Zusammenfassung | Call 19 | Projekt ID 7207

1. Projektziel

Wir sind ein Team von ForscherInnen an der Hochschule für Angewandte Wissenschaften St. Pölten und haben uns im Projekt HaSPI, Hate Speech Prevention Through Imitation, mit der Erkennung von Hate Speech in deutschsprachigen Online-Foren beschäftigt.
Dabei haben wir einen neuartigen Ansatz entwickelt und prototypisch implementiert: Anstatt über statisch gelabelte Datensätze Kommentare in Hate-Speech oder Nicht-Hate-Speech einzuordnen, versucht unser System menschliches Verhalten durch Nachahmung zu verstehen. Ein wesentlicher Vorteil dieses Ansatzes ist seine Nachvollziehbarkeit: Das Modell liefert nicht nur eine Entscheidung, sondern zugleich Informationen darüber, wie diese zustande kommt. Dadurch muss nicht erst im Nachhinein rekonstruiert werden, warum beispielsweise ein Beitrag als problematisch eingestuft oder entfernt wurde.
Evaluiert haben wir den Ansatz anhand des “One-Million-Posts”-Korpus der österreichischen Tageszeitung DER STANDARD evaluiert, der eine große Sammlung deutschsprachiger Online-Kommentare einschließlich menschlicher Moderationsentscheidung bereitstellt.
HaSPI richtet sich an ForscherInnen, die sich mit Hate-Speech-Erkennung, automatisierter Moderation und erklärbaren KI-Systemen beschäftigen, sowie an deutschsprachige Moderationsteams, die die Qualität von Online-Diskussionen verbessern und einen verantwortungsvollen gesellschaftlichen Diskurs fördern möchten.

Github Repository des Projekts: https://github.com/fhstp/HaSPI

2. Projektendergebnisse

1	Projektzwischenbericht	CC BY-SA 4.0	www.netidee.at/haspi
2	Projektendbericht	CC BY-SA 4.0	www.netidee.at/haspi
3	Entwickler:innen-Dokumentation	CC BY-SA 4.0	haspi.readthedocs.io
4	Anwender:innen-Dokumentation	CC BY-SA 4.0	https://github.com/fhstp/HaSPI/blob/main/README.md
5	Veröffentlichungsfähiger Einseiter / Zusammenfassung	CC BY-SA 4.0	www.netidee.at/haspi
6	Dokumentation Externkommunikation (als Teil des Endberichts)	CC BY-SA 4.0	www.netidee.at/haspi
7	Prototypische Softwareimplementierung: Datenaufbereitungs-Pipeline für den One-Million-Posts-Korpus	Apache-2.0	https://github.com/fhstp/haspi
8	Paket für Endanwender:innen API auf PyPi: * Interface für Hate Speech Detection * Interface für Erklärungen der Entscheidungen	Apache-2.0	https://pypi.org/project/haspi/

9	Code für Endanwender:innen API: HaSPI-Framework: Imitation-Learning-Pipeline für Hate-Speech-Erkennung (JAX-Implementierungen von IQ-Learn und Varianten; RL-Environment für Textklassifikation)	Apache-2.0	https://github.com/fhstp/haspi
10	Wissenschaftliche Publikation: „Context-Aware Content Moderation for German Newspaper Comments“ (iDSC 2025)	CC BY-SA 4.0	arxiv.org/abs/2505.20963

3. Geplante weiterführende Aktivitäten nach netidee-Projektende

Die Projektergebnisse werden nach Projektende vielfältig weiter genutzt. Zunächst befindet sich die Publikation zur Imitation-Learning-basierten Klassifikation in Arbeit, und soll bei einer einschlägigen Konferenz oder Fachzeitschrift veröffentlicht werden. Die im Projekt entwickelten Methoden sollen auch an der USTP in den Unterricht einfließen und in Bachelor- und Masterarbeiten weiter untersucht werden. Darüber hinaus haben wir im Projekt viele neue Richtungen zur Weiterentwicklung gefunden, darunter die etwa die Anwendung verschiedenster Methoden aus dem Bereich von Imitation Learning oder die Adaption der verwendeten Methoden auf die Diversität menschlichen Verhaltens. Diese werden in weiterführenden Projektanträgen aufgegriffen.

4. Anregungen für Weiterentwicklungen durch Dritte

Durch die Veröffentlichung unseres Frameworks ergeben sich für Dritte weitreichende Anknüpfungspunkte. Konkret würden wir empfehlen, die Generalisierbarkeit auf weitere deutschsprachige Plattformen und Korpora zu testen. Weiters wären Studien aus dem Bereich der sozialen Wissenschaften hilfreich, die den Nutzen der untersuchten Erklärbarkeitsmethode für Moderationsteams und deren Integration in bestehende Moderationsworkflows (z.B. als Vorfilter mit menschlicher Letztentscheidung) testen.
Auch die Weiterentwicklung in Richtung anderer Sequenzklassifikationsmethoden wie die Verbindung unserer Komponenten mit xLSTMs würde die Methode auch für Unternehmen ohne Zugriff auf leistungsfähige Hardware effizienter machen.