Classification of Toxic Comments

Zohreh Asaee

June 2022

DISCLAIMER AND WARNING: *Because of the nature of this project, this report contains words that are very offensive. Needless to say, it is not the intention of the author of this report, and this project, to offend anybody.*

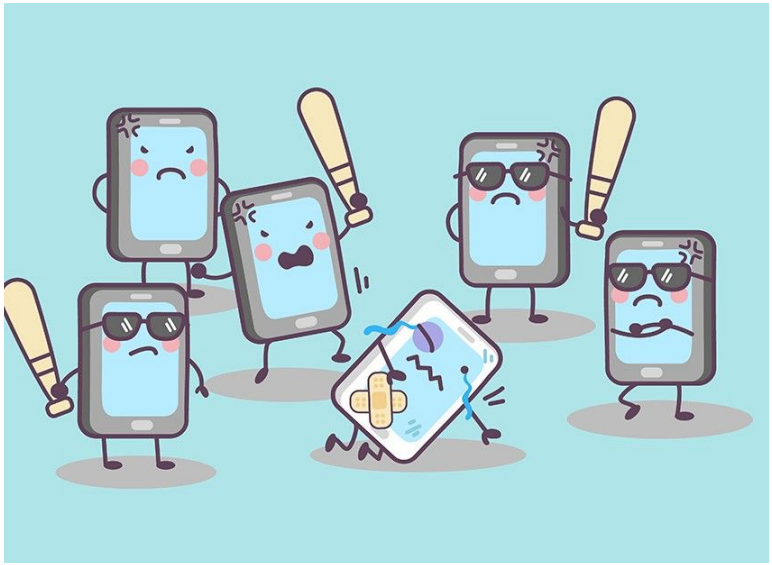
Motivation

- Social media has been growing fast and profound. The percentage of US adults who use social media increased from 5% in 2005 to 79% in 2019.

- The rapid growth of social media has raised serious concerns about the threat of abuse and harassment on these platforms.

Reference:
https://vijaygadepalli.medium.com/toxic-comments-classification-696603741872

Goals



Reference:
https://medium.com/@nehabhangale/toxic-comment-classification-models-comparison-and-selection-6c02add9d39f

- To investigate the percentage of toxic comments in the database and compare some of their features like sentence counts and word counts with the non-toxic comments
- To create a model to classify comments into non-toxic and toxic
- Perform an interpretability analysis and specify features impacting each class of non-toxic and toxic

Data Acquisition

Source of data: two datasets of training and test from Jigsaw toxic comment classification challenge[1] on Kaggle

Data - attributes:

- id
- comment text
- Toxic
- Severe toxic
- Obscene
- Threat
- Insult

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I ...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits in...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on improvement - I wondered if the section statistics should be later on, or a subsection of ""types of...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember what page that's on?	0	0	0	0	0	0

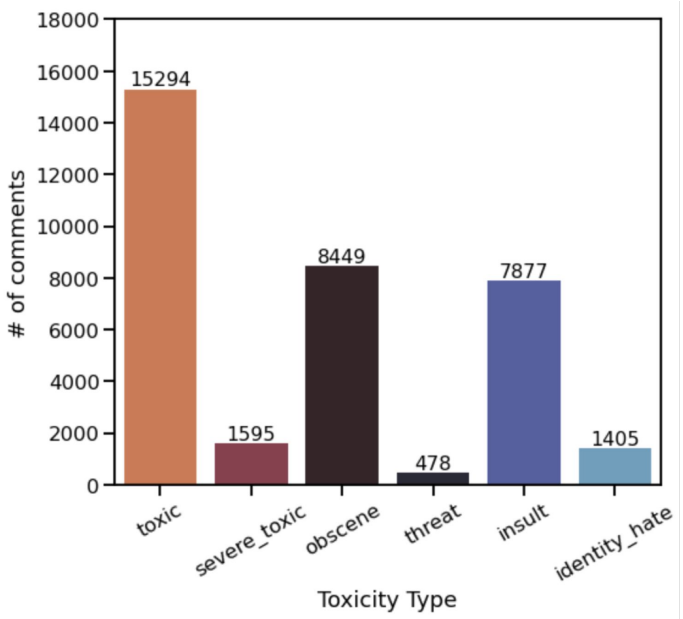
Train dataset: 159,571 rows
Test dataset: 63,978 rows

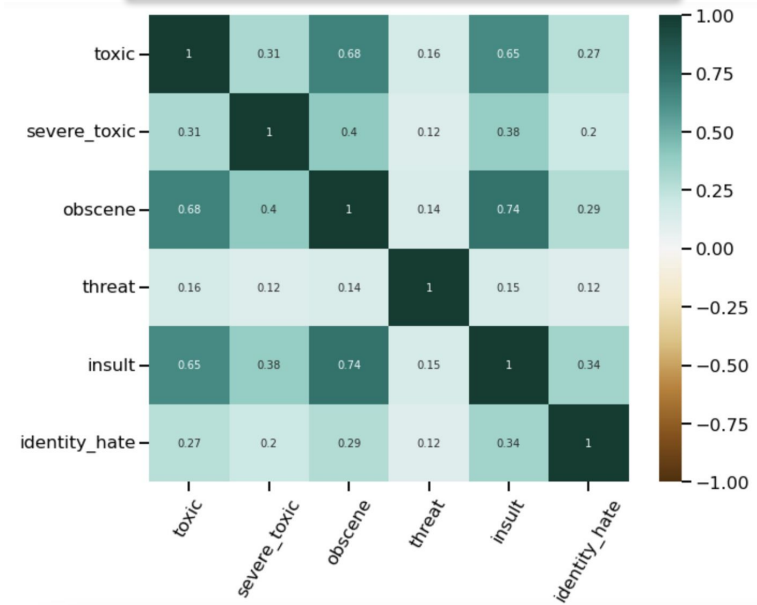
[1] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

EDA and Feature Engineering

- The highest correlation is between toxic, obscene, and insult classes.
- The threat category has the least correlation with the other categories.

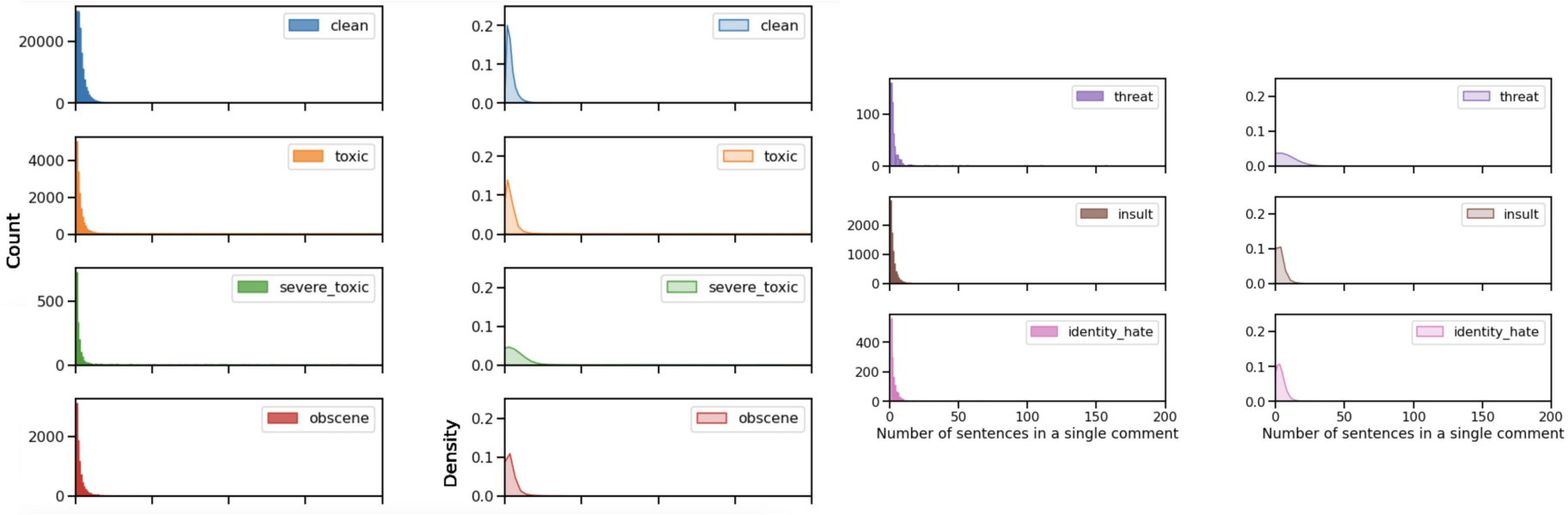
Non-toxic comments: 143,346
Toxic comments: 16,225





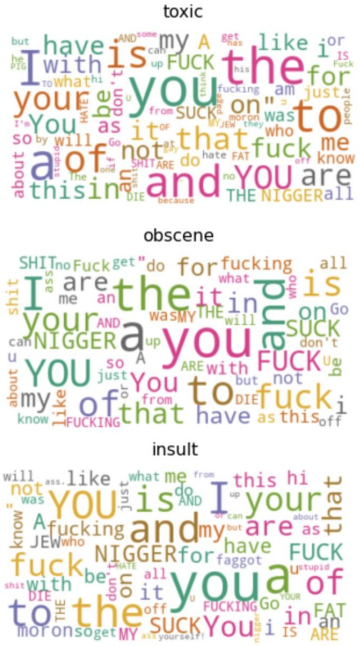
EDA and Feature Engineering - Sentence Counts

- All categories are positively skewed. In general, the toxic comments tend to be longer than the non-toxic comments.



EDA and Feature Engineering – Word Cloud

- The word cloud results reveal that the text is noisy, containing abbreviations, repetitions, and pause-filling words.
- We perform the following steps to process and clean the text:
  - remove special characters and HTML tags,
  - lemmatization,
  - remove stop words,
  - case conversion.



<u>DISCLAIMER AND WARNING</u>: *Because of the nature of this project, this report contains words that are very offensive. Needless to say, it is not the intention of the author of this report, and this project, to offend anybody.*

EDA and Feature Engineering – Word Cloud

Clean

toxic

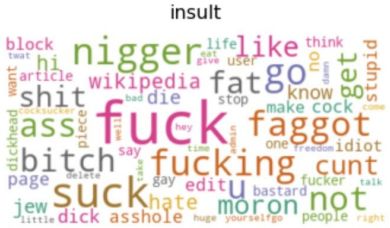
severe_toxic

obscene

identity_hate

threat

insult



DISCLAIMER AND WARNING: *Because of the nature of this project, this report contains words that are very offensive. Needless to say, it is not the intention of the author of this report, and this project, to offend anybody.*

Modeling – Baseline Model: *Logistic Regression*

	BOW Term Frequency	*TF-IDF*
Mean fit time	*10.43*	*3.05*
Mean test score	*0.64*	*0.68*

Confusion Matrix for the test dataset- BOW model

	precision	recall	f1-score	support
0	0.97	0.94	0.95	57298
1	0.57	0.73	0.64	6243
accuracy			0.92	63541
macro avg	0.77	0.84	0.80	63541
weighted avg	0.93	0.92	0.92	63541

Confusion Matrix for the test dataset - TF-IDF model

	precision	recall	f1-score	support
0	0.97	0.95	0.96	57298
1	0.62	0.76	0.68	6243
accuracy			0.93	63541
macro avg	0.80	0.85	0.82	63541
weighted avg	0.94	0.93	0.93	63541

Modeling – Extended Modeling

Classification Algorithms:

- Logistic Regression
- Random Forest Classifier
- XGBOOST Classifier
- LGBM Classifier
- Naive Bayes
- Ensemble Stacking Models

Imbalance Algorithms:

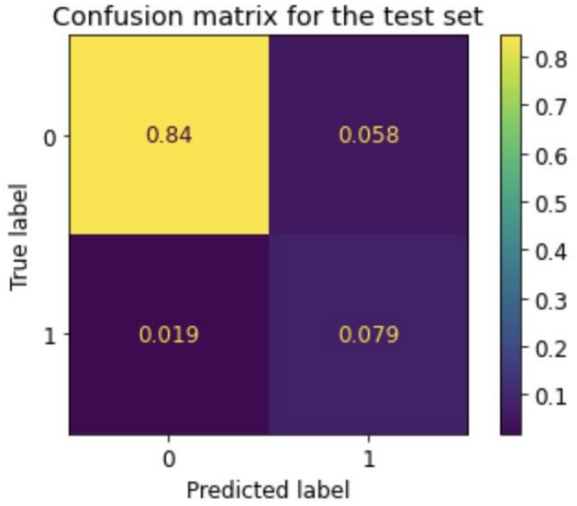
- Random Oversampling
- Synthetic Minority Oversampling (SMOTE)
- Random Undersampling
- Near Miss

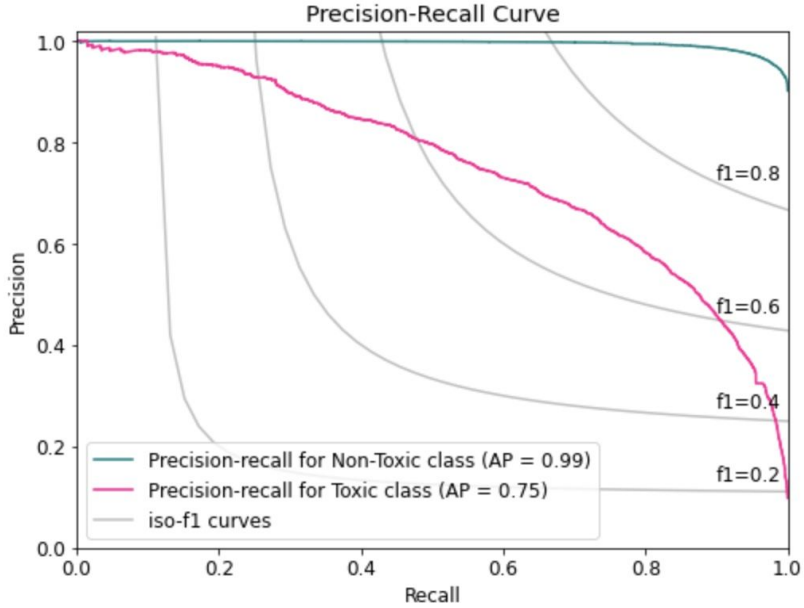
Hyperparameter Tuning:

- BayesSearchCV

Modeling – Extended Modeling

The LGBM model has the **highest *f1*-score** of 68% among all the trained models. In this model, the number of false positives is greater than false negatives.

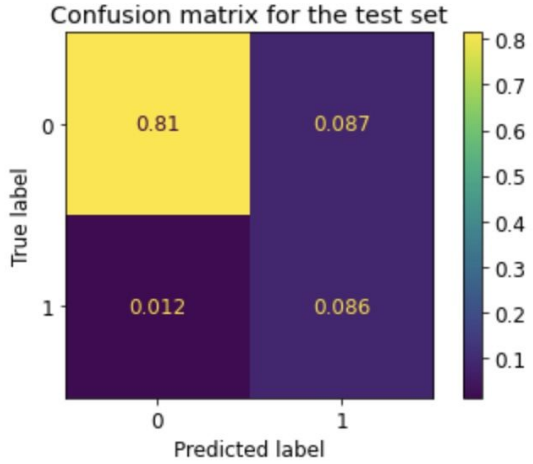


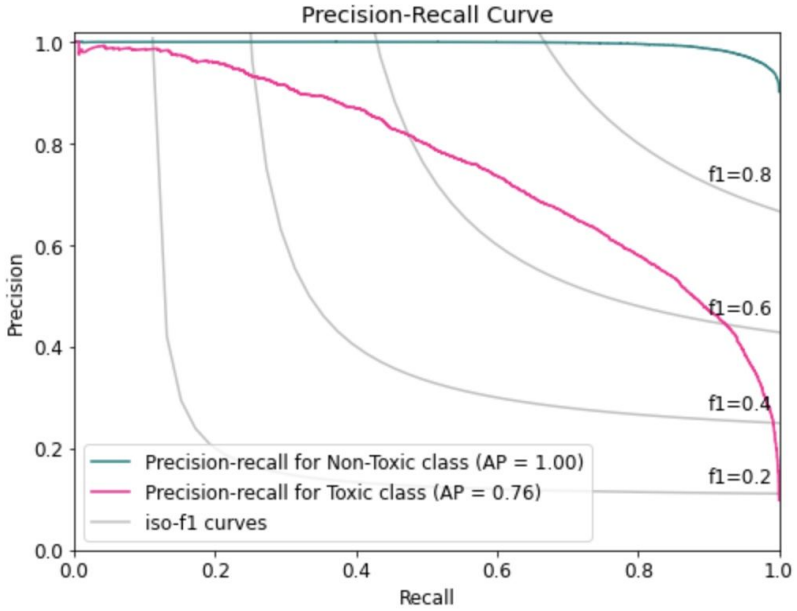


LGBM for the test dataset
F1-score: 0.68

Modeling – Extended Modeling

Across all the trained models, the **logistic regression models with resampled dataset** have the largest recalls. The logistic regression with the random undersampling technique has **86% recall**. The logistic regression and random undersampling model has higher false positives and lower false negatives compared to the LGBM model.

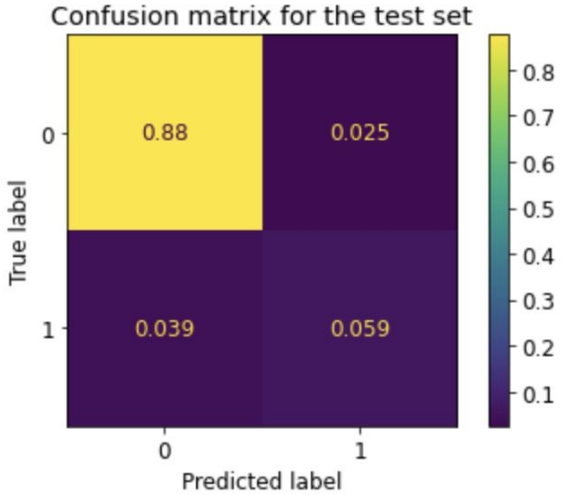


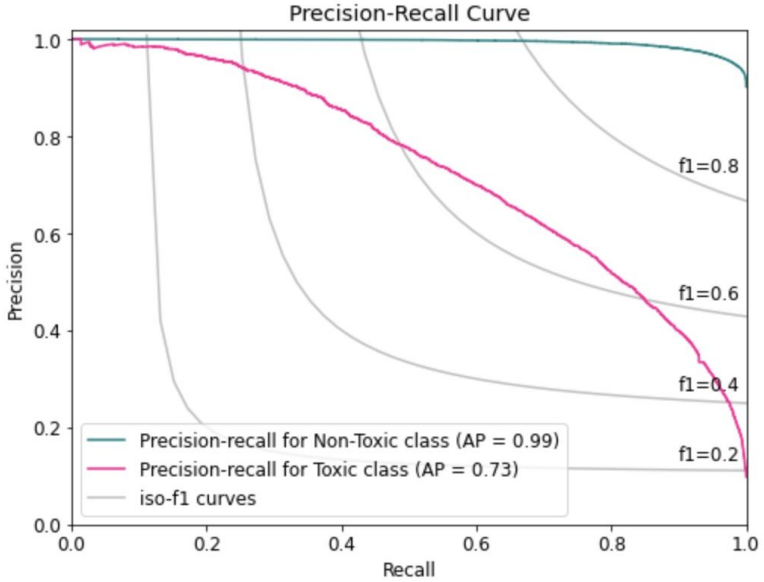


Logistic Regression with Random Oversampling for the test dataset - Recall: 0.86

Modeling – Extended Modeling

The **Naive Bayes model** has the **highest precision of 70%** amongst all trained models.





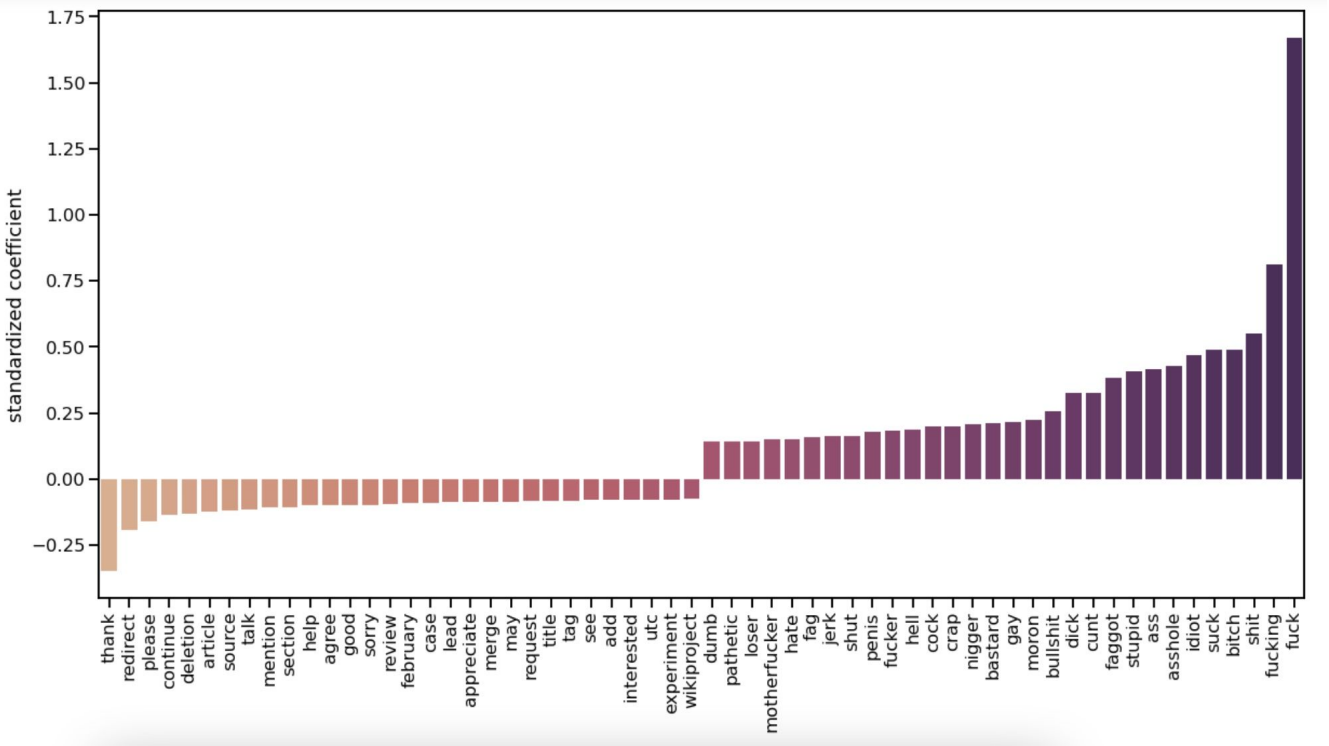
Naive Bayes Model
for the test dataset - Precision: 0.70

Feature Importance Analysis

Interpretable Model for the test dataset: Logistic Regression

The negative and positive coefficients belong to the non-toxic and toxic classes, respectively.

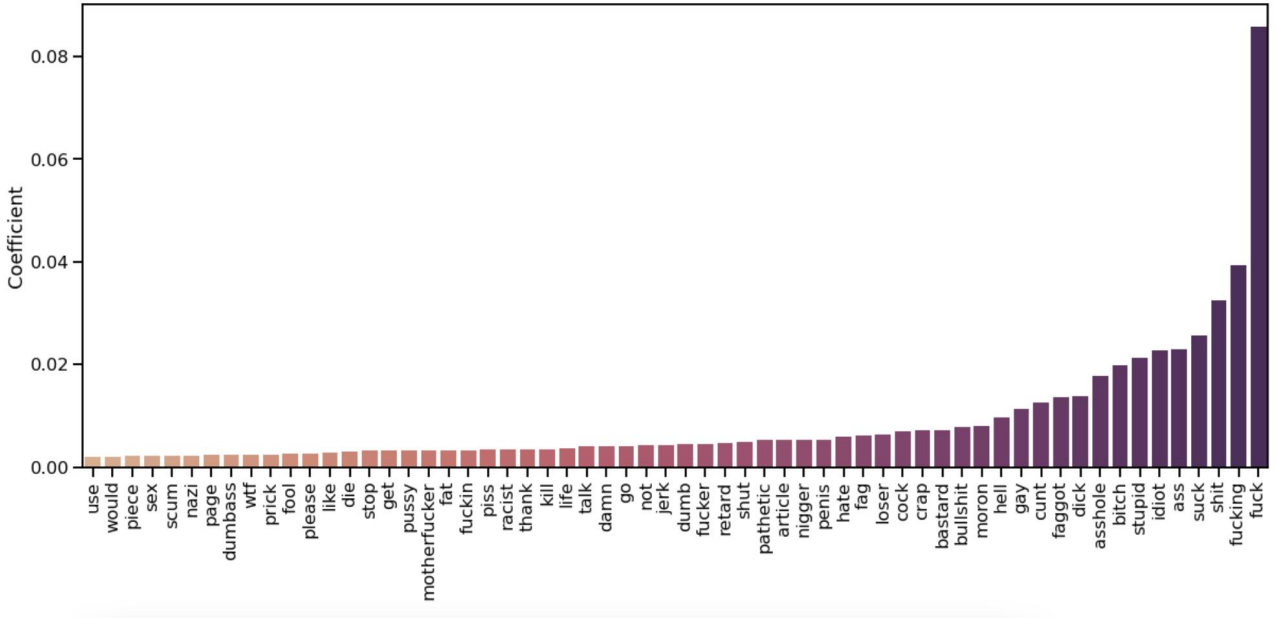
The coefficients of features impacting the toxic class are considerably larger than (three to five folds) those for the non-toxic class.



Feature Importance Analysis

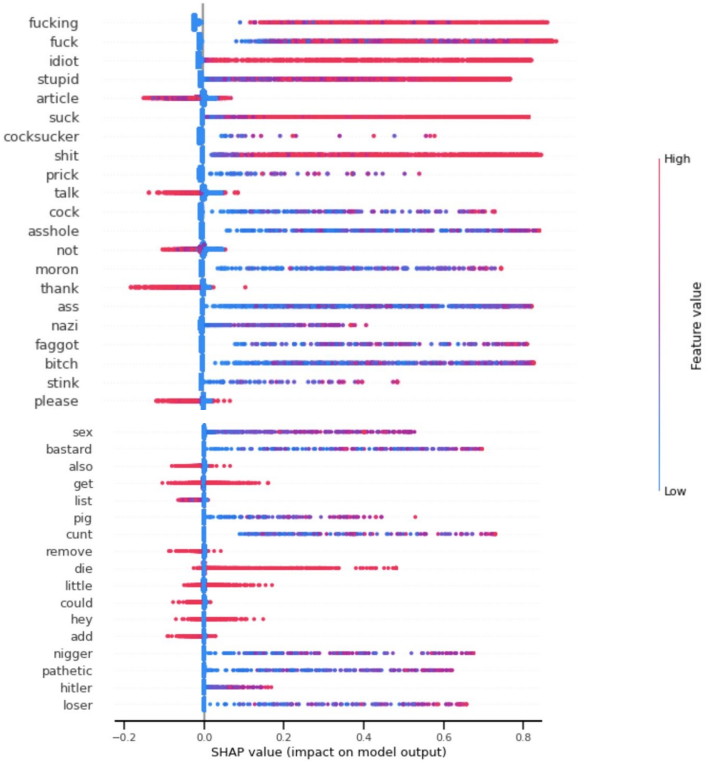
Interpretable Model for the test dataset: Random Forest

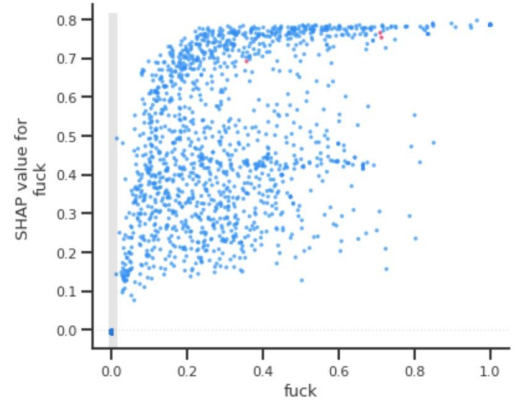
The features from the random forests model are not separated based on their impact on the class. However, the first few features with a considerable value impact detection of the toxic class.

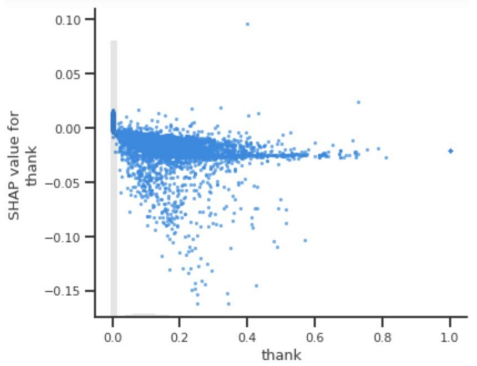


Feature Importance Analysis

Interpretable ML Model for the test dataset: Random Forest and SHAP Technique



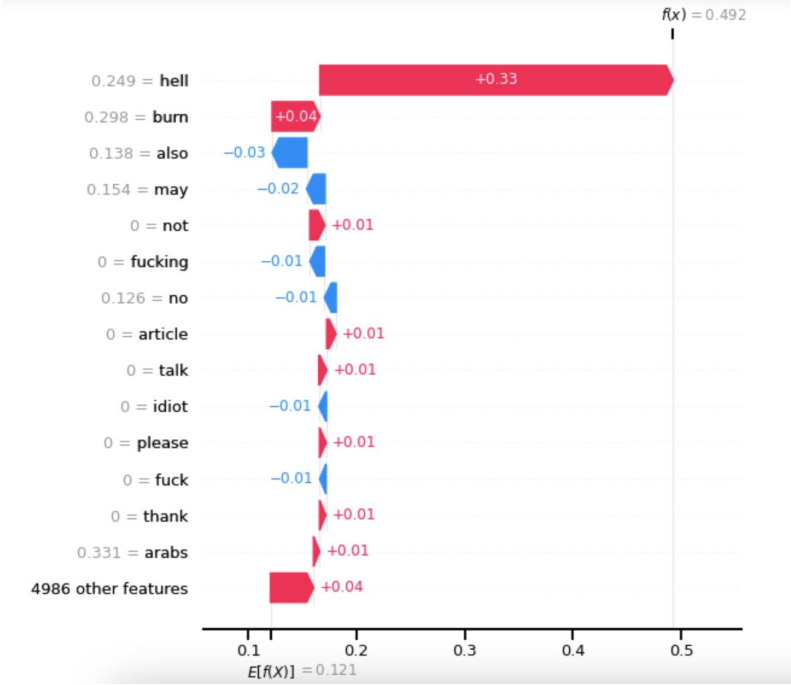




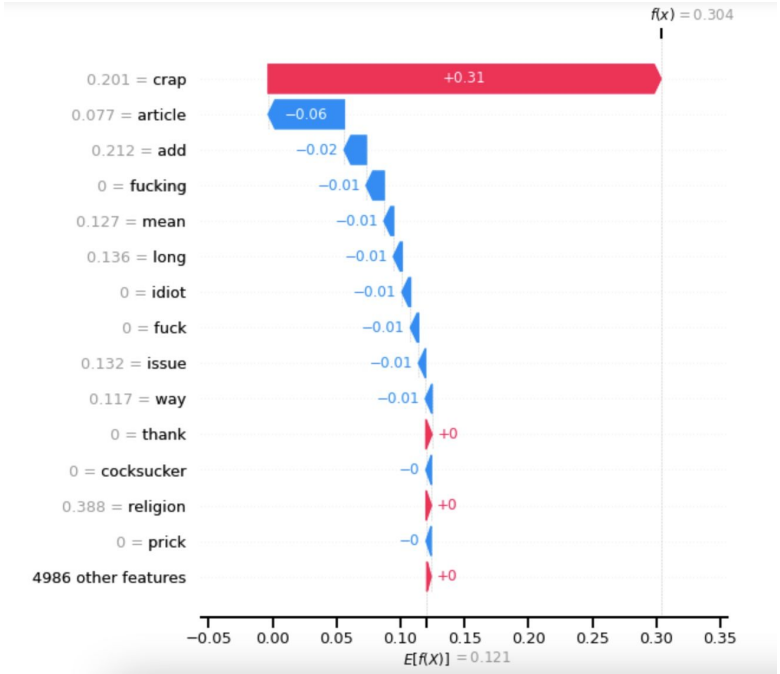
Feature Importance Analysis

Interpretable ML Model for the test dataset: Random Forest and SHAP Technique

```
arabs commit genocide iraq no protest europe may europe also burn hell
Class: 1
```



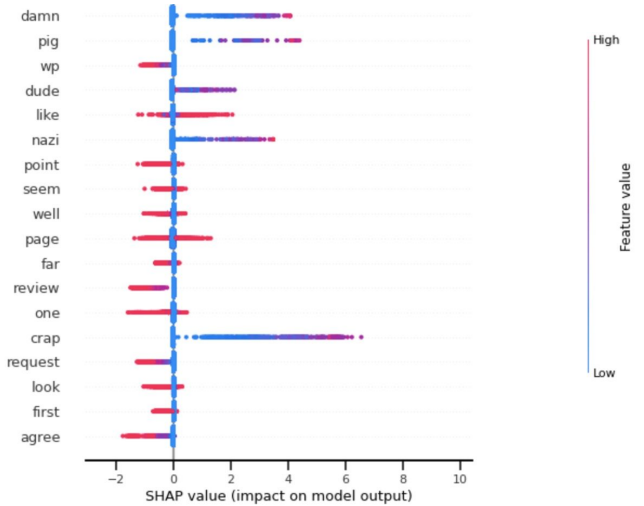

```
somebody invariably try add religion really mean way people invariably keep add religion samuel beckett infobox bothe
r bring long dead completely non existent influence issue flail make crap fly comparison explicit acknowledgement ent
ire amos oz article personally jewish category
Class: 0
```



Feature Importance Analysis

Interpretable ML Model for the test dataset: XGBOOST and SHAP Technique

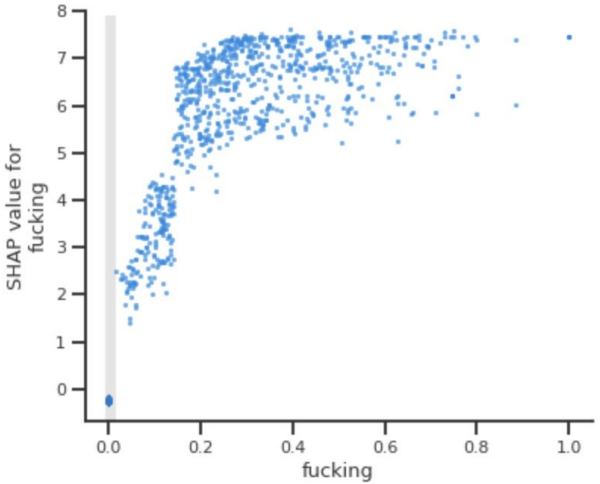
The positive SHAP values interpret the toxic class, whereas the positive values detect the non-toxic class. Similar to the random forests model, the SHAP value of features detecting the toxic class is considerably larger than those impacting the non-toxic class. The most important features of the XGBOOST model are similar to those of the random forests model.

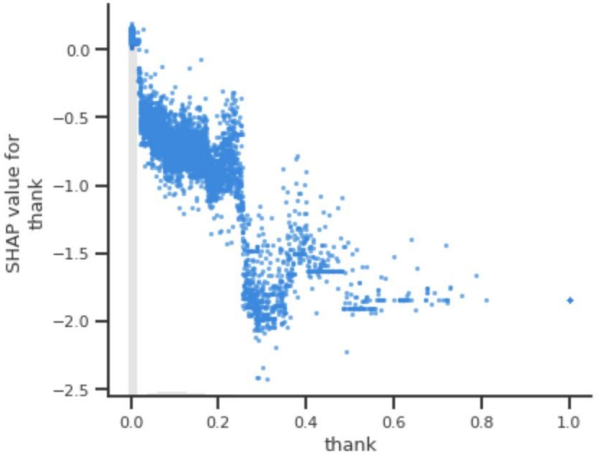


Feature Importance Analysis

Interpretable ML Model for the test dataset: XGBOOST and SHAP Technique

For both features, an increase in the feature value results in an increase in the SHAP value. Also, the pattern is not as sparse as in the previous analysis of the random forest model. Moreover, the zero feature value shifts the output towards the opposite side for both features.



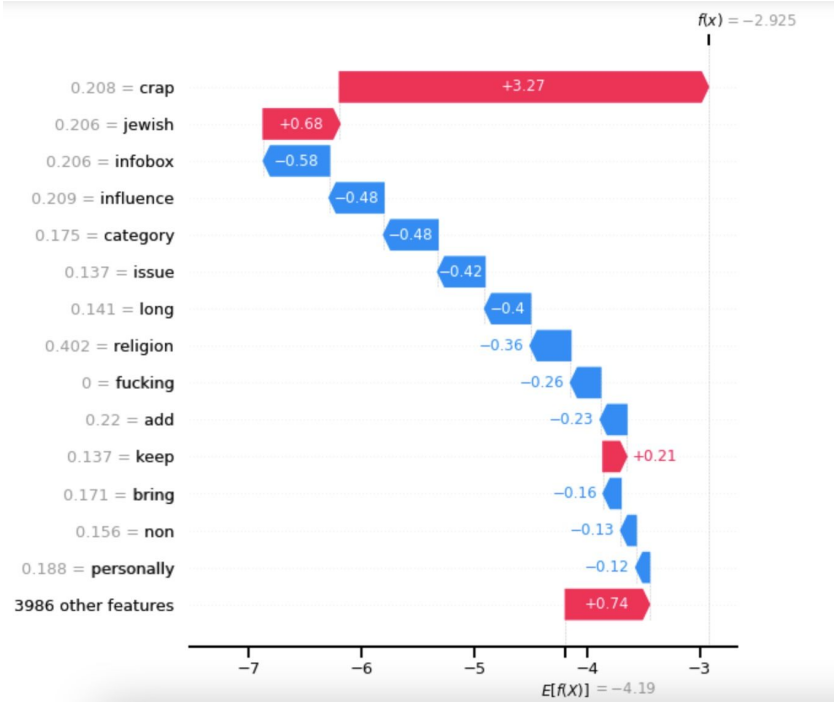


Feature Importance Analysis

Interpretable ML Model: XGBOOST and SHAP Technique

This comment is misinterpreted by the previous model (SHAP and random forest) as toxic. The plot reveals that the “cr-p” feature moves the output to the positive value; however, the other features shift it towards the negative SHAP value and interpret it as non-toxic.

somebody invariably try add religion really mean way people invariably keep add religion samuel beckett infobox bother bring long dead completely non existent influence issue flail make crap fly comparison explicit acknowledgement entire amos oz article personally jewish category
Class: 0



Conclusion

- several ML models (Logistic Regression, Random Forest, Naive Bayes, XGBOOST, LGBM Classifier, and Ensemble Stacking models) have been tested. Resampling techniques (Random Oversample, SMOTE, Random Undersample, and Near Miss) have been used to balance the dataset.

- Among all tested models, the LGBM classifier showed the best performance in terms of *f1*-score (0.68). However, the logistic regression in conjunction with random undersampling had 0.86 recall. Therefore, based on the importance of toxicity detection and application, either of these models is recommended.

- Finally, the interpretability analysis showed that the features (or words) impacting the model output as toxic have a considerably larger SHAP value and will help to interpret the toxicity of a comment. It was shown that the SHAP model could detect the toxicity of comments even with a small feature value.

Future Work

- As for future work, It is recommended to test models based on anomaly detection algorithms. We observed that the number of false positives was considerable, and an anomaly detection algorithm will help to identify the outliers and point out where an error is occurring.

- The other suggestion is performing the Recurrent Neural Network (RNN) model. The RNN-based models perform well for text processing problems.

Recommendations for Client

- In creating recommendations for our results, it is essential to understand the application of this specific platform and the age range of users. In platforms with youth age range users, our goal is to block toxic comments as many as possible. Our recommended model can block 86% percent of the toxic comments. However, only 58% of the blocked comments are toxic. It means that 42% of the blocked comments are clean and non-toxic. Moreover, our results prove a prevalence of some words with the toxicity of comments. The other recommendation can be blocking comments consisting of these words.

- If we are willing to detect toxic comments and block the corresponding accounts, it is essential to decrease the false detections since we don't want to block an account by mistake. On the other hand, it is crucial to detect toxic comments and block them. In this case, our recommended model can detect 81% of the toxic comments, while only 41% of the blocked comments are clean and non-toxic.