Harnessing Text Analytics to Shield Children from Online Toxicity: A Moderator Extension

INSY 669 Group Project Report

Atharva Vyas

Richard El Chaar

Hazel Foo

Margot Gerard

Introduction

The increasing prevalence of social media has created an environment where children are routinely exposed to harmful content. Despite efforts to implement content moderation, existing approaches often lack the sophistication needed to capture the nuanced forms of toxicity that exist in online discourse. Sarcasm, evolving slang, and implicit derogatory language frequently bypass standard filters, rendering children vulnerable to psychological distress. While many platforms deploy keyword-based filtering and basic AI-driven moderation tools, these often fail to differentiate between harmful and benign language effectively. Furthermore, automated moderation systems are susceptible to biases inherent in their training data, which may result in either over-censorship or an inability to detect disguised toxicity. Our project sought to address these limitations by developing a robust text analytics model capable of accurately classifying toxic comments and adapting to the complexities of online language. By integrating such a model into a Chrome extension, we envision a system that operates in real-time, minimizing children's exposure to harmful digital interactions while allowing for more refined content control.

Methodology

Our approach consisted of three phases: constructing and training text classifiers, finetuning model parameters to enhance performance, and validating the models on real-world, external datasets.

The first step involved curating a dataset from Kaggle's *Jigsaw Toxic Comment Classification Challenge*, comprising of 159,450 comments extracted from Wikipedia's talk pages. These comments were labeled according to 6 toxicity categories: toxic, severe toxic, obscenity, identity hate, insult, and threat. To efficiently explore different model architectures, the dataset was scaled down to 50,000 comments before investing the substantial resources required to train on the complete dataset. Given the imbalance in non-toxic comments, we ensured proportional representation of each category in the scaled dataset to prevent model bias. Preprocessing techniques, including tokenization, lemmatization, and sentiment analysis using VADER, were applied to optimize text representation for classification. To better understand the extent of toxicity, comments were reclassified into four bins: non-toxic where the comment was not given any toxicity label, mild toxicity where 1 toxic label was given, moderate toxicity for 2 labels and severe toxicity for 3 labels and above. The sentiment analysis revealed that sentiment scores ranged from –1 to 1 for most categories, reinforcing the notion that toxicity is often masked within seemingly neutral linguistic structures, necessitating a more advanced classification methodology rather than mere sentiment analysis alone.

Next, we implemented 4 models to develop our toxicity classifier, including Naïve Bayes (NB), Support Vector Machine (SVM), Gradient Boosting (GB), and a Multi-Layer Perceptron (MLP) Classifier. These models were trained using Term Frequency-Inverse Document Frequency and Count Vectorizer features, emphasizing the importance of word frequency distributions in determining toxicity. Each model underwent hyperparameter optimization through Randomized or Grid Search CV to finetune learning rates, depth constraints, feature selections, and regularization parameters. The comparative analysis of these models revealed key insights into their performance trade-offs.

Results & Discussion

NB, while computationally efficient, demonstrated a tendency to favor majority-class predictions, achieving an accuracy of 93% with an AUC-ROC of 90% (refer to Exhibit 1). However, its lower recall for toxic comments (69% vs 96% for non-toxic comments) indicated that it was unsuitable for real-world applications where false negatives could be particularly damaging. GB emerged as a strong contender with a 96% accuracy and an AUC-ROC of 96% (refer to Exhibit 2), demonstrating its capability to capture complex textual patterns. However, its higher computational cost raised concerns about scalability in a real-time application.

SVM struck a balance between computational efficiency and predictive power, yielding an accuracy of 95% and an AUC-ROC of 95% (refer to Exhibit 3). However, during hyperparameter tuning, the

optimal value for class weight was determined to be 'None'. This posed a potential concern regarding the model's sensitivity to the class imbalance in the dataset, where non-toxic comments dominated. A model without class weights might be biased towards the majority class, potentially underperforming in identifying toxic comments. Despite that, the tighter convergence in the SVM's learning curves compared to the other models suggests less overfitting and better generalization potential.

The MLP classifier showed initial promise, achieving 95% accuracy and a 94% AUC-ROC. Analysis of its performance on toxic comment identification revealed a reasonable balance between recall and precision. Specifically, the model achieved 71% recall and 77% precision for toxic comments, demonstrating its ability to capture a substantial portion of true positives while limiting false positives. However, it exhibited the largest difference between training and validation accuracy among the models evaluated. This persistent gap, as seen in the learning curves (refer to Exhibit 4), raised concerns about potential overfitting and uncertainty regarding its performance on unseen data.

To assess the robustness of these models, we conducted an external validation using 200 manually labeled Reddit comments sourced from *r/politics*. The real-world nature of this dataset, with its inherent variability and nuances in spoken language, posed significant challenges to classification accuracy. The results underscored the limitations of some models while highlighting the strengths of others. As anticipated from the learning curves, SVM outperformed its counterparts, achieving an F1 score of 0.72 and an AUC-ROC of 0.79 (refer to Exhibit 5), demonstrating its ability to maintain a balanced precision-recall trade-off. GB and NB models exhibited high precision but suffered from low recall, failing to identify a substantial proportion of toxic comments. The MLP, despite strong training performance, struggled with external validation data, highlighting the critical need to evaluate models on diverse, real-world datasets, not just controlled training environments.

Expected Impact

The practical implementation of our findings hinges on the development of a Chrome extension that integrates the most effective model, providing real-time toxicity detection for online content. While our current model demonstrates promising results, several improvements are necessary before deployment. Expanding the training dataset to include a broader spectrum of linguistic variations and cultural nuances will enhance the model's adaptability. Additionally, implementing an adaptive learning mechanism that incorporates user feedback—allowing parents and educators to review and adjust flagged content—will improve long-term accuracy. Monitoring false positives and false negatives will further refine the classification process, ensuring that safe content is not unnecessarily restricted while all toxic content is accurately identified.

Another critical component of the implementation strategy is refining contextual understanding within the model. Traditional machine learning approaches struggle to differentiate between content that is overtly toxic and content that contains complex, multi-layered interactions. Future iterations of the project should incorporate transformer-based architectures such as BERT or GPT to enhance contextual awareness and improve classification precision. Moreover, integrating multimodal analysis—considering images, videos, and emojis alongside textual content—would create a more comprehensive moderation system.

The implications of this project extend beyond content moderation. By leveraging advanced text analytics, we can contribute to the broader discourse on digital safety, encouraging platforms to adopt more sophisticated approaches to toxicity detection. The integration of AI-driven moderation tools into mainstream applications has the potential to reshape online interactions, fostering healthier digital spaces for younger audiences. Addressing toxicity at its core requires a combination of algorithmic precision and human oversight, reinforcing the importance of interdisciplinary collaboration between data scientists, educators, and policymakers. The findings from this study serve as a foundation for future research, paving the way for the next generation of intelligent, context-aware content moderation systems. Additionally, our work highlights the ethical considerations inherent in AI-driven moderation, urging stakeholders to establish transparent policies that balance protection with freedom of expression.

By iterating upon this research, we can move toward a digital ecosystem that prioritizes both safety and inclusivity.

Appendix

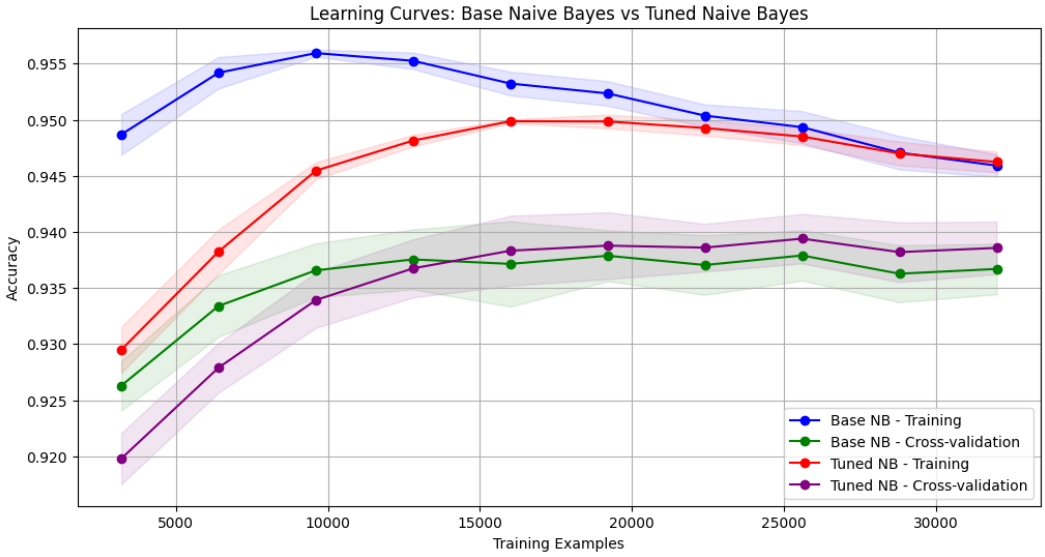

Exhibit 1: NB learning curves

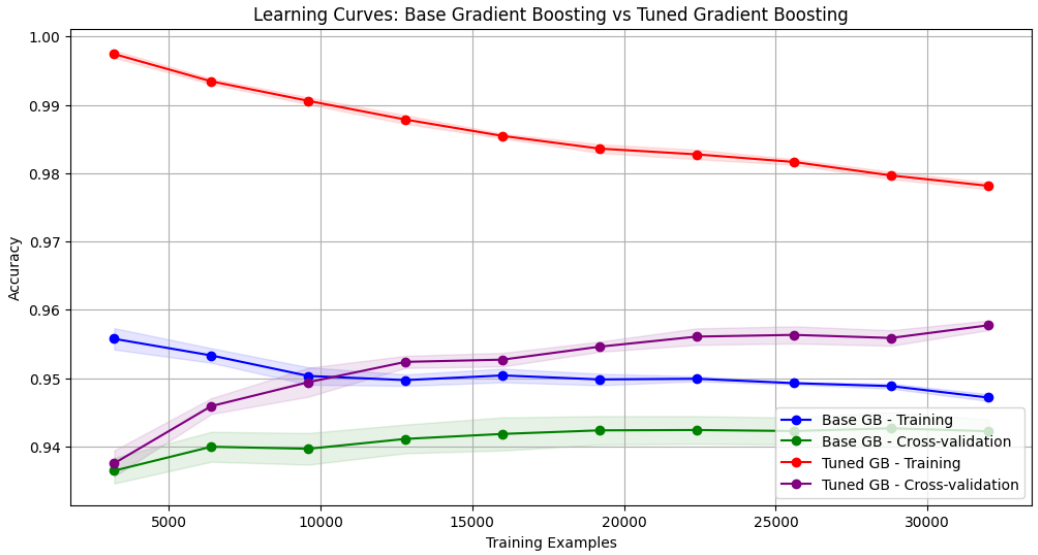

Exhibit 2: Gradient Boosting learning curves

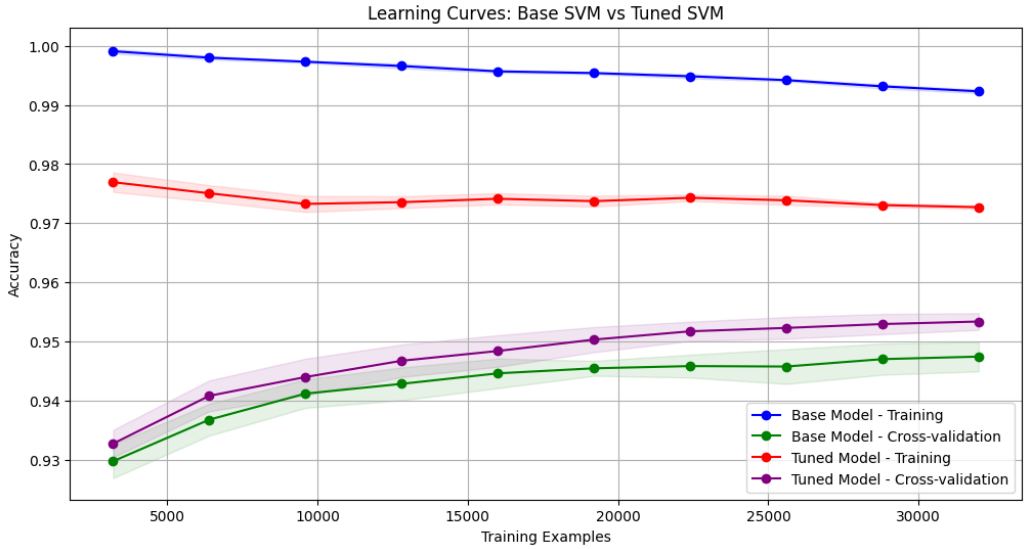

Exhibit 3: SVM learning curves

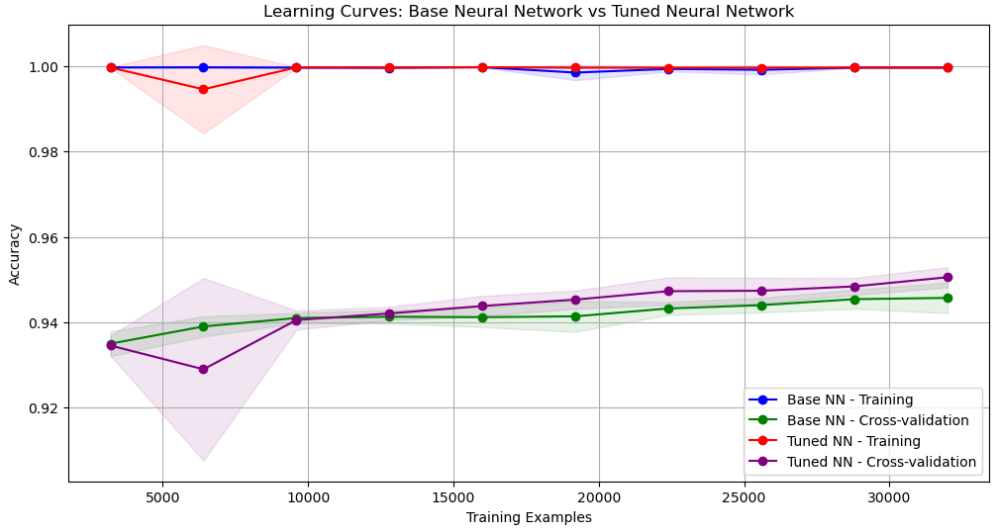

Exhibit 4: MLP learning curves

Metric	Naive Bayes	SVM	Gradient Boosting	Neural Network
F1 Score	0.39	0.72	0.55	0.16
AUC-ROC	0.64	0.79	0.69	0.51
Precision	0.47	0.93	1.00	0.14
Recall	0.33	0.58	0.38	0.21

Exhibit 5: Performance Metrics of the 4 models on reddit data