

E-DELIB – ”Powering up E-DELIBeration: Towards AI-supported moderation” Schlussbericht Teil II: Sachbericht *

Project Leader: Jun.Prof. Dr. Gabriella Lapesa
FKZ: 01IS20050
Project duration: 01.10.2020/31.12.2024

1 Einleitung

E-Demokratie ist die digital erweiterte und transformierte Version der direkten Demokratie. Digitale Innovationen ermöglichen es demokratischen Prozessen, eine breite Öffentlichkeit zu erreichen, was die Bürgerbeteiligung und neue Formen der Zusammenarbeit begünstigt. Das Internet spielt dabei natürlich eine Schlüsselrolle. Ein entscheidender Engpass ist jedoch die Notwendigkeit, groß angelegte Diskussionen unter den Bürgern zu optimieren, die Generierung von Ideen und Vorschlägen zu erleichtern und sicherzustellen, dass die Diskussion nicht nur respektvoll, sondern auch produktiv ist. Die Lösung ist Moderation (”Facilitation” in der sozialwissenschaftlichen Literatur), die jedoch nicht in großem Umfang und in hoher Qualität erreicht werden kann, wenn nur wenige menschliche Moderatoren die Diskussionen großer Menschenmengen optimieren müssen. Durch die Unterstützung der Moderation kann NLP eine entscheidende Rolle bei der groß angelegten E-Partizipation spielen. Genau das war das Ziel von E-DELIB: die Entwicklung von NLP-Methoden zur Unterstützung menschlicher Moderatoren bei der E-Partizipation. Die Forschungsagenda von E-DELIB hat sich auf drei Forschungsfragen konzentriert:

- RQ1 Was sind die empirischen Merkmale einer erfolgreichen E-Deliberation (d. h. einer Deliberation, die in einem Online-Kontext stattfindet)?
- RQ2 Wie agieren Moderatoren in der E-Deliberation?
- RQ3 Ist eine NLP-basierte Moderation möglich und was ist die beste Strategie, um sie in der realen E-Deliberation einzuführen?

*Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS20050 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin.

Um diese Forschungsfragen zu beantworten, wurden die Forschungsaktivitäten von E-DELIB in zwei große Untersuchungsbereiche unterteilt.

- **Bereich 1: Integration von Argument Mining und Deliberative Theory.** Argument Mining (AM) ist ein Teilbereich der natürlichen Sprachverarbeitung, der Argumentation aus der Perspektive der Begründung und Überzeugung untersucht. Die Deliberative Theorie ist ein Teilgebiet der Politikwissenschaft, das sich mit Entscheidungsfindung in der deliberativen Demokratie befasst. Entscheidend ist, dass beide Disziplinen dieselbe Frage gestellt haben: Was macht einen Beitrag zu einer Diskussion (z. B. einen Beitrag in einem Benutzerforum) gut? Interessanterweise war die Schnittstelle zwischen Argument Mining und Deliberative Theory zu Beginn von E-DELIB noch nicht erforscht worden: Durch die starke Betonung dieses interdisziplinären Aspekts haben sich E-DELIB und seine Mitglieder als Bezugspunkt in der AM-Community etabliert und einen fruchtbaren Austausch mit der sozialwissenschaftlichen Community ermöglicht. Aus projektspezifischer Sicht war die Integration von Argument Mining und Deliberative Theory entscheidend für die Beantwortung der Forschungsfragen RQ1 und RQ2 und floss schließlich, wie weiter unten ausführlich erläutert wird, in den Modellierungsansatz zur Beantwortung der Forschungsfrage RQ3 ein.
- **Bereich 2: Empirische Modellierung des Moderatorenverhaltens.** Die Verfügbarkeit von menschlich moderierten Daten, d. h. Datensätzen mit Diskussionen und Interventionen menschlicher Moderatoren, war für die Forschungsagenda von E-DELIB von entscheidender Bedeutung. Das Vorhandensein einer Moderatorintervention an einem bestimmten Punkt einer Diskussion ist die Art von "Moderationssignal", das E-DELIB in den Datensätzen erfassen und mit NLP-Modellen reproduzieren wollte. Im besten Fall möchten wir so viele von Menschen moderierte Daten wie möglich aus verschiedenen Domänen und in verschiedenen Sprachen haben, um daraus robuste Modelle erstellen zu können. Entscheidend ist, dass wir die Handlungen menschlicher Moderatoren (in der sozialwissenschaftlichen Literatur auch als "Interventionen" bezeichnet) in realen Diskussionsszenarien als theorieneutrale Komponente des "Moderationssignals" betrachten können (im Gegensatz zu den theoriegeleiteten Operationalisierungen, die Argument Mining und Deliberative Theory liefern), selbst wenn diese Moderatoren nach bestimmten Richtlinien geschult sind.

In Abschnitt 2 geben wir dem Leser einen Überblick über die im Rahmen des Projekts durchgeführten Forschungsarbeiten, wobei wir uns an einem chronologischen Kriterium orientieren und die Diskussion nach Jahren gliedern. Wenn wir auf Veröffentlichungen verweisen, zitieren wir diese anhand des Indexes, der in der diesem "Schlussbericht" beigefügten Publikationsliste aufgeführt ist.

Im Laufe des Projekts sind eine Reihe von Herausforderungen aufgetreten, sowohl aufgrund der rasanten Forschungsentwicklungen in diesem Bereich (z. B. der Aufstieg großer Sprachmodelle und das Potenzial zur Generierung von Moderationskommentaren)

als auch aufgrund logistischer Faktoren (z. B. Wechsel im Projektteam, verzögerte Datenbereitstellung durch Kooperationspartner). Diese Herausforderungen führten zu Aktualisierungen des Projektplans, die wir in Abschnitt 3 diskutieren.

Wir schließen mit Abschnitt 4, in dem ausdrücklich auf einige erforderliche Felder eingegangen wird, die in der offiziellen Vorlage für den "Schlussbericht" aufgeführt sind.

2 Überblick über die Projektarbeit

2.1 Erstes Jahr (Oktober 2020/Dezember 2021)

In unserem ersten Jahr¹ konzentrierten wir uns hauptsächlich auf RQ1 und RQ2, und zwar aus der stark interdisziplinären Perspektive, die das Markenzeichen von E-DELIB ist.

Bereich 1: Integration von Argument Mining und Deliberative Theory Anfang 2021 haben wir auf der wichtigsten NLP-Konferenz, der ACL, eine Umfrage veröffentlicht: "Towards Argument Mining for Social Good: A Survey" [1]. Unsere Umfrage zeichnet ein interdisziplinäres Bild von AM, wobei der Schwerpunkt auf seinem Potenzial zur Lösung von Problemen im Bereich der Sozial- und Politikwissenschaften liegt. Genauer gesagt konzentrieren wir uns auf die Herausforderungen von AM im Zusammenhang mit seinen Anwendungen in sozialen Medien und im mehrsprachigen Bereich und gehen dann zu dem viel diskutierten Begriff der Argumentationsqualität über. Wir schlagen eine neuartige Definition der Argumentationsqualität vor, die mit der Definition der deliberativen Qualität aus der sozialwissenschaftlichen Literatur integriert ist. Nach unserer Definition muss die Qualität eines Beitrags auf mehreren Ebenen bewertet werden: dem Beitrag selbst, seinem vorangehenden Kontext und den daraus resultierenden Auswirkungen auf die Entwicklung des bevorstehenden Diskurses. Letzteres hat innerhalb der Community nicht die verdiente Aufmerksamkeit erhalten. Schließlich definieren wir eine Anwendung von AM für das soziale Wohl: die (halb-)automatische Moderation, eine hochintegrative Anwendung, die (a) einen anspruchsvollen Testfall für den von uns vertretenen integrierten Qualitätsbegriff darstellt, (b) die empirische Quantifizierung der Argumentations-/Deliberationsqualität ermöglicht, um von den Entwicklungen in anderen NLP-Bereichen (z. B. Hassredeerkennung, Faktenprüfung, Debiasing) zu profitieren, und (c) dank ihrer realen Anwendung ein eindeutig vorteilhaftes Potenzial auf gesellschaftlicher Ebene hat.

Argumentationsqualität (aus der NLP-Community) und Deliberationsqualität (aus der sozialwissenschaftlichen Community) sind zentrale theoretische Konzepte für E-DELIB. Es überrascht nicht, dass bei konkreten Anwendungen der Engpass in der Verfügbarkeit von annotierten Daten liegt, mit denen Tools trainiert werden können, die Moderatoren unterstützen, indem sie darauf hinweisen, dass ein Nutzer einen Kommentar von geringer Qualität abgegeben hat. Einer der Hauptengpässe bei der Integration

¹Beachten Sie, dass die Postdoktorandin von Februar 2021 bis März 2023 im Elternzeit war, sodass E-DELIB ein Projektmitglied weniger hatte als ursprünglich geplant.

der Begriffe Argumentationsqualität und Deliberationsqualität ist das Ungleichgewicht in der Menge der verfügbaren Annotationen: Einerseits ist die Quantifizierung der Argumentationsqualität eine etablierte Aufgabe in der NLP, und Klassifikatoren sind in unterschiedlichen Granularitätsgraden verfügbar (Gesamtqualität vs. feinkörnige Aspekte); andererseits ist die bestehende Annotation zur deliberativen Qualität zeitaufwändig und wird von Experten durchgeführt, was in der Regel zu kleinen Datensätzen führt, die zudem unter einer starken Klassenungleichheit leiden. In der Forschungsarbeit, die in der Veröffentlichung "Scaling up Discourse Quality Annotation for Political Science" [4] (eingereicht im Oktober 2021, veröffentlicht im Jahr 2022) vorgestellt wird, konzentrieren wir uns auf die Verbesserung der automatischen Annotation der deliberativen Qualität. Die Skalierung solcher Annotationen mit automatischen Tools ist wünschenswert, aber sehr anspruchsvoll. Wir nehmen diese Herausforderung an und untersuchen verschiedene Strategien zur Verbesserung der Vorhersage von deliberativen Qualitätsdimensionen (Begründung, Gemeinwohl, Interaktivität, Respekt) in einem Standarddatensatz. Unsere Ergebnisse zeigen, dass einfache Datenvergrößerungstechniken das Datenungleichgewicht erfolgreich verringern. Klassifikatoren, die Annotationen zur Argumentationsqualität integrieren, wurden durch Transformer-basierte Modelle mit oder ohne Datenvergrößerung durchweg übertroffen.

Während die direkte Integration von Deliberative Quality und Argument Quality in [4] nicht erfolgreich war, ist es uns stattdessen gelungen, zwei spezifische Dimensionen von Argument Quality und Deliberative Quality zu integrieren, nämlich die Verwendung von Berichten über persönliche Erfahrungen und Geschichten in der Argumentation. Die Deliberative Theory hat eine klare Sicht auf die Auswirkungen dieses Phänomens: Geschichten kommen im deliberativen Diskurs häufig vor und wirken sich positiv auf die Qualität der Argumentation aus, in der sie vorkommen. Wie sieht es mit der alltäglichen Argumentation aus? Persönliche Erfahrungen oder Geschichten sind ebenso wirkungsvoll (leicht verständlich, steigern die Empathie) wie potenziell komplex (der Bericht kann gleichzeitig Behauptung und Beweis sein), und sie waren zum Zeitpunkt des Projektstarts in der AM noch weitgehend unerforscht. Unser Artikel "Reports of personal experiences and stories in argumentation: datasets and analysis" [3] (eingereicht im Oktober 2021, veröffentlicht im Jahr 2022) füllte diese Lücke. Unsere Herausforderung war die Knappheit an annotierten Daten: Unser Zielphänomen wird in der Deliberative Theory als "Storytelling" und im Argument Mining als "Testimony" annotiert, und unser Ziel war es, vorhandene Annotationen zu nutzen, um die Analyse skalieren zu können. Wir führten eine Reihe von domäneninternen und domänenübergreifenden Experimenten mit verschiedenen Datensätzen, Modellarchitekturen, Trainingskonfigurationen und Feinabstimmungsoptionen durch, die auf die beteiligten Domänen zugeschnitten waren, und konnten zeigen, dass eine gemischte Konfiguration (trainiert sowohl auf Storytelling- als auch auf Zeugnisannotationen) Klassifikatoren, die nur auf einer Art von Annotation trainiert wurden, deutlich übertraf.

Bereich 2: Empirische Modellierung des Moderatorenverhaltens Wie in der Einleitung erläutert, ist die Verfügbarkeit von Daten, die von Menschen moderiert

wurden, für die Untersuchung des Moderatorenverhaltens von entscheidender Bedeutung. Leider hatten wir aufgrund der in Abschnitt 3 beschriebenen Probleme im Jahr 2021 keinen Zugriff auf Daten unseres Kooperationspartners. Konkret bedeutete dies, dass wir uns auf den einzigen zu diesem Zeitpunkt öffentlich zugänglichen moderierten E-Deliberation-Datensatz konzentrieren mussten. Die in unserer Arbeit "Predicting moderation of deliberative arguments: Is argument quality the key?" [2] (eingereicht und veröffentlicht im Jahr 2021) unternimmt erste Schritte in Richtung einer (halb-)automatischen Moderation, indem sie modernste Klassifizierungsmodelle verwendet, um vorherzusagen, welche Beiträge moderiert werden müssen. Wir haben gezeigt, dass diese Aufgabe zwar zweifellos schwierig ist, die Leistung jedoch deutlich über dem Basiswert liegt. Wir haben weiter untersucht, ob die Qualität der Argumente ein wichtiger Indikator für die Notwendigkeit einer Moderation ist, und konnten überraschenderweise zeigen, dass auch hochwertige Argumente eine Moderation auslösen.

Um dem ressourcenarmen, mehrsprachigen Szenario gerecht zu werden, von dem wir annahmen, dass es die Moderationsdatensätze, mit denen wir arbeiten wollten, charakterisieren würde, führten wir Experimente zum sprachübergreifenden Transfer in ressourcenarmen Szenarien durch. Die Arbeit, in der diese Ergebnisse vorgestellt werden, "How to Translate Your Samples and Choose Your Shots? Analyzing Translate-train and Few-shot Cross-lingual Transfer" [7], wurde im Oktober 2021 eingereicht und 2022 veröffentlicht. Translate-Train oder Few-Shot Cross-Lingual Transfer können verwendet werden, um die Zero-Shot-Leistung von mehrsprachigen vortrainierten Sprachmodellen zu verbessern. Angesichts der geringeren Kosten und der höheren Verfügbarkeit von maschineller Übersetzung im Vergleich zu manueller professioneller Übersetzung ist es wichtig, Few-Shot und Translate-Train systematisch zu vergleichen, zu verstehen, wann jeweils welche Methode Vorteile bietet, und zu untersuchen, wie die zu übersetzenden Shots ausgewählt werden sollten, um den Few-Shot-Gewinn zu steigern. Diese Arbeit zielte darauf ab, diese Lücke zu schließen: Wir haben den Leistungsgewinn von Few-Shot im Vergleich zu Translate-Train anhand von drei verschiedenen Basismodellen und einer variierenden Anzahl von Beispielen für drei Aufgaben/Datensätze in 17 Sprachen verglichen und quantifiziert. Wir haben gezeigt, dass die Skalierung der Trainingsdaten mithilfe maschineller Übersetzung einen größeren Gewinn bringt als die Verwendung kleinerer (qualitativ hochwertigerer) Few-Shot-Daten. Wenn Few-Shot vorteilhaft ist, haben wir gezeigt, dass es zufällige Sätze von Beispielen gibt, die über alle Sprachen hinweg eine bessere Leistung erzielen, und dass sowohl die Leistung in Englisch als auch die maschinelle Übersetzung der Beispiele verwendet werden kann, um die zu übersetzenden Beispiele auszuwählen und so den Few-Shot-Gewinn zu steigern.

2.2 Zweites Jahr (2022)

Im Laufe des zweiten Jahres haben wir die Untersuchung von RQ1 und RQ2 fortgesetzt und erste Schritte in Richtung RQ3 unternommen.

Bereich 1: Integration von Argument Mining und Deliberative Theory Da beide Theorien Definitionen dafür liefern, was einen guten Beitrag zu einer Diskus-

sion ausmacht, kann ihre Integration für die Entwicklung von NLP-Methoden zur Unterstützung von Moderatoren von entscheidender Bedeutung sein. In dem Artikel "Bridging Argument Quality and Deliberative Quality Annotations with Adapters" [9] (eingereicht im Oktober 2022, veröffentlicht 2023) nutzen wir die Aufmerksamkeit, die sowohl die Argument-Mining- als auch die Sozialwissenschafts-Community der Definition und Annotation von Argumentqualität gewidmet haben, was zu einer großen Anzahl annotierter Ressourcen und einer Vielzahl von Argumentqualitätsdimensionen geführt hat. Unser Ziel ist es, ein besseres Verständnis dafür zu erlangen, wie die verschiedenen Aspekte der Qualität miteinander in Beziehung stehen. Wir setzen Adapter als Modellierungsstrategie ein, die a) die Vorhersage einzelner Qualitätsdimensionen durch Einbringen von Wissen über verwandte Dimensionen verbessern kann, b) effizient und modular ist und c) als Analysewerkzeug zur Untersuchung der Beziehungen zwischen verschiedenen Dimensionen dienen kann. Wir führen Experimente mit 6 Datensätzen und 20 Qualitätsdimensionen durch. Wir stellen fest, dass die meisten Dimensionen als gewichtete Kombination anderer Qualitätsaspekte gelernt werden können und dass für 8 Dimensionen die Adapterfusion die Qualitätsvorhersage verbessert.

Als weiteren Schritt zur Untersuchung des Einflusses persönlicher Erfahrungen und Erzählungen in der Argumentation wurde in der Veröffentlichung "StoryARG: a corpus of narratives and personal experiences in argumentative texts" (veröffentlicht im Jahr 2023, aber Ergebnis einer einjährigen Annotationsstudie) veröffentlicht. StoryARG wurde aus etablierten Korpora der computergestützten Argumentationsforschung und der Sozialwissenschaften sowie aus Kommentaren zu Artikeln der New York Times zusammengestellt und enthält umfassende Annotationen zur Verwendung von Storytelling für argumentative Zwecke. Unsere Analyse zeigt, wie subjektiv die Präferenzen für bestimmte Arten von Geschichten sind.

Bereich 2: Empirische Modellierung des Moderatorenverhaltens Im zweiten Teil des Artikels [9] kehrten wir zum Kerninteresse des Projekts zurück und zeigten die Vorteile der Integration von Argumentationsqualität und Deliberationsqualität zur Verbesserung der Leistung bei der Vorhersage von Moderatoreninterventionen in einem deliberativen Forum. Anhand derselben Datensätze und Datensplits, die wir in Artikel [2] verwendet haben, zeigen wir, dass ein solcher integrativer Ansatz die Standardvorhersagemodelle übertrifft und zusätzlich den wünschenswerten Vorteil bietet, dass er erklärbare Vorhersagen liefert (d. h. "dieser Kommentar muss aufgrund mangelnder Qualität in diesem bestimmten Aspekt moderiert werden").

Menschliche Moderatoren stehen je nach Plattform, auf der die Diskussion stattfindet, vor unterschiedlichen Aufgaben. In den Artikeln [2,9] haben wir uns auf ein Standardforum konzentriert, aber es gibt auch alternative Plattfortmtypen (und damit auch alternative Arten von Moderatoreneingriffen).

Argumentationskarten wie Kialo (kialo.com) strukturieren Diskurse in Knotenpunkten in einem Baum, wobei jeder Knotenpunkt ein Argument darstellt, das das übergeordnete Argument unterstützt oder ablehnt. Dieses Format ist verständlicher und weniger redundant als ein unstrukturiertes Format. Das Erkunden dieser Karten und das

Aufrechterhalten ihrer Struktur durch das Platzieren neuer Argumente unter geeigneten übergeordneten Argumenten ist für Benutzer mit großen Karten, wie sie in Online-Diskussionen typisch sind, eine größere Herausforderung. Dementsprechend ist dies eine typische Aufgabe für menschliche Moderatoren. In dem Artikel "Node Placement in Argument Maps: Modeling Unidirectional Relations in High & Low-Resource Scenarios" (eingereicht 2022, veröffentlicht 2023) haben wir die Aufgabe der Knotenplatzierung vorgestellt, um Nutzer (und Moderatoren) von Argumentationskarten zu unterstützen. Zusammen mit der Aufgabendefinition haben wir eine Obergrenze für die menschliche Leistungsfähigkeit festgelegt und Experimente mit Modellen unterschiedlicher Größe und Trainingsstrategien durchgeführt. Basierend auf einer Annotationsstudie heben wir die Mehrdeutigkeit der Aufgabe hervor, die sie sowohl für Menschen als auch für Modelle zu einer Herausforderung macht. Wir untersuchen die unidirektionale Beziehung zwischen Baumknoten und zeigen, dass die Kodierung eines Knotens in unterschiedliche Einbettungen für jeden der "Parent" und "Child" Fälle die Leistung verbessert. Darüber hinaus zeigen wir die Few-Shot-Effektivität unseres Ansatzes.

2.3 Drittes Jahr (2023)

Im dritten Jahr haben wir einen entscheidenden Schritt zur Lösung des Problems der geringen Ressourcenunität unternommen, indem wir einen neuen Moderationsdatensatz gesammelt haben, und wir haben weitere gezielte Schritte zur Erforschung der Schnittstelle zwischen Argument Mining und Deliberative Theory unternommen.

Bereich 1: Integration von Argument Mining und Deliberative Theory Im Jahr 2023 haben wir die Forschungslinie fortgesetzt, die sich mit dem Einfluss von Narrativen und persönlichen Erfahrungen in Benutzerforen befasst. In dem Artikel "Stories and personal experiences in the COVID-19 discourse" [18] (eingereicht im Oktober 2023, veröffentlicht 2024) verwenden wir den StoryARG-Datensatz [10], um Klassifizierungsmodelle zu trainieren, mit denen wir die Verwendung von Storytelling im COVID-19-Diskurs weiter untersuchen können.

Anschließend widmen wir uns der Untersuchung soziodemografischer Verzerrungen in Nutzerforen. In dem Artikel "Self-reported demographics and discourse dynamics in a persuasive online forum" [19] (eingereicht im Oktober 2023, veröffentlicht 2024) analysieren wir die Diskursdynamik, die durch soziodemografische Selbstoffenbarungen ausgelöst wird (z. B. "Als Frau denke ich, dass die beste Lösung für dieses Problem ... ist" ; "Als weißer Mann glaube ich, dass ..."). Diese Arbeit ist eine Zusammenarbeit mit A. Falenska (Universität Stuttgart, Gruppenleiterin am Interchange Forum for Reflection on Intelligent Systems) und ihre Erkenntnisse sind auch für die Moderation von entscheidender Bedeutung – da Selbstoffenbarungen die Nutzer bloßstellen und eine wichtige Aufgabe des Moderators darin besteht, sicherzustellen, dass die Diskussion höflich und respektvoll bleibt.

Unsere letzte Arbeit in diesem Bereich für 2023 ist ein Positionspapier, das sich mit der Anwendung von LLMs auf die Aufgabe der Bewertung der Argumentationsqualität befasst. Dieses Papier mit dem Titel "Argument Quality assessment in the

age of instruction-following Large Language Models” (eingereicht im Oktober 2023, veröffentlicht 2024) [20] ist das Ergebnis der Zusammenarbeit mit einer großen Gruppe internationaler Co-Autoren und zeigt die zentrale Rolle, die E-DELIB in der Community mittlerweile spielt.

Bereich 2: Empirische Modellierung des Moderatorenverhaltens Neben der Suche nach Modellierungslösungen für Szenarien mit geringen Ressourcen besteht eine naheliegende (aber kostspielige) Strategie darin, mehr Daten für das Zielphänomen zu sammeln und zu annotieren: In dem Papier ”Moderation in the Wild: Investigating User-Driven Moderation in Online Discussions” [17] haben wir diesen Ansatz verfolgt, indem wir das Phänomen der ”Benutzermoderation” untersucht haben, d. h. die Tatsache, dass in großen Online-Diskussionen oft Benutzer die Moderatorenrolle übernehmen und andere Benutzer mit Kommentaren ansprechen, die auf die Verbesserung der Diskussionsqualität abzielen (z. B. indem sie um Klarstellung oder weitere Belege bitten), und damit effektiv als Moderatoren fungieren. Durch eine groß angelegte Crowdsourcing-Studie und eine gründliche Analyse ermitteln wir die Eigenschaften der Nutzer Moderation und sammeln einen Datensatz, den wir zur Unterstützung der Vorhersage von Moderation weiter verwenden.

2.4 Viertes Jahr (2024)

Im Jahr 2024 hat sich der Schwerpunkt von E-DELIB hauptsächlich auf die Bearbeitung von RQ3 verlagert, wobei die Integration der neu gesammelten Ressourcen und der in unserer bisherigen Arbeit etablierten Methoden im Vordergrund steht.

Bereich 1: Integration von Argument Mining und Deliberative Theory Die Bewertung der Argumentationsqualität hängt von gut etablierten logischen, rhetorischen und dialektischen Eigenschaften ab, die unvermeidlich subjektiv sind: Es kann mehrere gültige Bewertungen geben, es gibt keine eindeutige Grundwahrheit. Diese Beobachtung steht im Einklang mit den jüngsten Entwicklungen im Bereich des maschinellen Lernens, die die Koexistenz unterschiedlicher Perspektiven begrüßen. Dieses Potenzial ist jedoch in der NLP-Forschung zur Argumentations- und Deliberationsqualität noch weitgehend unerforscht. Ein entscheidender Grund dafür scheint die noch unerforschte Verfügbarkeit geeigneter Datensätze zu sein. Der Artikel ”Towards a Perspectivist Turn in Argument Quality Assessment” [24] (eingereicht im Oktober 2024, veröffentlicht 2025) schließt diese Lücke durch eine systematische Überprüfung von Datensätzen zur Argumentations- und Deliberationsqualität, wobei ein besonderer Schwerpunkt auf den Merkmalen der Annotatoren liegt.

Ebenfalls im Jahr 2024 hat das Projektteam weitere Schritte bei der Untersuchung der Dynamik persönlicher Argumentation und des Storytelling unternommen. Konkret hat eine der Doktorandinnen, Frau Neele Falk, einen Forschungsaufenthalt an der University of Michigan absolviert und dort gemeinsam mit Prof. David Jurgens an einer Annotationsstudie gearbeitet, die eine der möglichen Ursachen für die subjektive Wahrneh-

mung und Interpretation von Argumenten untersucht: den Unterschied zwischen den moralischen und kulturellen Präferenzen der Verfasser und der Leser eines bestimmten Textes. Diese Zusammenarbeit wird derzeit in einer Veröffentlichung zusammengefasst (siehe Veröffentlichungsliste, Abschnitt 2, b).

Bereich 2: Empirische Modellierung des Moderatorverhaltens Das Jahr 2024 war für die Modellierung des Moderatorenverhaltens durch E-DELIB von entscheidender Bedeutung. Wir haben zwei groß angelegte Nutzerstudien durchgeführt, die sich auf die beiden grundlegenden Schritte im Moderationsworkflow konzentrierten: a) die Vorhersage des Moderationsbedarfs (muss dieser Kommentar moderiert werden?) und b) die Generierung des Moderationskommentars mit LLMs.

Was die Moderationsvorhersage betrifft, haben wir einen großen Annotationsaufwand betrieben, um PerspectiveMod zu sammeln, einen Datensatz zur Moderationsvorhersage, der in seiner absichtlichen Variation einzigartig ist, und zwar über (a) das Niveau der Moderationserfahrung, das in den Quelldaten eingebettet ist (professionelle vs. nicht-professionelle Moderationsumgebungen), (b) den Annotatorprofilen (Experten vs. geschulte Crowdworker) und (c) der Fülle der einzelnen Moderationsurteile, sowohl in Bezug auf detaillierte Kommentarmerkmale (aus der Argumentations- und Deliberationstheorie) als auch in Bezug auf die Darstellung der Individualität des Annotators (soziodemografische Merkmale und Einstellungen gegenüber der Aufgabe). Der entsprechende Artikel wird derzeit geprüft (siehe Publikationsliste, Abschnitt 2, b).

Was die Generierung von Moderationskommentaren betrifft, so nutzen wir in dem Artikel "It Is Not Only the Negative that Deserves Attention! Understanding, Generation & Evaluation of (Positive) Moderation" [21] (eingereicht 2024, veröffentlicht 2025) nutzen wir das Wissen, das in den Annotationsrichtlinien kodiert ist, nach denen Moderatoren geschult werden, und experimentieren mit der Generierung von Moderationskommentaren (z. B. wenn wir wissen, dass ein Kommentar aus einem bestimmten Grund moderiert werden muss, schlagen wir Moderatoren Texte vor, die sie für die Nutzer bearbeiten oder genehmigen können). Wir fördern das Verständnis von positiver Moderation, indem wir einen Datensatz mit mehreren Moderationsmerkmalen annotieren, z. B. Neutralität, Klarheit und Neugier. Wir extrahieren Anweisungen aus professionellen Moderationsrichtlinien und verwenden sie, um LLaMA zur Generierung einer solchen Moderation anzuregen. Darauf folgt eine umfassende Bewertung, die zeigt, dass Annotatoren die generierte Moderation höher bewerten als die professionelle Moderation, aber im paarweisen Vergleich dennoch leicht die professionelle Moderation bevorzugen, und dass LLMs als effiziente Alternative zur Schätzung der menschlichen Bewertung verwendet werden können. Die zweite Doktorandin des Projekts, Frau Iman Jundi, hat Ende 2024 einen Forschungsaufenthalt an der Universität Cambridge absolviert, um in Zusammenarbeit mit Prof. Andreas Vlachos das Generierungspotenzial von LLMs zur Umschreibung von Texten unter Reduzierung ihrer Medienvoreingenommenheit weiter zu erforschen.

Zuletzt hat der PI im Mai 2024 den DELITE2024-Workshop ("The First Workshop on Language-driven Deliberation Technology") mitorganisiert, der im Rahmen

der LREC-COLING-Konferenz in Turin stattfand (weitere Details zu diesem Workshop finden Sie im Abschnitt "Editorship" der Publikationsliste). Dieser Workshop konzentrierte sich genau auf das Thema des E-DELIB-Projekts und umfasste eine Präsentation der E-DELIB-Mitglieder, eine Podiumsdiskussion mit den akademischen Partnern von E-DELIB sowie mehrere externe Beiträge. Damit bot er eine hervorragende Gelegenheit zum Networking und ersetzte effektiv den internen Workshop, den wir im ersten Jahr des Projekts organisieren wollten (was aufgrund der COVID-Beschränkungen nicht möglich war).

3 Änderungen am ursprünglichen Projektplan

Wie in der Einleitung dieses Berichts bereits erwähnt, hat eine Vielzahl unterschiedlicher Faktoren im Laufe der Jahre dazu geführt, dass die Projektziele neu definiert werden mussten. In Abschnitt beschreibe ich diese Gründe und gehe anschließend darauf ein, wie die ursprünglich festgelegten Meilensteine erreicht wurden und wann dies geschah.

3.1 Gründe für Änderungen am ursprünglichen Plan

Erstens sollte das Projekt zwar am 1. Oktober 2020 beginnen (und hat auch begonnen), doch verzögerte sich die endgültige Genehmigung erheblich (bis Ende September 2020), sodass es einfach nicht möglich war, die Arbeitsverträge für die Projektmitarbeiter auszustellen. Infolgedessen begann der Postdoktorand einen Monat später, ein Doktorand 1,5 Monate später und der andere 3 Monate später. Das Projekt konnte erst drei Monate nach dem geplanten Start mit voller Personalstärke beginnen. Um die Auswirkungen der verspäteten Genehmigung so gering wie möglich zu halten, gelang es mir mit der Unterstützung (und großem Einsatz) der Verwaltung der Universität Stuttgart, meine PI-Stelle noch im Oktober anzutreten – und das war nur möglich, weil ich bereits am IMS angestellt war. Die ersten Monate des Projekts investierte ich dann in die Vorbereitung eines Übersichtsartikels auf der Grundlage des Antrags, der dann im Januar 2021 mit dem Projektteam fertiggestellt und im Mai 2021 auf der renommiertesten Konferenz für natürliche Sprachverarbeitung veröffentlicht wurde [1].

Zweitens wurde das Projekt von den COVID-Beschränkungen beeinträchtigt, die den Start der gemeinsamen Arbeit einer neu gegründeten Gruppe erschwerten und auch die Vernetzung weniger effektiv machten. Trotzdem hat die Gruppe im ersten Jahr große Fortschritte gemacht, wie die Veröffentlichungen auf den wichtigsten Konferenzen 2021 und 2022 und die Sichtbarkeit unserer Arbeit zeigen (z. B. meine Einladung zur Teilnahme an einer Podiumsdiskussion im Rahmen eines auf das Projektthema spezialisierten Workshops auf einer der wichtigsten Konferenzen in diesem Bereich bereits im Dezember 2021).

Drittens ging die Postdoktorandin wenige Monate nach Projektbeginn in Mutterschaftsurlaub. Trotz meiner Bemühungen (sowohl durch Anrufe als auch über Netzwerke und Kontakte) war es einfach nicht möglich, eine geeignete Ersatzkraft zu finden. Eine einjährige Stelle, die aus bürokratischen Gründen nur jemandem angeboten werden

konnte, der in Deutschland wohnt (oder bereit ist, nach Deutschland zu ziehen), war für das gesuchte Postdoc-Profil einfach nicht attraktiv genug.

Drittens habe ich ab Oktober 2023 eine Doppelstelle als Juniorprofessorin mit Tenure-Track in Responsible Data Science und Machine Learning an der HHU Düsseldorf und als Teamleiterin für Data Science Methods am Leibniz-Institut für Sozialwissenschaften (GESIS) in Köln angetreten. Auch hier war es nicht optimal, meine eigene Projektleitungsposition für einen relativ kurzen Vertrag (ein Jahr) aufzugeben, und ich beschloss stattdessen, die Finanzierung beizubehalten und sie für Folgendes zu verwenden: a) Einstellung einer wissenschaftlichen Mitarbeiterin, Carlotta Quensel einzustellen, die eigentlich bis zum Ende Teil des Projektteams sein sollte, aber eine Doktorandenstelle an anderer Stelle angeboten bekam und nur zwei Monate blieb (statt der geplanten sechs oder neun Monate, je nach kostenneutraler Verlängerung), und b) eine sechsmonatige kostenneutrale Verlängerung für das gesamte Projektteam zu unterstützen, in der Hoffnung, die Verzögerungen im Zusammenhang mit der Datenbereitstellung aufzuholen (leider wurde die Verlängerung nur für drei statt für sechs Monate gewährt). Es sei darauf hingewiesen, dass ich zwar seit Oktober 2023 nicht mehr bei E-DELIB beschäftigt bin, aber in den Verhandlungen über meine Anstellung bei GESIS klar zum Ausdruck gebracht habe, dass ich die Leitung von E-DELIB übernehmen und einen halben Tag pro Woche für das Projekt und Einzelgespräche aufwenden würde, was, wie die Veröffentlichungen zeigen, sehr gut funktioniert hat.

Schließlich war auch der wichtigste Kooperationspartner, ZebraLog, von der COVID-Krise betroffen. Der erhöhte Bedarf an der Organisation von Online-Beteiligungsprozessen führte zu einem Druck, das Portal neu zu strukturieren. Infolgedessen wurden die Daten, die wir zu Beginn des Projekts von ihnen erwartet hatten, tatsächlich erst im Frühjahr 2022 bereitgestellt (diese Verzögerung war auch auf die Notwendigkeit zurückzuführen, die rechtlichen Grundlagen für die Zusammenarbeit und den Datenaustausch zu schaffen). Selbst als wir die Daten schließlich erhielten, waren sie in einer viel geringeren Menge vorhanden, als wir für das Training von Machine-Learning-Modellen benötigt hätten. Wir haben dies durch zusätzliche Annotationsbemühungen zum Sammeln neuer Datensätze und durch Investitionen in die Grundlagenforschung des Projekts kompensiert, was, wie unsere Veröffentlichungen zeigen, sehr erfolgreich war.

3.2 Meilensteine

Im Folgenden beschreibe ich die ursprünglich im Antrag geplanten Meilensteine sowie wie und wann sie erreicht wurden.

Meilenstein 1: Erste Veröffentlichung des E-Deliberation-Repositoriums (ursprünglich für Dezember 2021 geplant) Dieser Meilenstein wurde stark durch die einjährige Abwesenheit des Postdoktoranden beeinträchtigt, der für die infrastrukturellen Aspekte des Projekts verantwortlich war. Die verzögerte Datenbereitstellung durch den Kooperationspartner (und generell die bereits im Abschnitt "Risikobewertung" des Vollertrags antizipierten Datenschutzprobleme) schränkten den Umfang der

Datenerhebung weiter ein. Aufbauend auf der ersten Veröffentlichung der Umfrage im Jahr 2021 [1] wurde dieser Meilenstein mit einer Veröffentlichung erreicht, in der die verfügbaren Datensätze gesammelt und für eine Reihe von Multi-Task-Lernversuchen verwendet wurden, die auch als Analyseinstrument zur Untersuchung der Eigenschaften der entsprechenden Annotationen dienen [9]. Die Einreichung des Artikels [9] im Oktober 2022 (veröffentlicht im Mai 2023) markiert das Erreichen des Meilensteins.

Meilenstein 2: Erste Veröffentlichung des Moderationsprototyps (ursprünglich für Dezember 2022 geplant) Dieser Meilenstein war vergleichsweise stark von der verzögerten Datenbereitstellung betroffen, wodurch dem Projektteam nur sehr geringe Datenmengen zur Verfügung standen als ursprünglich geplant. Wie oben erläutert, konzentrierte sich das Team jedoch auf die grundlegenden Forschungsaspekte der Modellierung der Moderation und auf die öffentlich zugänglichen Datensätze. Ein erster Artikel, der als erste Pilotstudie zu dieser Aufgabe diente, wurde im Juli 2021 eingereicht und im Dezember 2021 veröffentlicht [2]. Ein umfassendes Computermodell für die Moderationsaufgabe, das auf Multi-Task-Lernen basiert und alle im E-Deliberation-Repository (Meilenstein 1) gesammelten Datensätze integriert, wurde mit dem Artikel [9] (eingereicht im Oktober 2022, veröffentlicht im Mai 2023) veröffentlicht und freigegeben. Parallel dazu konzentrierte sich das Team auf die Grundlagenforschung zur Moderationsaufgabe in einer alternativen Plattform (Argument Maps, kialo.com). Der entsprechende Artikel [11] wurde im Oktober 2022 eingereicht, im Mai 2023 erneut eingereicht und im Juli 2023 veröffentlicht. Die Artikel [9] und [11] markieren das Erreichen von Meilenstein 2 im Mai 2023.

Meilenstein 3: E-Deliberation-Experiment (ursprünglich für Ende 2023 geplant) Aufgrund der Verzögerungen bei den Meilensteinen 1 und 2 sowie der Probleme mit dem Kooperationspartner war es nicht möglich, das E-Deliberation-Experiment, mit dem der Transfergedanke unseres Projekts effektiv umgesetzt werden sollte, innerhalb der Projektlaufzeit durchzuführen. Nach Rücksprache mit Kollegen des Düsseldorfer Instituts für Internetdemokratie an der Universität Düsseldorf, die über umfangreiche Erfahrungen mit kollaborativen E-Demokratie-Projekten wie E-DELIB verfügen, kam ich zu dem Schluss, dass ein vollwertiges Deliberationsexperiment in der Praxis, wie ich es ursprünglich geplant hatte, innerhalb einer Projektlaufzeit von vier Jahren einfach zu ambitioniert war (selbst unter perfekten Bedingungen ohne Verzögerungen durch COVID und all die oben genannten Probleme). Wie bereits im Abschnitt "Projektübersicht" erläutert, habe ich mich daher entschlossen, das E-Deliberation-Experiment in zwei Nutzerstudien umzuwandeln, die mit a) fachkundigen Moderatoren und b) regulären Nutzern durchgeführt werden. Was die Fachmoderatoren betrifft, so erwies sich die Suche nach ihnen für die Annotation trotz umfangreicher Bemühungen als äußerst schwierig, was den Fortschritt des entsprechenden Forschungs-Teilprojekts verlangsamte, das erstmals im Juni 2024 eingereicht und im Mai 2025 in einer vollständig überarbeiteten Fassung erneut eingereicht wurde (in Begutachtung, Publikationsliste, Abschnitt 2, a). Was die regelmäßigen Nutzer betrifft, so wurde die entsprechende Studie zur Unter-

suchung der Wahrnehmung von Moderationskommentaren, die von LLMs generiert wurden, zur Veröffentlichung im Dezember 2025 angenommen [21]. Meilenstein 3 wurde somit im Dezember 2024 abgeschlossen.

Meilenstein 4a: Endgültige Veröffentlichung des E-Deliberation-Repositorys (geplant für das Ende des Projekts, September 2024) Auf der Grundlage der mit dem Abschluss von Meilenstein 1 gesammelten Datenübersicht haben wir weitere Annotationsarbeiten durchgeführt, um Daten zur Unterstützung unserer Analysen zu sammeln und zu annotieren, und wir haben deren Ergebnisse wie oben beschrieben veröffentlicht. Darüber hinaus hat der PI zusammen mit seinen Mitarbeitern Anfang 2025 (Einreichung Oktober 2024) einen Artikel veröffentlicht, in dem die größte und umfassendste Sammlung von Datensätzen zur Argumentations- und Deliberationsqualität beschrieben wird, was den Abschluss von Meilenstein 4a markiert.

Meilenstein 4b: Endgültige Veröffentlichung des Moderator-Prototyps (geplant für das Ende des Projekts, September 2024) Die beiden für Meilenstein 3 durchgeführten Nutzerstudien enthalten eine Modellierungskomponente, die ebenfalls in den entsprechenden Artikeln beschrieben und veröffentlicht wird. Die erste Veröffentlichung, die derzeit geprüft wird [Veröffentlichungsliste, Abschnitt 2, a], befasst sich mit der Vorhersage des Moderationsbedarfs (muss dieser Kommentar moderiert werden?) und ihre Implementierungskomponente ist eine Erweiterung des bereits in [9] veröffentlichten Modells. Der zweite Artikel [11] befasst sich mit der Generierung von Moderationskommentaren und basiert auf großen Sprachmodellen, die auf Moderationsrichtlinien abgestimmt sind. Die beiden Veröffentlichungen markieren das Erreichen von Meilenstein 4b im Dezember 2024.

4 Weitere erforderliche Abschnitte

Wichtigste Punkte der numerischen Belege Neben den umfangreichen Veröffentlichungen, die in den vorangegangenen Abschnitten sowie in der diesem Bericht beigefügten Publikationsliste aufgeführt sind, ist ein sehr wichtiger Indikator für den Erfolg dieses Projekts die Sichtbarkeit, die die Gruppe durch die Etablierung eines "Markenzeichens" für das Phänomen der deliberativen Moderation erlangt hat, das im Bereich der NLP bislang weitgehend unerforscht war. So wurde der PI beispielsweise 2021, zu Beginn des Projekts, als Diskussionssteilnehmer zu einem Panel-Gespräch im Rahmen des Argument Mining-Workshops eingeladen (und wird seitdem regelmäßig eingeladen); darüber hinaus haben der PI und der Postdoktorand mehrere Kurse an internationalen Veranstaltungsorten unterrichtet (siehe Publikationsliste). Nicht zuletzt hat eine der Doktorandinnen im September 2024 ihre Dissertation erfolgreich verteidigt, während die andere Doktorandin gerade dabei ist, ihre Dissertation abzuschließen.

Notwendigkeit und Angemessenheit der durchgeführten Projektarbeit Ich habe bereits im vorigen Absatz die Aktualität und Neuartigkeit von E-DELIB erörtert.

Was die Angemessenheit betrifft, so kommt zu der begutachteten Veröffentlichung noch die Anerkennung durch die Fachwelt hinzu, mit 58 Zitaten für die Übersichtsarbeit, mit der das Projekt begann [1], sowie Zitaten unserer gesammelten Datensätze in speziellen Übersichtsarbeiten.²

Erwarteter Nutzen, insbesondere die Verwertbarkeit der Ergebnisse – einschließlich konkreter Pläne für die nahe Zukunft – im Hinblick auf den aktualisierten Verwendungsplan Damit die Ergebnisse für die Moderation unterstützt werden können, müssen weitere Softwareentwicklungsarbeiten zur Integration der Modelle für die beiden Aufgaben (Vorhersage des Moderationsbedarfs und Generierung des Moderationskommentars) durchgeführt werden.

Fortschritte auf dem Gebiet des Projekts, die dem Förderungsempfänger während der Durchführung des Projekts von anderen Organisationen mitgeteilt wurden Die Laufzeit des E-DELIB-Projekts fiel mit der explosionsartigen Verbreitung von LLMs sowie mit einem wachsenden Interesse an deren Einsatz für "Social Good"-Anwendungen zusammen. Zwar gab es mehrere Projekte und Veröffentlichungen, die sich mit der Unterstützung deliberativer Diskussionen befassten (allen voran die "Habermas Machine"³), doch ist mir kein Projekt bekannt, das eine echte Interdisziplinarität von E-DELIB mit einer Bandbreite von Methoden kombiniert, die von der Vorhersage über die Generierung bis hin zur Datenerfassung und Annotation reichen.

Veröffentlichungen oder geplante Veröffentlichungen der Ergebnisse Siehe das diesem Bericht beigefügte Dokument "Veröffentlichungen".

²Rositsa V Ivanova, Thomas Huber und Christina Niklaus. 2024. Let's discuss! Quality Dimensions and Annotated Datasets for Computational Argument Quality Assessment. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 20749–20779, Miami, Florida, USA. Association for Computational Linguistics.

³<https://www.science.org/doi/10.1126/science.adq2852>