Text Mining for Online Toxicity: Classification and Topic Modeling

Julius Maliwat 864520
Muluken Bogale Megersa 919654

Dataset Overview

223,549 user-generated comments annotated for research on toxic behavior detection.

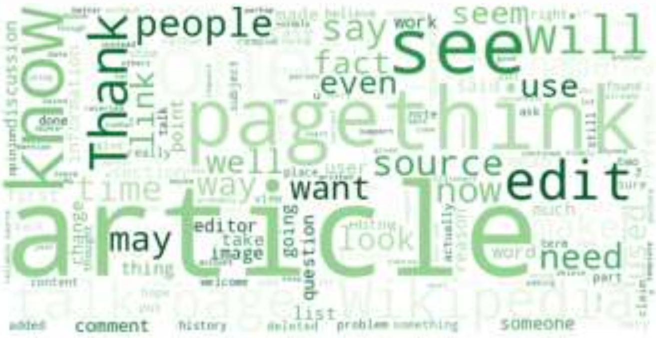
Six binary categories: Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate.

Multi-label nature: A single comment can belong to one or more toxicity categories.

89.95% of comments have no toxic labels.

Label	Freq. (%)
Toxic	9.57
Obscene	5.43
Insult	5.06
Identity Hate	0.95
Severe Toxic	0.88
Threat	0.31

Comparison of Non-Toxic and Censored Toxic Comments WordClouds



Research objectives

Text Classification

Classify user-generated comments into multiple toxicity categories to detect complex and overlapping toxic behaviors.

Topic Modeling

Identify thematic structures in toxic content to better understand patterns of online toxicity.

Preprocessing Pipeline

Minimal Preprocessing

Preserves the contextual structure for transformer models:

1. Remove URLs and mentions.
2. Convert text to lowercase.
3. Normalize whitespaces.

Comprehensive Preprocessing

Extends minimal preprocessing for traditional methods like TF-IDF:

1. Remove punctuation and numbers.
2. Eliminate stopwords.

Classification Methods

TF-IDF + Logistic Regression

Chosen as a simple and interpretable baseline to evaluate multi-label classification.

Text Representation:
TF-IDF generates a sparse matrix of word frequencies, limited to the top 10,000 terms for computational efficiency.

Model:
Logistic Regression applies a one-vs-rest strategy, leveraging class weights to handle class imbalance.

DistilBERT Fine-Tuning

Selected for its ability to capture contextual and nuanced toxic behaviors.

Text Representation:
DistilBERT produces contextual embeddings for tokens, encoding semantic and syntactic relationships.

Model:
The model is fine-tuned for multi-label classification, handling overlapping toxicity categories with Binary Cross-Entropy loss

Classification Results

The models were evaluated using **Mean ROC AUC** for overall performance and **F1-Score** for label-specific performance across six categories.

Mean ROC AUC

Method	Mean ROC AUC
DistilBERT	**0.985**
Logistic Regression	0.970

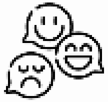
F1 Score per Label

Label	Logistic Regression	DistilBERT
Toxic	0.57	**0.68**
Obscene	0.57	**0.69**
Insult	0.50	**0.71**
Identity Hate	0.28	**0.62**
Severe Toxic	0.19	**0.39**
Threat	0.25	**0.52**

Topic Modeling Approach

Text Representation:
Text was represented using term frequencies, excluding rare and overly frequent terms, focusing on unigrams for simplicity and compatibility with LDA

Model:
Latent Dirichlet Allocation (LDA) was applied to uncover latent topics as probabilistic distributions over words.



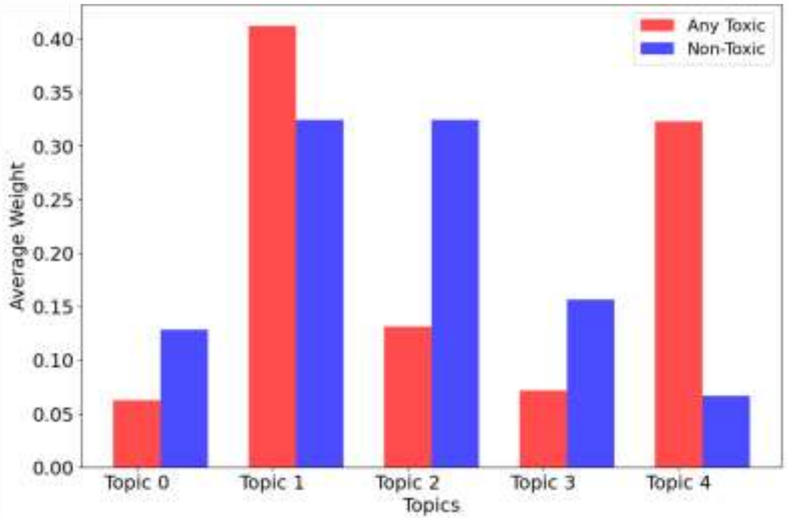
Global Dataset:
Includes all comments to identify broader themes and observe overlaps between toxic and non-toxic content

Toxic-Only Subset:
Latent Dirichlet Allocation (LDA) was applied to uncover latent topics as probabilistic distributions over words.

Results - Global Dataset

We evaluated LDA with 5 and 10 topics on the global dataset to observe trade-offs between coherence, perplexity, and topic diversity

Metric	5 Topics	10 Topics
Coherence	**0.754**	0.680
Perplexity	-7.727	**-7.779**
Diversity	**0.860**	0.830

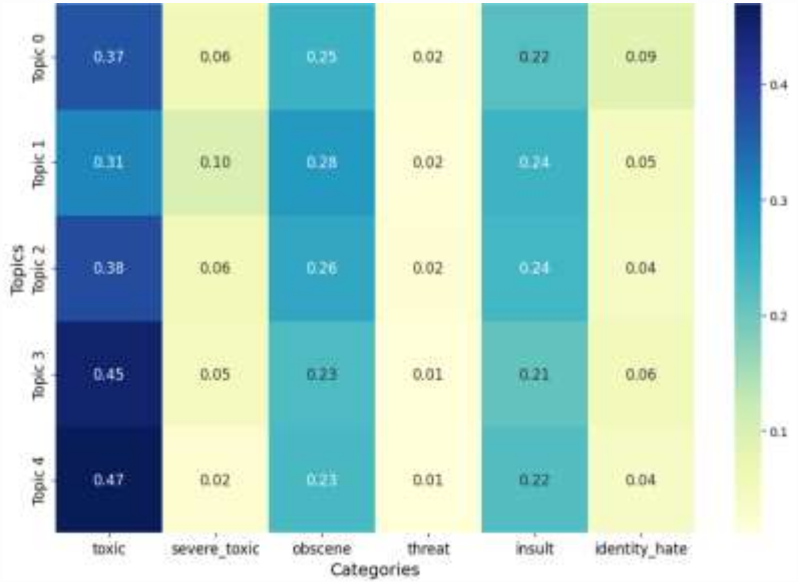


Average Topic Weights for Toxic and Non-Toxic Comments in the Global Dataset

Results - Toxic-Only Dataset

We evaluated LDA with 5 and 10 topics on the toxic-only dataset to analyze the challenges of modeling toxic themes and trade-offs between coherence, perplexity, and diversity.

Metric	5 Topics	10 Topics
Coherence	**0.490**	0.478
Perplexity	-7.049	**-7.061**
Diversity	0.920	**0.930**



Proportional distribution of toxicity categories across topics for the 5-topic model

Conclusions and Future Work

Key Takeaways

Classification
DistilBERT outperforms Logistic Regression, particularly on minority labels, while Logistic Regression remains a viable option for resource-constrained scenarios.

Topic Modeling
LDA struggled with overlapping themes in the global dataset and failed to differentiate toxicity-specific topics in the toxic-only subset, highlighting challenges in isolating distinct toxic behaviors

Future Work

Address Class Imbalance
Combine data augmentation and advanced transformer architectures like RoBERTa or T5 to enhance classification performance, particularly on imbalanced datasets and rare labels.

Improve Topic Modeling
Adopt contextual methods like BERTopic to improve thematic coherence and better handle overlapping toxic behaviors.

Thanks!

Do you have any questions?