Qeios

Peer Review

Review of: "ToxiLab: How Well Do Open-Source LLMs Generate Synthetic Toxicity Data?"

Harry Kabodha Kabodha[1]

1. Fu Foundation School of Engineering and Applied Sciences, Columbia University, United States

General Comments

This paper addresses a timely and critical topic in Natural Language Processing: the scarcity of high-quality data for toxic content detection and the potential of open-source Large Language Models (LLMs) to fill this gap. The authors present a well-organized study comparing prompt engineering against supervised fine-tuning (SFT) strategies.

The paper is structured logically, and the two-stage experimental design provides a clear narrative of the hypothesis testing. Furthermore, I commend the authors for explicitly acknowledging the ethical implications and potential risks associated with generating synthetic hate speech; this "Responsible AI" statement is essential for work in this domain.

However, while the premise is strong, the methodology and evaluation mechanisms rely on oversimplified architectures and binary frameworks that fail to capture the complexity of the subject matter.

Specific Critiques & Recommendations

1. Limitations of Binary Classification The current study relies heavily on binary labels (Positive/Harmful vs. Negative/Non-harmful). Toxicity in real-world scenarios is rarely black and white; it is highly contextual and exists on a spectrum. By forcing a binary rule, the study loses the depth required to evaluate whether the synthetic data captures nuanced toxicity (e.g., microaggressions, sarcasm, or dog whistles) versus overt hate speech.

doi.org/10.32388/SEPAAI

Recommendation: Future iterations should move away from strict binary classification in favor of multi-dimensional continuous scoring.

For example, the Perspective API approach utilizes probability scores (0.0 to 1.0) across distinct attributes like Severe Toxicity, Insult, and Identity Attack. Adopting a similar schema would allow for dynamic thresholding rather than a flat binary cutoff.

Additionally, implementing **Label Smoothing** (e.g., targets of 0.9 and 0.1 rather than 1 and 0) would prevent model overconfidence and better reflect the subjective nature of annotator disagreement.

2. Over-Simplicity of the Downstream Evaluator (MLP) The use of a Multi-Layer Perceptron (MLP) with two hidden layers as the primary downstream evaluation tool is a significant bottleneck. An MLP is too simple to effectively evaluate the semantic richness of data generated by complex LLMs like Mistral or Vicuna. If the synthetic data contains subtle contextual cues, a simple MLP is unlikely to capture them, potentially underreporting the efficacy of the fine-tuning.

Recommendation: The authors should employ architectures capable of modeling non-linear relationships and global corpus dependencies. Graph Neural Networks (GNNs) would be particularly effective here:

TextGCN: Constructing a word-document graph would allow the model to capture global word co-occurrence patterns, which is critical for detecting toxic terms that rely on specific community contexts.

3. Advanced Techniques for Handling Discreteness The generation and classification processes described appear somewhat rigid. To better handle the discrete nature of text generation and classification in a differentiable way, the methodology could be modernized.

Recommendation: I suggest exploring the use of the Gumbel-Softmax distribution. This would allow for the reparameterization of discrete distributions, enabling the model to learn smoother transitions between "toxic" and "non-toxic" states during training. This technique is particularly valuable when paired with the continuous scoring metrics suggested above.

4. Data Nuance and Diversity While the paper utilizes five datasets, the approach to these datasets feels aggregated and lacks distinct nuance handling. Simply mixing datasets (Data Mixing) improves generalization, but the paper does not sufficiently analyze why certain nuances are lost or retained.

Recommendation: A more rigorous breakdown of how the models handle specific linguistic nuances (slang, cultural context, obfuscation) is needed. The evaluation should prove that the synthetic data adds value to these "edge cases" rather than just bolstering the bulk detection of obvious slurs.

Conclusion

ToxiLab contributes valuable insights into the cost-effectiveness of fine-tuning open-source models for toxicity generation. However, to make the findings robust and applicable to state-of-the-art content moderation, the authors must move beyond binary classifications and simplistic evaluation models. Incorporating weighted scoring systems (like label smoothing or multi-attribute probability) and structural evaluation models (like GNNs or GATs) would significantly elevate the technical depth and real-world applicability of this research.

Declarations

Potential competing interests: No potential competing interests to declare.