Einsatz neuer Daten in Kommunikationszusammenhängen auf Unternehmensebene (EnDiKaU)

AP 2: Webseiten-basierte Erhebung bei allen hessischen Unternehmen mit Webauftritt zur Verfügbarkeit und Nutzung neuartiger Kommunikationsdaten

Technical Report

Justus-Liebig-Universität Gießen
Institut für Geographie
Professur für Wirtschaftsgeographie
Senckenbergstrasse 1
D-35390 Giessen, Germany

Prof. Dr. Stefan Hennemann
Phone +49 (0) 641-99-36222
Stefan.Hennemann@geogr.uni-giessen.de

Sina Sänger, M.Sc.
Phone +49 (0) 641-99-36224
Sina.Saenger@geogr.uni-giessen.de

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Executive Summary

1. Objective and Context

As part of the research project EnDiKaU, working package 2 explores the potential of unstructured web data and advanced Natural Language Processing (NLP) to generate novel economic indicators. Traditional statistical methods often lack the regional granularity required for targeted economic policymaking. To bridge this gap, this project utilizes website texts to automatically classify small and medium-sized enterprises (SMEs) in the German state of Hesse into specific sub-industries (Information Economy and Manufacturing) and to measure their latent "digital affinity" as an indicator of digital transformation that is difficult to capture through conventional secondary data.

2. Data and Methodology

The analysis relies on web data extracted from the Common Crawl corpus, focusing on active German domains with a legal imprint (Impressum). To minimize noise, the data processing specifically targeted highly informative subpages, such as "Home" and "About Us" sections.

The analytical core of the project is a highly efficient, three-stage Machine Learning pipeline utilizing the SetFit architecture based on the DistilBERT model, which is optimized for few-shot learning:

Step 1: Binary Pre-filtering. The model successfully separates companies belonging to the target sectors (Information Economy and Manufacturing) from non-relevant entities with an accuracy of 97.8%.

Step 2: Multiclass Sub-industry Classification. Relevant companies are assigned to specific sub-industries with an accuracy of 92.4%.

Step 3: Digital Affinity Classification. A specialized binary model identifies whether a company actively communicates digital transformation (e.g., "Cloud", "Industry 4.0"). This model accounts for industry-specific digital terminology and achieved an F1-Score of 97.4%.

3. Key Findings: Industry Structure and Regional Clusters

Out of approximately 185,000 analyzed domains in Hesse, roughly 43,000 (23.2%) were identified as relevant SMEs. The geographic distribution accurately reflects historically grown economic structures and clusters.

Urban Hubs: The Rhine-Main area (especially Frankfurt) and Kassel show the highest concentration of relevant businesses.

Regional Specializations: The NLP models successfully mapped regional clusters, such as the automotive industry in Fulda and North Hesse, the chemical and pharmaceutical sector in the Bergstraße district, and knowledge-intensive services (research, engineering) around Darmstadt.

4. Key Findings: Digital Affinity

Overall, 21.4% of the identified SMEs communicate a clear digital affinity on their websites. However, the analysis reveals a significant sectoral and regional divide.

The Sectoral Gap: Knowledge-intensive services exhibit high digitalization rates (e.g., ICT services at 59.2%, Marketing at 44.9%). In contrast, traditional manufacturing sectors show much lower outward digital visibility (Mechanical Engineering at 16.2%, Chemicals/Pharma at 6.4%).

The Impact of Regional Networks: A deeper analysis using the Location Quotient (LQ) reveals a crucial nuance. While traditional industries generally show lower digital affinity, companies within specific regional clusters (e.g., the automotive sector in Kassel or Groß-Gerau) demonstrate an above-average digital affinity compared to their peers in other regions. This strongly suggests that established regional transformation networks and cluster initiatives successfully foster digital modernization in traditional industries.

5. Limitations and Interpretation

While the methodology provides deep regional insights, the results must be interpreted with specific caveats.

Communicated vs. Internal Digitalization: The models measure visible digital affinity based on website texts. B2B companies (like machinery manufacturers) often digitize internal production and supply chain processes without advertising them on their websites, leading to a potential underestimation of their actual digital maturity compared to B2C businesses.

SME Focus: The focus on SMEs means the immense digital innovation and capital-intensive infrastructure of large corporations (e.g., in the pharmaceutical sector) are not reflected in these figures.

6. Conclusion

The project successfully demonstrates that web data and modern Transformer models can reliably map regional economic structures. Furthermore, the findings highlight that digital transformation is not a uniform, state-wide phenomenon, but rather a regionally and sectorally differentiated structural shift that is heavily supported by local industry competencies and innovation networks.

1. Einleitung

Im wissenschaftlichen Erkenntnisprozess sind zwei große methodologische Zugänge etabliert, die sich im Zusammenhang mit der empirischen Umsetzung in ein qualitatives Methodeninstrumentarium und ein quantitatives Methodeninstrumentarium einteilen lassen. Derzeit sind vor allem Mischformen als Mixed-Methods-Ansätze etabliert, die je nach Forschungsfrage zwischen qualitativen und quantitativen Elementen variieren, um sowohl breit angelegt verallgemeinern zu können als auch Details in der gebotenen Tiefe ergründen zu können. Seit wenigen Jahren beginnt sich mit der rasanten Entwicklung in der Algorithmik und der Datenverfügbarkeit ein weiteres Forschungsinstrumentarium zu etablieren. Der schier unendliche Bestand an (unstrukturierten) Textdaten, der sich seit der Entstehung des World Wide Web und anderer Datensysteme entwickelt hat, trifft auf Algorithmen und Software-Architekturen sowie auf eine stark wachsende Hardwarefähigkeit, um diese extreme Datenfülle zu strukturieren und zu analysieren.

Diese Entwicklung ist sowohl aus der Forschungsperspektive von hoher Relevanz, da völlig neue Indikatoren aus den Textdaten abgeleitet werden können, als auch aus der Unternehmenssicht eine wichtige Entwicklung, aus der neue Geschäftsmodelle und Steuerungsmöglichkeiten entstehen können. Insbesondere der Aspekt der Nutzung von Kommunikationsdaten für neue oder verbesserte Produkte und Dienstleistungen im Unternehmenszusammenhang stehen im Fokus des Forschungsprojekts EnDiKaU. Gleichzeitig ist es ein Ziel im Projekt, auch die neuen Perspektiven für die Erzeugung neuartiger Forschungsdaten und die umfangreiche, aber gleichzeitig detaillierte Analyse von Textdaten auszuloten.

Ein Vorteil der Erschließung der neuen Daten und Verfahren liegt in der Fähigkeit, eher latente, aber inhaltlich relevante Indikatoren in statistische Analysen zur Unternehmenstätigkeit und regionalen Entwicklungsprozessen einfließen zu lassen. Hierzu zählt die latente Konstruktion der Digitalaffinität von Unternehmen, die keine direkte Messmöglichkeit mit herkömmlichen Sekundärdaten bietet. Hierfür sind typischerweise Primärbefragungen nötig, wie sie auch im Projekt EnDiKaU durchgeführt wurden, aber hinsichtlich der branchenbezogenen und regionalen Differenzierung nicht hinreichend sind. Das statistische Bundesamt befragt ebenfalls regelmäßig Unternehmen zu ausgewählten IuK-Themen, aber auch hier sind hinsichtlich der regionalen Auflösung und projektspezifischer Branchenzusammenstellungen keine befriedigenden Erkenntnisse für den Forschungsprozess abzuleiten. Eine Herausforderung ist dabei, eine größere Stichprobe zu rekrutieren, um regionalisierte Daten zu gewinnen, wie sie für die regionalpolitische Steuerung nötig und von der (regionalen) Wirtschaftsförderung gewünscht sind.

An diesem Punkt setzt das Arbeitspaket an, das auf der Grundlage von Webseitendaten und neuen textanalytischen Methoden auf der kleinräumigen politisch-administrativen Ebene zusätzliche Erkenntnisse bezüglich der Branchenstrukturen und deren Digitalisierungsperspektiven erwarten lassen. Zweifelsfrei liegen hier neue Herausforderungen in der Analyse und der Abdeckung der Datenbasis, die es gilt zu diskutieren. Vor allem ist die bislang oft fehlende Validierung mit externen Datenquellen eine Herausforderung. Aus Forschungssicht soll das Projekt einen Beitrag für Hessen leisten, indem die Ergebnisse aus der Webdatenanalyse anschließend mithilfe qualitativer Verfahren und Unternehmensinterviews in Beziehung gesetzt werden. Vorlaufend profitiert der Analyseschritt von den Ergebnissen der quantitativen Unternehmensbefragung im Arbeitspaket 1.

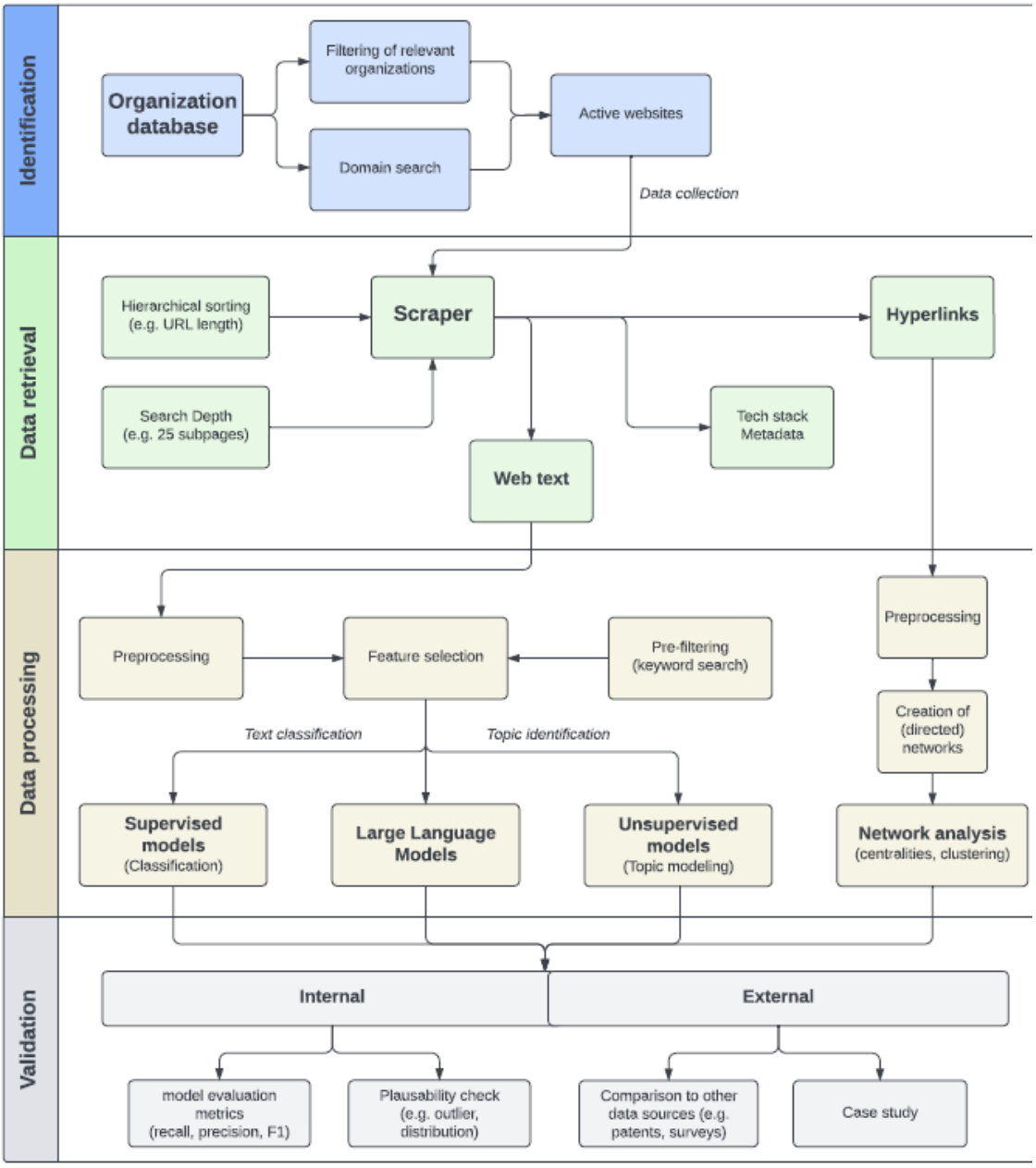
Da es sich bei dem Arbeitspapier um einen technischen Report handelt, wird anders als bei reinen wissenschaftlichen Abhandlungen, die den Fokus auf die Ergebnisdiskussion legen, insbesondere der methodisch-technische Ablauf stärker in den Fokus genommen. In den folgenden Abschnitten wird deshalb entlang der Prozessschritte, die in Abbildung 1 dargestellt sind, ein idealtypischer Ablauf einer Webdatenanalyse skizziert. Dies ist verbunden mit dem Ziel, Akteuren aus Politik und Wirtschaft eine nachvollziehbare Handreichung zu geben, eigene Analysen durchzuführen. Es wäre vermessen, ein vollständiges Handbuch der Webdatenanalyse anzubieten, da dafür auch die Innovationszyklen zu kurz sind und die Weiterentwicklungen zu dynamisch. Dennoch sollen die grundlegenden Definitionen und Erläuterungen der Algorithmen und Architekturen einen Einstieg in die Materie ermöglichen und Anregungen für den tieferen Einstieg geben. Im Anhang wird zudem eine kuratierte Liste mit genutzten und empfohlenen Paketen für die Programmiersprache Python zur umfassenden Bearbeitung der Pipeline vorgelegt, die als Ausgangspunkt für eine vertiefende Beschäftigung mit dem Thema dienen kann.

In Abbildung 1 sind die vier wesentlichen Schritte für den Aufbau und die Analyse von webbasierten Textdaten dargestellt. Nach der Identifikationsphase der zu analysierenden Organisationen und der Filterung aus kommerziellen oder selbst aufgebauten Datenbanken, steht eine Liste mit Domains aus der relevante und aktive Webseiten identifiziert werden. Daran schließt sich der eigentliche Datenbezug an (data retrieval). Dieser wird üblicherweise mittels eigener automatisierter Programme zur Sammlung von Webseitendaten (engl. Scraper) vorgenommen. In dieser Phase können sich unterschiedliche Strategien auszahlen, um effizient große und vor allem die richtigen Datenmengen zu beziehen (siehe unten). Am Ende dieser Phase stehen die Quelltexte der Webseiten, die im folgenden Schritt der Datenverarbeitung (data processing) für die eigentliche Textanalyse aufbereitet werden müssen. In dieser Phase stehen vor allem Maßnahmen zur Erhöhung der Textqualität im Fokus. Neben einer Säuberung und Bereinigung um nicht relevante Textbestandteile im engeren Sinne können hier auch weitere Filterschritte vorgenommen werden, die dazu beitragen, die zu verarbeitende Datenmenge weiter zu reduzieren. Dabei ist zu beachten, dass die Laufzeiten von rechenintensiven Modellierungen und Strukturierungen auf der Basis maschineller Lernverfahren bei großen Datenmengen selbst bei moderner Hardware und Parallelverarbeitung teilweise mehrere Tage oder gar Wochen andauern können. Es hat sich dabei gezeigt, dass es ein effizientes Vorgehen ist, zunächst mithilfe wenig rechenintensiven Verfahren den Datensatz schnell in seiner Größe zu reduzieren, um dann in folgenden Schritten mit rechenintensiven Verfahren eine höhere Verfahrensgeschwindigkeit bei großer Genauigkeit zu erreichen. Neben klassischen Auswahlverfahren wie regulären Ausdrücken oder vergleichenden Textoperationen mittels Wortlisten mit relevanten Schlüsselwörtern (Keywordlist), haben sich für solche Schritte insbesondere bei Produktivsystemen mittlerweile leistungsfähige Vektordatenbanken wie Milvus etabliert, die eine SQL-artige Abfrage auf große Textmengen ermöglichen. Diese Datenbanksysteme stellen jedoch hohe Anforderungen an die Hardware und sind für kleinere Projekte in der Umsetzung vergleichsweise komplex. Zur Datenverarbeitung gehören schließlich auch die eigentliche Modellierung und die darauf aufbauenden Verfahren, die z. B. Klassifikationsprobleme lösen oder Themenmodellierung ermöglichen. Eine Analysevariante, besonders mit Blick auf die für regionale Entwicklungsprozesse wichtige Vernetzung, nutzt Links aus den Webseiten, um diese zu einem Graphen zu verbinden und zu analysieren. Abschließend werden die Ergebnisse der Datenverarbeitung und -analyse validiert. Dies wird einerseits

systemintern durch explorative Verfahren aus der klassischen Inferenzstatistik (z. B. Verteilungsanalysen und Parametervergleiche) sowie über Kennzahlen, die sich aus dem maschinellen Lernprozess selbst ergeben, vorgenommen. Andererseits ist es wichtig, auch eine externe Validierung durchzuführen, also über Fallstudien und den Vergleich mit anderen, z. B. amtlichen Datenquellen die Ergebnisse zu plausibilisieren.

Im Folgenden wird entlang dieses schematischen Prozessablaufs zunächst ein genereller Überblick über die Terminologie und geeignete Verfahren gegeben, die dann im weiteren Verlauf der Ergebnispräsentation des Arbeitspakets 2 eingesetzt und vertieft diskutiert werden.

Abb. 1: Die WebAI-Pipeline: Von Unternehmenswebsites zu fundierten Erkenntnissen



Dahlke et al. 2025

2. Methodische Vorgehensweise

2.1. Datengrundlage, Extraktion und Verarbeitung von Webseitentexten

Im Projekt wurde auf eine Datenbank zurückgegriffen, die über mehrere Forschungsprojekte hinweg an der JLU in der Arbeitsgruppe Wirtschaftsgeographie entwickelt wurde. Die Datengrundlage für die hier verwendete Datenbank stammt aus dem Common Crawl. Common Crawl ist ein gemeinnütziges Projekt mit Unterstützung großer Technologiekonzerne, das regelmäßig große Teile des öffentlich zugänglichen Webs durchsucht (engl. crawling) und die gesammelten Daten frei zur Verfügung stellt. Ziel des Projekts ist es, Forschung, Entwicklung und Innovation im Bereich Webanalyse, Suchmaschinen, Sprachverarbeitung und Künstliche Intelligenz (KI) zu unterstützen.

Im Bereich der natürlichen Sprachverarbeitung (engl. natural language processing, NLP) fungiert Common Crawl als bedeutende Rohdatenquelle für Webkorpora und Trainingsdatensätze. Die Relevanz von Common Crawl für Large Language Modelle ergibt sich aus der enormen Skalierung, der sprachlichen Vielfalt und der Abdeckung realer Webkommunikation.

Die Daten werden seit 2013 gesammelt und monatlich in einem speziellen Archivformat WARC veröffentlicht. Das WARC-Datenformat enthält eine Reihe von Metadaten zum Such- und Speicherprozess (z. B. URI / URL, Zeitstempel) sowie die eigentlichen Webseitenquelltexte („payload“). Ergänzend gibt es extrahierte Metadaten und Textversionen, etwa in WAT- und WET-Dateien separat gespeichert, da die WARC-Dateien aufgrund ihrer Größe von bis zu 100 Terabyte (TB) pro Monat mit lokalen IT-Infrastrukturen nicht effizient durchsucht und analysiert werden können. Grundsätzlich sind die Daten komprimiert, in großen Crawl-Segmenten organisiert und über Cloud-Infrastrukturen wie Amazon S3 abrufbar[1] bzw. direkt über die Amazon Web Services in der Cloud zu verarbeiten. Einen Überblick über die bereitgestellten Dateiformate gibt Tabelle 1.

WARC – Web ARChive
Standardformat zur Archivierung von Webinhalten. Enthält HTTP-Responses, Header, HTML, Binärdaten und weitere Ressourcen. Wird häufig komprimiert gespeichert, z. B. als .warc.gz.
WAT – Web Archive Transformation
Enthält extrahierte Metadaten aus WARC-Dateien. Beispiele: Links, HTTP-Header, MIME-Typen, Crawl-Informationen. Nützlich für Webgraph-Analysen und Metadatenverarbeitung.
WET – WARC Encapsulated Text
Enthält extrahierten Klartext aus Webseiten. HTML-Struktur, Skripte und Formatierungen sind weitgehend entfernt. Besonders relevant für NLP und Sprachmodelltraining.
CDX / Indexformate
Indexdaten zur schnellen Suche nach URLs, Zeitpunkten, MIME-Typen oder Speicherorten. Erleichtern gezielten Zugriff auf einzelne Webseiten oder Domains.

Tab. 1: Dateiformate im CommonCrawl

[1] https://data.commoncrawl.org/crawl-data/index.html

Aus diesen Daten werden die Stammdaten aller deutschen Domains abgeleitet und in der Datenbank gespeichert. Es werden dabei nur diejenigen Domains berücksichtigt, die über ein Impressum nach §5 TMG verfügen und aus dem öffentlichen Internet ohne Zugangsbeschränkungen erreichbar sind.

Die Stammdatenbank enthält derzeit ca. 4,3 Mio. Domains und beinhaltet alle deutschen Domains, die in den Common Crawl Durchläufen seit 2018 erfasst wurden. Jeden Monat kommen zwischen 40.000 und 60.000 neue Domains hinzu, die über ein Impressum verfügen. Die Informationen aus dem Impressum werden mithilfe eines mehrstufigen Klassifikationsprozesses hinsichtlich relevanter Merkmale untersucht und strukturiert. Hierbei kommt eine Mischung aus regulären Ausdrücken (z. B. für die Extraktion von Postleitzahlen oder Umsatzsteuer-IDs) und maschinellen Lernverfahren zum Einsatz (z. B. für die variantenreicheren Bestandteile wie Straßennamen). Sofern eine vollständige Adressextraktion möglich war, werden die Adressbestandteile (Straßen, Ortsbezeichnungen) über OpenStreetMap-Daten für Deutschland validiert und mittels eines Fuzzy-Matching korrigiert (z. B. „Berleiner Platz“ zu „Berliner Platz“). Die korrigierte Adresse wird abschließend mittels einer lokalen Installation des Geocoders Nominatim einer Latitude und Longitude Koordinate zugeordnet. Zudem wird der amtliche Regionalschlüssel für Deutschland[2] gespeichert, um nachfolgende Analysen mit sekundärstatistischen Daten zu vereinfachen. Neben den Adressinformationen und den daraus abgeleiteten Attributen werden weitere Informationen gespeichert, die eine weitere Strukturierung der Daten, z. B. in Organisationsformen wie Unternehmen erleichtern. Hierzu zählen Angaben wie die Vereinsregisternummer oder die Handelsregisternummer.

Jede Domain der Stammdatenbank wird halbjährlich besucht, um relevante Unterseiten zu speichern und perspektivisch zeitliche Veränderungen analysieren zu können. Hierbei wird die Suche auf Unterseiten beschränkt, die analyserelevante Informationen enthalten. Dazu zählen Seiten die über aktuelle Nachrichten zur Organisation berichten (z. B. News, Aktuelles), die Organisationshistorie darstellen (z. B. über uns, Historie) oder über Projekte und Produkte berichten (z. B. unser Angebot). Hierfür werden die verlinkten URLs der Startseite sowie die zugehörigen menschenlesbaren Beschreibungen für die Links ausgewertet. Für jede Domain werden zwischen zehn und 20 Unterseiten im Quelltext gespeichert und bilden die *Contentdatenbank*, die textanalytisch und inhaltlich ausgewertet werden kann. Für das Projekt EnDiKaU wurde der Durchlauf des ersten Halbjahres 2025 für die Analyse verwendet.

An dieser Stelle ist es wichtig, Herausforderungen zu diskutieren, die die Stammdaten betreffen, um die daraus abgeleiteten Ergebnisse besser einordnen zu können. Webseitenbetreiber ergreifen zunehmend technische Maßnahmen, um einen automatisierten Abruf ihrer Webseiten zu unterbinden. Zugleich werden seitens großer Serverbetreiber und Dienstleister für die Internettechnologie Maßnahmen ergriffen, die die Belastung der Netze durch umfangreiches Crawling und Scraping verhindern sollen. Auch dies hat Auswirkungen auf die automatisierte Verarbeitung und den Abruf. Da die Anbieter großer KI-Modelle zunehmend die Netzinfrastruktur mit flächendeckenden Scraping-Läufen belasten, werden die Möglichkeiten für die Forschung zunehmend eingeschränkt, da es sehr aufwändig ist, trotz der Schutzmaßnahmen und unter Beachtung der guten Praxis im Internet, Analyseinhalte zu generieren. Dies gilt sowohl für die Datenbasis im Common Crawl als auch insbesondere für selbst durchgeführte Scrapings.

[2] https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/_inhalt.html

Damit entstehen potenziell systematische Verzerrungen in dem Unternehmensdatensatz. Große Unternehmen sind aufgrund ihrer Fähigkeiten ihre Server effektiver vor automatisierten Abrufen zu schützen, in der Datenbank leicht unterrepräsentiert. Dies hat für das Projekt EnDiKaU insofern kaum Relevanz, da das Projekt einen KMU Fokus hat, ist aber bei anderen Analysen zu berücksichtigen. Des Weiteren ist die Neigung eines Unternehmens, eine eigene Webpräsenz zu haben, abhängig von der Branche und der Unternehmensgröße. Auch das ist bei der Ergebnisinterpretation zu berücksichtigen. Regelmäßige Veröffentlichungen des Statistischen Bundesamtes[3], zuletzt von 2023 zeigen, dass Kleinstunternehmen (<10 Beschäftigte) generell unterrepräsentiert sind und insbesondere bei der Gruppe der Kleinstunternehmen der Energie- und Wasserversorgung, Entsorgung u.a. weniger als die Hälfte über eine eigene Webpräsenz verfügt (42%). Insgesamt zeigt sich in den Daten des Statistischen Bundesamtes jedoch, dass branchenunabhängig mindestens 90% der Unternehmen mit mindestens zehn Beschäftigen eine eigene Internetpräsenz betreiben.

Ein weiterer Aspekt der technischen Herausforderungen beim Bezug von Webseitentexten betrifft die Programmierung und die server- und clientenseitige Verarbeitung der Anfragen an einen Webserver. Die verwendeten Scraping-Frameworks sind standardmäßig nicht in der Lage, komplexe Javascript-basierte Webseiten abzurufen, die clientenseitig im Browser verarbeitet werden. Webseiten mit einem komplexen Aufbau sind insbesondere bei Großunternehmen und internationalen Konzernen üblich. Grundsätzlich dürfte jedoch ein Großteil der klein- und mittelständischen Unternehmen über das Vorgehen erfasst werden. Ein Abgleich der Zahl der Unternehmen in der Stammdatenbank mit der Zahl der Unternehmen auf Gemeindeebene nach Angaben des Statistischen Bundesamtes zeigt eine Korrelation von über 0,99.

2.2 Grundlagen der natürlichen Sprachverarbeitung (maschinelle Lernverfahren, Modellevaluation)

Für das bessere Verständnis der weiteren Analysen und Ergebnispräsentation in den nachfolgenden Kapiteln ist zunächst ein kurzer allgemeiner Einstieg in die Grundlagen der maschinellen Verarbeitung unstrukturierter Textdaten notwendig. Im Mittelpunkt steht der Einsatz sogenannter transformerbasierter Modelle zur Durchführung textbasierter Klassifikationsaufgaben. Die zugrundeliegende Algorithmik und Architektur der neuronalen Netze entsprechen dabei dem Prinzip, das auch aktuelle generative KI-Systeme, z. B. in Chatbots, einsetzen.

Die automatisierte Verarbeitung und Strukturierung von Texten wie im vorliegenden Projekt EnDiKaU in AP 2 beinhaltet oftmals Klassifikationsprobleme, die etwa in Form von Multiclass-Problemen (ein Textabschnitt wird genau einem von mehreren Attributen ausgezeichnet) oder Multilable-Problemen (ein Textabschnitt kann mehr als einem von mehreren Attributen zugeordnet werden) formuliert werden. Damit können vielfältige Aufgaben durchgeführt werden, um z. B. aus einer Beschreibung des Produktangebots eines Unternehmens die Zuordnung in einen bestimmten Wirtschaftszweig abzuleiten.

Der Verarbeitungsprozess umfasst dabei mehrere Schritte (Datenaufbereitung, Tokenisierung, Einbettung der Eingaben, Verarbeitung durch Transformer-Schichten sowie die anschließende Klassifikation und Evaluation), die das Ziel verfolgen, die

[3] https://genesis.destatis.de/datenbank/online/statistic/52911/table/52911-0004/table-toolbar

Ausgangstexte (Rohdaten) so vorzubereiten, dass ein möglichst gutes Klassifikationsergebnis erzielt werden kann.

Datenaufbereitung und Preprocessing

Zu Beginn werden unstrukturierte Textdaten in ein maschinenlesbares Format überführt. Zur Textbereinigung in Webdaten ist die Entfernung irrelevanter Bestandteile wie HTML-Tags, Sonderzeichen, mehrfacher Leerzeichen oder technischer Artefakte sinnvoll. Je nach Modell und Anwendungsfall kann eine Normalisierung der Groß- und Kleinschreibung erfolgen. Anschließend wird der Text in kleinere Einheiten (Tokens) zerlegt. Moderne Transformer-Modelle verwenden hierfür häufig Subword-Verfahren wie Byte-Pair Encoding, WordPiece oder SentencePiece. Diese Verfahren reduzieren das Problem unbekannter Wörter, sogenannter Out-of-Vocabulary-Tokens, und ermöglichen eine effiziente Abbildung morphologischer Strukturen. Die zu Token verarbeiteten Sequenzen werden um modellspezifische Steuerzeichen ergänzt. Bei den im Projekt EnDiKaU verwendeten BERT-basierten Modellen wird beispielsweise das Token [CLS] am Anfang einer Sequenz eingefügt, dessen finale Repräsentation häufig für Klassifikationsaufgaben genutzt wird. Das Token [SEP] dient der Trennung einzelner Sequenzen oder Satzpaare. Da Transformer-Modelle Eingaben fester oder begrenzter Länge verarbeiten, werden kürzere Sequenzen durch Auffüllen von Textdaten (Padding) ergänzt und längere Sequenzen gegebenenfalls gekürzt. Sogenannte Attention Masks markieren dabei, welche Tokens tatsächliche Eingabebestandteile sind und welche lediglich Padding darstellen. Die Schritte der Tokenerstellung laufen in den aktuellen Programmpaketen wie z. B. Python-Transformers / PyTorch weitgehend automatisiert ab.

Einbettung / Embedding

Nach der Tokenerstellung werden die Tokens in hochdimensionale Vektorräume projiziert. Dadurch entsteht eine numerische Repräsentation, die vom Modell verarbeitet werden kann. Jedes Token wird durch einen Vektor repräsentiert, der semantische und kontextuelle Informationen kodieren kann. Da Transformer-Architekturen im Gegensatz zu rekurrenten neuronalen Netzen keine inhärente Sequenzordnung besitzen, werden Positionsinformationen zusätzlich eingebracht. Dadurch kann das Modell die Reihenfolge der Tokens innerhalb einer Sequenz berücksichtigen. Bei Aufgaben mit mehreren Eingabesequenzen, etwa Satzpaar-Klassifikationen, können zusätzliche Segment- oder Token-Type-Embeddings verwendet werden, um die Zugehörigkeit einzelner Tokens zu unterschiedlichen Sequenzen zu kennzeichnen.

Diese Vorverarbeitungsschritte stellen sicher, dass der nachfolgende maschinelle Lernprozess effizient abläuft und sorgen bei inhaltlich hochwertigem Rohdatenmaterial für eine hohe Güte und Validität der Ergebnisse. Zum Einsatz in EnDiKaU kommen ausschließlich aktuelle Transformer-Architekturen, die im Folgenden in ihrer Funktionsweise grob dargestellt werden.

Modellarchitektur und Self-Attention-Mechanismus

Anders als klassische rekurrente neuronale Netze (wie LSTMs), die Text streng sequenziell verarbeiten, betrachten Transformermodelle die gesamte Eingabesequenz auf einmal. Ihre hohe Genauigkeit und exzellente Skalierbarkeit verdanken sie dem sogenannten Self-Attention-Mechanismus. Dieser bewertet die Relevanz jedes Wortes in Bezug auf alle anderen Wörter im Satz und versieht diese mit einer Gewichtung, wodurch das Modell kontextuelle Zusammenhänge hochgradig parallel erfasst. Um diese

Repräsentationen weiter zu verfeinern, kommt die Multi-Head Attention zum Einsatz: Durch die parallele Ausführung mehrerer Attention-Mechanismen kann das Modell gleichzeitig verschiedene linguistische Aspekte wie syntaktische Beziehungen oder semantische Ähnlichkeiten erlernen. Im Anschluss werden die so gewonnenen Repräsentationen durch Feed-Forward-Netzwerke geleitet. Diese vollvernetzten Schichten ermöglichen nichtlineare Transformationen und erhöhen die Ausdrucksstärke des Modells. Abgerundet wird die Architektur durch Residualverbindungen und Normalisierungsschichten, die den Informationsfluss durch das tiefe Netzwerk verbessern und den Trainingsprozess stabilisieren (vgl. Devlin et al., 2019; Sanh et al., 2019).

Es gibt eine Fülle vortrainierter transformerbasierter Sprachmodelle für nahezu jede Textdomäne (z. B. allgemeine Webtexte, Social-Media Texte, juristische Texte / Patentschriften), die als Grundlage für die nachfolgenden Klassifikationsaufgaben genutzt werden können. Um eine hohe Klassifikationsgüte zu erreichen, ist es entscheidend, dass die sprachlichen und strukturellen Merkmale der Zieldaten bereits im Pre-Training-Korpus des Modells repräsentiert sind. Dies ist im Projekt EnDiKaU der Fall, da Unternehmenswebseiten klassifiziert werden sollen und die meisten Sprachmodelle zu großen Teilen auf Common Crawl Daten, also Webdaten trainiert wurden.

Klassifikationsverfahren

Für die konkrete Klassifikationsaufgabe wird das vortrainierte Transformer-Modell um einen aufgabenspezifischen Klassifikationskopf (Classification Head) erweitert. Während das Modell für jedes Token eine kontextabhängige Repräsentation erzeugt, wird für die Klassifikation in der Regel die finale Repräsentation des [CLS]-Tokens extrahiert, da diese als aggregierte Zusammenfassung der gesamten Eingabesequenz dient. Dieser Vektor wird anschließend an den Klassifikations-Layer übergeben, der meist aus einer oder mehreren vollvernetzten Schichten besteht. Zur Erzeugung einer Wahrscheinlichkeitsverteilung über die Zielklassen kommt bei der Mehrklassenklassifikation eine Softmax-Funktion zum Einsatz, während bei binären oder Multi-Label-Problemen eine Sigmoid-Aktivierung verwendet wird.

Das eigentliche Training (Fine-Tuning) erfolgt durch die Minimierung einer geeigneten Verlustfunktion, wie etwa der Cross-Entropy Loss. Die Anpassung der Modellparameter geschieht dabei mittels Backpropagation. Der berechnete Vorhersagefehler wird dabei rückwärts durch das Netzwerk geleitet, um die Gewichte proportional zu ihrem Fehlerbeitrag zu aktualisieren. Bei diesem Prozess werden sowohl die Parameter des neuen Klassifikationskopfs als auch die des vortrainierten Basismodells auf die spezifische Zielaufgabe feinjustiert. Das synergetische Zusammenspiel all dieser Schritte sorgt für die in der Regel hervorragende Klassifikationsgüte von Transformer-Modellen.

Training und Evaluation

Zur robusten Bewertung der Modellleistung wird der Datensatz standardmäßig in Trainings-, Validierungs- und Testdaten unterteilt. Während das Trainingsset der eigentlichen Parameteroptimierung dient, wird das Validierungsset zur Hyperparameterabstimmung und Modellselektion herangezogen. Die finale Evaluation erfolgt ausschließlich auf dem Testset, also auf Daten, die das Modell zuvor nie gesehen hat. Dieses Vorgehen stellt sicher, dass eine eventuelle Überanpassung (Overfitting) an die Trainingsdaten zuverlässig aufgedeckt wird.

Die Wahl der geeigneten Evaluationsmetriken hängt maßgeblich von der spezifischen Klassifikationsaufgabe und der Verteilung der Klassen ab. Die Basis für die meisten Metriken bildet die Confusion Matrix (Wahrheitsmatrix). Diese tabellarische Darstellung vergleicht die tatsächlichen mit den vorhergesagten Klassen und unterteilt die Ergebnisse in True Positives (TP), True Negatives (TN), False Positives (FP) und False Negatives (FN). Sie ermöglicht eine detaillierte Fehleranalyse, da sofort ersichtlich wird, zu welchen Fehlertypen das Modell neigt. Bei einer Mehrklassenklassifikation wird die Matrix entsprechend der Klassenzahl erweitert.

Aus der Confusion Matrix lassen sich folgende zentrale Kennzahlen ableiten (vgl. Sammut & Webb, 2011):

Die Accuracy (Genauigkeit) gibt den Anteil aller korrekt klassifizierten Instanzen an der Gesamtzahl aller Instanzen an. Sie misst also, wie viele Vorhersagen insgesamt richtig waren. Eine Accuracy von 0,90 bedeutet, dass 90 % aller Fälle korrekt klassifiziert wurden. Bei unausgeglichenen Klassenverteilungen kann Accuracy irreführend sein. Wenn beispielsweise 95 % aller Fälle zur Klasse A gehören, kann ein Modell durch ständiges Vorhersagen von Klasse A bereits eine hohe Accuracy erreichen, ohne tatsächlich gut zu klassifizieren. Bei unausgeglichenen Klassenverteilungen ist insbesondere der F1-Score oder eine klassenweise Betrachtung von Precision und Recall aussagekräftiger als die reine Accuracy.

Die Precision (Genauigkeit der positiven Vorherrsagen) beschreibt, wie viele der als positiv vorhergesagten Fälle tatsächlich positiv sind. Eine Precision von 0,80 bedeutet, dass 80 % der positiven Vorhersagen tatsächlich korrekt sind. Precision ist besonders wichtig, wenn falsch-positive Vorhersagen problematisch sind, etwa wenn ein Modell eine sonstige Organisation als Unternehmen klassifiziert.

Der Recall (Sensitivität/Trefferquote) gibt an, welcher Anteil der tatsächlich positiven Fälle vom Modell korrekt gefunden wurde. Ein hoher Recall ist entscheidend, wenn das Übersehen von positiven Fällen kritisch ist (z. B. bei der medizinischen Diagnostik).

Der F1-Score ist als harmonisches Mittel aus Precision und Recall eine Zusammenfassung beider Metriken. Er eignet sich für unausgeglichene Klassenverteilungen, da ein hoher Wert nur erreicht wird, wenn das Modell sowohl bei FP als auch bei FN geringe Fehlerraten aufweist.

Für binäre Klassifikationsaufgaben wird ergänzend häufig die ROC-AUC herangezogen. Die ROC-Kurve (Receiver Operating Characteristic) visualisiert die Trennfähigkeit des Modells über verschiedene Schwellenwerte hinweg. Die AUC (Area Under the Curve) beziffert dabei die Fläche unter dieser Kurve. Ein Wert nahe 1,0 (z. B. 0,90) signalisiert eine sehr gute Fähigkeit des Modells, positive von negativen Fällen anhand seiner internen Wahrscheinlichkeitswerte zu unterscheiden.

2.3. Einordung der Unternehmen in Branchen (Modelltraining und Architektur, Evaluierung)

Basierend auf der Auswahl der Branchen für die Befragung zum Einsatz neuer Daten in Kommunikationszusammenhängen auf Unternehmensebene des ZEW erfolgt die Branchenabgrenzung der Informationswirtschaft und dem verarbeitenden Gewerbe nach der Klassifikation der Wirtschaftszweige (Ausgabe 2008). Ziel des Arbeitsschrittes ist die

automatisierte Einordnung/Klassifizierung der identifizierten hessischen Unternehmen in die Subbranchen der Informationswirtschaft und dem verarbeitenden Gewerbe auf Basis ihrer Webseitentexte, da bereits die Branchen einen Aufschluss über die Digitalaffinität der Unternehmen geben (Calvino et al., 2018). Hierfür wurden zwei Textklassifikationsmodelle auf Basis der SetFit-Architektur[4] entwickelt und trainiert, wobei die Klassifizierung auf Basis deutschsprachiger Branchentexte vorgenommen wird. Es wurde sich für ein zweistufiges Verfahren entscheiden: Zunächst erfolgt eine binäre Klassifikation, ob ein Unternehmenstext einer relevanten Branche zugeordnet werden kann. Nur positive Fälle werden im zweiten Schritt weiter den spezifischen Subbranchen zugewiesen. Diese sequenzielle Modellarchitektur wurde gewählt, da erste Versuche mit einer direkten Multilabel-Zuordnung in die Subbranchen ohne vorherige Filterung nach den relevanten Branchen zu einer zu geringen Accuracy zwischen 55% und 65% geführt haben. Des Weiteren war diese Form der Modellierung sehr komplex und daher entsprechend rechenintensiv. Durch die Vorfilterung wird die Komplexität reduziert und die Genauigkeit signifikant gesteigert.

Die Datengrundlage für das Training der Branchenklassifikatoren bildet ein manuell annotierter Datensatz aus deutschsprachigen Unternehmenswebseitentexten. Insgesamt umfasst der Datensatz 1.391 Textausschnitte, die auf Basis ihrer inhaltlichen Beschreibung einer Branche gemäß der Klassifikation der Wirtschaftszweige (WZ 2008) zugeordnet wurden. Die Texte stammen aus öffentlich zugänglichen Unternehmenswebseiten, insbesondere den Hauptseiten sowie „Über-uns"- und Unternehmensprofilseiten, da dort typischerweise die Kerntätigkeiten, Produkte, Dienstleistungen und Kompetenzen eines Unternehmens beschrieben werden. Dadurch enthalten die Trainingsdaten vor allem jene Informationen, die für die Identifikation der Branchenzugehörigkeit besonders relevant sind. Der Datensatz deckt sämtliche betrachteten Subbranchen der Informationswirtschaft und des verarbeitenden Gewerbes ab und umfasst zusätzlich eine Klasse „non-classified", welche Unternehmen außerhalb des Untersuchungsbereichs repräsentiert.

Die manuelle Annotation erfolgte auf Grundlage der in den Webseitentexten beschriebenen wirtschaftlichen Haupttätigkeit des Unternehmens. Durch die Verwendung realer Unternehmensbeschreibungen wird sichergestellt, dass die Trainingsdaten die sprachliche Vielfalt und die branchenspezifischen Formulierungen der späteren Anwendungsdaten möglichst realitätsnah abbilden. Gleichzeitig wird die Vergleichbarkeit zwischen Trainings- und Anwendungsdaten erhöht, da für beide Datensätze primär dieselben Seitentypen herangezogen werden.

Das erste Modell ist ein binäres Klassifizierungsmodell (SetFit, DistilBERT[5]), welches zunächst unterscheidet, ob der Webseitentext eines Unternehmens in eine der Subbranchen der Informationswirtschaft sowie dem verarbeitenden Gewerbe zugeordnet werden kann oder nicht. Somit werden zunächst alle relevanten Unternehmen der zu untersuchenden Branchen herausgefiltert. Methodisch wird ein SetFit-Ansatz angewendet, ein modernes Transfer-Learning-Verfahren das auf kontrastivem Lernen mit Sentence-Transformers basiert und insbesondere für effizientes Few-Shot-Learning, also dem maschinellen Lernen auf der Grundlage weniger ausgewählter Trainingsdaten, optimiert ist. Dieses Vorgehen ist deshalb geeignet, da kontrastives Lernen es ermöglicht, die semantische Struktur von Texten durch die Nutzung vortrainierter Transformer-

[4] https://huggingface.co/docs/setfit/index
[5] https://huggingface.co/docs/transformers/model_doc/distilbert

Modelle wie DistilBERT effizient zu modellieren und dabei eine hohe Performance erzielt (vgl. Gao et al., 2021a; Tunstall et al., 2022). Ziel ist es somit, im Sinne eines effizienten Workflows mit vergleichsweise wenig annotierten Daten robuste Klassifikationsentscheidungen zu ermöglichen. Dafür erfolgte die Aufteilung in Trainings-, Evaluations- und Testdaten im Verhältnis von 70 zu 15 zu 15 Prozent. Die Daten wurden anschließend so strukturiert, dass sie den Anforderungen des HuggingFace-Dataset-Formats entsprechen. Für die Modellarchitektur kommt SetFit zum Einsatz, ein auf Sentence-Embeddings basierender Klassifikationsansatz, der auf einem vortrainierten Transformer-Modell (hier: distilbert-base-german-cased) basiert. DistilBERT ist eine komprimierte Version von BERT, die durch geschickte Komprimierung ein kleineres Modell (40 % weniger Parameter) bei vergleichbarer Leistung (97 % der ursprünglichen Leistungsfähigkeit) bereitstellt und somit den Rechenaufwand reduziert (Sanh et al., 2019). Die Wahl dieses Modells begründet sich in seiner guten Verfügbarkeit für die deutsche Sprache und seiner Effizienz. Zusätzlich lag der Trainingsdatensatz mit 1.222 gelabelten Beispielen deutlich unter den Größenordnungen, die für klassische Transformer-Modelle wie BERT oder T5 notwendig wären.

Der Trainingsprozess gliedert sich in zwei Phasen: Zunächst erfolgt ein kontrastives Fine-Tuning, bei dem das Modell lernt, semantisch ähnliche Textpaare enger zusammenzuführen und unähnliche zu trennen. Dabei werden semantisch ähnliche und dissimilar gepaarte Textkombinationen generiert (insgesamt 449.192 Paarungen) und mit einem kontrastiven Loss (CoSENTLoss) optimiert. Die Wahl dieses Loss-Typs erfolgt automatisch im SetFit-Framework, wenn kein expliziter Wert angegeben wird, da dieser sich in vielen Anwendungsfällen als stabil und performant erwiesen hat. Im Anschluss wurde ein linearer Klassifikator auf den erzeugten Sentence-Embeddings trainiert, um die binäre Klassifikation durchzuführen und dabei zwischen relevanten und nicht relevanten Texten zu unterscheiden. Diese Zweiteilung erlaubt es, mit sehr wenigen gelabelten Beispielen eine zuverlässige Klassifikation zu erzielen, ohne komplexe Prompting-Techniken einsetzen zu müssen. Die Trainingskonfiguration umfasste zehn Epochen, eine Batchgröße von 32 sowie Evaluation und Modell-Speicherung nach jeder Epoche. Diese Parameter sind in der Literatur als geeignete Standardwerte bekannt, um eine gute Balance zwischen Trainingszeit, Ressourcenverbrauch und Vermeidung von Überanpassung zu gewährleisten (Tunstall et al., 2022). Darüber hinaus wurde das Modell so konfiguriert, dass nach Abschluss der Trainingsphase automatisch das Modell mit der besten Validierungsgenauigkeit geladen wurde. Die Evaluierung des finalen Modells erfolgte anhand eines separaten Testdatensatzes. Zur Bewertung der Qualität der Klassifikationsleistung wurden die etablierten Metriken Accuracy, Precision, Recall und F1-Score berechnet (Sammut & Webb, 2011). Da es sich bei einem SetFit-Modell um ein Low-Data-Szenario handelt, erfolgt die Evaluierung des finalen, nachtrainierten Modells anhand einer 5-fachen stratified Cross-Validation, bei der die Daten in fünf Folds aufgeteilt werden, wobei die Klassenverteilung in jedem Fold erhalten bleibt. Diese Vorgehensweise stellt sicher, dass auch seltene Klassen in allen Folds angemessen repräsentiert sind, zudem reduziert sie die Varianz, die durch einzelne zufällige Datenaufteilungen entstehen könnte (Gao et al., 2021b; Kohavi, 1995).

Ergebnisse

Der erste Trainingsdurchlauf erzielte eine Accuracy von 90,22 % (Precision = 89,66 %, Recall = 97,74 %, F1 = 93,53 %). Um die Leistungsfähigkeit weiter zu verbessern, wurde das Modell im Anschluss nachtrainiert, indem der Trainingsdatensatz erweitert wurde.

Für dieses Retraining wurde das bereits vortrainierte Modell erneut geladen und auf Basis der erweiterten Datenmenge weiter optimiert. Das finale, nachtrainierte Modell erreichte die nachfolgend dargestellten Evaluationsmetriken und weist damit eine sehr hohe Klassifikationsleistung auf, die eine zuverlässige Identifikation relevanter Branchenunternehmen gewährleistet:

Tab. 2: Durchschnittliche Metriken über 5 Folds

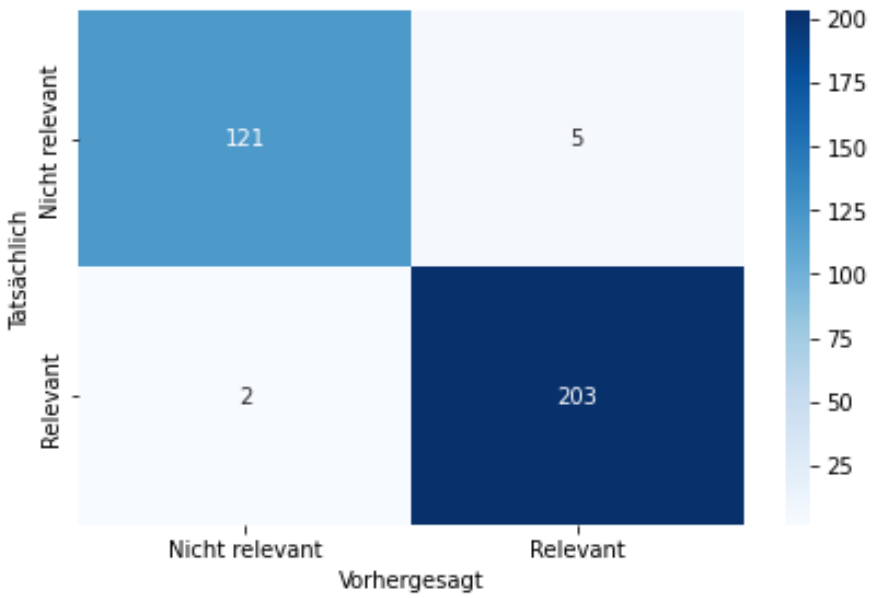
Metrik	**Mittelwert**	**Standardabweichung**
Accuracy	0.9777	0.0031
Precision	0.9752	0.0107
Recall	0.9893	0.0064
F1-Score	0.9821	0.0023

Die aggregierten Ergebnisse über die fünf Folds zeigen eine durchschnittliche Accuracy von 97,8 %, Precision von 97,5 %, Recall von 98,9 % und F1-Score von 98,2 %, wobei die geringe Standardabweichung (<1,1 %) auf eine sehr stabile Leistungsbewertung hinweist. Die hohe Sensitivität (Recall) des Modells ist besonders relevant, da Fehlklassifikationen zugunsten der Erkennung relevanter Branchentexte minimiert werden sollen. Gleichzeitig verdeutlicht der hohe F1-Score, dass eine gute Balance zwischen Sensitivität und Präzision erzielt wurde. Für den letzten Fold wurde exemplarisch ein Klassifikationsbericht erstellt, der nachfolgend die Vorhersagequalität im Detail zeigt. Zusätzlich legt die Konfusionsmatrix (Abb. 2) eine geringe Anzahl falsch negativer Klassifikationen dar, was eine konservative Modellstrategie unterstreicht.

Tab. 3: Klassifikationsbericht binäres Modell zur Branchenklassifikation (letzter Fold)

Klasse	**Precision**	**Recall**	**F1-Score**	**Support**
Nicht relevant	0.98	0.96	0.97	126
Relevant	0.98	0.99	0.98	205
Gesamt (Accuracy)			**0.98**	**331**
Macro Average	0.98	0.98	0.98	331
Weighted Avg	0.98	0.98	0.98	331

Abb. 2: Konfusionsmatrix Modell 1



Eigene Darstellung 2025

Aufbauend auf dem beschriebenen binären Klassifikationsmodell (SetFit, DistilBERT) folgt im nächsten Schritt die automatisierte Anwendung des Modells auf die Textdatenbestände hessischer Unternehmenswebseiten. Ziel dieses Arbeitsschritts ist die systematische Erkennung und Klassifikation jener Unternehmen, die auf Basis ihrer Webseitentexte den relevanten Branchen der Informationswirtschaft und des verarbeitenden Gewerbes zugeordnet werden können. Als Datenbasis werden zwei Arten von SQLite-Datenbanken herangezogen: (1) die Stammdatenbanken, welche grundlegende Unternehmensinformationen wie Domain, Standort und Klassifikationsmerkmale enthalten, sowie (2) die Content-Datenbanken, in denen die extrahierten Webseitentexte der jeweiligen Domains gespeichert sind. Beide Datenquellen sind in einer hierarchischen Verzeichnisstruktur abgelegt und werden im Skript iterativ verarbeitet. Zur Fokussierung auf den Untersuchungsraum Hessen werden ausschließlich jene Einträge berücksichtigt, deren Adresse in Hessen liegt.

Da die Textdaten auf Unternehmenswebseiten häufig heterogen und von unterschiedlicher inhaltlicher Relevanz sind, wird im Rahmen der Datenvorverarbeitung eine gezielte Filterung der URLs durchgeführt. Hierbei werden ausschließlich solche Seiten einbezogen, die typischerweise zentrale oder beschreibende Unternehmensinformationen enthalten, beispielsweise Hauptseiten (*index.html*, *home*) oder „Über-uns“-Seiten (*ueber-uns*, *unternehmen*, *about-us*). Diese Auswahl erfolgt über eine Heuristik zur Mustererkennung, die auf der Normalisierung und Analyse des URL-Pfads basiert. Die gezielte Fokussierung auf diese Seitentypen wird vorgenommen, da Angaben des Unternehmens zur eigenen Tätigkeit, Positionierung und Branchenzugehörigkeit vor allem auf Haupt- und „Über-uns“-Seiten zu finden sind. Auf diesen Seiten präsentieren Unternehmen in der Regel ihre Kernkompetenzen, Produkte und Dienstleistungen sowie eine Selbstdarstellung, die inhaltlich stark mit der Branchenidentität verknüpft ist. Durch diese Filterung wird sichergestellt, dass die Texte, die dem Klassifikationsmodell übergeben werden, tatsächlich Unternehmensinformationen enthalten und nicht etwa rechtliche Hinweise, Kontaktseiten

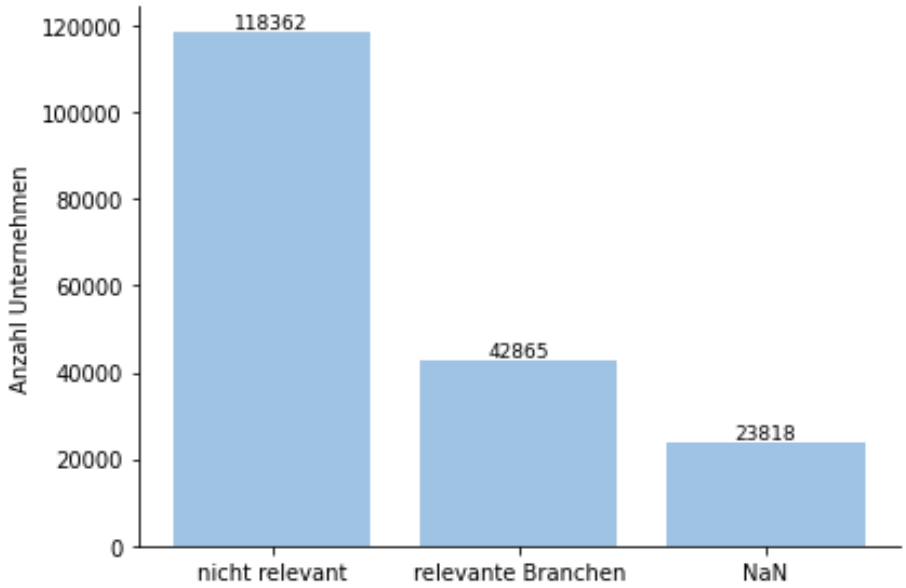
oder technische Inhalte. Obwohl modernste LLMs HTML-Quelltexte direkt verarbeiten können (Tan et al., 2025), ist eine sorgfältige Vorverarbeitung nach wie vor sehr wertvoll, insbesondere um die zu verarbeitende Textmenge zu reduzieren und damit die Recheneffizienz zu optimieren und die Rechenkosten zu senken. Encoder-basierte Modelle profitieren nach wie vor von strukturierten Vorverarbeitungsschritten, darunter Rauschunterdrückung, Entfernung von Boilerplate-Inhalten und Text-Chunking (Penedo et al., 2025; Wenzek et al., 2019). Dies liegt daran, dass diesen Modellen die umfangreichen kontextbezogenen Schlussfolgerungsfähigkeiten autoregressiver LLMs fehlen und sie für eine optimale Leistung stärker auf saubere, strukturierte Eingaben angewiesen sind (Devlin et al., 2019; Raffel et al., 2020). Eine effektive Vorverarbeitung zielt darauf ab, Rauschen zu reduzieren und die Daten zu strukturieren, ohne wichtige Informationen zu verlieren, d. h. die Menge an Daten zu reduzieren, die keine relevanten Informationen enthalten (z. B. Cookie- oder Datenschutztexte). Daher werden die üblichen Text-Vorverarbeitungsschritte ausgeführt. Die Texte der gefilterten URLs werden hinsichtlich HTML-Tags, E-Mail-Adressen, URLs und Telefonnummern bereinigt, um Störinformationen auszuschließen. Danach wird der Text normalisiert, tokenisiert und mithilfe der NLTK-Bibliothek um deutsche Stopwörter bereinigt. Um die inhaltliche Qualität sicherzustellen, werden nur Texte berücksichtigt, die eine Mindestanzahl an Wörtern aufweisen und deren durchschnittliche Wortlänge einen bestimmten Schwellenwert überschreitet. Auf diese Weise werden kurze oder semantisch nicht aussagekräftige Texte ausgeschlossen. Zusätzlich wird im Rahmen der Vorverarbeitung die Textlänge der Webseitentexte auf 150 Wörter reduziert. Diese Längenbegrenzung orientiert sich an der Struktur der Trainingsdaten, um eine hinreichende Übereinstimmung zwischen Trainings- und Anwendungsdaten zu gewährleisten und einen möglichen Distribution Shift zu reduzieren. Eine starke Abweichung hinsichtlich des Umfangs oder der Struktur der Eingaben kann die interne Repräsentationsbildung verändern und somit die Modellleistung beeinträchtigen. Zusätzlich reduziert dieses Vorgehen Rauschen und begrenzt den Einfluss einzelner überlanger Texte oder redundanter Textpassagen. Encoder-basierte Transformer-Modelle wie DistilBERT verarbeiten Eingaben in tokenisierter Form mit begrenzter Sequenzlänge; eine kontrollierte Textlänge trägt daher zur Stabilität der Repräsentationen bei (Devlin et al., 2019; Sanh et al., 2019). Durch diese Normalisierung wird die Varianz der Eingabedaten reduziert, ohne zentrale inhaltliche Informationen systematisch auszuschließen.

Nach Abschluss der Text-Vorverarbeitung wird das bereits beschriebene trainierte SetFit-Modell auf die bereinigten Texte angewendet. Für jede Unternehmensdomain werden die zugehörigen Textsegmente klassifiziert, wobei das Modell eine Wahrscheinlichkeitsverteilung über die beiden Klassen „relevant“ und „nicht relevant“ ausgibt. Das Ergebnis wird als Label (0 = nicht relevant, 1 = relevant) mit einer zugehörigen Konfidenz gespeichert. Da eine Domain mehrere Seiten enthalten kann, wird auf Domänenebene aggregiert: Eine Domain gilt als relevant, wenn mindestens eine ihrer Seiten eine positive Klassifikation erhält. Diese Aggregation reduziert die Wahrscheinlichkeit von Fehlklassifikationen durch zufällige Textfragmente und verbessert die Robustheit der Klassifikation. Die Klassifikationsergebnisse wurden im Anschluss mit den Unternehmensstammdaten zusammengeführt.

Abbildung 3 veranschaulicht die aggregierte Verteilung der Klassifikationslabels über alle untersuchten Domains hinweg. Von insgesamt rund 185.000 erfassten Unternehmensdomains wurden 42.865 (≈ 23,2 %) als relevant, d. h. den Branchen der

Informationswirtschaft oder des verarbeitenden Gewerbes zugehörig, klassifiziert. Ein Großteil der Unternehmen (118.362 bzw. ≈ 63,9 %) wurde dagegen als nicht relevant im Sinne des Forschungsprojekts eingestuft. Darüber hinaus konnten 23.818 Einträge (≈ 12,9 %) keiner Kategorie zugeordnet werden, was auf fehlende oder unvollständige Textinhalte in der zugrundeliegenden Datenbank zurückzuführen ist. Die Webseitentexte dieser 42.865 Unternehmen gelten sowohl als Grundlage zur weiterführenden detaillierten Klassifizierung in Branchen als auch zur Einordung im Hinblick auf Digitalisierung.

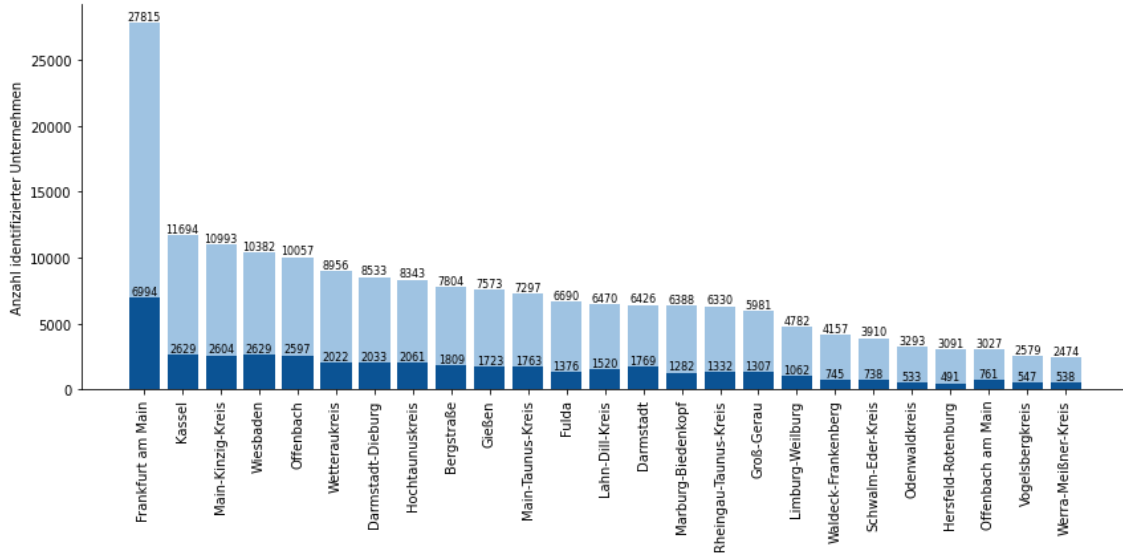
Abb. 3: Verteilung Unternehmen in Informationswirtschaft und verarbeitendem Gewerbe



Eigene Darstellung 2025

In Abbildung 4 wird die Verteilung aller identifizierten Unternehmen durch deren Unternehmenswebseiten (hellblaue Balken) und den Unternehmen in Informationswirtschaft und verarbeitendem Gewerbe (dunkelblaue Balken) nach Landkreisen dargestellt. Mit Abstand an der Spitze liegt Frankfurt am Main, wo 27.815 Unternehmenswebseiten erfasst wurden, darunter 6.994 relevante Fälle. Dies entspricht einem Anteil von rund 25,1% branchenrelevanter Unternehmen. Damit entfällt etwa 16 % aller hessischer Unternehmen in Informationswirtschaft und verarbeitendem Gewerbe auf den Frankfurter Wirtschaftsraum. Es folgen Kassel (11.694 identifizierte Unternehmen, 2.629 relevant, 22,5 %) sowie die Regionen Main-Kinzig-Kreis (10.993/2.604, 23,7 %) und Wiesbaden (10.382/2.597, 25,0 %). In einem Großteil der Kreise liegt der Anteil an Unternehmen in Informationswirtschaft und verarbeitendem Gewerbe stabil im Bereich von rund 20 – 25%. Nur in den ländlich geprägten und strukturschwachen Regionen wie dem Odenwaldkreis (3.091/533, 17,2 %), dem Landkreis Waldeck-Frankenberg (4157/745, 17,9 %) oder dem Schwalm-Eder-Kreis (3.910/738, 18,9 %) fällt der Anteil an Unternehmen in Informationswirtschaft und verarbeitendem Gewerbe geringer aus. Diese Werte verdeutlichen die bestehenden strukturellen Unterschiede zwischen urbanen und ländlichen Räumen in Bezug auf die Branchendichte.

Abb. 4: Unternehmen in Informationswirtschaft und verarbeitendem Gewerbe - Verteilung nach Kreisen



Eigene Darstellung 2025

Aufbauend auf der binären Vorselektion der branchenrelevanten Unternehmen erfolgt im nächsten Arbeitsschritt die feingranulare Zuordnung der identifizierten Unternehmen zu spezifischen Subbranchen der Informationswirtschaft und des verarbeitenden Gewerbes gemäß der Klassifikation der Wirtschaftszweige (WZ 2008). Methodisch wird für den zweiten Schritt der Einordung der Unternehmen in Branchen ein Multiclass-Modell mit einer SetFit Modellarchitektur und dem vortrainierten DistilBERT-Basismodell (distilbert-base-german-cased) verwendet. Die Datengrundlage für die Multiklassenklassifikation bildet ein manuell annotierter Trainingsdatensatz aus Unternehmenswebseitentexten, der gemäß der Klassifikation der Wirtschaftszweige (WZ 2008) auf Subbranchenebene gelabelt wurde. Um ein eindeutiges Lernsignal sicherzustellen, wurden ausschließlich Texte berücksichtigt, die genau einer Subbranche zugeordnet werden konnten. Datensätze mit Mehrfachlabels sowie Einträge, die zum Training des ersten Modells zur Unterscheidung in relevante vs. nicht relevante Branchen genutzt wurden, wurden im Rahmen der Datenaufbereitung entfernt. Dieses Vorgehen entspricht etablierten Empfehlungen im supervised machine learning, da mehrdeutige Labels die Modellkonvergenz und Klassifikationsgenauigkeit insbesondere bei Multiklassenproblemen negativ beeinflussen können (Goodfellow et al., 2016; Song et al., 2022). Die verbleibenden Subbranchenlabels wurden anschließend mittels Label-Encoding in numerische Klassen überführt. Die Aufteilung der Daten erfolgte stratifiziert in Trainings-, Evaluations- und Testdaten, um die Klassenverteilung über alle Teilmengen hinweg konsistent zu halten. Der Einteilung wurde im Verhältnis von 60 % Trainingsdaten, 20 % Evaluationsdaten und 20 % Testdaten vorgenommen.

Aufgrund der heterogenen Klassenverteilung auf Subbranchenebene wurde zusätzlich ein gezieltes Oversampling der Minderheitsklassen im Trainingsdatensatz durchgeführt. Dabei wurden Subbranchen mit geringer Fallzahl mittels zufälligem Resampling mit Zurücklegen so erweitert, dass jede Klasse mit mindestens 30 Trainingsbeispielen vertreten ist. Dieses Vorgehen reduziert die Verzerrung zugunsten häufiger Klassen und verbessert die Stabilität der Klassifikationsentscheidung insbesondere für seltene

Subbranchen (He & Garcia, 2009). Die Evaluations- und Testdaten blieben von diesem Schritt unberührt, um eine unverzerrte Evaluation sicherzustellen. Methodisch ist dieses Vorgehen insbesondere im Kontext von SetFit und kontrastivem Lernen sinnvoll, da eine zu geringe Repräsentation einzelner Klassen im Trainingsdatensatz zu instabilen Embedding-Strukturen und damit zu unsicheren Entscheidungsgrenzen führen kann. Durch das Oversampling wird sichergestellt, dass auch seltene Subbranchen im kontrastiven Fine-Tuning regelmäßig in positiven und negativen Textpaarungen berücksichtigt werden, was die Trennschärfe im semantischen Vektorraum erhöht. Zusätzlich wurde dem Risiko einer Überanpassung an spezifische sprachliche Muster einzelner Texte (Overfitting) bei sehr kleinen Klassen durch die Begrenzung auf maximal 30 Trainingsbeispiele pro unterrepräsentierter Subbranche entgegengewirkt. Auf diese Weise wird eine Balance zwischen ausreichender Klassenrepräsentation und der Vermeidung künstlich aufgeblähter, inhaltlich redundanter Trainingsdaten angestrebt (Chawla et al., 2002).

Das Trainingsverfahren folgt analog zum binären Modell einem zweistufigen SetFit-Ansatz. Zunächst findet ein kontrastives Fine-Tuning des Sentence-Transformer-Modells statt, bei dem semantisch ähnliche Textpaare näher im Embedding-Raum positioniert und dissimilar gepaarte Texte voneinander separiert werden. Anschließend wird ein linearer Klassifikator auf den erzeugten Sentence-Embeddings trainiert, um die Zuordnung zu den einzelnen Subbranchen vorzunehmen. Diese Architektur erlaubt es, auch bei Multiklassenproblemen mit vergleichsweise wenigen gelabelten Beispielen zuverlässige Klassifikationsentscheidungen zu treffen, ohne auf komplexe Prompting- oder Generative-Ansätze zurückgreifen zu müssen (Tunstall et al., 2022). Die Trainingskonfiguration umfasste sechs Epochen mit einer Batchgröße von 64. Nach jeder Epoche wurde eine Evaluation auf dem Validierungsdatensatz durchgeführt. Um Overfitting zu vermeiden und die Generalisierbarkeit des Modells zu maximieren, wurde automatisiert das Modell mit der besten Validierungsgenauigkeit gespeichert und nach Abschluss des Trainings als finales Modell geladen. Dieses Vorgehen entspricht etablierten Verfahren im Deep-Learning-basierten Textklassifikationssetting (Goodfellow et al., 2016).

Die Evaluierung des finalen Modells erfolgte zunächst auf einem separaten Testdatensatz mittels einer 5-fach stratifizierten Cross-Validation, der während des Trainingsprozesses nicht verwendet wurde. Die aggregierten Ergebnisse über alle fünf Folds zeigen eine durchschnittliche Accuracy von 92,4 %, eine Precision von 92,7 %, einen Recall von 92,4 % sowie einen F1-Score von 92,3 %, bei einer Standardabweichung von rund 2,6 %. Die vergleichsweise geringe Varianz der Ergebnisse über alle Folds hinweg weist auf eine insgesamt stabile und konsistente Modellleistung sowie eine gute Generalisierbarkeit der Subbranchenklassifikation hin (Tabelle 4).

Tab. 4: Durchschnittliche Metriken über 5 Folds

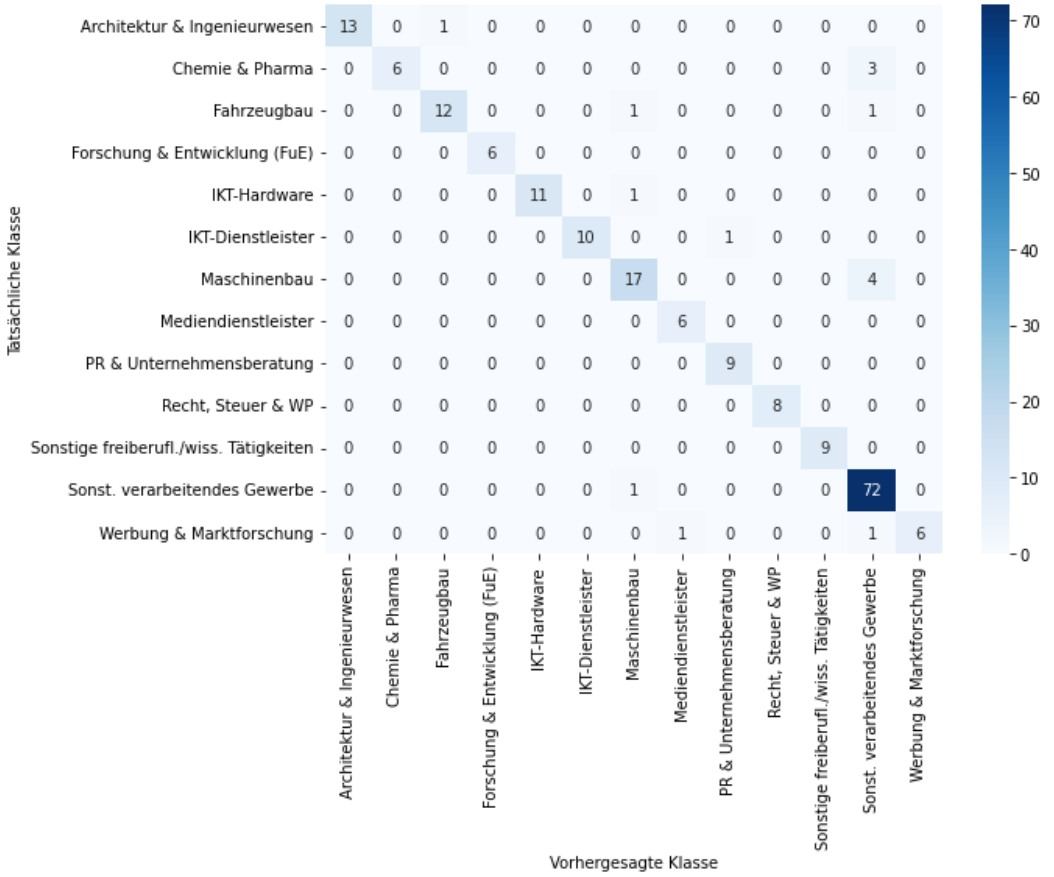
Metrik	Mittelwert	Standardabweichung
Accuracy	0.9243	0.0263
Precision	0.9267	0.0257
Recall	0.9243	0.0263
F1-Score	0.9231	0.0263

Der nachfolgende, exemplarisch für den letzten Fold ausgewiesene Klassifikationsbericht zeigt zudem, dass Subbranchen mit vergleichsweise homogener sprachlicher Ausprägung und klar abgrenzbaren Tätigkeitsprofilen, beispielsweise Forschung und Entwicklung (FuE), Rechts-, Steuerberatung und Wirtschaftsprüfung oder Architektur und Ingenieurwesen mit sehr hohen Precision- und Recall-Werten von nahezu 1,00 klassifiziert werden. Dies deutet darauf hin, dass diese Branchen auf Unternehmenswebseiten über konsistente terminologische Muster verfügen, die vom Modell zuverlässig erfasst werden können. Demgegenüber weisen einzelne thematisch breiter gefasste Subbranchen, wie etwa Maschinenbau, Chemie und Pharma, oder Werbung und Marktforschung geringere Recall-Werte auf. Dies bedeutet, dass ein Teil der tatsächlich zu dieser Subbranche gehörenden Texte vom Modell anderen, inhaltlich benachbarten Klassen zugeordnet wird. Inhaltlich lässt sich dies dadurch erklären, dass Unternehmenswebseiten in diesen Bereichen häufig technologieübergreifende oder anwendungsorientierte Beschreibungen enthalten, die semantische Überschneidungen mit angrenzenden Subbranchen aufweisen (z. B. Fahrzeugbau, sonstiges verarbeitendes Gewerbe, Mediendienstleister). Gleichzeitig bleiben die Precision-Werte in diesen Klassen überwiegend hoch, was darauf hindeutet, dass die vom Modell getroffenen positiven Klassifikationen in der Regel korrekt sind. Insgesamt bestätigt die Evaluation jedoch, dass das Modell eine zuverlässige und differenzierte Subbranchenzuordnung auf Basis von Unternehmenswebseitentexten ermöglicht und damit eine tragfähige Grundlage für weiterführende sektorale Analysen darstellt.

Tab. 5: Klassifikationsbericht Multiclass-Modell zur Branchenklassifikation (letzter Fold)

Klasse	**Precision**	**Recall**	**F1-Score**	**Support**
Architektur & Ingenieurwesen	1.00	0.93	0.96	14
Chemie & Pharma	1.00	0.67	0.80	9
Fahrzeugbau	0.92	0.86	0.89	14
Forschung & Entwicklung (FuE)	1.00	1.00	1.00	6
IKT-Hardware	1.00	0.92	0.96	12
IKT-Dienstleister	1.00	0.91	0.95	11
Maschinenbau	0.85	0.81	0.83	21
Mediendienstleister	0.86	1.00	0.92	6
PR & Unternehmensberatung	0.90	1.00	0.95	9
Recht, Steuer & WP	1.00	1.00	1.00	8
Sonstige freiberufl./wiss. Tätigkeiten	1.00	1.00	1.00	9
Sonst. verarbeitendes Gewerbe	0.89	0.99	0.94	73
Werbung & Marktforschung	1.00	0.75	0.86	8
Gesamt (Accuracy)			**0.93**	**200**
Macro Average	0.96	0.91	0.93	200
Weighted Avg	0.93	0.93	0.92	200

Abb. 5: Konfusionsmatrix Modell 2



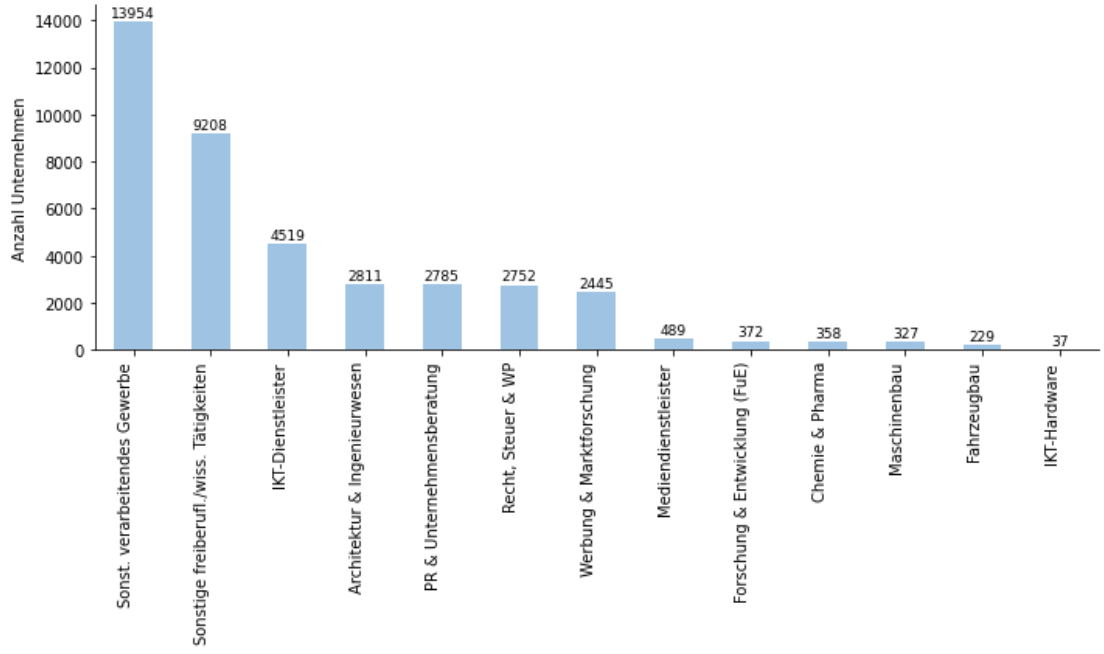
Eigene Darstellung 2025

Im Anschluss an das beschriebene Trainingsverfahren wird das trainierte Multiclass-Modell auf die zuvor als relevant identifizierten Unternehmenswebseitentexte angewendet, um eine differenzierte Einordnung über die Branchenstruktur der Informationswirtschaft und des verarbeitenden Gewerbes in Hessen zu erhalten. Dabei erzeugt das Modell für jeden Text eine Wahrscheinlichkeitsverteilung über alle definierten Subbranchen. Die finale Klassenzuweisung erfolgt über das Maximum der jeweiligen Klassenwahrscheinlichkeiten (argmax-Entscheidungsregel). Dieses Entscheidungsprinzip entspricht dem Maximum-a-posteriori-Kriterium und stellt im Multiclass-Setting ein etabliertes Standardverfahren dar. Zusätzlich wird für jede Vorhersage die höchste Klassenwahrscheinlichkeit als Konfidenzmaß gespeichert. Um Unsicherheiten explizit zu berücksichtigen, wird ein Schwellenwert von 0,45 definiert: Liegt die maximale Klassenwahrscheinlichkeit unterhalb dieses Wertes, wird die Zuordnung als „unsicher“ gekennzeichnet. Diese konservative Entscheidungsregel dient der Erhöhung der inhaltlichen Validität der resultierenden Branchenzuweisungen und reduziert potenzielle Fehlklassifikationen in semantisch ambivalenten Fällen. Empirisch betrifft dies jedoch lediglich rund 20 Unternehmen von insgesamt etwa 42.000 klassifizierten Fällen und ist damit quantitativ vernachlässigbar. Die sehr geringe Zahl unsicherer Zuordnungen spricht für eine insgesamt hohe Trennschärfe des Modells im Anwendungsdatensatz. Da einzelne Domains mehrere URLs beziehungsweise Textsegmente enthalten können, erfolgt im nächsten Schritt eine Aggregation auf Unternehmensebene. Hierbei wird pro Domain jene Klassifikation ausgewählt, die die

höchste Modellkonfidenz aufweist. Dieses Vorgehen basiert auf der Annahme, dass der Webseitentext mit der stärksten inhaltlichen Branchenprägung zugleich die eindeutigste semantische Repräsentation der unternehmerischen Tätigkeit darstellt, sodass die robusteste verfügbare Textrepräsentation als Entscheidungsgrundlage dient. Dennoch können Fehlklassifikationen nicht vollständig ausgeschlossen werden. Dies gilt insbesondere für Branchen mit ähnlichen Tätigkeitsprofilen oder überlappenden Begrifflichkeiten, wie beispielsweise Maschinenbau, Fahrzeugbau und sonstiges verarbeitendes Gewerbe. Solche Fehler wirken sich zwar aufgrund der hohen Gesamtfallzahlen vermutlich nur begrenzt auf die aggregierten Ergebnisse aus, können jedoch lokale oder branchenspezifische Analysen beeinflussen.

Abbildung 6 zeigt die aggregierte Verteilung der Subbranchen der Informationswirtschaft und des verarbeitenden Gewerbes aller untersuchten hessischen Unternehmen. Diese Verteilung reflektiert zum einen die Wirtschaftsstruktur Hessens mit einem hohen Anteil dienstleistungsorientierter und wissensintensiver Unternehmen, zum anderen aber auch potenzielle Unterschiede in der digitalen Sichtbarkeit und Webpräsenz verschiedener Branchen.

Abb. 6: Branchenverteilung der hessischen Unternehmen



Eigene Darstellung 2025

Zur Analyse regionaler Spezialisierungsmuster wird die zuvor ermittelte Branchenzuordnung auf Kreisebene aggregiert und für jede Gebietseinheit der relative Anteil der jeweiligen Subbranche an allen klassifizierten Unternehmen berechnet. Die Interpretation erfolgt bewusst auf Basis relativer Anteile, da sich die absoluten Unternehmenszahlen zwischen den Kreisen erheblich unterscheiden. Zugleich ist zu beachten, dass die Skalierung der Facet-Maps je nach Branche variiert, da einzelne Subbranchen stark unterschiedliche Gesamtfallzahlen aufweisen (vgl. Abb. 6). Während breit definierte Branchen wie das sonstige verarbeitende Gewerbe hohe absolute Werte erreichen, sind kapitalintensive oder hochspezialisierte Bereiche wie IKT-Hardware oder Fahrzeugbau deutlich seltener vertreten. Die relative Betrachtung erlaubt somit eine

vergleichbare Analyse struktureller Spezialisierungen, unabhängig von der Größenordnung der jeweiligen Branche. Die hessische Wirtschaftsstruktur ist insgesamt stark durch kleine und mittlere Unternehmen (KMU) geprägt. Da das zugrundeliegende Webscraping primär KMU im Bereich der Informationswirtschaft und des verarbeitenden Gewerbes erfasst, bildet die vorliegende Datengrundlage insbesondere die mittelständisch geprägte Branchenlandschaft ab. Die beobachteten regionalen Muster entsprechen in weiten Teilen bekannten industrie- und regionalökonomischen Strukturen Hessens.

Ein deutliches Spezialisierungsmuster zeigt sich im Landkreis Fulda im Bereich des Fahrzeugbaus. Die Region weist traditionell eine starke industrielle Basis auf, wobei konventionelle Technologien wie der Verbrennungsmotor nach regionalökonomischen Einschätzungen direkt oder indirekt etwa jeden siebten Arbeitsplatz beeinflussen. Die erhöhte relative Präsenz von Unternehmen des Fahrzeugbaus in der Klassifikation spiegelt diese historisch gewachsene Clusterstruktur wider. Netzwerke und regionale Wertschöpfungsketten tragen dazu bei, dass sich auch kleinere Zulieferbetriebe in der Region ansiedeln und die industrielle Spezialisierung verstärken (Region Fulda GmbH 2026).

Im Bereich Chemie und Pharmazie zeigt insbesondere der Landkreis Bergstraße eine erhöhte relative Konzentration. Die Wirtschaftsregion Bergstraße gilt als bedeutender Standort für Chemie, Pharma und Biotechnologie, geprägt von mittelständischen und familiengeführten Unternehmen sowie spezialisierten Nischenanbietern. Die räumliche Nähe zu Mannheim und Ludwigshafen – Standorte großer Chemiekonzerne – sowie die Einbettung in bestehende Wertschöpfungsnetzwerke begünstigen diese Spezialisierung. Auch die Rhein-Main-Region weist eine starke chemisch-pharmazeutische Prägung auf, wenngleich dort vor allem größere Unternehmen dominieren, die in der vorliegenden KMU-basierten Analyse nur eingeschränkt erfasst werden. Gleichwohl existiert insbesondere rund um den Industriepark Höchst ein dichtes Zulieferernetzwerk mittelständischer Unternehmen. Insgesamt bleibt der Anteil kleiner und mittlerer Unternehmen der Chemie- und Pharmaindustrie an der lokalen Unternehmenszahl zwar vergleichsweise gering (unter 2 %), der ökonomische Einfluss der Branche in Hessen gemessen an Umsatz und Wertschöpfung ist jedoch überproportional hoch. (Wirtschaftsförderung Bergstraße GmbH 2026a, b)

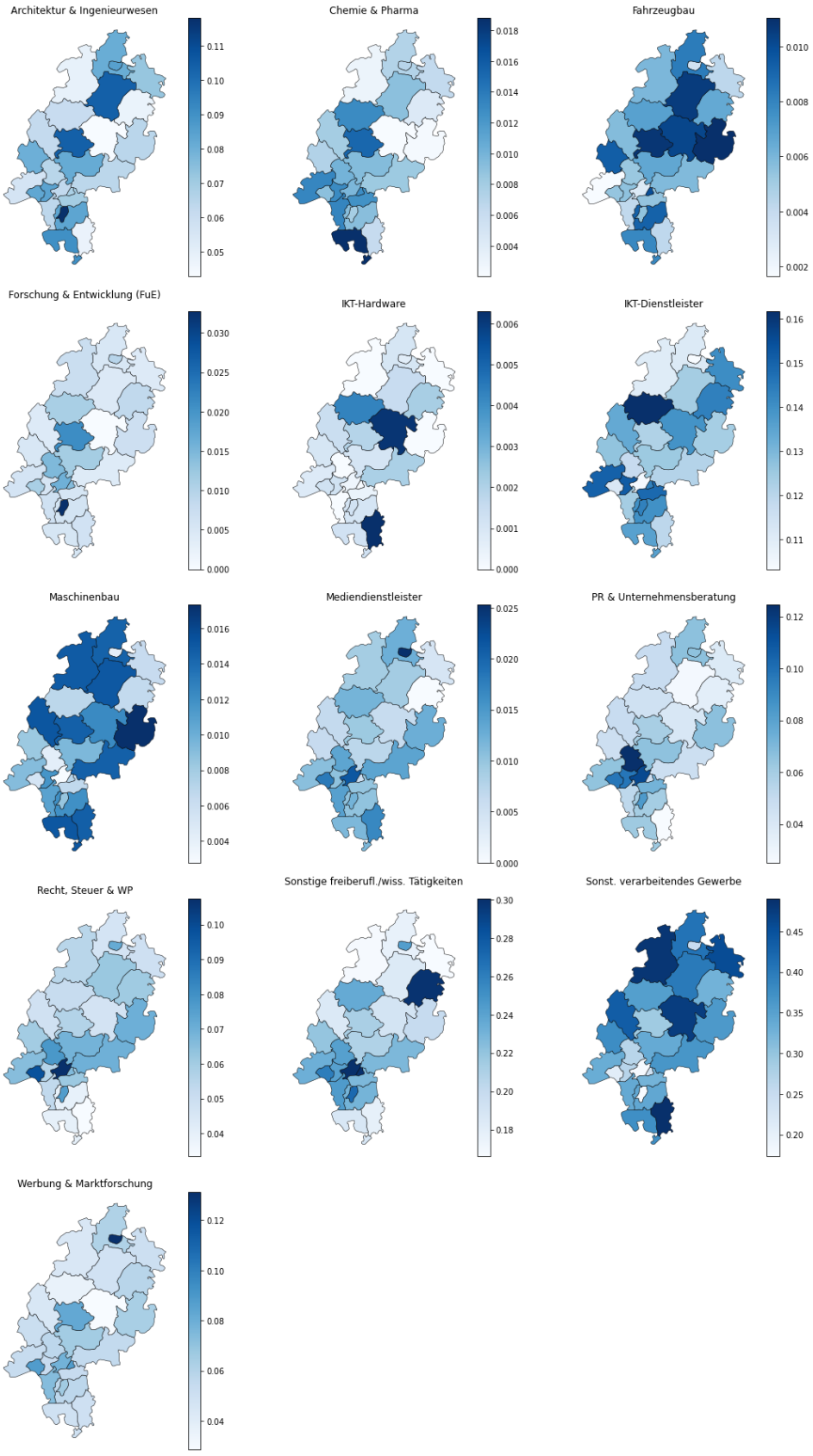
Eine weitere klare regionale Spezialisierung zeigt sich im Maschinenbau in Nordhessen, insbesondere im Raum Kassel. Die Konzentration wird durch regionale branchenspezifische Netzwerke wie dem „Die Maschinenbau Partner e. V." gestützt, die Kooperationen fördern und technologische Kompetenzen bündeln. Der Maschinenbau in Nordhessen ist häufig mittelständisch geprägt und auf Sondermaschinenbau sowie spezialisierte Produktionslösungen ausgerichtet. Die im Datensatz identifizierte relative Häufung entsprechender Unternehmen unterstreicht diese industrielle Schwerpunktbildung. Das sonstige verarbeitende Gewerbe stellt eine vergleichsweise sehr breit gefasste Kategorie dar, die von Holz- und Metallverarbeitung bis hin zur Herstellung von Textilien oder Spielwaren reicht. Entsprechend verteilt sich diese Branche flächendeckend über das Land, weist jedoch in Nordhessen eine erhöhte relative Bedeutung auf. Regionale Initiativen zur Bündelung metallverarbeitender Kompetenzen sowie die starke handwerklich-industrielle Tradition tragen zu dieser Ausprägung bei. Ein Beispiel dafür ist das im Kreis Waldeck-Frankenberg ansässige „MetallverarbeitungsCluster Waldeck-Frankenberg". Die breite Definition der Branche

erklärt zugleich ihre hohen absoluten Fallzahlen und die vergleichsweise geringe Spezialisierung einzelner Kreise im relativen Vergleich (HMWVW 2019).

Recht, Steuer- und Wirtschaftsprüfung sowie PR- und Unternehmensberatung konzentrieren sich erwartungsgemäß stark in der Region Frankfurt-Rhein-Main. Neben international tätigen Großkanzleien und Beratungshäusern ist hier eine Vielzahl kleiner und mittlerer Kanzleien sowie spezialisierter Beratungsunternehmen angesiedelt. Initiativen wie regionale Beratungsnetzwerke (ConsultingRegion FrankfurtRheinMain) fördern die Sichtbarkeit kleiner Anbieter und stärken die Vernetzung innerhalb der Branche. Die erhöhte relative Präsenz dieser wissensintensiven Dienstleistungen in Frankfurt, Wiesbaden und angrenzenden Kreisen reflektiert die Funktion des Rhein-Main-Gebiets als Finanz-, Dienstleistungs- und Verwaltungszentrum mit hoher Nachfrage nach spezialisierten Beratungsleistungen. In den Branchen Medienwirtschaft sowie Werbung und Marktforschung zeigen sich zwei Schwerpunkte: zum einen Frankfurt am Main und das gesamte Rhein-Main-Gebiet als bundesweit bedeutender Medien- und Verlagsstandort, zum anderen die Stadt Kassel als zentraler Medienstandort in Nordhessen (u.a. Sitz der Medienanstalt Hessen). Neben klassischen Werbeagenturen sind hier zunehmend digitale Marketingdienstleister und spezialisierte Online-Agenturen vertreten. Die Kombination aus regionaler Medienproduktion, kreativen Dienstleistungen und institutionellen Akteuren führt zu einer erhöhten relativen Konzentration dieser Branchen in urbanen Räumen. (HMWVW 2019, IHK Wiesbaden 2026)

Architektur- und Ingenieurwesen sowie Forschung und Entwicklung weisen eine deutliche Konzentration in Darmstadt auf. Als „Wissenschaftsstadt“ mit Technischer Universität und Hochschule bildet Darmstadt einen zentralen Standort für ingenieurwissenschaftliche Ausbildung, Forschung und technologieorientierte Unternehmensgründungen. Die enge Verzahnung von Wissenschaft, Wirtschaft und angewandter Forschung begünstigt die Ansiedlung entsprechender Unternehmen. Auch der Landkreis Gießen zeigt eine erhöhte relative Präsenz in diesen Bereichen, was unter anderem auf die Technische Hochschule Mittelhessen und die Justus-Liebig-Universität zurückzuführen ist. Die akademische Infrastruktur wirkt hier als wichtiger Standortfaktor für wissens- und forschungsintensive Tätigkeiten.

Abb. 7: Räumliche Branchenstruktur in hessischen Kreisen



Eigene Darstellung 2025

2.5 Digitalaffinität (Modelltraining und Architektur, Evaluierung)

Analog zu dem beschriebenen binären Klassifikationsmodell für die Branchen wurde ein binäres SetFit-Modell zur Bestimmung der Digitalaffinität der Unternehmen entwickelt. Ziel dieses Modells ist es, anhand der Webseitentexte systematisch zu erkennen, welche Unternehmen digitalisierte Geschäftsprozesse, IT-Infrastruktur oder digitale Interaktionsformen implementiert haben, somit einen gewissen Grad an Digitalisierung erreicht haben.

Bei der Erstellung des Trainingsdatensatzes wurde die branchenabhängige semantische Ausprägung digitaler Transformation durch unterschiedliche Schlagworte beachtet. So finden sich in der Informations- und Kommunikationstechnologie Begriffe wie *Cloud Computing, KI/AI-Tools* oder *Smart-Home-Produkte*, während im Maschinenbau oder Fahrzeugbau Konzepte wie *Industrie 4.0, Digitaler Zwilling* oder *Automatisierung* typisch sind. Die Digitalisierungsforschung zeigt, dass Unternehmen digitale Technologien nicht einheitlich, sondern branchen- und kontextspezifisch einsetzen, wodurch auch die sprachlichen Indikatoren digitaler Transformation stark variieren (Brennen & Kreiss, 2016; Legner et al., 2017). Diese branchenabhängige Semantik wurde explizit in der Datenaufbereitung berücksichtigt, um die Variabilität der Ausdrucksweise digitaler Transformation entsprechend abzubilden. Diese Vorgehensweise reduziert das Risiko von Fehlklassifikationen, insbesondere bei Low-Data-Szenarien, und gewährleistet eine robuste Modellleistung über unterschiedliche Branchen hinweg (Gao et al., 2021a; Tunstall et al., 2022).

Nachfolgend werden exemplarisch Schlagworte dargestellt, die die Heterogenität Abbildung der Digitalisierung in unterschiedlichen Sektoren auf Webseiten zeigt. Diese Gruppierung dient nicht der wissenschaftlichen Klassifikation der Branchen, sondern als semantische Grundlage für die Trainingsdaten, um sicherzustellen, dass das Modell unterschiedliche Ausdrucksformen digitaler Transformation in verschiedenen Sektoren robust erfassen kann:

Tab. 6: Branchenabhängige exemplarische Schlagworte

Branche	**Exemplarische Schlagworte**
Allgemein	Daten (-basiert/-gesteuert), digitale Geschäftsprozesse/Wertschöpfung/Plattform, KI/AI, Big Data
Rechts- und Steuerberatung, Wirtschaftsprüfung	Legal Tech, RA-Software, digitale Akten
IKT-Hardware und IKT-Dienstleister	IT-Infrastruktur, Software (-Lösungen), Cloud (-basiert/ Computing), Breitband/FFTx/Glasfaser
Maschinen- und Fahrzeugbau	Smart-Home-Produkte, Robotik, Sensoren, autonome Systeme, intelligente Geräte/Systeme
Werbung und Marktforschung	Social Media Marketing/Plattformen, Echtzeit-Daten, Big Data Analytics

Methodisch basiert das Modell wie beim Branchenklassifikator auf dem SetFit-Ansatz, der kontrastives Fine-Tuning mit Sentence-Embeddings nutzt. Da kontrastive Methoden besonders in Szenarien mit begrenzten annotierten Daten leistungsstark sind (Gao et al., 2021a; Gunel et al., 2021), ist dieses Verfahren gut geeignet, um semantische Variationen digitalisierungsbezogener Begriffe über Branchen hinweg abzubilden. Semantische Ähnlichkeiten zwischen Texten werden unabhängig von der spezifischen Wortwahl erkannt, und das Modell kann Differenzen zwischen hochdigitalisierten und weniger digitalisierten Unternehmen zuverlässig erfassen. Die Evaluierung des Modells erfolgte analog zur Branchenklassifikation über eine 5-fache stratified Cross-Validation, wodurch die Stabilität und Generalisierbarkeit der Klassifikation gesichert wird. Die durchschnittlichen Metriken der Validierung sind in Tabelle 6 dargestellt.

Tab. 7: Durchschnittliche Metriken über 5 Folds

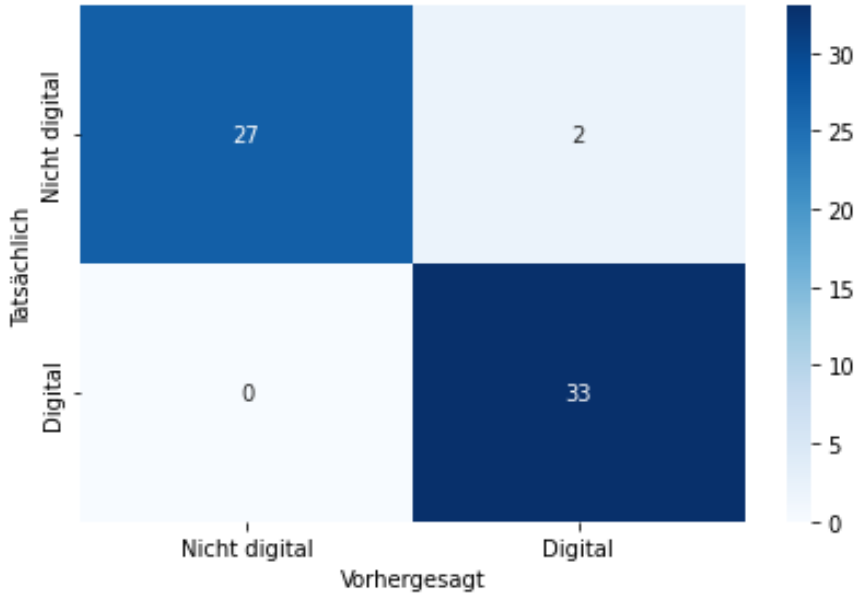
Metrik	**Mittelwert**	**Standardabweichung**
Accuracy	0.9713	0.0119
Precision	0.9498	0.0198
Recall	1.0000	0.0000
F1-Score	0.9742	0.0105

Die durchschnittlichen Metriken über alle fünf Folds zeigen eine insgesamt sehr hohe Modellqualität, wobei insbesondere der hohe F1-Score von 97,42% relevant ist. Insgesamt demonstrieren die Evaluierungsergebnisse, dass das SetFit-Modell den Digitalisierungsgrad aus Webseitentexten sehr zuverlässig identifizieren kann. Die Kombination aus hohem Recall, sehr guter Präzision und äußerst stabilen Ergebnissen über alle Folds hinweg unterstreicht, dass die kontrastive Architektur nicht nur robust gegenüber variierenden sprachlichen Ausdrucksformen ist, sondern auch in stark unausbalancierten Klassifikationsszenarien eine präzise Trennung zwischen digitalisierten und nicht digitalisierten Unternehmen ermöglicht.

Tab. 8: Klassifikationsbericht binäres Modell zur Klassifikation der Digitalaffinität (letzter Fold)

Klasse	**Precision**	**Recall**	**F1-Score**	**Support**
Nicht digital	1.00	0.93	0.96	929
Digital	0.94	1.00	0.97	33
Gesamt (Accuracy)			**0.97**	**66**
Macro Average	0.97	0.97	0.97	66
Weighted Avg	0.97	0.97	0.97	66

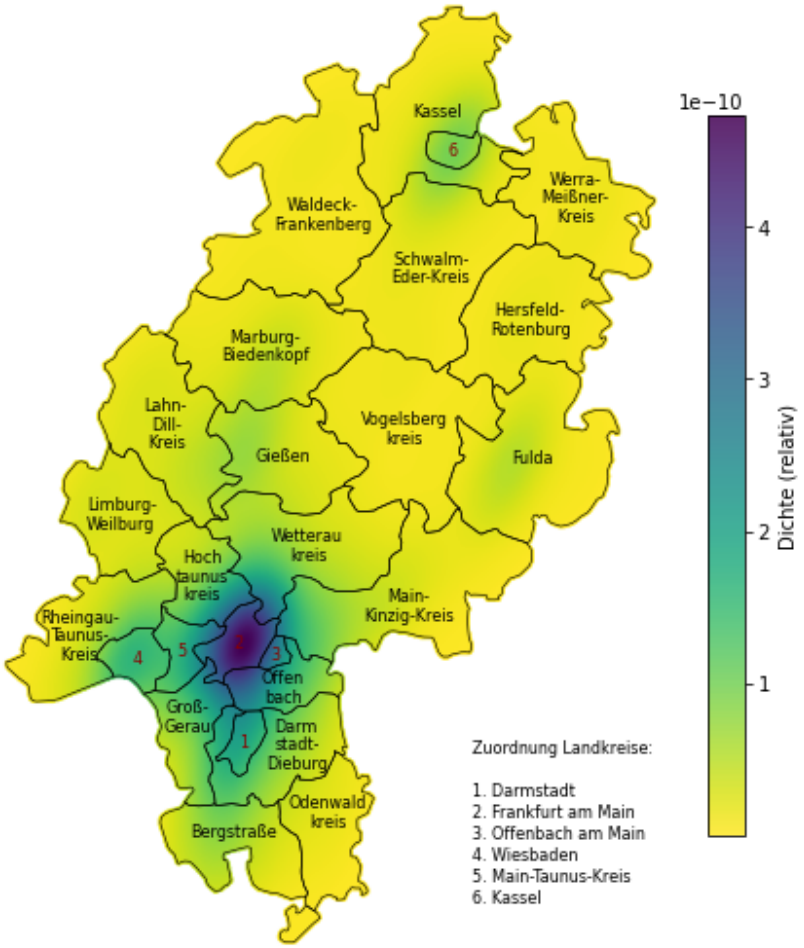
Abb. 8: Konfusionsmatrix Modell 3



Eigene Darstellung 2025

Insgesamt konnten auf Basis ihres Webseitentextes 9.178 von 42.865 Unternehmen als digital klassifiziert werden, was einem Anteil von 21,4% entspricht. Somit kommuniziert etwa jedes fünfte kleine und mittlere Unternehmen der Informationswirtschaft und des verarbeitenden Gewerbes in Hessen eine gewisse Digitalaffinität auf ihrer Webseite. Die geographische Verteilung der digitalisierten Unternehmen in Hessen zeigt ein deutlich räumlich konzentriertes Muster. Die Heatmap verdeutlicht, dass die größte Dichte digitaler Unternehmen im Rhein-Main-Gebiet auftritt, insbesondere in den urbanen Zentren Darmstadt (1), Frankfurt am Main (2), Offenbach am Main (3) und Wiesbaden (4). Diese Regionen weisen im Vergleich zum restlichen Landesgebiet signifikant höhere relative Dichtewerte auf und bilden einen klar erkennbaren Hotspot digitaler wirtschaftlicher Aktivität. Des Weiteren lässt sich eine höhere Dichte an digitalen Unternehmen entlang der Lahn (Wetzlar, Gießen, Marburg), sowie den Städten Fulda und Kassel feststellen. Insgesamt wird sichtbar, dass die Dichte digitaler Unternehmen vom urbanen und technologiegeprägten Rhein-Main-Gebiet ausstrahlt, während ländliche Räume strukturell weniger digital geprägte Unternehmenslandschaften aufweisen.

Abb. 9: Digitale Unternehmen in Hessen – Heatmap



Eigene Darstellung 2025

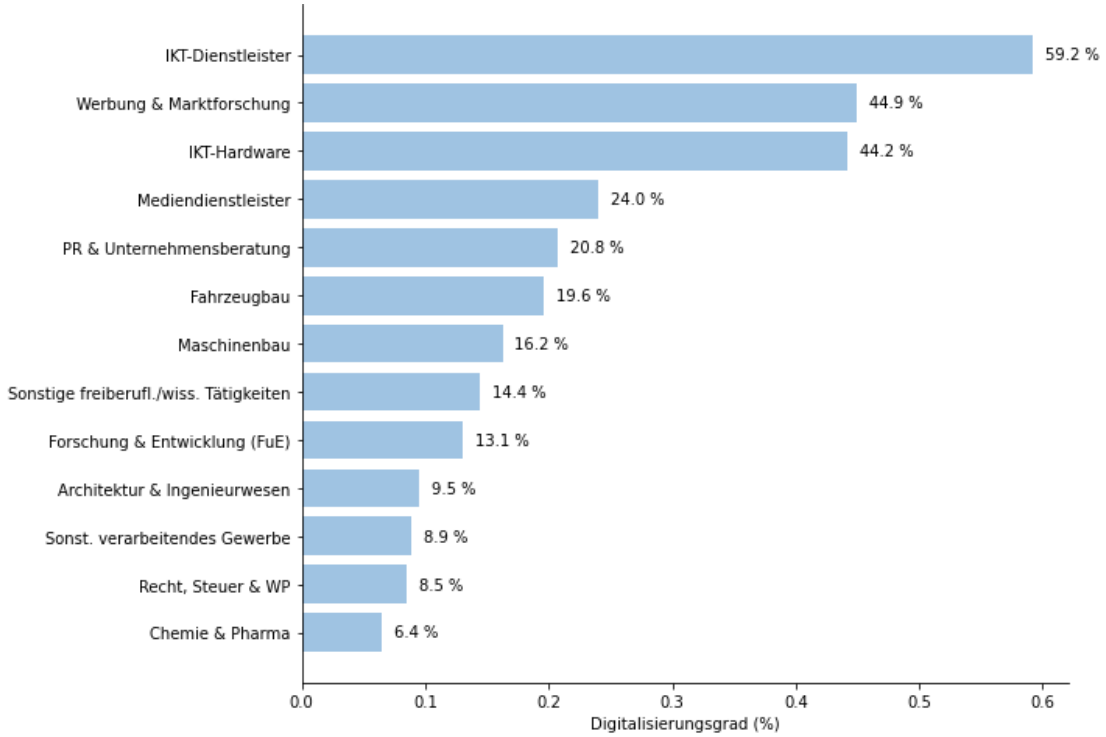
3. Ergebnisse

Aufbauend auf der zuvor beschriebenen mehrstufigen Klassifikationspipeline wird im nächsten Analyseschritt der Digitalisierungsgrad der identifizierten Branchen systematisch ausgewertet und in einen regionalökonomischen Kontext eingebettet. Ziel ist es, sowohl allgemeine Muster der Digitalisierung in den betrachteten Branchen als auch sektoral-räumliche Unterschiede in der digitalen Durchdringung hessischer Unternehmen der Informationswirtschaft und des verarbeitenden Gewerbes sichtbar zu machen. Der im Rahmen dieses Ansatzes operationalisierte Digitalisierungsgrad basiert auf der automatisierten Auswertung von Webseitentexten und misst somit primär die kommunizierte bzw. sichtbare Digitalisierung und nicht notwendigerweise den tatsächlichen Reifegrad interner digitaler Prozesse. Diese methodische Einschränkung ist bei der Einordnung der Ergebnisse zu berücksichtigen.

Die aggregierte Darstellung des Digitalisierungsgrads auf Branchenebene (Abb. 10) zeigt deutliche Unterschiede zwischen den betrachteten Sektoren. Dabei wird ein strukturelles

Gefälle hinsichtlich des Digitalisierungsgrads zwischen wissensintensiven Dienstleistungsbranchen und klassischen Industriebranchen deutlich. Einen vergleichsweise hohen Digitalisierungsgrad weisen die KMUs aus den Branchen IKT-Dienstleister (59,2 %), Werbung & Marktforschung (44,9 %) sowie IKT-Hardware (44,2 %) auf. Diese Ergebnisse sind konsistent mit der Erwartung, dass gerade informations- und wissensintensive Branchen eine hohe Affinität zu digitalen Technologien besitzen, denn deren Geschäftsmodelle basieren unmittelbar auf digitalen Technologien, datengetriebenen Prozessen und internetbasierten Dienstleistungen, welche auch nach außen kommuniziert werden. Demgegenüber weisen klassische Industriebranchen wie Maschinenbau (16,2 %), Fahrzeugbau (19,6 %) sowie insbesondere Chemie und Pharmazie (6,4 %) einen deutlich geringeren durchschnittlichen Digitalisierungsgrad auf. Diese Branchen befinden sich typischerweise in Transformationsprozessen, in denen digitale Technologien zunehmend in die klassische industrielle Produktion integriert werden (z. B. Industrie 4.0, datengetriebene Dienstleistungen), jedoch noch nicht vollständig durchdrungen sind. Insbesondere kleine und mittlere Unternehmen, auf denen die vorliegende Analyse primär basiert, verfügen oftmals über begrenzte finanzielle und personelle Ressourcen für umfassende Digitalisierungsinitiativen.

Abb. 10: Digitalisierungsgrad nach Branche

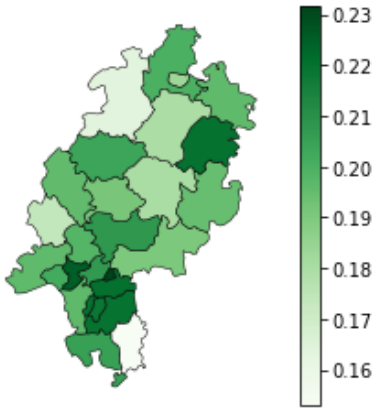


Eigene Darstellung 2025

Abbildung 12 zeigt die räumliche Differenzierung des Digitalisierungsgrades der Branchen auf Kreisebene und ermöglicht eine vertiefte Analyse regionaler Muster, wie digitalisiert eine bestimmte Branche in einem bestimmten Kreis ist. Die dargestellten Werten zeigen somit die relativen Anteile innerhalb eines Kreises, d. h. der Digitalisierungsgrad wird ins Verhältnis zur Anzahl der in dieser Branche im jeweiligen Kreis vorhandenen Unternehmen gesetzt. Bei der Interpretation dieser Abbildung ist jedoch zu berücksichtigen, dass die einzelnen Karten jeweils branchenspezifisch skaliert

sind. Das bedeutet, dass die Farbintensitäten innerhalb jeder Teilkarte relativ zur jeweiligen Branche normiert wurden und somit keine direkte Vergleichbarkeit der Farbwerte zwischen den Branchen gegeben ist. Eine dunklere Einfärbung signalisiert folglich lediglich einen höheren Digitalisierungsgrad innerhalb der jeweiligen Branche, nicht jedoch im branchenübergreifenden Vergleich. Kreise mit weniger als vier Unternehmen der Branche werden grau dargestellt, um zu große, relative Verzerrungen durch geringe Fallzahlen zu vermeiden. Abbildung 11 zeigt ergänzend den Anteil der digitalen Unternehmen in Informationswirtschaft und verarbeitenden Gewerbe je Kreis.

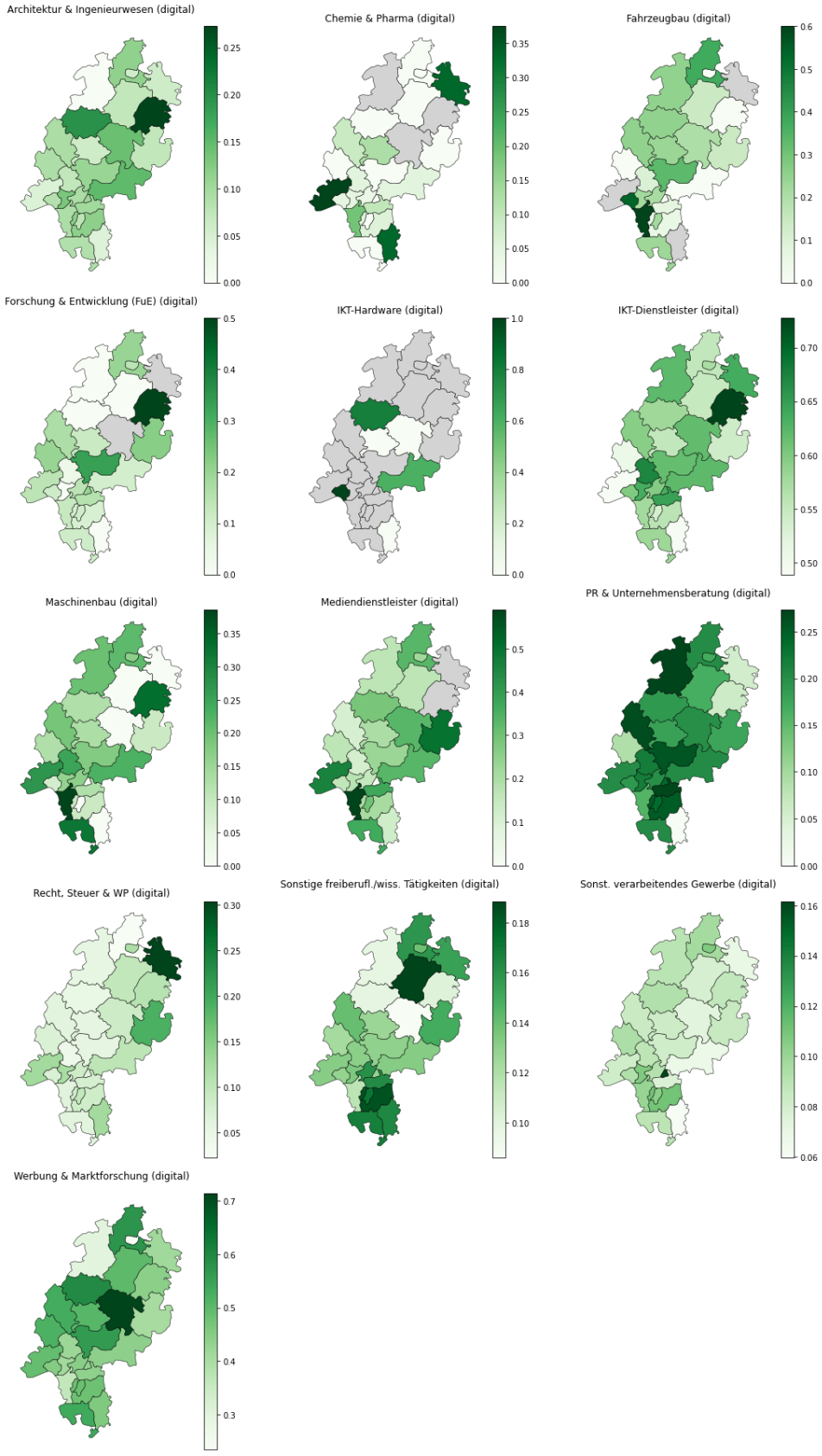
Abb. 11: Anteil digitaler Unternehmen je Kreis



Eigene Darstellung 2025

Die räumlich differenzierte Betrachtung verdeutlicht, dass der Digitalisierungsgrad der Branchen nicht homogen über Hessen verteilt ist, sondern klare regionale Muster aufweist. Dennoch zeigt sich insgesamt eine erhöhte digitale Durchdringung in den urban geprägten Regionen Hessens, insbesondere im Rhein-Main-Gebiet und Südhessen mit Ausnahme des Odenwaldkreises, der zu den strukturschwachen Regionen Hessens zählt (BMWE 2021; HMWVW 2021). Dies lässt sich durch eine hohe regionale Dichte an wissensintensiven Dienstleistungen, Forschungseinrichtungen sowie technologieorientierten Unternehmen zurückführen. Insbesondere die digitalaffinen Branchen wie IKT-Dienstleister, Werbung und Marktforschung, Mediendienstleister und PR und Unternehmensberatung (vgl. Abb. 10) konzentrieren sich mit hohen Digitalisierungsgraden räumlich im Rhein-Main-Gebiet und Südhessen. Diese Regionen sind somit nicht nur strukturell stark in digitalen Branchen vertreten (vgl. Abb. 7), diese Branchen weisen auch in den Regionen einen hohen relativen Digitalisierungsgrad auf. Gleichzeitig lassen sich bei den Branchen des produzierenden Gewerbes, welche einen insgesamt niedrigeren Digitalisierungsgrad aufweisen, regionale Unterschiede feststellen. So sind beispielsweise im Maschinenbau oder Fahrzeugbau trotz des im Durchschnitt geringeren Digitalisierungsniveaus einzelne Kreise mit vergleichsweise hohen Digitalisierungsgraden erkennbar. Diese konzentrieren sich in industriell geprägten Regionen in Nordhessen, wie dem Landkreis Kassel oder dem Landkreis Waldeck-Frankenberg, oder Südhessen, wie dem Landkreis Groß-Gerau oder dem Landkreis Bergstraße. Dabei ist auffällig, dass die höheren relativen Digitalisierungsgrade im Fahrzeug- und Maschinenbau mit in diesen Regionen bestehenden industriellen Clustern und Netzwerken korreliert. Dies deutet darauf hin, dass auch innerhalb traditioneller Industrien Transformationsprozesse stattfinden, die jedoch stärker cluster- und unternehmensspezifisch ausgeprägt sind. Im Vergleich dazu steht der Landkreis Fulda, der ebenfalls durch den Fahrzeug- und Maschinenbau geprägt ist (vgl. Abb. 7), jedoch in keiner beiden Branchen einen hohen Digitalisierungsgrad aufweist.

Abb. 12: Digitalisierungsgrad nach Branchen und Kreisen (relativ)

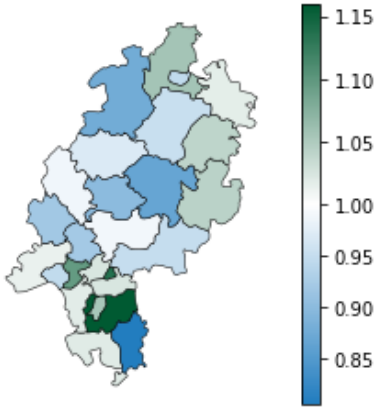


Eigene Darstellung 2025

Zur weitergehenden Einordnung der regionalen Digitalisierungsstrukturen wird ergänzend der Locationquotient (LQ) des Digitalisierungsgrades betrachtet (Abbildung 13 und 14). Der LQ misst hierbei, ob eine Branche in einer Region über- oder unterdurchschnittlich stark digitalisiert ist im Vergleich zur durchschnittlichen Digitalisierungsgrad Hessens. Dabei muss allerdings beachtet werden, dass keine Aussage über die allgemeine Wichtigkeit der Branche in der Region getroffen werden kann (Abb. 14). Werte größer als 1 signalisieren eine überdurchschnittliche Spezialisierung einer Region auf digitalisierte Unternehmen dieser Branche, während Werte kleiner als 1 auf eine unterdurchschnittliche Ausprägung hinweisen. Zur Verbesserung der Interpretierbarkeit wurden die Werte zusätzlich um den Mittelpunkt 1 normiert. Regionen ohne ausreichende Fallzahlen werden ausgegraut dargestellt, um Verzerrungen durch kleine Stichproben zu vermeiden.

Abbildung 13 zeigt zunächst den aggregierten, branchenübergreifend normalisierten LQ für Hessen, also die relative Spezialisierung der Kreise hinsichtlich digitalisierter Branchenstrukturen unter Berücksichtigung der jeweiligen Branchenbedeutung. Somit kann eine Aussage darüber getroffen werden, ob strategisch wichtige und bedeutende Branchen einer Region überdurchschnittlich digital geprägt sind. Berücksichtigt wird hierbei sowohl die regionale Branchenstruktur als auch die landesweite Bedeutung der jeweiligen Branchen. Hohe Werte weisen folglich auf Kreise hin, in denen digitale Unternehmen insbesondere in wirtschaftlich relevanten und strukturprägenden Branchen überrepräsentiert sind. Südhessen und das Rhein-Main-Gebiet können dabei als überdurchschnittlich digital geprägt eingeordnet werden. Diese Regionen verfügen nicht nur über eine hohe absolute Dichte digitaler Unternehmen, sondern weisen zugleich eine überdurchschnittliche Spezialisierung auf digitalisierte wissens- und technologieintensive Branchen auf (vgl. Abb. 7 und Abb. 10). Des Weiteren haben einzelne industriell geprägte Regionen wie beispielsweise der LK Kassel, der LK Hersfeld-Rothenburg oder der LK Fulda trotz insgesamt geringerer Digitalisierungsgrade der jeweiligen vorherrschenden Branchen einen LQ größer 1. Dies bedeutet, dass die für diese Regionen strukturell besonders bedeutenden Branchen im Vergleich zu denselben Branchen in anderen hessischen Kreisen überdurchschnittlich stark digitalisiert sind. Die Ergebnisse verdeutlichen somit, dass auch in traditionellen Industrie- und Produktionsregionen digitale Transformationsprozesse stattfinden, selbst wenn das allgemeine branchenspezifische Digitalisierungsniveau insgesamt niedriger ausfällt. Digitalisierung zeigt sich dort folglich weniger als breitflächiges Phänomen über alle Regionen hinweg, sondern als gezielte Modernisierung regionaler industrieller Kernkompetenzen und bestehender Wertschöpfungsstrukturen.

Abb. 13: Hessenweiter LQ, normalisiert mit der Branchenbedeutung



Eigene Darstellung 2025

Dies lässt sich durch den branchenspezifischen LQ auf Kreisebene (Abb. 14) konkretisieren. Im Gegensatz zu den vorherigen Karten des Digitalisierungsgrads wird

dargestellt, ob eine Branche relativ stärker oder schwächer digitalisiert ist als der hessische Durchschnitt dieser Branche (Hinweis: Keine Aussage über die allgemeine Wichtigkeit der Branche in der Region).

Wie bereits benannt, treten insbesondere in den industriellen Branchen, wie dem Fahrzeug- oder Maschinenbau räumliche Unterschiede hervor. Besonders deutlich wird dies im Fahrzeugbau, wo insbesondere der Landkreis Kassel sowie der Landkreis Groß-Gerau hohe Werte des Locationquotients aufweisen, was auf eine überdurchschnittliche Digitalaffinität der dort ansässigen Unternehmen des Fahrzeugbaus im Vergleich zu Fahrzeugbauunternehmen in anderen hessischen Kreisen hinweist. Dabei gehört die Branche in diesen Regionen zu den wichtigen Standpfeilern der Wirtschaftsstruktur (IHK Darmstadt Rhein Main Neckar 2026a, Wirtschaftsförderung Region Kassel GmbH 2026). Als möglicher Erklärungsansatz können die seit Jahren etablierten Cluster- und Netzwerkstrukturen der Automobilwirtschaft herangezogen werden. In der Region Kassel unterstützen beispielsweise die Transformationsnetzwerke MoWiN und TRegKS Unternehmen bei Digitalisierungs-, Innovations- und Transformationsprojekten und fördern Kooperationen zwischen Industrie, IT-Wirtschaft und Forschungseinrichtungen (HMWEVW 2026, MoWiN 2026, TRegKS 2026). Ziel dieser Initiativen ist es, die Region langfristig als innovativen Technologie- und Produktionsstandort der Mobilitätswirtschaft weiterzuentwickeln (TRegKS 2026). Ähnliche Unterstützungsstrukturen bestehen im Automotive Cluster RheinMainNeckar, das bereits seit 2003 Unternehmen der Automobilzulieferindustrie vernetzt. Die Clusterinitiative bündelt Kompetenzen spezialisierter Unternehmen der Automobilzulieferbranche und agiert als Koordinator und Moderator im Netzwerk (HMWVW 2026, IHK Darmstadt Rhein Main Neckar 2026a). Dabei wird auch spezielle Hilfestellung und Unterstützung für kleine und mittlere Automobilzulieferer in Hinblick auf die Digitalisierung angeboten, beispielsweise durch den „Baukasten für den Einstieg in die digitale Lieferkette“ (IHK Darmstadt Rhein Main Neckar 2026b). Die langjährige Etablierung solcher Netzwerkstrukturen könnte dazu beitragen, dass Unternehmen dieser Regionen bei der digitalen Transformation einen Vorsprung gegenüber vergleichbaren Zulieferstandorten ohne vergleichbar ausgeprägte Unterstützungsangebote aufweisen oder diese zumindest sichtbar in der Außendarstellung der Unternehmen verankert sind.

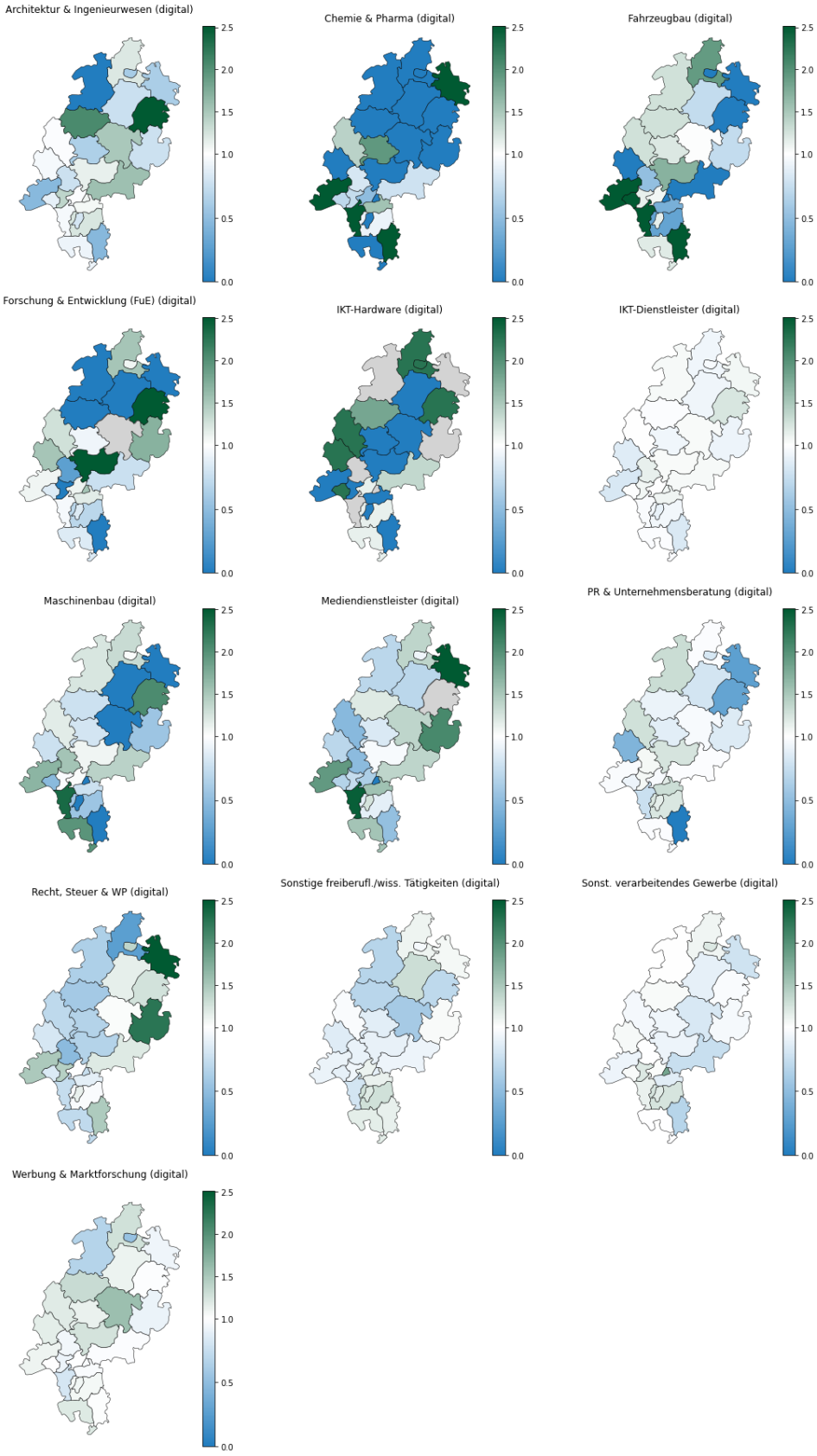
Auch in Mittelhessen weisen beispielsweise die Landkreise Gießen und Marburg-Biedenkopf erhöhte Locationquotients im Fahrzeugbau auf. Die Werte fallen geringer aus als in den stärker etablierten Automobilregionen Kassel oder Rhein-Main-Neckar, deuten jedoch bereits auf eine vergleichsweise überdurchschnittliche digitale Affinität der regionalen Unternehmen hin. Der überdurchschnittliche Locationquotient kann somit als Ausdruck eines beginnenden Transformationsprozesses und ein Indikator für vorhandene Entwicklungspotenziale und einer strukturellen Neuausrichtung der regionalen Fahrzeugindustrie interpretiert werden. Für die zukünftige Entwicklung könnte hierbei insbesondere das Transformationsnetzwerk TeamMit von Bedeutung sein. Das Ende 2022 gegründete Netzwerk unterstützt Automobilzulieferer bei den Herausforderungen des Strukturwandels und adressiert zentrale Zukunftsthemen wie Digitalisierung, Nachhaltigkeit, Fachkräftesicherung und die Entwicklung neuer Geschäftsmodelle (Regionalmanagement Mittelhessen 2025).

Im Gegensatz zu den Industriebranchen mit regionalen Clustern weisen die bereits stark digitalaffinen Branchen IKT-Dienstleister, Werbung und Marktforschung, Mediendienstleister sowie teilweise PR und Unternehmensberatung (vgl. Abb. 10) insgesamt deutlich geringere räumliche Unterschiede auf. Entsprechend sind nur geringe

Abweichungen des LQ der hessischen Kreise vorhanden. Die Ergebnisse deuten daher darauf hin, dass die Digitalaffinität in diesen Branchen weitgehend flächendeckend vorhanden ist und regionale Standortunterschiede eine geringere Rolle spielen als in den industriellen Branchen. Die Digitalisierung stellt somit in diesen Branchen kein regionales Differenzierungsmerkmal dar, sondern kann als grundlegende Voraussetzung des Geschäftsmodells verstanden werden.

Auch bei anderen wissensintensiven Dienstleistungsbranchen, die einen geringeren Digitalisierungsgrad aufweisen, zeigen sich wenig ausgeprägte regionale Unterschiede. Im Bereich Recht, Steuerberatung und Wirtschaftsprüfung konzentrieren sich überdurchschnittlich digitalisierte Unternehmen tendenziell auf das Rhein-Main-Gebiet und Südhessen. Gleichzeitig treten vereinzelt auch strukturschwächere Regionen hervor, etwa der Werra-Meißner-Kreis, in dem die Branche insgesamt nur einen kleinen Anteil an der regionalen Wirtschaftsstruktur ausmacht. Dennoch weisen die vorhandenen Unternehmen eine überdurchschnittliche digitale Affinität auf. Ein Beispiel hierfür ist eine regionale Steuerberatungskanzlei, die auf ihrer Unternehmenswebseite digitale Mandantenprozesse aktiv bewirbt und unter dem Konzept „Buchführung mit Zukunft“ vollständig digitale Austausch- und Dokumentationsprozesse anbietet. Solche Beispiele verdeutlichen, dass auch in peripheren Regionen einzelne Unternehmen digitale Vorreiterrollen übernehmen können.

Abb. 14: Locationquotient nach Branchen und Kreisen



Eigene Darstellung 2025

4. Diskussion

Die Ergebnisse verdeutlichen zunächst, dass die Digitalaffinität hessischer KMU stark branchenabhängig ausgeprägt ist. Insbesondere Unternehmen der Informationswirtschaft sowie wissensintensiver Dienstleistungen weisen deutlich höhere Digitalisierungsgrade auf als Unternehmen klassischer Industriebranchen. Die Plausibilitätsprüfung der Webseiten legt nahe, dass Digitalaffinität in der Informationswirtschaft und den wissensintensiven Dienstleistungen einen wesentlichen Bestandteil der Wertschöpfung bildet. Die Unternehmen dieser Branchen werben explizit mit digitalen Dienstleistungen, Softwarelösungen, Online-Marketing oder datenbasierten Geschäftsmodellen. Die vergleichsweise geringen räumlichen Unterschiede der LQ-Werte innerhalb dieser Branchen (Abb. 14) unterstützen diese Interpretation zusätzlich, sodass die Digitalaffinität als notwendige Grundvoraussetzung für die Marktteilnahme gilt und daher keinen regional differenzierenden Wettbewerbsfaktor darstellt. Im Gegensatz dazu zeigen sich im verarbeitenden Gewerbe deutliche regionale Unterschiede in der Digitalaffinität. Die Ergebnisse sprechen dafür, dass bestehende Netzwerke, Transformationsinitiativen und branchenspezifische Unterstützungsstrukturen die digitale Transformation von KMU begünstigen können. Regionen mit etablierten Clustern weisen nicht zwingend den höchsten Digitalisierungsgrad auf, zeigen jedoch häufig eine überdurchschnittliche Digitalaffinität ihrer strukturell bedeutenden Branchen.

Dabei muss allerdings beachtet werden, dass die Operationalisierung von Digitalaffinität auf den Inhalten von Unternehmenswebseiten basiert und somit primär die kommunizierte beziehungsweise sichtbare Digitalisierung erfasst. Somit können Unternehmen intern digitalisierte Prozesse implementiert haben, ohne diese auf ihrer Webseite explizit darzustellen. Umgekehrt können Unternehmen digitale Technologien intensiv bewerben, obwohl deren tatsächliche betriebliche Nutzung vergleichsweise begrenzt ist. Die Ergebnisse erlauben daher keine direkte Aussage über den tatsächlichen digitalen Reifegrad eines Unternehmens, sondern über die kommunikativ sichtbare Digitalaffinität. Die in Abbildung 7 beobachteten Unterschiede zwischen den Branchen spiegeln daher nicht ausschließlich reale Unterschiede in der Digitalisierung wider, sondern sind teilweise auch durch unterschiedliche Kommunikationsstrategien und Geschäftsmodelle geprägt. Branchen mit starkem Endkundenbezug (B2C), wie etwa Werbung und Marktforschung oder Mediendienstleistungen, weisen naturgemäß eine höhere digitale Sichtbarkeit auf. In diesen Sektoren ist die Außendarstellung zentraler Bestandteil des Geschäftsmodells, wodurch digitale Technologien, Plattformen und datengetriebene Ansätze aktiv auf Webseiten kommuniziert werden. Dies führt tendenziell zu höheren gemessenen Digitalisierungsgraden. Demgegenüber sind Branchen mit überwiegend unternehmensbezogenen Geschäftsmodellen (B2B), wie etwa Maschinenbau oder Teile des verarbeitenden Gewerbes, häufig deutlich zurückhaltender in der digitalen Selbstdarstellung. In diesen Sektoren spielt die Webseite oftmals eine geringere Rolle als Vertriebskanal oder Kommunikationsinstrument. Auch sind digitale Prozesse stärker intern, beispielsweise durch Prozessoptimierung, Automatisierung und datenbasierte Steuerung von Produktionsabläufen in Produktion, Logistik oder Supply-Chain-Management, verankert und damit weniger unmittelbar in öffentlich sichtbaren Webseitentexten abgebildet. Diese Formen der Digitalisierung werden durch den gewählten Messansatz nur eingeschränkt erfasst, was zu einer systematischen Unterschätzung des tatsächlichen Digitalisierungsgrades führen kann.

Zusätzlich ist ein struktureller Bias durch die Fokussierung auf kleine und mittlere Unternehmen zu berücksichtigen. In Branchen wie beispielsweise Chemie und Pharma

wird ein erheblicher Teil der tatsächlichen digitalen Innovationskraft durch große, kapitalintensive Unternehmen mit hochentwickelten digitalen Infrastrukturen, automatisierten Produktionsprozessen und datenintensiven Forschungsumgebungen getragen, die allerdings aufgrund des Analysefokusses auf KMU nicht erfasst sind. Der vergleichsweise niedrige gemessene Digitalisierungsgrad dieser Branche (6,4 %) ist daher nicht als Indikator für eine geringe technologische Leistungsfähigkeit zu interpretieren, sondern vielmehr als Ergebnis der eingeschränkten Sichtbarkeit digitaler Aktivitäten der KMU. Hinzu kommt die Kapitalintensität industrieller Sektoren, die zu längeren Investitionszyklen führt. Bestehende Produktionsanlagen werden über viele Jahre genutzt, sodass digitale Transformationen häufig schrittweise und inkrementell erfolgen. Insbesondere kleine und mittlere Unternehmen verfügen oftmals über begrenzte finanzielle und personelle Ressourcen für umfassende Digitalisierungsinitiativen.

Dennoch können die Ergebnisse als Indikator für die nach außen sichtbare, digitale Transformation interpretiert werden. Insbesondere ist dies für regionale Innovations- und Transformationsprozesse relevant, da digitale Sichtbarkeit ein wichtiger Bestandteil der Wettbewerbsfähigkeit und Außenwahrnehmung von Unternehmen geworden ist. Eine manuelle Nachprüfung zeigt, dass Unternehmen als digital klassifiziert werden, wenn Digitalisierung aktiv als Teil des Leistungsversprechens kommuniziert wird, etwa durch Begriffe wie „digitale Prozesse“, „Cloud-Lösungen“, „Unternehmen online“, „Industrie 4.0“ oder digitale Kundenplattformen. Gleichzeitig existieren auch Unternehmen, deren Webseiten technisch und gestalterisch wenig modern wirken, die jedoch dennoch eindeutig digitale Produkte oder Dienstleistungen anbieten. Beispiele hierfür finden sich insbesondere bei den IKT-Dienstleistern. Die gemessene Digitalaffinität muss daher im Kontext branchenspezifischer Wertschöpfungslogiken, Kommunikationsstrategien und struktureller Rahmenbedingungen betrachtet werden.

Insgesamt verdeutlicht die differenzierte Analyse, dass Digitalisierung nicht als flächendeckender und einheitlicher Transformationsprozess auftritt, sondern als regional und sektoral differenzierter Strukturwandel, der eng mit bestehenden Branchenkompetenzen, Innovationsnetzwerken und regionalen Entwicklungspfaden verknüpft ist.

5. Zusammenfassung

Der vorliegende Technical Report beschreibt die automatisierte Klassifizierung von hessischen KMU der Informationswirtschaft und des verarbeitenden Gewerbes sowie die Messung ihrer Digitalaffinität anhand von Webseitentexten. Um diese komplexe Aufgabe ressourceneffizient zu lösen, wurde eine mehrstufige Machine-Learning-Pipeline auf Basis der SetFit-Architektur mit dem Sprachmodell DistilBERT entwickelt. Das Modelltraining basiert auf manuell annotierten Texten, die jeweils zielgerichtet für die Klassifikationsaufgabe zusammengestellt wurden. Der Klassifikationsprozess erfolgt dabei sequenziell in mehreren Schritten. In einem ersten Schritt trennt ein binäres Modell branchenrelevante von nicht-relevanten Webseiten mit einer Genauigkeit (Accuracy) von 97,8 Prozent. Anschließend ordnet ein Multiclass-Modell die relevanten Unternehmen mit 92,4 Prozent Genauigkeit spezifischen Subbranchen zu. In einem dritten Schritt bewertet ein weiteres Modell die kommunizierte Digitalisierung der Unternehmen. Dabei berücksichtigt der Algorithmus branchenspezifische Terminologien, da sich digitale Transformation in der IT-Branche beispielsweise durch Begriffe wie „Cloud“ ausdrückt,

während im Maschinenbau Konzepte wie „Industrie 4.0“ dominieren. Auch dieses Modell erreicht mit einem F1-Score von 97,4 Prozent eine sehr hohe Zuverlässigkeit.

Die Anwendung der Pipeline auf rund 185.000 erfasste Domains zeigte, dass etwa 23 Prozent der Unternehmen den untersuchten Zielbranchen angehören. Geografisch konzentrieren sich diese stark auf das Rhein-Main-Gebiet, insbesondere Frankfurt, sowie den Raum Kassel. Die automatisierte Klassifikation spiegelt dabei historisch gewachsene Wirtschaftscluster präzise wider. So zeigt sich eine deutliche Konzentration des Fahrzeugbaus in Fulda und Nordhessen, die Chemie- und Pharmaindustrie bündelt sich an der Bergstraße, und wissensintensive Dienstleistungen wie Forschung und Entwicklung sowie Architektur konzentrieren sich im Raum Darmstadt.

Hinsichtlich der Digitalaffinität ergab die Analyse, dass rund 21,4 Prozent der relevanten KMU eine sichtbare digitale Ausrichtung aufweisen. Dabei tritt ein starkes sektorales Gefälle zutage. Während wissensintensive Dienstleister wie die IKT-Branche mit knapp 59 Prozent oder Marketingagenturen mit 45 Prozent hochgradig digitalisiert auftreten, weisen klassische Industriebranchen wie der Maschinenbau (16 Prozent) oder die Chemieindustrie (6 Prozent) deutlich geringere Werte auf. Räumlich betrachtet ballt sich die Digitalisierung erwartungsgemäß in den urbanen Zentren. Die tiefergehende Analyse mittels des Location Quotients (LQ) offenbart jedoch ein differenzierteres Bild. Traditionelle Industrieregionen, wie etwa die Fahrzeugbauregion in und um Kassel oder Groß-Gerau, sind innerhalb ihrer eigenen Branche überdurchschnittlich digitalisiert. Dieser Befund lässt sich unter anderem auf den Erfolg regionaler Transformationsnetzwerke und Cluster-Initiativen zurückführen, die den digitalen Strukturwandel im Mittelstand aktiv fördern.

Trotz dieser überzeugenden Ergebnisse, die durch externe Quelle validiert werden können, weist der Ansatz auch methodische Grenzen auf. Der gewählte Ansatz misst ausschließlich die nach außen kommunizierte Digitalisierung. Insbesondere im B2B-Sektor und in klassischen Industrien werden digitale Produktions- und Logistikprozesse oft intern optimiert, ohne werbewirksam auf der Webseite platziert zu werden. Darüber hinaus führt der Fokus auf KMU dazu, dass die enorme digitale Innovationskraft von Großkonzernen, beispielsweise in der Chemie- und Pharmaindustrie, in dieser Auswertung nicht vollumfassend abgebildet wird, was bei der Interpretation der sektoralen Digitalisierungsgrade zwingend berücksichtigt werden muss.

Quellenverzeichnis

Brennen, S., & Kreiss, D. (2016). Digitalization. The International Encyclopedia of Communication Theory and Philosophy. https://doi.org/10.1002/9781118766804.wbiect111.

Bundesministerium für Wirtschaft und Energie (BMWE) (2021). GRW-Fördergebiete 2022-2027 und Fördergebiete des GRW-Sonderprogramms „Beschleunigung der Transformation in den ostdeutschen Raffineriestandorten und Häfen“. https://www.bundeswirtschaftsministerium.de/Redaktion/DE/Downloads/grw-fordergebiete-2022-2027.html (abgerufen am 27.03.2026).

Calvino, F., Criscuolo, C., Marcolin, L. & Squicciarini, M. (2018). A taxonomy of digital intensive sectors. OECD Science. Technology and Industry Working Papers, No. 2018/14, OECD Publishing, Paris, https://doi.org/10.1787/f404736a-en.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357. https://doi.org/10.1613/jair.953

Dahlke, J., Schmidt, S., Lenz, D., Kinne, J., Dehghan, R., Abbasiharofteh, M., ... & Rammer, C. (2025). The WebAI paradigm of innovation research: Extracting insight from organizational web data through AI (No. 25-019). ZEW Discussion Papers.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Gao, T., Yao, X., & Chen, D. (2021a). SimCSE: Simple Contrastive Learning of Sentence Embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

Gao, T., Fisch, A., & Chen, D. (2021b). Making pre-trained language models better few-shot learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 3816-3830). https://doi.org/10.18653/v1/2021.acl-long.295

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. In: Adaptive computation and machine learning. MIT Press.

Gunel, B., Du, J., Conneau, A., & Stoyanov, V. (2020). Supervised contrastive learning for pre-trained language model fine-tuning. arXiv preprint arXiv:2011.01403.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hessisches Ministerium für Wirtschaft, Energie, Verkehr und Wohnen (HMWVW) (2019). Cluster- und Netzwerkinitiativen in Hessen. https://www.technologieland-hessen.de/mm/mm001/Cluster_und_Netzwerkinitiativen_in_Hessen_2019.pdf.

Hessisches Ministerium für Wirtschaft, Energie, Verkehr, Wohnen und ländlichen Raum (HMWVW) (2021). Regionale Förderung- Odenwald wird neues Fördergebiet. https://wirtschaft.hessen.de/presse/pressearchiv/odenwald-wird-neues-foerdergebiet#:~:text=Hessen%20kann%20strukturschwache%20Landesteile%20auch%20k%C3%BCnftig%20mit%20j%C3%A4hrlich%20%C3%BCber%20acht%20Mio.%20Euro%20unterst%C3%BCtzen. (abgerufen am 27.03.2026).

Hessisches Ministerium für Wirtschaft, Energie, Verkehr, Wohnen und ländlichen Raum (HMWVW) (2021). MoWiN.net e.V. https://www.technologieland-hessen.de/mowin-net (abgerufen am 01.06.2026).

Hessisches Ministerium für Wirtschaft, Energie, Verkehr, Wohnen und ländlichen Raum (HMWVW) (2026): Kurz-Info zu Automotive Cluster RheinMainNeckar. https://www.technologieland-hessen.de/clusterliste (abgerufen am 01.06.2026).

Industrie- und Handelskammer (IHK) Wiesbanden (2026). Consulting – Heimliche Beraterhauptstadt. https://www.ihk.de/wiesbaden/wirtschaftspolitik/branchen/consulting-1252616#:~:text=Die%20Wirtschaftsf%C3%B6rderungen%20aus%20Wiesbaden%2C%20Bad,mehr%20als%20500%20Unternehmen%20zusammengeschlossen. (abgerufen am 18.02.2026).

Industrie- und Handelskammer (IHK) Darmstadt Rhein Main Neckar (2026a): Automotive-Cluster RheinMainNeckar. https://www.ihk.de/darmstadt/servicemarken/ueber-uns/gremien-ihk-darmstadt/arbeits-und-netzwerke/automotive-cluster-rheinmainneckar-4914754 (abgerufen am 01.06.2026).

Industrie- und Handelskammer (IHK) Darmstadt Rhein Main Neckar (2026b): Baukasten für den Einstieg in die digitale Lieferkette. https://www.automotive-cluster.org/inhalte/ueber-uns/car4kmu/baukasten-fuer-den-einstieg-in-die-digitale-lieferkette-3866908 (abgerufen am 01.06.2026).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Legner, C., Eymann, T., Hess, T., Matt, C., Böhmann, T., Drews, P., Mädche, A., Urbach, N. & Ahlemann, F. (2017). Digitalization: opportunity and challenge for the business and information systems engineering community. Business & information systems engineering, 59(4), 301-308. https://doi.org/10.1007/s12599-017-0484-2

MoWiN.net e.V. (2026). Über uns. https://www.mowin.net/ueber-uns/ (abgerufen am 01.06.2026).

Penedo, G., Kydlíček, H., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L., & Wolf, T. (2024). The fineweb datasets: Decanting the web for the finest text data at scale. Advances in Neural Information Processing Systems, 37, 30811-30849. https://doi.org/10.52202/079017-0970

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1–67.

Region Fulda GmbH (2026). Netzwerke und Cluster. https://www.region-fulda.de/netzwerke-cluster/ (abgerufen am 28.02.2026).

Regionalmanagement Mittelhessen GmbH (2025): Netzwerk für Automobilzulieferer. https://mittelhessen.eu/teammit/ueber-teammit/ (abgerufen am 01.06.2026).

Sammut, C., & Webb, G. I. (Eds.). (2011). Encyclopedia of machine learning. Springer Science & Business Media.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J. G. (2022). Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems, 34(11), 8135-8153. https://doi.org/10.1109/TNNLS.2022.3152527

Tan, J., Dou, Z.,Wang,W.,Wang, M., Chen,W., &Wen, J.-R. (2025). Htmlrag: Html is better than plain text for modeling retrieved knowledge in rag systems. https://doi.org/10.1145/3696410.3714546

Transformationsnetzwerk Region Kassel (TRegKS) 2026: Gemeinsam Richtung morgen! https://tregks.de/ (abgerufen am 01.06.2026).

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. arXiv preprint arXiv:2209.11055.

Wenzek, G., Lachaux, M. A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020, May). CCNet: Extracting high quality monolingual datasets from web crawl data. In Proceedings of the twelfth language resources and evaluation conference (pp. 4003-4012).

Wirtschaftsförderung Bergstraße GmbH (2026a). Pharmazie und Biotechnologie. https://www.wirtschaftsregion-bergstrasse.de/Wirtschaft/Schluesselbranchen-Clusterinitiativen/Pharmazie-Biotechnologie (abgerufen am 23.02.2026).

Wirtschaftsförderung Bergstraße GmbH (2026b). Kunststoff-, Gummi- und Chemische Industrie. https://www.wirtschaftsregion-bergstrasse.de/Wirtschaft/Schluesselbranchen-Clusterinitiativen/Kunststoff-Gummi-Chemische-Industrie (abgerufen am 23.02.2026).

Wirtschaftsförderung Region Kassel GmbH (2026): Kasseler Firmen und Branchenfokus. https://wfg-kassel.de/wirtschaft/firmen-branchen/ (abgerufen am 01.06.2026).

Anhang

A. Kuratierte Liste von Frameworks, die bei der Analyse großer Textmengen eingesetzt werden.

1. NLP und Textverarbeitung

Framework / Bibliothek	Kurzbeschreibung	Link
spaCy	Leistungsfähige NLP-Bibliothek für Tokenisierung, POS-Tagging, Named Entity Recognition, Dependency Parsing und linguistische Pipelines. Besonders geeignet für produktive Anwendungen.	https://spacy.io/
NLTK	Klassische NLP-Bibliothek mit vielen Algorithmen, Korpora und Lehrmaterialien. Häufig in Lehre und Forschung eingesetzt.	https://www.nltk.org/
Gensim	Bibliothek für Topic Modeling, Word Embeddings und semantische Ähnlichkeitsanalysen, z. B. Word2Vec, Doc2Vec und LDA.	https://radimrehurek.com/gensim/
Stanza	NLP-Toolkit der Stanford NLP Group mit Modellen für viele Sprachen, u. a. Tokenisierung, Lemmatization, POS-Tagging und Dependency Parsing.	https://stanfordnlp.github.io/stanza/
Flair	NLP-Framework mit Fokus auf Sequenzklassifikation, Named Entity Recognition und kontextuellen Embeddings.	https://github.com/flairNLP/flair
TextBlob	Einfache Bibliothek für grundlegende NLP-Aufgaben wie Sentiment Analysis, Tokenisierung, POS-Tagging und Übersetzung. Gut geeignet für schnelle Prototypen.	https://textblob.readthedocs.io/
Sentence Transformers	Bibliothek zur Erzeugung semantischer Satz- und Dokument-Embeddings. Häufig genutzt für Ähnlichkeitssuche, Clustering, semantische Suche und Retrieval-Augmented Generation.	https://www.sbert.net/

2. Web Scraping und Crawling

Framework / Bibliothek	Kurzbeschreibung	Link
Requests	Sehr verbreitete Bibliothek für HTTP-Anfragen. Geeignet zum Abrufen von Webseiten, APIs und Dateien.	https://requests.readthedocs.io/
HTTPX	Moderne Alternative zu Requests mit Unterstützung für synchrone und asynchrone HTTP-Anfragen.	https://www.python-httpx.org/

aiohttp	Asynchrone HTTP-Client- und Server-Bibliothek. Besonders nützlich für paralleles Crawling und performante Webanfragen.	https://docs.aiohttp.org/
Scrapy	Umfangreiches Framework für Web Scraping und Web Crawling. Bietet Spider, Pipelines, Request-Handling, Middleware und Exportfunktionen.	https://scrapy.org/
Beautiful Soup	Bibliothek zum Extrahieren von Informationen aus HTML- und XML-Dokumenten. Häufig kombiniert mit Requests.	https://www.crummy.com/software/BeautifulSoup/
Selenium	Browser-Automatisierungsframework. Geeignet für dynamische Webseiten, die stark auf JavaScript basieren.	https://www.selenium.dev/
Playwright	Modernes Browser-Automatisierungsframework für Chromium, Firefox und WebKit. Sehr gut geeignet für JavaScript-lastige Webseiten und End-to-End-Tests.	https://playwright.dev/python/
MechanicalSoup	Kombiniert Requests und Beautiful Soup zur einfachen Interaktion mit Webseiten, Formularen und HTML-Dokumenten.	https://mechanicalsoup.readthedocs.io/

3. Parsing, Boilerplate Removal und Textextraktion

Framework / Bibliothek	**Kurzbeschreibung**	**Link**
Resiliparse	Python-Bibliothek für robustes HTML-Parsing, Web-Datenextraktion und Boilerplate Removal. Besonders relevant für die Verarbeitung großer Webkorpora.	https://resiliparse.chatnoir.eu/
lxml	Sehr schnelle Bibliothek für XML- und HTML-Parsing. Unterstützt XPath und XSLT und wird häufig in Scraping-Workflows verwendet.	https://lxml.de/
selectolax	Schneller HTML-Parser auf Basis von Modest/lexbor. Besonders nützlich für performantes Parsing großer Mengen an Webseiten.	https://github.com/rushter/selectolax
trafilatura	Bibliothek zur Extraktion von Haupttexten, Metadaten und Kommentaren aus Webseiten. Gut geeignet für Web-Corpus-Erstellung.	https://trafilatura.readthedocs.io/
jusText	Werkzeug zur Entfernung von Boilerplate-Inhalten wie Navigation, Werbung oder Footer aus HTML-Seiten.	https://github.com/miso-belica/jusText
readability-lxml	Implementierung eines Readability-Algorithmus zur Extraktion des zentralen Artikeltexts aus Webseiten.	https://github.com/buriy/python-readability

4. Machine Learning und Datenanalyse

Framework / Bibliothek	Kurzbeschreibung	Link
NumPy	Grundlage für numerisches Rechnen in Python. Bietet effiziente Arrays und mathematische Operationen.	https://numpy.org/
pandas	Standardbibliothek für Datenanalyse und Datenmanipulation mit tabellarischen Daten. Zentral für Preprocessing und explorative Datenanalyse.	https://pandas.pydata.org/
scikit-learn	Umfangreiche ML-Bibliothek für Klassifikation, Regression, Clustering, Feature Engineering, Modellselektion und Evaluation.	https://scikit-learn.org/
SciPy	Bibliothek für wissenschaftliches Rechnen, Statistik, Optimierung und lineare Algebra.	https://scipy.org/
PyTorch	Deep-Learning-Framework mit dynamischen Rechengraphen. Sehr verbreitet in Forschung und bei Transformer-/LLM-Anwendungen.	https://pytorch.org/
TensorFlow	Deep-Learning-Framework von Google für Training und Deployment neuronaler Netze.	https://www.tensorflow.org/
Keras	High-Level-API für neuronale Netze, heute eng mit TensorFlow integriert. Geeignet für schnelle Modellprototypen.	https://keras.io/
XGBoost	Effiziente Implementierung von Gradient Boosting. Sehr stark bei tabellarischen Daten und klassischen ML-Aufgaben.	https://xgboost.readthedocs.io/
LightGBM	Gradient-Boosting-Framework von Microsoft, optimiert für Geschwindigkeit und große Datensätze.	https://lightgbm.readthedocs.io/
imbalanced-learn	Erweiterung für scikit-learn zur Behandlung unausgeglichener Klassenverteilungen, z. B. durch Over- und Undersampling.	https://imbalanced-learn.org/
Optuna	Framework zur automatisierten Hyperparameteroptimierung. Unterstützt effiziente Suchverfahren und Pruning.	https://optuna.org/
MLflow	Plattform für Experiment Tracking, Modellverwaltung und Reproduzierbarkeit von ML-Workflows.	https://mlflow.org/

5. LLMs, Transformer und Hugging-Face-Ökosystem

Framework / Bibliothek	Kurzbeschreibung	Link
Transformers	Zentrale Hugging-Face-Bibliothek für Transformer-Modelle wie BERT, RoBERTa, GPT, T5, LLaMA und viele weitere. Unterstützt Training, Fine-Tuning und Inferenz.	https://huggingface.co/docs/transformers/

Hugging Face Hub	Plattform und Python-Bibliothek zum Laden, Teilen und Verwalten von Modellen, Datasets und Spaces.	https://huggingface.co/docs/huggingface_hub/
Datasets	Hugging-Face-Bibliothek zum Laden, Verarbeiten und Teilen großer Datensätze. Besonders nützlich für NLP- und ML-Pipelines.	https://huggingface.co/docs/datasets/
Tokenizers	Sehr schnelle Bibliothek für Tokenisierung, u. a. BPE, WordPiece und Unigram. Wird häufig mit Transformer-Modellen eingesetzt.	https://huggingface.co/docs/tokenizers/
Accelerate	Hugging-Face-Bibliothek zur Vereinfachung von Training und Inferenz auf CPU, GPU, Multi-GPU und TPU.	https://huggingface.co/docs/accelerate/
LangChain	Framework zur Entwicklung von LLM-Anwendungen, z. B. Chatbots, Agenten, Tool-Use und Retrieval-Augmented Generation.	https://www.langchain.com/
OpenAI Python SDK	Offizielle Python-Bibliothek zur Nutzung der OpenAI-API, z. B. für Chat Completion, Embeddings und Assistants.	https://github.com/openai/openai-python
Haystack	Framework für NLP- und LLM-basierte Suchsysteme, Question Answering und Retrieval-Augmented Generation.	https://haystack.deepset.ai/

6. Visualisierung und Evaluation

Framework / Bibliothek	**Kurzbeschreibung**	**Link**
Matplotlib	Standardbibliothek zur Erstellung statischer Diagramme und Visualisierungen in Python.	https://matplotlib.org/
Seaborn	Auf Matplotlib aufbauende Bibliothek für statistische Visualisierungen, z. B. Heatmaps, Verteilungen und Korrelationsplots.	https://seaborn.pydata.org/
Plotly	Bibliothek für interaktive Visualisierungen und Dashboards.	https://plotly.com/python/
Yellowbrick	Visualisierungsbibliothek für Machine-Learning-Modelle, z. B. Confusion Matrices, ROC-Kurven und Feature-Importances.	https://www.scikit-yb.org/
Evaluate	Hugging-Face-Bibliothek zur Berechnung standardisierter Evaluationsmetriken für ML- und NLP-Aufgaben.	https://huggingface.co/docs/evaluate/