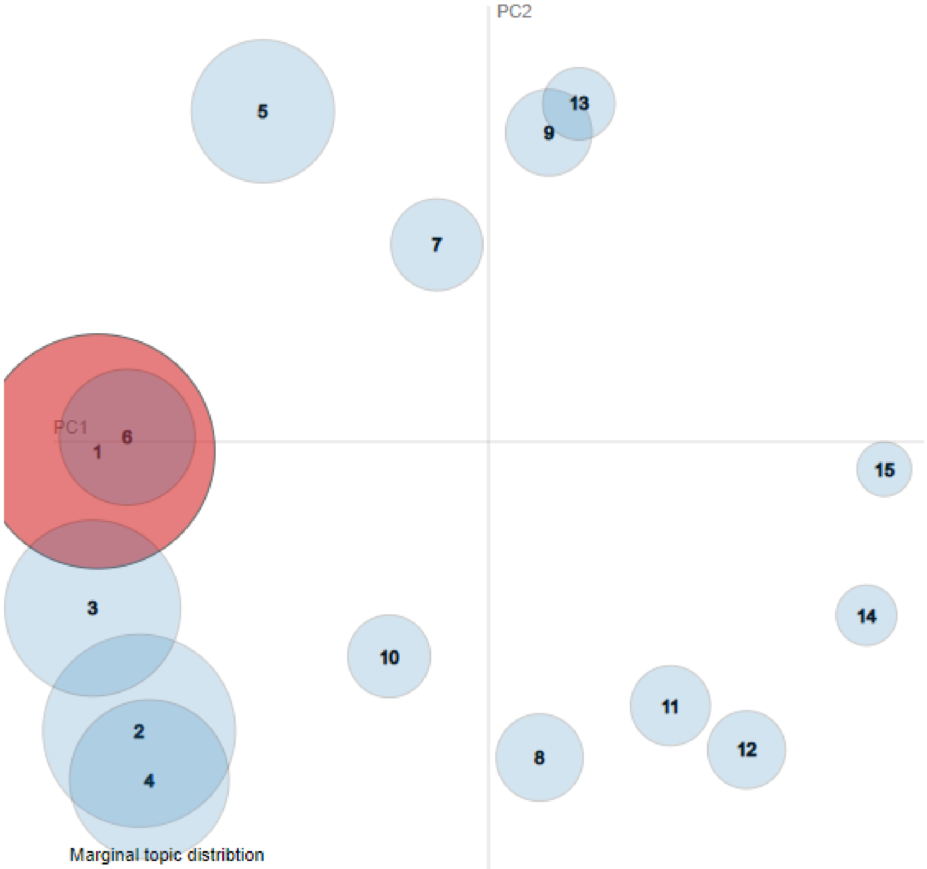




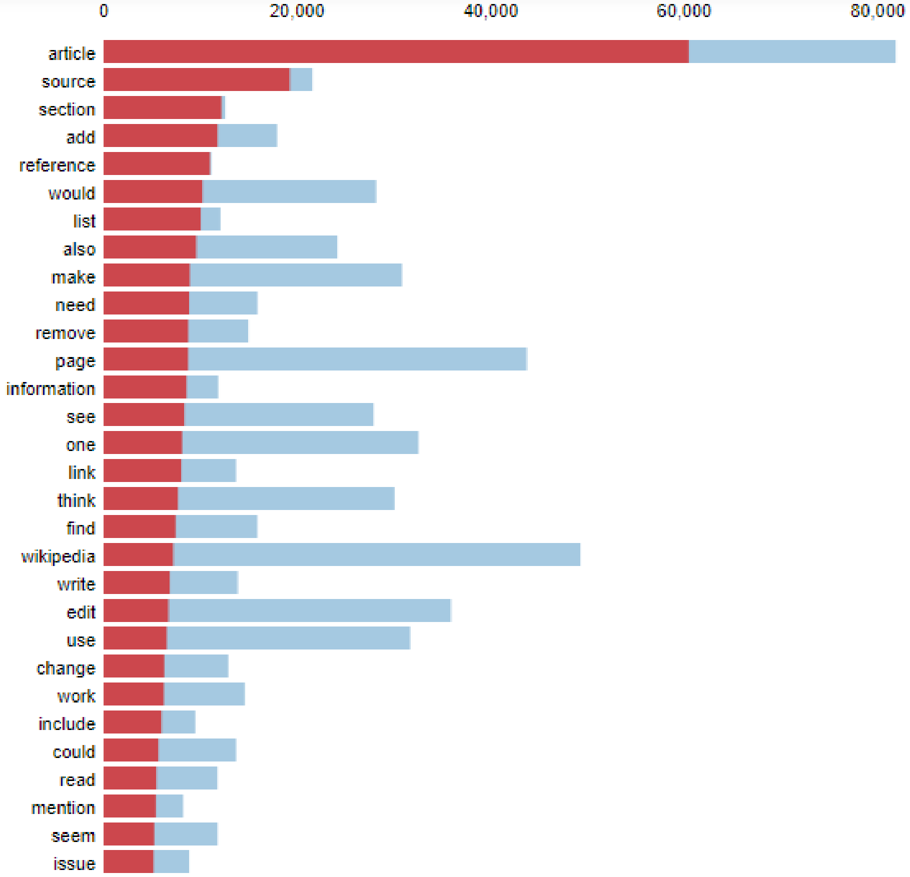
LDA Topic Modeling

- **Topic modeling** is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.
- **Latent Dirichlet Allocation** is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.
- **Gensim** is a popular open-source python toolkit for unsupervised topic modeling and natural language processing.
- **pyLDAvis** is an interactive LDA visualization python package. It is designed to help users interpret the topics in a topic model that has been to a corpus of text data. The package extracts information from a LDA topic model to inform an interactive web-based visualization.

Intertopic Distance Map



Top 30 Most Relevant Terms for Topic 1



TFIDF Vectorizer

(Term Frequency Inverse Document Frequency)

- This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction
- **Term Frequency (TF)**- The number of times a word appears in a document divided by the total number of words in the document. Every document has its own term frequency.

$$\boxed{tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}}$$

$$\boxed{idf(w) = log(\frac{N}{df_t})}$$

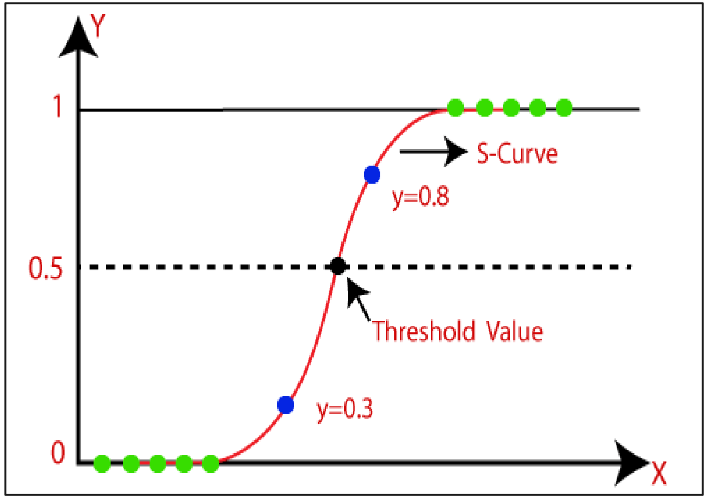
- ❖ **Inverse Document Frequency (IDF)**- The log of the number of documents divided by the number of documents that contain the word ***w***. Inverse data frequency determines the weight of rare words across all documents in the corpus.

$$\boxed{w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)}$$

- ❖ **TF-IDF** is simply the TF multiplied by IDF.

Classification Algorithms Used

Logistic Regression



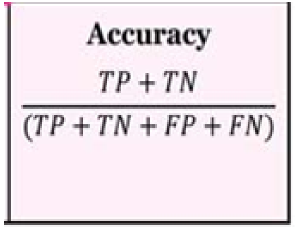
- ❖ Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable.
- ❖ The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

XGBoost

- ❖ XGBoost stands for “Extreme Gradient Boosting”.
- ❖ XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.
- ❖ It implements Machine Learning algorithms under the Gradient Boosting framework.
- ❖ It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.

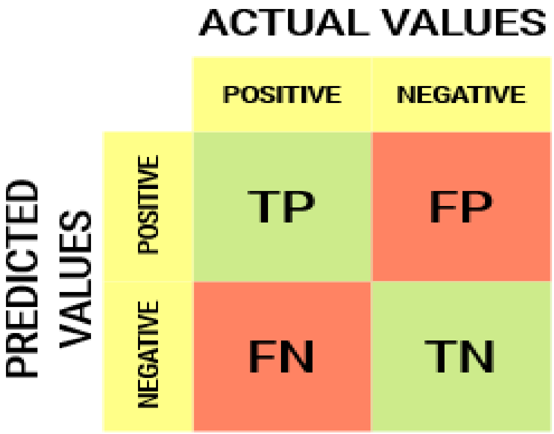
Metrics for Evaluation

- **Confusion Matrix -** It is a N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes.The matrix compares the actual target values with those predicted by the machine learning model.
- **Accuracy Score -** Classification accuracy is the ratio of correct predictions to total predictions made.
- **ROC-AUC Score** - The roc_auc_score always runs from 0 to 1, and is sorting predictive possibilities. 0.5 is the baseline for random guessing, so you want to always get above 0.5.



Accuracy

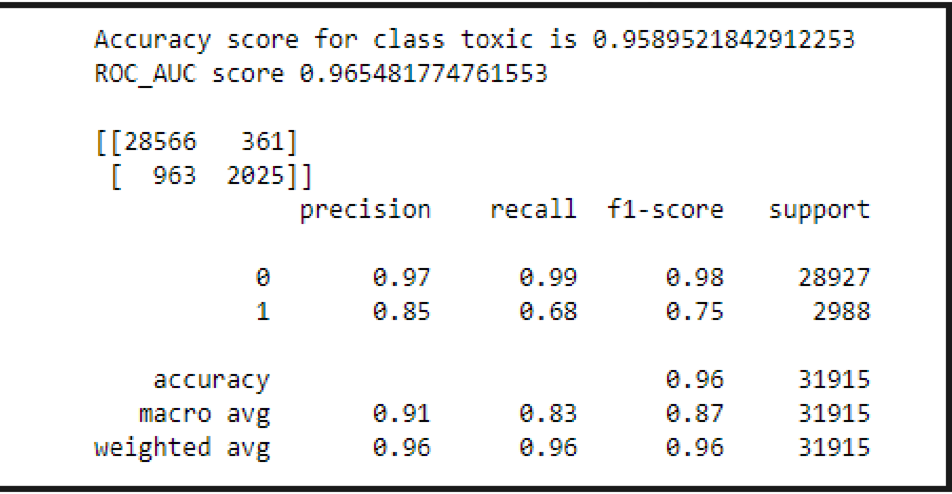
$$\frac{TP + TN}{(TP + TN + FP + FN)}$$



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Results

Logistic regression

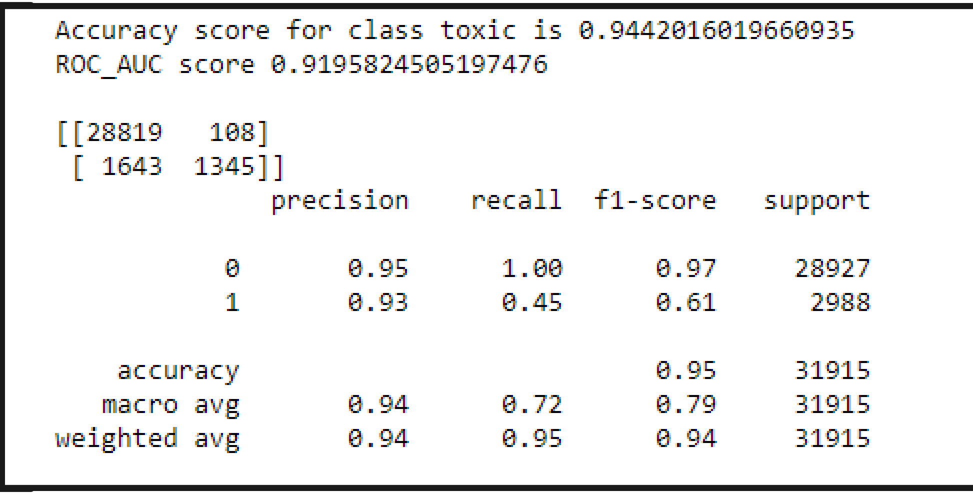


```
Accuracy score for class toxic is 0.9589521842912253
ROC_AUC score 0.965481774761553

[[28566   361]
 [  963  2025]]
              precision    recall  f1-score   support

           0       0.97      0.99      0.98     28927
           1       0.85      0.68      0.75      2988

    accuracy                           0.96     31915
   macro avg       0.91      0.83      0.87     31915
weighted avg       0.96      0.96      0.96     31915
```



```
Accuracy score for class toxic is 0.9442016919660935
ROC_AUC score 0.9195824505197476

[[28819   108]
 [ 1643  1345]]
              precision    recall  f1-score   support

           0       0.95      1.00      0.97     28927
           1       0.93      0.45      0.61      2988

    accuracy                           0.95     31915
   macro avg       0.94      0.72      0.79     31915
weighted avg       0.94      0.95      0.94     31915
```

Testing For YouTube Comments

We are using Selenium along with Google Chrome Driver to get comments from Youtube videos.

Youtube comments are dynamically loaded, which means that they are only visible when you scroll down the page. So we want a loop that will:

1. Scroll down
2. Wait for comments to appear
3. Scrape the comments
4. Repeat for whatever range we want.

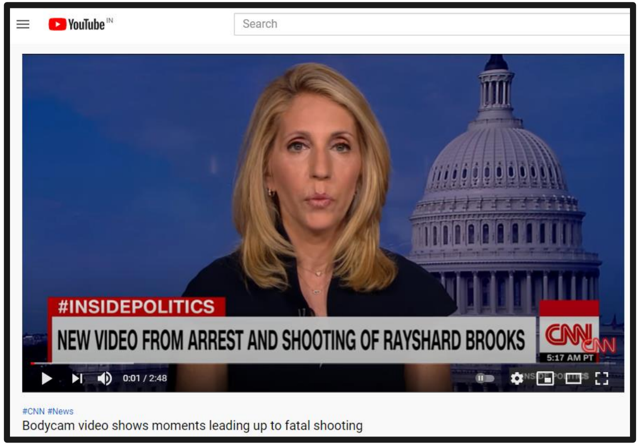
Access the URL you want with the **driver.get** function.

Scrape the comments by finding all the **#content-text** elements in the current viewed page.

We can then use our Classifier to predict if any toxic comments are present in the data.

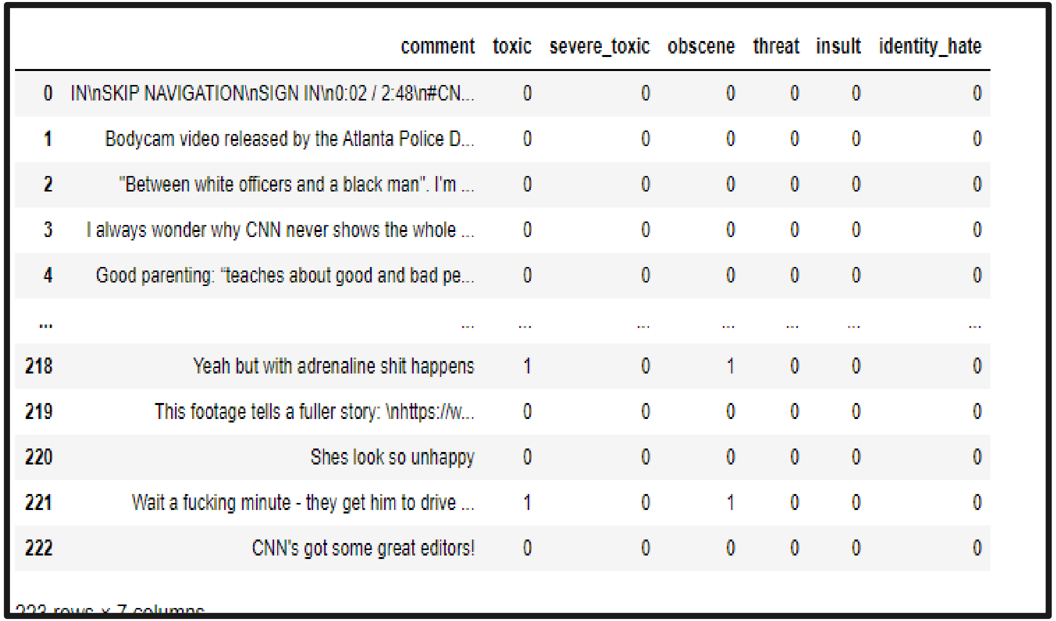
Example Output

URL : https://youtu.be/kuhhT_cBtFU



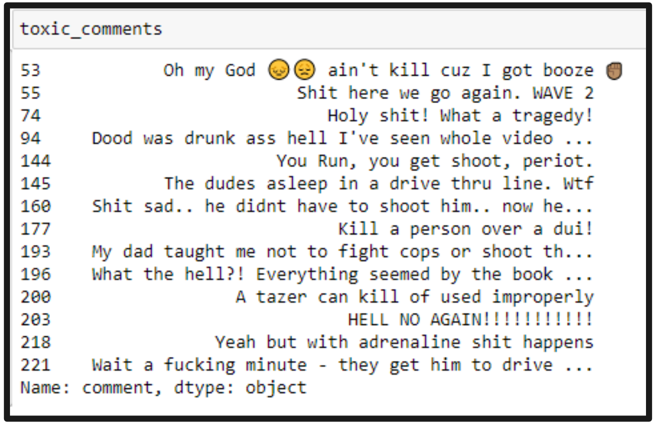
#CNN #News

Bodycam video shows moments leading up to fatal shooting



	comment	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	IN\nSKIP NAVIGATION\nSIGN IN\n0:02 / 2:48\n#CN...	0	0	0	0	0	0
1	Bodycam video released by the Atlanta Police D...	0	0	0	0	0	0
2	"Between white officers and a black man". I'm ...	0	0	0	0	0	0
3	I always wonder why CNN never shows the whole ...	0	0	0	0	0	0
4	Good parenting: "teaches about good and bad pe...	0	0	0	0	0	0
...	...	...	...	...	...	...	...
218	Yeah but with adrenaline shit happens	1	0	1	0	0	0
219	This footage tells a fuller story: \nhttps://w...	0	0	0	0	0	0
220	Shes look so unhappy	0	0	0	0	0	0
221	Wait a fucking minute - they get him to drive ...	1	0	1	0	0	0
222	CNN's got some great editors!	0	0	0	0	0	0

223 rows × 7 columns



```
toxic_comments

53             Oh my God 😔😞 ain't kill cuz I got booze ✊🏽
55                             Shit here we go again. WAVE 2
74                                Holy shit! What a tragedy!
94     Dood was drunk ass hell I've seen whole video ...
144                           You Run, you get shoot, periot.
145                The dudes asleep in a drive thru line. Wtf
160    Shit sad.. he didnt have to shoot him.. now he...
177                              Kill a person over a dui!
193    My dad taught me not to fight cops or shoot th...
196    What the hell?! Everything seemed by the book ...
200                    A tazer can kill of used improperly
203                               HELL NO AGAIN!!!!!!!!!!!
218                  Yeah but with adrenaline shit happens
221    Wait a fucking minute - they get him to drive ...
Name: comment, dtype: object
```

Conclusion

The classifiers we have used provide good accuracy for detecting toxic comments along with the levels of toxicity.

We can implement filters for popular social media sites including YouTube, Instagram and Facebook to detect and alert users if toxic comments are present in the comments section.

Practical Applications can include -

1. Semi-Automated Comment Moderation
2. Troll Detection

Future Work

Good explanations are essential in semi-automated comment moderation tools to help the moderators to make the right decision. For fully automated systems, explanations are even more critical. Moreover, with the growing number of comments on platforms without moderation, such as Facebook or Twitter, more automatic systems are needed.

Finding a balance between censorship and protecting individuals and groups on the web will be challenging. However, this challenge is not only a technical but also a societal and political one, with nothing less than democracy on the line.

References

1. Risch, Julian & Krestel, Ralf. (2020). Toxic Comment Detection in Online Discussions. 10.1007/978-981-15-1216-2_4.
2. Van Aken, B., Risch, J., Krestel, R., L¨oser, A.: Challenges for toxic comment classification: An in-depth error analysis. In: Proceedings of the Workshop on Abusive Language Online (ALW@EMNLP), pp. 33–42 (2018)
3. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International Conference on Web and Social Media (ICWSM), pp. 512–515 (2017)
4. Guberman, J., Schmitz, C., Hemphill, L.: Quantifying toxicity and verbal violence on twitter. In: Proceedings of the Conference on Computer Supported Cooperative Work (CSCW), pp. 277–280. ACM, New York, NY, USA (2016)

Thank You!