Qeios

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"

Parth Patwa[1]

1. University of California, Los Angeles, United States

Summary: The paper proposes a dataset of counterspeech to hatespeech in the Chinese language. The existing hatespeech dataset is filtered to retrieve hate content, which is then converted to counterspeech by LLMs and annotated by humans.

Strengths:

1. Useful dataset in counterspeech detection.
2. The paper tackles an important real-world problem.

Weaknesses:

1. No baseline is trained on the dataset.
2. Inter-annotated agreement is not provided.
3. Examples of the counterspeech generated are not analyzed.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/8U91C9