Qeios

Peer Review

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

Apoorva Singh[1]

1. Fondazione Bruno Kessler, Trento, Italy

This research tackles the challenge of toxicity classification in online content, especially in short texts like tweets. It recognizes that existing models often perpetuate biases due to annotator subjectivity, leading to inaccurate toxicity detection that disproportionately affects marginalized groups. To enhance fairness and robustness in toxicity classification, the authors propose a novel approach combining topic modeling with fine-tuning of large language models, specifically BERTweet and HateBERT.

The methodology utilizes the NLPositionality dataset of labeled toxic tweets and annotator demographic metadata. The authors applied data preprocessing techniques and used Latent Dirichlet Allocation (LDA) for topic clustering. They fine-tuned BERTweet and HateBERT on topic-specific subsets from the LDA to capture features relevant to toxicity classification. The results show that this fine-tuning method significantly outperformed traditional classification models, demonstrating the effectiveness of topic-informed strategies in reducing biases and improving accuracy in toxicity classification.

There are several shortcomings of the work that the authors need to address:

Generalization of Pre-trained Models: BERTweet was primarily pre-trained on general tweets, which may detract from its effectiveness in the specific context of toxicity classification. This lack of specialization can impact its performance on the task at hand.

Impact of Additional Labels on HateBERT: Including a third label in the fine-tuning process for HateBERT, which was trained on a binary toxicity classification dataset, may have contributed to high error rates, particularly for the new label.

doi.org/10.32388/TFKRCY

Potential for High False Positive Rates: The results revealed high false positive rates for both fine-tuned models, indicating that the models might not effectively distinguish between toxic and non-toxic content in certain contexts.

Limited Exploration of Generalization: Although the research touches upon generalization to other platforms, it lacks extensive analysis or validation on different datasets beyond the NLPositionality dataset.

Insufficient Addressing of Biases: While the work highlights the impact of biases introduced by annotator positionality, it does not thoroughly explore strategies to mitigate these biases within the model training processes.

Few other suggestions to improve the methodology:

1. Evaluate the methodology on multiple different datasets to include a wider variety of platforms and content types, which could help generalize the findings across different contexts of online communication.

2. Employ algorithmic techniques specifically designed to mitigate bias, such as re-weighting examples during training or using adversarial debiasing methods.

Declarations

Potential competing interests: No potential competing interests to declare.