Fact-Sheet zu unserer Methodik im Bericht zur digitalen Gewalt im Wahlkampf

Die Hate Aid Datenanalysen sind Teil eines Projekts, das zum Ziel hat, Hasskampagnen - insbesondere gegen Politiker*innen während der Bundestagswahl - frühzeitig zu erkennen. Die Analysen erheben dabei nicht den Anspruch einer wissenschaftlichen Studie.

Zur Automatisierten Auswertung von Hassrede wurde eine künstliche Intelligenz von Google namens „Perspective API“ eingesetzt. Bei der Perspective API wurden die Scores „Insult“ und „Threat“ verwendet, die beide erfahrungsgemäß sinnvolle Ergebnisse für diese Form der Auswertung liefern. Die Scores „Identity Attack“ und „Toxicity“ kamen an dieser Stelle nicht zum Einsatz. Dabei wurde ein Score von 0.85 als „Cutoff“ für Hass gewählt (Google empfiehlt 0.7).

Zusätzlich zu der KI kam außerdem eine manuelle Überprüfung zum Einsatz, um übermäßiges Auftreten von Falsch-Positiven oder falschen Zuordnungen zu Politiker*innen zu vermeiden. Diese ergab keine größeren Unterschiede zwischen den einzelnen Politiker*innen, die das Ergebnis hätten verändern können. Eine Ausnahme waren Tweets, die Alice Weidel erwähnten: Wie im Report angegeben, ergab deren manuelle Überprüfung, dass ein signifikanter Teil der Kommentare Alice Weidel zwar erwähnt, aber dabei andere Personen angreift.

Zusätzlich kam eine Wörterliste mit beleidigenden und extremistischen Ausdrücken zum Einsatz. Die Liste besteht aus einem begrenzten Wortschatz und ordnet dementsprechend weniger Kommentare als Hassrede ein als die KI. Dennoch waren die Proportionen bezüglich der einzelnen Spitzenkandidat*innen sehr ähnlich wie bei den KI-basierten Analysen, was erneut die Ergebnisse absichert. Zuordnungen von Beleidigungen zu einzelnen Spitzenkandidat*innen wurden ebenfalls nochmals manuell überprüft, und gegebenenfalls gestrichen, falls es sich nicht um Beleidigungen handelte. Ein Beispiel: Die Liste enthielt das Wort „Holocaust“, welches oft in Bezug auf den CDU-Kanzlerkandidten Armin Laschet auftauchte. Der Grund dafür waren allerdings keine Drohungen gegen Laschet, sondern Kritik an einem CDU-Werbespot im Holocaust-Mahnmal in Berlin. Daher wurde das Wort in dieser Konstellation im Report nicht als beleidigende oder verletzende Sprache aufgenommen.

Eine rechtliche Einordnung wurde manuell von unserer Rechtsabteilung vorgenommen. Die KI diente hier lediglich der Vorfilterung. Die Ergebnisse einer Stichprobe bestätigten, dass die KI bei allen Kandidat*innen anteilsmäßig gleich viele potentiell illegale Kommentare erkannt hat, was das Instrument erneut validiert.

Der Fokus auf Twitter ist der vorteilhaften Schnittstelle der Plattform zu verdanken. Andere Netzwerke können wir immer nur ausschnitthaft analysieren, während auf Twitter der gesamte Diskurs zu eine*r Kandidat*in ausgewertet werden kann.

Von verschiedenen rechtsextremen Akteuren ist bekannt, dass sie Hass und Einschüchterung strategisch und koordiniert zum Erreichen politischer Ziele und Kontrolle über den Diskurs einsetzen. Daher beobachten wir in unseren Analysen dezidiert eine rechte bis rechtsextreme Blase, um Hasskampagnen frühzeitig zu erkennen.