

Schlussbericht

Zuwendungsempfänger:

Förderkennzeichen

Bauhaus-Universität Weimar

01 IS 19059A

Vorhabenbezeichnung:

Agile-AI: Agile Entwicklung von Systemen der Künstlichen Intelligenz

Teilvorhaben C: Experiment Retrieval Engine (ERE)

Laufzeit des Vorhabens:

01.11.2019 bis 30.11.2022

Inhaltsverzeichnis

1	Kurzbericht	2
1.1	Aufgabenstellung	2
1.2	Wissenschaftlicher/technischer Stand an den angeknüpft wurde	2
1.3	Ablauf des Vorhabens	3
1.4	Wesentliche Ergebnisse der Teilvorhaben	4
1.5	Zusammenarbeit mit anderen Forschungseinrichtungen	4
2	Schlussbericht	6
2.1	Verwendung der Zuwendung und erzielte Ergebnisse im Einzelnen	6
2.2	Wichtigste Positionen des zahlenmäßigen Nachweises	9
2.3	Notwendigkeit & Angemessenheit der geleisteten Arbeit	9
2.4	Nutzen & Verwertbarkeit der Ergebnisse	10
2.5	Fortschritt auf dem Gebiet des Vorhabens an anderen Stellen	11
2.6	Veröffentlichungen der Ergebnisse	12
A	Erfolgskontrollbericht	13

1 Kurzbericht

1.1 Aufgabenstellung

Die Entwicklung von Softwaresystemen, die Methoden der künstlichen Intelligenz operationalisieren (KI-Systeme), wird hauptsächlich durch empirische Analysen auf großen Datenmengen (KI-Experimente) getrieben. Die Erfahrung zeigt hierbei, dass – abgesehen von wenigen Ausnahmen – Methoden und Experimente nicht anhand gegebener Anforderungen vollständig spezifiziert, implementiert und konfiguriert werden können, sondern empirisch, evolutionär und flexibel den Anforderungen mit geeigneten Methoden und Konzepten angenähert werden müssen. Dementsprechend war das Ziel von Agile-AI die Adaptionen von Softwareentwicklungsmethoden aus der „agilen Softwareentwicklung“ für die Entwicklung von KI-Systemen mit besonderem Fokus auf KI-Experimente. Der grundsätzliche Ansatz ist, dass agile Softwareentwicklung in kleineren Iterationen abläuft und dabei ständig Softwaresysteme verbessert und erweitert. Dieser Grundgedanke der Agilität soll dabei auf die Entwicklung von KI-Systemen übertragen werden. Schwerpunkt hierbei sind Experimente, die genau diese iterative Entwicklung von KI-Modellen in einer nachvollziehbaren, reproduzierbaren und optimierbaren Weise ermöglichen. Diesen Ansatz adressieren wir in drei Teilvorhaben (TV):

(TV:A) Experiment Specification Language (ESL). Spezifikation von KI-Experimenten mittels einer domänenspezifischen Sprache (DSL).

(TV:B) Experiment Execution Platform (EEP). Technologien Verteilte parallele Ausführung, Reproduzierbarkeit und Skalierbarkeit von Experimenten auf Cluster-Computern.

(TV:C) Experiment Retrieval Engine (ERE). Retrieval Modelle für Experimente, die anhand von Suchkriterien sowohl historische also auch laufende Experimentserien analysiert und filtert.

1.2 Wissenschaftlicher/technischer Stand an den angeknüpft wurde

Der Erfolg von KI und des maschinellen Lernens sowie der empirischen Forschung erhöht den Bedarf an reproduzierbaren Experimenten und deren Spezifikation [1]. Es gibt bereits eine Reihe von Experimentierumgebungen für die Entwicklung von KI-Systemen, die sich auf die Ausführung von Code konzentrieren wie bspw. Google Colab [2], oder umfassendere Experimentierplattformen wie Comet, Polyaxon und Cloudera.¹ Diese Experimentierplattformen bieten Funktionen zur Unterstützung der Versionierung, der Reproduzierbarkeit und der kollaborativen Arbeit an. Allerdings fehlt diesen Umgebungen eine einfache Sprache für die Experimentdefinition. Zwar gibt es Frameworks wie PyTorch Ax, DeepDIVA, BenchExec und Weevil [3],² die Experimente verwalten, deployen, reproduzieren und automatisieren können – diese bieten jedoch keine Möglichkeit, Domänenkonzepte zu abstrahieren.

Suchmaschinen, die spezifisch auf das Retrieval von Experimenten unter Einbeziehung von Eingabedaten, Metadaten und Ergebnisqualität ausgelegt sind, gibt es nach unserem Kenntnisstand bislang nicht. Die aktuelle Forschung im Bereich des Information Retrieval konzentriert sich vor allem auf Suchanwendungen für Endbenutzer, insbesondere die ad-hoc-Web Suche. Daneben gibt es noch Spezialsuchmaschinen bspw. für Patente und eDiscovery, Gesundheit, Enterprise Search, Digital Libraries sowie für Querschnittsthemen wie Multimedia-Retrieval und sprachübergreifendes Retrieval.

¹<https://www.comet.com>, <https://cloudera.com>, <https://polyaxon.com>

²<https://ax.dev/>, <https://diva-dia.github.io/DeepDIVAweb>, <https://github.com/sosy-lab/benchexec>

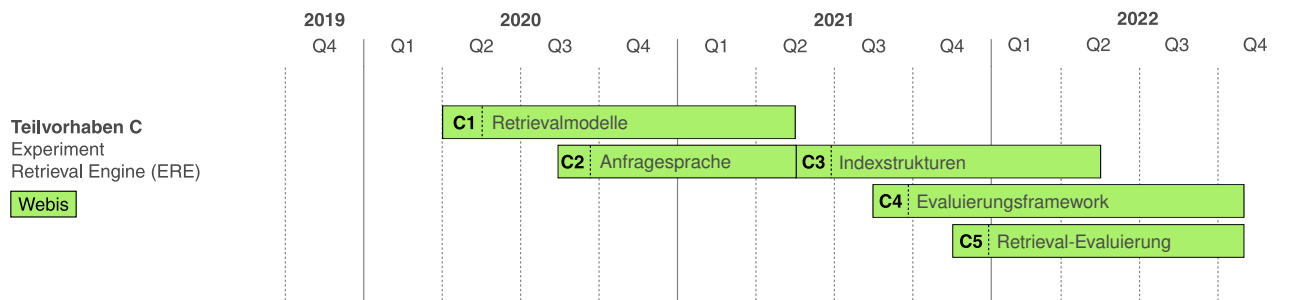


Abbildung 1: Zeitliche Einordnung der Arbeitspakete von Teilprojekt C: Experiment Retrieval Engine (ERE)

Verwandt zu unserem Vorhaben ist einzig der Bereich des Datensatz-Retrieval, bei dem es darum geht, wissenschaftliche Datensätze auffindbar zu machen. Im Zuge der Big-Data-Bewegung wurden und werden im wissenschaftlichen und industriellen Umfeld immer größere und komplexere Datensätze erstellt, teilweise veröffentlicht bzw. für eine zukünftige Nutzung archiviert. Die mittlerweile große Zahl von existierenden Datensätzen macht es einzelnen Forschern aber nahezu unmöglich, auf dem aktuellen Stand zu bleiben [4]. Allein aus ökonomischer Sicht (die Erstellung von Datensätzen ist üblicherweise mit hohem Aufwand verbunden) ist eine effektive Datensatzsuche eine wichtige Forschungsaufgabe, und vor allem Forscher in den Naturwissenschaften stehen vor diesem Akquiseproblem [5, 6, 7]. In den letzten Jahren hat man im Information Retrieval dieses Problem aufgegriffen [8], einschlägige Workshops veranstaltet [9] und das Problem auch aus informationswissenschaftlicher Sicht beleuchtet [10, 11]. Die Tatsache, dass die Suche nach Datensätzen auch für die Industrie von großer Bedeutung ist, hat Google kürzlich mit der Veröffentlichung einer dedizierten Datensatzsuchmaschine unterstrichen.³ Insgesamt befindet sich das Datensatz-Retrieval jedoch noch in den Kinderschuhen. Für die Retrieval Engine des Agile-AI-Labors sind weitere Analysen insbesondere dafür erforderlich, um herauszufinden, wie Teilmengen von Datensätzen gemäß Suchbedingungen, die in der Experimentsprache formuliert wurden, ad-hoc bzw. durch geschickte Vorindizierung vorgehalten werden können. Auch die Indizierung und Suche in Kombinationen aus textuellen und nicht-textuellen Daten (insbesondere numerischen Daten) ist ein offenes Forschungsproblem.

1.3 Ablauf des Vorhabens

Der Beginn der Arbeiten am Projekt musste auf 2020 verschoben werden, da die Ausschreibung und Suche nach geeigneten Mitarbeitern bzw. Mitarbeiterinnen erst nach Eingang des Zuwendungsbescheids im November 2019 gestartet werden konnte. Aufgrund des Fachkräftemangels in der IT-Branche in Deutschland wurde auch auf internationaler Ebene nach geeigneten Personen gesucht. Auch der Ausbruch der Corona-Pandemie erschwerte die Einstellung von Mitarbeitern, sodass einige Arbeiten verschoben werden mussten.

Das (virtuelle) Kick-Off-Meeting mit allen Projektbeteiligten fand im April 2020 statt. Aufgrund der universitären Vorgaben (Beschränkungen des Zugangs zu Universitätsbüros aufgrund der Corona-Pandemie) wurden die Meetings virtuell durchgeführt, was jedoch keine signifikante Einschränkung darstellte. Mit dem Start der Teilvorhaben wurde zudem ein zweiwöchentliches Meeting aller Projektbeteiligten inklusive aller PIs durchgeführt. Flankiert wurde dieser regelmäßige Austausch von

³<https://toolbox.google.com/datasetsearch>

bilateralen Treffen der Mitarbeiter, welche wöchentlich stattfanden. Diese Vorgehensweise und der Rhythmus der Treffen aller Projektbeteiligten und PIs wurde über den gesamten Berichtszeitraum fortgesetzt.

1.4 Wesentliche Ergebnisse der Teilvorhaben

Wesentliche wissenschaftliche und technische Ergebnisse des Projektes aus dem gesamten Berichtszeitraum aufgeschlüsselt bezüglich der Mitarbeit aller drei Teilvorhaben:

- Werkzeug zur statischen Quellcodeanalyse für KI-Experiment-Code (TV A)
- DSL als Konfigurationsnetzwerk realisiert, zur Erfassung und Überprüfung von Parameterwerten und Konfigurationen von Softwaresystemen und KI-Experiment-Settings (TV A)
- Sozio-technische Anti-Patterns bei der agilen Entwicklung von Softwaresystemen mit KI-Komponenten; extrahiert von internationalen technischen Leadern in der Industrie; (TV A).
- Einführung einer neuen Lehrveranstaltung („Software Engineering for AI“) basierend auf den Erkenntnissen des Projekts (TV A)
- Skalierbare Experiment-Plattform (Akronym: TIRA), die reproduzierbare Experimente mit automatisch generierter Dokumentation in Cloud-Umgebungen unterstützt (TV B)
- Organisation von internationalen wissenschaftlichen Veranstaltungen und Workshops, in denen KI-Experimente für spezifischen Aufgaben mit Hilfe der entwickelten Werkzeuge durchgeführt wurden (TV B).
- Entwicklung von TIREx, einer Experiment-Plattform für Information Retrieval, basierend auf der TIRA-Experiment-Plattform (TV B/C)
- Integration von Evaluationsmaßen für das Experiment- und Datensatz-Retrieval in die TIRA-Plattform (TV C)
- Entwicklung des Metadaten-Such- und Browsing-Werkzeugs „Information Retrieval Anthology“ sowie eines Teils der Spezifikation des Datensatz- und Experiment-Retrieval-Modells und der Anfragesprache (TV C)
- Mining von wissenschaftlichen Veröffentlichungen und Quellcode-Repositoryn für die Suche nach Datensätzen und Experimenten (TV C)
- Entwicklung eines Datensatz-Benchmarks zur Retrieval-Evaluation (TV C)
- Forschungsbeiträge, die auf hochrangigen Konferenzen und in vielzitierten Journalen veröffentlicht wurden und einen wichtigen Beitrag zur Grundlagenforschung und Weiterentwicklung in spezifischen Bereichen des Experimentierens mit KI-Verfahren leisten (TV A/B/C)

1.5 Zusammenarbeit mit anderen Forschungseinrichtungen

Im Rahmen des Projekts wurde mit mehreren Partnern kollaboriert, um Synergien zu nutzen und Projektergebnisse zu teilen:

- Universität Leipzig, Fakultät für Mathematik und Informatik, Text Mining and Retrieval Group (Verbundpartner)
- Universität Leipzig, Fakultät für Mathematik und Informatik, Chair of Software Systems (Verbundpartner)
- Universität Leipzig, ScaDS.AI Center for Scalable Data Science and Artificial Intelligence (Forschungskooperation)
- Martin-Luther-Universität Halle-Wittenberg, Big Data Analytics (Forschungskooperation)

2 Schlussbericht

2.1 Verwendung der Zuwendung und erzielte Ergebnisse im Einzelnen

Das Ziel von Agile-AI ist die Adaptionen von Softwareentwicklungsmethoden aus der „agilen Softwareentwicklung“ für die Entwicklung von KI-Systemen mit Fokus auf KI-Experimente. Der grundsätzliche Ansatz ist, dass agile Softwareentwicklung in kleineren Iterationen abläuft und dass dabei kontinuierlich das Softwaresystem verbessert und erweitert wird. Dieser Grundgedanke der Agilität soll auf die Entwicklung von KI-Systemen übertragen werden. Schwerpunkt hierbei sind Experimente, die genau diese iterative Entwicklung von KI-Modellen in einer nachvollziehbaren, reproduzierbaren und optimierbaren Weise ermöglichen. Um dieses Ziel zu erreichen, haben wir unser Vorhaben in drei Teilvorhaben gegliedert: (1) die Spezifikation von KI-Experimenten mittels einer domänenspezifischen Sprache (DSL), (2) die verteilte parallele Ausführung, Reproduzierbarkeit und Skalierbarkeit von Experimenten auf Cluster-Computern sowie (3) eine Retrieval Engine, die anhand von Suchkriterien sowohl historische also auch laufende Experimentserien analysiert und filtert.

Um eine erfolgreiche Bearbeitung der einzelnen Teilvorhaben zu gewährleisten, erfolgte eine vorhabenbezogene Ressourcenplanung in Form von Arbeitspaketen. Die Zielsetzung für den gesamten Berichtszeitraum besteht in der Bearbeitung der geplanten Arbeitspakete für die Teilvorhaben A, B und C. Dabei haben wir die aktuelle Entwicklung in der Praxis und Forschung im KI-Bereich so berücksichtigt, dass Anpassungen an den Herangehensweisen für die in den Arbeitspaketen beschriebenen Ziele direkt umsetzbar waren.

C1 Retrieval-Modelle für Experiment- und Datensatz-Retrieval

Mit dem Teilprojekt C1 wurde Forschung insbesondere auf dem Gebiet des Datensatz-Retrieval betrieben. Wesentlicher Bestandteil eines Experiments sind Daten, die als Input für das Experiment dienen und entweder für Experimente abgerufen oder simuliert werden können. Die für die auf der TIRA-Plattform laufenden KI-Wettbewerbe benötigten Daten wurden in dieser Phase ebenso gesammelt wie die Daten von Experimenten aus einer Vielzahl von Forschungsbereichen, in denen moderne Techniken des maschinellen Lernens eingesetzt werden. Für die Suche nach Experimenten und Datensätzen wurden mehrere populäre Datenquellen genutzt, wie PapersWithCode⁴ von Facebook MetaAI⁵, sowie offene wissenschaftliche Publikations-Repositories wie arXiv⁶ und die Sammelbände der wichtigsten Konferenzen des Fachbereichs.

Wir untersuchten, wie die Datensätze in der Forschung verwendet werden, wie Autoren sie nutzen, welche Erfahrungen sie damit gemacht und wie sie in ihren Publikationen darüber geschrieben haben. Wir führten verschiedene Experimente durch, um zu erfahren, wie das Datensatz-Retrieval verbessert werden kann – ähnlich wie bei der Online-Suche nach kommerziellen Produkten.

Zunächst grenzten wir das Forschungsproblem ein und konzentrierten uns auf den Bereich der natürlichen Sprachverarbeitung (NLP). Als vorbereitende Schritte wurde das Korpus der NLP-bezogenen Datensätze gesammelt. Zu diesem Zweck haben wir die wichtigsten Ressourcen in diesem Bereich analysiert, darunter European Language Association (ELRA),⁷ Linguistic Data Consortium (LDC),⁸

⁴www.paperswithcode.com

⁵<https://ai.facebook.com/>

⁶<https://arxiv.org/>

⁷<http://www.elra.info/en/>

⁸<https://www.ldc.upenn.edu/>

Language Resources monitoring (LRE Map)⁹ und andere. Schließlich wurde ein Korpus von 13.372 Einträgen von Metadaten über in der NLP-Forschung verwendete Datensätze erstellt. Gleichzeitig wurde das Korpus der ACL Anthologie-Publikationen extrahiert und aufbereitet. Auf diese Weise konnte eine Liste mit NLP-bezogenen Datensätzen und eine repräsentative Teilmenge von über 57.000 Publikationen erstellt werden. Anschließend wurde das gesamte Publikationskorpus analysiert und relevante Datensätze extrahiert, die später als Testimonials und Referenzen klassifiziert wurden. U.a. untersuchten wir, wie Datensätze in der Forschung verwendet wurden bzw. werden, wie die Autoren sie nutzen, welche Erfahrungen sie damit machen und was sie in ihren Veröffentlichungen darüber geschrieben haben. Wir führten eine Reihe von Experimenten durch, um herauszufinden, wie die Abfrage von Datensätzen verbessert werden kann – vergleichbar mit einer Suche nach kommerziellen Produkten im Internet. Das Papier, in dem die Ergebnisse beschrieben werden, wurde zur Veröffentlichung auf der 59. Jahrestagung der Association for Computational Linguistics und der 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021) [12] angenommen.

Unsere Forschungsanstrengungen gingen so weit, dass die Formulierung der grundlegenden Richtlinien für Datensatz-Retrieval-Modelle erfolgreich abgeschlossen war. Für die Implementierung eines Experiment-Retrieval-Modells einschließlich dessen Übertragung auf eine funktionierende Retrieval Engine mit API-Endpunkten zeigte sich jedoch, dass die Forschungslandschaft, auf der man hätte aufbauen können, bis dato zu unvollständig war und wir zunächst die entsprechende Grundlagenarbeit leisten mussten. Aus diesen Grund wurde beschlossen, die Arbeit an der API für die Retrieval Engine in den Hintergrund zu stellen, was Auswirkungen auf TV A, AP4 hatte.

C2 und C3: Anfragesprache für die ERE und Indexstrukturen

Mit den Zielen der Arbeitspakete C2 „Anfragesprache“ und C3 „Indexstrukturen“ wurde die Forschung im Bereich Datensatz- und Experiment-Retrieval weitergeführt. Dabei stand die Forschung im Bereich des Experiment-Retrievals im Vordergrund.

Um die spezifischen Merkmale eines Experiments zu analysieren, die für Experimentbeschreibungen für das Retrieval-Modell erforderlich sind, haben wir einen Korpus wissenschaftlicher Publikationen und zugehöriger Quellcode-GitHub-Repositories aus verschiedenen KI- und ML-Domänen erstellt, indem wir die Liste wissenschaftlicher Publikationen verglichen haben, die von dem bekannten “Papers with Code”-Service indexiert wurden[13]. Insgesamt wurden von uns 86.053 Repositories gesammelt, die über 700 Gigabyte an Quellcode-Dateien enthalten. Auf der Grundlage dieses umfangreichen Datensatzes wurde eine Analyse der Merkmale von Experimenten zum maschinellen Lernen durchgeführt, einschließlich (1) der Konfiguration der Umgebung, (2) der Programmiersprache und der Bibliotheken, (3) der Frameworks für maschinelles Lernen, (4) der Ansätze zur Verwaltung der Experimente und (5) der Repository-Metadaten wie der Anzahl der Mitwirkenden und der Zugehörigkeit der Autoren der Veröffentlichung (akademische Einrichtung, Industrie oder gemischt).

Die Veröffentlichung der oben genannten Ergebnisse wird derzeit vorbereitet.

Darüber hinaus verdient die Entwicklung des Metadaten-Such- und Browsing-Werkzeugs “Information Retrieval Anthology” (IR Anthology) [14, 15, 16] sowie eines Teils der Spezifikation des Datensatz- und Experiment-Retrieval-Modells und der Abfragesprache besondere Erwähnung: Das Hauptanliegen des vorgestellten Systems, eine umfassende Sammlung von Metadaten und Volltexten IR-

⁹<http://lremap.elra.info/>

bezogener Publikationen zu erstellen, unterstützt nicht nur die IR-Forschungsgemeinschaft, sondern stellt auch eine wertvolle Datenquelle für weitere Fortschritte in der Forschung des Projekts im Bereich Datensatz- und Experiment-Retrieval dar. Mit Hilfe von Informationsextraktion und Konzeptmodellierung auf der Grundlage der Veröffentlichungen in Sammlungen wie der ACL Anthology¹⁰ und der IR Anthology¹¹ wurden weitere Analysen von Aspekten der in der Literatur beschriebenen Experimente und Datensätze durchgeführt, die gleichzeitig die Möglichkeit boten, sich auf bestimmte Forschungsbereiche zu konzentrieren, in denen KI- und ML-Methoden sehr häufig eingesetzt werden. Diese Analysen bildeten die gemeinsame Basis für weitere Projektaufgaben in den Arbeitspaketen C2 und C3.

Die Entwicklung zusätzlicher Funktionen der TIRA-Plattform wurde durchgeführt, um die Integration von zuvor erstellten IR-Datensätzen zu ermöglichen, die effizient für die auf der Plattform gehosteten Experimente verwendet werden können. Diese Funktion ermöglicht es Forschern, ihre Lösungen einfach zu skalieren und mit hunderten von verfügbaren Datensätzen zu arbeiten, was die Reproduzierbarkeit und Robustheit der Experimente “out-of-the-box” verbessert.

Darüber hinaus wurde ein weiterer Ansatz für das Mining von Experiment- und Datensatz-Metadaten implementiert, bei dem die Techniken der Inhaltsextraktion eingesetzt wurden, um Open-Linked-Daten aus umfangreichen Web-Korpora wie CommonCrawl¹² und WebArchive¹³ zu sammeln. Diese strukturierten Daten, die genau definierte Vokabularien wie schema.org¹⁴ verwenden, beschreiben die auf einer Webseite veröffentlichten Ressourcen und machen sie maschinenlesbar. Sie ermöglichen den Aufbau von domänenspezifischen Wissensgraphen und, was für die Ziele des Projekts besonders wichtig war, die Anreicherung von Informationen für Indexstrukturen zum Abrufen von Datensätzen und Erfahrungen. Entlang der “Linked Data”-Hauptinhaltsinformationen, die in diesem Fall wertvoll sein könnten, wurde im Rahmen dieser Aufgabe ein umfangreicher Vergleich von Hauptinhaltsextraktionsalgorithmen durchgeführt [17].

C4 und C5: Evaluierungsframework und Retrieval-Evaluierung

Um die Effektivität der zu entwickelnden Retrieval- und Indizierungsmethoden zu bewerten und den Entwicklungsprozess der Experiment Retrieval Engine als Ganzes steuern zu können, musste ein Evaluierungs-Framework für Experiment- und Dataset-Retrieval geschaffen werden. Unser Ansatz zur Lösung dieser Aufgaben war wie folgt: (1) Wir integrierten die wohldefinierten Information Retrieval Evaluationsmetriken in die TIRA Plattform, um eine automatische Evaluation derjenigen Experimente zu ermöglichen, die auf TIRA laufen und (2) entwickelten – soweit uns bekannt ist – den ersten Retrieval-Benchmark für Datensätze überhaupt.

Um ein Evaluierungs-Framework für die Abfrage von Datensätzen und Experimenten zu schaffen, wurde die API der TIRA-Plattform um die Unterstützung für `ir_datasets` und `ir_measures`¹⁵ erweitert. Diese Funktionalität erlaubt es, alle wichtigen Evaluierungsmetriken mit wenigen Zeilen Quellcode in allen Information Retrieval Experimenten, die auf TIRA laufen, zur Verfügung zu haben[18].

Die Bewertung der Abfrageeffektivität anhand von Testsammlungen ist eine im Bereich des Infor-

¹⁰<https://aclanthology.org/>

¹¹<https://ir.webis.de/anthology/>

¹²<https://commoncrawl.org/>

¹³<https://web.archive.org/>

¹⁴<https://schema.org/>

¹⁵<https://ir-datasets.com/ir-measures.html>

mation Retrieval weit verbreitete Methodik. Da die Bewertung der Relevanz durch den Menschen einer der Hauptkosten für den Aufbau einer Testsammlung ist, sowohl in Form von Zeit als auch in Form von Geld, haben wir einen neuartigen Ansatz für die Erstellung einer Benchmark für das Retrieval von Datensätzen vorgeschlagen, bei dem eine sorgfältig kuratierte Sammlung von Diskussionssträngen aus spezialisierten Foren von Open-Data-Forschungsgemeinschaft wie opendata.stackexchange.com¹⁶ und der “datasets”-Forschungsgemeinschaft auf Reddit¹⁷ verwendet wird. Der entwickelte Benchmark besteht aus einer standardisierten “Dokumentensammlung” (Datensatz), einem standardisierten Satz von Testanfragen (Abfragen oder Beschreibungen von Datensatzanforderungen) und Goldstandard-Relevanzbewertungen einer signifikanten Teilmenge der Sammlung für jede Testanfrage. Ein wissenschaftlicher Artikel zum vorgeschlagenen Datenabruf-Benchmark ist derzeit in Vorbereitung.

2.2 Wichtigste Positionen des zahlenmäßigen Nachweises

Tabelle 1 zeigt die wichtigsten Positionen des zahlenmäßigen Nachweises des Teilvorhabens. Die Angaben zur Zuwendung beziehen sich auf die Konkretisierung des Projektplans vom November 2019.

Position	Posten	Ausgaben (bis 2022)	Zuwendung
0812	Personalmittel	139.098,16	133.724,00
0822	Hilfskräfte	11.591,95	11.822,50
0846	Dienstreisen	0,00	5.800,00
0850	Gegenstände über 410 EUR	290.898,54	290.203,00
Gesamt		441.588,65	441.549,50

Tabelle 1: Wichtigste Position des zahlenmäßigen Nachweises (Bauhaus-Universität Weimar)

2.3 Notwendigkeit & Angemessenheit der geleisteten Arbeit

Die auf dem Gebiet der KI-Technologien in Deutschland erreichten Innovationen sind weiter unter starkem Druck durch internationale Konkurrenz aus Europa, Amerika und Asien. Oft werden KI-Durchbrüche durch bloßes Brute-Force-Experimentieren erreicht. Das heißt, dass automatisierte Systeme, bspw. Google’s AutoML, massiv-parallel Experimente auf der hauseigenen Cloud-Infrastruktur ausführen, bis eine wesentliche Verbesserung erreicht wurde. Hier haben deutsche Forschungseinrichtungen und Firmen im KI-Bereich Nachteile, die wir im Projekt Agile-AI abfedern wollen.

Die Nutzung einer wie in diesem Projekt beschriebenen Experimentplattform trägt dazu bei, die Entwicklung von KI-Systemen mit höchster Qualität und Zuverlässigkeit massiv zu beschleunigen. Dies wird einen klaren Standortvorteil für Deutschland darstellen. Nur führende KI-Systeme werden es zukünftig erlauben, innovative Produktideen auch weiterhin in Arbeitsplätze in Deutschland umzumünzen. Der gegenwärtige Stand der Praxis auf diesem Gebiet wird aus softwaretechnischer Sicht dieser Herausforderung nicht gerecht. Defizite liegen in erster Linie in Trial-and-Error-Ansätzen zur Verbesserung der Güte von Modellen, der fehlenden Reproduzierbarkeit und Versionierung, sowie

¹⁶<https://opendata.stackexchange.com/>

¹⁷<https://www.reddit.com/r/datasets/>

in der stark beschränkten Skalierbarkeit von Experimenten, die mit manuell nachgelagerter Analyse durchgeführt werden.

Unser interdisziplinär aufgestelltes Projekt im Bereich der Informatik war ideal geeignet, Ergebnisse aus den drei Disziplinen Softwaretechnik, Retrieval und Forschungsdatenmanagement zum Ausgleich dieser Defizite zusammenzuführen. Aus praktischer Sicht stellen sich ähnliche Herausforderungen in allen Branchen im KI-Bereich. Deshalb bot dieses Projekt die einmalige Chance, Ergebnisse zu großen Teilen in mehreren Anwendungsbereichen zum Einsatz zu bringen.

Die Zuwendung war insbesondere notwendig, um die interdisziplinäre Kooperation sowie die branchenübergreifende Einsatzfähigkeit zu unterstützen. Die Ergebnisse des Projekts werden zur weiteren Exponierung des Wissenschafts- und Wirtschaftsstandorts Deutschland in den strategisch wichtigen Gebieten der KI- und Software-Technologien beitragen.

Die im Projektplan definierten Teilvorhaben wurden unter Berücksichtigung der aktuellen Entwicklung in der Praxis und Forschung im KI-Bereich geplant und durchgeführt, sodass Anpassungen an den Herangehensweisen für die in den Arbeitspaketen beschriebenen Ziele direkt umsetzbar waren. Die geleisteten Projektarbeiten sind aufgrund der Personalmittel, Sachmittel und Beschäftigungsentgelten angemessen. Mit Blick auf den Forschungsplan änderte sich die Zeitplanung aufgrund der verzögerten Mitarbeiter Einstellungen, der Corona-Pandemie und des Universitätswechsels von Prof. Siegmund. Der erfolgreiche Abschluss des Projekts konnte dennoch gewährleistet werden.

2.4 Nutzen & Verwertbarkeit der Ergebnisse

Die Verwertungsplanung im Projekt Agile-AI hat drei wesentliche Komponenten: Unterstützung des KI-Forschungsstandorts Deutschland, Verbesserung der Ausbildung und Lehre von KI-Experten sowie die Verwertung in der Wirtschaft durch Training und Vorträgen der im Projekt erzielten wesentlichen Einsichten und Methodiken.

Die im Antrag beschriebenen Meetups und Vorträge sind ein wesentlicher Eckpunkt im Training von Softwareentwicklern. Aufgrund der Corona-Pandemie während des Berichtszeitraums konnten Meetups und Vorträge leider nur begrenzt stattfinden. Es wurden Gespräche mit Organisatoren geführt, um virtuelle Meetings abzuhalten. Ein solches virtuelles Meeting fand im Rahmen der Feature-Oriented Software Development (FOSD) Forschungsgemeinschaft im April 2021 statt. In diesem Meeting wurde die Idee des Konfigurationsnetzwerks (A2 und A3) vorgestellt und wertvolles Feedback für die weitere Entwicklung des Netzwerkes eingeholt. Des Weiteren wurden im Rahmen der Veranstaltungen „Imitatoren des Menschlichen“ an der Universität Halle und der Konferenz STUTS/TACOS an der Universität Leipzig Vorträge zur den den PAN-Workshops zugrundeliegenden Technologien gehalten. Darin wurden auch die durch TIRA ermöglichten reproduzierbaren Experimente im Bereich der PAN Workshop-Serie erläutert. Wir waren außerdem Gastgeber für den deutschen Teil eines internationalen Events, das von dem renommierten AI-Unternehmen Hugging Face – die 2022 ML Demo.cratization Tour with Hugging Face – veranstaltet wurde. Im Rahmen dieser Veranstaltung demonstrierte einer der Experten den Einsatz des unternehmenseigenen Ecosystems, mithilfe dessen KI-Demos eingerichtet und durchgeführt werden können. Das Ziel dieser KI-Demos ist es, den Output eines Machine-Learning-Experiments oder Modells an die Öffentlichkeit oder Kollegen zu vermitteln.

Hinsichtlich der Lehre an der Universität Leipzig und der Bauhaus-Universität Weimar fanden und finden bereits mehrere Vorlesungen und Seminare statt, wie beispielsweise „Information Retrieval“,

“Advanced Information Retrieval”, “Advanced Natural Language Processing”, “Foundations of Machine Learning” und “Big Data and Language Technologies”, in denen wesentliche Konzepte des Projekts an Studierende vermittelt und auch praktisch bei der Durchführung von KI-Experimenten erprobt werden. Darüber hinaus wird die Experimentierplattform TIRA auch erfolgreich an der Martin-Luther-Universität Halle-Wittenberg sowie an der Universität Padua, Italien, verwendet. In diesen Veranstaltungen wurde TIRA im Rahmen von Vorträgen den Lehrenden und Studierenden näher gebracht. Darüber hinaus konnte die neuartige Lehrveranstaltung “Software Engineering for AI-Enabled Software Systems” konzipiert und durchgeführt werden, die genau an der Projektschnittstelle zwischen agilen Entwicklungsprozessen, Experimentdurchführung von KI und Produktion von KI-Software steht. Durch den Transfer unserer Erkenntnisse in den Lehrplan schaffen wir eine hervorragende Perspektive für Studierende auf dem Arbeitsmarkt und gehen damit gleichzeitig ein wichtiges Problem im Bereich der sozio-technischen Herausforderungen an: den Mangel an qualifiziertem Personal an der Schnittstelle von Künstlicher Intelligenz und Software Engineering.

Bezüglich der Unterstützung des KI-Forschungsstandorts fand und findet weiterhin eine institutionelle Kooperation der PIs Potthast und Siegmund mit dem Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI)¹⁸ statt. Diese Kooperation ermöglicht eine direkte Verwertung der Arbeiten und Ergebnisse des Agile-AI Projektes.

2.5 Fortschritt auf dem Gebiet des Vorhabens an anderen Stellen

Im Berichtszeitraum sind mehrere Publikationen von Dritten mit inhaltlicher Überschneidung zu diesem Projekt veröffentlicht worden, die bei der Bearbeitung Berücksichtigung fanden. Eine Auswirkung auf den Arbeitsplan oder das Projektziel hatte das jedoch nicht. Die nachfolgende Aufstellung ist eine chronologisch geführte Auflistung relevanter Publikationen.

- Marius Schlegel, Kai-Uwe Sattler (2023). Management of machine learning lifecycle artifacts: A survey. *ACM SIGMOD Record*, 51(4):18–35. [19]
- Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, Emine Yilmaz (2022). Fostering coopetition while plugging leaks: The design and implementation of the MS MARCO leaderboards. *SIGIR 2022*. 2939–2948. [20]
- Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, Stefan Wagner (2022). Software engineering for ai-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(2):1–59. [21]
- Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering*, pages 413–425, 2022. [22]
- Xinyu Zhang and Nandan Thakur and Odunayo Ogundepo and Ehsan Kamaloo and David Alfonso-Hermelo and Xiaoguang Li and Qun Liu and Mehdi Rezagholizadeh and Jimmy Lin (2022). Making a MIRACL: Multilingual information retrieval across a continuum of languages. *arXiv:2210.09984*. [23]

¹⁸<https://www.scads.de/de/>

- Gökrem Giray. A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software*, 180:111031, 2021. [24]
- Tetsuya Sakai and Sijie Tao and Zhaohao Zeng and Yukun Zheng and Jiaxin Mao and Zhumin Chu and Yiqun Liu and Maria Maistro and Zhicheng Dou and Nicola Ferro, et al. (2020). Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) task. *NTCIR 2020*. [25]
- Timo Breuer and Philipp Schaer and Narges Tavakolpoursaleh and Johann Schaible and Benjamin Wolff and Bernd Müller (2019). STELLA: Towards a framework for the reproducibility of online search experiments. *OSIRRC at SIGIR 2019*. 8–11. [26]

2.6 Veröffentlichungen der Ergebnisse

Die im Projekt erarbeiteten Publikationen konnten auf international hochrangigen Konferenzen und in wichtigen Zeitschriften platziert werden.

Publikationen

- Theresa Elstner and Frank Loebe and Yamen Ajjour and **Christopher Akiki** and Alexander Bondarenko and Maik Fröbe and Lukas Gienapp and **Nikolay Kolyada** and Janis Mohr and Stephan Sandfuchs and Matti Wiegmann and Jörg Frochte and Nicola Ferro and Sven Hofmann and **Benno Stein** and Matthias Hagen and **Martin Potthast**. Shared Tasks as Tutorials: A Methodical Approach. In *Proceedings of the Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 23)*. [27]
- Maik Fröbe and Matti Wiegmann and **Nikolay Kolyada** and Bastian Grahm and Theresa Elstner and Frank Loebe and Matthias Hagen and **Benno Stein** and **Martin Potthast**. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*. [28]
- **Sebastian Simon** and **Nikolay Kolyada** and **Christopher Akiki** and **Martin Potthast** and **Benno Stein** and **Norbert Siegmund**. Exploring Hyperparameter Usage and Tuning in Machine Learning Research. In *2nd International Conference on AI Engineering – Software Engineering for AI (CAIN), May, 2023*. CAIN. [29]
- Maik Fröbe and Jan Heinrich Reimer and Sean MacAvaney and Niklas Deckers and Simon Reich and Janek Bevendorff and **Benno Stein** and Matthias Hagen and **Martin Potthast**. The Information Retrieval Experiment Platform. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. [18]
- **Nikolay Kolyada**, **Christopher Akiki**, **Sebastian Simon**, Nicole Heinemann, **Norbert Siegmund**, **Martin Potthast**, and **Benno Stein**. Experiment Tracking Frameworks for Machine Learning: A Case Study in Research. (manuscript in preparation)
- **Nikolay Kolyada** and **Martin Potthast** and **Benno Stein**. Beyond Metadata: What Paper Authors Say About Corpora They Use. *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*. [12]
- **Nikolay Kolyada** and **Martin Potthast** and **Benno Stein**. Bootstrapping a Dataset Retrieval Benchmark. (manuscript in preparation)

Kooperationstreffen & Vorträge

- Vortrag auf dem Wikipedia Text Reuse: Within and Without (December 2019, DZHW Workshop, Berlin)
- Vortrag auf dem Multi-Document Information Consolidation (April 2019, Dagstuhl Seminar)
- Vortrag auf dem Algorithms for Authorship Analytics (May 2019, Digital Plato, Herrenhausen Palace)
- The Democratization of AI (November 2019, IFK, Wien)
- Vortrag auf dem 50. Jahrestagung der Gesellschaft für Informatik, INFORMATIK 2020,
- Vortrag auf dem Perceived Limits in Information Retrieval (September 2022, CLEF'22, Bologna)
- Keynote auf Touché: Argument Retrieval (September 2022, Touché @ CLEF'22, Bologna)

A Erfolgskontrollbericht

Der Erfolgskontrollbericht liegt dem Projektträger als separate Anlage vor.

Literatur

- [1] Babatunde K Olorisade, Pearl Brereton, and Peter Andras. Reproducibility in machine learning-based studies: An example of text mining.
- [2] Tiago Carneiro, Raul Victor Medeiros Da Nóbrega, Thiago Nepomuceno, Gui-Bin Bian, Victor Hugo C De Albuquerque, and Pedro Pedrosa Reboucas Filho. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *IEEE Access*, 6:61677–61685, 2018.
- [3] Yanyan Wang, Matthew J Rutherford, Antonio Carzaniga, and Alexander L Wolf. Automating experimentation on distributed testbeds. In *Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering*, pages 164–173, 2005.
- [4] V. M. Megler and David Maier. When Big Data Leads to Lost Data. In *Proceedings of the 5th Ph.D. Workshop on Information and Knowledge*, PIKM '12, pages 1–8, New York, NY, USA, 2012. ACM.
- [5] D. Maier and V. Megler. Data Near Here: Bringing Relevant Data Closer to Scientists. *Computing in Science and Engineering*, 15(03):44–53, may 2013.
- [6] Lucila Ohno-Machado, Susanna-Assunta Sansone, George Alter, Ian Fore, Jeffrey Grethe, Hua Xu, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Anupama E Gururaj, Elizabeth Bell, Ergin Soysal, Nansu Zong, and Hyeon eui Kim. Finding Useful Data Across Multiple Biomedical Data Repositories Using Datamed. *Nature Genetics*, 49:816–819, 2017.
- [7] Young M. Park, Silvano Squizzato, Nicola Buso, Tamer Gur, and Rodrigo Lopez. The Ebi Search Engine: Ebi Search as a Service—Making Biological Data Accessible for All. *Nucleic Acids Research*, 45(W1):W545–W549, 05 2017.
- [8] Christine L. Borgman, Andrea Scharnhorst, and Milena S. Golshan. Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse. *CoRR*, abs/1802.02689, 2018.
- [9] Paul T. Groth, Laura Koesten, Philipp Mayr, Maarten de Rijke, and Elena Simperl. DATA: SEARCH'18 - Searching Data on the Web. *CoRR*, abs/1805.11883, 2018.

- [10] Kathleen Gregory, Helena Cousijn, Paul T. Groth, Andrea Scharnhorst, and Sally Wyatt. Understanding Data Retrieval Practices: A Social Informatics Perspective. *CoRR*, abs/1801.04971, 2018.
- [11] Kathleen Gregory, Paul T. Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. Searching Data: A Review of Observational Data Retrieval Practices in Selected Disciplines. *JASIST*, 70(5):419–432, 2019.
- [12] Nikolay Kolyada, Martin Potthast, and Benno Stein. Beyond Metadata: What Paper Authors Say About Corpora They Use. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pages 5085–5090. ACL-IJCNLP, August 2021.
- [13] Nikolay Kolyada, Christopher Akiki, Sebastian Simon, Nicole Heinimann, Norbert Siegmund, Martin Potthast, , and Benno Stein. Experiment Tracking Frameworks for Machine Learning: A Case Study in Research, May 2023. Manuscript in preparation.
- [14] Martin Potthast, Benno Stein, and Matthias Hagen. The Information Retrieval Anthology 2021: Inaugural Status Report and Challenges Ahead. *SIGIR Forum*, 55(1), June 2021.
- [15] Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, and Matthias Hagen. The Information Retrieval Anthology. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2550–2555. ACM, July 2021.
- [16] Martin Potthast, Sebastian Günther, Janek Bevendorff, Jan Philipp Bittner, Alexander Bondarenko, Maik Fröbe, Christian Kahmann, Andreas Niekler, Michael Völske, Benno Stein, and Matthias Hagen. The Information Retrieval Anthology. In *Lernen. Wissen. Daten. Analysen. - LWDA 2021*, September 2021.
- [17] Janek Bevendorff, Sanket Gupta, Johannes Kiesel, and Benno Stein. An Empirical Comparison of Web Content Extraction Algorithms. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. ACM, July 2023.
- [18] Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. The Information Retrieval Experiment Platform. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*. ACM, July 2023.
- [19] Marius Schlegel and Kai-Uwe Sattler. Management of machine learning lifecycle artifacts: A survey. *ACM SIGMOD Record*, 51(4):18–35, 2023.
- [20] Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Fostering coopetition while plugging leaks: The design and implementation of the MS MARCO leaderboards. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2939–2948. ACM, 2022.
- [21] Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, and Stefan Wagner. Software engineering for ai-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(2):1–59, 2022.
- [22] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering*, pages 413–425, 2022.
- [23] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamaloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Making a MIRACL: multilingual information retrieval across a continuum of languages. *CoRR*, abs/2210.09984, 2022.

- [24] Görkem Giray. A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software*, 180:111031, 2021.
- [25] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, et al. Overview of the ntcir-15 we want web with centre (www-3) task. *Proceedings of NTCIR-15. to appear*, 2020.
- [26] Timo Breuer, Philipp Schaer, Narges Tavakolpoursaleh, Johann Schaible, Benjamin Wolff, and Bernd Müller. STELLA: towards a framework for the reproducibility of online search experiments. In Ryan Clancy, Nicola Ferro, Claudia Hauff, Jimmy Lin, Tetsuya Sakai, and Ze Zhong Wu, editors, *Proceedings of the Open-Source IR Replicability Challenge co-located with 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, OSIRRC@SIGIR 2019, Paris, France, July 25, 2019*, volume 2409 of *CEUR Workshop Proceedings*, pages 8–11. CEUR-WS.org, 2019.
- [27] Theresa Elstner, Frank Loebe, Yamen Ajjour, Christopher Akiki, Alexander Bondarenko, Maik Fröbe, Lukas Gienapp, Nikolay Kolyada, Janis Mohr, Stephan Sandfuchs, Matti Wiegmann, Jörg Frochte, Nicola Ferro, Sven Hofmann, Benno Stein, Matthias Hagen, and Martin Potthast. Shared Tasks as Tutorials: A Methodical Approach. In *Thirteenth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 23)*. EAAI, 2023.
- [28] Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, pages 236–241, Berlin Heidelberg New York, April 2023. Springer.
- [29] Sebastian Simon, Nikolay Kolyada, Christopher Akiki, Martin Potthast, Benno Stein, and Norbert Siegmund. Exploring Hyperparameter Usage and Tuning in Machine Learning Research. In *2nd International Conference on AI Engineering - Software Engineering for AI (CAIN 2023)*. CAIN, May 2023.