CLAIM

Hybrid AI Models for Claim Detection and Verification

Programm / Ausschreibung	DST 24/26, DST 24/26, AI Ökosysteme 2024: AI for Tech & AI for Green	**Status**	laufend
Projektstart	01.05.2025	**Projektende**	30.04.2028
Zeitraum	2025 - 2028	**Projektlaufzeit**	36 Monate
Keywords	Fine-tuned Language Models for Claim Verification; Knowledge Graph Construction and Evolution; Visual Exploration of AI Model Outputs; Science Journalism; Mental Health Stigmatization		

Projektbeschreibung

Das CLAIM-Projekt adressiert wachsende Bedenken bezüglich falscher und irreführender Behauptungen in globalen Netzwerken. CLAIM ist ein menschzentrierter Ansatz zur Entwicklung hybrider KI-Modelle, um dieses Problem zu bekämpfen. Dabei werden Endbenutzer in zwei besonders betroffenen Domänen gleich zu Beginn miteinbezogen: Wissenschaftsjournalismus und die Stigmatisierung psychischer Erkrankungen.

Das Projekt wird Arbeitsabläufe zur Überprüfung von Behauptungen optimieren. Interaktive Werkzeuge sollen Journalisten, Forschern, Pädagogen und anderen Autoren dabei helfen, Misinformation und Stigmatisierung entgegenzuwirken. Während große Sprachmodelle (LLMs) beeindruckende Fähigkeiten in der Verarbeitung menschlicher Sprache gezeigt haben, neigen sie immer noch zu sogenannten Halluzinationen, da ihnen ein objektiver Sinn dafür fehlt, was nützlich, real oder wahr ist.

CLAIM wird diese Herausforderung durch die iterative Entwicklung hybrider KI-Modelle adressieren. Eine Kombination symbolischer und subsymbolischer Komponenten, auf Basis vertrauenswürdiger Quellen trainiert, soll die faktische Genauigkeit verbessern und Halluzinationen reduzieren. Wissensgraphen werden den Modellen Alltagswissen und Fachkenntnisse vermitteln, ergänzt durch Retrieval Augmented Generation (RAG) für den Zugriff auf vertrauenswürdige Informationen. Für die subsymbolische Embedding-Schicht wird CLAIM Wissensdestillationstechniken nutzen, um kleine Sprachmodelle für spezifische Aufgaben zu trainieren und damit die notwendige Rechenleistung und den Energieverbrauch erheblich zu reduzieren.

Die KI-Modelle von CLAIM werden Behauptungen erkennen, klassifizieren und verifizieren. Visualisierungen werden es Nutzern ermöglichen, die Ursprünge und Verbreitung falscher Behauptungen zu verfolgen und vertrauenswürdige Quellen zu identifizieren, um diese zu widerlegen. Diese Visualisierungen werden in drei Anwendungen integriert: (i) ein “Claim Explorer” zur Verfolgung von Trends und Zusammenhängen, (ii) ein Editor mit generativen Funktionen basierend auf vertrauenswürdigen Quellen, und (iii) eine Chrome-Extension zur Prüfung von Behauptungen in Echtzeit. Die Methoden zur

Erkennung, Klassifizierung und Verifizierung von Behauptungen werden auch als Datendienste verfügbar sein, um die Integration mit externen Datenräumen und Anwendungen zu erleichtern.

Automatische Klassifikation in IPTC-NewsCodes und Nachhaltige Entwicklungsziele (SDGs) wird es Analysten ermöglichen die ganzheitlichen Auswirkungen irreführender Behauptungen auf verschiedene SDGs einzuschätzen. Die KI-Modelle werden zeigen, wo solche Behauptungen entstehen, wie sie sich verbreiten und mit welchen Strategien man ihnen am besten begegnet. Der menschzentrierte und interdisziplinäre Charakter des Projekts spiegelt sich auch in der Zusammensetzung des Konsortiums wider. webLyzard technology arbeitet seit mehr als 15 Jahren an der Schnittstelle von KI, Nachhaltigkeit und Klimakommunikation und wird die technische Entwicklung des Projekts leiten. In Zusammenarbeit mit dem Deep-Tech-Startup Storypact wird es Autoren generative, auf vertrauenswürdigen Quellen basierende, KI-Funktionen bereitstellen. Zu den Anwendungspartnern zählen DER STANDARD, ein österreichisches Nachrichtenmedium, bekannt für seine fundierte Recherche und Berichterstattung über anspruchsvolle Themen, sowie die Medizinische Universität Graz, die als wissenschaftlicher Partner Stigmen rund um psychische Erkrankungen bekämpft.

Abstract

The CLAIM project addresses growing concerns around false and misleading claims in global information networks. CLAIM is a human-centered effort to develop hybrid AI models to combat this problem, involving end-users in two use cases particularly affected by misinformation (false information spread inadvertently) and disinformation (spread with the intent to mislead): science journalism and stigma around mental health disorders. The project will streamline claim-checking workflows when reporting on complex and often contested issues and provide tools for journalists, mental health researchers, educators and other content creators to counter claims that lead to stigmatization and false beliefs.

While Large Language Models (LLMs) have demonstrated superior capabilities to deal with the subtle nuances of human language, they remain prone to so-called hallucinations as they lack an objective sense of what is useful, real, or true. CLAIM will address this shortcoming by iteratively developing hybrid AI models combining symbolic and subsymbolic components, training on content from trusted sources to improve factual accuracy and reduce hallucinations, evolving knowledge graphs, which in turn provide the models with common sense and domain knowledge, and augmenting via Retrieval Augmented Generation (RAG) to access recent and trusted information. For the subsymbolic embeddings layer, CLAIM will use knowledge distillation techniques to train smaller language models for specific tasks, significantly reducing their computational requirements and energy footprint.

CLAIM’s AI models will be explainable and cross-lingual. They will detect, classify and verify claims. Interactive visualizations will help users track the origins and propagation of false claims and identify trusted sources with verified information that refutes these claims. CLAIM will integrate these visualizations into three Web applications that support all phases of the content production workflow: (i) a visual claim explorer to track evolving claims and their context, (ii) a text editor with generative features guided by trusted evidence sources, and (iii) a browser extension for real-time claim testing. The claim detection, classification and verification methods will be available as data services to facilitate their integration with external data spaces and applications. Through their ability to classify claims by IPTC NewsCodes and Sustainable Development Goals (SDGs), these services will highlight interdependencies and enable analysts to estimate the impact of a misleading claim across the SDGs.

The hybrid AI models developed and evaluated in CLAIM will reveal where false and misleading claims originate, how they spread, and how journalists and content creators can mitigate their impact. The project’s human-centered and interdisciplinary nature is reflected in the composition of its consortium. webLyzard technology has worked at the intersection of AI, sustainability and climate change communication for more than 15 years. It will lead the project’s technical development, collaborating with the deep tech startup Storypact to provide content authors with generative AI features guided by trusted sources. The use case partners include the STANDARD, a major Austrian news medium known for in-depth research and accurate coverage of complex issues, and the Medical University of Graz as the scientific partner leading the work on health literacy and destigmatization strategies.

Projektkoordinator

- webLyzard technology gmbh

Projektpartner

- STANDARD Verlagsgesellschaft m.b.H.
- Storypact GmbH
- Medizinische Universität Graz