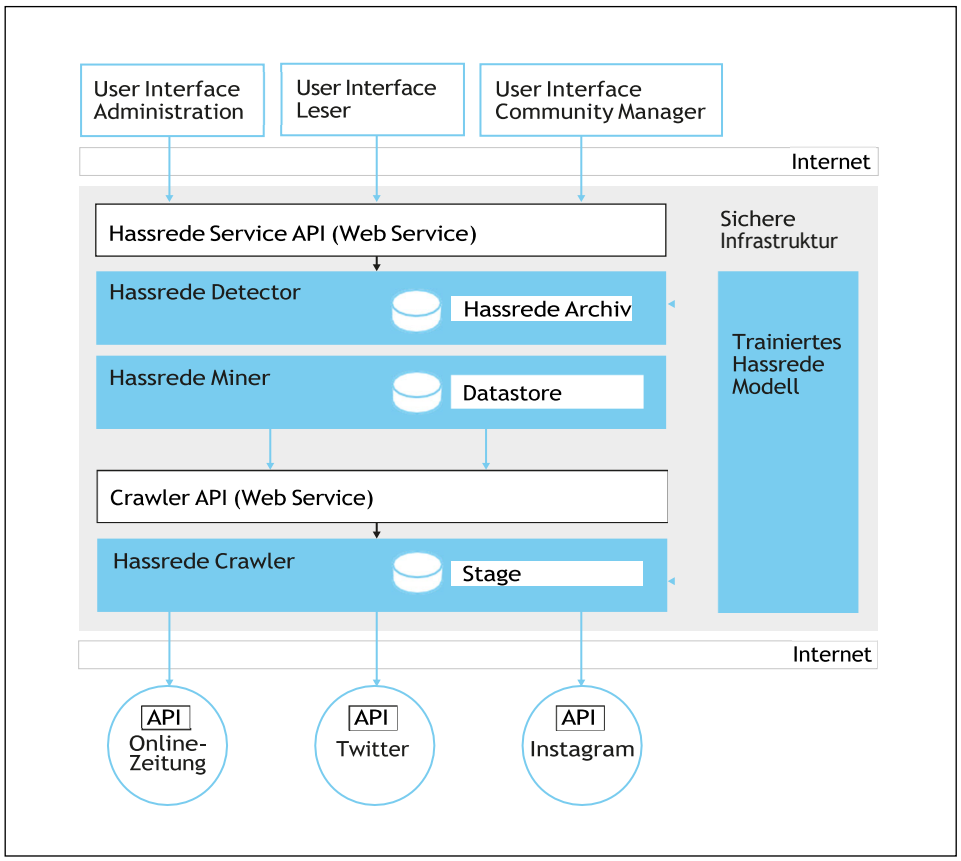
HSDetector

Automatisiertes Erkennungsverfahren für Hassrede in Online-Portalen



Projektname

HSDetector – Entwicklung eines automatisierten Erkennungsverfahrens für Hassrede in Online-Portalen auf der Basis eines nachtrainierbaren Classifiers

Projektleiter

Prof. Dr.-Ing. Peter Mandl

Forschungsbereich

Automatisierte Erkennung von Hassrede über Machine Learning und Natural Language Processing.

Forschungsthema

Kontextsensitive Erkennung von Hassrede auf deutschsprachigen Webseiten.

Kurzbeschreibung des Themas für den Forschungsmaster

Im Forschungsprojekt HSDetector wird ein internet-basierter Erkennungsdienst entwickelt, der Hassrede in Leserkommentaren von deutschen Online-Zeitungen, aber auch in Zukunft generell für beliebige Webseiten, mithilfe eines lernfähigen, nachtrainierbaren Erkennungsverfahrens mit hoher Wahrscheinlichkeit erkennt. Das konkrete Forschungsthema für das Masterstudium soll bei der Klassifizierung von Hassrede zusätzlich zu den Texten weitere Meta- bzw. Kontextinformationen in der Erkennung von Hassrede berücksichtigen.

Hintergrund

Hassrede (Hate Speech) ist inzwischen ein verbreite- tes, im gesellschaftlichen Fokus stehendes Phänomen, das zunehmend das Sicherheitsgefühl insbesondere von Personen des öffentlichen Lebens nachhaltig nega- tiv beeinflusst. Online-Content mit Hassrede wie Volks- verhetzungen, Beleidigungen, Verleumdungen und Morddrohungen stachelt Menschen auf und dient im Einzelfall sogar als Motivation oder Auslöser von poli- tisch motivierten Straftaten bis hin zu Anschlägen auf Personen oder Sachen. Hassrede ist stark kontextab- hängig und kann sich im Laufe der Zeit verändern.

Ziele

Auch Online-Zeitungen müssen Vorsorge treffen, um Hassrede in ihrem Online-Content, vor allem in Leser-Kommentaren, zu vermeiden. Ziel des Forschungsvor- habens HSDetector ist es, einen neuen Dienst zu entwi- ckeln, der Hassrede in Online-Zeitungen mit hoher Wahrscheinlichkeit entdeckt. Das neue Verfahren fin- det Hassrede in Webseiten auf der Basis eines lern- fähigen Erkennungsverfahrens mithilfe eines nachtrai- nierbaren Classifiers, der auch Veränderungen in der Ausdrucksweise lernen kann.

Ausblick

Es wird angestrebt, die Ergebnisse aus dem Projekt zu einem webbasierten HatesSpeech Detection Service weiterzuentwickeln, der in Redaktions- systeme von Online-Zeitungen aber auch in andere Content-Management-Systeme eingebunden werden kann.

Eingesetzte Methoden, Technologien und Werkzeuge

CNN, Support Vector Machine, logistische Regression, Random Forest, Bag-of-words, TF-IDF und Google BERT, elasticsearch, Python, Scrapy

Publikationen

Januzaj, E., Weber, M., Keller, M.-E., Auch, M., Mandl, P., CoSim: An Approach to Calculate Complex Object Similarity. 23rd International Conference on Information Integration and Web Intelligence, 2021