Erstellung eines Benchmark-Datensatzes für eine robuste deutsche Hassredeerkennung

Zusammenfassung

Autoren:
Gerold Schneider, Janis Goldzycher Universität Zürich

16. April 2024

Modelle zur Erkennung von Hassrede sind nur so gut wie die Daten, auf denen sie trainiert werden. Datensätze aus sozialen Medien weisen systematische Lücken und Verzerrungen auf, was zu unzuverlässigen Modellen mit vereinfachten Entscheidungsgrenzen führt. Adversarial-Datensätze, die durch Ausnutzung von Modellschwächen gesammelt werden, versprechen, dieses Problem zu beheben. Die Sammlung von Adversarial-Daten kann jedoch langsam und kostspielig sein, und die Kreativität einzelner Annotatoren ist begrenzt. In diesem Projekt stellen wir GAHD vor, einen neuen deutschen Adversarial-Hate-Speech-Datensatz mit ca. 11.000 Beispielen. Während der Datenerhebung erforschen wir neue Strategien zur Unterstützung von Annotatoren, um effizienter vielfältigere gegnerische Beispiele zu erstellen und eine manuelle Analyse der Meinungsverschiedenheiten der Annotatoren für jede Strategie bereitzustellen. Unsere Experimente zeigen, dass der resultierende Datensatz selbst für hochmoderne Modelle zur Erkennung von Hassreden eine Herausforderung darstellt und dass das Training mit GAHD die Robustheit des Modells deutlich verbessert. Darüber hinaus stellen wir fest, dass die Kombination mehrerer Unterstützungsstrategien am vorteilhaftesten ist. Wir stellen GAHD öffentlich zur Verfügung unter https://github.com/jagol/gahd.