Introduction

Online comment sections often contain toxic and abusive language, especially in discussions around controversial or political topics. Understanding how such language appears and differs across discussion contexts is important for analyzing online behavior and moderation challenges.

This project uses the Civil Comments dataset, which contains millions of online comments annotated with toxicity and identity-attack scores. The analysis begins with basic data cleaning and exploratory analysis to measure the prevalence of toxic and bigoted language. Since explicit news outlet labels are not available, comments are grouped into clusters based on textual similarity to approximate different discussion spaces.

Toxicity levels are compared across these clusters using descriptive statistics and hypothesis testing. Finally, a supervised machine learning model is trained to predict whether a comment is toxic based only on its text, in order to evaluate how well toxicity can be learned from language patterns.

Data Source

The data used in this project comes from the Civil Comments dataset, available through Hugging Face under google/civil_comments. The dataset contains approximately 1.8 million English-language online comments, each annotated with continuous scores for toxicity and related attributes such as insults, threats, and identity attacks. The data was loaded directly using the datasets Python library, and only the training split was used since it already includes all relevant labels. No API access or web scraping was required.

Before analysis, several preprocessing steps were applied. First, comments with missing or empty text fields were removed to ensure that all remaining observations contained valid textual content. Next, the dataset was reduced to include only the text and toxicity-related columns required for the analysis. To simplify interpretation, two binary indicators were created from the continuous labels: a comment was classified as toxic if its toxicity score was at least 0.5, and as bigoted if its identity-attack score was at least 0.5. This threshold was chosen to separate clearly abusive comments from neutral or ambiguous ones.

Because of the large size of the dataset, a random sample of 100,000 comments was drawn for exploratory analysis, clustering, and machine learning experiments. All subsequent analyses in this report are based on this sampled dataset.

Exploratory Data Analysis

After preprocessing, exploratory data analysis was conducted to understand the overall distribution of toxic and identity-based abusive language in the dataset.

Using the sampled dataset of 100,000 comments, approximately **7.9%** of comments were classified as **toxic** based on the toxicity threshold, while about **0.8%** were classified as **bigoted** (identity-based attacks). This indicates that while general toxicity is relatively common in online comments, explicit attacks targeting identity groups occur less frequently.

To better visualize these proportions, a bar chart was created comparing the fraction of toxic and bigoted comments.

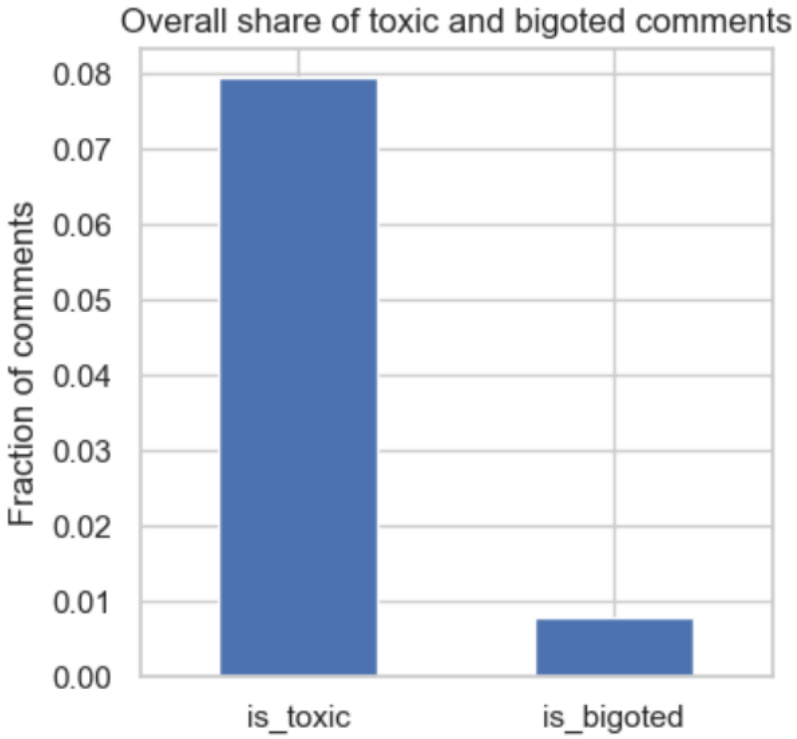

Figure 1: Proportion of toxic and bigoted comments in the sampled dataset.

Clustering Analysis

The dataset used in this project does not contain explicit identifiers for news outlets or discussion communities. To address this limitation, comments were grouped based on **textual similarity** in order to approximate **different discussion spaces**. The idea behind this approach is that comments discussing similar topics or themes are likely to belong to similar conversational contexts.

To represent the textual content numerically, **TF–IDF (Term Frequency–Inverse Document Frequency)** was computed from the comment text. Comments were grouped into **20 clusters** using the K-means clustering algorithm. The number of clusters was chosen as for balance. Fewer clusters would oversimplify discussion types, while many more clusters would make interpretation difficult.

After clustering, toxicity-related statistics were computed for each group, including the number of comments, the proportion of toxic comments, the proportion of bigoted comments, and the average toxicity score. The results show noticeable variation across clusters. Some clusters exhibit toxicity rates well above the dataset average, while others remain comparatively low.

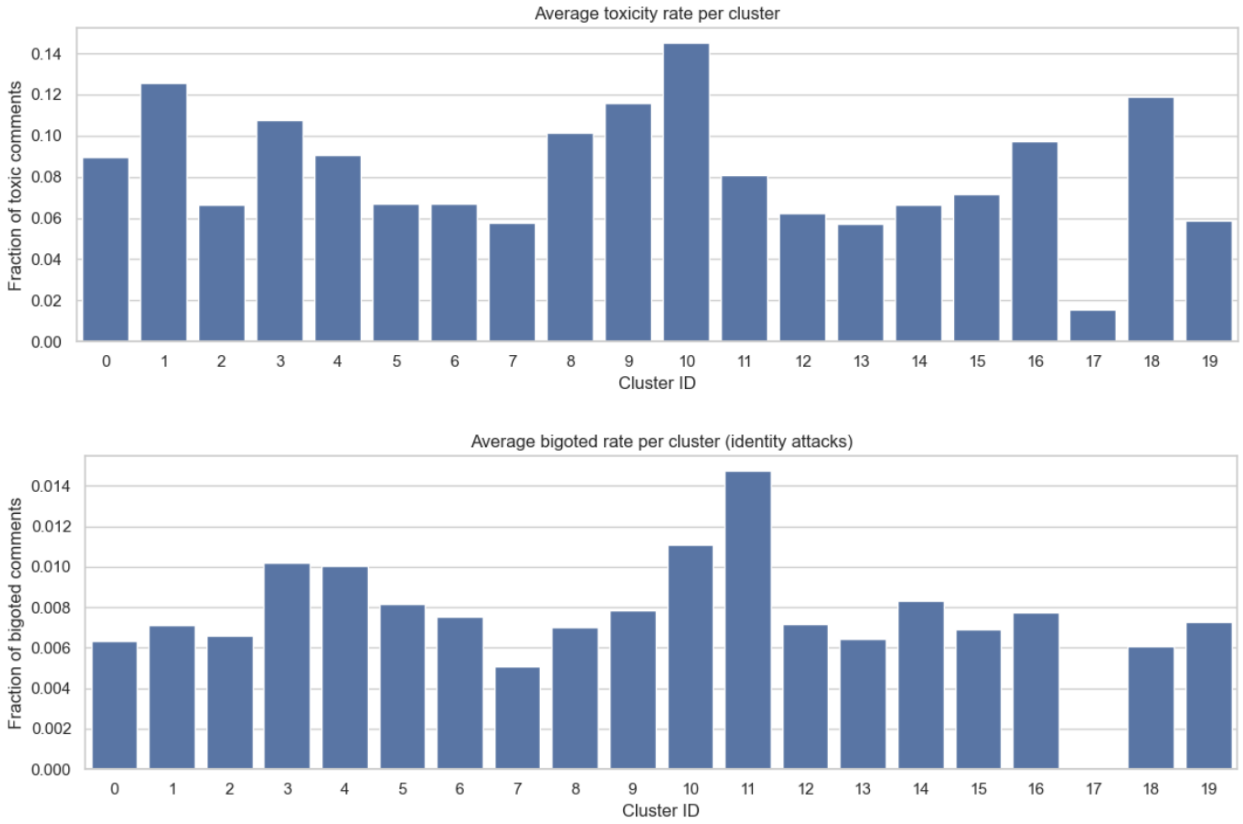

Figure 2: Average toxicity and bigoted rate per cluster visualized

Visualizations of toxicity and bigoted-language rates across clusters reveal that a small number of clusters account for disproportionately high levels of abusive language. These clusters are likely associated with more contentious or polarized discussions. In contrast, clusters with lower toxicity tend to reflect more neutral or conversational language.

Overall, the clustering analysis suggests that toxic language is not evenly distributed across discussions. Instead, it appears to be concentrated in specific types of comment groups, motivating further statistical testing and deeper inspection of cluster content.

Hypothesis Testing

The clustering analysis suggested that toxicity levels vary across different groups of comments. To determine whether these observed differences are statistically significant, hypothesis testing was conducted using toxicity scores.

The primary hypothesis test used was a **one-way Analysis of Variance (ANOVA)**.

The null hypothesis: all clusters have the same mean toxicity score

Alternative hypothesis: at least one cluster differs in mean toxicity.

Toxicity scores of comments were grouped by their assigned cluster, and ANOVA was applied to compare the group means.

The ANOVA test produced an **F-statistic of 60.15 with a p-value less than 0.001**, providing strong evidence against the null hypothesis. This result indicates that toxicity is not evenly distributed across clusters and that some discussion groups are significantly more toxic than others.

To further illustrate this difference, a **two-sample t-test** was performed comparing the cluster with the highest toxicity rate to the cluster with the lowest toxicity rate. The test yielded a highly significant result ($p \ll 0.001$), confirming that the difference between these two clusters is substantial and not due to random variation.

Together, these statistical tests support the conclusion that discussion context, approximated by text-based clustering, is significant in shaping the level of toxic language present in online comments.

Supervised Machine Learning

In addition to clustering and statistical analysis, a supervised machine learning approach was used to evaluate whether toxic language can be predicted directly from comment text. The task was framed as a **binary classification problem**, where the goal is to predict whether a comment is toxic based on its textual content alone.

The dataset was split into training (**80%**) and testing (**20%**) sets using a stratified split to preserve the proportion of toxic comments in both sets. Text was converted into numerical features using **TF–IDF**. A **logistic regression classifier** was chosen as a baseline model due to its simplicity, efficiency, and interpretability in text classification tasks.

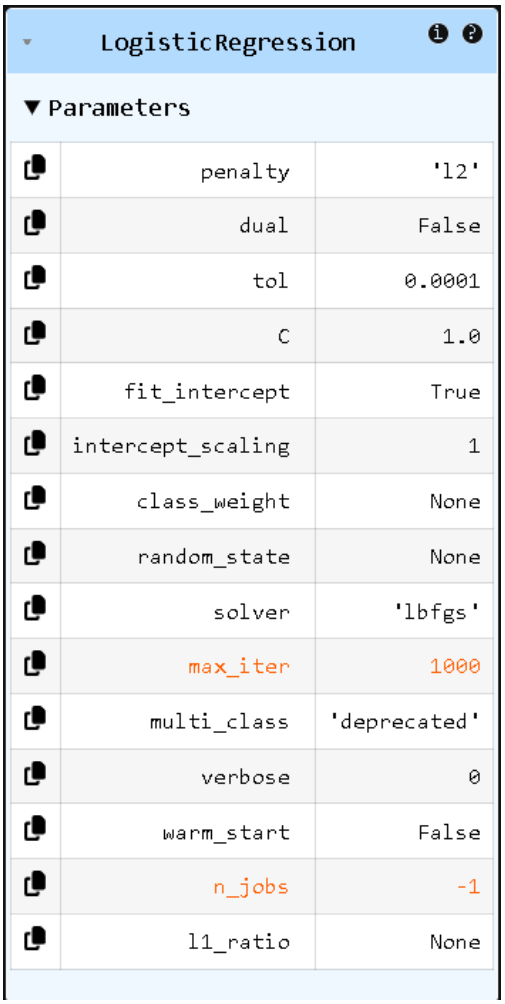

Figure 3: Logistic Regression Classifier

Model performance was evaluated using precision, recall, F1-score, and ROC-AUC. While the model achieved high overall accuracy, accuracy alone is not sufficient due to the imbalanced nature of the dataset. The classifier showed **high precision for toxic comments**, meaning that when it labeled a comment as toxic, it was usually correct. However, recall for the toxic class was lower, indicating that some toxic comments were not detected. Despite this, the model achieved a **ROC-AUC score above 0.9**, demonstrating strong overall separation between toxic and non-toxic comments.

```
              precision    recall  f1-score   support

       False      0.933     0.998     0.965     18412
        True      0.897     0.171     0.287      1588

    accuracy                          0.933     20000
   macro avg      0.915     0.584     0.626     20000
weighted avg      0.930     0.933     0.911     20000

ROC-AUC: 0.903

array([[18381,    31],
       [ 1317,   271]])
```

Figure 4: ROC-AUC results

To better understand the model's decisions, the learned feature weights were examined. Words such as *"stupid," "idiot," "ignorant,"* and *"pathetic"* strongly increased the probability of a comment being classified as toxic, while words such as *"thanks," "good,"* and *"excellent"* were associated with non-toxic comments. These results align with intuitive expectations and suggest that the model captures meaningful linguistic patterns related to toxicity.

```
Top features pushing toward TOXIC:
stupid 18.70197761564193
idiot 11.240103143771043
idiots 10.631166287556146
```

```
Top features pushing toward NON-TOXIC:
thanks -1.9799938445242462
question -1.8620222056343774
thank you -1.7303742442998475
```

Figure 5: Top 3 words associated with TOXIC and NON-TOXIC clusters

Overall, the supervised learning results show that toxic language can be effectively learned from word usage alone, supporting the findings from the earlier exploratory and clustering analyses.

Conclusions

7. Discussion

The results of this project consistently show that toxic language in online comments is not randomly distributed but is instead strongly influenced by the type of discussion taking place. The exploratory analysis revealed that while most comments are non-toxic, a noticeable minority

contain abusive language, highlighting the relevance of toxicity as an ongoing issue in online discourse.

The clustering analysis demonstrated substantial variation in toxicity levels across discussion groups. Clusters characterized by politically charged language and references to public figures exhibited higher rates of toxic and identity-based abusive comments, while clusters containing more neutral or conversational language showed lower toxicity. Hypothesis testing confirmed that these differences are statistically significant, reinforcing the idea that discussion context plays a crucial role in shaping online behavior.

The supervised machine learning results further support these findings. The logistic regression model successfully learned linguistic patterns associated with toxicity, achieving strong overall discrimination between toxic and non-toxic comments. The words most strongly associated with toxic predictions were explicit insults and derogatory terms, while non-toxic predictions were linked to polite or neutral language. This alignment between statistical patterns, clustering behavior, and model interpretation strengthens confidence in the conclusions.

Taken together, these results suggest that certain topics—particularly political discussions—are more likely to attract hostile language. The combination of exploratory analysis, clustering, hypothesis testing, and supervised learning provides a coherent picture of how toxicity manifests in large-scale online comment data.

Limitations and Future Work

This project has several limitations that should be considered when interpreting the results. First, the dataset does not include explicit identifiers for news outlets or discussion communities. As a result, clustering based on textual similarity was used as an approximation of discussion spaces. While this approach captures thematic differences, it does not perfectly reflect real-world communities or platform structures.

Additionally, the analysis focuses solely on English-language comments and does not consider temporal factors such as changes in toxicity over time. Extending the analysis to include multilingual data or time-based trends could provide further insight into how toxic language evolves.

Future work could also explore more advanced machine learning models, such as transformer-based language models, to improve classification performance. Finally, incorporating datasets with explicit outlet or community labels would allow for a more direct comparison between different sources and provide a clearer understanding of how toxicity varies across platforms.