merz medien + erziehung
zeitschrift für medienpädagogik

merz | medien + erziehung | Arnulfstraße 205 | 80634 München
| fon 089.68989120 | merz@jff.de | www.merz-zeitschrift.de

Lara Moritz: Challenges for Toxic Comment Classification

Beitrag aus Heft »2019/03 Digitalität. Religion. Pluralismus«

Aus dem Forschungsprojekt NOHATE zum Thema Hate Speech liegen erste Ergebnisse im Bereich der Klassifizierung von Hasskommentaren im Internet durch intelligente Filtersysteme (Classifier) vor. Zwischenfazit der Untersuchungen ist, dass ein Erkennen von Hasskommentaren nur durch eine Kombination von Classifiern gewährleistet ist, die unter anderem beleidigende Sequenzen in langen Kommentaren, falsch geschriebenen Wörtern sowie Redewendungen erkennen sollen.

Das Ziel war, Hasskommunikation in Sozialen Medien, Online- Foren und Kommentarbereichen auf seine (Früh-)Erkennbarkeit, Ursachen und Dynamiken sowie auf potenzielle Deeskalationsmöglichkeiten zu untersuchen.

Es zeigte sich, dass Kommentare vorwiegend fälschlicherweise als unbedenklich eingestuft wurden, wenn beleidigende Kommentare ohne Schimpfwörter verfasst wurden (50 %) und der Kontext der Kommentare unberücksichtigt blieb. Weitere Fehlerquellen liegen in Fehlinterpretationen von Aussagen mit orthografischen Mängeln oder Slang-Wörtern. Somit erfolgt die Klassifizierung dieser Kommentare als Hate Speech bei 30 Prozent unbegründet. Weiterhin werden in rhetorischen Fragen häufig (20 %) keine Beleidigungen erkannt (21 %).

Umgekehrt kategorisieren Classifiers Kommentare als Hate Speech, obwohl ein Schimpfwort lediglich im Kontext eines nicht-toxischen Kommentars oder sogar einer Entschuldigung verwendet wurde (60 %). Zudem können intelligente Filtersysteme Zitate und Referenzen nur begrenzt korrekt einordnen, was in 17 Prozent der Fälle die falsche Eingruppierung eines Kommentares erklärte.

Die Ergebnisse des NOHATE-Projekts beruhen auf einer Analyse von 200 Kommentaren, die von Filtersystemen falsch kategorisiert wurden. Zuvor wurden verschiedene Ansätze von Filtersystemen unter anderem für die Erkennung von Beleidigungen in falsch geschriebenen oder abgekürzten Wörtern sowie in komplexen Kontextinformationen getestet.

Das noch bis 2020 angelegte Forschungsprojekt NOHATE der Freien Universität Berlin, der Beuth Hochschule für Technik Berlin und VICO Research & Consulting GmbH wird im Rahmen der Fördermaßnahme „Zusammenhalt stärken in Zeiten von Krisen und Umbrüchen“ vom Bundesministerium für Bildung und Forschung gefördert.

www.das-nettz.de

kopaed verlagsgmbh| Arnulfstraße 205 | 80634 München
| fon 089.688 900 98 | info@kopaed.de | www.kopaed.de