Sachbericht zum Verwendungsnachweis

Vorhabenbezeichnung
Verbundprojekt: KI-unterstütztes Assistenzsystem für die Crowdsourcing-basierte Erkennung von über digitale Plattformen verbreiteter Desinformation - noFake

Förderkennzeichen
16KIS1519

Laufzeit des Vorhabens Berichtszeitraum
01.12.2021 -bis- 31.01.2025

Zuwendungsempfänger
Technische Universität Berlin

Antragstellerin: Prof. Dr. Dorothea Kolossa
Technische Universität Berlin
Fakultät IV: Elektrotechnik und Informatik
Sekretariat EN 3
Einsteinufer 17
D-10587 Berlin

Kurzbericht

Aufgabenstellung und Stand der Technik

Im Rahmen des noFake-Projekts wurden KI-Methoden und eine neue Plattform – das *Faktenforum* – entwickelt, um die Zusammenarbeit von Bürger-Journalisten, dem professionellen Faktencheck- und Journalisten Team von CORRECTIV und von KI-Tools bei der Erkennung und Bekämpfung von Desinformation zu ermöglichen. Ausgangspunkt war die Beobachtung, dass sich Desinformation trotz hoher gesellschaftlicher Aufmerksamkeit schnell und vielfältig verbreitet und dabei sowohl demokratische Prozesse als auch die persönliche Sicherheit in der Gesellschaft gefährdet. Da rein technische Lösungen nicht in der Lage sind, falschen Inhalte zuverlässig zu erkennen, setzte das Projekt auf die Kombination der Stärken menschlicher Urteilskraft und maschineller Analyse.

Ziel war es, Methoden und eine Plattform zu schaffen, um große Datenmengen effizient zu verarbeiten, verdächtiges Text- und Bildmaterial zu identifizieren, Verbindungen zu ähnlichen Inhalten aufzuzeigen, sowie Quellen und Verbreitungswege aufzubereiten. Für diese Zwecke sollten im Rahmen des Teilprojekts der TU Berlin multimodale KI-Methoden entwickelt und eingesetzt werden, um professionelle Journalisten und Bürgerjournalisten in ihrer Zusammenarbeit an der Aufklärung von Desinformation zu unterstützen. Besonderes Augenmerk galt den rechtlichen und ethischen Rahmenbedingungen: Um sowohl die Meinungsfreiheit als auch die notwendige Sorgfalt bei der Auswahl und Bewertung möglicher Desinformation zu gewährleisten, wurden sämtliche Entwicklungs- und Arbeitsprozesse unter juristischer Begleitung gestaltet.

Ablauf des Vorhabens

Das Projekt wurde erfolgreich in einem interdisziplinären Konsortium umgesetzt, das Fachkompetenzen aus der forensischen Linguistik (RUB-DFL), dem Journalismus und Fact-Checking (CORRECTIV) sowie dem Medien- und Internetrecht (TU Dortmund) mit den Erfahrungen der TU Berlin zu multimodalem maschinellem Lernen aus Text, Bild und Graphen-Informationen bündelte.

Die Zusammenarbeit mit unseren Projektpartnern war fruchtbar und für alle Seiten sehr produktiv. Wir waren vor allem mit dem CORRECTIV-Team in ständigem Kontakt und profitierten von dessen einzigartiger Expertise im Bereich Fact-Checking. Die Partner an der Ruhr-Universität Bochum stellten uns qualitativ hochwertige und gut strukturierte Faktencheck- und Nutzerabfragedaten zur Verfügung, während das Team der TU Dortmund aus der Medienjuristischen Perspektive die relevanten Aspekte des Datenschutzes und des neuen KI-Gesetzes analysierte und im Konsortium Antworten auf diesbezüglich aufkommende Fragen im Detail aufbereitete, insbesondere im Hinblick auf die Datenerfassung und -aufbewahrung.

Wegen einer Verzögerung bei der Einstellung von Softwareentwicklern für den Aufbau der Plattform ging das Faktenforum erst zum Ende des Projekts, am 6. Dezember 2024 online. Dies machte es notwendig, dass alle unsere multimodalen Lernverfahren zunächst auf anderen Daten entwickelt wurden und dass die Integration vieler der Verfahren in die Plattform noch aussteht. Um in dieser Situation hochwertige multimodale Modelle zu trainieren, wurden in großem Umfang für die Desinformationserkennung relevante Datensätze kreiert bzw. kuratiert und in weiten Teilen der wissenschaftlichen Öffentlichkeit zur Verfügung gestellt.

Ergebnisse des Projekts

An der TU Berlin gliederten sich die Arbeiten in vier Arbeitspakete (AP): In AP1 bereiteten wir in Zusammenarbeit mit unseren Projektpartnern die Ausgestaltung der Bürgerjournalistenplattform vor, mit einem Fokus auf der Akquise und Aufbereitung der Daten und der Planung der zu integrierenden KI-Module. Die zusätzliche Datensammlung konzentrierte sich vor allem auf Social-Media-Daten auf Telegram. Zu Beginn des Projekts waren noch viele weitere APIs für Forschungszwecke verfügbar (z. B. Twitter und Reddit), die sukzessive geschlossen wurden und uns so daran hinderten, einen aussagekräftigeren Datensatz zusammenzustellen, der verschiedene Demografien repräsentiert. Zur gleichen Zeit begann Telegram jedoch, eine viel wichtigere Rolle im deutschen Informationsraum zu spielen, und ist besonders bei Wählern im rechten politischen Spektrum beliebt. Ebenfalls innerhalb von AP1 erstellten wir Lernmaterial zu den KI-Tools für die Community.

Im Rahmen von AP2 konzentrierten wir uns zunächst auf die Entwicklung textbasierter Methoden zur Erkennung von Anzeichen von Desinformation. Dabei betrachteten wir sowohl linguistische als auch statistische Merkmale von Sprache, und untersuchten die Möglichkeit einer effektiven Fusion dieser Informationen. Neben der Erkennung generierten Textes spielte hierbei das *Claim-Matching* eine wichtige Rolle, um bei Nutzeranfragen schnell bestehende Faktenchecks finden und vorschlagen zu können.

Mit AP3 verfolgten wir das wissenschaftlich herausfordernde Ziel, auch sogenannte *multimodale* Desinformation zu erkennen, also Desinformation, die mehrere Modalitäten—wie Text, Bilder, Audio- und Videodaten—umfasst, und damit komplexere Analysen benötigt. Hierfür wurde auch eine Graphdatenbank für Social-Media-Daten aufgebaut und veröffentlicht, um so ein Modell zu trainieren, das neben Text auch Verbindungsinformationen aus sozialen Netzwerken berücksichtigt. Es ließ sich zeigen, dass hierdurch wesentliche Verbesserungen der Erkennungsrate von Desinformation erreicht werden.

Schließlich konzentrierte sich AP4 auf die Verbesserung der Plattform, wie z.B. die Integration visueller Komponenten für die KI-Tools, die Konzeption, Integration und Entwurf von Gamification-Elementen zur Verbesserung der Nutzerbindung und eines Human-In-the-Loop-Ansatzes zur systematischen und kontinuierlichen weiteren Verbesserung des Systems.

Die Ergebnisse des Projekts lassen sich insgesamt in zwei Aspekte fassen: Wissenschaftliche Erkenntnisse für eine verbesserte Desinformationserkennung und eine technische Realisierung einer Plattform zur gemeinschaftlichen Erkennung von Desinformation.

In wissenschaftlicher Hinsicht wurden neue Detektoren für multimodale Desinformation entwickelt. Sämtliche auf Seiten der TU Berlin vorgesehenen und geplanten KI-Tools für die Plattform wurden dabei vorbereitet und den Projektpartnern zur Verfügung gestellt. Diese Tools sind wissenschaftlich evaluiert, präzise, verallgemeinerbar, gut kalibriert und mit erklärbaren KI-Funktionen gepaart.

Als wesentliches Gesamtergebnis entstand schließlich das CORRECTIV.Faktenforum, in dem professionelle Fact-Checker und Laie vernetzt werden, um das Prinzip und zentrale Arbeitsschritte des Fact-Checkings der Öffentlichkeit zugänglich zu machen. Hierfür wurden auch Lehrmaterialien entwickelt, um eine breite Öffentlichkeit zur Zusammenarbeit an der Aufklärung von Desinformation zu befähigen und die Medienkompetenz nachhaltig zu stärken.