Relieving Bias in Hate Speech Detection with a Small Number of Expert Annotations: A Prompt-based Learning Approach

Hate speech is a major problem on social media platforms. Automatic hate speech detection methods relying on machine learning models, which learn from manually labeled datasets, have been proposed in both academia and industry. However, there is increasing evidence that hate speech detection datasets labeled by general annotators (e.g., amateurs or MTurk workers) contain systematic bias, as they cannot effectively consider language use differences among different speakers. When such biased datasets are used to train machine learning models, the resulting models will also be biased. Unlike general annotators, experts can produce much less biased annotations. However, expert annotations cannot be efficiently obtained in large quantity. This paper bridges the gap by adopting a few-shot learning method for hate speech detection using a small number of expert annotations. We propose a novel design that uses contrastive learning and prompt-based learning based on large language models, incorporating a group estimator, a pair generator, and knowledge injection. Using real-world Twitter posts written by African American English speakers and other racial groups as an example, extensive experiments were conducted to demonstrate the superior performance of the proposed method. The proposed approach was also evaluated on data in the LGBTQ+ community and achieved consistent results. The study has important academic and practical implications for hate speech detection and large language models.