Qeios

Peer Review

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

Rui Song[1]

1. Jilin University, China

A work that is fairly good but has obvious flaws utilized topic information in the text toxicity detection task and demonstrated model recognition capabilities comparable to GPT-4. However, there are still the following shortcomings that need improvement:

1. The authors seem to intentionally mislead readers with the title, suggesting that their method is based on large models, while in fact, no improvements have been made to the large model itself. I recommend clearly distinguishing between the concepts of pre-trained models and large models.
2. If the authors can demonstrate that the topic-based method still effectively enhances the performance of large models (such as GPT-4), then this issue would no longer be a concern.
3. The authors should provide a more detailed introduction to their model in the introduction section, rather than briefly mentioning it in one sentence. This would help readers form their first impression of the method while reading the introduction.
4. Similarly, the authors do not provide enough detail in the Methods section. How does the topic model work? Is it used to fine-tune BERT with topic labels first, or is it trained jointly with toxicity classification? I suggest that the authors include more formal descriptions here to fully present the specific details of their work.
5. The authors should consider citing more recent works, such as:
    1. Data-Centric Explainable Debiasing for Improving Fairness in Pre-trained Language Models
    2. Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation
    3. Measuring and mitigating language model biases in abusive language detection

 doi.org/10.32388/X8RY1P

Based on my experience, if the authors can fully address the above issues, then it may be considered for acceptance after revisions.

Declarations

Potential competing interests: No potential competing interests to declare.