Human-in-the-Loop ML Project Proposal Template

Rishi More

1 Setting

The rapid expansion of online communities has revealed significant limitations in current toxic content detection systems. Traditional models, such as Bidirectional Long Short-Term Memory networks (BiLSTMs), are effective at identifying explicit toxicity but often struggle with nuanced, context-dependent harmful content across diverse platforms. Recent advancements have focused on transformer-based architectures Kamphuis [2024] and ensemble methods Mazari et al. [2024], achieving high accuracy on benchmarks. However, these models face issues like computational intensity and insensitivity to contextual nuances. Large Language Models (LLMs) also encounter challenges in understanding community-specific content, particularly with context-dependent language and implicit social norms Naveed et al. [2024]. Integrating methods that leverage community feedback is vital, as highlighted by Valmeekam et al. (2023), who discuss LLMs' limitations in planning and reasoning Valmeekam et al. [2023].

Approaches involving human-in-the-loop machine learning enable systems to adapt to evolving community standards by learning from human interactions. Fine-tuning a BiLSTM model using Behavior Cloning (BC) or Direct Preference Optimization (DPO) offers an efficient way to incorporate community feedback into context-aware toxicity detection. While BC trains the model to imitate behaviors labeled from community data, DPO optimizes the model based on preference comparisons between comments. This project focuses on comparing BC and DPO to determine which method more effectively captures context-specific nuances and enhances the detection of subtle toxicity, improving alignment with community preferences.

2 Methods and Research Question

The project compares the performance of Direct Preference Optimization (DPO) and Behavior Cloning (BC) in enhancing context-aware toxicity detection using a BiLSTM model. Both approaches leverage Reddit comment interactions, utilizing upvote and downvote ratios as indicators of community feedback for toxicity classification. The lightweight BiLSTM model processes input sequences bidirectionally, capturing contextual information from both past and future tokens, which is crucial for understanding nuanced language in online interactions.

In the BC approach, the model is trained and supervised using labels derived directly from Reddit voting patterns. Comments with higher upvote ratios are labeled as non-toxic, while those with higher downvote ratios are labeled as toxic. The model learns to map input sequences to these labels, effectively imitating the community's behavior as reflected in the voting data. The training objective is to minimize the binary cross-entropy loss between the predicted probabilities and the true labels:

$$\mathcal{L}_{\text{BC}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \tag{1}$$

On the other hand, the DPO approach optimizes the model to align with community preferences without the need for explicit labeling. Preference pairs (x_i^+, x_i^-) are created based on voting patterns: preferred comments (x_i^+) have higher upvote ratios, indicating alignment with community standards, while less preferred comments (x_i^-) have higher downvote ratios, suggesting less acceptance. The

model learns a scoring function $s_\theta(x)$, where θ represents the model parameters and x is the input comment. The training objective is to maximize the likelihood that preferred comments receive higher scores than less preferred ones:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log\sigma(s_\theta(x_i^+) - s_\theta(x_i^-)) \tag{2}$$

where σ is the sigmoid function, x_i^+ is the preferred comment, and x_i^- is the less preferred comment in the i_{th} pair.

Integrating voting patterns allows both models to adapt to evolving community standards without full retraining. This comparison aims to determine if DPO's direct utilization of preference data provides a more nuanced understanding of context and community norms compared to BC's imitation of labeled behaviors.

3 Experiments and Expected Outcomes

The experiments aim to compare the effectiveness of Direct Preference Optimization (DPO) and Behavior Cloning (BC) in enhancing context-aware toxicity detection using a BiLSTM model. Two primary datasets will be utilized: the Kaggle Toxic Comment Classification Challenge dataset, containing 160,000 Wikipedia comments labeled for toxicity, and a web-scraped dataset of approximately 100,000 Reddit comments, where upvote and downvote ratios serve as implicit feedback signals to derive community preferences. Initially, the BiLSTM model will be trained using SGD on the Kaggle dataset to establish a baseline for explicit toxicity detection. Following this, the model will be fine-tuned using either BC or DPO with the Reddit dataset. In the BC approach, Reddit comments will be labeled as toxic or non-toxic based on their upvote/downvote ratios, and the model will learn to predict these labels through supervised learning. In the DPO approach, the model will learn directly from pairs of comments, optimizing to prefer comments with higher upvote ratios without the need for explicit labels.

The experimental hypothesis is that fine-tuning the BiLSTM model using DPO will achieve comparable or superior accuracy to BC in both explicit and contextual toxicity detection while maintaining computational efficiency. Specifically, the model is expected to achieve an F1-score of over 85% for explicit toxicity detection and over 80% for contextual toxicity, with training time reduced to less than 50% of comparable transformer models.

The expected outcome is that the DPO approach will more effectively adapt the model to community-specific standards using implicit feedback, achieving high accuracy in toxicity detection while maintaining computational efficiency. This will enable the model to adapt to community norms without relying on extensive retraining or explicit moderator actions, thus demonstrating the potential advantages of DPO over BC in context-aware toxicity detection.

References

Michiel Kamphuis. Tiny-toxic-detector: A compact transformer-based model for toxic content detection, 2024. URL https://arxiv.org/abs/2409.02114.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. Bert-based ensemble learning for multi-aspect hate speech detection. *Cluster Computing*, 27(1):325–339, Feb 2024. ISSN 1573-7543. doi: 10.1007/s10586-022-03956-x. URL https://doi.org/10.1007/s10586-022-03956-x.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. URL https://arxiv.org/abs/2307.06435.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change, 2023. URL https://arxiv.org/abs/2206.10498.