Project Machine Learning
— Milestone 2 —

[Imene Ben Ammar, Julian Dobler, Yannik Queisler]

December 15, 2024

Abstract

This project aims to fine-tune BERT for toxic comment classification, addressing the challenges of identifying harmful online content. Baseline methods such as logistic regression, support vector machines, and random forest were implemented and evaluated on the Civil Comments dataset as part of this milestone. These methods served as a reference for understanding the dataset and establishing performance benchmarks for BERT. Key findings revealed that logistic regression with preprocessing performed best among the baselines, achieving high-weighted F1 and AUC-ROC scores. However, precision and recall for the minority toxic class (C1) remained suboptimal. The dataset's significant class imbalance was identified as the primary challenge, emphasizing the need for advanced methods to improve performance. Future work will involve extending baseline evaluations to additional datasets (e.g., Jigsaw Toxicity and SST-2) and fine-tuning BERT for comparative analysis. The code for this project is publicly available on GitHub[1].

Contents

[1] https://github.com/devWhyqueue/pml-bert

1 Introduction

The exponential growth of social media platforms has led to unprecedented levels of online engagement, but it has also introduced significant challenges, particularly concerning the prevalence of toxic comments. Toxic comments have become a major issue, characterized by rudeness, disrespect, and a tendency to disrupt or drive participants out of discussions. [1]

Their presence exacerbates the problem of cyberbullying, with studies showing that approximately 15% of teens have experienced online bullying, which is strongly linked to psychological problems such as depression and anxiety. [2]

Effective moderation is essential, but manual approaches are not scalable given the sheer volume of user-generated content. This makes automated toxic comment classification a critical task for ensuring safer and more inclusive online spaces.

Machine learning (ML) and deep learning methods have emerged as promising tools to tackle this problem. Classical approaches like naive Bayes and support vector machines (SVMs) have been employed as benchmarks, while more sophisticated models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and BERT have demonstrated considerable success. [1]

The focus of this report is on achieving the first milestone in this project, which involves implementing baseline methods for toxic comment classification. This milestone is designed to familiarize us with the datasets, establish performance baselines, create a prototype BERT model, and identify suitable evaluation metrics. The subsequent sections of this report will address these objectives in the order outlined, providing a foundation for further advancements in the project.

2 Results

This section presents the evaluation of our baselines' performances on the Civil Comments data set. The evaluation is conducted using multiple standard metrics to assess its effectiveness in classifying toxic and non-toxic comments. Note that this dataset poses a binary classification task. Therefore, category differences were no issue, and adaption of classic binary validation metrics was not necessary.

We include the metrics accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These metrics are widely used in the literature, allowing us to ensure that our results are comparable with existing studies. [1, 3, 4]

The results highlight the method's strengths and provide insights into areas requiring further improvement.

2.1 Evaluation metrics

The metrics are defined as follows.

Accuracy measures the proportion of all classification instances that are classified correctly:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{1}$$

Where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

Precision indicates the proportion of predicted positive instances that are positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

Recall measures the amount of correctly positive classified instances proportionate to the true amount of positive classifications:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

F1-score represents the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

Finally, the area under the receiver operating characteristic curve (**AUC-ROC**) measures the ability of the model to distinguish between classes across all classification thresholds. It is computed as:

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{5}$$

where TPR (true positive rate) and FPR (false positive rate) are defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{6}$$

We evaluate the performances using precision, recall, and F1-score specifically for the toxic class (Class 1) because it represents the primary focus in toxic comment classification. The ability to correctly identify and classify toxic comments is critical, as these instances typically have a disproportionate impact on user experience and platform safety. By focusing on these metrics for Class 1, we ensure the evaluation reflects the model's effectiveness in addressing the key challenge of identifying toxicity.

Additionally, we include macro-averaged F1 and weighted-averaged F1 scores to provide a holistic evaluation across all comment classes. The macro-averaged F1 treats all classes equally, offering insights into how well the model performs across all types of comments, regardless of class imbalance. The weighted-averaged F1 adjusts for class distribution, ensuring that the evaluation accounts for the prevalence of non-toxic comments, which are the majority class.

Finally, overall accuracy is reported as a general performance metric for completeness, though its limitations in imbalanced datasets necessitate reliance on the other metrics for a more comprehensive understanding of the model's ability to detect toxic comments.

2.2 Findings

Table 1 shows the results of the grid search over baseline methods and hyper-parameters pre-processing and proportion of positive samples. They reveal important insights into the performance of the classifiers and the challenges inherent in the task.

Table 1: Classifier performance on the Civil Comments validation dataset

Method	**Prep.**	**Pos.**	**Precision**	**Recall**	**F1**	**Macro**	**Weighted**	**Acc.**	**AUC**
		Prop.	**(C1)**	**(C1)**	**(C1)**	**F1**	**F1**		**ROC**
Naïve Bayes	True	0.1	**0.86**	0.15	0.25	0.61	0.93	0.95	0.88
		0.25	0.58	0.39	0.47	0.72	0.94	0.95	0.88
		0.5	0.14	**0.85**	0.25	0.53	0.77	0.69	0.86
	False	0.1	0.83	0.18	0.29	0.63	0.93	0.95	0.87
		0.25	0.52	0.44	0.48	0.73	0.94	0.94	0.87
		0.5	0.14	**0.85**	0.24	0.52	0.76	0.68	0.86
Support Vector Machine	True	0.1	0.52	0.60	0.56	0.76	0.95	0.94	0.89
		0.25	0.37	0.72	0.49	0.72	0.92	0.91	0.90
		0.5	0.21	0.82	0.34	0.61	0.86	0.81	0.89
	False	0.1	0.51	0.59	0.55	0.76	0.94	0.94	0.88
		0.25	0.38	0.68	0.49	0.72	0.93	0.92	0.89
		0.5	0.25	0.76	0.37	0.64	0.88	0.85	0.88
Logistic Regression	True	0.1	0.67	0.43	0.52	**0.75**	**0.95**	**0.95**	**0.90**
		0.25	0.50	0.55	**0.53**	**0.75**	0.94	0.94	**0.90**
		0.5	0.22	0.78	0.34	0.62	0.87	0.82	0.88
	False	0.1	0.70	0.37	0.48	0.73	0.95	0.99	0.89
		0.25	0.56	0.47	0.51	0.74	0.94	0.95	0.89
		0.5	0.27	0.68	0.39	0.66	0.90	0.87	0.87
Random Forest	True	0.1	0.47	0.61	**0.53**	**0.75**	0.94	0.94	0.88
		0.25	0.37	0.73	0.49	0.72	0.92	0.91	0.90
		0.5	0.22	0.81	0.35	0.62	0.86	0.82	0.89
	False	0.1	0.45	0.61	0.52	0.74	0.94	0.93	0.87
		0.25	0.37	0.69	0.49	0.72	0.92	0.91	0.88
		0.5	0.22	0.75	0.34	0.62	0.87	0.83	0.85

Logistic Regression, when combined with preprocessing and evaluated on positive sample proportions of 0.1 and 0.25, demonstrated the best performance among all tested methods. These configurations achieved the highest overall scores across critical metrics, including weighted F1, accuracy, and AUC-ROC. The evaluation on the test set confirmed these findings, with all three metrics reaching satisfactory levels ($\geq$ 0.9), indicating robust performance in distinguishing toxic and non-toxic comments. This consistency suggests that overfitting is not a significant concern in these configurations, further reinforcing their reliability and generalizability.

Despite the overall strong performance, the precision, recall, and F1 scores for the positive class (C1) remained suboptimal, peaking at only 0.53. This limitation suggests that the model struggles with effectively identifying and classifying toxic comments. The implications of these results highlight the need for further investigation into strategies to improve the model's sensitivity to toxic content, such as better handling of class imbalances, enhancing feature representation, or exploring alternative architectures.

Preprocessing had no significant or consistent impact across all metrics. While it slightly enhanced performance in certain configurations, the benefit was not universally observed across all classifiers or dataset splits. This suggests that preprocessing alone is insufficient to address the challenges posed by the dataset's inherent characteristics, such as class imbalance or feature sparsity.

The imbalance of the dataset emerged as the most critical factor influencing classifier performance. Higher positive proportions led to noticeable declines in precision and overall F1 scores, underscoring the challenges of effectively handling the minority toxic class. These results emphasize the importance of addressing imbalance, potentially through oversampling, undersampling, or class-weighted learning techniques, to achieve more reliable toxic comment classification.

3 Discussion

Logistic Regression, particularly with preprocessing and optimized positive sample proportions (0.1 and 0.25), proved to be the most reliable baseline, achieving high weighted F1, accuracy, and AUC-ROC scores, with minimal signs of overfitting as performance remained consistent across validation and test sets. This indicates that the model is learning generalizable patterns effectively. However, the approach struggles to identify toxic comments accurately, with precision, recall, and F1 scores for the toxic class (C1) remaining unsatisfactory, peaking at only 0.53.

Additionally, TF-IDF vectorization was found to be unsuitable, confirming its limitations in capturing the complex linguistic features needed for this task. Despite preprocessing providing marginal improvements in some configurations, it did not show consistent benefits across all metrics, highlighting the challenges posed by the dataset's significant class imbalance and the need for more sophisticated approaches to address these limitations.

These results align with findings in the wider body of research on toxic comment classification. As observed by Gladwin et al., traditional techniques like TF-IDF for feature extraction often fail to capture the complex semantic and syntactic nuances necessary for this task, particularly in datasets with significant class imbalance. This reinforces the need for more advanced text representation techniques, such as word embeddings or transformer-based embeddings, to better capture the intricacies of online discourse. [4]

3.1 Task complexity

The Civil Comments dataset presents a unique combination of challenges that make the task of toxic comment classification non-trivial:

- **Class Imbalance:** Toxic comments represent a small fraction of the dataset, resulting in models that struggle to generalize well for the minority class (C1). As shown in our results, while overall accuracy and weighted F1 scores are satisfactory, precision and recall for the toxic class are not.

- **Linguistic Variability:** Toxic comments often involve subtle linguistic cues, sarcasm, or coded language, making them hard to detect with traditional methods.

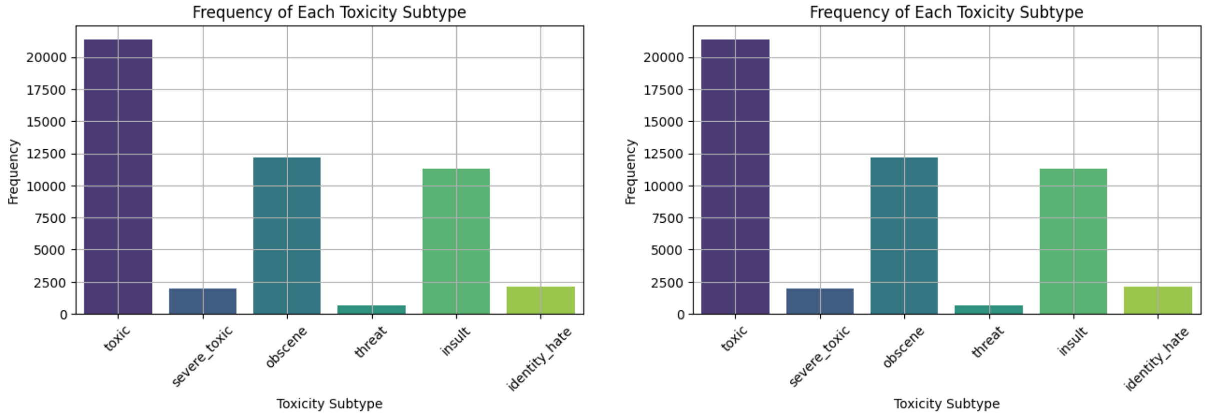
- **Practical Implications:** Despite these challenges, achieving even moderate success in this task is critical, as the insights provided by the final model can help identify harmful content on online platforms and foster safer digital spaces.

Based on our findings, achieving satisfactory performance on the Civil Comments dataset is possible but requires careful handling of class imbalance and feature representation. In a business context, deploying a model with these capabilities could have significant value for content moderation on social media platforms, forums, and other online communities. However, improving precision and recall for toxic comments is essential before practical implementation to avoid misclassifications that could undermine user trust.

3.2 Future directions

Several promising avenues for future work can be explored to enhance toxic comment classification. First, applying the baseline methods to additional datasets, such as Jigsaw Toxicity and SST-2, would enable a broader evaluation of their effectiveness and provide a foundation for comparing the performance of BERT against the baselines across all three datasets. Furthermore, exploring a non-binary classification format, such as toxicity subtype classification, may yield deeper insights into the nuances of toxic language and offer more granular moderation capabilities. Finally, fine-tuning our pre-trained BERT model on these datasets and comparing its performance to the current baseline methods could provide more satisfactory results.

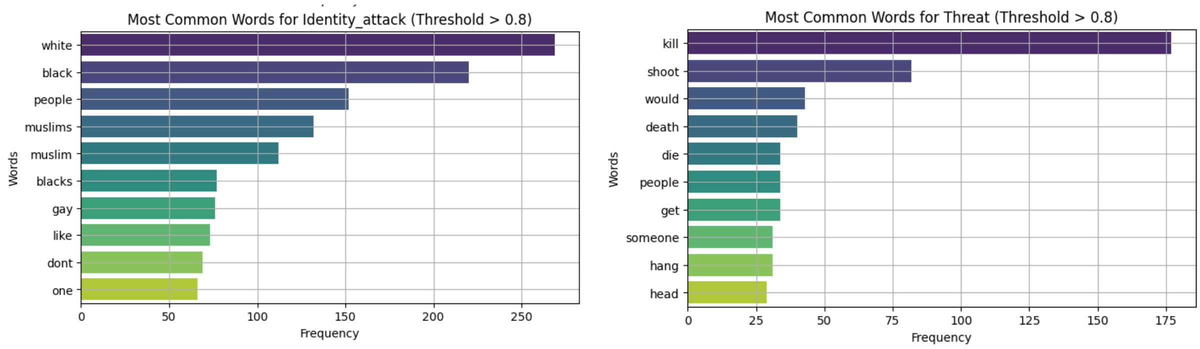
A Visualizations



(a) The Civil Comments dataset.

(b) The Jigsaw Toxicity dataset.

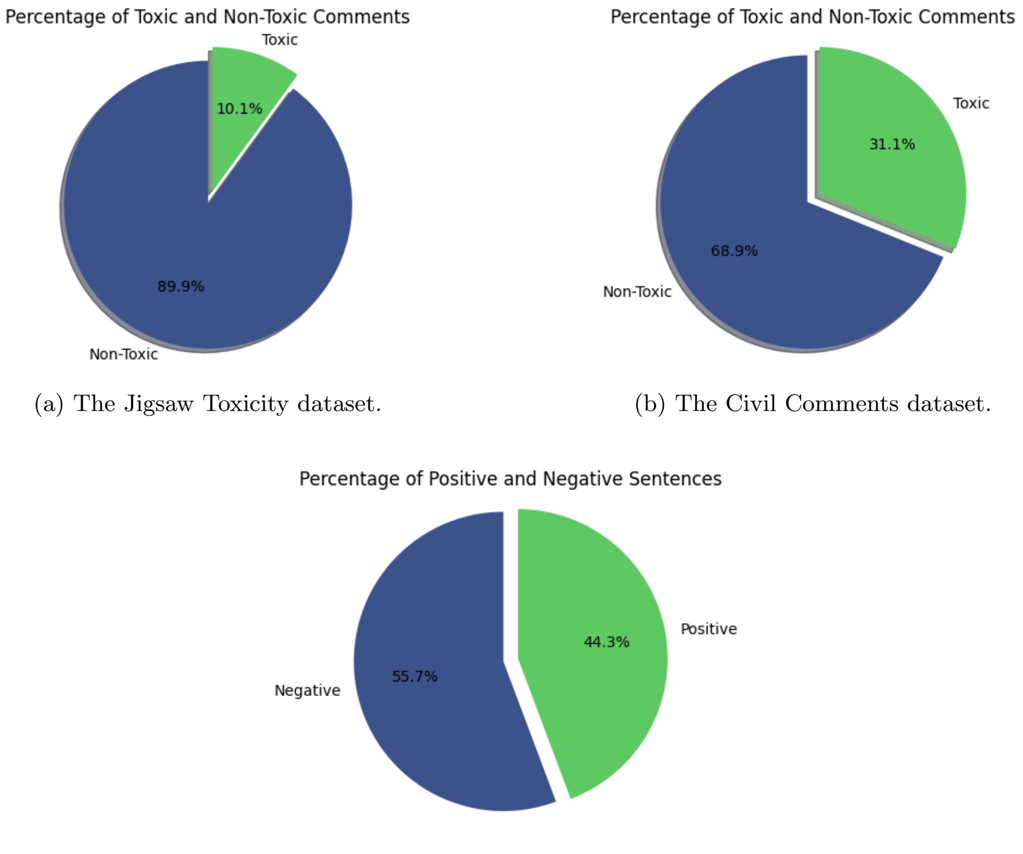
Figure 1: Toxicity frequency plots.



(a) identity_attack subtype.

(b) Threat subtype.

Figure 2: Example of word frequency plots for the Civil Comments datasets for two toxicity subtypes.



(a) The Jigsaw Toxicity dataset.

(b) The Civil Comments dataset.

(c) The SST-2 dataset.

Figure 3: Toxicity/Sentiment percentage.

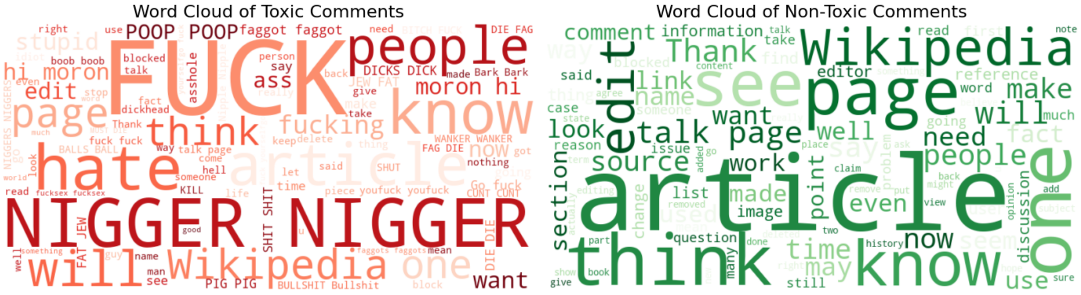
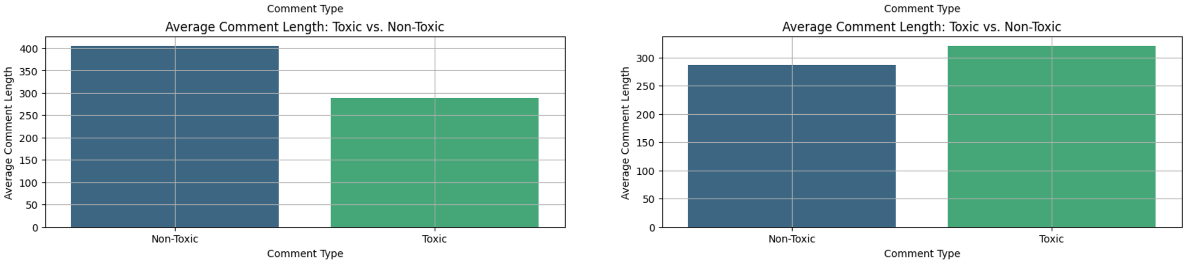


Figure 4: Example of Word Clouds for the Jisaw Toxicity dataset.



(a) The Jigsaw Toxicity dataset.

(b) The Civil Comments dataset.

Figure 5: Average comment length.

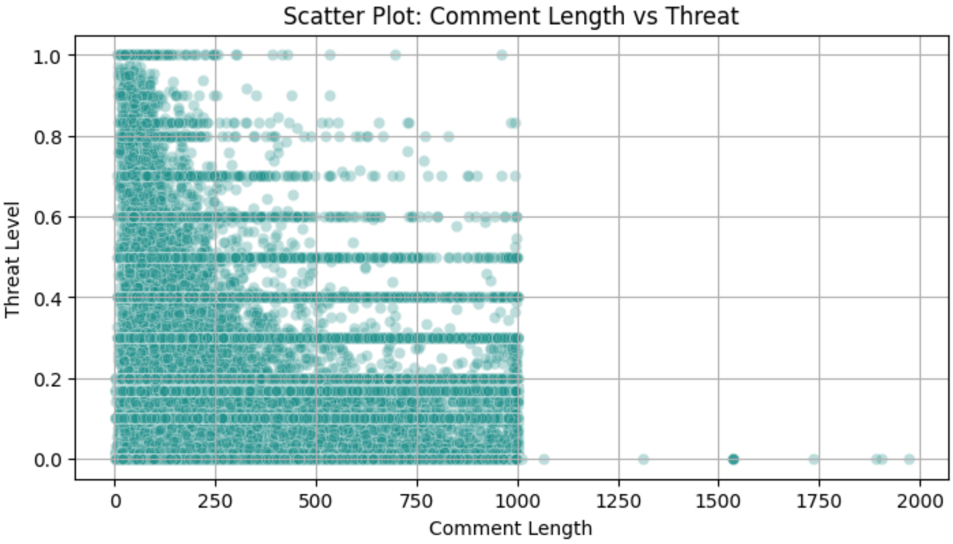


Figure 6: Toxicity subtype vs. comment length example for the Civil Comments dataset.

References

[1] D. Andročec, "Machine learning methods for toxic comment classification: A systematic review," *Acta Universitatis Sapientiae, Informatica*, vol. 12, no. 2, pp. 205–216, 2020.

[2] S. Zaheri, J. Leath, and D. Stroud, "Toxic comment classification," *SMU Data Science Review*, vol. 3, no. 1, p. 13, 2020. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol3/iss1/13

[3] C. Duchene, H. Jamet, P. Guillaume, and R. Dehak, "A benchmark for toxic comment classification on civil comments dataset," *arXiv preprint arXiv:2301.11125*, 2023.

[4] I. Gladwin, E. V. Renjiro, B. Valerian, I. S. Edbert, and D. Suhartono, "Toxic comment identification and classification using BERT and SVM," in *2022 8th International Conference on Science and Technology (ICST)*. IEEE, 2022, pp. 1–6.