

Peer Review

Review of: "PANDA – Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to Create the First Chinese Counterspeech Dataset"

Akihiro Fujii¹

1. Hosei University, Japan

The paper introduces a new labeled Chinese hate speech dataset and proposes new evaluation metrics for assessing the quality and effectiveness of counterspeech in Chinese, providing support for future research in this research domain. However, I didn't see the apparent novelty of the work.

In section 2.5, "Current Evaluation Metrics," the authors mention several existing evaluation methods. However, these methods are general-purpose metrics for NLP-related downstream tasks. Given the unique challenges of hate speech detection, I believe these methods are not suitable for comparison in this section. Moreover, the authors seem to overlook some of the established hate speech detection metrics, which I find concerning.

I also have some doubts about the fairness and validity of the manual annotation process. The paper mentions that this process was conducted by a small group of annotators, but it lacks a clear explanation of how potential biases were mitigated, which raises concerns about the reliability of the annotations.

The authors mention that there is much more research on hate speech in English compared to Chinese and suggest applying existing methods to a Chinese corpus to compare the performance. While this may yield interesting results, it would be more convincing if the authors conducted both before-and-after comparisons to evaluate whether the existing methods are well-suited for Chinese.

I believe the paper lacks a comparative review of existing research from various perspectives, such as datasets, evaluation methods, detection approaches, and the differences between Chinese and English hate speech. This baseline omission makes the paper's contribution appear less clear.

I also did not see any significant difference between the proposed annealing and ranking mechanisms and traditional methods. The paper would benefit from a clearer explanation of how these mechanisms offer improvements.

Lastly, the paper's format should be carefully rechecked to ensure consistency throughout the manuscript.

Declarations

Potential competing interests: No potential competing interests to declare.