Sachbericht zum Verwendungsnachweis „Automatische Anonymisierung von Gerichtsentscheidungen für E-Justice und Legal-Tech (AnGer)“
Projekt 16KISA111 – FAU Erlangen-Nürnberg

— Teil I: Kurzbericht —

1 Aufgabenstellung

Das Forschungsprojekt **AnGer** beschäftigte sich mit der **automatischen Anonymisierung von Gerichtsurteilen**. Ziel war die Entwicklung und fundierte Evaluation von Verfahren, mit denen personenbezogene und andere sensible Informationen in Gerichtsurteilen zuverlässig erkannt und anonymisiert werden können. Für die Veröffentlichung von Gerichtsurteilen ist eine korrekte Anonymisierung rechtlich zwingend (vgl. DSGVO). Daher liegt der Fokus auf einem hohen **Recall**: nahezu alle zu anonymisierenden Textstellen, insbesondere Hochrisikostellen wie Personennamen und Adressen, müssen erkannt werden. Die **Precision** ist weniger kritisch; überflüssige Maskierungen sind akzeptabel, solange Lesbarkeit und inhaltlicher Zusammenhang erhalten bleiben.

Das Projekt knüpft an die Ergebnisse des **Vorgängerprojekts LeAK** (2020–2022) an. Eine zentrale Erkenntnis aus LeAK war, dass die Erkennung zu anonymisierender Textstellen effektiv durch Finetuning von Large Language Models (**LLMs**) gelöst werden kann, hierfür jedoch ein **umfangreicher** und qualitativ **hochwertiger Goldstandard** sorgfältig annotierter Trainingsdaten zwingend erforderlich ist. In LeAK konnte für zwei **Domänen** (aufgefasst als Kombination von Rechtsgebiet und gerichtlicher Instanz) ca. 99% Recall für Hochrisikostellen bei ebenfalls hoher Precision erreicht werden, nämlich amtsgerichtliche Urteile im *Miet-* und *Verkehrsrecht*. LeAK zeigte aber auch, dass nahezu perfekte Ergebnisse eben nur von hochspezialisierten Modelle in eng abgegrenzten Domänen erreicht werden können. Vor diesem Hintergrund bestand die zentrale Forschungsaufgabe von AnGer in der **Übertragbarkeit auf weitere Domänen**. Dazu mussten entsprechende Goldstandards aufgebaut und manuell annotiert werden. Diese bilden die Grundlage für Training und Evaluation der entwickelten Verfahren zur Domänenanpassung.

2 Wissenschaftlicher und technischer Stand

Für die vollautomatische Anonymisierung von Gerichtsurteilen liegen in den letzten Jahren nur **wenige belastbare Forschungsergebnisse** vor. Ein relevantes Vorhaben ist das Projekt **JANO** der Bundesländer Baden-Württemberg und Hessen. Dort wird eine **semi-automatische Anonymisierung** eingesetzt, bei der menschliche Bearbeitungsschritte weiterhin erforderlich sind. Ähnliche Ansätze werden auch in der Forschung verfolgt, bspw. durch das BMBF-geförderte Projekt HILANO. Sie können zwar die manuelle Anonymisierung beschleunigen (und bisweilen auch deren Qualität erhöhen), erreichen jedoch nicht die Skalierbarkeit vollautomatischer Verfahren, die für eine breite und systematische Veröffentlichung gerichtlicher Entscheidungen notwendig ist. Innerhalb des Forschungsnetzwerks hat sich das Projekt **MEDINYM** mit der **Anonymisierung medizinischer Texte** beschäftigt. Auch hier konnte durch Finetuning von LLMs 99% Recall erzielt werden. Die Ergebnisse zeigen wie unser Forschungsprojekt, dass LLMs grundsätzlich für Anonymisierungsaufgaben hervorragend geeignet sind.

Zunehmend werden **kommerzielle Lösungen** zur **KI-basierten Anonymisierung** angeboten, etwa von k2view oder Elephant Labs. Für diese Systeme liegen jedoch **kaum wissenschaftliche Evaluationen** vor, sodass ihre tatsächliche Leistungsfähigkeit schwer abschätzbar ist. Eine geplante Evaluation der Elephant-Labs-Lösung in Kooperation mit AnGer wurde seitens des Unternehmens nicht fortgeführt. Insgesamt zeigt der aktuelle Stand, dass zwar verschiedene Ansätze existieren, aber weiterhin erheblicher Forschungs- und Evaluierungsbedarf besteht –

insbesondere für robuste, skalierbare und domänenübergreifend einsetzbare Verfahren zur vollautomatischen Anonymisierung sensibler Textdaten.

3 Ablauf des Vorhabens

AnGer startete im Sommer 2023 nach Verzögerungen bei der Einstellung der Mitarbeiter:innen. Ursprünglich war ein Goldstandard aus vier Rechtsgebieten mit jeweils bis zu 2 Mio. Token geplant, der insbesondere auch Urteile von Landgerichten enthalten sollte; die Datenlieferung verzögerte sich jedoch bis Juli 2025. Stattdessen begann das Team mit der Annotation von **OLG-Entscheidungen** aus **elf Rechtsgebieten** – eine erheblich schwierigere Aufgabe sowohl für die Erstellung des Goldstandards als auch für die automatische Anonymisierung. Neben zusätzlichem Aufwand für die Anpassung der **Annotationsrichtlinien** an elf neue Rechtsgebiete war die **manuelle Annotation** erheblich zeitaufwändiger, obwohl sie durch ein weiterentwickeltes **Annotationstool** unterstützt werden konnte. So konnte im Projektverlauf ein **OLG-Goldstandard** im Umfang von **4,7 Mio. Token** erstellt werden. Durch vollständige manuelle **Pseudonymisierung** ist dieser Goldstandard auch außerhalb geschützter Räume nutzbar.

Für die **automatische Anonymisierung** wurde planmäßig auf die LeAK-Modelle aufgebaut; eine Domänenanpassung per **Continual Learning** erwies sich für die Übertragbarkeit auf OLG-Urteile als besonders effektiv. Eine fundierte Evaluation mit Kreuzvalidierung, Lernkurven und unserem eigens entwickelten Python-Modul CLUEval gewährleistet belastbare Ergebnisse. Namen, Adressen und Datumsangaben können darüber hinaus automatisch durch **realistische Surrogate** maskiert werden. Zusätzlich wurde die weitere **Übertragbarkeit** auf Daten von Landgerichten sowie auf Zwischenverfügungen von Registergerichten (in Kooperation mit dem Projekt DIREGA) geprüft. Eine Kooperation mit dem Universitätsklinikum Erlangen erweitert das Anwendungsspektrum auf **medizinische Texte** und wird derzeit auch nach Projektende fortgesetzt. Schließlich konnte mit Hilfe der in AnGer trainierten Modelle erstmals systematisch die **Anonymisierungsqualität veröffentlichter Urteile** überprüft werden.

4 Wesentliche Ergebnisse

Der finale **OLG-Goldstandard** umfasst 1.477 Urteile aus elf Rechtsgebieten mit insgesamt ca. 4,7 Mio. Token, darunter auch drei der vier ursprünglich geplanten Rechtsgebiete (*Bausachen*, *Familiensachen* und *Handelssachen*). Ergänzend steht ein **LG-Goldstandard** mit 102 landgerichtlichen Urteilen aus den Rechtsgebieten *Verkehrsunfallsachen* und *Bau-/Architektensachen* zur Verfügung, der im Nachgang noch auf 300 Urteile und 1 Mio. Token erweitert werden soll.

Ausgangspunkt der **automatischen Anonymisierung** war ein auf dem AG-Goldstandard von LeAK neu trainiertes Modell, welches dort einen Hochrisiko-Recall von 99,34% erreicht. Bei der domänenübergreifenden Anwendung auf OLG-Urteile ging diese auf 95,76% zurück. Zentrales Projektergebnis ist ein mit Continual Learning auf dem OLG-Goldstandard domänenangepasstes Modell, welches **in zehn Rechtsgebieten mehr als 99 % Recall auf Hochrisikostellen** erreicht. Lediglich *Immaterialgüter* sind schwierig zu anonymisieren (z.B. wg. Songtiteln und Autorennamen). Wir sehen dieses **AnGer Foundation Model** als ein hochwertiges Basismodell für die vollautomatische Anonymisierung von Gerichtsentscheidungen. Es kann per Finetuning an weitere Instanzen und Rechtsgebiete angepasst werden kann, was auf dem LG-Goldstandard erfolgreich getestet wurde (über 98,5% Hochrisiko-Recall nach Finetuning auf einer kleinen Datenmenge).

Eine **Risikoabschätzung** zeigt, dass unsere vollautomatische Anonymisierung innerhalb der Trainingsdomänen mindestens ebenso gute Ergebnisse erzielt wie die gängige Praxis der manuellen Anonymisierung. Während AG-Urteile durch die Maskierung von Hochrisikostellen zuverlässig geschützt sind, benötigen Urteile höherer Instanzen oft eine **tiefe Anonymisierung**, die auch weitere identifizierende Merkmale umfasst und von unserem Modell ebenfalls geleistet wird.