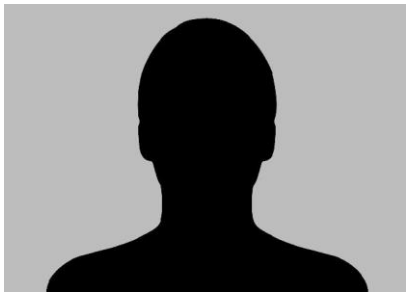
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

ICP Institute of Computational Physics

Entwicklung einer Softwarekomponente zur automatischen Erkennung rassistischer oder beleidigender Userkommentare



Diplomierende
Simon Breiter
Julien Gong-za Wenger

Dozierende
Pius von Däniken
Markus Roos
Mark Cieliebak

Bild klein 1.

Hatespeech im Internet ist heute ein verbreitetes Phänomen, das einerseits eine psychische Belastung für Betroffene darstellt und andererseits Betreiber von Internetplattformen vor die Herausforderung stellt, Kommentare mit hohem Aufwand zu moderieren. Eine automatische Erkennung solcher Kommentare würde daher einen grossen Mehrwert gegenüber dem heutigen Stand bieten.

Das Ziel dieser Arbeit war, mittels Machine Learning einen solchen Ansatz zu entwickeln. Dazu sollte ein Modell trainiert werden, welches zwischen Hatespeech und neutralen Kommentaren unterscheiden kann. Wir erklären dafür die nötigen theoretischen Grundlagen und zeigen auf, wie die Datensätze bereinigt und aufbereitet wurden und welche Schritte unternommen wurden, um diese mit TF-IDF zu vektorisieren. Für die Klassifikation der Daten wurde anschliessend eine Support-Vector-Machine verwendet.

Zur Einordnung unseres Ansatzes trainierten wir ein Modell auf dem Twitter-Datensatz von HASOC FIRE 2019 und verglichen unsere Resultate mit den Teilnehmenden. Unser Modell erzielte dabei einen Macro F1-Score von 0.57, was etwa dem der teilnehmenden Teams entsprach. Dieses Ergebnis deutet darauf hin, dass auch mit einem vergleichsweise einfachen Ansatz ein ähnlicher Macro F1-Score erzielt werden kann.

Die zweite Frage, die uns beschäftige, war, wie gut ein trainiertes Modell auf einem ihm unbekannten Datensatz einer fremden Quelle funktioniert. Dazu trainierten wir ein Modell mit Twitter- und SRF-Daten und wendeten dieses Modell auf einen Datensatz des Industriepartners an. Dieses Resultat verglichen wir anschliessend mit einem Modell, welches nur auf den Daten des Industriepartners trainiert und getestet wurde. Das Modell, welches mit Twitter-Daten trainiert wurde, erzielte auf dem Datensatz des Industriepartners einen Macro F1-Score von 0.50. Dies ist zwar besser als die Ausgangslage mit einem Macro F1-Score von 0.46, aber betrachtet man den F1-Score klassenweise, so unterlag das Modell auf Klasse Hatespeech dem Basis-Modell. Die Frage kann sich damit noch nicht abschliessend beantworten lassen und es braucht dafür noch mehr Experimente mit einem grösseren Basis-Datensatz.