Intelligentes Web Crawling für die industrielle Trendanalyse: Eine skalierbare KI-gestützte Architektur

Richard Zowalla[1], Jan Mackensen[1], Meng Jin[1], Kristian Schaefer[1], Safa Omri[1], Jens Neuhüttler[1]

[1] Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO, richard.zowalla@iao.fraunhofer.de, jan.mackensen@iao.fraunhofer.de, meng.jin@iao.fraunhofer.de, kristian.schaefer@iao.fraunhofer.de, safa.omri@iao.fraunhofer.de, jens.neuhuettler@iao.fraunhofer.de

Zusammenfassung

Die rechtzeitige Erkennung neuer Trends ist für strategische Vorausschau und industrielle Forschung von entscheidender Bedeutung. Herkömmliche Web-Crawler sind zwar bei der Erfassung großer Datenmengen effektiv, verfügen jedoch häufig nicht über semantische Filter-funktionen, was zur Anhäufung von irrelevanten oder wenig wertvollen Informationen führt. Um diese Limitation zu beheben, wird in diesem Artikel eine fokussierte Web-Crawling-Architektur vorgestellt, die Large Language Models (LLMs) für die dynamische Bewertung und Priorisierung von Inhalten integriert. Das System basiert auf einem verteilten Framework unter Verwendung von Apache StormCrawler, Apache Storm, Playwright und OpenSearch. LLMs sind in den Crawling-Prozess integriert, um bei Bedarf Relevanzbewertungen durchzuführen und den Crawler zu Inhalten mit hoher thematischer Relevanz zu leiten. Dieser Ansatz reduziert Datenrauschen und erhöht die Effizienz der webbasierten Informationssammlung bei gleichzeitiger Verringerung der Rechenkomplexität und der Kosten. Über die technische Infrastruktur hinaus beschreibt der Artikel die analytischen Methoden, die auf die gesammelten Inhalte angewendet werden, darunter Topic Modelling zur Trenderkennung und Clustering-Algorithmen zur thematischen Strukturierung. Das Ergebnis ist eine modulare, skalierbare Pipeline, die unstrukturierte Webdaten in strukturierte Erkenntnisse umwandeln kann und da-mit fortschrittliche Anwendungen in der strategischen Frühwarnung für aufkommende Trends und für die Technologieplanung unterstützt. Durch die Integration von LLMs in das fokussierte Web-Crawling leistet diese Arbeit einen methodischen Beitrag in den Bereichen Voraus-schau, Innovationsmanagement und datengesteuerte strategische Analyse. Alle Komponenten und Techniken werden vor dem Hintergrund des aktuellen Stands der Technik kontextualisiert, um ihren Mehrwert und ihren neuartigen Beitrag zum Fachgebiet hervorzuheben.

Schlüsselworte

Web-Crawler, Large Language Models, Topic Modelling, Foresight, Trend Analysis

Intelligent Web Crawling for Industrial Trend Analysis: A Scalable AI-Driven Architecture

Abstract

Timely identification of emerging trends is essential for strategic foresight and industrial research. Conventional web crawlers, though effective at large-scale data collection, often lack semantic filtering capabilities, leading to the accumulation of contextually irrelevant or low-value information. To address this limitation, this paper presents a focused web crawling architecture that integrates Large Language Models (LLMs) for dynamic content evaluation and prioritization. The system is built on a distributed framework using Apache StormCrawler, Apache Storm, Playwright, and OpenSearch. LLMs are integrated into the crawling process to perform on-demand relevance estimation, guiding the crawler toward content with high thematic relevance. This approach reduces data noise and increases the efficiency of web-based information gathering, while reducing computational complexity and cost. Beyond infrastructure, the paper details the analytical methods applied to the collected content, including topic modeling for trend detection and clustering algorithms for thematic structuring. The result is a modular, scalable pipeline capable of transforming unstructured web data into structured insights, supporting advanced applications in strategic early warning for emerging trends and technology planning. By integrating LLMs into focused web crawling, this work contributes methodological improvements to the fields of foresight, innovation management and data-driven strategic analysis. All components and techniques are contextualized against the current state of the art to highlight their added value and novel contribution to the field.

Keywords

Web Crawler, Large Language Models, Topic Modelling, Foresight, Trend Analysis

1 Einleitung

Die strategische Vorausschau und frühzeitige Identifikation technologischer Trends bilden heute einen entscheidenden Wettbewerbsvorteil für Unternehmen und Forschungseinrichtungen [RK 2018, HG 2015]. In einer zunehmend digitalisierten Wirtschaft mit verkürzten Technologiezyklen stehen Organisationen vor der Herausforderung, aus der exponentiell wachsenden Informationsflut des Internets relevante Erkenntnisse zu extrahieren [PFN 2023, DSL+ 2025, MSA+ 2025].

Vor diesem Hintergrund gewinnen datengetriebene Methoden zur Trendanalyse und strategischen Früherkennung zunehmend an Bedeutung [Roh14], da klassische Verfahren wie Expertenbefragungen, Patentanalysen oder Marktstudienangesichts der Datenmenge und Dynamik an ihre Grenzen stoßen. Die automatisierte Erfassung und Analyse offener Webdaten versprechen daher einen erheblichen Mehrwert – vorausgesetzt, es gelingt, die Relevanz der Inhalte effizient zu bewerten und inhaltlich präzise einzuordnen. Herkömmliche Web-Crawler sind jedoch primär auf die breite Sammlung von Informationen ausgelegt und verfügen oft nicht über hinreichende semantische Filtermechanismen. In der Folge kommt es zu einer hohen Rate irrelevanter oder nur marginal nützlicher Inhalte, was die nachgelagerte Analyse erschwert und unnötige Rechenressourcen bindet.

Zur Überwindung dieser Einschränkung wird in der vorliegenden Arbeit ein neuartiger Ansatz vorgestellt, der fokussiertes Web-Crawling mit Generativer Künstlicher Intelligenz (GenAI) – insbesondere Large Language Models (LLMs) – kombiniert. Ziel ist es, thematisch hochrelevante Inhalte automatisiert zu identifizieren, zu priorisieren und in strukturierter Form zur Verfügung zu stellen. Die entwickelte Architektur basiert auf einem verteilten Framework mit Apache StormCrawler und Playwright. Das System integriert LLMs zur dynamischen Relevanzbewertung während des Crawling-Prozesses. Damit werden Inhalte mit hoher thematischer Dichte gezielt angesteuert, Datenrauschen (irrelevante Informationen) reduziert und der Wirkungsgrad der webbasierten Informationsgewinnung erheblich gesteigert.

Durch die Kombination von fokussiertem Web-Crawling mit modernen Topic-Modelling-Techniken wie BERTopic [Gro22] sowie der LLM-gestützten semantischen Analyse entsteht eine skalierbare, modular aufgebaute Pipeline, die unstrukturierte Webdaten in verwertbare Erkenntnisse überführt. Neben der technischen Architektur werden im weiteren Verlauf dieser Arbeit auch methodische Aspekte wie Klassifikation, Themenextraktion, Evaluation sowie die Visualisierung relevanter Trends detailliert behandelt. Auf diese Weise leistet die Arbeit einen Beitrag zur Weiterentwicklung datenbasierter Strategiewerkzeuge im Bereich der industriellen Vorausschau und trägt zur methodischen Integration generativer KI in reale Informationssysteme bei.

1.1 Stand der Forschung

1.1.1 Fokussiertes Web-Crawling

Das indexierte Web, d.h. der Teil des Internets, der durch öffentliche Suchmaschinen indexiert ist, wurde 2015 von van der Bosch et al. auf circa 47 Milliarden Webseiten geschätzt [vBK16] und hat sich seitdem weiter vergrößert. Angesicht dieser Datenmenge ist es notwendig, fokussiertes Web-Crawling einzusetzen, um themenspezifische Informationen effizient erfassen zu können. Hierfür nutzt ein fokussierter Web-Crawler die Themenlokalität des Webs aus [Dav00, Men04, Men05], indem die thematische Nähe von Webseiten durch Priorisierung der zu erfassenden Webseiten maximal ausgenutzt wird [KBR17].

Fokussierte Web-Crawler werden für unterschiedlichste Zwecke eingesetzt. Rheinländer et al. setzten fokussiertes Web-Crawling unter Verwendung eines Naive Bayes Klassifikators für die gezielte Sammlung von englischsprachigen biomedizinischen Webinhalten ein und erreichten eine Harvest Rate (HR; Anteil themenrelevanter Seiten an allen gecrawlten Webeiten) von 38% [RLK+16]. Zowalla et al. hingegen verwendeten einen Support Vector Machine (SVM) Ansatz zur Sammlung von deutschsprachigen Gesundheitswebseiten und erreichten eine HR von 21,27% [ZWP20]. In 2023 untersuchten Sakai et al. ein System zur automatischen Erkennung von japanischen „Fake Shops“ unter Verwendung klassischer Verfahren aus dem Bereich des Maschinellen Lernens und erreichten eine HR von mehr als 20% [STK+23]. [RN23] verwendeten Support Vector Regression (SVR) für ihren fokussierten Web-Crawler und erreichten eine HR von 37%.

Bisher kommen dabei vornehmlich klassische Verfahren des maschinellen Lernens zum Einsatz. Der Einsatz von GenAI nimmt jedoch langsam zu – insbesondere im Bereich der Textextraktion und des Parsings werden LLMs häufig eingesetzt [BRR+24, AW24, Hua25]. Für die Relevanzschätzung von Webseiten finden LLMs hingegen bislang kaum Anwendung, vermutlich aufgrund noch unzureichender Performance bei der thematischen Präzisierung und Effizienz im Vergleich zu etablierten, leichter gewichtigen Verfahren. Herkömmliche Suchmaschinen wie Google bieten zudem keine vollständige Abdeckung themenspezifischer Inhalte und schränken den automatisierten Zugriff technisch sowie lizenzrechtlich ein. Zudem fehlt Transparenz über Indexierungs- und Rankingmechanismen, was eine gezielte Relevanzbewertung erschwert. Daher ist es notwendig, fokussiertes Web-Crawling selbst umzusetzen, um eine gezielte, steuerbare und umfassende Sammlung von relevanten Web-Inhalten zu gewährleisten. Aus diesem Grund wird in dieser Studie ein entsprechender Ansatz vorgestellt, um LLMs gezielt für die Relevanzbewertung beim fokussierten Web-Crawling einzusetzen.

Zusammenfassend ist festzuhalten, dass fokussiertes Web-Crawling zwar erfolgreich eingesetzt wird, aber bisher vor allem auf klassischen Verfahren aus dem Bereich des Maschinellen Lernens basiert. Moderne KI-Methoden wie LLMs wurden hier noch kaum integriert.

1.1.2 KI-gestützte Trendanalysen

Trendscouting umfasst alle Aktivitäten, die sich mit dem Sammeln und Auswerten von Informationen beschäftigen, um frühzeitig Veränderungen in Technologien und Themengebieten zu erkennen [Roh14]. Als Teil des Corporate Foresight werden die gewonnenen Informationen genutzt, um eine zukunftsorientierte Unternehmensstrategie zu entwickeln [Roh14]. KI spielt eine große Rolle beim Trendscouting, da durch den Einsatz von KI-Methoden Informationen effizienter dargestellt werden können sowie die Sammlung und Analyse großer Datenmengen automatisiert werden kann. Dies übertrifft die Möglichkeiten der etablierten, klassischen Trendscouting-Methoden, wie beispielsweise der Expertenbefragung [FNB24].

Eine Anwendung von KI im Trendscouting ist die Vorhersage zukünftiger Marktentwicklungen anhand von Zeitreihendaten. Lesmana et al. zeigen, dass sich Zeitreihendaten von Nachrichtenagenturen mit KI-Modellen besser vorhersagen lassen als mit traditionellen Ansätzen [LWN+24]. Im Trendscouting werden in der Regel jedoch Analysen von Textinhalten durchgeführt. Schuh et al. demonstrieren, wie KI dazu eingesetzt werden kann, technologische Begriffe in Texten zu identifizieren und einem Technologieradar zuzuordnen. [SHS+21]. Die verbreitetste KI-Methode zur Textanalyse für Trendscouting ist das Topic Modeling zur Identifikation von Themen.

Eine der meistgenutzten Topic-Modelling-Techniken ist die Latent Dirichlet Allocation (LDA) [BNJ03]. LDA wird beispielsweise zur Analyse von Trends in Twitter-Daten [RFB+21] oder zur Ermittlung von Trends in politischen Debatten und Magazinen eingesetzt [DSL+23]. Eine neuere Methode des Topic-Modellings ist BERTopic, bei der Embedding-Technologien zusammen mit Clustering-Verfahren eingesetzt werden, um Dokumente in passende Themengebiete zu clustern [Gro22]. Jin et al. nutzen BERTopic, um in F&E-Förderungsdatenbanken Trends in der Förderung emissionsfreien Bauens zu entdecken [JB23].

Auch Large Language Models (LLMs) finden im KI-gestützten Trendscouting Anwendung. Besonders relevant ist die Retrieval-Augmented Generation (RAG), bei der LLMs mit Textdatenbanken verknüpft werden, um themenspezifische Analysen und Zusammenfassungen zu erstellen [HSM24]. In Kombination mit Topic Modeling lassen sich so tiefere Einblicke in einzelne Themenbereiche gewinnen [KW:24].

Insgesamt dient KI im Trendscouting vor allem der Analyse von Textdaten. Die Datenerhebung erfolgt hingegen meist ohne KI – entweder selektiv [LWN+24] oder durch Schlüsselwortsuche bei großen Datenmengen [DSL+23, JB23, KW:24].

1.2 Ziele und Forschungsfragen

In dieser Studie liegt der Fokus auf der Entwicklung eines KI-gestützten, fokussierten Web-Crawlers zur themenspezifischen Erfassung und Analyse von Webinhalten. Ziel ist es, das Potenzial von LLMs für die Relevanzbewertung im Crawling-Prozess zu demonstrieren und so die Effizienz bei der Datenerhebung im Kontext von Trend- und Technologiescouting zu erhöhen. Konkret verfolgt die Arbeit vier Ziele:

1) Demonstration der Einsatzmöglichkeiten von LLMs zur dynamischen Relevanzbewertung und thematischen Steuerung eines fokussierten Web-Crawlers.
2) Entwicklung einer flexiblen, konfigurierbaren Crawler-Architektur für kundenspezifische Anwendungsfälle
3) Kombination von Topic-Modelling und RAG-Ansätzen zur Analyse und Strukturierung gesammelter Inhalte
4) Darstellung eines prototypischen Systems zur kontinuierlichen, themenspezifischen Beobachtung des Webs für die technologische Vorausschau

Darüber hinaus liefert die Arbeit Erkenntnisse zur Integration heterogener Datenquellen und gibt Einblicke in die technischen und konzeptionellen Herausforderungen beim Aufbau einer skalierbaren Analysepipeline. Soweit den Autoren bekannt, existieren bislang keine vergleichbaren Ansätze, die LLMs in fokussierte Web-Crawler zur Trendidentifikation integrieren.

2 Methodik

2.1 Systemüberblick

Das entwickelte Software-System folgt einer Pipeline-Architektur und besteht aus verschiedenen Komponenten, um den Trendscouting Prozess abzubilden. Der Pipeline-Prozess lässt sich in drei Phasen gliedern und besteht aus den nachfolgenden Elementen:

1) Datenerfassung: In diesem Schritt werden relevante Daten (z.B. Webseiten, Literaturdatenbanken) mittels des fokussierten Web-Crawlers SPIDERWISE erfasst.
2) Themenidentifikation: Anschließend werden die Daten mithilfe von Methoden wie Topic Modelling analysiert, um spezifische Themen und Trends zu erkennen.
3) Präsentation: Die ermittelten Themen und Trends werden dem Nutzenden über ein Dashboard dargestellt.

2.2 Fokussiertes Web-Crawling

2.2.1 Web-Crawling-Prozess

Ein Web-Crawler traversiert den gerichteten Graphen des Webs [BP12]. Ausgehend von einer Menge an Startadressen („seeds“), erfasst dieser die jeweiligen Webseiten. Wenn der Download erfolgreich ist, wird die Struktur der jeweiligen Webseite zerlegt und Referenzen auf weitere Webseiten (Hyperlinks bzw. Unified Ressource Locators (URLs)) extrahiert. Diese Hyperlinks werden dann analysiert und in eine prioritätssortierte Warteschlange, die sog. Frontier, eingefügt [BP12, DS11, CvD99]. Der Web-Graph wird dann Schritt für Schritt über die in der Frontier enthaltenen URLs besucht. Dieser Vorgang wird wiederholt, bis die Frontier leer ist oder der Web-Crawl manuell gestoppt wird.

Auf Grund der Größe des Webs [vBK16] muss man sich auf einen bestimmten Themenbereich festlegen, um die Erfassung zu beschleunigen. In diesem Zusammenhang besucht ein fokussierter Web-Crawler nur diejenigen ausgehenden Links einer Webseite, die für das gegebene Thema relevant zu sein scheinen. Um festzustellen, ob ein Link relevant ist oder nicht, wird davon ausgegangen, dass Webseiten desselben Themas höchstwahrscheinlich mit anderen Webseiten desselben Themas verlinkt sind [Dav00, Men04]. Die Relevanz einer Webseite wird häufig über Textklassifikationsverfahren bestimmt [MMB14, RN23, STK+23]. Während dabei traditionell klassische Methoden des Maschinellen Lernens wie SVMs zum Einsatz kommen, zeigen wir erstmals, dass auch LLMs hierfür effektiv genutzt werden können. Eine Menge von Klassifikatoren und Heuristiken werden dann eingesetzt, um irrelevante Inhalte während des Web-Crawls herauszufiltern und den extrahierten URLs auf der Grundlage des Klassifikationsergebnisseines eine Priorität zuzuweisen.

2.2.2 Systemarchitektur und Prozessablauf

Angesichts der Größe des Webs ist es naheliegend, dass die sequenzielle Verarbeitung einer solchen Datenmenge einen enormen Zeitaufwand und entsprechende finanzielle Ressourcen erfordern würde. Aus diesem Grund ist eine verteilte Systemarchitektur notwendig, um Ergebnisse verfügbar zu machen, bevor sich das Web merklich verändert hat. Dies ist insbesondere für die Erkennung von Trends und deren Überwachung von Bedeutung. Daher muss eine solche Architektur so konzipiert sein, dass sie Tausende von Threads zum parallelen Abrufen von Webseiten verarbeiten kann und dabei auch mit dynamischen Web-Inhalten umgehen kann. Neben der Effizienz in Bezug auf den Durchsatz sollte ein Crawler auch die Crawler-Ethik beachten [Liu11, Eic95], d.h. sich an das Robots-Exclusion-Protokoll zu halten und den Ziel-Webserver nicht mit zu vielen Anfragen in kurzer Zeit zu überlasten. Aus diesem Grund ist die Implementierung von künstlichen Verzögerungen zwischen Anfragen an denselben Webserver obligatorisch. Darüber hinaus muss ein Web-Crawler robust gegenüber Spidertraps, d.h. Webseiten mit Programmfehlern oder dynamisch generierten URLs, die zu einer Endlosschleife des Web-Crawlers führen [Liu11], sein. Zudem muss der HTML-Parser mit invalider Syntax oder ungültigem Markup sowie binären oder dynamischen Inhalten umgehen [Liu11]. Darüber hinaus müssen die Textextraktionskomponenten Boilerplate-Erkennung in geeigneter Weise handhaben.

Unser fokussierter Web-Crawler *SPIDERWISE* wurde auf Grundlage des Open-Source Frameworks Apache StormCrawler (SC) entwickelt. SC ist ein Software Development Kit für den Aufbau von skalierbaren Web-Crawlern auf Grundlage des Stream Processing Frameworks Apache Storm. SC verfügt jedoch nicht über fertige Komponenten für fokussiertes Web-Crawling, bietet jedoch die Möglichkeit, eigene Erweiterungen und Konfigurationsoptionen hinzuzufügen. Aus diesem Grund wurde SC für den beschriebenen Anwendungsfall erweitert. Bild 1 zeigt die Architektur unseres fokussierten Web-Crawlers.

SC stellt eine rekursive Web-Crawler-Architektur zur Verfügung. Ein *Seed-Injector* liest zunächst ungesehene URLs (Seeds) aus einer Textdatei ein und fügt diese der Frontier hinzu. Im nächsten Schritt werden bisher ungesehen URLs durch eine definierte Menge von *Spouts* aus der *Frontier* ausgelesen. Um den Regeln der Web-Crawler Ethik gerecht zu werden, werden

diese URLs basierend auf ihrem aufgelösten Hostnamen dedizierten Cluster-Knoten zugewiesen und an die jeweiligen *Fetcher* weitergeleitet. Letztere laden die Webseiten herunter und leiten diese an Parser zur Link- und Textextraktion weiter; ungesehene URLs werden der *Frontier* hinzufügt. Werden in diesem Prozess binäre Inhalte (wie z.B. MS Word, PDF) erkannt, werden diese an einen Apache Tika basierten Parser zur Textextraktion weitergeleitet. Wird durch eine Heuristik erkannt, dass eine Webseite dynamischen Inhalt bereitstellt, d.h. Javascript Ausführung für das Laden von Inhalten benötigt, wird *Playwright* für das browsernahe Rendering einsetzt. Anschließend wird der Inhalt an sogenannte *Indexer* gesendet, die den Volltext in einem OpenSearch Cluster speichern.

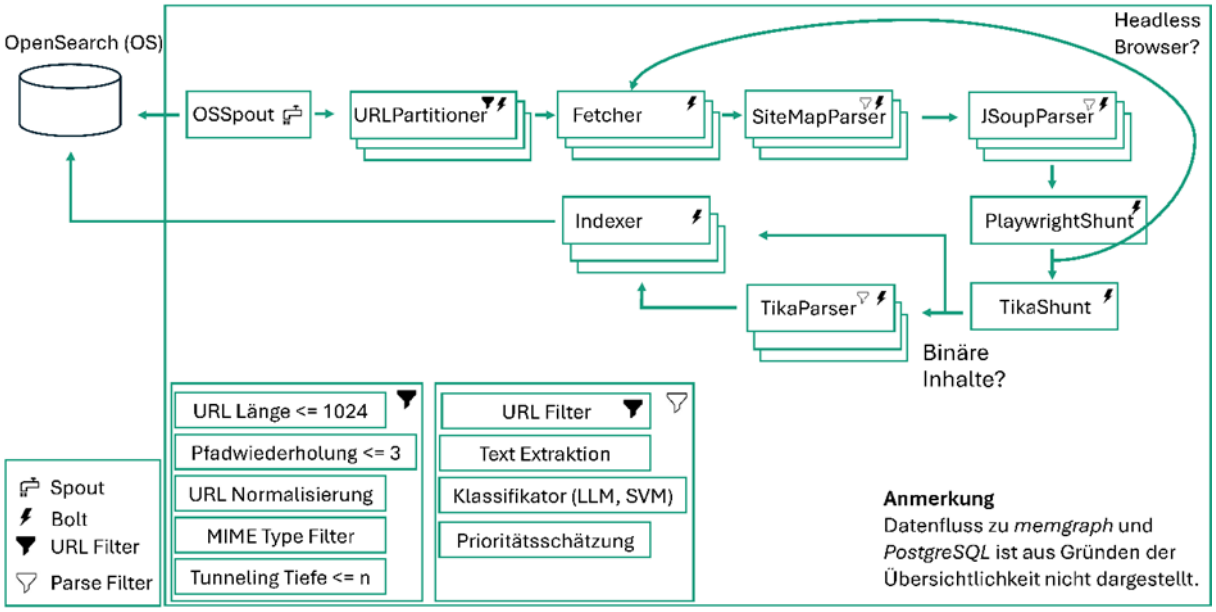

Bild 1: Architektur eines fokussierten Web-Crawlers, basierend auf Apache StormCrawler. Spouts geben Daten aus (hier: URLs), Bolts verarbeiten Daten (d.h. herunterladen, parsen und speichern die extrahierten Inhalte). Bolts können mit URL Filtern oder Parse Filtern erweitert werden

SC wurde durch das Hinzufügen von verschiedenen Bolts und Filter-Komponenten erweitert. Nachdem eine Webseite erfolgreich geparsed wurde, wird der Rohtext mit Hilfe von Boilerplate-Erkennung und XPath-Ausdrücken extrahiert. Anschließend wird er von einer Text-Klassifikationspipeline verarbeitet, um die Relevanz für das freikonfigurierbare Themengebiet zu berechnen. Wenn eine Webseite als relevant eingestuft wird, wird sie für die weitere Verarbeitung vorgemerkt. Als nächstes wird jeder URL, die auf der gegebenen Webseite enthalten ist, ein Prioritätswert zwischen 0 und 127 zugewiesen. Dies geschieht anhand (a) der Klassenwahrscheinlichkeit der aktuellen Webseite, (b) einer Überprüfung, ob die extrahierte URL auf denselben Hostname verweist, (c) der textuellen Beschreibung dieses Links und der Link-URL selbst unter Verwendung eines n-Gram-Ansatzes. Eine hohe Priorität garantiert eine frühere Verarbeitung, um die Themenlokalität bestmöglich auszunutzen.

Zusätzlich dazu wurde eine „soft focused crawling"-Strategie mit Tunneling implementiert. Ziel dieser Strategie ist es, den Web-Crawler nicht bei der ersten irrelevanten Webseite (z.B. Portalstartseite) zu stoppen, sondern die weitere Verarbeitung erst nach *n* weiteren Schritten abzubrechen. Hierzu verfolgt eine spezielle Filterkomponente die Crawl-Tiefe und stoppt nach

vorgegeben *n* irrelevanten Webseiten auf dem jeweiligen Pfad. Irrelevante Webseiten werden nicht indexiert. Um den Web-Graphen des gewählten Themas aufzubauen, verfolgt ein spezieller *Bolt* alle besuchten und entdeckten Links und fügt diese in eine *memgraph*-Graphdatenbank ein. Auf diese Weise entsteht ein themenspezifischer, host-aggregierter Web-Graph für weitere Analysen. Für Statistiken und Metriken, die mit dem Web-Crawl zusammenhängen, aktualisiert ein weiterer *Bolt* kontinuierlich den Fortschritt in einer PostgreSQL-Datenbank. Der gesamte Prozess wird wiederholt, bis die Frontier leer ist oder manuell durch den Benutzer gestoppt wird.

2.2.3 Systemumgebung

Der in dieser Studie eingesetzte fokussierte Web-Crawler wurde auf einem Cluster bestehend aus acht virtuellen Maschinen im Frankfurter Rechenzentrum der IONOS AG betrieben. Die entsprechenden Dienste werden benötigt, um die erfassten Webseiten im laufenden Betrieb zu verarbeiten, zu analysieren sowie die einzelnen Web-Crawler Prozess zu koordinieren. Für die LLM-gestützte Inferenz wurde der IONOS Model Hub eingesetzt.

2.3 Textklassifikation

2.3.1 Support Vector Machines

SVMs sind ein etabliertes Verfahren aus dem Bereich des Maschinellen Lernens, welches in dieser Arbeit zur Textklassifikation analog zu [Joa98, ZWP20] eingesetzt wird. Die zu klassifizierenden Webseiten werden zunächst von nicht relevanten Textbestandteilen (z.B. Navigationselementen), bereinigt. Im nächsten Schritt wird der jeweilige Text tokenisiert und normalisiert. Die die verbleibenden Tokens werden auf ihre Stammform reduziert, um morphologische Variationen zu minimieren. Tokens, die Stoppwörter enthalten (z.B. "und", "oder", "das"), werden entfernt. Im nächsten Schritt werden die vorverarbeiteten Inhalte gemäß dem Bag-of-Words Modell in Vektoren überführt. Zur Reduktion der Dimensionalität wird in dieser Arbeit das Feature-Selection-Verfahren *Information Gain* [YP97] eingesetzt; die Termgewichtigung erfolgt mit *tfc* [SB88]. Als Softwarebibliothek für SVMs wurde *LIBSVM* eingesetzt [CL 2011].

Sowohl KI-relevante Webseiten als auch Webseiten ohne KI-relevante Inhalte werden auf diese Weise verarbeite, um einen SVM-Klassifikator zu trainieren. Die Qualität des Klassifikators wird mit anerkannten Metriken wie *Accuracy, Precision* und *Recall* gemessen.

2.3.2 Large Language Models

LLMs können vielfältige Aufgaben im Bereich der Textverarbeitung übernehmen, einschließlich Textklassifikation [VS25]. Ein zentraler Vorteil von LLMs gegenüber klassischen Verfahren wie SVMs ist, dass kein spezielles Training und damit kein aufwändiges Labeln von Daten erforderlich ist. Dies geht jedoch mit höheren Inferenzzeiten und Ressourcenaufwänden einher. Für den Einsatz eines LLMs zur Textklassifikation wird ein passender Prompt formuliert, der

die Aufgabe beschreibt. Dabei stehen verschiedene Methoden der Prompt-Gestaltung zur Verfügung: (1) *Zero-Shot Prompting*, bei dem lediglich die Aufgabe übermittelt wird [GHS+23], (2) *Few-Shot Prompting*, bei dem zusätzlich zur Aufgabenbeschreibung einige Beispiele bereitgestellt werden [KGR+22], (3) *Chain-of-Thought-Prompting (CoT)*, bei dem das LLM durch schrittweises Denken zur Lösung geführt wird [WWS+22] und (4) *CARP-Prompting*, bei dem eine strukturierte Anleitung zur Reduktion von Mehrdeutigkeiten eingesetzt wird [SLL+23].

Analog zur Vorgehensweise bei SVMs werden die Webseiteninhalte zunächst bereinigt und anschließend durch das LLM klassifiziert. In dieser Arbeit wurde *Zero-Shot-Prompting* gewählt, da es bei vergleichbarer Leistung die geringste Komplexität, die schnellste Antwortzeit und die geringsten Kosten verursacht.

2.4 Topic Modelling

Analog zu [Gro22] und [JB23] wird *BERTopic* als Topic Modelling Methode eingesetzt. Die Integration aktueller LLMs gewährleistet eine zeitnahe Erkennung neuer Begriffe und möglicher Trends. Zudem unterstützt *BERTopic* mehrsprachige Analysen und bietet durch GPU-Beschleunigung eine effiziente Verarbeitung großer Datenmengen [LK24]. Die dynamische Anpassung der zu ermittelnden Themenanzahl ermöglicht einen flexiblen, datengetriebenen Ansatz für die Themenextraktion.

Unser methodischer Ansatz basiert auf einem mehrstufigen Ablauf (siehe Bild 2): Zunächst erfolgt die Textbereinigung (analog wie bei SVMs). Zur Reduktion der rechnerischen Komplexität wird das Vokabular auf Nomen beschränkt. Zusätzlich bleiben nur Wörter mit einer Mindesthäufigkeit von fünf im Textkorpus erhalten.

Anschließend werden die Texte mittels Sentence-Transformers [LZH+20] in numerische Repräsentationen (Embeddings) umgewandelt. Mittels *UMAP* [MHM18] wird die Dimensionalität dieser Embeddings weiter reduziert. Dieser Schritt ist essenziell, um die hochdimensionalen numerischen Repräsentationen der Texte auf eine niedrigere Dimension zu transformieren und somit die nachfolgende Clusterbildung zu erleichtern. Anschließend wird *HDBSCAN* [MHA17] eingesetzt, um die transformierten Embeddings zu clustern.

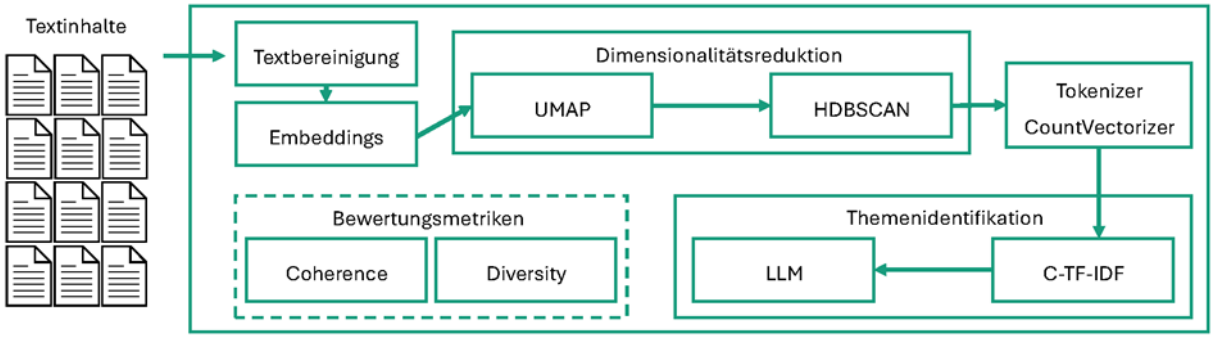

Bild 2: Ablauf Themenidentifikation mittels BERTopic (in Anlehnung an [Gro22])

Nach der Tokenisierung der Texte werden mit C-TF-IDF themenspezifische Schlüsselwörter aus den Clustern extrahiert [Gro22]. Für jedes Thema werden dabei die charakteristischsten und häufigsten Wörter identifiziert, die es am besten repräsentieren. Zur eindeutigen Bestimmung der Bedeutung und des inhaltlichen Fokus eines jeden Themas werden diese Schlüsselwörter zusammen mit den zugehörigen Originaltexten an *LLaMA* 3.1 8B [GDJ+24] übergeben. Hier erfolgt eine LLM-basierte Themenidentifikation.

Für diesen Schritt wurde ein spezifischer Prompt entworfen, der aus einem System-Prompt (SP) und einem Main-Prompt (MP) besteht:

- SP*: You are a classifier tasked with labelling a given list of keywords. Respond with a single label (under 5 words) that best describes the topic. Do not provide more than one label! Avoid explanations, alternatives, or any additional responses.*
- MP: *Respond with exactly one short label of at most 5 words for the topic. I have a set of keywords that represent the topic of these documents: '[DOCUMENTS]'. The topic is described by these keywords: '[KEYWORDS]'.*

Auf diese Weise wird gewährleistet, dass jedes Topic ein prägnantes und eindeutiges Label erhält, das die inhaltliche Essenz des jeweiligen Clusters genau widerspiegelt. Auf diese Weise wird zudem eine manuelle, menschliche Interaktion zur Benennung der Wortcluster vermieden.

Abschließend werden die extrahierten Themen anhand der Metriken *Coherence* und *Diversity* evaluiert. *Coherence* misst die semantische Ähnlichkeit der Wörter innerhalb eines Themas und bewertet dessen inhaltliche Konsistenz [Mim11]. Diversity hingegen gibt an, wie gut sich die einzelnen Topics inhaltlich voneinander abgrenzen [CS21]. In unserem Anwendungsfall liegt der Fokus auf einer hohen *Diversity*, um feine, neu entstehende Trends zu entdecken, die nur in wenigen ausgewählten Texten vorkommen, aber für das Trendscouting von Bedeutung sind.

2.5 Datenakquise

2.5.1 Akquise von Seeds

Für fokussiertes Web-Crawling ist die Auswahl geeigneter, themenspezifischer Seeds essenziell. Zu diesem Zweck wurde im Rahmen dieser Studie die Software *Seed me!* entwickelt. Die in Java geschrieben Software stellt konfigurierbare Suchanfragen an das Google-Such-API und bildet damit ein Standardverfahren zur Generierung von Seeds ab [RLK+16, VBD+16, PÖ11].

Bei der *Erfassung* von wissenschaftlicher, grauer Literatur nutzen wir die von der Literaturdatenbank arXiv zur Verfügung gestellten Kategorien, um nur Inhalte zu erfassen, die KI-Thematiken adressieren. Dabei erfassen wir die Kategorien „KI" (cs.AI), „Machine Learning" (cs.LG), „Computer Vision" (cs.CV) und „Computational Linguistics" (cs.CL). Zur Erfassung der PDF-Dokumente wird *SPIDERWISE* eingesetzt.

2.5.2 Datensätze für Maschinelles Lernen

Im Vorfeld wurde ein eintägiger Web-Crawl durchgeführt, bei dem ein *LLaMA 3.1 8B* Model zur Relevanzschätzung eingesetzt wurde. Dieses Model wurde über den IONOS Model Hub bereitgestellt. Die bei diesem Web-Crawl gesammelten Daten wurden anschließend sowohl für den Trainings- als auch für den Testkorpus genutzt.

2.5.2.1 Trainingskorpus

Der verwendete Trainingskorpus für die SVM besteht aus 5.000 Dokumenten und wurde nach den Prinzipien aus [DS11] und [DS11, WD13] konstruiert. Die entsprechende Klassenzugehörigkeit wurde a-priori durch einen Web-Crawl mit LLM-Unterstützung (*LLaMA 3.1 8B*) ermittelt und stichprobenhaft überprüft.

2.5.2.2 Testkorpus

Für die Erstellung des Testkorpus wurden 500 zufällige Texte aus diesem initialen Web-Crawl ausgewählt. Um die Validität der Testdaten zu gewährleisten, wurde bei der Auswahl der Texte darauf geachtet, dass keine Überschneidungen mit dem Trainingskorpus bestehen. Die Texte wurden manuell von sechs Personen mit KI-Fachwissen bezüglich ihrer KI-Relevanz annotiert und anschließend in einer Peer-Session durch zwei Personen kuratiert und zusammengeführt. *Fleiss-Kappa* lag bei 0,58, während das *Percent Agreement* 72,2 % betrug. Die Annotation und Kuration erfolgte in der Software *INCEpTION* [KBB+18]. Basierend auf diesem Datensatz wurde ein ausgeglichener Datensatz bestehend aus 117 Dokumenten pro Kategorie mit einer Gesamtgröße von 234 Dokumenten erstellt.

3 Ergebnisse

3.1 Seeds

Im Zeitraum 06.2024 bis 04.2025 wurden kontinuierlich Seed URLs über Seed me! erzeugt. Als Anfragen gegen das Google Search API wurden die nachfolgenden Begriffe verwendet: (a) artificial intelligence trends, (b) AI future predictions, (c) AI impact on industries, (d) AI in healthcare trends, (e) AI in finance trends, (f) AI technology advancements, (g) AI investment trends, (h) AI regulatory landscape, (i) AI ethical considerations, (j) AI job market trends.

Für diese Studie wurde die kostenfreie Version der Google Search API eingesetzt, die maximal zehn Treffer pro Anfrage zurückliefert und auf 100 Suchanfragen in einem 24 Stundenzeitfenster limitiert ist. Nach Bereinigung von Dubletten und Hostaggregation verblieben 3870 Seeds, die für die Datenakquise eingesetzt wurden.

3.2 Performance Textklassifikation

Der SVM-Textklassifikator wurde mit n = 5.000 Features trainiert und anhand des Testkorpus evaluiert. Die Ergebnisse sind in Tabelle 1 dargestellt. Der Klassifikator erreichte eine Precision von 0,95, einen Recall von 0,70 und eine Accuracy von 0,78. 39,3 % (46/117) der KI-bezogenen Webseiten wurden von der SVM fälschlicherweise als nicht relevant klassifiziert. Andererseits wurden 4,2 % (5/117) der nicht relevanten Webseiten als KI-relevant klassifiziert.

Der LLM-basierte Klassifikator erreichte eine Precision von 0.88, einen Recall von 0,75 und eine Accuracy von 0,80. 28,2 % (33/117) der KI-bezogenen Webseiten wurden vom LLM fälschlicherweise als nicht relevant klassifiziert. Andererseits wurden 11,1 % (13/117) der nicht relevanten Webseiten als KI-relevant klassifiziert.

Tabelle 1: Evaluationsmetriken für den vorgestellten Testkorpus

	Baseline					
	Relevant	Nicht relevant	Summe	Accuracy	Precision	Recall
SVM	-	-	-	0,78	0,95	0,70
Relevant	112	5	117	-	-	-
Nicht relevant	46	71	117	-	-	-
Summe	158	76	234	-	-	-
LLaMA3.1 8B	-	-	-	0,80	0,88	0,75
Relevant	104	13	117	-	-	-
Nicht relevant	33	84	117	-	-	-
Summe	137	97	234	-	-	-

3.2.1 Web-Crawler-Performance

In einem experimentellen Web-Crawl mit einer Dauer von 168 Stunden wurden insgesamt 1.491.973 KI-relevante Webseiten identifiziert. Die durchschnittliche Harvest Rate in diesem Zeitraum lag bei 0,458. Der fokussierte Web-Crawler erreiche dabei eine Downloadrate zwischen 7 und 12 Webseiten pro Sekunde.

3.3 Themen und Trends

Um eine geeignete Anzahl an Themen zu bestimmen, wurde *BERTopic* mit variierenden minimalen Themengrößen ausgeführt und jeweils die *Diversity* und *Coherence* betrachtet. Bei einer minimalen Themenanzahl von 30 erreichen sowohl *Diversity* als auch *Coherence* ein hohes Niveau. Ab höheren Werten nimmt die Diversität deutlich ab. Daher wurde für die weitere Analyse eine minimale Themenanzahl von 30 für die weiteren Analysen gewählt.

Zur vertieften und benutzerfreundlichen Analyse wurden die Ergebnisse in einem interaktiven Dashboard visualisiert. *Bild 3* zeigt die entwickelte Webanwendung. Auf der linken Seite zeigt eine Themenkarte der generierten Hauptthemen, darunter die Top 5: *Privacy and Security Challenges, Natural Language Processing Techniques, Graph Neural Networks, Causal Graph Analysis* und *Quantum Computing Technology*. In der unteren linken Ecke wird zusätzlich die zeitliche Entwicklung der Themenanzahl dargestellt.

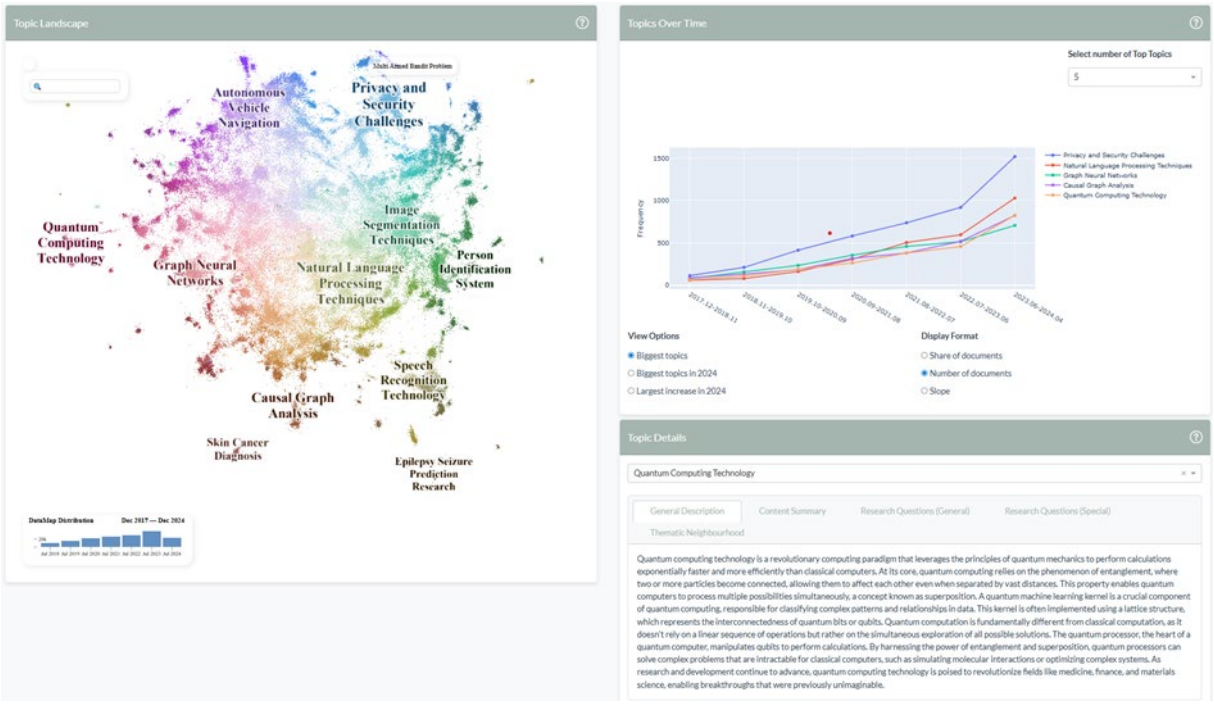

Bild 3: Trendscouting Dashboard als Web-Anwendung

Im oberen rechten Bereich wird eine zeitliche Analyse aller generierten Themen dargestellt, die Veränderungen in ihrer Relevanz über den Zeitverlauf sichtbar macht. Nutzer können dabei eine beliebige Anzahl an Top-Themen auswählen. Die Optionen *Biggest Topics, Biggest Topics in 2024* und *Largest Increase in 2024* ermöglichen eine differenzierte Betrachtung der wichtigsten und am stärksten wachsenden Themen. Letztere Darstellung, basierend auf der Steigung (Slope), zeigt, dass die meisten der im Jahr 2024 am stärksten wachsenden Themenim Bereich Natural Language Processing (NLP) liegen. Die vorderen Plätze belegen u. a. Natural Language Processing Technology, Language Model Evaluation und *Language Safety and Toxicity*. Dies spiegelt die zunehmende Aufmerksamkeit gegenüber GenAI bzw. LLMs in der breiten Öffentlichkeit wider.

Im unteren rechten Bereich werden detaillierte Beschreibungen zu jedem Thema präsentiert. Dabei werden sowohl die Inhalte der Veröffentlichungen als auch die Schlüsselwörter der Themen verwendet, um mit Hilfe eines LLMs verschiedene Aspekte der Themen im Detail zu beleuchten. Ergänzend wird für jedes Thema die thematische Nachbarschaft visualisiert, die Zusammenhänge zu verwandten Themen aufzeigt und so eine kontextuelle Einordnung des jeweiligen Themas ermöglicht.

4 Diskussion & Ausblick

4.1 Zusammenfassung

Die vorliegende Arbeit zeigt, dass fokussiertes Web-Crawling unter Einsatz von LLMs zur Textklassifikation Fortschritte bei der effizienten Erfassung und Analyse themenspezifischer Webinhalte sowie bei der frühzeitigen Erkennung und Beobachtung von Trends ermöglicht.

Die Ergebnisse zeigen zudem, dass traditionelle maschinelle Lernverfahren wie SVMs eine hohe Präzision erzielen, allerdings auf gut annotierte Trainingsdaten angewiesen sind und dennoch Schwächen im Recall aufweisen. Im Gegensatz dazu benötigen LLMs keine umfangreichen Trainingsdaten, bieten eine vielversprechende Alternative und liefern akzeptable Ergebnisse. Vor diesem Hintergrund erscheint ein hybrider Ansatz besonders vielversprechend, bei dem LLMs zur effizienten Datengenerierung eingesetzt werden, um darauf aufbauend leistungsstarke SVM-Modelle zu trainieren.

4.2 Limitationen

Für diese Arbeit bestehen mehrere Einschränkungen. Die verfügbare Datenmenge zur Validierung des SVM-Modells und des LLM ist vergleichsweise gering. Obwohl die Testdaten manuell kuratiert wurden und aus authentischen Webquellen stammen, wodurch eine hohe Qualität gegeben ist, begrenzt die geringe Größe des Korpus die Aussagekraft der Klassifikator-Performance. Sowohl die SVM- als auch die LLM-Modelle zeigten eine ähnliche Genauigkeit, unterscheiden sich jedoch in den Inferenzzeiten. Dies macht den Ansatz vielversprechend, zunächst LLMs für die Datengenerierung einzusetzen und darauf aufbauend effizientere SVM-Modelle zu trainieren, um den Durchsatz des Systems zu steigern. Zudem stellen API-Limitationen beim Einsatz externer LLM-Dienste sowie der vergleichsweise hohe Rechenaufwand und die Kosten im Vergleich zu SVM-Modellen weitere Einschränkungen dar, die insbesondere bei skalierter Anwendung berücksichtigt werden müssen.

Die Crawling-Dauer von etwa einer Woche ist zudem kurz, insbesondere im Hinblick auf die Erfassung von Zeitreihendaten. Eine längere Beobachtungsperiode wäre notwendig, um aussagekräftigere und belastbarere Aussagen zur Entwicklung und Leistungsfähigkeit des Systems über die Zeit hinweg treffen zu können. Darüber hinaus könnte, wie von [VBD+16] vorgeschlagen, eine tiefgreifende Analyse des durch das Web-Crawling entstandenen host-aggregierten Web-Graphs durchgeführt werden, um die Ergebnisse insbesondere mit Hinblick auf eine repräsentative Themenerfassung weiterhin methodisch abzusichern. Dennoch liefern die bisherigen Ergebnisse bereits erste positive Hinweise auf die Stabilität und Effektivität des vorgestellten Ansatzes.

Eine systematische Validierung der identifizierten Themen und Trends, etwa durch Expertenbefragungen, wurde bislang noch nicht durchgeführt. Dieser Schritt ist jedoch entscheidend für zukünftige Untersuchungen, um die Generalisierbarkeit und Validität des Ansatzes im Vergleich zu etablierten Verfahren fundiert zu bestätigen. Zudem könnte die Auswahl der Seed-Seiten das Themenspektrum nur unzureichend abdecken, da diese auf spezifischen Abfragen

gegenüber Google basieren. Dadurch wird die Vielfalt der erfassten Themen und somit die Repräsentativität des Datensatzes möglicherweise eingeschränkt.

4.3 Vergleich zu Vorarbeiten

Fokussierte Web-Crawler werden in verschiedenen Anwendungsbereichen eingesetzt, wobei die erzielte HR je nach Methode und Zielgebiet variieren. Rheinländer et al. erreichten mit einem Naive Bayes Klassifikator bei der Sammlung englischsprachiger biomedizinischer Webinhalte eine HR von 38 % [RLK+16]. Zowalla et al. setzten einen SVM Ansatz für deutschsprachige Gesundheitswebseiten ein und erzielten eine HR von 21,2 % [ZWP20]. Sakai et al. erreichten bei der automatischen Erkennung japanischer „Fake Shops“ mit klassischen maschinellen Lernverfahren eine HR von über 20 % [STK+23], während [RN23] mit SVR eine HR von 37 % berichteten.

Unsere Ergebnisse weisen eine Harvest Rate von 45,8 % auf, was im typischen Bereich vergleichbarer fokussierter Web-Crawler liegt. Trotz möglicher Einflussfaktoren, die die Harvest Rate beeinflussen könnten, unterstreicht dies die Effektivität unseres Ansatzes.

Die vorliegende Arbeit zeigt zudem, dass LLMs im Kontext des Trendscoutings nicht nur – wie häufig in der Literatur betont – zur Analyse von Daten, sondern auch in Verbindung mit gezieltem Web-Crawling zur Datenerhebung eingesetzt werden kann. Die Validität unseres Ansatzes spiegelt sich auch in der durch *BERTopic* ermöglichten Interpretierbarkeit der gesammelten Daten wider, die mit jener von Studien vergleichbar ist, in denen die Datenerhebung ohne den Einsatz von KI erfolgte [KW:24, JB23].

4.4 Ausblick und zukünftige Forschungsperspektiven

Ein zentraler nächster Schritt besteht in der systematischen Validierung der identifizierten Themen und Trends mittels etablierter Verfahren, wie beispielsweise Expertenbefragungen oder Benchmark-Vergleichen mit traditionellen Methoden. Diese Validierung ist entscheidend, um die Generalisierbarkeit und Verlässlichkeit unseres KI-gestützten Ansatzes im Vergleich zu klassischen Verfahren belegen zu können. Darüber hinaus bietet der Einsatz von *Active Learning* Methoden mit LLMs [XMX+25] großes Potenzial, den fokussierten Web-Crawling-Prozess kontinuierlich zu verbessern. Durch gezielte menschliche Rückmeldungen können die Prompts der generativen Modelle und die nachgeschalteten SVMs iterativ optimiert werden. Ein solcher zyklischer Lernprozess – LLM zur Datengenerierung, SVM zum effizienten Klassifizieren, Rückkopplung der Ergebnisse zur weiteren Verfeinerung der LLMs – ermöglicht eine adaptive und zunehmend präzisere Datenakquise und Klassifikation.

Der Einsatz von LLMs beim fokussierten Web-Crawling eröffnet zudem die Perspektive, wesentlich flexibler und kundenspezifischer auf unterschiedliche Anforderungen reagieren zu können. Durch die Fähigkeit, semantisch relevante Inhalte gezielt zu erkennen und dynamisch auf neue Themenlagen einzugehen, lassen sich sowohl die Effizienz der Datenerfassung als auch die Qualität der Analyse steigern. Dies macht diesen Ansatz besonders attraktiv für Anwendungen im Trendscouting und der technologischen Vorausschau, bei denen Aktualität und Themenvielfalt von hoher Bedeutung sind.

Zukünftige Forschungen sollten daher neben der Validierung auch die Erweiterung des fokussierten Web-Crawling über längere Zeiträume hinweg in den Fokus nehmen, um belastbare Zeitreihendaten zu generieren und so die Entwicklung und Identifikation von Trends noch besser abzubilden. Zudem könnte die Diversifizierung der Seed-Seiten und die Integration weiterer Datenquellen dazu beitragen, die Repräsentativität und Breite der erfassten Themen zu erhöhen.

Literatur

[AW24] AHLUWALIA, A.; WANI, S.: Leveraging Large Language Models for Web Scraping. arXiv, 2024

[BNJ03] BLEI, D. M.; NG, A. Y.; JORDAN, M. I.: Latent dirichlet allocation. Journal of machine Learning research, (3)Jan, 2003, S. 993–1022

[BP12] BRIN, S.; PAGE, L.: Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer Networks, (56)18, 2012, S. 3825–3833

[BRR+24] BENFENATI, D.; RINALDI, A.; RUSSO, C.; TOMMASINO, C.: GenCrawl: A Generative Multimedia Focused Crawler for Web Pages Classification: Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. 16th International Conference on Knowledge Discovery and Information Retrieval, 17.11.2024 - 19.11.2024, Porto, Portugal, SCITEPRESS - Science and Technology Publications, 2024, S. 91–101

[CL 2011] CHANG, C.-C.; LIN, C.-J.: LIBSVM. ACM Transactions on Intelligent Systems and Technology, (2)3, 2011, S. 1–27

[CS21] CHAUHAN, U.; SHAH, A.: Topic modeling using latent Dirichlet allocation: A survey. ACM Computing Surveys (CSUR), (54)7, 2021, S. 1–35

[CvD99] CHAKRABARTI, S.; VAN DEN BERG, M.; DOM, B.: Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks, (31)11-16, 1999, S. 1623–1640

[Dav00] DAVISON, B. D.: Topical locality in the Web. In: Yannakoudakis, E.; Belkin, N. J.; Leong, M.-K.; Ingwersen, P. (Hrsg.): Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR00: 23rd ACM International SIGIR Conference on Research and Development in Information Retrieval, 24 07 2000 28 07 2000, Athens Greece, ACM, New York, NY, USA, 2000, S. 272–279

[DS11] DOBBIN, K. K.; SIMON, R. M.: Optimally splitting cases for training and testing high dimensional classifiers. BMC medical genomics, (4), 2011, S. 31

[DSL+ 2025] DAHLKE, J.; SCHMIDT, S.; LENZ, D.; KINNE, J.; DEHGHAN, R.; ABBASIHAROFTEH, M.; SCHÜTZ, M.; KRIESCH, L.; HOTTENROTT, H.; KANILMAZ, U. N.; GRASHOF, N.; HAJIKHANI, A.; LIU, L.; RICCABONI, M.; BALLAND, P.-A.; WÖRTER, M.; RAMMER, C.: The WebAI Paradigm of Innovation Research: Extracting Insight from Organizational Web Data Through AI – The WebAI Paradigm of Innovation Research: Extracting Insight from Organizational Web Data Through AI, 2025

[DSL+23] DUMBACH, P.; SCHWINN, L.; LÖHR, T.; ELSBERGER, T.; ESKOFIER, B. M.: Artificial intelligence trend analysis in German business and politics: a web mining approach. International Journal of Data Science and Analytics, 2023

[Eic95] EICHMANN, D.: Ethical Web agents. Computer Networks and ISDN Systems, (28)1-2, 1995, S. 127–136

[FNB24] FERRÀS, X.; NYLUND, P.; BREM, A.: Connecting The (Invisible) Dots. In: Chesbrough, H.; Radziwon, A.; Vanhaverbeke, W.; West, J. (Hrsg.): The Oxford Handbook of Open Innovation. Oxford University Press, 2024, S. 519–532

[GDJ+24] GRATTAFIORI, A.; DUBEY, A.; JAUHRI, A.; PANDEY, A.; KADIAN, A.; AL-DAHLE, A.; LETMAN, A.; MATHUR, A.; SCHELTEN, A.; VAUGHAN, A.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024

[GHS+23] GRETZ, S.; HALFON, A.; SHNAYDERMAN, I.; TOLEDO-RONEN, O.; SPECTOR, A.; DANKIN, L.; KATSIS, Y.; ARVIV, O.; KATZ, Y.; SLONIM, N.; EIN-DOR, L.: Zero-shot Topical Text Classification with LLMs - an Experimental Study. In: BOUAMOR, H.; PINO, J.; BALI, K. (Hrsg.): Findings of the Association for Computational Linguistics: EMNLP 2023. Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, Association for Computational Linguistics, Stroudsburg, PA, USA, 2023, S. 9647–9676

[Gro22] GROOTENDORST, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794, 2022

[Gro22] GROOTENDORST, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022w

[HG 2015] HINES, A.; GOLD, J.: An organizational futurist role for integrating foresight into corporations. Technological Forecasting and Social Change, (101), 2015, S. 99–111

[HSM24] HAN, B.; SUSNJAK, T.; MATHRANI, A.: Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview. Applied Sciences, (14)19, 2024, S. 9103

[Hua25] HUANG, C.-J.: The Synergy of Automated Pipelines with Prompt Engineering and Generative AI in Web Crawling. arXiv, 2025

[JB23] JIN, B.; BAE, Y.: Prospective Research Trend Analysis on Zero-Energy Building (ZEB): An Artificial Intelligence Approach. Sustainability, (15)18, 2023, S. 13577

[Joa98] JOACHIMS, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Carbonell, J. G.; Siekmann, J.; Goos, G.; Hartmanis, J.; Van Leeuwen, J.; Nédellec, C.; Rouveirol, C. (Hrsg.): Machine Learning: ECML-98. Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, S. 137–142

[KBB+18] KLIE, J.-C.; BUGERT, M.; BOULLOSA, B.; ECKART DE CASTILHO, R.; GUREVYCH, I.: The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Zhao, D. (Hrsg.): Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Santa Fe, New Mexico, 2018, S. 5–9

[KBR17] KUMAR, M.; BHATIA, R.; RATTAN, D.: A survey of Web crawlers for information retrieval. WIREs Data Mining and Knowledge Discovery, (7)6, 2017

[KGR+22] KOJIMA, T.; GU, S. S.; REID, M.; MATSUO, Y.; IWASAWA, Y.: Large Language Models are Zero-Shot Reasoners. arXiv, 2022

[KW:24] KUMAR, D.; WEISSENBERGER-EIBL, M.; UNAV: Artificial Intelligence Driven Trend Forecasting: Integrating BERT Topic Modelling and Generative Artificial Intelligence for Semantic Insights. Fraunhofer-Gesellschaft, 2024

[Liu11] LIU, B.: Web Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011

[LK24] LEE, Y.-G.; KIM, S.: A Comparative Study on Topic Modeling of LDA, Top2Vec, and BERTopic Models Using LIS Journals in WoS. Journal of the Korean Society for Library and Information Science, (58)1, 2024, S. 5–30

[LWN+24] LESMANA, R.; WIJAYA, I.; NABILA, E. A.; AGUSTIAN, H.; AUDIAH, S.; FATURAHMAN, A.: Enhancing Market Trend Analysis Through AI Forecasting Models. International Journal of Cyber and IT Service Management, (4)2, 2024, S. 105–113

[LZH+20] LI, B.; ZHOU, H.; HE, J.; WANG, M.; YANG, Y.; LI, L.: On the sentence embeddings from pre-trained language models. arXiv preprint arXiv:2011.05864, 2020

[Men04] MENCZER, F.: Lexical and semantic clustering by Web links. Journal of the American Society for Information Science and Technology, (55)14, 2004, S. 1261–1269

[Men05] MENCZER, F.: Mapping the Semantics of Web Text and Links. IEEE Internet Computing, (9)3, 2005, S. 27–36

[MHA17] MCINNES, L.; HEALY, J.; ASTELS, S.: hdbscan: Hierarchical density based clustering. J. Open Source Softw., (2)11, 2017, S. 205

[MHM18] MCINNES, L.; HEALY, J.; MELVILLE, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018

[Mim11] MIMNO, D., WALLACH, H., TALLEY, E., LEENDERS, M., & MCCALLUM: Optimizing semantic coherence in topic models, 2011

[MMB14] MEUSEL, R.; MIKA, P.; BLANCO, R.: Focused Crawling for Structured Data. In: Li, J.; Wang, X. S.; Garofalakis, M.; Soboroff, I.; Suel, T.; Wang, M. (Hrsg.): Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. CIKM '14: 2014 ACM Conference on Information and Knowledge Management, 03 11 2014 07 11 2014, Shanghai China, ACM, New York, NY, USA, 2014, S. 1039–1048

[MSA+ 2025] MATAMOROS-ECHEVERRIA, K. L.; SANCHEZ-LEON, E. R.; ARISTEGA-ZUÑIGA, A. A.; CÁRDENAS-RODRÍGUEZ, M. M.; PERALTA-GAMBOA, D. A.: The Impact of Artificial Intelligence on the Digital Economy: Advances and Challenges - A Systematic Literature Review. Journal of Posthumanism, (5)2, 2025

[PFN 2023] PLEKHANOV, D.; FRANKE, H.; NETLAND, T. H.: Digital transformation: A review and research agenda. European Management Journal, (41)6, 2023, S. 821–844

[PÖ11] PRASATH, R.; ÖZTÜRK, P.: Finding Potential Seeds through Rank Aggregation of Web Searches. In: Kuznetsov, S. O.; Mandal, D. P.; Kundu, M. K.; Pal, S. K. (Hrsg.): Pattern Recognition and Machine Intelligence. Lecture Notes in Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, S. 227–234

[RFB+21] RODRIGUES, A. P.; FERNANDES, R.; BHANDARY, A.; SHENOY, A. C.; SHETTY, A.; ANISHA, M.: Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques. Wireless Communications and Mobile Computing, (2021)1, 2021

[RK 2018] ROHRBECK, R.; KUM, M. E.: Corporate foresight and its impact on firm performance: A longitudinal analysis. Technological Forecasting and Social Change, (129), 2018, S. 105–116

[RLK+16] RHEINLÄNDER, A.; LEHMANN, M.; KUNKEL, A.; MEIER, J.; LESER, U.: Potential and Pitfalls of Domain-Specific Information Extraction at Web Scale. In: Özcan, F.; Koutrika, G.; Madden, S. (Hrsg.): Proceedings of the 2016 International Conference on Management of Data. SIGMOD/PODS'16: International Conference on Management of Data, 26 06 2016 01 07 2016, San Francisco California USA, ACM, New York, NY, USA, 2016, S. 759–771

[RN23] RAJIV, S.; NAVANEETHAN, C.: An Optimal Topic Centric Crawler for Acquiring Bio-medical Themes Utilizing Gaussian Support Vector Regression. SN Computer Science, (4)6, 2023

[Roh14] ROHRBECK, R.: Trend Scanning, Scouting and Foresight Techniques. In: Gassmann, O.; Schweitzer, F. (Hrsg.): Management of the Fuzzy Front End of Innovation. Springer International Publishing, Cham, 2014, S. 59–73

[SB88] SALTON, G.; BUCKLEY, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management, (24)5, 1988, S. 513–523

[SHS+21] SCHUH, G.; HICKING, J.; STROH, M.-F.; BENNING, J.; GNANARAJ, C.: FEASIBILITY Analysis of Entity Recognition as a Means to Create an Autonomous Technology Radar. Hannover publish-Ing, 2021

[SLL+23] SUN, X.; LI, X.; LI, J.; WU, F.; GUO, S.; ZHANG, T.; WANG, G.: Text Classification via Large Language Models. arXiv, 2023

[STK+23] SAKAI, K.; TAKESHIGE, K.; KATO, K.; KURIHARA, N.; ONO, K.; HASHIMOTO, M.: An Automatic Detection System for Fake Japanese Shopping Sites Using fastText and LightGBM. IEEE Access, (11), 2023, S. 111389–111401

[VBD+16] VIEIRA, K.; BARBOSA, L.; DA SILVA, A. S.; FREIRE, J.; MOURA, E.: Finding seeds to bootstrap focused crawlers. World Wide Web, (19)3, 2016, S. 449–474

[vBK16] VAN DEN BOSCH, A.; BOGERS, T.; KUNDER, M. DE: Estimating search engine index size variability: a 9-year longitudinal study. Scientometrics, (107), 2016, S. 839–856

[VS25] VAJJALA, S.; SHIMANGAUD, S.: Text Classification in the LLM Era—Where do we stand? arXiv, 2025

[WD13] WEI, Q.; DUNBRACK, R. L.: The role of balanced training and testing data sets for binary classifiers in bioinformatics. PloS one, (8)7, 2013, e67863

[WWS+22] WANG, X.; SCHUURMANS, D.; BOSMA, M.; ICHTER, B.; XIA, F.; CHI, E. H.; LE, Q. V.; ZHOU, D.: Chain-of-thought prompting elicits reasoning in large language models: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc, Red Hook, NY, USA, 2022

[XMX+25] XIA, Y.; MUKHERJEE, S.; XIE, Z.; WU, J.; LI, X.; APONTE, R.; LYU, H.; BARROW, J.; CHEN, H.; DERNONCOURT, F.; KVETON, B.; YU, T.; ZHANG, R.; GU, J.; AHMED, N. K.; WANG, Y.; CHEN, X.; DEILAMSALEHY, H.; KIM, S.; HU, Z.; ZHAO, Y.; LIPKA, N.; YOON, S.; HUANG, T.-H. K.; WANG, Z.; MATHUR, P.; PAL, S.; MUKHERJEE, K.; ZHANG, Z.; PARK, N.; NGUYEN, T. H.; LUO, J.; ROSSI, R. A.; MCAULEY, J.: From Selection to Generation: A Survey of LLM-based Active Learning, 2025

[YP97] YANG, Y.; PEDERSEN, J. O.: A comparative study on feature selection in text categorization. Icml, 412-420, 1997, S. 35

[ZWP20] ZOWALLA, R.; WETTER, T.; PFEIFER, D.: Crawling the German Health Web: Exploratory Study and Graph Analysis. Journal of medical Internet research, (22)7, 2020, e17853

Autoren

Dr. Richard Zowalla ist wissenschaftlicher Mitarbeiter im Team Cognitive Service im Forschungsbereich Kognitive Dienstleistungssysteme am Fraunhofer IAO in Heilbronn. Er promovierte als Dr. sc. hum. im Bereich Medizinische Informatik an der Universität Heidelberg im Bereich des fokussierten Web-Crawling. Zu seinen Kernkompetenzen zählen Softwareentwicklung, Web Data Mining sowie der Einsatz von Open Source Software. Zudem ist er Mitglied der Apache Software Foundation und der aktuelle Chair des Apache StormCrawler Projekts.

Jan Mackensen ist wissenschaftlicher Mitarbeiter im Team Automated Service Interactions im Forschungsbereich Kognitive Dienstleistungssysteme am Fraunhofer IAO in Heilbronn. Er hat einen Masterabschluss (M. Sc.) in Cognitive Science und spezialisiert sich auf Machine Learning, Künstliche Intelligenz, Natural Language Processing sowie Kognitionswissenschaft.

Jin Meng ist wissenschaftliche Mitarbeiterin im Team Cognitive Distribution Systems im Forschungsbereich Kognitive Dienstleistungssysteme am Fraunhofer IAO in Heilbronn. Sie verfügt über einen Masterabschluss (M. Sc.) und bringt ihre Erfahrung in den Bereichen Machine Learning, Deep Learning sowie Datenanalyse und -visualisierung ein die Forschung ein.

Kristian Schaefer ist wissenschaftlicher Mitarbeiter im Team Cognitive Service Technologies im Forschungsbereich Kognitive Dienstleistungssysteme am Fraunhofer IAO in Heilbronn. Er verfügt über einen Masterabschluss (M. Sc.) und ist spezialisiert auf Softwarearchitekturen im Bereich IoT und Cloud. Seine Kernkompetenzen umfassen IoT-/Cloud-Anwendungen, Datenverarbeitung und -analysen, Softwareentwicklung sowie Embedded Systems.

Dr. Safa Omri leitet das Team für Cognitive Service Technologies im Forschungsbereich Kognitive Dienstleistungssysteme am Fraunhofer IAO in Heilbronn. Sie ist auf Smart Scheduling, angewandte KI für industrielle Systeme sowie die digitale Service-Transformation spezialisiert. Safa Omri promovierte in Mathematik und Informatik am Karlsruher Institut für Technologie

(KIT), wobei sie sich auf qualitätsbewusste Testfall Priorisierung mithilfe datengetriebener und Reinforcement Learning-basierter Methoden in industriellen Umgebungen konzentrierte.

Dr. Jens Neuhüttler leitet den Forschungsbereich Kognitive Dienstleistungssysteme am Fraunhofer IAO in Heilbronn und Stuttgart. Er promovierte in Ingenieurswissenschaften an der Universität Stuttgart und war Visiting Researcher an der Universität Cambridge. Seit mehr als zehn Jahren beschäftigt er sich in angewandten Forschungs- und Beratungsprojekten mit der systematischen Entwicklung innovativer, datenbasierter Dienstleistungen und Geschäftsmodelle in verschiedenen Branchen.