Building a Benchmark Dataset for Robust German Hate Speech Detection

Summary

Authors:
Gerold Schneider, Janis Goldzycher
Department of Computational Linguistics
University of Zurich

16 April 2024

Hate speech detection models are only as good as the data they are trained on. Datasets sourced from social media suffer from systematic gaps and biases, leading to unreliable models with simplistic decision boundaries. Adversarial datasets, collected by exploiting model weaknesses, promise to fix this problem. However, adversarial data collection can be slow and costly, and individual annotators have limited creativity. In this project, we introduce GAHD, a new German Adversarial Hate speech Dataset comprising ca. 11k examples. During data collection, we explore new strategies for supporting annotators, to create more diverse adversarial examples more efficiently and provide a manual analysis of annotator disagreements for each strategy. Our experiments show that the resulting dataset is challenging even for state-of-the-art hate speech detection models, and that training on GAHD clearly improves model robustness. Further, we find that mixing multiple support strategies is most advantageous. We make GAHD publicly available at https://github.com/jagol/gahd.