



Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen
(VIKING)

Teilvorhaben: Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den
transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation

Förderkennzeichen (FKZ) **13N16244**

Abschlussbericht **Universität der Bundeswehr München**

Laufzeit 01.01.2022 – 31.03.2025

Erstellt am 09.09.2025

Gefördert durch:



Bundesministerium
für Forschung, Technologie
und Raumfahrt

der Bundeswehr
Universität  **München**

Projektbeteiligte:

- Prof. Dr. Michaela Geierhos (1) / Projektleitung
- Falk Maoro, M. Sc. (1) / wissenschaftlicher Mitarbeiter
- Josua Köhler, M. Sc. (1) / wissenschaftliche Hilfskraft
- Philipp Hellwig (1) / studentische Hilfskraft

Institutionen:

1. Forschungsinstitut CODE, Universität der Bundeswehr München

Inhaltsverzeichnis

| | | |
|------------|---|-----------|
| I | KURZDARSTELLUNG | 3 |
| 1 | Aufgabenstellung | 3 |
| 2 | Ablauf | 3 |
| 3 | Ergebnisse | 4 |
| II | EINGEHENDE DARSTELLUNG | 6 |
| 1 | Ausgangsfragen und Zielsetzung des Projekts | 6 |
| 2 | Entwicklung der durchgeführten Arbeiten..... | 6 |
| 2.1 | Anforderungsspezifikation..... | 6 |
| 2.2 | Erstellung von Benchmark-Datensätzen & Modell-Training | 8 |
| 2.3 | Entwicklung von Debiasing-Methoden für Sprachmodelle | 14 |
| 2.4 | Generierung lokaler Erläuterungen zu Klassifikationsergebnissen..... | 16 |
| 2.5 | Demonstrator-Entwicklung..... | 18 |
| 2.6 | DIN-SPEC..... | 21 |
| 3 | Wichtigste Positionen des zahlenmäßigen Nachweises..... | 21 |
| 4 | Notwendigkeit und Angemessenheit der geleisteten Arbeit | 21 |
| 5 | Stellungnahme bezüglich der wirtschaftlichen Verwertbarkeit | 22 |
| 6 | Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen..... | 22 |
| 7 | Veröffentlichungen..... | 22 |
| III | ERFOLGSKONTROLLBERICHT (NICHT ÖFFENTLICH)..... | 23 |
| IV | BERICHTSBLATT..... | 24 |
| V | DOCUMENT CONTROL SHEET | 25 |
| VI | LITERATURVERZEICHNIS | 26 |

I Kurzdarstellung

1 Aufgabenstellung

Das Verbundprojekt VIKING (Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen) beschäftigte sich mit der interdisziplinären Erforschung und Implementierung von Lösungen zur Messung und Optimierung der Genauigkeit, Nachvollziehbarkeit und Robustheit vertrauenswürdiger KI in der polizeilichen Anwendung.

Im Teilvorhaben *Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation* hat sich die Universität der Bundeswehr München damit beschäftigt, wie die Auswertung großer Textdatensätze durch vertrauenswürdige Sprachmodelle unterstützt werden kann. In der Polizeiarbeit müssen große Datenmengen analysiert werden, um relevante Informationen für die Ermittlungen zu finden. Das hohe Datenvolumen stellt dabei eine große Herausforderung dar, die sich nicht effizient per Hand lösen lässt. Deshalb rückt der Einsatz von KI in den Fokus. Beim Einsatz von Sprachmodellen in polizeilichen Einsatzszenarien bestehen verschiedene Risiken für Individuen und die Gesellschaft. Für diese Risiken müssen Lösungen gefunden werden, um einen vertrauenswürdigen Einsatz zu gewährleisten. So können die Fehler von Modellentscheidungen weitreichende Folgen haben. In diesem Teilprojekt wurden daher Verfahren zur Herstellung der Vertrauenswürdigkeit beim Einsatz von Sprachmodellen untersucht.

Zu den Aufgabenstellungen gehörten die Identifikation von Anwendungsszenarien, die Erstellung von Datensätzen, die Untersuchung von Bias und Debiasing sowie die Generierung von Erklärungen für Modellentscheidungen. In Kombination sollten so relevante Daten für polizeiliche Anwendungsszenarien zusammengestellt werden, auf Basis derer Sprachmodelle zur Textklassifikation trainiert werden sollten. Diese Sprachmodelle sollten dazu dienen, Verfahren zur Analyse und Minderung von Bias in Daten und Modellen zu untersuchen. Um Anwendern, Betroffenen oder weiteren involvierten Personen solcher Modelle bzw. Systeme die Modellentscheidungen nachvollziehbar zu machen, sollten Verfahren zur Generierung von Erklärungen untersucht bzw. entwickelt werden.

2 Ablauf

Das Teilprojekt „Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation“ ist Teil des Arbeitspakets 5 des Verbundprojekts VIKING. In diesem Arbeitspaket wurden folgende Teilarbeitspakete (TAP) von der Universität der Bundeswehr München bearbeitet:

- 5.1 Erstellung von Benchmark-Datensätzen & Modell-Training
 - 5.1.1 Erstellung von Benchmark-Datensätzen
 - 5.1.2 Modell-Training
- 5.2 Entwicklung von Debiasing-Methoden für Sprachmodelle
 - 5.2.1 Algorithmische Experimente
 - 5.2.2 Untersuchung des Daten-Bias
 - 5.2.3 Abdeckung und Adaptivität
- 5.4 Generierung lokaler Erläuterungen zu Klassifikationsergebnissen
 - 5.4.1 Zerlegungsmethoden und Template-basierte Erläuterungen
 - 5.4.2 Integration ins Framework explAIner
 - 5.4.3 Determiniertheit von Erklärungen

○ 5.4.4 Beitrag zur Demonstrator-Entwicklung
 Neben den für die Textauswertung spezifisch geplanten Teilarbeitspaketen wurden auch Beiträge für weitere Arbeitspakete (AP) des Verbundprojekts erarbeitet:

- AP1 Bedarfsanalysen und Szenario-bezogene Spezifikation
 - 1.1 Bedarfsanalysen
 - 1.2 Methodenspezifikation
 - 1.3 Spezifikation Demonstratoren
- AP8 Methodenspektrum und Verfahrensempfehlungen
 - 8.1 Übertragbarkeit zwischen Anwendungen
 - 8.2 Integriertes, interdisziplinäres Framework
 - 8.3 Standardisierung
 - 8.4 Evaluierung der Demonstratoren

Abbildung 1 zeigt die Projektübersicht der für die Textauswertung geplanten Arbeitspakete sowie den Meilenstein im 18. Projektmonat. Am 19. Juni 2023 wurden dem Projektträger im Rahmen des Verbund- und Meilensteintreffens die Ergebnisse der Meilensteine laut Teilvorhabenbeschreibung vorgestellt. Dieser bestätigte, dass alle Meilensteinziele erfüllt wurden.

| VIKING Projektplan | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|------------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|--|--|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | | |
| AP1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP1.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP1.2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP1.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP5.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP5.2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP5.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP5.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP8 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP8.1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP8.2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP8.3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| AP8.4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Meilenstein M18 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Abbildung 1: Projektübersicht

3 Ergebnisse

Zu Beginn des Projekts wurden gemeinsam mit den polizeilichen Anwendern relevante Problemstellungen in der Polizeiarbeit identifiziert. Im Anschluss wurden geeignete Datensätze gesammelt bzw. erstellt. Auf Basis dieser Datensätze wurden verschiedene Textklassifikationsmodelle für Sequenzklassifikation, Tokenklassifikation und Relationsextraktion trainiert. Diese Modelle dienten als Grundlage für die folgende Untersuchung von Bias, Debiasing und Erklärbarkeit.

Die verwendeten bzw. entwickelten Datensätze, Modelle und Algorithmen wurden in zwei Demonstratoren integriert. Diese bieten polizeilichen Anwendern verständliche und intuitive Interaktionsmöglichkeiten, um mit den Anwendungsszenarien und Lösungen zu experimentieren. Insbesondere wurde ein umfangreiches System zur sogenannten semantischen Modellierung von Polizeiberichten entwickelt. Dieses ermöglicht die Strukturierung komplexer Berichtsinhalte durch Informationsextraktionsverfahren und bietet Anwendern die Möglichkeit, Informationen schnell zu verstehen und zu nutzen. Darüber

hinaus wurden Visualisierungen von Bias-Evaluations- und Debiasing-Verfahren sowie lokalen Erläuterungen implementiert.

Da die in diesem Teilvorhaben verwendeten bzw. erarbeiteten Modelle und Verfahren auf nicht-polizeilichen Stellvertreterdaten basieren, ist eine Übertragung der Ergebnisse und Verfahren auf andere Daten für die produktive Anwendung im polizeilichen Kontext notwendig. Deshalb wurde ein Fokus auf den Wissenstransfer mit polizeilichen Anwendern gelegt. Zu diesem Zweck wurden regelmäßige Gespräche, Diskussionen und Präsentationen der Arbeitsergebnisse mit polizeilichen Anwendern sowie mit Partnern anderer Arbeitspakete aus technischen, ethischen und juristischen Bereichen durchgeführt. So konnten Erkenntnisse im Konsortium geteilt werden, die die Entwicklung zukünftiger Lösungen positiv beeinflussen werden.

II Eingehende Darstellung

1 Ausgangsfragen und Zielsetzung des Projekts

Im Projekt „Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen“ (VIKING) wurde die Implementierung von Lösungen zur Messung und Optimierung der Genauigkeit, Nachvollziehbarkeit und Robustheit vertrauenswürdiger KI in der polizeilichen Anwendung behandelt. In dem hier betrachteten Teilprojekt zur Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Einsatz bei Sicherheitsbehörden zur Textklassifikation wurden ähnliche Lösungen für die KI-basierte Textauswertung erforscht und implementiert. Das Teilprojekt fokussierte sich auf vier Ziele und Aufgabenbereiche zur Herstellung vertrauenswürdiger Textklassifikationsverfahren:

- Teilziel 1: Als Grundlage zur Entwicklung und Evaluation der Verfahren in diesem Projekt sollte ein deutschsprachiger Benchmark-Datensatz erstellt werden, der die Anforderungen von Sicherheitsbehörden widerspiegelt und für das Training von KI-Modellen zur Textauswertung verwendet werden kann. Dieser Datensatz sollte vielfältige Textgattungen, Formate und Inhalte aufweisen, um repräsentativ für die vielfältigen Auswertungsszenarien der Polizeiarbeit zu sein und verschiedene Problemstellungen beleuchten zu können.
- Teilziel 2: Basierend auf dem Benchmark-Datensatz sollte die Eignung neuronaler Netze zur Textklassifikation untersucht werden. Von besonderem Interesse war die Performance solcher Modelle, die Anpassbarkeit im Sinne von Overfitting bzw. Underfitting bei ungleich verteilten Trainingsdaten und die Adaptionfähigkeit auf andere Fragestellungen der Textauswertung anhand von exemplarischen Problemstellungen der polizeilichen Anwender.
- Teilziel 3: Bias in Daten und Modellen kann beim Einsatz datenbasierter Modelle in realen Einsatzszenarien zu verschiedenen Risiken führen. Bias kann sich auf die Leistung des Modells auswirken, beispielsweise im Hinblick auf spezifische, über- oder unterrepräsentierte Klassen oder auf sensible Attribute wie Geschlecht, Alter oder Herkunft, die in den Daten enthalten sind. Um diese Risiken zu quantifizieren und zu minimieren, sollten Verfahren zur Evaluierung von Bias in Daten und Modellvorhersagen sowie Debiasing-Verfahren untersucht und entwickelt werden.
- Teilziel 4: Um Nachvollziehbarkeit von KI-Entscheidungen zu gewährleisten, sollten anwendergerechte, lokale Erläuterungen der Klassifikationsergebnisse generiert werden. Dazu sollten Visualisierungen und Interaktionsmöglichkeiten in einer Demonstrationsoberfläche integriert werden, die es Menschen erlaubt, die Entscheidungen eines Modells nachzuvollziehen.

2 Entwicklung der durchgeführten Arbeiten

Zur Erreichung der definierten Teilziele wurden verschiedene Projektphasen durchlaufen, auf die im Folgenden näher eingegangen wird.

2.1 Anforderungsspezifikation

Zu Beginn des Projekts hat sich die Universität der Bundeswehr München (UniBw M) mit der Definition konkreter Anwendungsszenarien und Datenanforderungen befasst. Diese wurden gemeinsam mit den polizeilichen Anwendern (Bundeskriminalamt, Landeskriminalamt Nordrhein-Westfalen, Landeskriminalamt Baden-Württemberg, Polizeipräsidium München)

und Partnern aus den Bereichen Ethik und Recht (Internationales Zentrum für Ethik in den Wissenschaften Tübingen, Hochschule für Wirtschaft und Recht Berlin) diskutiert, sondiert und verfeinert. In mehreren Gesprächen wurden Ermittlungsszenarien, Vorgehensweisen, Analyseaspekte, Problemstellungen und Herausforderungen der polizeilichen Textauswertung mit der UniBw M besprochen. Auf dieser Grundlage wurden mögliche Methoden sowie öffentlich verfügbare Datensätze recherchiert, die in diesem Zusammenhang anwendbar bzw. nutzbar sein könnten. Grundsätzlich sollen Textklassifikationsverfahren verwendet werden, um Texten oder Textbausteinen vordefinierte Klassen zuzuordnen. Diese zugeordneten Klassen sind strukturierte Daten, die Informationen über die Inhalte der Texte liefern und somit weitergehende Automatisierungen wie Auswertungen oder Analysen ermöglichen. Im Bereich der Textklassifikation lassen sich verschiedene Aufgabenstellungen durch Sprachmodelle lösen. Im Folgenden werden die in der Anforderungsspezifikation ausgewählten Aufgabenstellungen für das Projekt vorgestellt.

Sequenzklassifikation

Bei der Sequenzklassifikation wird einem gesamten Text ein oder mehrere vordefinierte Labels zugeordnet. Wird dies für eine Menge von Texten durchgeführt, so lässt sich diese Menge anhand der vergebenen Labels filtern und durchsuchen. Bei der Klassifikation wird zwischen drei Typen unterschieden: Binäre, Multiclass- und Multilabel-Klassifikation. Erstere weist einem Text genau eins von zwei möglichen Labels zu. Bei der Multiclass-Klassifikation wird einem Text genau ein Label aus einer Menge möglicher Labels zugewiesen. Die Multilabel-Klassifikation ermöglicht dagegen die Zuweisung mehrerer Labels gleichzeitig. Da in der Polizeiarbeit große Datensammlungen nach relevanten Informationen untersucht werden müssen, ist eine automatisierte Klassifikation der Texte zur Ermöglichung einer Vorfilterung, bspw. nach Delikttypen, eine Arbeitsentlastung für die manuelle Untersuchungsarbeit.

Named Entity Recognition

Die zweite in diesem Projekt geplante Aufgabe ist die Named Entity Recognition bzw. Token-Klassifikation. Dabei werden Textsequenzen in Wörter bzw. Tokens (Wortbausteine) aufgeteilt, bevor jedem Token genau ein Label zugewiesen wird. Dieses Verfahren ermöglicht es, den Textbausteinen gewisse Informationsklassen zuzuordnen. Beispielsweise kann so erkannt werden, ob es sich bei einem Begriff um einen Namen, eine Ortsangabe oder eine Zeitangabe handelt. Dies ist besonders dann hilfreich, wenn nach spezifischen Inhalten oder Inhaltstypen in Texten gesucht wird. In der Polizeiarbeit werden Texte häufig manuell nach Informationsschnipseln durchsucht. Besonders bei vielen zu analysierenden Texten hilft das Verfahren dabei, den Anwender bei dem Finden solcher Informationen zu unterstützen.

Relationship Extraction

Die dritte Klassifikationsaufgabe, die durch Sprachmodelle gelöst werden soll, ist die Relationship Extraction, die mit den Ausgaben der Named Entity Recognition interagiert. Die im vorangegangenen Schritt erkannten Entitäten liegen zwar aufgelistet vor, enthalten jedoch keine Informationen über Zusammenhänge zwischen den Entitäten. Die extrahierten Entitätsinformationen können zwar für verschiedene Filter- und Suchvorgänge verwendet werden, reichen aber alleine nicht aus, damit ein Mensch die Vorgänge des zugehörigen Texts versteht. Durch das Erkennen und Speichern vordefinierter Beziehungs-Klassen zwischen den Entitäts-Klassen lassen sich mittels der Relationship Extraction Abläufe und Zusammenhänge extrahieren und als strukturierte Informationen speichern. So kann beispielsweise erfasst werden, welche Person sich zu welchem Zeitpunkt an welchem Ort befand, wie viele

unterschiedliche Personen in einem Text erwähnt werden oder welche beschreibenden Attribute einzelnen Personen in einem Text zugeordnet werden.

2.2 Erstellung von Benchmark-Datensätzen & Modell-Training

Für die Lösung der vorgestellten Aufgaben durch Sprachmodelle sind Trainingsdaten notwendig, die Texte und dazugehörige Labels enthalten. Die Modelle verarbeiten während des Trainings die Texte, berechnen Vorhersagen über die zugehörigen Labels und werden bei Fehlern durch Lernalgorithmen so angepasst, dass die Wahrscheinlichkeit für solche Fehler bei zukünftigen Vorhersagen verringert werden. Die ursprüngliche Zieldefinition bzw. Anforderung des Projekts, verschiedene polizeilich relevante Textsorten in verschiedenen Sprachen, u. a. mit Echtdateien, für spezifische polizeilich relevanten Klassifikationsaufgaben in einem großen Benchmark-Datensatz zu sammeln, der für das Training der Sprachmodelle und die darauffolgenden Forschungsaspekte zum Debiasing und zur Erklärbarkeit genutzt werden kann, erfordert die Verfügbarkeit und anschließende Sammlung, Verarbeitung und etwaige Annotation von Textdaten. Von den polizeilichen Projektpartnern konnten während des Projekts weder annotierte noch rohe Textdaten bereitgestellt werden. Stattdessen wurde ausschließlich in öffentlichen Quellen nach möglichen Daten recherchiert.

Im Rahmen der Textkorpusrecherche wurden öffentliche Datensätze aus verschiedenen Textquellen und Textformaten zusammengetragen. Dazu wurde eine Tabelle mit ca. 30 Datensätzen erstellt, in der die Datensätze nach Größe, Textformat, Sprache und Klassifikationsszenario (z. B. binäre Klassifikation, Multilabel-Klassifikation, Multiclass-Klassifikation) kategorisiert und gefiltert wurden. Ein Großteil der verfügbaren Datensätze bezieht sich auf die Identifikation und Klassifikation von Hassrede und Toxizität in sozialen Medien, wie X (ehemals Twitter) und Facebook. Zu den Datensätzen existieren bereits zahlreiche wissenschaftliche Veröffentlichungen sowie etablierte und performante Methoden. Darüber hinaus wird im Rahmen der polizeilichen Textauswertung gemäß den Diskussionen während der Anforderungsdefinition kein Monitoring von Social Media zur Erkennung von u. a. Hassrede durchgeführt. Aufgrund der geringen erwarteten Forschungserkenntnisse anhand dieser Daten und des fehlenden Bezugs zu polizeilichen Analyseaspekten wurde entschieden, diese Datensätze größtenteils nicht für Modelltrainings und die weiteren Forschungsaspekte in diesem Projekt zu verwenden.

Im Folgenden werden die Datensätze aufgeführt, die in diesem Teilprojekt hauptsächlich untersucht bzw. erstellt und verwendet wurden.

PAN12: Sexual Predator Identification

Der einzige Datensatz aus dieser Sammlung, der aufgrund seines thematischen Bezugs zur Polizeiarbeit für die weitere Verwendung eingeplant wurde, heißt „*PAN12: Sexual Predator Identification*“ [1]. Der im Rahmen eines Shared Task veröffentlichte Datensatz enthält englischsprachige Chatverläufe aus verschiedenen Quellen, in denen mutmaßliche Sexualstraftäter identifiziert werden müssen und liefert so ein breites Spektrum an sprachlichen und inhaltlichen Merkmalen. Die Verfasser der Chatnachrichten wurden von den Herausgebern pseudonymisiert, so dass keine Rückschlüsse auf die Identität der Personen möglich sind. Darüber hinaus werden die Pseudonyme in einer Liste aufgeführt, die den Sexualstraftätern zugeordnet werden können. In dem Shared Task sollten zunächst Sexualstraftäter identifiziert und anschließend besonders auffällige Textnachrichten erkannt werden. Die unter den Teilnehmern des Shared Tasks erfolgreichste Lösung definierte drei aufeinanderfolgende Klassifikationsaufgaben: *Suspicious Conversation Identification*, *Victim from Predator Disclosure* und *Suspicious Line Identification* [2]. Alle drei Aufgaben sind binäre Klassifikationsaufgaben, die sich jeweils auf andere Textsequenzen beziehen. Da in jeder

Konversation eine verdächtige Person vorhanden sein kann, klassifiziert erstere den gesamten Konversationsverlauf als eine Textsequenz und beurteilt, ob eine verdächtige Person vorhanden ist. Ist eine verdächtige Person vorhanden, so wird bei der *Victim from Predator Disclosure* klassifiziert, welche Verdächtige und welche Opfer sind. Dazu werden für jeden Konversationsteilnehmer alle Nachrichten zu einer Sequenz zusammengesetzt und klassifiziert. Wurde jemand als verdächtig klassifiziert, so wird die *Suspicious Line Identification* durchgeführt. Bei dieser werden alle Nachrichten der als verdächtig klassifizierten Person einzeln klassifiziert, um zu erkennen, welche Nachrichten besonders auffällig bzw. verdächtig sind. Somit bietet der Datensatz aufgrund der strafrechtlich bzw. ethisch relevanten Inhalte, der vielfältigen Textquellen und Gesprächsthemen, sowie der geringen Anzahl an Beispielen mit verdächtigen Texten, eine interessante und herausfordernde Aufgabenstellung, die deshalb in diesem Teilprojekt erneut aufgegriffen wird.

Zur Ermittlung von Performance-Benchmarks wurden die drei zuvor benannten Aufgabenstellungen übernommen und Verfahren zur Lösung mittels aktueller Sprachmodelle entwickelt. Die Datensätze für alle drei Aufgaben sind jeweils in Trainings- und Validierungsdaten aufgeteilt. Mit den Trainingsdaten wurden in diesem Projekt verschiedene Sprachmodelle trainiert und die Performance anhand der Validierungsdaten gemessen. Für die *Suspicious Line Identification* sind keine Trainingsdaten verfügbar. Deshalb wurden die Testdaten in Trainings- und Validierungsdaten aufgeteilt, um auch diese Aufgabe mittels überwachten Lernens zu lösen. Die Ergebnisse des BERT-Modells [3] mit der höchsten Performance, dem `google-bert/bert-large-uncased`¹, sind in Tabelle 1 aufgeführt.

Tabelle 1: Evaluationsergebnisse der Sexual Predator Identification mit dem `google-bert/bert-large-uncased` Modell auf den Validierungsdatensätzen der drei Aufgaben. Für die *Suspicious Line Identification* wurde der Testdatensatz zu 80 % für das Training und zu 20 % für die Validierung verwendet.

| Accuracy | F1 | Precision | Recall |
|---|-------|-----------|--------|
| Suspicious Conversation Identification | | | |
| 0,998 | 0,986 | 0,984 | 0,988 |
| Victim from Predator Disclosure | | | |
| 0,904 | 0,906 | 0,892 | 0,919 |
| Suspicious Line Identification | | | |
| 0,961 | 0,620 | 0,640 | 0,601 |

Presseportal-Datensatz

Die weiteren Datensätze liefern keine polizeilich relevanten Inhalte oder enthalten keine für das Training von KI-Modellen notwendigen Annotationen. Von den polizeilichen Projektpartnern konnten ebenfalls keine Daten zur Verfügung gestellt werden. Da folglich keine den spezifischen Anforderungen entsprechenden Daten zur Verfügung standen, konnte kein inhaltlich relevanter Benchmark-Korpus in deutscher Sprache erstellt werden. Stattdessen wurde in Absprache mit den Projektpartnern entschieden, sogenannte Surrogatdaten zu verwenden, die für die Methodenentwicklung nutzbar sind, für den behördlichen Einsatz aber nachträglich ausgetauscht werden müssten. Hierfür wurden Presseartikel aus der Rubrik *Blaulicht* der Website [Presseportal.de](https://www.presseportal.de)² ausgewählt. Die Datenerhebung erfolgte mittels Webscraping, wobei die Berichtstexte, Überschriften, Herausgeber, Veröffentlichungszeitpunkte, Kontaktdaten sowie redaktionell annotierte Themen und Orte erfasst wurden. In den Berichtstexten werden polizeilich relevante Sachverhalte, wie z. B. Delikte, Einsätze, Fahndungen oder Unfälle sachlich beschrieben. Die Inhalte können somit als

¹ Verfügbar unter <https://huggingface.co/google-bert/bert-large-uncased>.

² Verfügbar unter <https://www.presseportal.de/blaulicht/>

Stellvertreterdaten genutzt werden, um strukturierte Daten mittels Informationsextraktion durch Sprachmodelle zu erzeugen. Eine Herausforderung bei der Verwendung dieser Daten für Sprachmodelle ist, dass keine Annotationen für Klassifikationen vorhanden sind. Deshalb wurden manuell Annotationen für drei verschiedene Klassifikationsaufgaben erstellt.

Aufgabe 1: Delikttypenklassifikation

Eine wichtige Aufgabe bei der Durchsuchung großer Datenmengen ist die Filterung von Texten anhand vordefinierter Attribute. Im Falle der Presseportal-Berichte ist eine Filterung anhand der vorhandenen Delikttypen interessant. Dazu wurden vier Delikttypen in Absprache mit den polizeilichen Anwendern des VIKING-Projekts definiert: *Betäubungsmittel*, *Gefahr für Leib und Leben*, *Waffen* und *Anderes*. Für diese prototypische Umsetzung decken diese vier Klassen die wichtigsten Inhalte ab und bieten eine klare Abgrenzbarkeit für die nicht polizeilich geschulten Annotatoren. Es wurden 758 Beispiele manuell annotiert, von denen 70 % für das Training und 30 % für die Validierung verwendet wurden. Die Verteilung der Klassen ist in Tabelle 2 zu finden. Diese Datenbasis wurde genutzt, um verschiedene Sprachmodelle nachzutrainieren.

Tabelle 2: Klassenverteilung der Delikttypen über Trainings- und Validierungssplits sowie den gesamten Datensatz.

| Kategorie | Training | | Validierung | | Summe | |
|---------------------------|----------|--------|-------------|--------|-------|--------|
| | Wahr | Falsch | Wahr | Falsch | Wahr | Falsch |
| Betäubungsmittel | 299 | 232 | 115 | 112 | 414 | 344 |
| Waffen | 106 | 425 | 36 | 191 | 142 | 616 |
| Anderes | 188 | 343 | 101 | 126 | 289 | 469 |
| Gefahr für Leib und Leben | 164 | 367 | 74 | 153 | 238 | 520 |

Die Ergebnisse des Modells mit der stärksten Performance (deepset/gbert-large³) sind in Tabelle 3 aufgeführt.

Tabelle 3: Evaluationsergebnisse des Delikttypklassifikations-Modells (deepset/gbert-large) auf dem Validierungsdatensatz.

| Label | F1-Score | Precision | Recall |
|---------------------------|----------|-----------|--------|
| Macro | 0,83 | 0,89 | 0,78 |
| Micro | 0,84 | 0,91 | 0,78 |
| Betäubungsmittel | 0,99 | 0,99 | 0,98 |
| Waffen | 0,86 | 0,86 | 0,86 |
| Anderes | 0,66 | 0,82 | 0,55 |
| Gefahr für Leib und Leben | 0,80 | 0,89 | 0,73 |

Aufgabe 2: Named Entity Recognition

Klassifizierte Delikttypen enthalten nur wenige Informationen über die Inhalte der Polizeiberichte. Die Berichte enthalten unstrukturierte Informationen über Personen, Personengruppen, Objekte, Orte, Zeitpunkte und Beziehungen durch Assoziationen oder Handlungen in den beschriebenen Ereignissen. Um diese unstrukturierten Informationen eines Texts zu verstehen, müssen Ermittler die Texte aufmerksam lesen und je nach Komplexität evtl. Notizen und Diagramme erstellen. Dieser Prozess ist zeitaufwändig und erfordert die Konzentration eines Ermittlers auf nur einen Text. Eine Methode, um diesen Prozess zu automatisieren, ist das Klassifizieren von Textsegmenten in vordefinierte Informationskategorien. So können spezifische Informationen gesucht oder automatische Analysen durchgeführt werden.

Die Klassifikation solcher Textsegmente wird Named Entity Recognition bzw. Token Klassifikation genannt. Dabei wird eine Textsequenz zunächst in eine Liste kleinerer Segmente, sog. Tokens, überführt. Schließlich werden diese einzeln klassifiziert und jeweils

³ Verfügbar unter <https://huggingface.co/deepset/gbert-large>

zusammengehörige Tokens zu sog. Spans zusammengefügt. Einem Span (ein Begriff aus ein oder mehreren Wörtern) wird somit eine Informationsklasse, z. B. Ortsangabe oder Personennamen, zugewiesen. Das bekannteste Klassifikationsschema für Named Entity Recognition umfasst die Klassen LOC, MISC, ORG und PER (für Ortsangabe, Anderes, Organisation und Person) [4]. Bei der Analyse der Polizeiberichte ist vor allem aufgefallen, dass keine Personennamen in den Texten vorhanden sind. Weiterhin ist die Klasse MISC uneindeutig, sodass der Informationsgehalt für automatische Analysen gering ist. Zudem enthalten die Texte weitere Informationstypen, die von dem Schema nicht erfasst werden. Deshalb wurde ein erweitertes Klassifikationsschema erstellt, mit dem anschließend Polizeiberichte manuell annotiert wurden. In Tabelle 4 ist das Klassifikationsschema mit Kurzbeschreibungen aufgeführt. Es wurden 751 Polizeiberichte mit Annotationen für die Named Entity Recognition versehen. Die Verteilung der Klassen ist in Tabelle 5 angegeben.

Tabelle 4: Erläuterung der möglichen Klassifikationstypen für die Named Entity Recognition.

| Entity | Beschreibung |
|------------------|--|
| LOCATION | Jede Form von Ortsangaben, wie z. B. Länder, Städte, Straßen, Bereiche an öffentlichen oder privaten Plätzen. |
| ORGANIZATION | Institutionen, Gruppierungen, formelle Organisationen. |
| PERSON-REFERENCE | Referenzierende Bezeichnungen von Personen oder Personengruppen, wie z. B. „Er“, „Sie“, „die Täter“, „die Mitarbeiterin“ |
| COUNT | Begriffe zur Zählung oder Zahlen inkl. möglicher Einheitsangaben. |
| OBJECT | Objekte, wie z. B. Waffen, Betäubungsmittel, Fahrzeuge oder Kleidung. |
| PROPERTY | Beschreibende Begriffe, meist Adjektive, wie z. B. Farben, Beschreibungen von Formen, Größen oder Altersangaben. |
| TIME | Zeit- und Datumsangaben. |

Tabelle 5: Klassenverteilung der Trainings- und Validierungsdatensätze für die Named Entity Recognition.

| Label | Training | Validierung | Summe |
|------------------|----------|-------------|-------|
| COUNT | 4.795 | 607 | 5.402 |
| LOCATION | 3.281 | 386 | 3.667 |
| OBJECT | 4.748 | 500 | 5.248 |
| ORGANIZATION | 1.400 | 156 | 1.556 |
| PERSON-REFERENCE | 8.383 | 1.060 | 9.443 |
| PROPERTY | 4.126 | 366 | 4.492 |
| TIME | 2.129 | 203 | 2.332 |

Aufgrund der häufig sehr ähnlich formatierten Zeit- und Datumsangaben in den Polizeiberichten wurde sich entschieden, die Klassifikation von TIME-Spans durch einen regelbasierten Ansatz lösen zu lassen. Dies wurde mit der Open Source Bibliothek *duckling*⁴ umgesetzt, die vordefinierte und erweiterbare Regeln zur Extraktion verschiedener Textdatentypen, wie z. B. URLs, E-Mails oder Telefonnummern, bereitstellt. Die weiteren annotierten Span-Klassen wurden genutzt, um Sprachmodelle zur Token-Klassifikation zu trainieren. Tabelle 6 zeigt die Ergebnisse des nachtrainierten mLUKE-large Modells⁵ [5].

⁴ Verfügbar unter <https://github.com/facebook/duckling>

⁵ Verfügbar unter <https://huggingface.co/studio-ousia/mluke-large>

Tabelle 6: Evaluationsergebnisse des Token-Klassifikations-Modells (studio-ousia/mluke-large) auf den Validierungsdaten

| Label | Tag-Typ | F1-Score | Precision | Recall |
|------------------|---------|----------|-----------|--------|
| Macro | - | 0,86 | 0,87 | 0,85 |
| Micro | - | 0,96 | 0,96 | 0,96 |
| O-Tag | O | 0,98 | 0,98 | 0,98 |
| LOCATION | B | 0,76 | 0,76 | 0,76 |
| | I | 0,83 | 0,91 | 0,77 |
| ORGANIZATION | B | 0,93 | 0,92 | 0,94 |
| | I | 0,86 | 0,90 | 0,82 |
| PERSON-REFERENCE | B | 0,92 | 0,93 | 0,94 |
| | I | 0,92 | 0,95 | 0,89 |
| OBJECT | B | 0,86 | 0,86 | 0,87 |
| | I | 0,87 | 0,89 | 0,86 |
| COUNT | B | 0,86 | 0,88 | 0,84 |
| | I | 0,89 | 0,93 | 0,84 |
| PROPERTY | B | 0,70 | 0,63 | 0,79 |
| | I | 0,76 | 0,71 | 0,81 |

Aufgabe 3: Relationship Extraction

Die in der vorausgehenden Aufgabe, der Named Entity Recognition, erkannten Informationen liegen zwar strukturiert vor, sind aber im Informationsgehalt limitiert. Jede extrahierte Entität enthält isoliert betrachtet nur Informationen über die Position innerhalb der Textsequenz, den Begriffstext und die dazugehörige Informationsklasse. Zusätzlich dazu sind für Ermittlungen die Beziehungen zwischen den erkannten Entitäten interessant. Um auch diese automatisiert zu erkennen, können Sprachmodelle eingesetzt werden, um zwischen zwei Entitäten innerhalb einer Textsequenz eine mögliche Beziehung zu klassifizieren. Diese Relationship Extraction wird für jede mögliche Beziehungskombination aller Entitäten eines Polizeiberichts durchgeführt, um die extrahierten Entitäten mit zusätzlichen Kontextinformationen anzureichern und weitere Analysen und Visualisierungen zu ermöglichen.

Passend zum erweiterten Annotationsschema für die Named Entity Recognition wurde deshalb ein Annotationsschema für die Relationship Extraction definiert. Dieses erlaubt die Erkennung der wichtigsten Beziehungen zwischen den Entitäten, die in den Polizeiberichten vorkommen. Allerdings besteht dabei kein Anspruch darauf, die Gesamtheit aller potentiellen Beziehungen der Realität abzubilden, da dies die Komplexität sowohl der Annotationsaufgabe als auch der Sprachmodellierungsaufgabe erheblich steigern würde. Insofern dient die Verwendung des Schemas in Anbetracht der Stellvertreterdaten als Machbarkeitsstudie.

Es wurden insgesamt fünf Beziehungen definiert, die zwischen verschiedenen Span-Klassen verwendet werden können. Dabei gilt zwischen zwei Spans jeweils einer als Quell-Span und einer als Ziel-Span. Gehört ein *PROPERTY* bspw. zu einem *OBJECT*, dann gilt das *PROPERTY* als Quell-Span und das *OBJECT* als Ziel-Span. Die Beziehung wird dann durch das Label *BELONGSTO* definiert. Die nächste Beziehung, *COUNTS*, geht immer von *COUNT*-Spans aus und ist vorhanden, wenn z. B. Personen, Objekte oder Gruppen gezählt werden. Dies ermöglicht unter anderem das strukturierte Speichern von Angaben bzgl. Gruppengrößen oder Mengenangaben. Die nächste Beziehung, *STAY*, verbindet Personen, Personengruppen oder Organisationen mit Orts- und Zeitangaben. So können Aufenthalte identifiziert und gespeichert werden. Weiterhin besteht die *CONTAINS*-Beziehung, wenn eine Personen-Referenz eine andere Personen-Referenz oder wenn ein Objekt ein anderes Objekt enthält. Die letzte Beziehung, die *COREF*-Beziehung, wird verwendet, wenn sich mehrere Spans auf dieselbe Entität beziehen. Dies ist der Fall, wenn z. B. eine verdächtige Person von mehreren

Begriffen (z. B. „Er“, „der Täter“, „dieser“, ...) im Text referenziert wird. Alle möglichen Beziehungen zwischen Quell- und Ziel-Spans sind in Tabelle 7 zu finden.

Tabelle 7: Mögliche Beziehungen zwischen Quell-Spans und Ziel-Spans.

| Quell-Span | Ziel-Span | | | | | | |
|------------------|-----------|-------------------|-------------------|-------------------|-------|-------------------|----------|
| | COUNT | OBJECT | PERSON-REFERENCE | ORGANIZATION | TIME | LOCATION | PROPERTY |
| COUNT | | COUNTS | COUNTS | COUNTS | | COUNTS | |
| OBJECT | | COREF CONTAINS | BELONGSTO | | | | |
| PERSON-REFERENCE | | | COREF CONTAINS | | STAY | STAY | |
| ORGANIZATION | | | CONTAINS | COREF CONTAINS | STAY | STAY | |
| TIME | | | | | COREF | | |
| LOCATION | | | | | | COREF CONTAINS | |
| PROPERTY | | BELONGSTO | | | | BELONGSTO | |

Unter Verwendung dieses Annotationsschemas wurden 751 Polizeiberichte annotiert. Dabei wurden 85 % der Beispiele für Trainings- und 15 % für Validierungszwecke verwendet. Die Klassenverteilung ist in Tabelle 8 zu sehen.

Tabelle 8: Klassenverteilung der Trainings- und Validierungsdatensätze für die Relationship Extraction.

| Beziehung | Training | Validierung | Summe |
|-----------|----------|-------------|-------|
| BELONGSTO | 7.142 | 987 | 8.129 |
| CONTAINS | 2.943 | 541 | 3.484 |
| COREF | 7.146 | 1.481 | 8.627 |
| COUNTS | 3.871 | 723 | 4.594 |
| STAY | 3.350 | 660 | 4.010 |

Dieser Datensatz wurde ebenfalls für das Training von Sprachmodellen verwendet. Die Ergebnisse des Relationship Extraction Modells (mLUKE-large⁶) sind in Tabelle 9 zu sehen.

Tabelle 9: Evaluationsergebnisse des Relationship Extraction-Modells (studio-ousia/mluke-large) auf den Validierungsdaten.

| Label | F1 Score | Precision | Recall |
|-----------|----------|-----------|--------|
| Macro | 0,77 | 0,77 | 0,77 |
| Micro | 0,96 | 0,96 | 0,96 |
| BELONGSTO | 0,74 | 0,74 | 0,75 |
| CONTAINS | 0,52 | 0,55 | 0,49 |
| COREF | 0,85 | 0,81 | 0,89 |
| COUNTS | 0,97 | 0,97 | 0,97 |
| NONE | 0,98 | 0,98 | 0,98 |
| STAY | 0,55 | 0,56 | 0,54 |

⁶ Verfügbar unter <https://huggingface.co/studio-ousia/mluke-large>

Abbildung 2 zeigt einen beispielhaften Polizeibericht, in welchem Entitäten markiert und Beziehungen zwischen den markierten Entitäten visualisiert wurden. Die Kombination aus den beiden Aufgabenstellungen der Named Entity Recognition und Relationship Extraction erlaubt somit eine Darstellung mittels Graphen, welche die Hauptakteure und Zusammenhänge eines Texts komprimiert verständlich macht. So können große Textdatensätze automatisiert von Sprachmodellen verarbeitet werden. Die Ergebnisse können Anwendern in geeigneten Nutzeroberflächen bereitgestellt werden, um präzise Informationen effektiv zu finden, ein schnelles Textverständnis komplexer Zusammenhänge aufzubauen und die Inhalte mit anderen Ermittlungsinformationen zu verknüpfen.

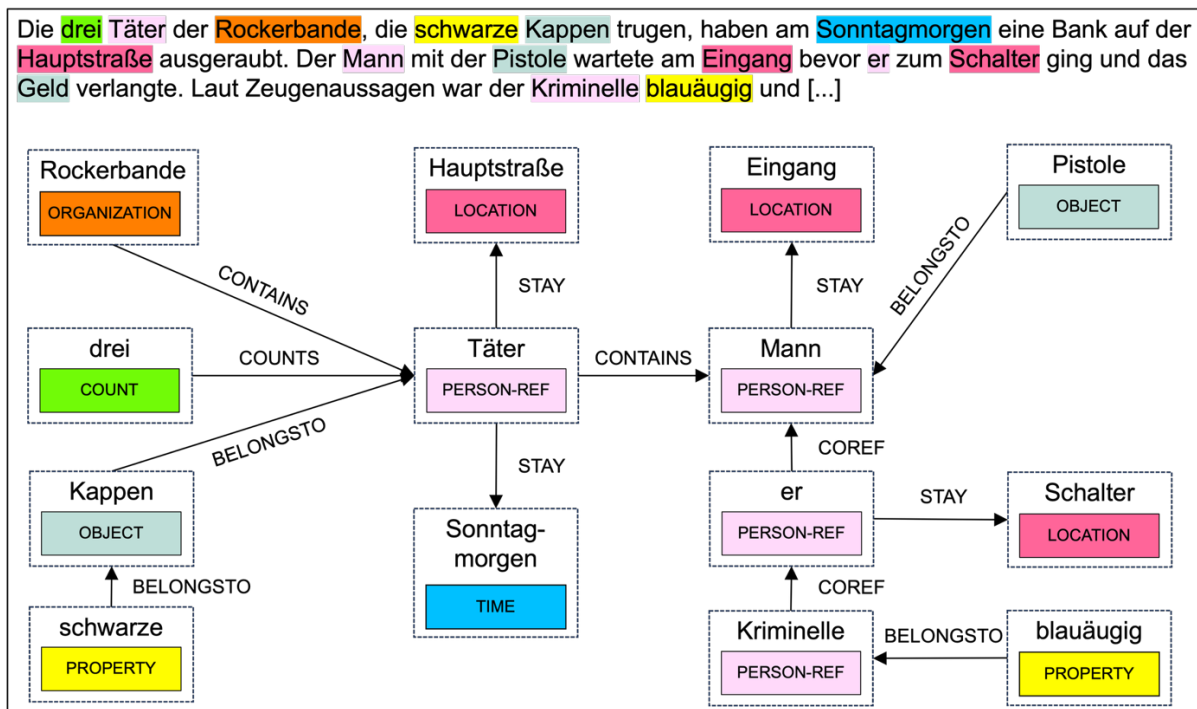


Abbildung 2: Polizeibericht mit annotierten Entitäten und Beziehungen.

Die Analyse der Polizeiberichte durch die Verwendung der Sprachmodelle kann als *Semantische Modellierung* bezeichnet werden. Diese kombiniert die drei Aufgabenstellungen, Delikttypklassifikation, Named Entity Recognition und Relationship Extraction, verbindet die Ein- bzw. Ausgaben mit zusätzlichen Metadaten aus den Berichten oder Geographie-Datenbanken und erstellt so ein strukturiertes Datenmodell, das als Grundlage für Visualisierungen dient. Zudem können die gespeicherten Daten nach verschiedenen Kriterien gefiltert oder automatisiert analysiert werden, um die Ermittlungsarbeit zu unterstützen und etwaige Kreuztreffer nach der Verlinkung zu anderen strukturierten Daten offenzulegen.

2.3 Entwicklung von Debiasing-Methoden für Sprachmodelle

Das Verhindern von Bias erfordert zunächst die Messung des Vorhandenseins von Bias. Bias eines Modells kann in zwei Kategorien unterteilt werden, für die jeweils unterschiedliche Methoden zur Evaluation bestehen: intrinsischer Bias und extrinsischer Bias [6]. Transformer-basierte Encoder-Modelle, wie BERT, RoBERTa oder LUKE werden mittels selbstüberwachtem Lernen auf großen Datensätzen vortrainiert und im zweiten Schritt auf spezifische Aufgaben unter Verwendung kleinerer Datensätze nachtrainiert [3], [5], [7]. Dieses zweistufige Lernen, das sogenannte Transferlernen, ermöglicht die Einführung von Bias in verschiedenen Phasen an verschiedenen Stellen innerhalb des Modells. Insbesondere während des Vortrainierens erlernt das Modell implizite Assoziationen zwischen sämtlichen sprachlichen Konzepten, die

grundsätzlich durch die in den Daten vorkommenden Verteilungen solcher Konzepte beeinflusst werden. Verzerrungen in den Daten wirken sich zunächst vor allem auf die numerischen Repräsentationen aus, die von solchen Modellen erstellt werden, um Textsequenzen zu modellieren. Die Vektorrepräsentationen, sogenannte Embeddings, erfassen Sprache und deren Bedeutung in mehrdimensionalen Räumen. So liegt impliziter Bias dann vor, wenn bspw. weibliche Namen geringere Distanzen zu negativ konnotierten Wörtern haben, als männliche Namen.

Extrinsischer Bias wird insbesondere während des Nachtrainierens erlernt. Beim Nachtrainieren der genannten Modelle wird dem Modell eine zusätzliche Schicht, z. B. zur Klassifikation der Stimmung innerhalb eines Texts, hinzugefügt. Die Parameter dieser Schicht werden entsprechend der Aufgabe und der Trainingsdaten optimiert, sodass das Modell ungesehene Texte mit hoher Genauigkeit klassifizieren kann. Wird die Klassifikationsgüte nun in Bezug zu sensiblen Attributen, wie z. B. dem Geschlecht, der Herkunft oder dem Beruf gesetzt, dann besteht extrinsischer Bias, sobald die Klassifikationsgüte sich für verschiedene Ausprägungen eines sensiblen Attributs unterscheidet.

Intrinsischer Bias

Zur Evaluation von Bias in den von uns verwendeten Encoder-basierten Sprachmodellen wurden zunächst bestehende Evaluationsmethoden aus der Literatur recherchiert, ausgewählt und an die bestehenden Modelle angepasst. Die *Projection*-Methode [8] misst, welches von zwei Zielkonzepten einem Attributkonzept näherliegt, indem sie die Lage ihrer Vektorrepräsentationen im semantischen Raum vergleicht. Ausgangspunkt sind Satz- oder Wortlisten, die jeweils ein Zielkonzept und das Attributkonzept repräsentieren. Diese Listen werden mithilfe des zu evaluierenden Sprachmodells in Vektoren (Embeddings) überführt und anschließend zu jeweils einem mittleren Konzeptvektor aggregiert. Der Attributvektor wird dann auf die Verbindung zwischen den beiden Zielvektoren projiziert, um zu bestimmen, welches Zielkonzept dem Attributkonzept semantisch näher liegt. Nutzt man beispielsweise zwei Geschlechter (männlich, weiblich) als Zielkonzepte und eine charakterliche Eigenschaft (wie z. B. fröhlich), dann kann berechnet werden, ob das Sprachmodell unter Verwendung der jeweiligen Textsequenzen eines der beiden Geschlechter mit der Eigenschaft stärker assoziiert. Der *Word Embedding Association Test* (WEAT) bzw. *Sentence Embedding Association Test* (SEAT) basiert ebenfalls auf Sequenzlisten, deren Embeddings und der Berechnung von Distanzen in Vektorräumen [9], [10]. Auch hier werden Listen mit Textsequenzen erstellt, die für zwei Zielkonzepte (z. B. männlich, weiblich) und hier für zwei Attributkonzepte (z. B. kompetent, inkompetent) verwendet werden. Die Embeddings aller Listen werden durch ein Sprachmodell berechnet und statistische Assoziationstests durchgeführt, die Aussagen darüber erlauben, ob und wie stark Assoziationen innerhalb der Embeddings bezogen auf die definierten Konzepte vorhanden sind.

Ein weiterer Test zur Messung von intrinsischem Bias eines Modells ist die *CrowS-Pairs*-Methode [11]. Diese vergleicht die Wahrscheinlichkeiten, dass ein Modell einen stereotypischen oder anti-stereotypischen Satz generiert. Für diese Evaluation wird nicht nur die Encoder-Schicht des Modells, sondern auch die für das Vortraining verwendete Masked-Language-Modeling-Schicht verwendet, die für Wortlücken einer Textsequenz eine Wahrscheinlichkeitsverteilung für in die Lücken passende Wörter generiert. Nutzt man zwei Sätze, die sich sprachlich nur marginal unterscheiden, aber eine gegensätzliche Bedeutung haben, dann wird für jedes Wort der Sequenzen die Generierungswahrscheinlichkeit berechnet und miteinander verglichen, um so schließlich Rückschlüsse auf das Vorhandensein von Bias zu ziehen. Daraus folgt, dass für den Test Satz-Templates benötigt werden, die Stereotypen aufweisen, auf die das Modell getestet werden soll.

Extrinsischer Bias

Extrinsischer Bias wird ersichtlich, sobald Sprachmodelle in konkreten Anwendungen gewisse Voreingenommenheit aufweisen. Insofern ist das Evaluieren dessen und die Methodenauswahl abhängig vom Anwendungskontext. Für die in diesem Projekt relevanten Klassifikationsaufgaben ist vor allem die Betrachtung der Klassifikationsperformance im Hinblick auf sensible Attribute ausschlaggebend. Dafür sind allerdings nicht nur Eingabetexte und deren Klassen-Labels, sondern auch Annotationen der sensiblen Attribute notwendig. Beispielsweise könnten Arztbewertungen durch Sprachmodelle mit einer Anzahl Sterne (0-5) versehen werden. Soll nun z. B. ein Geschlechterbias evaluiert werden, dann wird für jede Bewertung eine Annotation des Geschlechts des Arztes bzw. der Ärztin benötigt. Besteht für den Klassifikator eine höhere Genauigkeit für ein Geschlecht als für das andere, dann liegt ein extrinsischer Bias vor.

Da für die vorliegenden Datensätze und die darin enthaltenen Textsequenzen kein direkter Bezug zu einzelnen Personen bzw. Personengruppen und keine Annotationen sensibler Attribute vorliegen, ist eine Messung extrinsischen Bias nicht direkt möglich. Um trotz der Datenlage solche Evaluationen durchzuführen, wurde ein Template-basierter Ansatz entwickelt, um Wörter einer Ausprägung eines sensiblen Attributs mit solchen einer Ausprägung auszutauschen. So wurden für die Evaluationsdaten dupliziert, um bspw. Daten nur mit weiblichen Personen und nur mit männlichen Personen zu erzeugen, bevor die Klassifikationsperformance des Modells auf den zwei Datensätzen miteinander verglichen wird. Aufgrund der unzureichenden Datenlage und dem Szenario der Klassifikation von Polizeiberichten, bei denen nicht nur eine Person im Vordergrund steht, konnte dabei kein signifikanter Bias identifiziert werden.

Im Kontext aspektbasierter Stimmungsanalyse von Arztbewertungen wurde während des Projekts eine Veröffentlichung zur Erkennung von Bias anhand der Named Entity Recognition Aufgabe erarbeitet [12]. Dort wurden sowohl intrinsische als auch extrinsische Bias-Tests durchgeführt und Korrelationen zwischen den Ausgaben der Tests überprüft. Es konnte auch dort kein signifikanter Bias festgestellt werden, sodass auch ein Entfernen von Bias durch Debiasing-Methoden erschwert wurde.

Für das Debiasing von Sprachmodellen wurde das *SENT-DEBIAS*-Verfahren integriert [13]. Das Verfahren verwendet ebenfalls Satz- bzw. Wortlisten, für die vom jeweiligen Sprachmodell Embeddings erstellt werden. Auch hier wird zwischen zwei Konzepten (z. B. männlich, weiblich) eine Ebene im mehrdimensionalen Raum identifiziert, die im Anschluss von den Ausgaben der Embedding-Schicht des Modells subtrahiert wird, um so den intrinsischen Bias zu verringern.

Die beschriebenen Methoden zur Messung von Bias und zum Debiasing wurden in einen Demonstrator integriert, bei dem Modelle, Methoden und Methoden-bezogene Templates ausgewählt werden können (siehe 2.5).

2.4 Generierung lokaler Erläuterungen zu Klassifikationsergebnissen

Lokale Erläuterungen von zur Klassifikation fähigen Sprachmodellen werden in der Regel durch Attributionen generiert. Diese erlauben die Berechnung des Einflusses von Eingabefaktoren (Wörter bzw. Tokens) auf die Ausgaben eines Modells. Dazu bestehen verschiedene Algorithmen, die sogenannte Attributions berechnen und dazu entweder auf die Gradienten und Modellparameter zurückgreifen, oder direkt interpretierbare Surrogatmodelle trainieren, die Erklärungen für die Klassifikation einzelner Beispiele liefern.

Es wurde eine Vielzahl an Algorithmen anhand unserer Modelle getestet, um die Qualität der lokalen Erläuterungen zu evaluieren. Zu den Algorithmen zählen unter anderem SHAP [14],

Integrated Gradients [15], Saliency [16], DeepLIFT [17] und LIME [18]. Ein Beispiel für die lokale Erläuterungen des BERT-Modells zur Klassifikation der Delikttypen ist in Abbildung 3 dargestellt. Die Algorithmen berechnen die Attribution von Tokens auf die binäre Klassifikation eines Labels. Obwohl das Modell eine Multilabel-Klassifikation generiert, beziehen sich die Attributionen hier nur auf das Label *Drogen*. Grün markierte Tokens haben einen positiven Einfluss auf die Vorhersage, dass Drogen im Text vorhanden sind, rote haben dagegen einen negativen Einfluss. Die Farbintensität deutet auf die Intensität der Attribution hin. In diesem visualisierten Vergleich von drei Algorithmen ist zu sehen, dass die Algorithmen jeweils stark abweichende Attributionen berechnen. Insofern stellt sich die Frage, welcher Algorithmus bzw. welche Erläuterung für ein Beispiel und ein Modell ausgewählt werden sollte.

DeepLIFT

[CLS]Polizeibeamte haben am Mittwoch (23.03.2022) einen 29-jährigen Mann vorläufig festgenommen, der im Verdacht steht, mit Rauschgift gehandelt zu haben. Im Rahmen eines anderen Ermittlungsverfahrens ergab sich der Verdacht, dass der 29-Jährige im letzten Jahr mit Marihuana gehandelt haben soll. Bei der Wohnungsdurchsuchung fanden die Beamten neben 20 Marihuanapflanzen, über zwei Kilogramm Marihuana sowie mehrere Hundert Euro vermutliches Dealergeld. Der Mann wurde nach Abschluss der polizeilichen Maßnahmen wieder auf freien Fuß gesetzt.[SEP]

Integrated Gradients

[CLS]Polizeibeamte haben am Mittwoch (23.03.2022) einen 29-jährigen Mann vorläufig festgenommen, der im Verdacht steht, mit Rauschgift gehandelt zu haben. Im Rahmen eines anderen Ermittlungsverfahrens ergab sich der Verdacht, dass der 29-Jährige im letzten Jahr mit Marihuana gehandelt haben soll. Bei der Wohnungsdurchsuchung fanden die Beamten neben 20 Marihuanapflanzen, über zwei Kilogramm Marihuana sowie mehrere Hundert Euro vermutliches Dealergeld. Der Mann wurde nach Abschluss der polizeilichen Maßnahmen wieder auf freien Fuß gesetzt.[SEP]

GradientSHAP

[CLS]Polizeibeamte haben am Mittwoch (23.03.2022) einen 29-jährigen Mann vorläufig festgenommen, der im Verdacht steht, mit Rauschgift gehandelt zu haben. Im Rahmen eines anderen Ermittlungsverfahrens ergab sich der Verdacht, dass der 29-Jährige im letzten Jahr mit Marihuana gehandelt haben soll. Bei der Wohnungsdurchsuchung fanden die Beamten neben 20 Marihuanapflanzen, über zwei Kilogramm Marihuana sowie mehrere Hundert Euro vermutliches Dealergeld. Der Mann wurde nach Abschluss der polizeilichen Maßnahmen wieder auf freien Fuß gesetzt.[SEP]

Abbildung 3: Vergleich von drei lokalen Erläuterungen anhand eines Modells und Texts bei der Klassifikation des Labels 'Drogen'. Rote Markierungen signalisieren negativen Einfluss auf das Label, grüne Markierungen signalisieren positive Attributionen auf das Label.

Der Vergleich von Erklärungen und Erklärbarkeitsalgorithmen kann anhand verschiedener Eigenschaften vorgenommen werden. Dazu zählt unter anderem die Komplexität der Berechnung, die je nach Modellkomplexität und verfügbarer Ressourcen großen Einfluss auf die Geschwindigkeit hat. Zusätzlich dazu bestehen Algorithmen zur Evaluation der Erklärungsgüte mittels verschiedener Metriken. Dazu zählen z. B. Faithfulness [19], Truthfulness [20] oder Ranked Faithful Truthfulness [21]. Diese berechnen z. B. die Stabilität der Ausgaben der Erklärbarkeitsalgorithmen, wenn Änderungen der Eingabesequenzen

vorgenommen werden. Je nach Anwendungsdomäne, Daten, Modell und Algorithmus sind deshalb jeweils Vergleiche verschiedener Algorithmen durchzuführen, um die für den Kontext optimal passenden Algorithmen zu identifizieren.

2.5 Demonstrator-Entwicklung

Die Entwicklung von Demonstrationsoberflächen, mit denen Anwender die Daten, Modelle, Algorithmen, Erklärungen und Visualisierungen erkunden können, wurde in diesem Teilprojekt auf zwei voneinander unabhängige Web-basierte Demonstratoren aufgeteilt. Der erste Information-Extraction-Demonstrator ist für die Exploration der Textklassifikation und Generierung von lokalen Erläuterungen konzipiert. Der zweite Bias-Demonstrator bietet dagegen ausschließlich Funktionen zur Evaluation und Mitigation von intrinsischem und extrinsischem Bias über eine simplere Web-Oberfläche an.

Information-Extraction-Demonstrator

Der Information-Extraction-Demonstrator des Teilprojekts verfügt über mehrere Ansichten für verschiedene Anwendungen.

Die erste Anwendung ist zur Analyse und Inferenz des Presseportal-Datensatzes mittels der *Semantischen Modellierung* konzipiert. In Abbildung 4 auf Seite 14 ist ein Bildschirmfoto der Nutzerschnittstelle zu sehen. Nutzer können zunächst den Datensatz und den Polizeibericht auswählen. Zusätzlich dazu werden entweder die annotierten Labels angezeigt oder die Modellvorhersagen ad-hoc angefragt und in das Nutzerinterface geladen. Für jede der drei Aufgabenstellungen des Presseportal-Datensatzes, *Delikttypklassifikation*, *Named Entity Recognition* und *Relationship Extraction* besteht eine Komponente zur Visualisierung der Annotationen bzw. Modellvorhersagen.

Die erste Komponente zeigt die *Delikttypklassifikation*, bei der die Label für die vier Delikttypen entweder als wahr oder falsch oder im Falle von Modellvorhersagen als Ausgabe-wahrscheinlichkeiten angezeigt werden.

Die zweite Komponente dient der *Named Entity Recognition*, die den Haupttext des Polizeiberichts inklusive der markierten Entitäten anzeigt. Die nebenstehende Legende erlaubt das Auswählen einzelner Entitätstypen, die daraufhin im Polizeibericht besonders herausgestellt werden, um den Nutzern ein schnelles Navigieren und Verstehen des Inhalts zu ermöglichen.

In der dritten Komponente für die *Relationship Extraction* werden die Entitäten und deren Beziehungen zueinander in Graphenform visualisiert. Jede Entität gilt dabei als ein Knoten und jede Beziehung als eine Kante zwischen zwei Knoten. In der aggregierten Ansicht werden Entitäten, die durch COREF-Relationen miteinander verbunden sind, zu einem Cluster zusammengefasst, um weitergehende Relationen zu dem Cluster zuzuordnen und die Komplexität somit zu verringern.

Weiterhin ist es möglich, Knoten in der Graphen-Komponente oder Entitäten in der Named-Entity-Recognition-Komponente auszuwählen, woraufhin eine Detailansicht der Entität geöffnet wird, in der weitere extrahierte Informationen angezeigt werden.

Die letzten beiden Komponenten zeigen die Metadaten eines Polizeiberichts und eine interaktive Karte mit den im Bericht vorkommenden Orten an. Für die Visualisierung der Orte werden die extrahierten *Location*-Entitäten einer Geolocation-API übergeben, die zu gefundenen Orten entsprechende Detailinformationen und Koordinaten bereitstellt.

Semantische Modellierung

test_data 43 / 51

Select a Crime Type Annotation: 109 (prediction) Delete Predict Crime Types | Select an NER Annotation: 41 (dataset) Delete Predict NER | Select a Relation Extraction Annotation: 41 (dataset) Delete Predict Relations

Text Classification

🔪 Betäubungsmittel **98.77%** |
 🔫 Waffen **0.96%** |
 👁️ Anderes **35.17%** |
 🚨 Gefahr für Leib und Leben **3.30%**

Named Entity Recognition

Die **Polizei** hat **am Mittwochvormittag**, **20.10.2021**, im Bereich **Bunzt** den **betrunkenen** Fahrer **eines** Kleintransporters angehalten. Zeugen war zuvor **der schwankende Gang** des Mannes **auf dem Weg zum Fahrzeug** aufgefallen. Kurz darauf **gegen 12:10 Uhr** hielt **eine** Polizeistreife den **gesuchten** Transporter an der **Peter-Krall-Straße** an. **Ein** **freiwilliger** **Atemalkoholtest** bestätigte den Verdacht, dass **der 29-Jährige** **erheblich alkoholisiert** war. Außerdem besitzt er **keine** **gültige** **Fahrerlaubnis**. **Ein** **Arzt** entnahm **dem 29-Jährigen** **eine** **Blutprobe**. Die **Polizei** hat gegen **ihn** **Strafanzzeige** wegen Trunkenheit im Verkehr und Fahren ohne Fahrerlaubnis gestellt. Zusätzlich wird gegen **die Halterin** des **Fahrzeugs** wegen des Zulassens des Fahrens ohne Fahrerlaubnis ermittelt.(j)

TIME

COUNT

REF_PERSON

LOCATION

OBJECT

PROPERTY

ORGANISATION

Relation Extraction

Toggle Aggregation | 🔄 Reset Layout | 🔍 Reset Zoom

Metadata

Headline: POL-MG: Polizei stoppt betrunkenen Transporter-Fahrer
 URL: <http://www.presseportal.de/blaulicht/pm/30127/5051739>
 Date: 2021-10-20 16:19:53
 Police: Polizei Mönchengladbach
 State: Nordrhein-Westfalen
 Locations: ["Mönchengladbach"]
 Tags: Fahren unter Einfluss psychoaktiver Substanzen, Blutentnahme, Führerschein, POL, Strafanzzeige, Fahrer, Fahren ohne Fahrerlaubnis, MG, Transport, Sicherheitskräfte, Kriminalität, Transportunglück, Gesetz, Drogenkriminalität, Justiz, Kriminalität, Essen und Trinken, Polizei, Polizei Nordrhein-Westfalen

Map

Abbildung 4: Information Extraction Demonstrator zur Semantischen Modellierung von Polizeiberichten

Neben der Anwendung zur *Semantischen Modellierung* bietet der Information Extraction Demonstrator eine Oberfläche zur Analyse des Sexual Predator Identification Datensatzes. Hier können die Beispiele des Datensatzes in einem Chat-Interface durchsucht werden.

Die dritte Anwendung dieses Demonstrators dient dem Experimentieren mit im Projekt verwendeten Erklärbarkeitsalgorithmen. Unter Verwendung eines Sprachmodells zur binären Klassifikation von Stimmungen in Texten, können Erklärbarkeitsalgorithmen ausgewählt, deren Parameter eingestellt, Texte eingegeben und schließlich durch die Berechnung der Erklärungen mit den Algorithmen experimentiert werden. Nutzer bekommen die durch die Attributionsalgorithmen generierten Attributionsen auf Token-Ebene visualisiert, um den Einfluss der Tokens auf die jeweilige Klassifikation nachzuvollziehen. Weiterhin werden

Algorithmen zum Vergleich der Erklärbarkeitsalgorithmen bzw. deren Ergebnisse anhand des gewählten Beispiels berechnet und in einer sortierbaren Tabelle angezeigt. Die Ansicht des Demonstrators zur Untersuchung von Erklärbarkeitsalgorithmen ist in Abbildung 5 dargestellt.

VIKING Demonstrator

- Semantische Modellierung
- Sexual Predator Identification
- Attribution

Attribution

Input: The movie was extremely boring.

Compute Attributions

Algorithms

- Deconvolution
- DeepLiftSHAP
- DeepLift
- Gradient SHAP
- Guided Backprop
- Guided GradCam
- InputXGradient
- Layer Integrated Gradients
- Integrated Gradients
- Saliency

Parameters

All Algorithms

Model Label: NEGATIVE POSITIVE

Threshold: 0

Deconvolution, DeepLift, Integrated Gradients, Saliency

Explanations

| Algorithm | Attributions | Faithfulness | Truthfulness | Faithful Truthfulness | Ranked Faithful Truthfulness | Non Zero Weights |
|----------------------|---|--------------|--------------|-----------------------|------------------------------|------------------|
| Saliency | [CLS] the movie was extremely boring. [SEP] | 0.0555 | 0.8333 | 0.0556 | 0.0555 | 1 |
| Integrated Gradients | [CLS] the movie was extremely boring. [SEP] | 0.0555 | 0.8333 | 0.0555 | 0.0555 | 1 |
| Deconvolution | [CLS] the movie was extremely boring. [SEP] | 0.0001 | 0.3333 | -0.0555 | -0.0555 | 1 |
| DeepLift | [CLS] the movie was extremely boring. [SEP] | 0.0001 | 0.5 | -0.0554 | -0.0554 | 1 |

Abbildung 5: Demonstrator zum Experimentieren mit Erklärbarkeitsalgorithmen

Bias-Demonstrator

Dieser webbasierte Demonstrator erlaubt es einem Nutzer, Bias-Evaluations- oder Debiasing-Algorithmen anhand ausgewählter oder selbst erstellter Templates und Modelle zu testen. Hierfür gibt es für jeden Algorithmus eine Seite, auf der neben detaillierten Erläuterungen des Algorithmus zunächst eine Auswahl der Templates und Modelle, sowie eine Konfiguration möglicher Parameter verfügbar ist. Sämtliche Beispiele in den Templates können durchsucht werden, um die Bedeutung des Tests nachzuvollziehen. Mit der gewählten Konfiguration können die Ergebnisse ad-hoc berechnet und visualisiert werden. Zu den Ergebnissen bestehen Erläuterungen zur erleichterten Interpretierbarkeit der Ergebnisse. In Abbildung 6 ist die Oberfläche des Bias-Demonstrators und ein Ergebnis eines CrowS-Pairs-Tests zu sehen.

Der Demonstrator bietet drei Methoden für die Evaluation von intrinsischem Modell-Bias, eine für die Evaluation von extrinsischem Bias und eine Methode zum Debiasing von intrinsischem Bias an.

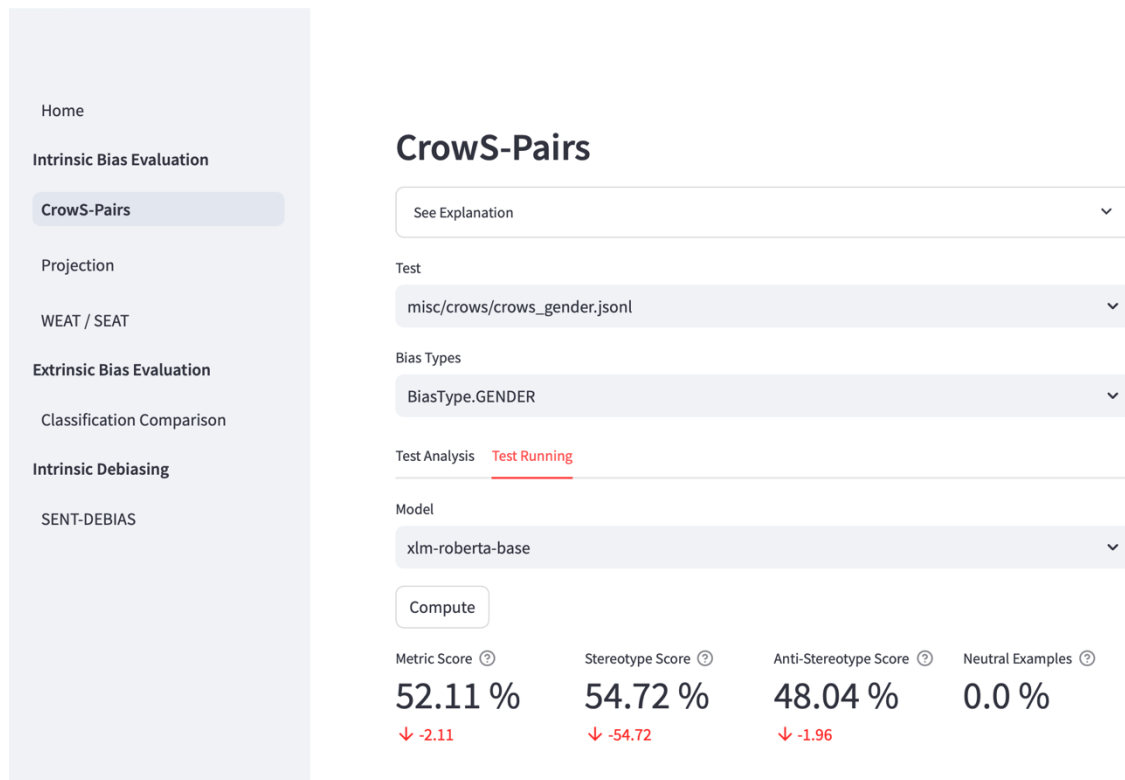


Abbildung 6: Bias-Demonstrator: Darstellung der Ergebnisse des CrowS-Pairs Tests bzgl. Geschlechts-Bias.

2.6 DIN-SPEC

Im Rahmen der Standardisierungsaktivitäten des Arbeitspakets 8 des VIKING-Konsortiums wurde die Entwicklung einer DIN-SPEC initialisiert. Auch die UniBw M hat an dieser Initiative teilgenommen und Anforderungen für die Vertrauenswürdigkeit im Kontext der Textauswertung entwickelt und beigetragen, die sich auf technische, ethische und rechtliche Aspekte fokussieren. So wurde im Jahr 2025 das DIN-SPEC-Dokument veröffentlicht [22].

3 Wichtigste Positionen des zahlenmäßigen Nachweises

Im Projekt VIKING waren die wesentlichen Kosten die Personalkosten (0812 und 0820). Zusätzliche Kosten entstanden gemäß Gesamtfinanzierungsplan durch Bewirtungs- und Open-Access-Kosten (beides 0843) sowie Reisen zu Verbundtreffen und Konferenzen (beides 0846). Entsprechende Reiseberichte liegen vor.

4 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Der Verlauf der Arbeit im Projekt folgte grundsätzlich der im Projektantrag formulierten Planung. Durch die um einen Monat verzögerte Personaleinstellung und die Herausforderungen der Datenbeschaffung, gab es leichte Verzögerungen, weshalb eine kostenneutrale Verlängerung des Projekts um drei Monate stattfand. Dabei wurden alle im Arbeitsplan formulierten Aufgaben erfolgreich bearbeitet.

5 Stellungnahme bezüglich der wirtschaftlichen Verwertbarkeit

Die im Projekt entwickelten Verfahren, Prototypen bzw. Erkenntnisse wurden, sofern möglich, veröffentlicht. Da insbesondere die Prototypen auf Stellvertreterdaten bzw. entsprechend angepassten Problemstellungen basieren, ist ein direkter Einsatz im polizeilichen Kontext nicht möglich und nicht geplant. Es besteht die Möglichkeit einer Adaption und Wiederverwendung verschiedenster Komponenten.

6 Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Es sind keine Fälle bekannt geworden, wo vergleichbare Arbeiten parallel zu den im Projekt durchgeführten Aufgaben mit vergleichbaren Resultaten bei anderen Stellen durchgeführt wurden.

7 Veröffentlichungen

1. Kersting, J., Maoro, F., & Geierhos, M. (2023). Towards comparable ratings: Exploring bias in German physician reviews. *Data & Knowledge Engineering*, 148, 102235. <https://doi.org/10.1016/j.datak.2023.102235>
2. Maoro, F., Vehmeyer, B., & Geierhos, M. (2024). Leveraging Semantic Search and LLMs for Domain-Adaptive Information Retrieval. In A. Lopata, D. Gudonienė, & R. Butkienė (Hrsg.), *Information and Software Technologies* (S. 148–159). Springer Nature. https://doi.org/10.1007/978-3-031-48981-5_12
3. Maoro, F., & Geierhos, M. (2024). Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen: Wie kann Künstliche Intelligenz in der Polizeiarbeit unterstützend eingesetzt werden und dabei sowohl fair als auch nachvollziehbar sein? *Kongress KI@HSBI2023 Solutions im Fokus*, 1, 22–23. <https://doi.org/10.60802/sidas.2024.1>
4. Maoro, F., & Geierhos, M. (2024). FICODE at GermEval 2024 GerMS-Detect closed ST1 & ST2: Ensemble- and Transformer-Based Detection of Sexism and Misogyny in German Texts. In B. Krenn, J. Petrak, & S. Gross (Hrsg.), *Proceedings of GermEval 2024 Task 1 GerMS-Detect Workshop on Sexism Detection in German Online News Fora (GerMS-Detect 2024)* (S. 21–25). Association for Computational Linguistics. <https://aclanthology.org/2024.germeval-2.3/>
5. Fischer, M. T., Schlegel, U., Keim, D. A., Altmann, S., Grote, C., Reuter, P., Coleman, G., Geierhos, M., Maoro, F., Kluin, M., Weinbruch, M., Aden, H., Kleemann, S., Tahraoui, M., Louban, A., Arndt, M., Schönrock, S., Brandner, L. T., Hirsbrunner, S. D., ... Yilmaz, Yusuf. (2025). *Anforderungen an vertrauenswürdige KI-Methoden in polizeilichen Anwendungen* (No. DIN SPEC 91517:2025-05). DIN Media GmbH. <https://doi.org/10.31030/3612025>
6. Maoro, F., & Geierhos, M. (2025). Contestable AI for criminal intelligence analysis: Improving decision-making through semantic modeling and human oversight. *Frontiers in Artificial Intelligence*, 8. <https://doi.org/10.3389/frai.2025.1602998>

III Erfolgskontrollbericht (nicht öffentlich)

– separate Anlage beim Projektträger –

IV Berichtsblatt

| | | |
|--|--|---------------------------------------|
| 1. ISBN oder ISSN | 2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht | |
| 3. Titel Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen (VIKING) – Teilvorhaben: Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation | | |
| 4. Autor(en) [Name(n), Vorname(n)] Geierhos, Michaela Maoro, Falk | 5. Abschlussdatum des Vorhabens März 2025 | 6. Veröffentlichungsdatum |
| | 7. Form der Publikation | |
| | 8. Durchführende Institution(en) (Name, Adresse) Universität der Bundeswehr München Forschungsinstitut CODE Werner-Heisenberg-Weg 39 85579 Neubiberg | |
| 12. Fördernde Institution (Name, Adresse) Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) 53170 Bonn | | 9. Ber. Nr. Durchführende Institution |
| | | 10. Förderkennzeichen 13N16244 |
| | | 11. Seitenzahl 22 |
| 16. Zusätzliche Angaben | | 13. Literaturangaben 22 |
| | | 14. Tabellen 9 |
| | | 15. Abbildungen 5 |
| 17. Vorgelegt bei (Titel, Ort, Datum) | | |
| 18. Kurzfassung Die polizeiliche Praxis zeigt, dass der Einsatz von Künstlicher Intelligenz (KI) einerseits die Ermittlungen beschleunigen und vereinfachen kann, andererseits aber auch verschiedene Risiken für den Ermittlungserfolg beinhaltet. So können unausgewogene Trainingsdatensätze, in denen beispielsweise demographische Häufigkeiten verzerrt abgebildet sind (Bias), zu fehlerhaften Ergebnissen von KI-Lösungen führen. Ein weiteres Risiko besteht in der fehlenden Nachvollziehbarkeit und mangelnden Transparenz der Ergebnisse komplexer KI. Die Europäische Kommission hat eine Verordnung für vertrauenswürdige KI vorgeschlagen. Die darin erhobenen Anforderungen an Genauigkeit, Nachvollziehbarkeit und Robustheit von KI-Systemen werfen aber erhebliche wissenschaftlich-technische Fragen auf: Wie können Anforderungen technisch realisiert, rechtlich-ethisch sichergestellt und objektiv gemessen werden? Deshalb ist das wissenschaftlich-technische Gesamtziel von VIKING die Erforschung und Implementierung von Lösungen zur Messung und Optimierung der Genauigkeit (Debiasing), Nachvollziehbarkeit (Erklärbarkeit) und Robustheit (Angriffsfestigkeit) zum Einsatz vertrauenswürdiger KI in der polizeilichen Anwendung. Das Teilvorhaben „Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation“ widmet sich der Erforschung vertrauenswürdiger KI-Methoden zur Textklassifikation und komplementiert somit die Forschung zur Gesichts- und Sprechererkennung sowie zur Objektdetektion der anderen Teilvorhaben von VIKING. Das wissenschaftlich-technische Ziel dieses Teilvorhabens ist vor allem die Erforschung und Implementierung von Lösungen zur Herstellung von Nachvollziehbarkeit (Erklärbarkeit) bei der KI-basierten Textauswertung in der polizeilichen Anwendung, um letztendlich mehr Transparenz beim Anwender zu schaffen. | | |
| 19. Schlagwörter Vertrauenswürdige Künstliche Intelligenz, Nachvollziehbarkeit, Angriffsfestigkeit, Debiasing, Transparenz, Standardisierung | | |
| 20. Verlag | 21. Preis | |

V Document Control Sheet

| | | |
|--|---|-------------------------------|
| 1. ISBN or ISSN | 2. type of document (e.g. report, publication) Report | |
| 3. title Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen (VIKING) – Teilvorhaben: Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation | | |
| 4. author(s) (family name, first name(s)) Geierhos, Michaela Maoro, Falk | 5. end of project March 2025 | 6. publication date |
| | 7. form of publication | |
| | 8. performing organization(s) (name, address) Universität der Bundeswehr München Forschungsinstitut CODE Werner-Heisenberg-Weg 39 85579 Neubiberg | |
| 12. sponsoring agency (name, address) Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) 53170 Bonn | | 9. originator's report no. |
| | | 10. reference no. 13N16244 |
| | | 11. no. of pages 22 |
| 16. supplementary notes | | 13. no. of references 22 |
| | | 14. no. of tables 9 |
| | | 15. no. of figures 5 |
| 17. presented at (title, place, date) | | |
| 18. abstract Police practice shows that the use of artificial intelligence (AI) can accelerate and simplify investigations on the one hand, but also entails various risks for the success of investigations on the other. For example, unbalanced training datasets in which demographic frequencies are distorted (bias) can lead to incorrect results from AI solutions. Another risk is the lack of traceability and transparency of the results of complex AI. The European Commission has proposed a regulation for trustworthy AI. However, the requirements it sets out for the accuracy, traceability, and robustness of AI systems raise significant scientific and technical questions: How can requirements be technically implemented, legally and ethically ensured, and objectively measured? Therefore, the overall scientific and technical goal of VIKING is to research and implement solutions for measuring and optimizing accuracy (debiasing), traceability (explainability), and robustness (attack resistance) for the use of trustworthy AI in police applications. The subproject “Erklärbarkeit vertrauenswürdiger KI-Sprachmodelle für den transparenten Gebrauch bei Sicherheitsbehörden zur Textklassifikation” is dedicated to researching trustworthy AI methods for text classification and thus complements the research on face and speaker recognition as well as object detection in the other VIKING subprojects. The scientific and technical goal of this subproject is primarily to research and implement solutions for establishing traceability (explainability) in AI-based text evaluation in police applications, with the ultimate aim of creating greater transparency for users. | | |
| 19. keywords trustworthy artificial intelligence, traceability, resilience to attacks, debiasing, transparency, standardization | | |
| 20. publisher | 21. price | |

VI Literaturverzeichnis

- [1] G. Inches und F. Crestani, „Overview of the international sexual predator identification competition at PAN-2012“, in *CLEF 2012 evaluation labs and workshop – working notes papers, 17-20 september, rome, italy*, P. Forner, J. Karlgren, und C. Womser-Hacker, Hrsg., CEUR-WS.org, Sep. 2012. [Online]. Verfügbar unter: <http://www.clef-initiative.eu/publication/working-notes>
- [2] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y-Gómez, und L. Villaseñor-Pineda, „Two-step approach for effective detection of misbehaving users in chats—notebook for PAN at CLEF 2012“, in *CLEF 2012 evaluation labs and workshop – working notes papers, 17-20 september, rome, italy*, P. Forner, J. Karlgren, und C. Womser-Hacker, Hrsg., CEUR-WS.org, Sep. 2012. [Online]. Verfügbar unter: <http://ceur-ws.org/Vol-1178>
- [3] J. Devlin, M.-W. Chang, K. Lee, und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, und T. Solorio, Hrsg., Minneapolis, Minnesota: Association for Computational Linguistics, Juni 2019, S. 4171–4186. doi: 10.18653/v1/N19-1423.
- [4] E. F. Tjong Kim Sang und F. De Meulder, „Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition“, in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, S. 142–147. Zugegriffen: 27. Februar 2025. [Online]. Verfügbar unter: <https://aclanthology.org/W03-0419/>
- [5] I. Yamada, A. Asai, H. Shindo, H. Takeda, und Y. Matsumoto, „LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention“, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, und Y. Liu, Hrsg., Online: Association for Computational Linguistics, Nov. 2020, S. 6442–6454. doi: 10.18653/v1/2020.emnlp-main.523.
- [6] P. Delobelle, E. Tokpo, T. Calders, und B. Berendt, „Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models“, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, und I. V. Meza Ruiz, Hrsg., Seattle, United States: Association for Computational Linguistics, Juli 2022, S. 1693–1706. doi: 10.18653/v1/2022.naacl-main.122.
- [7] Y. Liu u. a., „RoBERTa: A Robustly Optimized BERT Pretraining Approach“, *CoRR*, Bd. abs/1907.11692, 2019, [Online]. Verfügbar unter: <http://arxiv.org/abs/1907.11692>
- [8] N. Sevim, F. Şahinuç, und A. Koç, „Gender bias in legal corpora and debiasing it“, *Nat. Lang. Eng.*, Bd. 29, Nr. 2, S. 449–482, 2023, doi: 10.1017/S1351324922000122.
- [9] A. Caliskan, J. J. Bryson, und A. Narayanan, „Semantics derived automatically from language corpora contain human-like biases“, *Science*, Bd. 356, Nr. 6334, S. 183–186, 2017, doi: 10.1126/science.aal4230.
- [10] C. May, A. Wang, S. Bordia, S. R. Bowman, und R. Rudinger, „On Measuring Social Biases in Sentence Encoders“, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, und T. Solorio, Hrsg., Minneapolis, Minnesota: Association for Computational Linguistics, Juni 2019, S. 622–628. doi: 10.18653/v1/N19-1063.

- [11] N. Nangia, C. Vania, R. Bhalerao, und S. R. Bowman, „CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models“, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, und Y. Liu, Hrsg., Online: Association for Computational Linguistics, Nov. 2020, S. 1953–1967. doi: 10.18653/v1/2020.emnlp-main.154.
- [12] J. Kersting, F. Maoro, und M. Geierhos, „Towards comparable ratings: Exploring bias in German physician reviews“, *Data Knowl. Eng.*, Bd. 148, S. 102235, 2023, doi: <https://doi.org/10.1016/j.datak.2023.102235>.
- [13] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, und L.-P. Morency, „Towards Debiasing Sentence Representations“, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schlueter, und J. Tetreault, Hrsg., Online: Association for Computational Linguistics, Juli 2020, S. 5502–5515. doi: 10.18653/v1/2020.acl-main.488.
- [14] S. M. Lundberg und S.-I. Lee, „A unified approach to interpreting model predictions“, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, S. 4768–4777.
- [15] M. Sundararajan, A. Taly, und Q. Yan, „Axiomatic Attribution for Deep Networks“, in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, Juli 2017, S. 3319–3328. Zugegriffen: 19. Dezember 2024. [Online]. Verfügbar unter: <https://proceedings.mlr.press/v70/sundararajan17a.html>
- [16] K. Simonyan, A. Vedaldi, und A. Zisserman, „Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps“. 2014. [Online]. Verfügbar unter: <https://arxiv.org/abs/1312.6034>
- [17] A. Shrikumar, P. Greenside, und A. Kundaje, „Learning important features through propagating activation differences“, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, in ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, S. 3145–3153.
- [18] M. Ribeiro, S. Singh, und C. Guestrin, „“Why Should I Trust You?”: Explaining the Predictions of Any Classifier“, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, J. DeNero, M. Finlayson, und S. Reddy, Hrsg., San Diego, California: Association for Computational Linguistics, Juni 2016, S. 97–101. doi: 10.18653/v1/N16-3020.
- [19] M. Du, N. Liu, F. Yang, S. Ji, und X. Hu, „On Attribution of Recurrent Neural Network Predictions via Additive Decomposition“, in *The World Wide Web Conference*, in WWW ’19. New York, NY, USA: Association for Computing Machinery, Mai 2019, S. 383–393. doi: 10.1145/3308558.3313545.
- [20] I. Mollas, N. Bassiliades, und G. Tsoumakas, „LioNets: a neural-specific local interpretation technique exploiting penultimate layer information“, *Appl. Intell.*, Bd. 53, Nr. 3, S. 2538–2563, Feb. 2023, doi: 10.1007/s10489-022-03351-4.
- [21] N. Mylonas, I. Mollas, und G. Tsoumakas, „An attention matrix for every decision: faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification“, *Data Min. Knowl. Discov.*, Bd. 38, Nr. 1, S. 128–153, Jan. 2024, doi: 10.1007/s10618-023-00962-4.
- [22] M. T. Fischer u. a., *Anforderungen an vertrauenswürdige KI-Methoden in polizeilichen Anwendungen*, DIN SPEC 91517:2025-05, Berlin., 2025. doi: 10.31030/3612025.