Analyse von Hatespeech unter Berücksichtigung der strafrechtlichen Relevanz

Melanie Siegel, Jonathan Baum, Julia Brahms, Raphael Jährling, Julia Klassen, Christian Mina, Michael Seib, Kaleab Solomon, Christian Stute
Hochschule Darmstadt
melanie.siegel@h-da.de

In den vergangenen Jahren hat sich das Internet zu einem bedeutenden Kommunikationsmedium entwickelt, das Menschen die Möglichkeit bietet, ihre Meinungen und Ansichten auf vielfältige Weise auszudrücken. Allerdings hat dieser digitale Aspekt auch Nachteile: Die Verbreitung von Hatespeech. Diese Form der Online-Diskriminierung stellt nicht nur eine Bedrohung für das soziale Miteinander dar, sondern kann auch strafrechtliche Konsequenzen nach sich ziehen.

In einem Masterprojekt wurde das Phänomen Hatespeech zunächst unter dem Aspekt der strafrechtlichen Relevanz untersucht. Dadurch, dass es sich bei Hatespeech um ein interdisziplinäres Problem handelt, gibt es verschiedene Definitionen, die nicht vollständig konsistent sind. Die Analyse wurde begleitet von einer ethischen Betrachtung der Problematik. Eine automatische Erkennung von strafrechtlich relevanten Kommentaren kann dabei helfen, diese konsequent zu verfolgen.

Daher haben Studierende maschinelle Lernverfahren entwickelt, um Social-Media-Daten auf Hatespeech mit strafrechtlicher Relevanz zu prüfen. Voraussetzung dafür ist eine umfassende Analyse des Forschungsstands. Für das Projekt wurde der Datensatz des DeTox-Forschungsprojektes von 2022 verwendet (Demus et al 2022). In diesem Datensatz sind Hatespeech-Daten unter anderem nach strafrechtlicher Relevanz in Bezug auf das deutsche Strafrecht annotiert.

Die Studierendengruppe entwickelte Modelle nach dem Supervised-Learning-Ansatz mit Feature-Extraktion, die auf die Daten angewendet und evaluiert wurden. Die sorgfältige Auswahl relevanter Features und die Evaluierung der Modelle waren dabei entscheidend.

Die Methode Support Vector Machine (SVM) erreichte bei der Erkennung von Hatespeech einen F1-Score von 72 %. Für die Erkennung von strafrechtlicher Relevanz erzielte die Methode Stochastic Gradient Descent (SGD) einen F1-Score von 81 %.

References: • Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. Detox: A comprehensive dataset for German offensive language and conversation analysis. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Seattle, Washington, 2022.