Text Mining for Online Toxicity: Insights from Classification and Topic Modeling

Julius Maliwat 864520[a], Muluken Bogale Megersa 919654[a]

[a] *University of Milano-Bicocca, Data Science*

Abstract

This study explores two key challenges in text mining using the Jigsaw Toxic Comment Classification dataset: multi-label classification and topic modeling. For classification, Logistic Regression with TF-IDF representation was compared against fine-tuned DistilBERT. While DistilBERT achieved superior performance in Mean ROC AUC and F1 scores across six toxic categories, including minority labels such as 'threat' and 'identity hate', this came at the cost of greater computational requirements.

For topic modeling, Latent Dirichlet Allocation (LDA) was applied to the global dataset and a toxic-only subset. While the global analysis revealed overlapping themes, the toxic-only analysis highlighted challenges in distinguishing topics. Metrics such as coherence and perplexity provided quantitative insights into the trade-offs between interpretability and granularity.

Despite promising results, challenges such as severe class imbalance and limited topic interpretability were observed. Future work could explore advanced transformer architectures and hybrid topic modeling approaches to further improve performance and insights into online toxicity.

Contents

1. Introduction

The proliferation of toxic comments in online platforms has necessitated the development of advanced techniques for detecting and analyzing harmful content. This study addresses two key challenges in text mining: classifying toxic comments into predefined categories and identifying latent thematic structures within the dataset. By leveraging a combination of traditional and transformer-based methods, the study aims to provide a comprehensive exploration of predictive modeling and unsupervised learning techniques applied to online toxicity.

1.1. Objectives and Methodology Overview

This study focuses on two main tasks. The first task is multi-label text classification, where the goal is to categorize comments into six toxic categories: toxic, severe toxic, obscene, threat, insult, and identity hate. To achieve this, traditional machine learning methods (TF-IDF with Logistic Regression) and transformer-based models (DistilBERT fine-tuning) were employed. Performance was evaluated using mean ROC AUC and F1 score metrics.

The second task is topic modeling, aimed at uncovering thematic patterns in the dataset. Latent Dirichlet Allocation (LDA) was used to identify latent themes, analyzing both the entire dataset and a filtered subset of toxic comments. Model evaluation was based on coherence, perplexity, and diversity metrics.

This methodology integrates predictive and unsupervised approaches to address the complexities of classifying and analyzing toxic content.

1.2. Dataset Overview

The dataset used in this study is the Jigsaw Toxic Comment Classification dataset, designed to support research in text mining and toxicity detection (1). It comprises a training set of 159,571 user-generated comments and a test set of 63,978 comments, resulting in a total of 223,549 samples. Each comment is annotated with six binary labels indicating the presence or absence of specific types of toxic behavior: toxic, severe toxic, obscene, threat, insult, and identity hate.

A significant majority of the comments (89.95%) are non-toxic, containing no active toxicity labels. The remaining 10.05% exhibit at least one form of toxicity, highlighting the dataset's substantial class imbalance. Among the toxic comments, the 'toxic' label is the most frequent, present in 9.57% of the entire dataset, while 'threat' is the least frequent, appearing in only 0.31% of samples. Table 1 provides a detailed breakdown of label frequencies.

Table 1: Distribution of toxicity labels in the dataset.

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

The imbalance between non-toxic and toxic comments, combined with the rarity of certain labels like 'threat', presents significant challenges for multi-label classification. Models must effectively handle the imbalance to avoid bias towards the majority class while accurately identifying minority labels. Additionally, the overlap between labels, such as 'toxic' co-occurring with 'obscene' and 'insult', complicates the task further by requiring models to capture nuanced patterns in the data.

These characteristics make the dataset a challenging benchmark for evaluating the effectiveness of text mining techniques, particularly in multi-label classification scenarios with sparse and imbalanced data.

2. Preprocessing

To prepare the dataset for text classification and topic modeling, we implemented two distinct preprocessing pipelines tailored to the requirements of the selected text representation methods. A minimal preprocessing approach was applied for contextual embeddings, while a comprehensive preprocessing pipeline was designed for traditional feature extraction methods.

2.1. Minimal Approach

The minimal preprocessing pipeline was designed to retain the contextual integrity of the text, ensuring compatibility with pre-trained contextual embeddings such as DistilBERT. The following steps were applied to each text sample:

- URL Removal: All hyperlinks and web addresses were removed using regular expressions to eliminate non-informative content.
- Mention Removal: Mentions (e.g., `@username`) were stripped to avoid user-specific tokens that influence the model.
- Lowercasing: All text was converted to lowercase to standardize input and align with the requirements of uncased language models.

This minimal preprocessing approach preserves key syntactic and semantic features while removing noise, ensuring optimal input quality for contextualized models.

2.2. Comprehensive Approach

The comprehensive preprocessing pipeline was implemented to prepare the text for traditional feature extraction methods such as TF-IDF. This approach aimed to reduce noise and standardize the input data, enabling efficient feature extraction. Steps already described in the Minimal Approach, such as URL removal, mention removal, and lowercasing, were also applied. Additionally, the following transformations were performed:

- Punctuation removal: All punctuation marks were removed to reduce the sparsity in the feature space.
- Hashtag removal: Hashtags were cleaned by removing the # symbol while retaining the associated text.
- Number removal: All numerical values were stripped to focus on textual content.
- Whitespace normalization: Extra whitespaces, newlines, and tabs were consolidated into single spaces.
- Stop-word removal: Non-informative words were excluded using the NLTK English stop-word list.

This pipeline ensured that only relevant textual information was retained, enhancing the compatibility of the dataset with vectorization techniques like TF-IDF.

3. Text Classification

The objective of the text classification task was to assign toxic comments to one or more of six predefined categories: toxic, severe toxic, obscene, threat, insult, and identity hate. To address this multi-label classification problem, two distinct pipelines were developed. The first utilized TF-IDF as a text representation technique combined with Logistic Regression, providing a simple and interpretable baseline. The second leveraged DistilBERT, a transformer-based model, to generate contextualized embeddings, followed by fine-tuning for multi-label classification. The models were evaluated using mean ROC AUC and F1 score metrics, offering a quantitative comparison of their performance across all labels. This section outlines the methodology, including text representation, modeling approaches, evaluation metrics, and results, highlighting the trade-offs between traditional and transformer-based pipelines.

3.1. Text Representation

Text representation plays a pivotal role in determining the performance of machine learning models for text classification. In this study, two distinct methods were employed to convert textual data into numerical representations. The first method utilized TF-IDF, a traditional vectorization technique designed to capture term importance in the corpus, offering a sparse and interpretable feature set. The second method leveraged DistilBERT, a transformer-based model, to generate contextualized embeddings that encode semantic dependencies in the text. Each approach was tailored to the specific requirements of the corresponding classification pipeline and is detailed in the following subsections.

3.1.1. TF-IDF Representation

TF-IDF (Term Frequency-Inverse Document Frequency) was employed as a traditional vectorization technique to convert text data into numerical features (2). This method captures the importance of terms by balancing their frequency within a document and across the entire corpus. The implementation included the following key steps:

- Vectorization: The TF-IDF vectorizer was configured with a maximum of 10,000 features, limiting the vocabulary to the most informative terms. This reduced sparsity and improved computational efficiency.

- Training and Test Transformation: The vectorizer was fitted on the training dataset to learn term weights and then applied to transform both training and test sets into sparse matrices.

The resulting TF-IDF representation provides a sparse feature matrix that is well-suited for linear models like Logistic Regression, balancing interpretability and computational feasibility.

3.1.2. DistilBERT Representation

DistilBERT, a lightweight transformer-based model, was employed to generate contextualized embeddings for text representation (3). Each input sequence was tokenized and either truncated or padded to a fixed length of 128 tokens, ensuring a balance between computational efficiency and the ability to capture sufficient context. The DistilBERT base model, initialized with pretrained weights, provided robust semantic and syntactic representations tailored to the specific task. These embeddings, which incorporate contextual information across the input sequence, served as the input features for downstream classification.

3.2. Models

Two distinct models were employed to address the multi-label classification task. Logistic Regression was used as a baseline model due to its simplicity and interpretability, paired with the TF-IDF representation for input features. In contrast, DistilBERT, a transformer-based model, was fine-tuned to generate contextual embeddings and directly classify comments into multiple categories. This section details the configuration and rationale behind the use of these models for the toxic comment classification task.

3.2.1. Logistic Regression

Logistic Regression was employed as the baseline model for multi-label classification (4). The model was configured using the One-vs-Rest strategy, which trains a separate binary classifier for each of the six target categories. The following key parameters were used:

- Regularization: L2 regularization was applied by default to manage overfitting and ensure the model generalized effectively to unseen data.

- Class Weighting: The `balanced` option was enabled to address class imbalance by assigning weights inversely proportional to class frequencies in the training data.

- Iterations: The maximum number of iterations was set to 1,000, ensuring convergence for the dataset's size and complexity.

This configuration allowed the model to effectively process the high-dimensional, sparse feature matrices generated by TF-IDF and provided a robust and interpretable baseline for evaluating transformer-based models.

3.2.2. DistilBERT Fine-Tuning

The DistilBERT model, initialized with `distilbert-base-uncased` pretrained weights, was fine-tuned for the multi-label classification of toxic comments. A classification head, consisting of a fully connected layer with six output neurons and a sigmoid activation function, was added to map the contextual embeddings to the target categories. The model was fine-tuned using Binary Cross-Entropy with Logits as the loss function, which is suitable for multi-label tasks.

To ensure efficient training and convergence, the AdamW optimizer was employed with a learning rate of 2×10^{-5}. The training process was performed over 3 epochs, balancing performance and computational cost. This setup allowed DistilBERT to adapt its parameters to the specific requirements of the dataset while leveraging its pre-trained knowledge to capture semantic nuances effectively.

3.3. Metrics

The performance of the classification models was evaluated using two complementary metrics, each addressing different aspects of the task. The Mean Column-Wise ROC AUC provides an aggregated measure of the model's ability to separate positive and negative classes across all labels, serving as the primary metric. The F1 Score, calculated for each label, offers a detailed analysis of the balance between precision and recall, particularly relevant given the class imbalance in the dataset. Each metric will be discussed in the following subsections, including its theoretical foundation and relevance to the multi-label classification task.

3.3.1. Mean Column-Wise ROC AUC

The Mean Column-Wise ROC AUC was employed as the primary evaluation metric to measure the overall performance of the models in the multi-label classification task (5). This metric calculates the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) for each individual label, reflecting the model's ability to distinguish between positive and negative classes across various threshold settings.

In a multi-label setting, the column-wise ROC AUC is computed for each label individually, and the mean is taken to obtain an aggregated score. This approach ensures that each label is equally represented in the final metric, making it robust to class imbalance and allowing for a balanced evaluation across all categories.

Mathematically, the column-wise ROC AUC for a given label l is defined as:

$$\text{AUC}_l = \int_0^1 \text{TPR}_l(\text{FPR}_l)\, d\text{FPR}_l$$

where TPR_l and FPR_l are the True Positive Rate and False Positive Rate for label l, respectively. The mean column-wise ROC AUC is then given by:

$$\text{Mean AUC} = \frac{1}{L} \sum_{l=1}^{L} \text{AUC}_l$$

where L is the total number of labels.

This metric provides a global measure of the model's ability to distinguish between classes while accounting for the multi-label nature of the task. Its robustness to class imbalance makes it particularly suitable for multi-label classification problems with imbalanced and sparse label distributions.

3.3.2. F1 Score

The F1 Score was employed as a complementary evaluation metric to assess the balance between precision and recall for each label (6). This metric is particularly suitable for multi-label classification tasks with imbalanced datasets, as it captures the trade-offs between identifying true positives and avoiding false positives. It is defined as the harmonic mean of precision (P) and recall (R):

$$F1_l = 2 \cdot \frac{P_l \cdot R_l}{P_l + R_l}$$

where:

$$P_l = \frac{\text{TP}_l}{\text{TP}_l + \text{FP}_l}, \quad R_l = \frac{\text{TP}_l}{\text{TP}_l + \text{FN}_l}$$

In this study, F1 Scores were calculated for the six labels ('toxic', 'severe toxic', 'obscene', 'threat', 'insult', and 'identity hate'). This label-specific evaluation highlights the models' ability to handle frequent as well as rare categories, such as 'threat' and 'severe toxic', where class imbalance poses significant challenges. By examining F1 Scores across labels, we gain insights into the strengths and weaknesses of each model in managing both precision and recall.

3.4. Results

This section presents the performance of the classification models, comparing their effectiveness across the chosen evaluation metrics. The analysis highlights key differences in model behavior, focusing on their ability to handle the multi-label nature of the task and class imbalance. Additionally, insights derived from the results are discussed, providing a deeper understanding of the strengths and limitations of each approach in the context of toxic comment classification.

3.4.1. Comparison of model performances

The performance of the models was evaluated based on Mean Column-Wise ROC AUC and F1 Score across all six labels. Table 2 summarizes the Mean ROC AUC scores, highlighting that DistilBERT outperformed Logistic Regression with a score of 0.985 compared to 0.970. This result underscores the advantages of transformer-based models in capturing complex relationships within textual data.

Table 2: Comparison of Mean ROC AUC between models.

Model	Mean ROC AUC
Logistic Regression	0.970
DistilBERT	0.985

For F1 Score, Table 3 provides a per-label comparison. DistilBERT demonstrated higher F1 Scores across all categories, particularly excelling in labels such as 'insult' (0.71 vs. 0.50) and 'identity hate' (0.62 vs. 0.28). While Logistic Regression achieved competitive recall, its lower precision resulted in reduced F1 Scores, especially for 'severe toxic' and 'threat'. These results highlight the limitations of traditional linear models in multi-label tasks with significant class imbalance.

Table 3: Comparison of F1 Score per label between models.

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

In terms of efficiency, Logistic Regression was significantly faster in both training and inference compared to DistilBERT. While DistilBERT required more computational resources and longer training times due to its transformer-based architecture, the superior performance justifies its use in applications where accuracy is critical. Conversely, Logistic Regression's efficiency makes it a practical choice for scenarios with limited computational resources or strict time constraints.

Overall, the results illustrate the superiority of DistilBERT in both overall performance and label-specific classification, making it a robust choice for toxic comment classification, albeit at a higher computational cost.

3.4.2. Analysis and Insights

The superior performance of DistilBERT can be attributed to its ability to capture semantic and syntactic relationships within the text through contextualized embeddings. This advantage is particularly evident in labels such as 'identity hate' and 'threat', where the patterns of toxicity are more nuanced and less frequent. Logistic Regression, relying on linear combinations of TF-IDF features, struggles to model these complexities, leading to its comparatively lower performance.

The class imbalance within the dataset further amplified these differences. While both models exhibited trade-offs between precision and recall, DistilBERT maintained a better balance, especially for underrepresented labels. Logistic Regression achieved high recall for minority classes such as 'severe toxic' but at the cost of lower precision, resulting in inflated predictions and reduced overall F1 Scores.

From the results, both models faced challenges in accurately classifying the 'severe toxic' label, which is characterized by its low representation in the dataset. Despite DistilBERT's superior overall performance, its F1 Score for 'severe toxic' remained relatively low at 0.39, indicating difficulties in capturing the nuanced patterns of this label. Logistic Regression, while achieving high recall for 'severe toxic', suffered from extremely low precision, resulting in an F1 Score of just 0.19.

4. Topic Modeling

The objective of the topic modeling task was to uncover latent thematic structures in the dataset, offering insights into the underlying topics present in both toxic and non-toxic comments (7). Two distinct analyses were conducted: the first applied to the entire dataset and the second restricted to comments labeled as toxic. Text representation was performed using the Bag of Words (BoW) approach, which provides a straightforward and interpretable representation of text by capturing term frequencies without applying weighting schemes. This choice ensures compatibility with the probabilistic nature of the Latent Dirichlet Allocation (LDA) algorithm, which was employed as the modeling approach. LDA leverages its probabilistic framework to identify coherent topics across the dataset.

The models were evaluated using coherence, perplexity, and diversity metrics to assess the quality of the generated topics. This section outlines the methodology, detailing the text representation process, modeling approach, evaluation metrics, and results for both datasets. The comparative analysis highlights thematic differences between toxic and non-toxic content and evaluates the effectiveness of LDA in extracting meaningful patterns.

4.1. Text Representation

Text representation for the topic modeling task was performed using the Bag of Words (BoW) approach. This method was chosen for its simplicity and compatibility with probabilistic models like Latent Dirichlet Allocation (LDA). BoW captures the frequency of terms in each document without incorporating term weighting schemes such as TF-IDF, which can distort the probabilistic nature of LDA by altering term distributions.

Two distinct BoW representations were generated for different analytical objectives, both derived from the combined dataset, which included all comments from training and testing splits.

The first representation utilized the entire dataset without filtering by label, ensuring a comprehensive analysis of all comments. To improve the quality of the representation and ensure that only relevant terms were included, terms that appeared either too infrequently or too frequently across documents were excluded. Specifically, terms that appeared in fewer than 5 documents (low document frequency) or in more than 50% of the documents (high document frequency) were removed. This filtering process helps to eliminate both overly specific terms that do not generalize well and overly common terms that do not provide distinguishing power. The vocabulary was limited to the 10,000 most informative terms, and unigrams were selected as the n-gram range to balance computational efficiency with interpretability.

The second representation focused on comments with at least one positive label across the six predefined categories ('toxic', 'severe toxic', 'obscene', 'threat', 'insult', 'identity hate'). Similar to the first representation, filtering was applied to remove terms with low or high document frequency, and the same vocabulary size and n-gram range were used to maintain consistency.

4.2. Methodology

This section outlines the methodological approach used for topic modeling, focusing on the application of Latent Dirichlet Allocation (LDA) to extract latent thematic structures from the dataset. First, the theoretical foundation of LDA is presented, describing its probabilistic framework for identifying topics within text data. Subsequently, the experimental configuration is detailed, including the parameter settings and the distinct setups used for the global dataset and the toxic-only subset.

4.2.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic generative model widely used for topic modeling, enabling the identification of latent thematic structures within a collection of documents (8). It represents each document as a distribution over a fixed number of topics, and each topic as a distribution over terms. This generative framework captures the underlying relationships between words and documents, facilitating the extraction of coherent and interpretable topics from textual data.

In this study, LDA was selected as the modeling approach for its ability to uncover latent themes in text data, making it particularly suited for analyzing both general and toxic-specific content. The flexibility of LDA allows for varying the number of topics to tailor the granularity of the extracted themes, adapting to the distinct characteristics of each dataset.

4.2.2. Experimental Configuration

The experimental configuration for the topic modeling task involved the application of Latent Dirichlet Allocation (LDA) to two distinct datasets: the global dataset (comprising all comments) and the toxic-only subset (restricted to comments with at least one positive label across the six predefined categories). Each dataset was analyzed using different configurations to evaluate the impact of the number of topics on the quality and interpretability of the extracted themes.

For the global dataset, LDA models were trained with 5 and 10 topics to explore both coarse-grained and fine-grained thematic structures. Each model was trained using 5 passes over the corpus and 10 iterations per pass, balancing computational cost with convergence quality.

For the toxic-only subset, LDA models were trained with 10 and 15 topics to capture the more nuanced thematic structures inherent in toxic content. Due to the smaller size and higher complexity of this subset, 20 iterations per pass were used to ensure the stability of the results.

All models employed symmetric priors for the document-topic (α) and topic-word (β) distributions, with hyperparameter optimization performed automatically during training.

4.3. Metrics

The evaluation of the topic modeling results was based on three metrics: coherence, perplexity, and diversity. Each metric captures a distinct aspect of model performance, from semantic consistency to statistical fit and topic uniqueness. The following subsections provide a detailed explanation of each metric and its relevance to this study.

4.3.1. Coherence

Coherence is a metric designed to evaluate the semantic consistency of terms within a topic (9). It measures how frequently terms in a topic co-occur in the same documents across the corpus, providing an interpretable way to assess the quality of the extracted topics. A high coherence score indicates that the terms in a topic are closely related and likely to form a meaningful theme, whereas a low coherence score suggests that the topic is less interpretable.

In this study, topic coherence was computed using the C_v metric, which combines measures of term similarity based on a sliding window, word co-occurrence statistics, and a probabilistic distribution of terms. This metric was chosen due to its balance between interpretability and robustness across different datasets and model configurations.

4.3.2. Perplexity

Perplexity is a statistical metric used to evaluate the goodness-of-fit of a topic model. It measures how well the model predicts the observed data, with lower perplexity scores indicating a better fit. Mathematically, perplexity is the exponential of the average negative log-likelihood of the test set given the trained model.

In this study, perplexity was employed to compare LDA models trained on both the global and toxic-only datasets. It complemented the coherence metric by providing an additional perspective on model performance, particularly in assessing how well each configuration captured the underlying data distribution.

4.3.3. Diversity

Diversity measures the distinctiveness of terms across topics, ensuring that the extracted themes capture unique and non-overlapping information. A high diversity score indicates that the terms defining different topics have minimal overlap, enhancing the interpretability and utility of the model. Conversely, low diversity may signal redundancy, where multiple topics capture similar content.

In this study, diversity was quantified as the proportion of unique terms across all topics relative to the total number of terms. This metric complements coherence and perplexity by focusing on the distinctiveness of topics rather than their internal consistency or statistical fit. It was particularly useful for evaluating configurations with a higher number of topics, where the risk of redundancy increases.

4.4. Results

This section presents the results of the LDA models applied to the global and toxic-only datasets. The analysis includes evaluations of coherence, perplexity, and diversity metrics for different numbers of topics, as well as a qualitative exploration of the extracted topics. For the global dataset, results are reported for models with 5 and 10 topics, while for the toxic-only dataset, models are analyzed using the same configurations of 5 and 10 topics. A comparative analysis highlights differences in thematic structures and model performance between the two datasets.

4.4.1. Global Dataset Results

The analysis of the global dataset using LDA focused on configurations with 5 and 10 topics, evaluated through coherence, perplexity, and diversity metrics. Table 4 summarizes the metrics for each configuration. The 5-topic model achieved a coherence score of 0.754, a perplexity of -7.727, and a diversity of 0.860, while the 10-topic model reported slightly lower coherence (0.680), better perplexity (-7.779), but a reduced diversity (0.830).

Table 4: Evaluation metrics for the global dataset.

Number of Topics	Coherence (C_v)	Perplexity	Diversity
5	0.754	-7.727	0.860
10	0.680	-7.779	0.830

In terms of topic interpretation, clear distinctions emerged between toxic and non-toxic content. Figure 1 illustrates the average topic weights for toxic and non-toxic comments. Topic 1 exhibited a significantly higher weight for toxic comments, suggesting its association with explicit language and offensive terms. In contrast, Topic 2 was predominantly associated with non-toxic content, reflecting its focus on neutral discussions. Topics 0, 3, and 4 demonstrated mixed distributions, highlighting overlaps between toxic and non-toxic themes.

These results suggest that the 5-topic model provides a more interpretable thematic structure due to its higher coherence and better diversity. The 10-topic model, on the other hand, captures finer-grained distinctions with improved perplexity but sacrifices diversity. The overlap in mixed topics, such as 0, 3, and 4, reflects the nuanced nature of the dataset, where certain themes are shared across both toxic and non-toxic content.

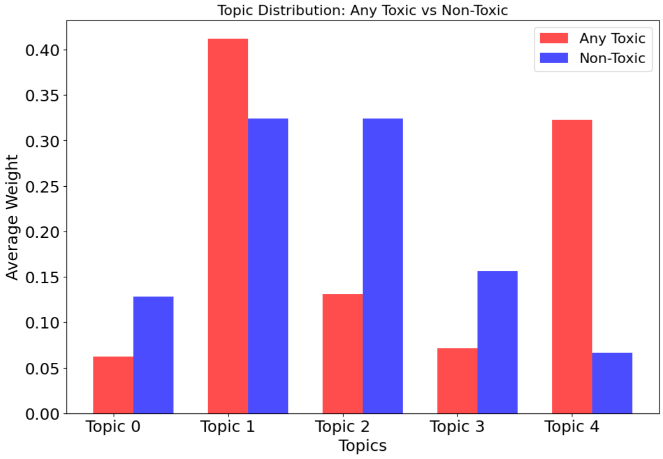

Figure 1: Average Topic Weights for Toxic and Non-Toxic Comments in the Global Dataset

4.4.2. *Toxic-Only Dataset Results*

The analysis of the toxic-only dataset was performed using LDA models with 5 and 10 topics. Evaluation metrics, including coherence, perplexity, and diversity, are presented in Table 5. The 5-topic model achieved a coherence score of 0.490, a perplexity of -7.049, and a diversity of 0.920, while the 10-topic model exhibited slightly lower coherence (0.478), but better perplexity (-7.061) and diversity (0.930). These results highlight the trade-off between interpretability and granularity, with the 5-topic model providing simpler and more interpretable themes, while the 10-topic model captures finer distinctions at the cost of clarity.

Table 5: Evaluation metrics for the toxic-only dataset.

Number of Topics	Coherence (C_v)	Perplexity	Diversity
5	0.490	-7.049	0.920
10	0.478	-7.061	0.930

Further analysis was conducted to examine the relationship between topics and toxicity categories (e.g., ‘toxic‘, ‘severe toxic‘, ‘obscene‘, etc.). Figure 2 shows the distribution of categories across topics for the 5-topic model. The heatmap reveals that all topics exhibit a similar distribution of toxicity categories, suggesting that LDA struggles to segregate specific toxic behaviors into distinct topics. This indicates that toxic language often overlaps across categories, making it challenging to isolate unique thematic structures.

The results indicate that both the 5-topic and 10-topic models struggled to provide coherent and well-separated topics. While the 5-topic model offers a marginally higher coherence score, it fails to distinctly categorize toxic behaviors. Similarly, the 10-topic model provides slightly better diversity but does not overcome the fundamental challenge of distinguishing between overlapping toxic categories. These findings highlight the limitations of LDA in modeling complex and interrelated toxic behaviors, suggesting that alternative modeling approaches may be necessary for better differentiation.

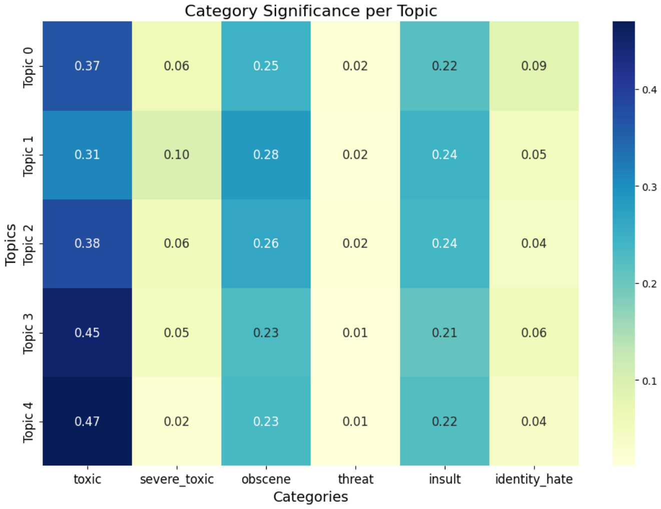

Figure 2: Proportional distribution of toxicity categories across topics for the 5-topic model (toxic-only dataset).

5. Conclusions

This study explored two fundamental tasks in text mining using the Jigsaw Toxic Comment Classification dataset: text classification and topic modeling. By employing a combination of traditional machine learning methods and state-of-the-art transformer-based models, as well as leveraging unsupervised learning for thematic analysis, the study aimed to address the challenges of detecting and analyzing toxic content in online platforms.

For the text classification task, DistilBERT demonstrated superior performance over Logistic Regression, achieving higher Mean ROC AUC and F1 scores across all toxic categories. Its ability to capture nuanced semantic patterns made it particularly effective for minority classes such as ‘threat‘ and ‘identity hate‘. However, this performance came at the cost of greater computational requirements compared to the lightweight and efficient Logistic Regression model, which still provided a competitive baseline.

In the topic modeling task, Latent Dirichlet Allocation (LDA) was applied to both the global dataset and a subset of toxic comments. While the global analysis revealed overlapping themes across toxic and non-toxic content, the toxic-only analysis highlighted challenges in thematic differentiation, with no clear separation between categories. Metrics such as coherence and perplexity provided quantitative insights into the trade-offs between interpretability and granularity of topics.

Despite the promising results, this study faced certain limitations. The severe class imbalance within the dataset posed challenges for both classification and topic modeling. Additionally, the interpretability of topics, particularly in the toxic-only analysis, was limited, suggesting that LDA might not fully capture the complexity of toxic language.

Future work could explore advanced transformer architectures such as RoBERTa or T5, which may further improve classification accuracy. Techniques such as data augmentation, semi-supervised learning, or specialized preprocessing pipelines could also help mitigate the effects of class imbalance. For topic modeling, hybrid approaches that combine LDA with neural embeddings or methods like BERTopic could offer better thematic differentiation and interpretability.

References

[1] cjadams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, nithum, and W. Cukierski, "Toxic comment classification challenge." https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge, 2017. Kaggle.

[2] S. Qaiser and R. Ali, "Text mining: use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018.

[3] A. F. Adoma, N.-M. Henry, and W. Chen, "Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 117–121, IEEE, 2020.

[4] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

[5] S. Narkhede, "Understanding auc-roc curve," *Towards data science*, vol. 26, no. 1, pp. 220–227, 2018.

[6] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *Australasian joint conference on artificial intelligence*, pp. 1015–1021, Springer, 2006.

[7] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[9] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, 2015.