

DA-VBB: Digitaler Assistent zum Verstehen Behördlicher Bescheide

Vorhabensbeschreibung zum studentischen Forschungsvorhaben
im Rahmen des Software Campus

Anna Filighera

7. Juni 2021

Antragssteller (Projektrahmen-Vorhaben)	Technische Universität Darmstadt
Projektleitung (Mikroprojekt)	Anna Filighera Technische Universität Darmstadt Fachbereich Elektrotechnik und Informationstechnik Multimedia Communications Lab - KOM Rundeturmstr. 10 64283 Darmstadt anna.filighera@kom.tu-darmstadt.de + 49 (6151) 16 - 20466
Akademischer Betreuer	Prof. Dr.-Ing. Ralf Steinmetz Technische Universität Darmstadt ralf.steinmetz@kom.tu-darmstadt.de
Industriepartner	Dr. Peter Schichtel IAV GmbH Trippstadterstraße 122, 67663 Kaiserslautern, Germany peter.schichtel@iav.de
Beginn des Mikroprojektes Laufzeit des Mikroprojektes	1. April 2021 2 Jahre

Inhaltsverzeichnis

1	Aufgabenstellung und Motivation	3
1.1	Schwerpunkte und Ziele	3
1.2	Wissenschaftliche und/oder technische Ziele des Vorhabens	4
1.3	Bezug des Vorhabens zu förderpolitischen Zielen / Förderprogramm	5
2	Stand der Wissenschaft und Technik	6
2.1	Feedbacksysteme	6
2.2	NLP mit deutschen Rechtsdokumenten	7
3	Partner und bisherige Arbeiten	8
3.1	Universität/Forschungseinrichtung	8
3.2	Unternehmen	9
3.3	Beziehung Universität - Unternehmen	9
4	Ausführliche Beschreibung des Arbeitsplans	10
4.1	Arbeitspaket 1: Projektmanagement	10
4.2	Arbeitspaket 2: Erstellung eines Korpus aus Verwaltungsakten	10
4.3	Arbeitspaket 3: Erstellung eines Fragenkatalogs	11
4.4	Arbeitspaket 4: Analyse und Evaluation des erstellten Korpus	12
4.5	Arbeitspaket 5: Konzeption des Feedbackmodells	12
4.6	Arbeitspaket 6: Implementation des Kategorisierung- und Feedbackmodells	13
4.7	Arbeitspaket 7: Evaluation des Kategorisierung- und Feedbackmodells	13
4.8	Zeitplanung und Meilensteine	14
4.9	Finanzplanung	14
4.9.1	Geplante Personalmittel	15
4.9.2	Geplante Reisemittel	16
4.9.3	Sonstige allgemeine Verwaltungsausgaben	17
5	Verwertungsplan	18
5.1	Wirtschaftliche Erfolgsaussichten	18
5.2	Wissenschaftlich-technische Erfolgsaussichten	18
5.3	Wissenschaftliche und wirtschaftliche Anschlussfähigkeit	19

1 Aufgabenstellung und Motivation

Fragen und Antworten sind seit jeher ein zentraler Bestandteil der Wissensakquise. Sei es um Verständnis in Prüfungen zu ermitteln oder Denkprozesse anzuregen, Lernenden Fragen zu stellen und ihre Antworten zu evaluieren ist ein essenzieller Bestandteil des pädagogischen Alltags. Die Evaluation von Antworten durch Experten kann allerdings sehr zeit- und kostenaufwendig sein, weshalb sich in den letzten Jahren oft einfach korrigierbare Aufgabenformate, wie z. B. Multiple-Choice Fragen, durchgesetzt haben. Diese geschlossenen Aufgabenformate sind jedoch zeitaufwendig zu erstellen und auf Grund ihres limitierten Formats nicht für alle Lernszenarien geeignet. Deshalb werden in diesem Projekt automatische Evaluationsmethoden für Freitextantworten, auch als *automatic short answer grading (ASAG)* [1] bekannt, untersucht. Hierbei ist es das Ziel, frei-verfasste Antworten auf inhaltliche Korrektheit, Vollständigkeit und Relevanz zu der gestellten Frage zu prüfen.

Automatic short answer grading ermöglicht es unter anderem, das Verständnis des Lernenden, welches in gegebenen Antworten zum Ausdruck kommt, mit Referenzmaterial abzugleichen und so Missverständnisse oder Lücken aufzudecken. Das macht es besonders attraktiv als Hilfestellungen für Bürger im Umgang mit Verwaltungsakten. Die mangelnde Verständlichkeit von amtssprachlichen Dokumenten wird schon seit dem 18. Jahrhundert diskutiert [2] und bleibt bis heute ein ernstzunehmendes Problem. Missverständnisse und die daraus resultierenden Handlungen können schwerwiegende Konsequenzen nicht nur für den betroffenen Bürger selbst sondern auch für die Gesellschaft als Ganzes haben. Neben Bußgeldern und dem Durchsetzen rechtswidriger Verwaltungsakte, leidet auch die bürgerseitige Akzeptanz wirksamer Bescheide bei mangelndem Verständnis, insbesondere der im Verwaltungsakt gegebenen Begründung. Ein möglicher Lösungsweg besteht darin, Verwaltungsakte in Leichte Sprache zu übersetzen, wie es im Modellprojekt “Übersetzung von Verwaltungsakten in Leichte Sprache”¹ der Evangelische Stiftung Volmarstein getan wurde. Dies ist allerdings mit enormen manuellem Übersetzungsaufwand verbunden. Ebenso ist das Bereitstellen von Bürgerbeauftragten mit Personalkosten verbunden, die eine ausreichende Unterstützung aller Bürger undurchführbar machen.

1.1 Schwerpunkte und Ziele

In diesem Projekt wird ein zu den bisher besprochenen Lösungen komplementäres Unterstützungssystem entwickelt, welches Bürger automatisiert zu den wichtigen Punkten ihres Verwaltungsaktes befragt, gegebene Antworten evaluiert und entsprechendes Feedback liefert. Dieser Ablauf ist schematisch in Abbildung 1 zu sehen. Hierbei soll der Fokus des Projekts auf der automatischen Textverarbeitung liegen, mit der wichtige Informationen des Bescheides extrahiert und Antworten der Bürger mit den extrahierten Daten abgeglichen werden, um hilfreiches Feedback zu generieren. Da zu unserem besten Wissen kein solches System für deutschsprachige Verwaltungsakte existiert, gilt es in diesem Projekt neuartige Herausforderungen zu überwinden. Zum einen besteht zur Zeit kein Datensatz aus Verwaltungsakten und entsprechenden Fragen und Antworten, um ein solches System zu trainieren. Zum anderen ist es fragwürdig, in wie weit die vortrainierten state-of-the-art Natural Language Processing (NLP) Modelle, wie beispielsweise T5 [3], mit deutscher Verwaltungssprache umgehen können. Darüber hinaus ist das automatische Generieren von korrektem und hilfreichem Feedback zu Freitextantworten, im Gegensatz zu Programmcode und mathematischen Formeln, ein wenig erforschtes Problem.

Hieraus leiten sich die folgenden Forschungsfragen für dieses Projekt ab:

- **F1:** Wie können wir automatisiert alle für den Bürger wichtigen Informationen aus einem Verwaltungsakt und gegebenenfalls verwandten Rechtsdokumenten extrahieren?
- **F2:** Wie können wir mit einem manuell erstellten Fragenkatalog die für das Verständnis des Bescheides nötigen Informationen und Zusammenhänge eindeutig und begrifflich abfragen?

¹Übersetzung von Verwaltungsakten in Leichte Sprache: <https://www.sw-nrw.de/foerderung/geoerderte-projekte/menschen-mit-behinderung/modellprojekt-uebersetzung-von-verwaltungsakten-in-leichte-sprache/>

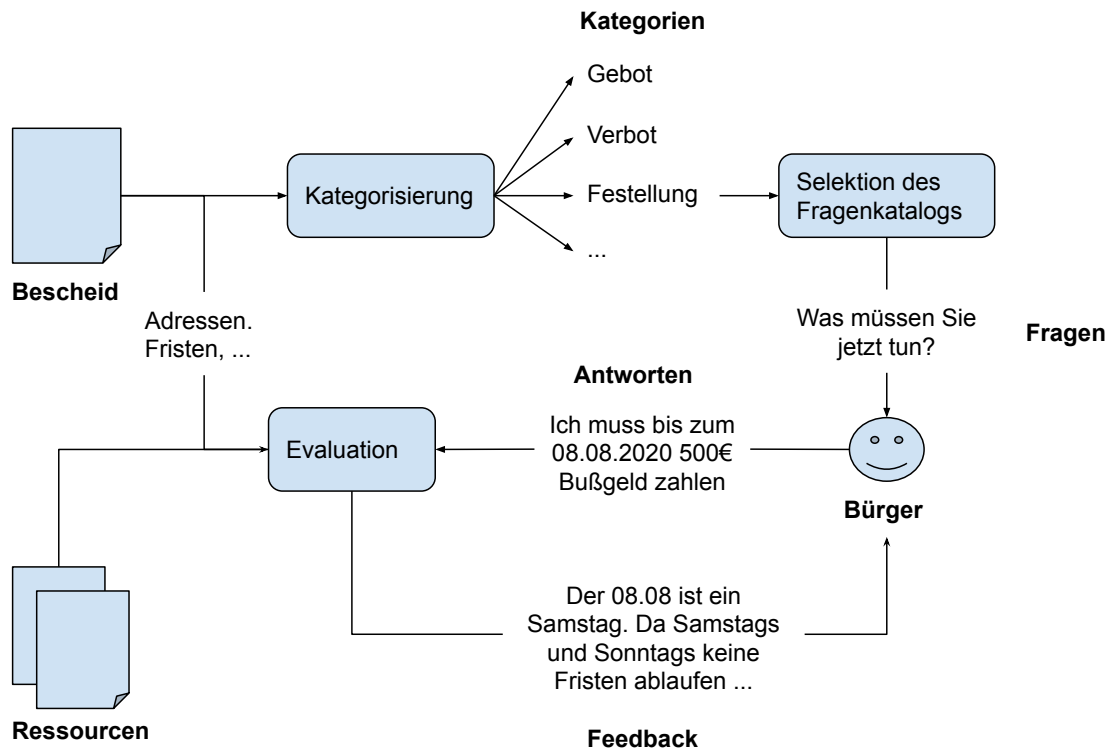


Abbildung 1: Schematische Darstellung der Anwendung zur Inferenzzeit. “Ressourcen” stellen hierbei juristische Dokumente, wie Gesetze oder Verordnungen, dar, die gegebenenfalls für das Verständnis des Bescheids von Bedeutung sind.

- **F3:** Wie können wir automatisiert korrektes und hilfreiches Feedback zu gegebenen Antworten generieren?

1.2 Wissenschaftliche und/oder technische Ziele des Vorhabens

Die übergreifende Vision dieses Vorhabens ist es, eine interaktive App zu schaffen, mit der Bürger erhaltene Verwaltungsakte besser verstehen lernen. Mit dieser App könnten Bürger einen zugestellten Bescheid fotografieren und anschließend die zur weiteren Vorgehensweise nötige Unterstützung erhalten. Dabei errungenes Wissen könnte ihnen, im Kontrast zu Übersetzungen in Leichte Sprache, nicht nur für den konkreten Bescheid selbst sondern für alle künftigen ähnlichen Situationen behilflich sein.

Dieses Projekt dient hier als Proof of Concept, um der App unterliegende Technologien auf ihre Reife und Nutzbarkeit in diesem Szenario zu prüfen (**Z3**) und die notwendigen Voraussetzungen zu schaffen (**Z1, Z2**). Hierbei werden nicht alle für eine solche App notwendigen Gebiete, wie beispielsweise Datenschutz, im Rahmen dieses Projekts thematisiert, sondern der Fokus auf die technische Umsetzbarkeit gelegt. Die in diesem Projekt definierten Arbeitspakete lassen sich den folgenden Zielen zuordnen:

- **Z1:** Erstellung eines Korpus aus Verwaltungsakten, die manuell mit einer Kategorie (Verbot, Ver-

sagung, Feststellung, etc.) versehen werden. Hier existiert schon eine kleine Sammlung von Musterbescheiden der Nürnberger Bundesagentur für Arbeit, die es zu annotieren und nach Möglichkeit zu erweitern gilt.

- **Z2:** Manuelle Erstellung eines kategorieabhängigen Fragenkatalogs, der für das Verständnis des Bescheides kritische Aspekte klar und verständlich abfragt. Zunächst ist hier eine Anforderungsanalyse vorgesehen, in der ermittelt wird, welche Aspekte eines Bescheides für den Bürger kritisch sind und wie diesbezügliche Fragen formuliert werden müssen, um leicht verständlich und eindeutig beantwortbar zu sein. Darüber hinaus soll der aus **Z1** resultierende Korpus mit Kennzeichnungen der für die Antwort relevanten Textpassagen in den Verwaltungsakten und gegebenenfalls weiteren Gesetzen und Verordnungen erweitert werden. Die Fragen und gekennzeichneten Antworten sollen in Hinblick auf ihre Verständlichkeit und Beantwortbarkeit qualitativ evaluiert werden.
- **Z3:** Abhängig vom Umfang und Diversität des in **Z1** und **Z2** erstellten Korpus sollen passende Modelle zur Kategorisierung der Bescheide und Evaluation der Bürgerantworten selektiert und trainiert werden. Vorzugsweise wird hier ein Multi-Task Modell wie T5 gewählt, um die erwartete Datenknappheit mit *Transfer Learning* zwischen verwandten Aufgaben zu mitigieren. Gegebenfalls werden hierfür zusätzlich große Sammlungen unannotierter Rechtsdokumente für ein *Pre-training* des Modells benötigt. Die verwendeten Modelle gilt es zusätzlich quantitativ und qualitativ zu evaluieren.

1.3 Bezug des Vorhabens zu förderpolitischen Zielen / Förderprogramm

Der Software Campus ist in die Aktivitäten der EIT ICT Labs eingebettet und fördert die Ausbildung zukünftiger IT-Führungskräfte in Deutschland. Die in Trainings erworbenen Fertigkeiten werden im Rahmen dieses Projekts praktisch angewandt und festigen so erworbenes Wissen für eine nachhaltige Qualifizierung.

Das geplante Vorhaben findet sich als Beispiel dafür “Künstliche Intelligenz in die Anwendung (zu) bringen” in der Hightech-Strategie 2025 der Bundesregierung wieder. Zum einen erfolgt ein Technologietransfer zwischen NLP Forschung und Industrie, in diesem Fall repräsentiert durch IAV. Zum anderen legt dieses Projekt die Technologiegrundsteine für eine intelligente und automatisierte Anwendung, die die Kompetenz von Bürgern im Umgang mit behördlichen Bescheiden schult. Dies hilft nicht nur Bürgern, für die Verwaltungssprache undurchsichtig und unverständlich ist, sondern auch der Gesellschaft im Allgemeinen, da größeres Verständnis oft mit einer höheren Akzeptanz einhergeht. Aufkommende Spannungen zwischen Behörden und Bürgern werden gemindert, wenn Missverständnisse aufgeklärt und unnötige Strafen, beispielsweise für das unbeabsichtigte Versäumen von Fristen, verhindert werden können. Selbstbewusste und kompetente Bürger sind darüber hinaus eher in der Lage fehlerhafte Verwaltungsakte zu identifizieren und entsprechende Maßnahmen zu ergreifen, z.B. Widerspruch einzulegen. Die Kommunikation wird auch verwaltungsseitig erleichtert, wenn beide Parteien ein grundsätzliches Verständnis der rechtlichen Lage haben und einfache Fragen schon im Vorhinein mit Hilfe der Anwendung geklärt werden.

2 Stand der Wissenschaft und Technik

Dieses Kapitel fasst den Stand der Wissenschaft und Technik der für dieses Projekt relevante Bereiche zusammen. Hierbei werden sowohl (2.1) Feedbacksysteme als auch (2.2) deutsche Sprachverarbeitung in der Rechtsdomäne betrachtet.

2.1 Feedbacksysteme

Feedback generierende Systeme für geschlossene Aufgabenformate und formale Domänen sind wohl erforscht. So wird beispielsweise in modernen Programmierumgebungen statische Programmanalyse verwendet, um häufig auftretende Fehler, wie Zugriffe auf ungültige Speicheradressen, automatisch zu erkennen, die betroffene Stelle im Code zu markieren und die passende manuell erstellte Fehlermeldung/Warnung anzuzeigen. Auch in der Lehre werden vergleichbare Systeme eingesetzt, um automatisiert Feedback zu Mathematik oder Programmieraufgaben zu geben [4; 5]. Diese Systeme machen sich die formale Struktur ihrer Eingabedaten zu nutzen, die bei Freitextantworten nicht gegeben ist.

Beim automatischen korrigieren von Freitextantworten müssen Systeme in der Lage sein, mit sprachlicher Variation, Mehrdeutigkeiten und fehlerhafter Grammatik oder Rechtschreibung umzugehen. *Automatic short answer grading* Modelle werden darauf trainiert, unabhängig von Sprachstil Antworten auf ihre inhaltliche Korrektheit, Vollständigkeit und Relevanz zur gegebenen Frage zu prüfen. Hier kommen verschiedene Ansätze zum Einsatz. Eine Gruppe Systeme verwendet manuell erstellte Regeln und Bewertungsschemata, die sie auf Antworten anwenden. So kann man beispielsweise definieren welche Worte oder Teilworte in welcher Reihenfolge in einer Antwort erscheinen müssen [6], damit sie als richtig gilt. Solche Systeme finden sich auch auf Lernplattformen wie Moodle². Die manuelle Erstellung solcher Muster ist allerdings sehr zeitaufwendig und resultierende Muster haben nur eine begrenzte Abdeckung.

Alternative Ansätze nutzt diverse *Clusteringverfahren*, um bei eine Menge Antworten nach semantischer Ähnlichkeit zu gruppieren. Antworten in einem Cluster erhalten meist manuell eine gemeinsame Bewertung und Feedback [7]. So wird manueller Korrekturaufwand gemindert, da sich die Lehrperson nur wenige Antworten pro Cluster ansehen muss, statt alle Antworten individuell zu evaluieren. Allerdings eignen sich solche Verfahren nur, wenn vorhandene Antworten sich sinnvoll in Cluster aufteilen lassen. Dies ist viel eher der Fall, wenn homogene Gruppen von Lernenden, wie z.B. Schulklassen, ähnliche Lernressourcen nutzen oder dem gleichen Unterricht folgen. So kristallisieren sich natürlicherweise ähnliche Fehlvorstellungen in Antworten heraus, die man in Clustern auffangen kann. Möchte man allerdings ein System für heterogene Gruppen, wie beispielsweise alle deutschen Bürger, konzipieren, ist diese inhärente Struktur des Antwortraumes nicht zwangsweise gegeben. Darüber hinaus können zur Laufzeit Antworten nur automatisiert evaluiert werden, wenn sie sich einem der vorhandenen Cluster zuordnen lässt.

Die letzte Gruppe Verfahren umfasst *supervised learning* Machine Learning Modelle. Diese werden mit einer Menge manuell evaluierter Antworten darauf trainiert, die Bewertung ungesehener Antworten vorherzusagen. Hier werden in den letzten Jahren Modelle, die auf manuell erstellten Textrepräsentationen bzw. Features basieren, von *Deep Learning* in Sachen Prädiktionskraft dominiert. Häufig wird das Evaluieren von Antworten als *text similarity* oder *textual entailment* Problem zwischen einer oder mehreren Referenzantworten und der gegebenen Antwort betrachtet. Moderne *automatic short answer grading* Modelle schneiden in manchen Domänen gut im Vergleich mit menschlichen Bewertern ab [8]. Hierbei sind sie allerdings sowohl auf das Vorhandensein von Referenzantworten als auch eine ausreichende Datenmenge angewiesen. Im Szenario dieses Projekts sind allerdings keine Referenzantworten gegeben, da die nötigen Informationen erst zu Inferenzzeit im jeweiligen Bescheid vorliegen. Eine Innovation dieses Projektes besteht also darin, Referenzantworten zu Inferenzzeit aus dem gegeben Bescheid zu extrahieren oder alternativ den gesamten Bescheid als Referenz zu betrachten. Darüber hinaus profitieren Neuronale Netze davon, große Mengen unannotierter Daten zum Erlernen eines "generellen Sprachverstehens" nutzen zu können. Dies nennt sich *pre-training* und involviert meist das maskieren und vorhersagen zufälliger Worte

²Moodle: <https://moodle.de/>

in Textsequenzen oder das ordnen von Sätzen. Da vorhandene *pre-trained* Modelle meist auf Texten von Wikipedia, Zeitungen und dem Internet allgemein trainiert sind, muss die Anwendbarkeit auf deutsche Verwaltungssprache zunächst im Rahmen dieses Projekts untersucht und gegebenenfalls neu trainiert werden.

2.2 NLP mit deutschen Rechtsdokumenten

Wie schon in der Motivation aufgegriffen, ist die deutsche Verwaltungssprache sowie die juristische Fachsprache für Laien intransparent und schwer verständlich. Dies wurde bisher in diversen Projekten thematisiert, die sich mit der manuellen Aufbereitung der für die Öffentlichkeit relevanten Dokumente beschäftigten. Ein Beispiel hierfür ist das schon erwähnte Projekt “Übersetzung von Verwaltungsakten in Leichte Sprache” der Evangelische Stiftung Volmarstein. In diesem Projekt wurde untersucht, wie Verwaltungsakte in Leichte Sprache übersetzt werden und dementsprechend für Menschen mit Lernschwierigkeiten zur Verfügung gestellt werden können. Dies unterscheidet sich von unserem Projekt in folgenden Aspekten:

- Die Transformation von Verwaltungsakten in ein verständliches Format erfolgt manuell statt automatisiert.
- Die Zielgruppe sind Menschen mit Lernschwierigkeiten statt alle Bürger.
- Es wird keine Kompetenz im Umgang mit Bescheiden in Verwaltungssprache geschult, was eine Übersetzung von Verwaltungsakten in Leichte Sprache auch in Zukunft notwendig macht.

Das unter Horizon 2020 geförderte EU-Projekt “Lynx”³ strebt eine Sammlung von Clouddiensten an, die Unternehmen bei der Verwaltung von Compliance-Dokumenten unterstützen. Hierbei sollen diverse multilinguale Rechtsdokumente in einem “Legal Knowledge Graphen” verlinkt und darauf basierend Dienste, wie beispielsweise das Zusammenfassen oder Vorschlagen von Dokumenten, angeboten werden. Hier unterscheiden sich sowohl die Zielgruppe als auch das Szenario von unserem Projekt. Über diese Projekte hinaus gibt es wissenschaftliche Arbeiten, die sich mit der automatisierten Sprachverarbeitung deutscher Rechtsdokumente beschäftigen. So werden beispielsweise Korpora aus deutschen Gerichtsurteilen veröffentlicht, die mit “Named Entity” Annotationen versehen sind [9] oder Tools, die die automatische Sprachverarbeitung rechtlicher Dokumente vereinfachen [10].

³Lynx: <https://lynx-project.eu/>

3 Partner und bisherige Arbeiten

Dieses Vorhaben ist in die Förderung im *Software Campus* mit der *IAV GmbH Ingenieurgesellschaft Auto und Verkehr* als Industriepartner sowie der *Technischen Universität Darmstadt (TUDa)* als akademischer Partner integriert. IAV wird hierbei durch Bereitstellung diverser Weiterbildungsmaßnahmen als Mentor der Projektleiterin agieren und nicht selbst gefördert. Das Projekt wird am Multimedia Communications Lab (KOM) der Technischen Universität Darmstadt durchgeführt. KOM ist dem Fachbereich Elektrotechnik und Informationstechnik zugehörig und wird von *Prof. Dr. Ralf Steinmetz* geleitet. Die wissenschaftliche Mitarbeiterin *Anna Filighera* wird die Leitung des Projekts übernehmen und ihre Erfahrung im Bereich *automatic short answer grading* einbringen [11; 12].

Im Folgenden werden der Industriepartner IAV, der akademische Partner TU Darmstadt mit den relevanten Fachbereichen, sowie deren Beziehungen untereinander vorgestellt.

3.1 Universität/Forschungseinrichtung

Die *TU Darmstadt*⁴ zeichnet sich seit ihrer Gründung im Jahre 1877 durch herausragende Forschung und Lehre aus. Dies spiegelt sich nicht nur in ihren Auszeichnungen im Rahmen der Exzellenzinitiativen in den Jahren 2007 und 2016 wieder, sondern auch durch regelmäßige Präsenz in den allgemeinen Medien. International ist die Technische Universität Darmstadt ebenfalls für ihre Exzellenz bekannt, was sich unter anderem durch großes Interesse ausländischer Studierender äußert. Die TU Darmstadt ist zudem Teil des Universitätsbundes TU9 German Institutes of Technologies, einem Zusammenschluss der neun führenden deutschen Technischen Universitäten. Darüber hinaus ist sie die erste autonome Universität Deutschlands. Mit etwa 25.000 Studierenden in 13 Fachbereichen mit mehr als 100 Studiengängen und 5.000 Beschäftigten, davon ca. 300 Professoren, ist die TU Darmstadt eine mittelgroße Universität in Deutschland. Das Forschungsprofil der TU Darmstadt wird geprägt durch die folgenden sechs Profilbereiche und ein Profilhema: (1) Cybersicherheit, (2) Internet und Digitalisierung, (3) Teilchenstrahlen und Materie, (4) Thermo-fluids & Interfaces, (5) Energiesysteme der Zukunft, (6) Vom Material zur Produktinnovation und (7) Computational Engineering. In all diesen Bereichen kooperiert die TU Darmstadt mit mehr als 100 Partneruniversitäten weltweit.

Der *Fachbereich Elektrotechnik und Informationstechnik*⁵ ist ein hoch angesehener Fachbereich mit drei Forschungsschwerpunkten. “Information & Communication Technology”, einer der Schwerpunkte, ist verwandt mit der Informatik. Dieser Schwerpunkt beinhaltet Elemente der Systemtheorie, die Charakterisierung elektronischer Bauelemente, Netze und relevante Anwendungen im Bereich Informationsübertragung und -verarbeitung. Er beschäftigt sich mit Datentechnik, Hochfrequenz- und Nachrichtentechnik sowie Photonik. Der Forschungsschwerpunkt ist durch eine intensive Kooperation mit der Informatik gekennzeichnet.

Der *Fachbereich Informatik*⁶ der TU Darmstadt gehört in diversen Rankings, wie beispielsweise CHE⁷, zu den Top-Informatikfachbereichen deutschlandweit. Der Fachbereich beschäftigt viele herausragende Wissenschaftler und Wissenschaftlerinnen, die in den folgenden Bereichen agieren: (1) Computational Engineering und Robotik, (2) IT-Sicherheit, (3) Massiv Parallele Softwaresysteme, (4) Netze und Verteilte Systeme, (5) Sprach- und Wissensverarbeitung und (5) Visual Computing. Die Fraunhofer-Institute für Graphische Datenverarbeitung und Sicherheit in der Informationstechnik, zusammen mit dem international sichtbaren Kompetenzzentrum für Cybersicherheit (CRISP) sowie die enge Verbindung mit der IT-Stadt Darmstadt und der IT-Region Rhein-Main ermöglichen dem Fachbereich, ein ideales Umfeld für innovative Forschung und gute Lehre zu bieten.

⁴Technische Universität Darmstadt (TUDa) <http://www.tu-darmstadt.de>

⁵Fachbereich Elektrotechnik und Informationstechnik der TU Darmstadt <https://www.etit.tu-darmstadt.de>

⁶Fachbereich Informatik der TU Darmstadt <http://www.informatik.tu-darmstadt.de>

⁷CHE Ranking <https://ranking.zeit.de/che/de>

Das *Multimedia Communications Lab*⁸ (KOM) wird von Prof. Dr. Ralf Steinmetz geleitet und ist sowohl dem Fachbereich Elektrotechnik und Informationstechnik als auch der Informatik der TU Darmstadt zugeordnet. Die folgenden Forschungsschwerpunkte finden sich in KOM: (1) Multimedia Technologies & Serious Games, (2) Knowledge & Educational Technologies und (3) Adaptive Communication Systems. KOM beschäftigt momentan 17 Wissenschaftliche Mitarbeiter, 7 Post-Docs und 11 administrative Mitarbeiter.

In der Knowledge Media Gruppe des Multimedia Communications Lab werden Erfahrungen zum anwendungsorientiertem Einsatz von NLP und ML in Lernszenarien gebündelt. Dies spiegelt sich unter anderem in den erfolgreich durchgeführten Projekten wieder, wie die LOEWE Linie 3 Projekte TexSaS und ZuMaP, das BMBF und ESF gefördertes Projekt KeaP digital, sowie das DFG Projekt „Design and Evaluation of new mechanisms for crowdsourcing as emerging paradigm for the organization of work in the Internet“. Darüber hinaus ist die Knowledge Media Gruppe im Innovationsforum für “(Trusted) Learning Analytics” der hessischen Hochschulen involviert.

3.2 Unternehmen

Der ungeförderte Industriepartner IAV zählt zu einem der weltweit führenden Engineering-Dienstleistern im Automobilbereich und arbeitet als solcher mit allen großen Namen der Automobil- und Forschungswelt zusammen. Selbstgegebenes Ziel des Unternehmers ist die Entwicklung der Mobilität der Zukunft. Hierzu schlägt IAV die Brücke zwischen der Automotive- und IT-Welt, zwischen Hardware und Software sowie zwischen Produkten und Service-Dienstleistungen. Weltweit arbeiten bei IAV mehr als 8.000 Mitarbeiter an 25 Standorten. 16 hiervon liegen allein in Deutschland. Sie zeichnen sich durch eine unmittelbare Nähe zu den Kunden aus und decken so alle nationalen sowie internationalen Hotspots der globalen Automobilentwicklung ab. IAV zählt heute zu einem der größten Engineering-Unternehmen weltweit.

3.3 Beziehung Universität - Unternehmen

Im Rahmen des Software Campus wird die Kooperation zwischen Industrie- und Forschungspartner in regelmäßigem Austausch über den Projektstatus, Vorgehensweise sowie Erfahrungsgewinne bei der Umsetzung der Projektidee bestehen. Hierbei fließt die Universität vor allem mit ihrer Erfahrung mit Sprachverarbeitungssystemen und der Betreuung studentischer Hilfskräfte ein. IAV wird das Mentoring der Projektleiterin übernehmen und auch im Rahmen des Projekts die nötige verwaltungsrechtliche Expertise einbringen, beispielsweise bei der Erstellung des Fragenkatalogs.

⁸Multimedia Communications Lab (KOM) <https://www.kom.tu-darmstadt.de>

4 Ausführliche Beschreibung des Arbeitsplans

Das Projekt lässt sich in 7 Arbeitspakete und 3 Meilensteine untergliedern. Der Projektstart ist für den 1. April 2021 angesetzt und die Laufzeit beträgt 24 Monate. Insgesamt werden im Rahmen dieses Projekts ein Wissenschaftlicher Mitarbeiter (WiMi) zu 10 Personenmonaten (PM) und drei studentische Hilfskräfte (HiWi) zu je 3,5 PM beschäftigt. Hierbei umfasst ein PM eine Vollbeschäftigung von 40 Stunden pro Woche für 4 Wochen. Die 7 Arbeitspakete sind in den folgenden Sektionen im Detail beschrieben, lassen sich allerdings folgendermaßen zusammenfassen:

Die organisatorischen und administrativen Tätigkeiten des Projektleiters, wie das Betreuen des Personals und das Schreiben der Berichte, finden sich in Arbeitspaket 1 - "Projektmanagement". Dieses Paket ist ungefördert. Arbeitspaket 2 - "Erstellung eines Korpus aus Verwaltungsakten" korrespondiert zu **Z1** und schafft die Datengrundlage für dieses Projekt. Hier werden reale Bescheide und Musterbescheide zusammengetragen und annotiert. **Z2** wird in Arbeitspaket 3 - "Erstellung des Fragenkatalogs" realisiert. Hier werden zunächst kritische Aspekte - aus Sicht der Bürger - der Bescheide identifiziert und verständliche Fragen zu den Aspekten formuliert. Daraufhin werden für die Antworten relevante Textstellen in den Bescheiden und gegebenenfalls weiteren juristischen Dokumenten markiert. Die Ergebnisse von Arbeitspaket 2 und 3 werden anschließend in Arbeitspaket 4 - "Analyse und Evaluation des erstellten Korpus" qualitativ evaluiert. Hierbei werden eventuelle Schwachstellen oder Lücken des Korpus identifiziert und der Korpus entsprechend modifiziert.

Nachdem die Datengrundlage geschaffen wurde, werden nun in Arbeitspaket 5 - "Konzeption des Feedbackmodells aus dem Korpus resultierende Möglichkeiten erwogen und entsprechende Modelle für das Feedbacksystem konzipiert. Dieses wird dann in Arbeitspaket 6 - "Implementation des Kategorisierungs- und Feedbackmodells" prototypisch umgesetzt und in Arbeitspaket 7 - "Evaluation des Kategorisierungs- und Feedbackmodells" sowohl qualitativ als auch quantitativ evaluiert. Ein Gantt-Diagramm des geplanten Projektverlaufs findet sich in Abbildung 2. Die Zuteilung der PM zu den einzelnen Arbeitspaketen findet sich in Tabelle 1.

	AP1	AP2	AP3	AP4	AP5	AP6	AP7	Summe PM
PL	(3)	0	0	0	0	0	0	(3)
WiMi	0	1	1	2	1,5	2,5	2	10
HiWi#1	0	2	1,5	0	0	0	0	3,5
HiWi#2	0	1	1,5	1	0	0	0	3,5
HiWi#3	0	1	1,5	1	0	0	0	3,5

Tabelle 1: Zuteilung des Personals in Personenmonaten (PM) zu den Arbeitspaketen. Die PM des Projektleiters (PL) sind ungefördert.

4.1 Arbeitspaket 1: Projektmanagement

Dieses Arbeitspaket wird von der ungeförderten Projektleiterin bearbeitet und beeinflusst dementsprechend die Kostenrechnung nicht. Es umfasst die organisatorischen und administrativen Aufgaben, wie das Schreiben der Berichte, das Betreuen des Personals und die Absprache mit dem Industriepartner, die während der gesamten Projektlaufzeit anfallen. Darüber hinaus beinhaltet es regelmäßige Fortschrittskontrollen, um die Passung des Zeitplans kontinuierlich zu überprüfen und Probleme frühzeitig zu identifizieren. Insgesamt werden für dieses Arbeitspaket 3 ungeförderte PM eingeplant. Ergebnisse dieses Pakets sind die zu erstellenden Berichte sowie Publikationen.

4.2 Arbeitspaket 2: Erstellung eines Korpus aus Verwaltungsakten

Für dieses Paket ist der wissenschaftliche Mitarbeiter mit einem PM sowie 3 studentische Hilfskräfte mit insgesamt 4 PM vorgesehen. Hierbei sind die Aufgaben des wissenschaftlichen Mitarbeiters zum

	Q2 2021			Q3 2021			Q4 2021			Q1 2022			Q2 2022			Q3 2022			Q4 2022			Q1 2023		
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A1																								
A2																								
A3																								
A4																								
A5																								
A6																								
A7																								
MS1																								
MS2																								
MS3																								

Abbildung 2: Gantt-Diagramm der Arbeitspakete A1-A7 und der Meilensteine MS1-MS3. Insgesamt beträgt die Projektlaufzeit 24 Monate.

einen Kontakt mit Behörden aufzunehmen, um Musterbescheide zu erlangen, sowie datenschutzkonforme Möglichkeiten zu ergründen, reale Bescheide zu nutzen. Beispielsweise könnten identifizierende oder persönliche Informationen aus Bescheiden geschwärzt werden, um sie anonymisiert dem Korpus hinzufügen zu können. Zum anderen ist es die Aufgabe des wissenschaftlichen Mitarbeiters, die studentischen Hilfskräfte in den Annotationsprozess einzuweisen und entsprechende Richtlinien und Kategorien für das Annotieren zu definieren. Hierbei wird relevante Literatur gesichtet, um den Best Practices im Bereich der Datenannotation folgeleisten zu können. Dies ist essentiell, da das gesamte Projekt auf der Qualität des annotierten Korpus beruht. Einer der drei studentischen Hilfskräfte wird mit einem PM den wissenschaftlichen Mitarbeiter bei seinen Aufgaben unterstützen und die Digitalisierung der Bescheide in ein für das weitere Projekt hilfreiche Format sicherstellen. Darüber hinaus werden alle drei studentischen Hilfskräfte die tatsächliche Annotation der Bescheide vornehmen. Hierfür sind drei parallele Annotatoren notwendig, um die Objektivität und Verlässlichkeit der Annotationen sicherzustellen und die Berechnung von Qualitätsmetriken, wie der Interrater-Reliabilität, zu ermöglichen. Sollten sich die Annotatoren bei einem Bescheid nicht einig sein, kann dieser im gesamten Team diskutiert werden, um einen Konsens zu erreichen.

Das Ergebnis dieses Pakets ist dementsprechend eine digitale Sammlung kategorisierter Bescheide, die veröffentlicht und so vielen Anwendungen als Datengrundlage dienen kann. Die Objektivität der Kategorisierung kann hierbei auf Grund der parallelen Annotation quantifiziert werden.

4.3 Arbeitspaket 3: Erstellung eines Fragenkatalogs

Dieses Arbeitspaket ist eine Vertiefung und Erweiterung der Bescheidsammlung aus Arbeitspaket 2. Es beginnt sobald die ersten Bescheide vorliegen mit einer Anforderungsanalyse. Hier wird **F2** thematisiert, also versucht zu ergründen, welche Informationen eines Bescheids für das weitere Handeln des Bürgers von Bedeutung sind und wie diese eindeutig und verständlich abgefragt werden können. Insbesondere wird hier die juristische Expertise des Industriepartners konsultiert, sowie relevante Literatur gesichtet. Dies wird vorrangig vom wissenschaftlichen Mitarbeiter (1PM) mit der Unterstützung der studentischen Hilfskräfte (je 0,5 PM) getrieben. Aus der Anforderungsanalyse werden anschließend analog zu Arbeitspaket 2 Annotationsrichtlinien und Guidelines zum Formulieren von Fragen definiert.

Je nach Umfang und Diversität der in Arbeitspaket 2 gesammelten Bescheide, wird nun der Fokus entweder auf eine bestimmte, ausreichend repräsentierte Kategorie von Bescheiden gelegt oder zu jeder Kategorie nur die essentiellen Fragen formuliert. Für das Formulieren der Fragen und das entsprechende

markieren der relevanten Textpassagen sind die studentischen Hilfskräften zu je einem PM vorgesehen. Während die Fragen gemeinsam erarbeitet werden, erfolgt das Markieren der relevanten Textpassagen parallel. Dies geschieht wie in Arbeitspaket 2 parallel, um die Objektivität der Annotationen bewerten zu können. Das Ergebnis dieses Arbeitspaketes ist eine Sammlung von Fragen, die je nach Kategorie des Bescheides die für den Bürger relevanten Aspekte klar und verständlich abfragen. Darüber hinaus ist der Korpus aus Arbeitspaket 2 für jede Frage mit relevanten Textpassagen annotiert.

4.4 Arbeitspaket 4: Analyse und Evaluation des erstellten Korpus

Ziel dieses Arbeitspaketes ist es, die Qualität des erstellten Korpus zu messen und sicherzustellen. Dies erfolgt zum einen kontinuierlich in der letzten Hälfte von Arbeitspaket 3 und zum anderen in einer abschließenden Evaluation nachdem der Korpus vollständig erstellt ist. Die kontinuierliche Evaluation dient dem Zweck, erstellte Fragen auf ihre Verständlichkeit und Präzision zu prüfen und gegebenenfalls zu überarbeiten. Hierbei werden die Fragen in Benutzerstudien Bürgern präsentiert und ihr Verständnis sowohl durch Beantwortung der Frage als auch durch explizite Angaben in einem Fragebogen erfasst. Die Bürgerantworten werden in der Auswertung manuell auf ihre Richtigkeit und Vollständigkeit geprüft, um zum einen Datenpunkte für das Feedbacksystem zu sammeln und zum anderen einen Indikator für das Verständnis zu haben, der nicht auf Selbsteinschätzung beruht. Das Design der Studie und die Erstellung des Fragebogens obliegt dem wissenschaftlichen Mitarbeiter (1 PM). Unter die Designentscheidungen der Studien fällt hierbei auch, ob sie beispielsweise auf Amazon Mechanical Turk ⁹, Social Media oder in Präsenz statt finden werden. Die Durchführung und Auswertung der Studien werden nach Anleitung des Mitarbeiters 2 studentische Hilfskräfte übernehmen, die ebenfalls die Überarbeitung der Fragen vornehmen (je 1 PM).

Die finale Evaluation dient dem Zweck, die Qualität des Korpus zu erfassen. Dies ist wichtig für die Veröffentlichung des Korpus und die weitere Nutzung. Nur, wenn der Korpus eine ausreichende Größe erreicht, sinnvolle Fragen enthält und die nötige Interrater-Reliabilität erfüllt, macht es Sinn weitere Komponenten auf den Korpus aufzubauen. Hierfür ist eine Benutzerstudie geplant, die vollständig vom wissenschaftlichen Mitarbeiter betreut wird (1 PM), um die nötige Wissenschaftlichkeit und Repräsentativität zu gewährleisten. Um eine ausreichende Stichprobe der Gesellschaft zu erlangen, wird hierbei ein finanzielles Budget eingeplant, um Studienteilnehmer entsprechend ihres Aufwandes entlohnen zu können. Das Ergebnis dieses Arbeitspaketes ist ein im Vergleich zu dem Ergebnis aus Arbeitspaket 3 verbesserter Korpus und ein Datasheet, welches relevante Metriken des Korpus, wie z.B. Klassenverteilungen, demographische Informationen der Annotatoren und Interrater-Reliabilitäten enthält.

4.5 Arbeitspaket 5: Konzeption des Feedbackmodells

Dieses Arbeitspaket wird vollständig von einem wissenschaftlichen Mitarbeiter mit 1,5 PM bearbeitet, da es tiefgreifende Kenntnisse im NLP und Machine Learning Bereich erfordert. Hier wird mit Hilfe der in Arbeitspaket 4 durchgeführten Analyse abgewogen, welche Machine Learning Modelle sich für den Korpus eignen. Vor allem Faktoren wie die Anzahl an Datenpunkten pro Kategorie und der Umfang an relevanten Textpassagen pro Frage spielen hier eine wichtige Rolle. Ebenfalls ist hier zu beachten, ob die in den Benutzerstudien gesammelten Bürgerantworten ausreichen, um einen *supervised* Ansatz zu trainieren, oder ein *unsupervised* Ansatz gewählt werden muss. Hierbei ist eine bessere Prädiktionskraft bei *supervised* Ansätzen zu erwarten. Darüber hinaus gilt es zu untersuchen, inwiefern verwandte Daten und Aufgaben genutzt werden können, um eventuelle Datenknappheit auszugleichen.

Je nach Korpus muss auch in Erwägung gezogen werden, inwiefern ein Feedbacksystem, welches textuelles Feedback generiert, trainierbar ist. Alternativ kann hier im Falle einer unzureichenden Datenlage auf Modelle zurückgegriffen werden, die entweder inhärent verständlich für Menschen sind, wie Entscheidungsbäume oder regelbasierte Systeme oder Methoden aus dem Feld der *Explainable AI* verwendet werden, um beispielsweise ein Highlighting für den Bürger einzublenden, welches anzeigt welche Teile seiner Antwort sich positiv oder negativ auf die Entscheidung des Modells ausgewirkt haben. Nehmen wir

⁹Amazon Mechanical Turk: <https://www.mturk.com/>

beispielsweise an, dass der Bürger eine Antwort geliefert hat, die teilweise richtig und teilweise falsch ist. So würde das Modell seine Antwort mit den Referenzdokumenten vergleichen und die Antwort als falsch klassifizieren. Nun könnte man dem Bürger markieren, welche Teile seiner Antwort zu der Klassifikation beigetragen haben und wie stark ihr Einfluss war. Insgesamt gibt es viele Möglichkeiten ein solches Feedbacksystem zu designen und das Ergebnis dieses Arbeitspakets ist eine Einschränkung der Möglichkeiten auf vielversprechende Optionen. Hierbei sind erste Experimente zur Einschätzung der Durchführbarkeit diverser Ansätze möglich.

4.6 Arbeitspaket 6: Implementation des Kategorisierung- und Feedbackmodells

Das Ergebnis dieses Arbeitspaketes ist eine prototypische Open Source Implementation des trainierten Kategorisierungs- und Feedbackmodells. Dies umfasst das Programmieren der Trainings- und Modellskripte, sowie das eigentliche Training des Modells. Hierbei gilt es ein oder mehrere in Arbeitspaket 5 identifizierte Ansätze praktisch umzusetzen. Dieses Arbeitspaket wird ausschließlich von einem wissenschaftlichen Mitarbeiter bearbeitet, da es zum einen Erfahrung im Umgang mit Machine Learning Frameworks wie NLTK ¹⁰, Tensorflow ¹¹ oder PyTorch ¹² und die nötige Expertise im wissenschaftlichen Arbeiten erfordert, um beispielsweise das Training nicht mit Wissen aus dem Testdatenset zu kontaminieren. Zum anderen wäre es nicht zeiteffizient eine studentische Hilfskraft in alle Erwägungen aus Arbeitspaket 5 einzuarbeiten und die notwendigen NLP Grundlagen sicherzustellen.

Für den wissenschaftlichen Mitarbeiter sind in diesem Paket 2,5 PM vorgesehen. Teile der Implementierung können hierbei gegebenenfalls in Lehrveranstaltungen wie Praktika unter Aufsicht des Mitarbeiters an Studenten ausgelagert werden.

4.7 Arbeitspaket 7: Evaluation des Kategorisierung- und Feedbackmodells

In diesem Paket gilt es das/die in Arbeitspaket 6 erstellte/n Modell/e quantitativ und qualitativ zu evaluieren. Hierbei reicht es für die Klassifikation der Kategorie der Bescheide aus, ein Testdatenset aus während des Trainings ungesehenen Bescheiden zurückzuhalten und die Prädiktionskraft auf dem Testdatenset quantitativ zu ermitteln. Für die Feedbackgenerierung ist dies allerdings nicht ausreichend. Automatische Evaluationsmetriken wie BLEU [13] für textuell generiertes Feedback eignen sich nur bedingt, um die Qualität von gegebenen Feedback einzuschätzen, da solche Metriken lediglich die Wortüberlappung zwischen dem generierten Text und einem Referenztext berücksichtigen. Aus diesem Grund ist hier ebenfalls eine Studie mit menschlichen Teilnehmern geplant. Hier ist es mit 2 PM die Aufgabe des wissenschaftlichen Mitarbeiters relevante Qualitätskriterien für Feedback zu recherchieren, für die Studie zu definieren und daraus einen Fragebogen für eine Benutzerstudie zu konzipieren. Bei dieser Studie wird ebenfalls im Projektverlauf unter Berücksichtigung der aktuellen Situation um COVID-19 entschieden, ob diese mit Crowdsourcing, Social Media oder vor Ort durchgeführt wird.

Dieses Arbeitspaket läuft zeitweise parallel zu Arbeitspaket 6 damit gewonnene Erkenntnisse aus den ersten Evaluationen noch in die Entwicklung des Systems einfließen können. Hierbei ist es wichtig, das Testdatenset erst für die Evaluation des finalen Modells zu verwenden, da sonst Informationen über das Testdatenset in die Entwicklung des Modells einfließen und so keine aussagekräftige Evaluation der Generalisierbarkeit des Modells durchgeführt werden kann. Hier kann bei ausreichender Datenlage ein Teil des Trainingssets zu Entwicklungszwecken abgespalten werden, um verschiedene Ansätze oder Konfigurationen des Modells vergleichen zu können. Sollten nicht ausreichend Daten zur Verfügung stehen, kann hier auf Kreuzvalidierungsverfahren zurückgegriffen werden [14].

Das Ergebnis dieses Pakets ist im schlechtesten Fall eine Indikation, inwiefern ein solches System realisierbar ist und im besten Fall eine repräsentative Validierung der Effektivität des Systems, das Verständnis

¹⁰NLTK: <https://www.nltk.org/>

¹¹Tensorflow: <https://www.tensorflow.org/>

¹²PyTorch: <https://pytorch.org/>

von Bürgern in Bezug auf erhaltene Bescheide zu erhöhen.

4.8 Zeitplanung und Meilensteine

Die Projektlaufzeit beträgt insgesamt 24 Monate und beginnt am 1. April 2021. Damit während des Projektverlaufs der Fortschritt gemessen werden kann, sind 3 Meilensteine definiert, deren zeitliche Einordnung in Abbildung 2 abgebildet ist. Eine Übersicht der Meilensteine findet sich in Tabelle 2. Im folgenden Abschnitt werden die Meilensteine im Detail definiert.

Meilenstein MS1 Der erste Meilenstein gilt als erreicht, wenn die Erstellung des Korpus abgeschlossen und damit die Grundlage für dieses Projekt geschaffen ist. Hierbei umfasst der Korpus mindestens die schon vorhandenen Bescheide der Nürnberger Bundesagentur für Arbeit, die mit ihren Kategorien annotiert sind. Darüber hinaus ist für mindestens eine Kategorie ein Fragenkatalog erstellt und zusammen mit Kennzeichnungen relevanter Textstellen für die Antworten in den Korpus eingepflegt. Bei Vollendung dieses Meilensteins liegt also ein veröffentlichbarer Korpus vor. Unter der Annahme, dass der Korpus für eine Veröffentlichung ausreichende Größe erlangt, ist zur Vollendung dieses Meilensteins ein Data-Set erstellt, welches die demographischen Informationen der Annotatoren und weitere Qualitätsmetriken des Korpus enthält. Dies ist wichtig, damit vorhandener Bias von anderen Wissenschaftlern identifiziert und auf dem Korpus erzielte Ergebnisse entsprechend interpretiert werden können. Die Erreichung dieses Meilensteins ist für den 9. Projektmonat geplant.

Meilenstein MS2 Der zweite Meilenstein ist für den 14. Projektmonat angedacht. Er ist erfüllt, wenn die Analyse des Korpus abgeschlossen und entsprechende Entschlüsse in Bezug auf den weiteren Projektverlauf daraus gezogen sind. Hierbei sind Umfang, Diversität und Verlässlichkeit der Annotationen für die Konzeption des Feedbacksystems in Betracht gezogen worden. Der Lösungsraum für das Feedbacksystem wurde in Pilotversuchen exploriert und mindestens ein Feedbacksystem wurde konzipiert, welches in den folgenden Monaten praktisch umgesetzt werden kann.

Meilenstein MS3 Der letzte Meilenstein ist zu Projektende zu erreichen. Hier sind alle Arbeitspakete abgeschlossen. Das in Meilenstein MS2 erarbeitete Konzept wurde implementiert und sowohl mit vom Training zurückgehaltenen Daten als auch mit Bürgern evaluiert. Hierbei sind mindestens Indikatoren für die Realisierbarkeit des konzipierten Feedbacksystems resultiert. Im besten Fall wird allerdings eine aussagekräftige und repräsentative Validierung oder Widerlegung der Effektivität des Systems, Bürgerverständnis in Bezug auf Bescheide zu erhöhen, angestrebt. Alle in diesem Projekt erzielten Erkenntnisse lassen sich nun in einem umfassenden Beitrag veröffentlichen. Dies schließt eine frühere Veröffentlichung von interessanten Erkenntnissen nicht aus.

Monat	Meilenstein / Beschreibung
9	MS1: Korpuserstellung abgeschlossen, Datengrundlage für das Projekt geschaffen
14	MS2: Lösungsraum auf Grund der Datenlage bestimmt, Konzipierung des Feedbackmodells abgeschlossen
24	MS3: Implementierung abgeschlossen, qualitative und quantitative Evaluation des gewählten Ansatzes abgeschlossen

Tabelle 2: Übersicht der Meilensteine

4.9 Finanzplanung

Dieses Kapitel schlüsselt alle in diesem Projekt geplanten Kosten auf. Eine Übersicht aller Kosten findet sich in 3, alle Einzelposten werden in den entsprechenden Abschnitten detaillierter aufgeführt. Die sonstigen unmittelbaren Vorhabenskosten beinhaltenden hierbei die erwartenden Kosten der geplanten Benutzerstudien. Insgesamt werden für dieses Projekt 99.878,60 € beantragt.

Kategorie	Ausgabe	Kalkulation	Preis	Kosten
Personalausgaben				86.379,60 €
Mitarbeiter	WiMi	10 PM	5.973,48 €/PM	59.734,80 €
Hilfskräfte	HiWis	10,5 PM	2.537,60 €/PM	26.644,80 €
Dienstreisen				5.499,00 €
Industriepartner & SWC	Gesamtkosten	8 Stück	316,00 €/Stück	2.528,00 €
Konferenzen	Gesamtkosten			2.971,00 €
Benutzerstudien Kosten				8.000,00 €
Summe				99.878,60 €

Tabelle 3: Auflistung der einzelnen Positionen und der zu erwartenden Ausgaben

4.9.1 Geplante Personalmittel

In diesem Projekt sollen ein wissenschaftlicher Mitarbeiter für 10 PM und drei studentische Hilfskräfte für je 3,5 PM eingestellt werden. Die Tätigkeit der studentischen Hilfskräfte beläuft sich, wie auch in Tabelle 1 zu sehen, primär auf die Erstellung des Korpus aus Verwaltungsakten. In diesem Rahmen annotieren sie nach einer Einweisung durch den wissenschaftlichen Mitarbeiter in Arbeitspaket 2 Bescheide mit ihren rechtlichen Kategorien (Gebot, Verbot, etc.), erstellen verständliche Fragen zu Bescheiden einer Kategorie und markieren in rechtlichen Dokumenten und den Bescheiden selbst für die Beantwortung der Frage relevante Textpassagen. Zwei der drei studentischen Hilfskräfte helfen anschließend bei der Analyse und Evaluation des erstellten Korpus. Hier sind sie für die Durchführung und Auswertung der durch den wissenschaftlichen Mitarbeiter geplanten Benutzerstudie und die der Ergebnisse entsprechende Überarbeitung des Fragenkatalogs verantwortlich. All diese Aufgaben eignen sich sehr gut für studentische Hilfskräfte, da in kurzen Einweisungen und mit guten Anleitungen alle für die Aufgabe wichtigen Informationen vermittelt werden können. Vorwissen in Verwaltungsrecht ist hier nützlich, aber bei umfassenden Annotationsguidelines nicht zwingend notwendig.

Die Aufgaben des wissenschaftlichen Mitarbeiters hingegen profitieren stark von tiefgehender Expertise und Erfahrung. So ist der wissenschaftliche Mitarbeiter in den Arbeitspaketen 2-4 für die Literaturrecherche, die Anforderungsanalyse, die Erstellung der Annotationsguidelines und das Design der Benutzerstudien verantwortlich. Die Literaturrecherche ist nötig, um folgende Schritte zu informieren. Es ist essentiell, dass eine darin ausgebildete Person die Anforderungsanalyse und die Erstellung der Annotationsguidelines übernimmt, da das gesamte Projekt auf dem erstellten Korpus aufbaut und kleine Fehler die Authentizität und Verlässlichkeit der Annotationen invalidieren können. Erfahrung und Vorwissen helfen hier beispielsweise missverständliche Formulierung und systematischen Bias zu minimieren. Ähnlich ist es bei der Benutzerstudie nötig, wissenschaftliche Standards einzuhalten damit erzielte Ergebnis auch Aussagekraft besitzen. Darüber hinaus obliegt es dem wissenschaftlichen Mitarbeiter weitere Bescheide zu erlangen, sodass zum einen Datenschutz Richtlinien eingehalten werden und zum anderen das Projekt im Kontakt mit Außenstehenden gebührend repräsentiert wird.

Die Arbeitspakete 5-7 werden ausschließlich von einem wissenschaftlichen Mitarbeiter bearbeitet, da sie Machine Learning und NLP Expertise erfordern. So muss der Stand des Wissens bekannt und ein tiefgehendes Verständnis relevanter Methoden vorhanden sein, um für dieses Projekt und vor allem für den erstellten Korpus passende Verfahren auszuwählen und anzuwenden. Darüber hinaus ist Programmiererfahrung mit den entsprechenden Machine Learning und NLP Bibliotheken nötig, um eine effiziente und fehlerfreie Implementierung zu ermöglichen und wissenschaftliche Fallstricke, wie das unbeabsichtigte

Ziel, Ort	An-/ Abreise	Gebühr	Über- nachtung	Tage- geld	Kosten pro Reise	#	Gesamt
Trainings, ?	100 €	0 €	160 €	56 €	316 €	6	1.896 €
SWC Summit, Berlin	100 €	0 €	160 €	56 €	316 €	1	316 €
IAV, Gifhorn	100 €	0 €	160 €	56 €	316 €	1	316 €
EC-TEL 2022, ?	220 €	650 €	610 €	257 €	1.737 €	1	1.737 €
DELFI 2022, ?	180 €	500 €	400 €	154 €	1.234 €	1	1.234 €
Summe							5.499 €

Tabelle 4: Übersicht der Reisekostenkalkulation.

Einbringen von Information aus dem für die Evaluation zurückgehaltenen Testset in die Modellselektion, zu vermeiden. Sollten sich hier unkritische Teilaspekte der Implementation auftun, steht es dem wissenschaftlichen Mitarbeiter frei, diese in Form von Praktikas oder Abschlussarbeiten von Studenten umsetzen zu lassen.

4.9.2 Geplante Reisemittel

Im Rahmen dieses Projekts sind ein internationaler Konferenzbesuch, ein nationaler, Trainings, eine Reise zum Projektpartner und der Besuch des Software Campus Summits in Berlin geplant. Insgesamt fallen hierfür 5.499 € Reisekosten an, die sich wie in Tabelle 4 dargestellt aufschlüsseln. Die Kosten für die einzelnen Posten errechnen sich wie folgt:

Trainings: Zur Weiterbildung der Projektleiterin sind im Rahmen des Software Campus sechs deutschlandweite Trainingsveranstaltungen vorgesehen. Da die Orte der jeweiligen Veranstaltungen zum Zeitpunkt des Verfassens dieser Vorhabensbeschreibung noch nicht veröffentlicht sind, werden für die An- und Abreisen durchschnittlich 100€ Kosten angenommen. Ein Training ist für zwei Tage konzipiert, sodass je zwei Übernachtungen eingeplant werden müssen. Die Übernachtungspauschale nach HRKG beträgt für Reisen innerhalb Deutschlands 80 €. Das Tagegeld nach HRKG beträgt 28 € für volle Tage und 14 € für Teiltage (mehr als 8 Stunden). Dementsprechend ergeben sich 28€ für den ersten Veranstaltungstag und je 14 € für den An- und Abreisetag für insgesamt 56 € Tagegeld. Für jedes einzelne Training fallen also 316 € Reisekosten an, was in Gesamtkosten in Höhe von 1.896 € für alle Trainings resultiert.

SWC Summit: Auf dem Software Campus Summit in Berlin werden die Projektergebnisse präsentiert. Die Kalkulation der Kosten ergibt sich analog zu den Kosten eines Trainings.

Treffen mit dem Industriepartner: Da Gifhorn ungefähr 400 km vom Projektsitz in Darmstadt entfernt ist, macht es Sinn Inhalte, für die ein Treffen in Präsenz sinnvoll ist, zu bündeln und ein größeres, anderthalbtägiges Treffen statt mehreren kleineren zu planen. Die Errechnung der Gesamtkosten ergibt sich dementsprechend analog zu der Berechnung der anderen nationalen Reisen. Regelmäßige Absprachen werden Online stattfinden.

Konferenzen: Die letzten Kosten entfallen auf den Besuch einer nationalen und einer internationalen Konferenz. Auf Grund der Forschung mit deutschsprachigen Texten in einem Lernszenario bieten sich hier die European Conference on Technology Enhanced Learning (EC-TEL 2022) und die Fachtagung Bildungstechnologien der Gesellschaft für Informatik (DEFLI 2022) an. Beide Konferenzen gelten in der betreffenden Forschungsgemeinschaft als reputabel und sind vor allem für ihre wertvollen Communities bekannt. Da für keine der beiden Konferenzen bisher der Standort feststeht, nehmen wir für die Kalkulation die Standorte aus 2019 an (in 2020 fanden beide Konferenzen Online statt). Die DELFI 2019 fand in Berlin und die EC-TEL 2019 in Delft in den Niederlanden statt. Die Konferenzgebühr, sowie die An-

Studie	# Fragen	Min/Frage	# Teilnehmer	# Min Gesamt	€/Min	Kosten
# 1	40	25	10	10.000	0,20	2.000 €
# 2	10	30	100	30.000	0,20	6.000 €
Summe						8.000 €

Tabelle 5: Berechnung der Kosten für die Benutzerstudien

und Abreisekosten orientieren sich ebenfalls an den Kosten, die beim Besuch der Konferenzen in 2019 angefallen sind. Die Übernachtungs- und Tagegeldkosten der DELFI errechnen sich durch 5 Nächte je 80 € und 5 volle Tage je 28 € sowie einem Teiltag zu 14 €. Analog sind 5 Nächte je 122 € (ARVVwV), 5 volle Tage je 39 € und ein Teiltag zu 31 € in die Berechnung der EC-TEL 2022 Übernachtungs- und Tagegeldkosten eingeflossen.

4.9.3 Sonstige allgemeine Verwaltungsausgaben

Die 8.000 € auf diesem Posten sind für die in diesem Projekt notwendigen Benutzerstudien reserviert. Die Kosten dieses Postens sind schwer abzuschätzen, da die genaue Anzahl an zu evaluierenden Fragen erst im Rahmen des Projekts bekannt sein wird. Für die Kalkulation schätzen wir 40 Fragen in der ersten Studie und 10 Fragen in der zweiten Evaluation. Diese Schätzung erlaubt für jede finale Frage 4 Iterationen zur Verfeinerung der Klarheit, Präzision und Verständlichkeit. Sollte die tatsächliche Fragenanzahl im Laufe des Projekts abweichen, kann die Zahl der Studienteilnehmer entsprechend reduziert oder erhöht werden. Des Weiteren nehmen wir einen Stundenlohn von 12 € für Studienteilnehmer an. Sollten die Benutzerstudien auf Amazon Mechanical Turk stattfinden, entspricht dies einem Stundenlohn von 10 € und 20 % Gebühr an Amazon. Die Kalkulation der Kosten ist in Tabelle 5 zu sehen. Pro Frage werden einem Studienteilnehmer in Studie #1 25 Minuten zur Beantwortung der Frage und Ausfüllen des Fragebogens eingeräumt. Da je Bescheid mehrere Fragen zu evaluieren sind, wird auch ein Anteil der Lesezeit des Bescheids eingerechnet. In Studie #2 sind zusätzlich 5 Minuten für das Lesen des systemgegebenen Feedbacks angedacht. Da das Ziel der ersten Studie in der iterativen Entwicklung und internen Validierung der Fragen liegt, sind hier 10 Evaluierende je Frage ausreichend. Die Teilnehmerzahl wird in der zweiten Studie auf 100 erhöht, um eine größere Aussagekraft der erzielten Ergebnisse zu erreichen.

5 Verwertungsplan

Dieses Kapitel stellt Verwertungsmöglichkeiten der in diesem Projekt erzielten Ergebnisse nach Projektende dar. Dies geschieht zum einen aus wirtschaftlicher (5.1) und wissenschaftlicher (5.2) Perspektive. Zum anderen wird in Sektion 5.3 auf weiterführende Arbeiten, Projekte und Ideen eingegangen.

5.1 Wirtschaftliche Erfolgsaussichten

Insbesondere zwei der in diesem Projekt untersuchten Technologien, die automatische Sprachverarbeitung deutscher Verwaltungssprache und das Feedbacksystem, sind aus wirtschaftlicher Perspektive attraktiv. Generell können NLP Modelle, die auf deutsche Verwaltungssprache spezialisiert sind, eingesetzt werden, um die Verarbeitung rechtlicher Dokumente zu unterstützen. Zeitersparnisse durch die Übernahme von Teilschritten oder Vorverarbeitung rechtlicher Dokumente zur Erleichterung des menschlichen Gebrauchs sind besonders wertvoll, da vor allem teure Arbeitszeit hoch qualifizierter Juristen eingespart werden kann. In diesem Projekt erzielte Erkenntnisse können durch die Kooperation mit IAV direkt genutzt werden, um weitere Arbeiten auf dem Weg zum automatisierten Rechtsverständnis zu informieren. Dabei können nicht nur methodische und fachliche Erkenntnisse von Nutzen sein, sondern auch Know-How über eine produktive Gestaltung des interdisziplinären Austauschs zwischen rechtlichen Experten und Informatikern.

Das Feedbacksystem bietet ebenfalls vielfältige Möglichkeiten Kosten einzusparen und neue Märkte zu erschließen. Kommerzielle automatische Korrektursysteme existieren bereits, sind allerdings in den angebotenen Aufgabenformaten und Domänen sowie den möglichen Feedbacks beschränkt. In diesem Projekt erzielte Erkenntnisse könnten hier existierende Beschränkungen lockern oder gar beseitigen. Je nach Qualität des in diesem Projekt erstellten Korrektur- und Feedbacksystems ist auch ein direkter Einsatz des Modells in einer “Verstehe deine Bescheide”-App denkbar. Werden solche Systeme in Weiterbildungen oder Trainings eingesetzt, können Kosten durch Lehrpersonal reduziert oder die Qualität der Lehre erhöht und dadurch besser qualifiziertes Personal ausgebildet werden. Dieses Projekt kann auch IAV bei der Entscheidung, solche Systeme für interne Lehrgänge einzusetzen, als Proof of Concept dienen.

Da sowohl die in diesem Projekt erzielten Erkenntnisse in Form von Publikationen als auch erstellter Code und die trainierten Modelle als Open Source verfügbar gemacht werden, ist eine weitere Verwendung durch andere Firmen und Wissenschaftler möglich.

5.2 Wissenschaftlich-technische Erfolgsaussichten

Dieses Projekt deckt drei wesentliche Beiträge zum wissenschaftlichen Stand der Technik in seiner Zielsetzung ab. Der erste Beitrag besteht in einem frei verfügbaren Korpus aus kategorisierten Bescheiden, dazugehörigen Verständnisfragen und eine Sammlung manuell evaluierter Bürgerantworten. Dieser kann anderen Firmen und Wissenschaftlern als Grundlage für eigene Projekte und Studien dienen. Beispielsweise sind die gegebenen Bürgerantworten relevant für Studien, die allgemein die Kommunikation zwischen Bürgern und Behörden untersuchen. Erkenntnisse über den Umgang mit deutscher Verwaltungssprache in modernen NLP-Modellen bilden den zweiten Beitrag. Diese sind wertvoll für alle Entwickler von Systemen und Wissenschaftlern in dieser Domäne. Der letzte Beitrag besteht in der Untersuchung möglicher Feedbacksysteme, die über den Stand der Technik hinausgehen. Im Gegensatz zu existierenden Ansätzen werden hier Freitextantworten evaluiert, ohne Referenzantworten oder Referenzmodelle der Domäne zu besitzen und daraus Feedback für den Bürger generiert.

Die in diesem Projekt gestellten Forschungsfragen sind sowohl innovativ als auch für die technologiegestützte Bildungsforschungsgemeinschaft von höchster Relevanz. Erzielte Ergebnisse lassen sich dementsprechend in nationalen und internationalen Konferenzen publizieren sowie für die Promotion der Projektleiterin verwenden. Des Weiteren findet im Rahmen dieses Projekts ein interdisziplinärer Austausch zwischen den Rechtswissenschaften und der Informatik statt, aus dem sowohl neue Forschungs- und Projektideen als auch langfristige Forschungskontakte resultieren können.

5.3 Wissenschaftliche und wirtschaftliche Anschlussfähigkeit

Dieses Projekt kann als Grundlage für diverse weiterführende Arbeiten dienen. Zum einen sind Arbeiten denkbar, die unmittelbar das in diesem Projekt erstellte, prototypische Feedbackmodell weiterentwickeln. Es kann beispielsweise um weitere Fragen, Trainingsdaten und die damit einhergehende Steigerung der Prädiktionskraft oder umfassendere Feedbackgenerierungsmethoden erweitert werden. Darüber hinaus könnte es in einer realen Applikation integriert und so in einem Echtweltszenario erprobt werden. Hierfür ist es zunächst nötig das System bis zur Produktreife zu verfeinern und um weitere Komponenten, wie eine automatische Texterkennung aus einem Foto des Bescheides, zu erweitern. Im Rahmen einer solchen praktischen Umsetzung könnten ebenfalls tiefergehende und größere Benutzerstudien durchgeführt werden, die beispielsweise den langfristigen Lerneffekt durch das Verwenden eines solchen System und den Wissenstransfer auf andere Rechtsdokumente untersuchen.

Teile des Systems und gewonnene Erkenntnisse können auch in verwandten Bereichen, wie beispielsweise Betriebsinternen Schulungen, weiterentwickelt und verwendet werden. Bei ausreichender Präzision und Generalisierbarkeit des Feedbackmodells ist auch ein unterstützender Einsatz und dementsprechend weiterführende Evaluation im Übungsbetrieb diverser Lehrveranstaltungen, zum Beispiel "Kommunikationsnetze I", denkbar. Hierfür ist wahrscheinlich ein erneutes Trainieren des Modells nötig, um auch in neuen Domänen verlässlich einsetzbar zu sein. Letztlich sind direkt anschließende weiterführende Projekte je nach gewonnenen Erkenntnissen möglich.

Literatur

- [1] S. Burrows, I. Gurevych, and B. Stein, “The eras and trends of automatic short answer grading,” *International Journal of Artificial Intelligence in Education*, vol. 25, no. 1, pp. 60–117, 2015.
- [2] M. Blaha, “Nur für Eingeweihte? Das Amt und seine Sprache,” 2017, [Online; accessed 5-August-2020]. [Online]. Available: <https://www.bpb.de/apuz/245595/nur-fuer-ingeweihte-das-amt-und-seine-sprache?p=1>
- [3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *arXiv e-prints*, 2019.
- [4] S. H. Edwards, “Using test-driven development in the classroom: Providing students with automatic, concrete feedback on performance,” in *Proceedings of the international conference on education and information systems: technologies and applications EISTA*, vol. 3. Citeseer, 2003.
- [5] S. Parihar, Z. Dadachanji, P. K. Singh, R. Das, A. Karkare, and A. Bhattacharya, “Automatic grading and feedback using program repair for introductory programming courses,” in *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education*, 2017, pp. 92–97.
- [6] A. Willis, “Using nlp to support scalable assessment of short free text responses,” in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, pp. 243–253.
- [7] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, “Automatic short answer grading and feedback using text mining methods,” *Procedia Computer Science*, vol. 169, pp. 726–743, 2020.
- [8] C. Sung, T. I. Dhamecha, and N. Mukhi, “Improving short answer grading using transformer-based pre-training,” in *International Conference on Artificial Intelligence in Education*. Springer, 2019, pp. 469–481.
- [9] E. Leitner, G. Rehm, and J. Moreno-Schneider, “A dataset of german legal documents for named entity recognition,” *arXiv preprint arXiv:2003.13016*, 2020.
- [10] J. Moreno-Schneider, G. Rehm, E. Montiel-Ponsoda, V. Rodriguez-Doncel, A. Revenko, S. Karampatakis, M. Khvalchik, C. Sageder, J. Gracia, and F. Maganza, “Orchestrating nlp services for the legal domain,” *arXiv preprint arXiv:2003.12900*, 2020.
- [11] L. Camus and A. Filighera, “Investigating transformers for automatic short answer grading,” in *International Conference on Artificial Intelligence in Education*. Springer, 2020, pp. 43–48.
- [12] A. Filighera, T. Steuer, and C. Rensing, “Fooling automatic short answer grading systems,” in *International Conference on Artificial Intelligence in Education*. Springer, 2020, pp. 177–190.
- [13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [14] S. Arlot, A. Celisse *et al.*, “A survey of cross-validation procedures for model selection,” *Statistics surveys*, vol. 4, pp. 40–79, 2010.