Qeios

Peer Review

Review of: "Leveraging Large Language Models and Topic Modeling for Toxicity Classification"

Bedour Alrashidi[1]

1. Department of Computer Science and Information, Computer Science and Engineering College, University of Ha'il, Ha'il 55211, Saudi Arabia, University of Hail, Saudi Arabia

This article explores the impact of topic modeling on the performance of fine-tuned language models for toxicity classification. The authors fine-tune BERTweet and HateBERT on the NLPositionality dataset, employing Latent Dirichlet Allocation (LDA) to cluster data into topics. The results indicate that topic-specific fine-tuning improves F1 scores compared to general fine-tuning or relying on state-of-the-art classifiers like GPT-4, PerspectiveAPI, and RewireAPI.

Some modifications need to be considered to improve clarity, methodological justification, and discussion depth in the article. In the background section, redundant wording should be revised for clarity, such as changing "neural networks, especially LLMs, often exhibit biases caused by biases in their training data" to "neural networks, particularly LLMs, often exhibit biases due to biased training data." Additionally, a sentence discussing GPT-generated toxicity should be improved from "Studies indicate that significantly more toxic language can be generated using GPT by assigning its persona" to "Studies indicate that assigning personas to GPT models can lead to the generation of more toxic language." More recent research on the limitations of toxicity detection in LLMs should be incorporated. In the methodology section, the choice of LDA for topic modeling should be justified by comparing it with alternative methods like BERTopic or NMF, and additional details on model hyperparameters (e.g., batch size, dropout) should be included. For the results and discussion, a key sentence should be refined for readability, changing "state-of-the-art large language models exhibit significant limitations in accurately detecting and interpreting text toxicity contrasted with earlier methodologies" to "state-of-the-art LLMs struggle with accurately detecting and interpreting text toxicity compared to earlier methods." Expanding the discussion with error analysis, including case studies of misclassified examples, would

doi.org/10.32388/NSFUQZ

strengthen the findings, and the demographic bias analysis should explain why certain groups had higher or lower F1 scores. In the conclusion, a statement on future research should be revised for better impact: "Future research should focus on mitigating the biases present in widely used models like GPT, as their increasing popularity raises significant concerns" could be improved to "Future research should focus on bias mitigation in widely used models like GPT, as their growing adoption poses ethical concerns." Additionally, practical implementations, such as integrating fine-tuned models into content moderation systems, should be suggested. Overall, these modifications would enhance the article's clarity, coherence, and impact.

Declarations

Potential competing interests: No potential competing interests to declare.