

Teil I: Kurzbericht (Murmuras GmbH)

FKZ 16KIS1525

Laufzeit: 01.01.2022 – 30.06.2025

Darstellung des Projektergebnisses

Im Verbundprojekt DeFaktS – *Detecting and Explaining Fake News in Text-based Social Media* – entwickelte Murmuras als Industriepartner die mobile Datenerfassungs- und Demonstrator-Infrastruktur für die Erkennung und Erklärung von Desinformation in Messenger-Diensten und sozialen Netzwerken. Ziel des Gesamtvorhabens war ein datenschutzkonformes End-to-End-System, das reale Kommunikationsdaten erfasst, mit KI-Verfahren analysiert und Ergebnisse für Nutzer*innen nachvollziehbar erklärt.

Murmuras trug dazu bei, indem (1) eine Android-basierte Screen-Analyse-Technologie auf Basis des Accessibility Service entwickelt, (2) eine serverseitige Datenpipeline inklusive Studien- und Annotationsplattform bereitgestellt und (3) ein Technologie-Demonstrator („DEFAKTRON 9000“) umgesetzt wurde, der die vom FZI entwickelten Klassifikations- und XAI-Komponenten in eine nutzbare Smartphone-Anwendung integriert.

Der Demonstrator wurde im Rahmen des DeFaktS-Roundtables im März 2024 vorgestellt, in Workshops erprobt und für Folgeeinsätze vorbereitet. Nutzer*innen konnten in Echtzeit markieren lassen, welche Inhalte auf ihrem Bildschirm als potenzielle Desinformation eingestuft wurden, erklärende Hervorhebungen sehen und über Feedbackformulare Rückmeldungen an die Modelle geben.

Ursprüngliche Aufgabenstellung und wissenschaftlicher Stand

Ausgangspunkt war die Förderbekanntmachung „Forschung Agil – Erkennen und Bekämpfen von digitalen Desinformationskampagnen“ des BMBF. Sie fordert neue Methoden, um Desinformation zu erkennen, ihre Wirkmechanismen zu verstehen und ihr entgegenzuwirken – insbesondere in verschlüsselten, mobilen Kommunikationsumgebungen, die sich klassischen Plattform- und Fact-Checking-Ansätzen entziehen.

Murmuras knüpfte fachlich an frühere Arbeiten zur Smartphone-Datenerhebung („Smartphone Sensing“, Psychoinformatik) an, bei denen App-Nutzung, Kontexte und

Verhaltensmuster kontinuierlich erfasst und wissenschaftlich ausgewertet werden. Das Unternehmen bringt eine erprobte Plattform für mobile Datenerhebung, Studienmanagement und Datenanalyse in Kooperation mit mehreren Universitäten ein.

Während internationale Desinformationsforschung bisher vor allem öffentliche Plattformen wie Twitter im Blick hatte, zielte DeFaktS auf halböffentliche und geschlossene Räume (z. B. Telegram-Kanäle, Messenger-Chats). Hierfür musste zunächst geklärt werden, wie Daten rechtssicher erhoben, pseudonymisiert und in KI-taugliche Formate überführt werden können.

Ablauf des Vorhabens

Zu Projektbeginn wurden gemeinsam mit der Universität Marburg (MAR) und dem FZI Anforderungen an Datenmanagement, Datenschutz und Studienorganisation spezifiziert. Murmuras passte seine Studienplattform an, um die Rekrutierung, Betreuung und Incentivierung realer Teilnehmender sowie die Verwaltung mehrstufiger Studien zu unterstützen.

Im Anschluss entwickelte Murmuras eine Android-App, die über den Accessibility Service Inhalte aus Ziel-Apps (WhatsApp, Telegram, Facebook, Instagram, Twitter) auslesen und für die weitere Verarbeitung strukturieren konnte. Die Daten wurden über eine sichere Backend-Infrastruktur an Server von Murmuras übertragen, dort für ETL-Prozesse vorbereitet und über eine API den Projektpartnern bereitgestellt.

Eine erste Feldstudie mit „In-the-wild“-Datenerhebung zeigte, dass zur Erreichung der vorgesehenen Datenmengen eine deutlich größere Stichprobe und mehr hochaktive Gruppen notwendig gewesen wären und dass sich die Ziel-Apps auf unterschiedlichen Geräten teils inkonsistent verhielten. In Abstimmung mit der MAR wurde daher auf serverseitige Erhebung über Telegram-Bots und Twitter-APIs umgestellt, was die systematische und skalierbare Datensammlung ermöglichte.

Murmuras richtete anschließend eine Doccano-Instanz ein, über die die gesammelten Texte durch Annotator*innen von FZI, MAR und Universität Bonn feingranular gelabelt wurden. Die resultierenden Datensätze flossen in die Entwicklung des DeFaktS-Korpus ein, der u. a. in der Publikation „DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing“ (LREC-COLING 2024) beschrieben wurde.

Parallel entwickelte Murmuras den Technologie-Demonstrator „DEFAKTRON 9000“. Dieser nutzt die Google-Assistant-Schnittstelle und kann per Power-Button aufgerufen werden, um den aktuellen Bildschirminhalt zu erfassen, an die vom FZI bereitgestellte Klassifikations-API zu senden und die dort erzeugten Labels und Erklärungen direkt auf dem Gerät anzuzeigen. Nutzer-Feedback wird über in die App eingebettete Fragebögen zurückgespielt.

In der späteren Projektphase wurde der Demonstrator in Labor- und Kurzstudien mit vorkonfigurierten Geräten evaluiert; Murmuras stellte dazu die App und die Studienlogistik bereit. Die Evaluation zeigte eine hohe Nutzbarkeit und Akzeptanz der App, aber auch den Wunsch nach stärker faktenbasierten Erklärungen – eine wichtige Erkenntnis für zukünftige Weiterentwicklungen und die Kombination mit Large Language Models.

Wesentliche Ergebnisse und Zusammenarbeit

Aus Sicht von Murmuras wurden die geplanten Arbeitspakete vollständig und – nach einem leichten Verzug zu Projektbeginn – im ursprünglichen Kosten- und Zeitrahmen umgesetzt. Im Verbund mit MAR, FZI und Liquid Democracy entstand:

- eine praxiserprobte Screen-Analyse-Infrastruktur für Android-Smartphones,
- eine skalierbare Datenpipeline mit Annotations- und Studienmanagementplattform,
- der erste umfangreiche, feingranular annotierte deutschsprachige Desinformationsdatensatz aus Messenger-Kommunikation,
- sowie ein funktionsfähiger Technologie-Demonstrator, der KI-basierte Desinformationsklassifikation und XAI-Erklärungen direkt in der alltäglichen Smartphone-Nutzung erfahrbar macht.

Die Zusammenarbeit mit den akademischen Partnern war eng und iterativ: Murmuras stellte die technische Infrastruktur und praktische Expertise in mobiler Datenerhebung bereit, während MAR und FZI die wissenschaftliche Modellierung, Annotation und Evaluation verantworteten und Liquid Democracy die Integration in Beteiligungsplattformen vorantrieb.
