Fine-Tuning Transformers for Toxicity Detection: A Sentiment-Informed Approach

Melody Masis

December 8, 2024

Abstract

Detecting toxic comments in online forums is crucial for fostering healthy discussions, especially in domains like personal health, where users may share sensitive information. This research explores the progression from traditional machine learning approaches, such as logistic regression, to advanced transformer-based models like DistilBERT and RoBERTa. To address the challenges of class imbalance and the nuanced nature of toxicity, it incorporates sentiment analysis as an additional feature and leverages a knowledge distillation framework to train a sentiment-enhanced student model.

The sentiment-aware student model outperformed its teacher (toxicity-only DistilBERT) in precision, recall, and F1-score, particularly for the minority toxic class. This demonstrates the potential of sentiment weighting to improve performance in imbalanced datasets. Visual analyses, including confusion matrices and error breakdowns, illustrate the significant reduction in false negatives and false positives achieved by the student model. Furthermore, this study highlights the importance of efficient memory usage and optimization for deploying transformer models in resource-constrained environments.

This work underscores the value of sentiment-informed training in improving toxicity detection while offering a scalable approach for real-world applications. Future efforts will focus on further refining the model and expanding its application to other domains with nuanced and imbalanced data.

1 Introduction

After soft launching the website healthbetweenus.org I became deeply aware that opening a forum intended for users to talk about personal health matters would require careful moderation, working full time and pursuing a masters in data science would leave me little time to dedicate to overseeing and manually flag any potential toxic comments for removal. This motivated me to advance research in toxicity detection using machine learning via natural language processing.

Leveraging a quintessentially classic dataset from Kaggle for toxicity detection, this documents the journey from a plain regression model to comparing different transformer model performance, and the creativity required to optimize these resource expensive models tracking memory usage and testing knowledge distillation as a potential way to be more effective with GPU constraints.

2 A Journey from Logistic Regression to DistilBERT

2.1 Dataset Preparation

All dataset from Kaggle's Toxic Comment Classification Challenge were loaded as is and train_df was split into training train_data and validation sets validation_data with 20% reserved for validation.

2.1.1 Dataset Statistics:

Label counts visualized for both the training and test datasets, focusing on the toxic and sub-labels (severe_toxic, obscene, threat, insult and identity hate) revealed an imbalance where the majority of the comments were not labeled as toxic, for example only 6090 out of the 153164 rows in the test data set or 3.9% of total. Rows with toxic labels were further classified as obscene, or an insult:

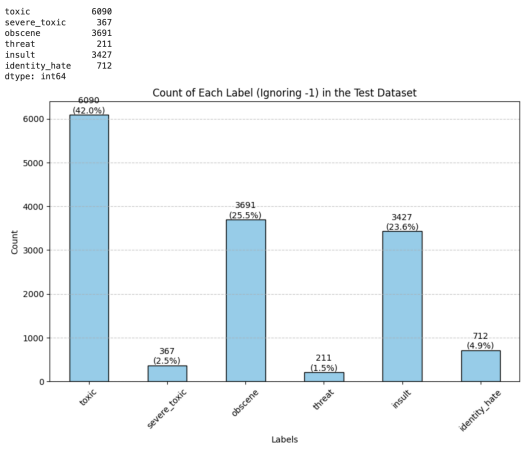

Figure 1: Toxic Sub Labels as % of Total Toxic Comments

2.1.2 Preprocessing:

- Used TF-IDF to convert comment_text into numerical features.
- Features (X_train, X_test) and target labels (y_train, y_test) were prepared for machine learning models.
- Removed rows with invalid or unwanted toxic label values (-1) from the labeled test data.

2.2 Baseline Model - Fine-tuning

Trained a logistic regression model to classify toxic vs. non-toxic comments and achieved high accuracy for the majority (non-toxic) class with 99% precision. However, saw relatively poor precision (41%) for the minority (toxic) class due to class imbalance.

2.2.1 Performance Metrics:

Achieved overall accuracy of about 86%. Confusion matrix in analyzes the model's strengths and weaknesses.

- **Strengths:** The model is excellent at identifying non-toxic comments, as seen by the high precision (0.99) and F1 score (0.92) for class 0. The recall for toxic comments (1) is also high (0.90), meaning the model is good at finding toxic comments.
- **Weaknesses**: The precision for toxic comments is low (0.41), indicating the model misclassified many non-toxic comments as toxic. This can be typical for imbalanced datasets where the majority class dominates the model's focus.

2.2.2 Improvement Ideas

- **Class Imbalance Handling:** Oversample the toxic class (1) using SMOTE or undersample the non-toxic class (0). This is not done in this research.
- **Explore Advanced Models:** Use transformer-based models like RoBERTa or DistilBERT for better understanding of the nuanced text. This is the next step completed in this research.

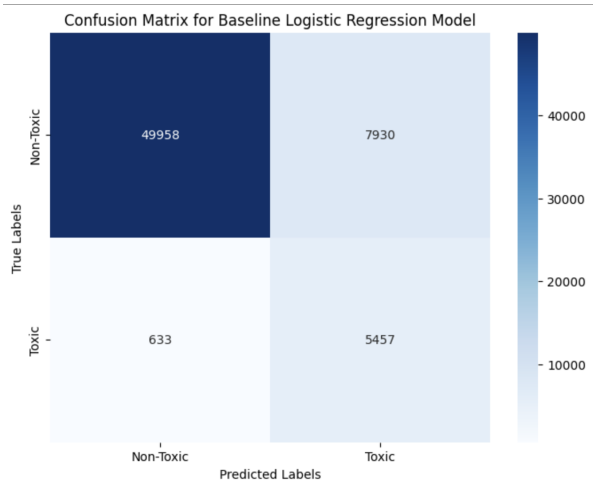

Figure 2: Class Imbalance visible in Confusion Matrix

3 Advanced Transformer Models

3.1 DistilBERT:

Chose this model as it is a smaller, faster, and lighter version of BERT achieved through knowledge distillation. It retains 97% of BERT's language understanding capabilities while being 60% faster and 40% smaller. [SDCW19]

Finetuned the distilbert-base-uncased model for toxicity classification using Hugging Face Transformers library, tokenizing data and loading it into a Hugging Face Dataset.

3.1.1 DistilBERT Results

Class 0 (Non-Toxic): Precision was 0.89, when the model predicts "non-toxic," it is correct 89% of the time. Recall was 0.78, out of all the true "non-toxic" comments, the model correctly identifies 78% of them. F1 Score was 0.83, this is the harmonic mean of precision and recall, indicating solid performance for this class.

Class 1 (Toxic): Precision: 0.05 - When the model predicts "toxic," it is correct only 5% of the time. Recall: 0.10, out of all the true "toxic" comments, the model only identifies 10% of them correctly. F1-Score: 0.06, this very low score reflects the poor balance between precision and recall for the toxic class.

Overall: Accuracy: 0.72, the model correctly classifies 72% of all comments, which is largely due to its strong performance on the "non-toxic" class. Weighted Avg F1 Score: 0.76, this shows an imbalance, with the model being heavily biased toward predicting "non-toxic."

Challenges: too much memory allocation and mismatched tensor shapes, ended up using **8.52 MB** of additional memory during inference.

3.2 RoBERTa Toxicity Classifier:

As an second experimentation option for comparison, loaded and applied s-nlp roberta_toxicity_classifier as a baseline transformer model, fine-tune it on the toxic comment dataset to establish a benchmark for accuracy, F1-score, and computational cost.

3.2.1 RoBERTa Results

Class 0 (Non-Toxic): Precision was 0.98, when the model predicts "non-toxic," it is correct 98% of the time. Recall was 0.96, out of all the true "non-toxic" comments, the model correctly identifies 96% of them. And F1-Score was 0.97, very high performance for this class.

Class 1 (Toxic): Precision was 0.69, when the model predicts "toxic," it is correct 69% of the time. Recall was 0.85, out of all the true "toxic" comments, the model identifies 85% of them correctly. And with F1-Score at 0.76, this balanced score reflects strong performance for the toxic class.

Overall: Accuracy was more than satisfactory, the model correctly classifies 95% of all comments. Weighted Avg F1 Score was 0.95, this shows balanced performance across both classes, with a slight bias toward the "non-toxic" class (which is expected due to the class imbalance). **Memory:** Implemented batch inference for scalability and monitored GPU memory usage, totaling **9.81 GB** of additional memory during this inference process, which was more expensive than DistilBERT, well worth the expense given the results.

3.3 ToxicChat (T5 Model):

Attempted batch processing using the lmsys/toxicchat-t5-large-v1.0 model [LWT+23] as it appear as state of the art innovation on toxicity detection models. Focused on inference efficiency and memory management. Results were not finalized due to runtime and performance constraints, inference was aborted due to high computational expense and documentation did not appear up to date.

3.4 Transformer Model Performance Comparison

Both models perform much better on the "non-toxic" class (Class 0) than the "toxic" class (Class 1), which is likely due to class imbalance in the dataset.

Table 1: Comparison of DistilBERT and RoBERTa Performance

Model	Class	Precision	Recall	F1-Score
DistilBERT	0	0.89	0.78	0.83
DistilBERT	1	0.05	0.10	0.06
DistilBERT	Accuracy		0.72	
DistilBERT	Macro Avg	0.47	0.44	0.45
DistilBERT	Weighted Avg	0.81	0.72	0.76
RoBERTa	0	0.98	0.96	0.97
RoBERTa	1	0.69	0.85	0.76
RoBERTa	Accuracy		0.95	
RoBERTa	Macro Avg	0.84	0.91	0.87
RoBERTa	Weighted Avg	0.96	0.95	0.95

DistilBERT vs. RoBERTa: DistilBERT struggles significantly with identifying "toxic" comments (low precision and recall). RoBERTa performs substantially better, especially on the "toxic" class, with an F1-Score of 0.76 compared to DistilBERT's 0.06.

Overall Performance: RoBERTa is far superior, with an overall accuracy of 0.95 and a balanced weighted F1-Score. DistilBERT has acceptable accuracy (0.72) but fails to generalize well to the minority "toxic" class. Next step is to improve Distilbert by including to fine tune using Sentiment Analysis.

4 Sentiment Analysis Aware Models

4.1 The Value of Sentiment Analysis

The next step in this research was to incorporate using sentiment information in order to improve the toxicity detection task, since it is possible for a user to mask toxic words by substitutions, but it is harder to mask the sentiment of a message. [BMST+23] Curiously, when detecting sentiment analysis, only the labels obscene and insult were classified as negative:

4.2 Sentiment as Feature in Baseline Model

The sentiment analysis improves the regression analysis performance dramatically. Including sentiment as a feature provides better discrimination between toxic and non-toxic comments, removes the invalid class (-1) issue and results in a more balanced model, especially for the toxic class.

Inclusion of Sentiment Improves Performance: With sentiment, the accuracy increases dramatically from 0.39 to 0.87. The precision and recall for toxic comments (Class 1) both improve

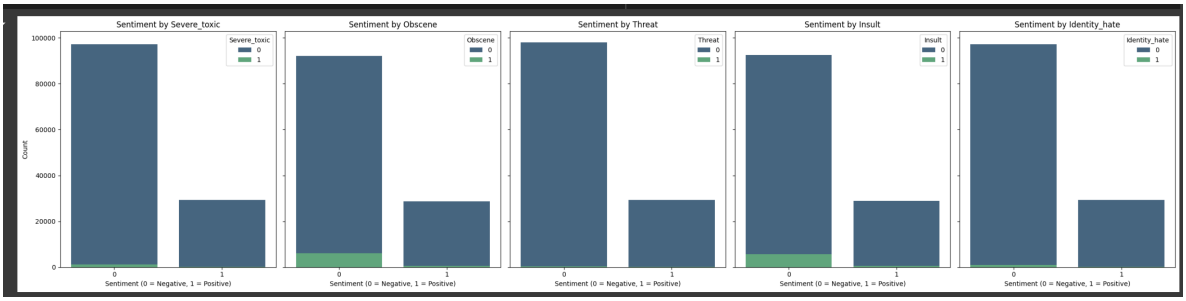

Figure 3: Sentiment by Sub-Labels

significantly when sentiment is included (Precision: 0.41 vs. 0.17, Recall: 0.90 vs. 0.69). The F1 score for toxic comments also improves (0.56 vs. 0.27).

Handling of Invalid Class: Without sentiment, the model includes invalid class predictions (-1), leading to poor performance and a 0.00 F1 score for that class. With sentiment, the focus is on valid labels (0 and 1), resulting in better overall metrics.

Class 0 (Non-Toxic): Both precision and recall are higher with sentiment included, resulting in better F1 scores for non-toxic comments (0.92 vs. 0.60).

Overall Balance: The weighted average and macro averages for precision, recall, and F1-score are significantly better with sentiment included.

4.3 Adding Sentiment to Transformer Model

Incorporating sentiment analysis into the DistilBERT model did not improve its ability to identify toxic comments. The metrics for both toxic (Class 1) and non-toxic (Class 0) comments are identical with and without sentiment.

This suggests that sentiment analysis may not contribute additional discriminatory power for the task of toxicity detection in this setup. The model most likely is already be capturing features correlated with sentiment implicitly, making the explicit addition redundant.

4.4 Distillation Experiment - Teacher/Student Model

Continuing to leverage this newly trained Distilbert Model, performed a distillation experiment with Sentiment Weighting to see if there was additional upside in efficiency for a transformer model in a resource-constrained environment [ANLQ23].

4.4.1 Results and Conclusion

Table 2: Comparison of Teacher Model (Toxicity-Only DistilBERT) and Student Model (Sentiment-Enhanced DistilBERT)

Metric	Teacher Model (Toxicity-Only DistilBERT)	Student Model (Sentiment-Enhanc
Precision (Toxic)	0.05	0.56
Recall (Toxic)	0.10	0.87
F1-Score (Toxic)	0.06	0.68
Accuracy	0.72	0.92
True Negatives (TN)	45,309	53,667
False Positives (FP)	12,579	4,221
True Positives (TP)	614	5,322
False Negatives (FN)	5,476	768

Comparison Table The student model vastly outperforms the teacher model across all metrics, showing the benefit of sentiment-enhanced training. The confusion matrix comparison highlights fewer false negatives and false positives in the student model.

- The teacher model struggles with identifying toxic comments, as reflected by the low precision, recall, and F1-score for the toxic class.
- Accuracy is high due to the class imbalance favoring the majority non-toxic class.
- The student model significantly improves upon the teacher model, especially for the toxic class.
- Higher precision, recall, and F1-score for the toxic class indicate that sentiment-enhanced training helps detect toxicity more effectively.
- Accuracy improvement is evident, with better discrimination of both classes.

4.4.2 Key Takeaways:

- Incorporating sentiment data in the student model enhances its ability to detect toxicity, resulting in better precision, recall, and overall F1-score.
- The student model demonstrates improved generalization with fewer errors compared to the teacher model.
- Training a smaller model (student) using knowledge distillation and sentiment-aware weighting is effective for tasks involving subtle and nuanced predictions like toxicity detection.

5 Future Work

5.0.1 Next Steps:

- Further fine-tune the student model to reduce errors for positive sentiment cases. Test RoBERTa as the teacher model.
- Explore additional features (e.g., context or user behavior) to further improve toxic class predictions.
- Apply this methodology to other domains with imbalanced data and nuanced prediction requirements.

References

[ANLQ23] Ahlam Husni Abu Nada, Siddique Latif, and Junaid Qadir. Lightweight toxicity detection in spoken language: A transformer-based approach for edge devices. *arXiv preprint arXiv:2304.11408*, 2023.

[BMST+23] Andrea Bonetti, Marcelino Martínez-Sober, Julio C. Torres, Jose M. Vega, Sebastien Pellerin, and Joan Vila-Francés. Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks. *Applied Sciences*, 13(10):6038, 2023. © 2023 by the authors. Distributed under the terms of the Creative Commons Attribution (CC BY) license.

[LWT+23] Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv preprint arXiv:2310.17389*, 2023.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC2 Workshop*, 2019.