netidee

PROJEKTE

HaSPI - Hate Speech Prevention Through Imitation

Zwischenbericht | Call 19 | Projekt ID 7207

Inhalt

1 Einleitung

Das Projekt HaSPI – Hate Speech Prevention Through Imitation (Projekt-ID: 7207, Call #19, Förderjahr 2024) widmet sich der Bekämpfung von Hate Speech im deutschsprachigen Internet. Trotz zahlreicher technischer Ansätze bleibt Hate Speech ein drängendes Problem, insbesondere da viele bestehende Lösungen entweder plattformunabhängig oder auf englischsprachige Inhalte beschränkt sind.
Mit HaSPI wollen wir einen innovativen Ansatz validieren und prototypisch implementieren, der auf Imitation Learning basiert. Ziel ist es, menschliche Moderationsentscheidungen nachzuahmen und dadurch die automatisierte Erkennung und Moderation von Hassrede zu verbessern.
Als Datenbasis dient der „One Million Posts“-Korpus der Tageszeitung DER STANDARD, der eine reichhaltige Sammlung deutschsprachiger Online-Kommentare bietet. Durch die Analyse von Nutzerverhalten und thematischem Kontext entwickeln wir in HaSPI Modelle, die Hate Speech nicht nur erkennen, sondern auch nachvollziehbar begründen können.
Das übergeordnete Ziel von HaSPI ist die Entwicklung eines Open-Source-Frameworks, das speziell auf die Moderation deutschsprachiger Inhalte ausgerichtet ist. Neben der Validierung des Imitation-Learning-Ansatzes steht die Erklärbarkeit der Entscheidungen im Fokus, um Vertrauen und Transparenz in automatisierte Moderationssysteme zu fördern.

2 Status der Arbeitspakete

2.1 Arbeitspaket 1 - AP1: Detailplanung und Formales am Projektstart

Arbeitspaket 1 wurde bereits mit April 2025 erfolgreich abgeschlossen. Die folgenden Arbeiten wurden in diesem initialen Arbeitspaket durchgeführt:

- der Vertrag wurde unterschrieben,
- der Detailprojektplan wurde erstellt und abgenommen,
- eine detaillierte Liste der geplanten Projektergebnisse mit Lizenz und Ort der öffentlichen Bereitstellung wurde erstellt und abgenommen,
- die Projekt-Website ging in Betrieb und ein erster Blogeintrag wurde erstellt und die erste Förderrate wurde beantragt.

2.2 Arbeitspaket 2 - Projektmanagement, Outreach und Dissemination

Dieses Arbeitspaket umfasst die organisatorische Koordination des Projekts sowie die externe Kommunikation und wissenschaftliche Dissemination. Zu den Tätigkeiten zählen:

- Erstellung von Blogposts zum Projektfortschritt
- Zwischenbericht (M6–M9)
- Endbericht (M12)

- Laufende wissenschaftliche Publikationen

Im Berichtszeitraum konnten mehrere Disseminationsmaßnahmen erfolgreich umgesetzt werden:

- **Wissenschaftliche Publikation**:
  Das Paper *"Context-Aware Content Moderation for German Newspaper Comments"* wurde veröffentlicht und auf der **iDSC-Konferenz 2025** an der FH Salzburg von Felix Krejca präsentiert. Die Arbeit untersucht automatisierte Moderationsverfahren für deutschsprachige Kommentarbereiche unter Berücksichtigung kontextueller Informationen wie Nutzerhistorie und Artikelthemen.
  Preprint verfügbar unter: https://arxiv.org/abs/2505.20963
- **Präsentation am Forschungsforum 2025**:
  Das Projekt wurde im Rahmen des Forschungsforums 2025 vorgestellt. Das Forschungsforum der österreichischen Fachhochschulen fand am 7. und 8. Mai 2025 an der FH Campus Wien. Weitere Infos und Link zur Präsentation:
  https://www.netidee.at/haspi/haspi-forschungsforum-2025
- **Blogposts zur Projektkommunikation**:
  Zur Sichtbarmachung des Projekts wurden Blogbeiträge veröffentlicht, darunter:
  - *"Hate doesn't only speak English"*: Thematisiert die Herausforderungen bei der Erkennung von Hassrede in deutschsprachigen Online-Kommentaren.
    Link: https://www.netidee.at/haspi/hate-doesnt-only-speak-english
- **Zwischenbericht**:
  Der Zwischenbericht wurde fristgerecht erstellt und dokumentiert den Fortschritt in allen Arbeitspaketen.

Es gab keine wesentlichen Abweichungen vom ursprünglichen Arbeitsplan. Alle geplanten Disseminationsmaßnahmen konnten wie vorgesehen umgesetzt werden.

2.3 Arbeitspaket 3 - Datenaufbereitung und RL-Umgebung

Zu Beginn des Arbeitspakets erfolgte eine systematische Literatursuche, um aktuelle Forschungsarbeiten im Bereich *Imitation Learning* zu identifizieren. Hierfür kamen die Datenbanken Google Scholar, IEEE Xplore, ACM Digital Library und Springer Link zum Einsatz. Berücksichtigt wurden Publikationen ab dem Jahr 2024, die jeweils 50 relevantesten Treffer wurden in Zotero importiert. Nach der Dublettenbereinigung verblieben 149 Einträge, von denen nach Sichtung von Titeln und Abstracts 48 Arbeiten für eine vertiefte Analyse ausgewählt wurden. Ergänzend wurde das IQ-Learn-Paper, auf dessen Konzept das Projekt aufbaut, in die Analyse aufgenommen. Über Google Scholar und Semantic Scholar wurden Publikationen identifiziert, die dieses Paper zitieren, und hinsichtlich ihrer Relevanz geprüft. Auf diese Weise kamen 59 weitere Arbeiten hinzu. Nach Dublettenbereinigung ergab sich eine finale Literaturliste mit 91 Publikationen. Zusätzlich wurden fünf weitere Arbeiten aus den Introduction- und Related Work-Abschnitten der gesichteten Publikationen ergänzt.

Diese Literaturrecherche bildete die Grundlage für die konzeptionelle und technische Weiterentwicklung des Projekts. Einzelne Ansätze aus den identifizierten Arbeiten wurden gesondert getestet und ein prototypisches Reinforcement-Learning-Environment zur automatisierten Textgenerierung entwickelt. Dieses diente vor allem dazu, Anforderungen an die Modellarchitektur und Trainingsumgebung besser einschätzen zu können. Aufbauend darauf wurde der IQ-Learn-Algorithmus zunächst in PyTorch implementiert, um das Environment zu testen. Dabei wurden Speicherrestriktionen im sogenannten Replay-Buffer identifiziert, die die Verwendung mit den Daten im Vollumfang erschweren würden. Um diese Einschränkungen zu adressieren, wurde der Soft Actor-Critic (SAC)-Algorithmus in JAX neu implementiert und mit der PyTorch-Version verglichen.

Die Experimente zeigten einen rund 40-fachen Geschwindigkeitszuwachs in typischen Benchmark-Umgebungen bei der Verwendung von JAX. Aufgrund dieser deutlichen Effizienzsteigerung erfolgte eine vollständige Re-Implementierung sowohl des Text-Environments als auch des IQ-Learn-Algorithmus in JAX. Diese Umstellung brachte in ersten Tests unter Realbedingungen eine Beschleunigung um den Faktor 16 (ca. 75 Trainingsschritte pro Sekunde vs. ca. 1250 Trainingsschritte pro Sekunde auf einer Nvidia RTX 3090) sowie eine verbesserte Kontrolle über den Speicherverbrauch, wodurch JAX als bevorzugtes ML-Framework etabliert wurde.

Parallel dazu wurde eine Datenaufbereitungs-Pipeline für den One-Million-Posts Korpus, auf den unser Klassifikationsansatz trainiert werden soll, erstellt. In dieser werden die Userkommentare bereinigt und in Trainings- und Testdatensätze unterteilt. Anschließend werden vortrainierte Open-Source-Transformer-Modelle (z.B. GBERT-Large) eingesetzt, um Vektorrepräsentationen der Kommentare zu generieren. Zur effizienten Weiterverarbeitung im Imitation Learning wird zudem eine Dimensionsreduktion dieser Vektoren durchgeführt.

Als Vorbereitung auf den Einsatz rekurrenter Architekturen (LSTM/xLSTM) wurde zudem die sogenannte Lambda-Discrepancy für das zugrundeliegende Lernproblem implementiert. Erste Ergebnisse blieben vorerst hinter den Erwartungen zurück, jedoch wurden kritische Implementierungsaspekte identifiziert, die im Rahmen von Arbeitspaket 4 aufgegriffen werden können.

Insgesamt konnte durch die bisherigen Arbeiten ein umfangreicher und kuratierter Literaturkorpus aufgebaut, ein leistungsfähiges RL-Environment entwickelt und die Grundlage für die Umsetzung komplexerer Agenten geschaffen werden. Besonders hervorzuheben ist die signifikante Verbesserung der Effizienz durch den Wechsel von PyTorch zu JAX. Kleinere Probleme wie die etwas enttäuschenden ersten Ergebnisse bei der Lambda-Discrepancy, beeinträchtigen den Projektfortschritt nicht wesentlich.

Größere Abweichungen vom ursprünglichen Projektplan traten nicht auf. Die Umstellung auf JAX war zwar nicht vorgesehen, stellte sich jedoch als strategisch sinnvolle Anpassung heraus. Sie trägt wesentlich dazu bei, die geplanten Forschungsziele effizienter zu erreichen und die kommenden Arbeitspakete – insbesondere zur Entwicklung rekurrenter Agenten – auf einer soliden technischen Basis aufzubauen.

2.4 Arbeitspaket 4 - Implementierung und Evaluierung

Im Rahmen von Arbeitspaket 4 wurde auf den in der vorangegangenen Phase gewonnenen Erkenntnissen aufbauend mit der vollständigen Implementierung der Imitation-Learning-Pipeline begonnen. Dabei wird der Fokus unter anderem auf eine effiziente Nutzung der neuen internen Server-Infrastruktur gelegt, die auch Finetuning und Training von LLMs zu diesem Zweck ermöglicht. Parallel dazu wurde mit der Implementierung einer xLSTM-Architektur in JAX begonnen, um die Leistungsfähigkeit rekurrenter Modelle in komplexen Imitation-Learning-Szenarien als Alternative zu Transformern zu evaluieren. Zudem wurde die theoretische Grundlage zur Anwendung von Imitation Learning für Klassifikationsaufgaben ausgearbeitet, als Vorbereitung auf die geplante wissenschaftliche Publikation der Projektergebnisse.
Aufgrund der kurzen Laufzeit des Arbeitspakets gab es noch keine größeren Erfolge oder Probleme, auch Abweichungen vom Projektplan gibt es keine.

2.5 Arbeitspaket 5 - Dokumentation und Formales am Projektende

(Dieses Arbeitspaket wurde noch nicht begonnen.)

3 Umsetzung Förderauflagen

(In der Fördervereinbarung sind keine Förderauflagen vorgesehen.)

4 Zusammenfassung Planaktualisierung

(Wir hatten bislang noch keinen Bedarf für Anpassungen im Projektplan.)

5 Öffentlichkeitsarbeit/ Vernetzung

Im Rahmen der Öffentlichkeitsarbeit wurden mehrere Maßnahmen zur Sichtbarmachung des Projekts HaSPI umgesetzt. Dazu zählen:

- **Wissenschaftliche Dissemination**:
  Präsentation des Papers *"Context-Aware Content Moderation for German Newspaper Comments"* auf der iDSC-Konferenz 2025 an der FH Salzburg.
  Preprint: https://arxiv.org/abs/2505.20963
- **Projektpräsentation am Forschungsforum 2025**:
  Vorstellung des Projekts und seiner Ziele auf einer nationalen Forschungsplattform.
  Beitrag: https://www.netidee.at/haspi/haspi-forschungsforum-2025
- **Blogposts zur Projektkommunikation**:
  Veröffentlichung von Beiträgen auf der Netidee-Plattform, z. B. *"Hate doesn't only speak English"*.
  Blog: https://www.netidee.at/haspi/hate-doesnt-only-speak-english

- **Social Media Aktivitäten**:
  Das Projekt sowie zentrale Ergebnisse werden aktiv über LinkedIn und Instagram beworben. Zusätzlich erfolgte eine Verbreitung über die offiziellen Kanäle der FH St. Pölten, um Reichweite und Vernetzung zu erhöhen.

Für die weitere Projektzeit sind weitere Beiträge auf Social Media sowie weitere wissenschaftliche Publikationen geplant.

6 Eigene Projektwebsite

Derzeit wird keine eigene öffentliche Projektwebsite zusätzlich zur Netidee-Projektseite betrieben. Die Software/Prototype-Entwicklung erfolgt über ein privates GitHub-Repository, das zur internen Koordination und Versionsverwaltung genutzt wird.
Sobald ein erster Release und ein entsprechender Prototyp verfügbar sind, ist die Veröffentlichung einer eigenen Website geplant – voraussichtlich direkt über GitHub Pages.