DS-UA 202 Group Project Final Report

Boyoon Han and Daphne Ozkan

May 9, 2025

1 Background

a.

The Automated Decision System (ADS) we audited is a multi-label text classification model designed to detect toxic and inappropriate comments online. The model was submitted to the Jigsaw Unintended Bias in Toxicity Classification competition on Kaggle. Its stated purpose is to identify various types of harmful content such as toxicity, severe toxicity, obscenity, threats, insults, and identity-based hate speech, enabling content moderation systems to better protect users while preserving free expression.

The system has multiple goals: (1) accurately detect and classify different types of toxic content, and (2) do so without disproportionately flagging neutral or positive comments that mention protected identities (e.g., "I am Muslim").

b.

Because the system has multiple goals, it introduces a trade-off between precision and fairness. A system optimized solely for accuracy may have high recall but risk over-flagging benign identity-related comments, while a fairness-optimized system might miss harmful content in order to avoid appearing biased.

2 Input and Output

a.

The model uses the Jigsaw Toxic Comment Classification dataset, which includes 159,571 comments from Wikipedia talk pages. Each comment is labeled with six binary target variables: `toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, and `identity_hate`. Comments are provided in raw text form, and each entry also has a unique ID. There are no missing values in the dataset. This dataset was originally curated for a Kaggle competition and reflects typical online user-generated content.

b.

The model uses both text-based features and numerical features:

- `comment_text` (string): Raw user comment. No missing values.
- `total_length` (integer): Number of characters in the comment. Right-skewed distribution with a long tail.

- uppercase (integer): Count of uppercase letters.
- exclamation_punctuation (integer): Count of exclamation marks.
- num_punctuation (integer): Count of general punctuation characters.
- num_symbols (integer): Count of special characters.
- num_words (integer): Count of words in the comment. Added for exploratory profiling. Right-skewed.

The dataset is notably imbalanced, with 89.83% of comments labeled as non-toxic. The next most frequent label combinations occur in fewer than 4% of examples, and many combinations appear in less than 0.01% of the data. For instance, several rare toxic combinations involving five or six labels occur in fewer than 0.002% of instances. This class imbalance poses a risk of biased model behavior, particularly when learning from underrepresented co-occurring toxic labels. Additionally, only a subset of the test set is used for official evaluation, which may impact how models are validated and selected.

To better understand the input space, we profiled the numerical features using histograms and examined their relationships using a Spearman correlation heatmap. A high correlation (0.99) was found between num_words and total_length, which could lead to multicollinearity. We included num_words for analysis and dropped it before modeling, as done in the original notebook.

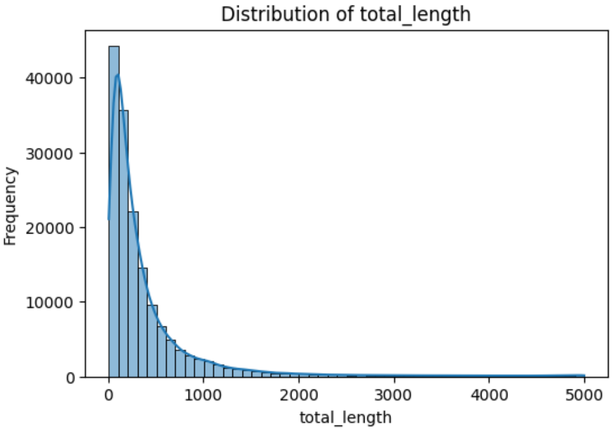
We also applied PCA and t-SNE to TF-IDF-transformed text data to assess whether toxic and non-toxic comments could be linearly or non-linearly separable in reduced-dimensional space. As shown in Figures **??** and **??**, while t-SNE revealed some local structure among toxic comments, both methods showed substantial overlap between classes. This highlights the complexity of the classification task and the challenges in cleanly separating toxic content from benign comments using feature space alone.

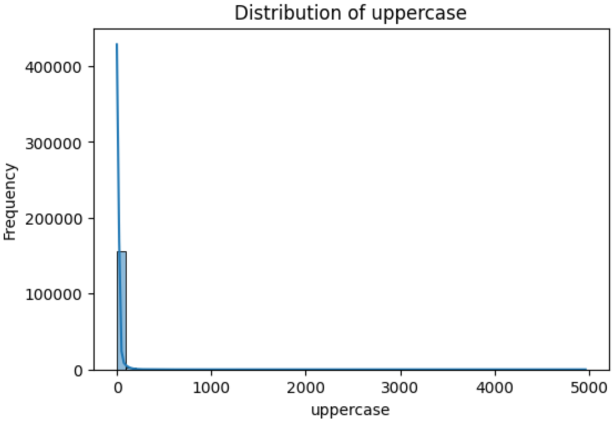
To better understand the feature space, we profiled each numerical input using histograms and explored pairwise relationships with a Spearman correlation heatmap. We found a near-perfect correlation ($\rho = 0.99$) between num_words and total_length. Although the original notebook had already dropped num_words to avoid multicollinearity, we temporarily reintroduced it for correlation analysis. As expected, it was highly redundant with total_length and was removed again before modeling.

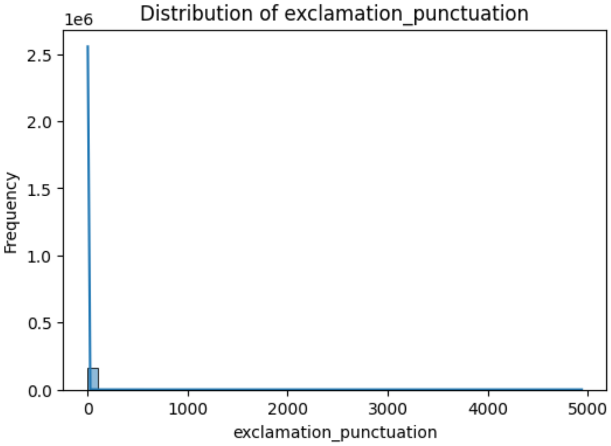
To assess potential separability in the text space, we applied two dimensionality reduction techniques, PCA and t-SNE, on TF-IDF vector representations of the comments. As shown in the figures below, PCA reveals a dense overlapping cluster with a slight separation of some toxic comments, while t-SNE creates a more circular projection with a small central region of toxic comments. These visualizations confirm that toxic and non-toxic comments do not form clearly separable clusters, which reflects the difficulty of the classification task and the nuanced nature of online toxicity.

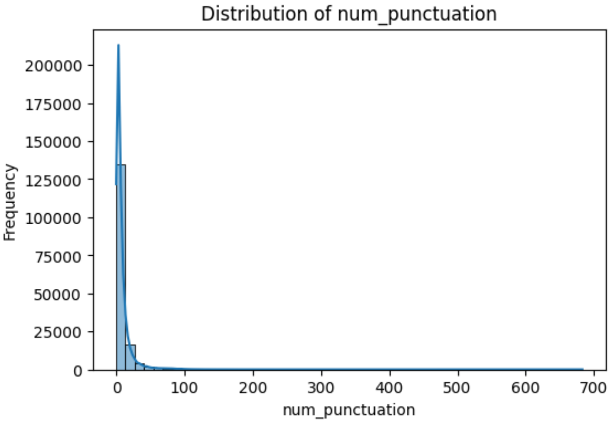
c.

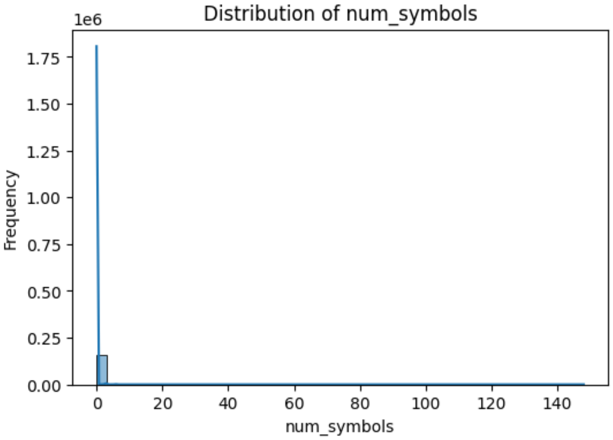
The model outputs a probability in [0, 1] for each of the six toxic comment labels. A threshold of 0.5 is applied to determine binary classifications. Since this is a multi-label classification task, a comment can receive multiple positive labels. The outputs are interpretable as the model's confidence in the presence of each type of toxicity in the input comment.

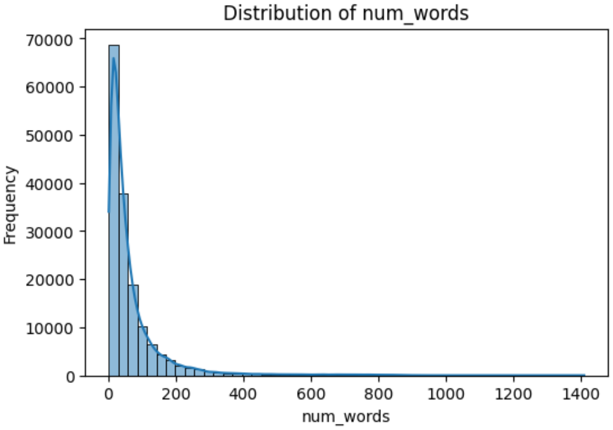

Distribution of total_length
Frequency
40000
30000
20000
10000
0
0
1000
2000
3000
4000
5000
total_length

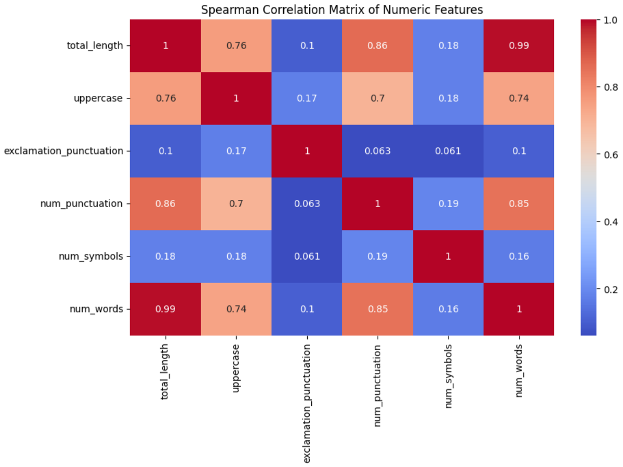

Distribution of uppercase
Frequency
400000
300000
200000
100000
0
0
1000
2000
3000
4000
5000
uppercase

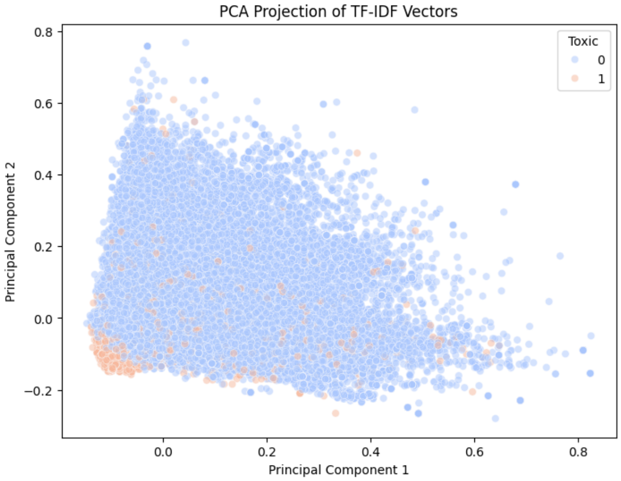

Distribution of exclamation_punctuation
1e6
Frequency
2.5
2.0
1.5
1.0
0.5
0.0
0
1000
2000
3000
4000
5000
exclamation_punctuation

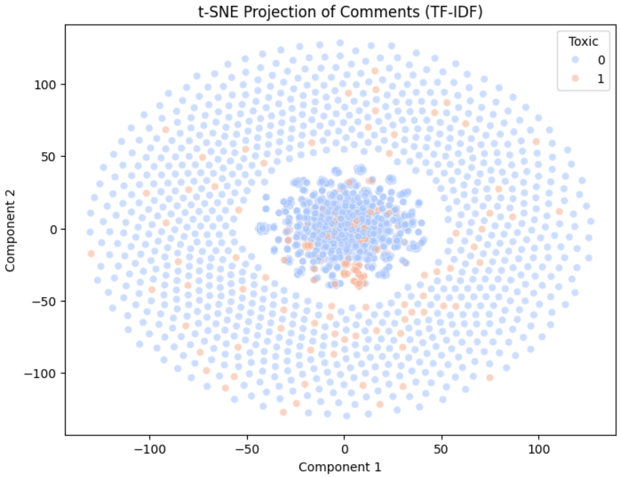

Distribution of num_punctuation
Frequency
200000
175000
150000
125000
100000
75000
50000
25000
0
0
100
200
300
400
500
600
700
num_punctuation

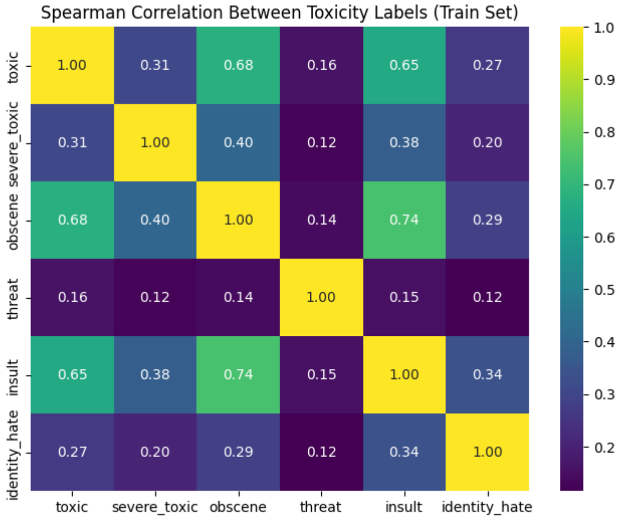

Distribution of num_symbols
1e6
Frequency
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
100
120
140
num_symbols

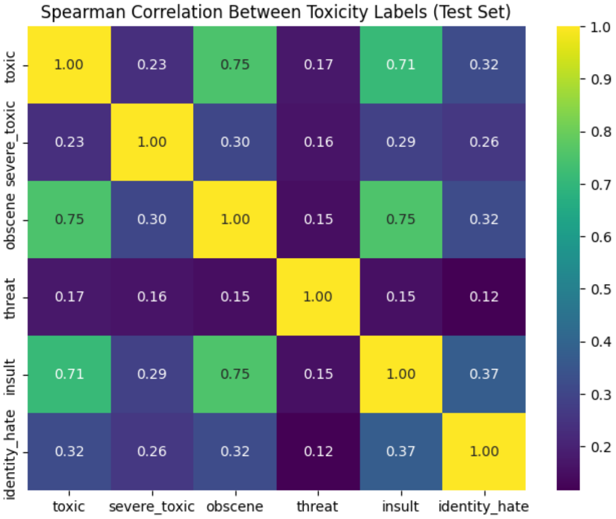

Distribution of num_words
Frequency
70000
60000
50000
40000
30000
20000
10000
0
0
200
400
600
800
1000
1200
1400
num_words


Spearman Correlation Matrix of Numeric Features
total_length
uppercase
exclamation_punctuation
num_punctuation
num_symbols
num_words
1 0.76 0.1 0.86 0.18 0.99
0.76 1 0.17 0.7 0.18 0.74
0.1 0.17 1 0.063 0.061 0.1
0.86 0.7 0.063 1 0.19 0.85
0.18 0.18 0.061 0.19 1 0.16
0.99 0.74 0.1 0.85 0.16 1
total_length
uppercase
exclamation_punctuation
num_punctuation
num_symbols
num_words
1.0
0.8
0.6
0.4
0.2

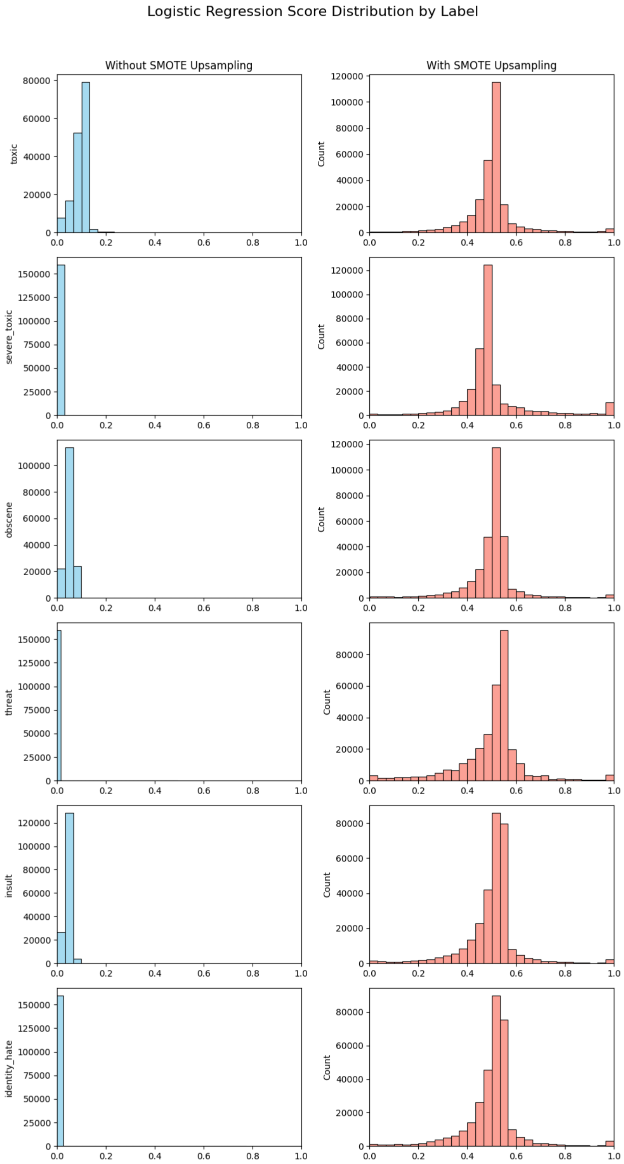

PCA Projection of TF-IDF Vectors
Toxic
0
1
Principal Component 2
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
Principal Component 1

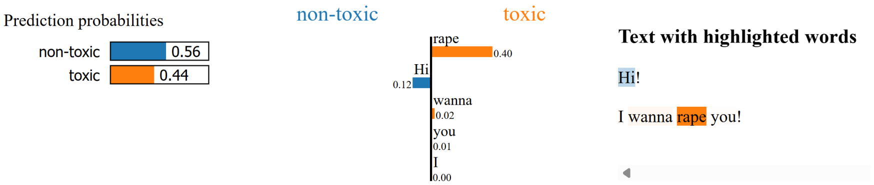

t-SNE Projection of Comments (TF-IDF)
Toxic
0
1
Component 2
100
50
0
−50
−100
Component 1
−100
−50
0
50
100


Spearman Correlation Between Toxicity Labels (Train Set)
toxic
severe_toxic
obscene
threat
insult
identity_hate
1.00 0.31 0.68 0.16 0.65 0.27
0.31 1.00 0.40 0.12 0.38 0.20
0.68 0.40 1.00 0.14 0.74 0.29
0.16 0.12 0.14 1.00 0.15 0.12
0.65 0.38 0.74 0.15 1.00 0.34
0.27 0.20 0.29 0.12 0.34 1.00
toxic
severe_toxic
obscene
threat
insult
identity_hate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2



	toxic	severe_toxic	obscene	threat	insult	identity_hate
toxic	1.00	0.23	0.75	0.17	0.71	0.32
severe_toxic	0.23	1.00	0.30	0.16	0.29	0.26
obscene	0.75	0.30	1.00	0.15	0.75	0.32
threat	0.17	0.16	0.15	1.00	0.15	0.12
insult	0.71	0.29	0.75	0.15	1.00	0.37
identity_hate	0.32	0.26	0.32	0.12	0.37	1.00

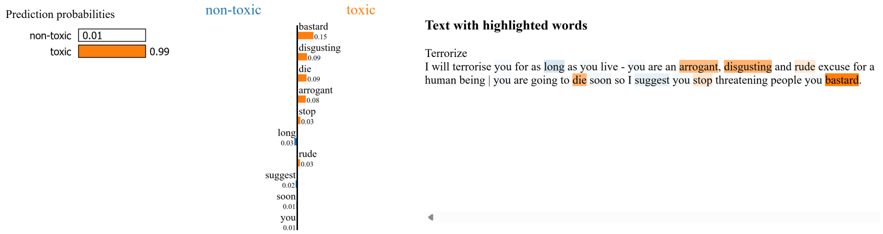




Text with highlighted words

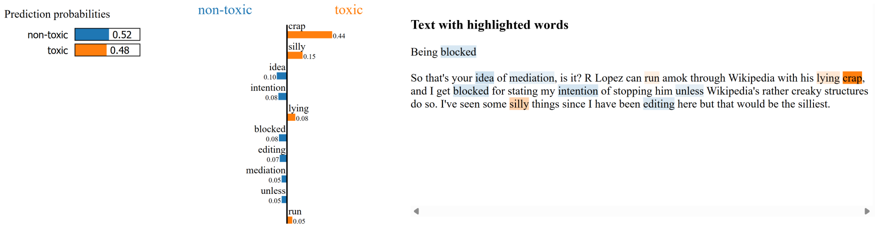
Hi!

I wanna rape you!



Text with highlighted words

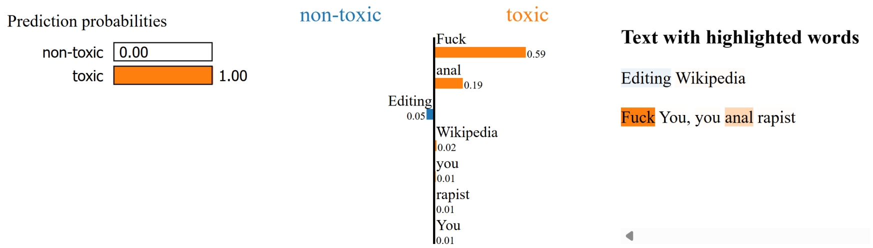
Terrorize
I will terrorise you for as long as you live - you are an arrogant, disgusting and rude excuse for a human being | you are going to die soon so I suggest you stop threatening people you bastard.



Text with highlighted words

Being blocked

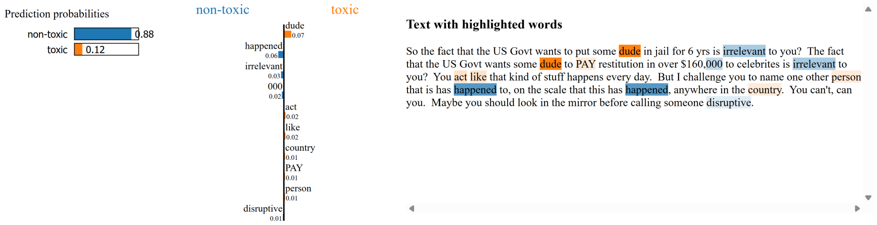
So that's your idea of mediation, is it? R Lopez can run amok through Wikipedia with his lying crap, and I get blocked for stating my intention of stopping him unless Wikipedia's rather creaky structures do so. I've seen some silly things since I have been editing here but that would be the silliest.



Text with highlighted words

Editing Wikipedia

Fuck You, you anal rapist



Text with highlighted words

So the fact that the US Govt wants to put some dude in jail for 6 yrs is irrelevant to you? The fact that the US Govt wants some dude to PAY restitution in over $160,000 to celebrites is irrelevant to you? You act like that kind of stuff happens every day. But I challenge you to name one other person that is has happened to, on the scale that this has happened, anywhere in the country. You can't, can you. Maybe you should look in the mirror before calling someone disruptive.

3. Implementation and Validation

We primarily use the notebook submission by collaborators Catherine(owner) and Egor Petrucho (editor) in response to the Kaggle competition as a guideline for our model implementation. The pipeline consists of preprocessing the text, transforming it with a TF-IDF vectorizer, and applying Logistic Regression in a multi-label classification setting using a binary relevance approach. Based on our research goals, we also introduced several improvements to the original implementation.

The raw comment text is first preprocessed by lowercasing, removing punctuation and numbers, and tokenizing the text. A TF-IDF vectorizer is then applied to convert the processed text into numerical feature vectors for classification. The TF-IDF value $w_{i,j}$ for word j in document i is calculated as:

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

where $tf_{i,j}$ is the frequency of term j in document i, N is the total number of documents, and df_j is the number of documents containing term j. This weighting scheme emphasizes rare but informative words that contribute strongly to classification.

An initial analysis revealed severe class imbalance: approximately 89.83% of comments were labeled as non-toxic across all categories. To mitigate this, we applied the Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates synthetic examples of the minority class by interpolating between an instance and one of its nearest neighbors in feature space. This helps to balance the dataset and improve the classifier's ability to detect underrepresented toxic categories.

Two versions of the model were trained for comparison: one using the original dataset, and one using the SMOTE-balanced dataset. In both cases, we applied a binary relevance approach, treating each of the six target labels (`toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, `identity_hate`) as independent binary classification problems. This resulted in a total of 12 Logistic Regression models (6 labels × 2 datasets). We deliberately chose binary relevance over classifier chains to maintain independence between classifiers and ensure consistent performance across labels.

Finally, we applied LIME (Local Interpretable Model-Agnostic Explanations) to visualize and interpret individual model predictions. LIME helped reveal which features (words) contributed most to classification decisions, providing valuable insights into the decision-making process and highlighting potential bias patterns.

4. Outcomes

We selected multiple metrics to evaluate the performance of our models on the test set: accuracy, recall, precision, false-negative rate (FNR), and false-positive rate (FPR). Two versions of the Logistic Regression models were trained: one on the original imbalanced dataset and one using SMOTE to address class imbalance. The results of both models are shown below.

Model Performance Without Upsampling (Original Data)

Label	Accuracy	Recall	Precision	FNR	FPR
toxic	0.9052	0.0186	0.7082	0.9814	0.0008
severe_toxic	0.9900	0.0238	0.5205	0.9762	0.0002
obscene	0.9472	0.0077	0.6075	0.9923	0.0003
threat	0.9970	0.0000	0.0000	1.0000	0.0000
insult	0.9507	0.0070	0.5851	0.9930	0.0003
identity_hate	0.9912	0.0014	0.2222	0.9986	0.0000

Model Performance With SMOTE Upsampling

Label	Accuracy	Recall	Precision	FNR	FPR
toxic	0.5548	0.7060	0.1396	0.2940	0.4613
severe_toxic	0.8922	0.4483	0.0420	0.5517	0.1033
obscene	0.5038	0.7554	0.0764	0.2446	0.5103
threat	0.4863	0.7950	0.0046	0.2050	0.5146
insult	0.5004	0.7479	0.0704	0.2521	0.5125
identity_hate	0.5071	0.7395	0.0131	0.2605	0.4950

From the comparison, we observe that recall increased dramatically and false-negative rates decreased substantially after applying SMOTE. However, this came at the cost of lower accuracy and precision, and much higher false-positive rates. The SMOTE model better detected rare toxic labels but also produced more false alarms. This highlights the trade-off between sensitivity and precision.

To assess fairness, we measured FNR and FPR for all labels. We found that `threat` comments had the highest false-negative rate in the original model, while `toxic` comments had the highest false-positive rate. These results were consistent with known challenges in detecting rare and ambiguous toxic content.

We further applied LIME to interpret individual predictions. LIME helped identify which words contributed most to each prediction. We found that both models correctly flagged explicit offensive terms, but the SMOTE model produced broader detection patterns and flagged more borderline cases. As a result, the model without upsampling is more conservative and minimizes false positives, while the model with SMOTE is more aggressive and reduces false negatives.

In conclusion, both models demonstrate interpretable decision-making. The choice between them depends on platform goals: if reducing harmful content exposure is prioritized, the SMOTE model is preferable; if minimizing over-flagging and promoting free expression is more important, the original model is better suited.

5. Summary

The data used to train our toxic comment classification system was largely appropriate but had notable limitations. The dataset contained a significant class imbalance, with the vast majority of comments labeled as non-toxic. As a result, we employed SMOTE upsampling to improve recall and reduce false negatives for underrepresented labels. While this helped the model detect more harmful comments, it also introduced a higher rate of false positives. <<33 Overall, the Logistic Regression models we implemented produced acceptable performance across multiple evaluation metrics, including accuracy, recall, precision, false-negative rate, and false-positive rate. We selected these metrics to capture both the overall predictive ability and the fairness of the system. Our use of LIME further enhanced model interpretability by allowing us to understand how individual predictions were made.

The stakeholders of this ADS include both platform users and platform administrators. Users who post content may prefer an ADS with a low false-positive rate to reduce the chance of their benign comments being flagged. In contrast, platform administrators and users from marginalized groups may prioritize minimizing false negatives to ensure that offensive content is reliably detected and removed. Our dual evaluation of models with and without SMOTE reflects these competing priorities.

In terms of accuracy, our model performs well but this metric is not fully representative due to the imbalance in the dataset. The system still shows weaknesses in detecting the most severe or rare forms of toxicity, including threatening and identity-hate comments, which raises fairness concerns. These limitations indicate that additional work is needed to provide adequate protection for vulnerable groups.

Given the current results, we would be cautious about deploying this ADS in industry or the public sector without further improvements. While it demonstrates reasonable performance and interpretability, the risk of misclassifying critical edge cases remains high.

We recommend several improvements: expanding the dataset with more examples of rare toxic behaviors, improving the balance of labels, and potentially exploring more advanced models beyond Logistic Regression. Additional research on more sophisticated fairness-aware algorithms would also help reduce disparities across subgroups. With these changes, this ADS could evolve into a more reliable and equitable tool for content moderation.