Projekt DeTox: Detektion von Toxizität und Aggressionen in Internet-Beiträgen und -Kommentaren

Jonas Pitz
Hochschule Darmstadt
jonas.pitz@h-da.de

Die Forschungsgruppe DeTox[1] ist eine Kooperation der Hochschule Darmstadt, des Fraunhofer Instituts für Sichere Informationstechnologie SIT und der Abteilung Cyber- und IT-Sicherheit des Hessen Cyber Competence Centers (H3C)[2]. Das Ziel ist die Entwicklung einer neuartigen und nachhaltigen Strategie für eine automatisierte Detektion von Toxizität und Aggressionen in Beiträgen und Kommentaren im Netz.

Mithilfe eines anonymisierten Datensatzes, den wir von der Meldestelle „Hass gegen Hetze"[3] des H3C zur Verfügung gestellt bekommen sowie einem eigens kreierten Twitter-Datensatz, arbeiten wir an State-of-the-Art Modellen zur Klassifikation von Hasskommentaren.

Dafür haben wir ein feingliedriges Annotationsschema und ein eigenes Annotationstool entwickelt, anhand derer die Kommentare in vielen verschiedenen Kategorien von Annotatoren bewertet werden. Dazu gehören unter anderem die Einstufung nach Hate Speech, strafrechtlicher Relevanz, Sentiment, Extremismus, Gefahr und Toxizität.

Mit diesen Daten trainieren wir verschiedene ML-Modelle um eine möglichst breite und akkurate Einstufung neuer Kommentare zu erzielen. Ziel ist es, Hilfestellungen für die Mitarbeiter des H3C zu entwickeln.

Zu unseren Tätigkeiten zählt auch die Teilnahme an Shared Tasks und Konferenzen wie zum Beispiel der GermEval 2021 Subtask zur Toxic Comment Classification (Schütz et al 2021).

References: Schütz, M., Demus, C., Pitz, J., Probol, N., Siegel, M., & Labudde, D. (2021). DeTox at GermEval 2021: Toxic Comment Classification. *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, 54–61. https://aclanthology.org/2021.germeval-1.8

[1] https://projects.fzai.h-da.de/detox/, zuletzt aufgerufen am 10.12.2021
[2] https://innen.hessen.de/Sicherheit/Cyber-und-IT-Sicherheit/Cybersicherheit, zuletzt aufgerufen am 10.12.2021
[3] https://projects.fzai.h-da.de/detox/, zuletzt aufgerufen am 10.12.2021