Bericht zur Machbarkeitsstudie

# Identifizierung von Querschnittsthemen in Projekten der Direkten Projektförderung des BMBF

**Bastian Krieger, Christian Rammer, Patrick Breithaupt**
ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung

Mannheim, März 2020

ZEW

**Ansprechpartner**

Bastian Krieger

Forschungsbereich
Innovationsökonomik und
Unternehmensdynamik

L 7, 1 · 68161 Mannheim

Postfach 10 34 43
68034 Mannheim

E-Mail krieger@zew.de
Telefon +49 621-1235-376
Telefax +49 621-1235-170

# Inhalt

**Das Wichtigste in Kürze ..... 3**

**Executive Summary ..... 5**

**1 Aufgabenstellung und Zielsetzung ..... 7**

**2 Datenbasis ..... 9**

**3 TexAn-Textfeldanalyse ..... 11**

3.1 TexAn-Analyse zum Querschnittsthema Digitalisierung ..... 11

3.2 TexAn-Analyse zum Themenfeld Künstliche Intelligenz ..... 21

3.3 TexAn-Analyse zum Querschnittsthema Soziale Innovationen ..... 25

**4 Analyse mittels maschinellem Lernen zum Themenfeld Künstliche Intelligenz ..... 30**

**Fazit ..... 34**

**5 Anhang ..... 36**

5.1 Unterkategorien der Digitalisierung ..... 36

5.2 Code zur TexAn-Analyse ..... 37

**Abbildungsverzeichnis ..... 43**

**Tabellenverzeichnis ..... 44**

**Verzeichnis der Übersichten ..... 44**

**Verzeichnis der Boxen ..... 44**

## Das Wichtigste in Kürze

Die Machbarkeitsstudie untersuchte, inwieweit **Querschnittsthemen** der Forschungsförderung durch eine **semantische Analyse** von Vorhabenbeschreibungen mit hinreichender Genauigkeit identifiziert werden können. Es wurden zwei Querschnittsthemen betrachtet:

- **Digitalisierung** (inkl. des Teilgebiets **Künstliche Intelligenz**)
- **Soziale Innovationen**

Die semantische Analyse wurde mit Hilfe eines Textanalyseprogramms vorgenommen, das Wörter, Wortkombinationen sowie den Abstand zwischen diesen berücksichtigt. Die Ergebnisse wurden manuell überprüft und für Verbesserungen des Programms genutzt. Datenbasis waren Kurzbeschreibungen (bis zu 3.000 Wörter) zu 70.460 vom **BMBF** im Zeitraum **2005 bis 2018 geförderten Forschungs- und Entwicklungsvorhaben**.

Das Ergebnis zum Querschnittsthema **Digitalisierung** ist gemischt. Eine eindeutige Zuordnung von Vorhaben zu diesem Thema stellte sich als nicht praktikabel dar, da es viele Vorhaben in einem Grenzbereich gab und das Themenfeld insgesamt sehr breit gestreut ist. Es wurde daher eine restriktive und eine weniger restriktive Abgrenzung von Digitalisierung umgesetzt. Auf Basis der restriktiveren Abgrenzung wurden fast 18.000 Vorhaben mit bewilligten Fördermittel von insgesamt ca. 8,7 Mrd. EUR (d.h. durchschnittlich rund **1.300 Vorhaben und rund 620 Mio. EUR pro Jahr**) dem Themenfeld Digitalisierung zugeordnet. Bei einer breiteren Abgrenzung ergeben sich rund 23.400 Vorhaben und 11,3 Mrd. EUR Fördermittel (d.h. knapp 1.700 Vorhaben und rund 810 Mio. EUR pro Jahr). Der Anteil der vom BMBF geförderten Vorhaben, die dem Querschnittsthema Digitalisierung zugeordnet wurden, stieg für die restriktivere Abgrenzung **von 18 % (2005) auf 28 % (2018)** und für die weniger restriktive von 25 auf 38 %. Zum Vergleich: Im Förderbereich "Informations- und Kommunikationstechnologien" der Leistungsplansystematik des Bundes (der häufig herangezogen wird, um Förderaktivitäten im Bereich Digitalisierung zu erfassen) wurden pro Jahr weniger als 600 Vorhaben mit Fördermitteln von knapp 250 Mio. EUR gefördert. Auf Basis der **semantischen Analyse** ergibt sich somit schon bei einer restriktiven Abgrenzung eine **mehr als doppelt so hohe Förderaktivität** im Bereich Digitalisierung.

Vorhaben im Querschnittsthema Digitalisierung sind in **allen Förderbereichen** der Leistungsplansystematik anzutreffen. Besonders hoch ist ihr Anteil - neben dem Bereich Informations- und Kommunikationstechnologien - in den Förderbereichen "FuE zur Verbesserung der Arbeitsbedingungen und im Dienstleistungssektor", "Zivile Sicherheitsforschung", "Produktionstechnologien" und "Innovationsrelevante Rahmenbedingungen, Querschnittsaktivitäten". Auf den letztgenannten Förderbereich entfällt sogar eine größere Anzahl an Vorhaben, die dem Querschnittsthema Digitalisierung zugeordnet wur-

den, als auf den Bereich "Informations- und Kommunikationstechnologien". Dies unterstreicht die Bedeutung, beim Thema Digitalisierung über den klassischen IKT-Bereich hinauszugehen.

Für das Themenfeld **Künstliche Intelligenz** ergibt die semantische Analyse eine Gesamtzahl von mehr als **3.000 Vorhaben** im Zeitraum 2005 bis 2018 mit bewilligten Fördermitteln von **1,4 Mrd. EUR**. Die Höhe der jährlich bewilligten Fördermittel stieg von 20 bis 30 Mio. EUR im Zeitraum 2005-2007 auf ca. 300 Mio. EUR im Jahr 2018 kräftig an. Der Anteil der dem Themenfeld Künstliche Intelligenz zugeordneten Vorhaben an allen vom BMBF geförderten Vorhaben stieg von 1 % im Jahr 2007 auf fast 10 % im Jahr 2018 steil an. Vorhaben zum Thema Künstliche Intelligenz finden sich in fast allen Förderbereichen, besonders häufig in den Bereichen "Innovationsrelevante Rahmenbedingungen, Querschnittsaktivitäten", "Informations- und Kommunikationstechnologien", "Produktionstechnologien", Gesundheitsforschung, Gesundheitswirtschaft" und "FuE zur Verbesserung der Arbeitsbedingungen und im Dienstleistungssektor".

Das Ergebnis zum Querschnittsthema **Soziale Innovationen** erbrachte erheblich geringere Fallzahlen. Für den Zeitraum 2005-2018 konnten insgesamt 127 Vorhaben mit Fördermitteln von zusammen 101 Mio. EUR identifiziert werden. Es zeigt sich eine ansteigende Tendenz der geförderten Vorhaben in diesem Themenfeld. Die meisten Vorhaben zu Sozialen Innovationen finden sich in den Förderbereichen "Innovationsrelevante Rahmenbedingungen, Querschnittsaktivitäten" sowie "Klima, Umwelt, Nachhaltigkeit", gefolgt von "FuE zur Verbesserung der Arbeitsbedingungen und im Dienstleistungssektor" sowie "Geistes-, Wirtschafts- und Sozialwissenschaften".

Zusätzlich zur semantischen Analyse wurde für das Teilgebiet Künstliche Intelligenz auch ein **Ansatz des Maschinellen Lernens** getestet. Dieser Ansatz wäre deutlich weniger arbeitsintensiv als semantische Analysen. Hierfür wurde ein neuronales Netzwerk mit einem Trainingsdatensatz trainiert, der die über die semantische Analyse dem Themenfeld Künstliche Intelligenz zugeordneten Vorhaben enthält. Ein Teil dieser Vorhaben wurde dabei nicht als Trainingsdaten, sondern zur Evaluation der Zuverlässigkeit des Ansatzes genutzt. Das **Ergebnis ist nicht zufriedenstellend**, da nur 68 % der über die semantische Analyse dem Thema Künstliche Intelligenz zugeordneten Vorhaben auch über das maschinelle Lernen zugeordnet werden, während 22 % der über den maschinellen Ansatz als Künstliche Intelligenz klassifizierten Vorhaben keine solchen sind. Ansätze des maschinellen Lernens sind somit nicht gut geeignet, um Vorhaben auf Basis von Kurzbeschreibungen automatisiert Querschnittsthemen zuzuordnen. Dies liegt u.a. wohl an den relativen kurzen Texten der Abstracts, einem relativ hohen Standardisierungsgrad der Abstracts und einer zu niedrigen Anzahl von Testdatensätzen, um die Vorteile von maschinellem Lernen effizient nutzen zu können.

## Executive Summary

This feasibility study explores the possibility of using **semantic analysis** of abstracts of research projects to assign projects to **cross-cutting areas** with a reasonable degree of accuracy. The study focuses on two areas:

- **Digitalisation** (incl. the sub-areas of **Artificial Intelligence**)
- **Social Innovation**

The semantic analysis is performed using a text analyses programme which considers words and word combinations, and the distance between them. Results are manually checked and used to improve the programme code. The analysis rests on extended abstracts (up to 3,000 words) from 70,460 **R&D projects** funded by the **Federal Ministry of Education and Research (BMBF)** in the time period **2005 to 2018**.

The results for the area **digitalisation** are mixed. It was not possible to unambiguously assign projects to this area, as there were too many border-line cases, and delineating the digitalisation is difficult given the complexity of technologies and applications. For that reason, a narrow and a broad demarcation of digitalisation was applied. Based on the narrow approach, almost 18,000 projects with a total amount of public funding of €8.7b have been assigned to the area of digitalisation (which is equal to about **1,300 projects and €620m of funds per year**). The broader approach yields appr. 23,400 projects and €11.3b of public money (almost 1,700 projects and €810m per year on average). The share of projects in the area of digitalisation in the total number of R&D projects funded by BMBF raised from **18% (2005) to 28% (2018)** based on the narrow approach, and from 25% to 38% based on the broader approach. These figures are **more than twice as high** as compared to the results one obtains when analysing the funding area "information and communication technologies" of the Federal Government's taxonomy for assigning funding activities to research and technology fields (*Leistungsplansystematik* - LPS). In this area, the annual average number of funded projects is about 600 projects and the amount of funding about €250m per year.

R&D projects assigned to the area of digitalisation can be found in **all research and technology fields** of the LPS. High shares of projects related to digitalisation are reported for the fields "information and communication technologies", "R&D to improve working conditions / R&D in services", "civil security research", "production technologies" and "innovation-related framework conditions, cross-cutting activities". The latter field comprises a higher absolute number of projects in the area of digitalisation as compared to the area of "information and communication technologies", stressing the importance of taking a broader view when looking on government funding activities related to digitalisation.

For the area of **Artificial Intelligence** (AI), the study identified over **3,000 projects** funded during 2005 and 2018 with about **€1.4b of public money**. The amount of public funding

per years steadily increased from 20 to 30 million in the years 2005 to 2007 to almost €300m in 2018. The share of projects assigned to AI in the total number of R&D projects funded by the BMBF climbed from 1% in 2007 to almost 10% in 2018. Projects related to AI can be found in almost all research and technology fields. The largest numbers of projects are in the areas "innovation-related framework conditions, cross-cutting activities", "information and communication technologies", "production technologies", "health research and health sector" and "R&D to improve working conditions / R&D in services".

In the area of **social innovation**, only a small number of projects were found. During 2005 and 2018, a total of 127 projects that received €101m of public funding have been identified. There is a clear upwards trend in projects related to social innovation over time. Most projects are found in the research areas "innovation-related framework conditions, cross-cutting activities" and "climate, environment, sustainability", followed by "R&D to improve working conditions / R&D in services" and "humanities, economics and social sciences".

In addition to the semantic analysis, a **machine learning approach** was applied in the area of artificial intelligence in order to examine the feasibility of a more automated way of identifying cross-cutting areas. A neural network was trained based on data from the projects that have been assigned to AI through the semantic analysis, setting aside a fraction of these project for evaluating the accuracy of the machine learning approach. The **result is unsatisfactory**. The neural network assigned only 68% of all AI projects correctly as AI, whereas 22% of all projects assigned to AI by the neural network were not classified as AI by the semantic analysis. Machine learning approaches do not seem to be useful for this particular problem of assigned research projects to cross-cutting areas based on project abstracts. Among others, extended abstracts represent rather short and relatively standardised texts while the number of observations is too small for an efficient use of machine learning tools.

# 1 Aufgabenstellung und Zielsetzung

Bei den Nutzern des Informationssystems PROFI - Fördermittelgeber, Politik und allgemeine Öffentlichkeit - besteht immer wieder der Bedarf, Aussagen über die FuE-Fördertätigkeit des Bundes in neu aufkommenden Themenfeldern oder in Themenfeldern, die über die Grenzen einzelner Forschungsbereiche und Technologien hinausgehen, treffen zu können. Allein mit Hilfe der Leistungsplansystematik[1] ist dies für solche *Querschnittsthemen* meist nicht möglich. Ein Beispiel hierfür ist das Thema Digitalisierung und Verfahren der Künstlichen Intelligenz. Die Digitalisierung spielt in verschiedensten Themenfeldern eine wichtige Rolle, zu denen nicht nur die Informations- und Kommunikationstechnologien zählen, sondern auch Produktionstechnologien, Fahrzeug- und Verkehrstechnologien sowie die Bereiche Klima, Umwelt, Nachhaltigkeit und Gesundheitsforschung, Gesundheitswirtschaft. Künstliche Intelligenz unterstützt Ärzte bei ihren Diagnosen, Smart-Meter managen den Stromverbrauch von Haushalten und Supercomputer ermöglichen eine schnellere Verarbeitung jeglicher Informationen. Die Anzahl der Anwendungen ist groß und eine klare Zuordnung der Digitalisierung zu einzelnen Schwerpunkten der Leistungsplansystematik nicht möglich. Auch lässt die hohe Geschwindigkeit der technologischen Entwicklung im Bereich Digitalisierung keine langfristige Anpassung der Leistungsplansystematik zu. Themen die heute noch eine wichtige Rolle spielen, können in wenigen Jahren bereits wieder redundant sein, während völlig neue Themen hinzugekommen sind.

Die vorliegende Machbarkeitsstudie prüft, inwieweit solche Querschnittsthemen durch eine semantische Analyse der Beschreibungen geförderter FuE-Vorhaben mit hinreichender Genauigkeit identifiziert werden können und welcher Aufwand notwendig ist, um eine solche Analyse umzusetzen.

Die Machbarkeitsstudie umfasst folgende Schritte:

- Grundlage der Textanalyse bilden die Abstracts der seit **2005 vom BMBF geförderten Vorhaben**. Die vom BMBF bereitgestellten Daten werden aufbereitet und durch weitere Informationen aus der PROFI-Datenbank (siehe Box 1) ergänzt, insbesondere durch Informationen über die Leistungsplansystematik, die Finanzierungszeiträume sowie die Förderdauern und Bewilligungssummen der einzelnen Vorhaben.
- Mit Hilfe eines Textanalyseprogramms wird nach Schlagwörtern und Schlagwortkombinationen gesucht (**semantische Analyse**), die auf ein Querschnittsthema hinweisen, oder Hinweise darauf geben, dass trotz Vorliegens eines Schlagworts

[1] Die Leistungsplansystematik ist eine Systematik zur Zuordnung von FuE-Förderungen zu Forschungsthemen, vgl. Abschnitt 2.

das Projekt nicht in das entsprechende Themenfeld gehört. Es werden drei Themenfelder betrachtet:

- *Abschnitt 3.1:* Erstes Querschnittsthema ist die **Digitalisierung**. Ausgangspunkt bildet eine existierende Analyse zu digitalen Geschäftsmodellen. Diese wurde vom ZEW erstellt und anhand von Geschäftstätigkeitsbeschreibungen junger Unternehmen kalibriert. Innerhalb dieser Studie wird die Anwendbarkeit der Analyse auf Vorhabenbeschreibungen überprüft und weiterentwickelt.
- *Abschnitt 3.2:* Innerhalb des Querschnittsthemas "Digitalisierung" wird die Kategorie "**Künstliche Intelligenz**" gesondert betrachtet. Hierfür wird ein eigenes Textfeldanalyseprogramm entwickelt und implementiert.
- *Abschnitt 3.3:* Als zweites Querschnittsthema wird "**Soziale Innovationen**" untersucht. Hierfür wird ein neues Textfeldanalyseprogramm entwickelt und implementiert.

- Zuletzt wird ein weiteres Analysetool basierend auf Methoden des **Maschinellen Lernens** getestet. Dieses Analysetool hat den Vorteil, dass die Zuordnung von Vorhaben zu Querschnittsthemen bzw. Themenfeldern automatisiert vorgenommen werden kann, was den Bearbeitungsaufwand im Vergleich zu einer semantischen Analyse deutlich reduziert. Hierzu wird auf die Ergebnisse aus Abschnitt 3.2 zu Förderungen im Bereich Künstliche Intelligenz zurückgegriffen. Es wird getestet, inwiefern ein trainierter, automatisierter Algorithmus geeignet ist, um dieses Themenfeld in den Abstracts von geförderten Vorhaben zu identifizieren.

# 2 Datenbasis

Datengrundlage der Machbarkeitsstudie sind alle vom BMBF geförderten Forschungs- und Entwicklungsvorhaben (im Folgenden kurz: Vorhaben), die in den Jahren 2005 bis 2018 im Rahmen der Direkten Projektförderung bewilligt wurden. Dabei handelt es sich um 70.460 Vorhaben. Die Fördersumme (bewilligte Mittel) für diese Vorhaben beläuft sich auf ca. 63,8 Mrd. EUR.

Zu jedem Vorhaben wurden dem ZEW vom BMBF Titel und Kurzbeschreibung zur Verfügung gestellt. Diese Informationen wurden um weitere Informationen aus der PROFI-Datenbank (siehe Box 1), die dem ZEW im Rahmen des Projekts „Monitoring der Beteiligung von KMU an der Direkten Projektförderung“ vorliegen, ergänzt. Dazu zählen u.a. die Zuordnung zur Leistungsplansystematik, die Bewilligungssumme und der Vorhabenzeitraum.

**Box 1: PROFI-Datenbank**

Die Datenbank "Projektförder-Informationssystem" (PROFI) ist ein Instrument für die Abwicklung von Zuwendungen und Aufträgen durch den Bund. Die Datenbank enthält Informationen zu den einzelnen geförderten oder beauftragten Vorhaben. Zu geförderten FuE-Vorhaben liegt u.a. eine Kurzbeschreibung des Vorhabens (Abstract) vor. Der Text wird von den Zuwendungsempfängern erstellt und mit der Antragstellung vorgelegt. Jedes Vorhaben ist außerdem einer Nummer der Leistungsplansystematik (siehe Box 2) zugeordnet.

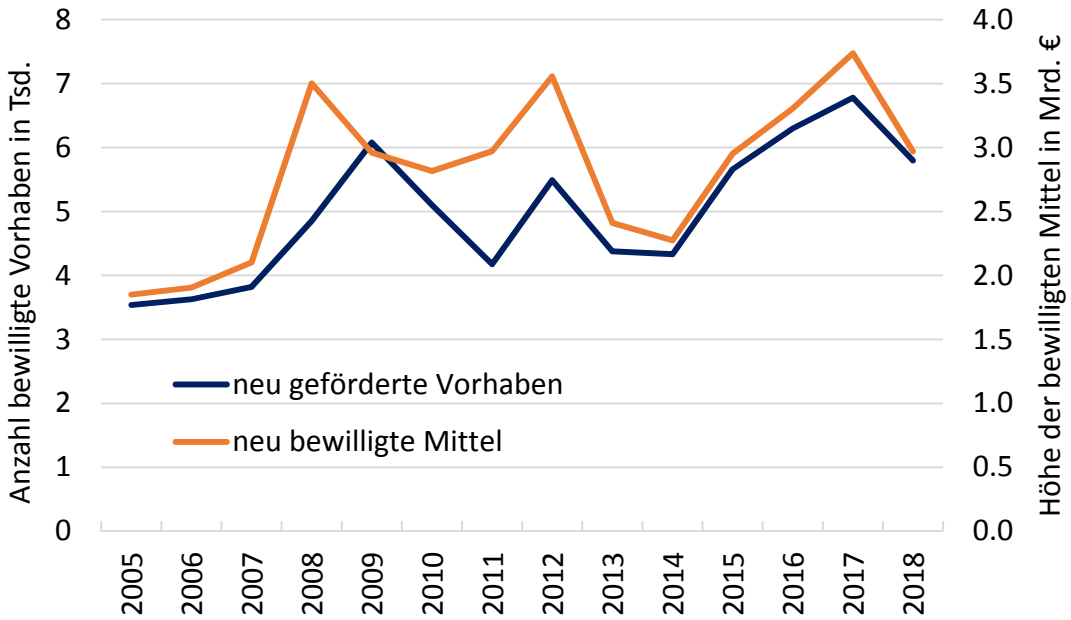
Abbildung 1 zeigt die Anzahl der pro Jahr neu bewilligten Vorhaben und die Höhe der bewilligten Mittel. Sowohl die Anzahl der Vorhaben wie die Bewilligungssumme nahmen im betrachteten Zeitraum zu. Die Anzahl der bewilligten Vorhaben stieg von 3.537 im Jahr 2005 auf 5.798 im Jahr 2018, die Höhe der bewilligten Mittel nahm von rund 1,9 Mrd. EUR im Jahr 2005 auf 3,0 Mrd. EUR im Jahr 2018 zu.[2]

Die verwendete Datengrundlage deckt nicht alle Förderbereiche der Leistungsplansystematik (siehe Box 2) ab, da die Projektförderung des BMBF nicht alle Förderbereiche umfasst. Nicht abgedeckt sind in dieser Studie die Förderbereiche D (Ernährung, Landwirtschaft, Verbraucherschutz), H (Fahrzeug- und Verkehrstechnologien inkl. maritimer Technologien), I (Luft- und Raumfahrt) und Q (Innovationsförderung des Mittelstands). Für andere Förderbereiche wie z.B. E (Energieforschung und Energietechnologien) und

[2] Diese Zahlen enthalten nicht den Förderbereich T "Förderorganisationen, hochschulbezogene Sonderprogramme".

N (Raumordnung und Stadtentwicklung, Bauforschung) repräsentieren die in diese Studie einbezogenen Vorhaben nur einen kleineren Teil der gesamten Förderaktivitäten des Bundes.

**Abbildung 1: Anzahl und bewilligte Mittel vom BMBF geförderter Vorhaben* 2005-2018**




**ohne Leistungsplanbereich T.*
*Quelle: PROFI-Datenbank, Berechnungen des ZEW.*

**Box 2: Leistungsplansystematik**

Die Leistungsplansystematik des Bundes gruppiert die Forschungsausgaben des Bundes nach forschungsthematischen Gesichtspunkten. Sie unterscheidet dabei übergeordnete Forschungsbereiche (Förderbereiche), die in Forschungsschwerpunkte (Förderschwerpunkte) und weiter in Unterklassen unterteilt sind. Mit der Leistungsplansystematik werden die FuE-Ausgaben des Bundes unabhängig vom finanzierenden Ressort einzelnen Forschungsthemen zugeordnet. Jedes geförderte Vorhaben wird dabei einer Leistungsplan-Nummer zugeordnet. Da die Zuordnung nach dem Schwerpunktprinzip erfolgt, ist eine Mehrfachzuordnung zu verschiedenen Förderbereichen nicht möglich. Dies kann insbesondere bei interdisziplinär ausgerichteten Vorhaben zu Unschärfen führen. Zudem sind Querschnittsthemen kaum über die Leistungsplansystematik abbildbar.[3]

[3] Vgl. BMBF (2018), Bundesbericht Forschung und Innovation 2018. Datenband, S. 146. Berlin.

# 3 TexAn-Textfeldanalyse

Die in dieser Studie durchgeführten semantischen Analysen ("Textfeldanalysen") nutzen eine vom ZEW entwickelte Software, die die Bezeichnung "TexAn – Text Analyser" führt. Diese Software erlaubt es, Texte nach Schlagwörtern, Wortteilen und Schlagwortkombinationen (unter Berücksichtigung der Distanz zwischen Wörtern sowie positiver und negativer Beziehungen) zu durchsuchen und zu vom Anwender definierten Klassen zuzuordnen. Im Rahmen dieser Machbarkeitsstudie wird der aus Titeln und Abstracts eines Vorhabens gebildete Textkorpus klassifiziert.

Innerhalb der Textanalyse ist es wichtig, die Anzahl an fehlerhaften Klassifizierungen zu minimieren. Zum einen sollte es möglichst wenig "False-Negatives" geben, sprich möglichst wenig Vorhaben sollen beispielsweise als nicht-digital eingestuft werden, obwohl sie eigentlich dem Bereich Digitalisierung zuzurechnen sind. Zum anderen sollten auch möglichst wenig "False-Positives" entstehen, sprich möglichst wenig Vorhaben sollten als digital eingestuft werden, obwohl sie es nicht sind. In Bezug auf die Machbarkeitsstudie würde eine zu große Menge von False-Negatives eine Unterschätzung der Förderaktivitäten bedeuten, eine zu große Menge von False-Positives eine Überschätzung. Demnach ist die Klassifizierung eines Querschnittsthemas mit Hilfe der Textfeldanalyse nur machbar, wenn eine hinreichend niedrige Anzahl von beiden Arten von Fehlern erreicht wird. Das Austarieren beider Arten von Fehlern und die wiederholte Anpassung der Textfeldanalyse ist daher ein Schlüsselelement des Projekts. Die entwickelten konkreten TexAn-Analysecodes zu den Querschnittsthemen Digitalisierung, künstliche Intelligenz und Soziale Innovationen befinden sich im Anhang.

## 3.1 TexAn-Analyse zum Querschnittsthema Digitalisierung

Für die Textfeldanalyse zum Querschnittsthema Digitalisierung wurde zunächst untersucht, ob die zuvor vom ZEW durchgeführte semantische Analyse zu digitalen Geschäftsmodellen junger Unternehmen[4] geeignet ist, um geförderte Vorhaben des BMBF im Themenfeld Digitalisierung zu erkennen. Hierzu betrachten wir den Anteil der mit dieser Analyse dem Querschnittsthema Digitalisierung zugeordneten Vorhaben innerhalb einer Zusammenstellung von 4.351 Vorhaben aus verschiedenen Leistungsplanklassen mit einem besonders hohen Potenzial, digitale Themen zu beinhalten (siehe Tabelle 1). Identifiziert wurde diese Liste an Leistungsplanklassen, durch einen manuellen Themenabgleich der Leistungsplanklassen anhand ihres Titels, mit den Themen der digitalen Geschäftsfeldanalyse des ZEWs. In Leistungsplanklassen mit einem auffällig geringen Anteil

[4] Dabei wurden die Geschäftstätigkeitsbeschreibungen von Unternehmen, die in der Datenbank des Mannheimer Unternehmenspanels vorlagen, analysiert. Geschäftstätigkeitsbeschreibungen stellen relativ stark standardisierte Texte dar, in denen die zentralen Tätigkeiten und Marktangebote eines Unternehmens beschrieben werden.

wurde eine Stichprobe von potenziell inkorrekten negativ klassifizierten Vorhaben (False-Negatives) gezogen und manuell überprüft. Ziel der Überprüfung war es, notwendige Anpassungen in der Textfeldanalyse zu identifizieren. Anschließend wurden diese Anpassungen vorgenommen und eine erneute Analyse durchgeführt. Dieses Vorgehen wurde so lange wiederholt, bis die Anzahl negativer Fehlklassifikationen sehr gering war.

Im Anschluss wurde die auf diese Weise erstellte Textfeldanalyse auf alle vorliegenden 70.460 Vorhaben angewendet. An dieser Stelle war es notwendig, eine Überprüfung auf inkorrekt positiv klassifizierte Vorhaben (False-Positives) durchzuführen. Hierzu wurde eine Stichprobe von als "digital" klassifizierten Vorhaben aus den verschiedenen Leistungsplanklassen manuell begutachtet und die Textfeldanalyse auch hier im Falle fehlerhafter Klassifikationen angepasst. Anschließend wurde die veränderte Textfeldanalyse erneut durchgeführt und auf die gleiche Weise erneut auf inkorrekte positive Klassifikationen untersucht. Auch wurde an dieser Stelle die Veränderung des Anteils der dem Querschnittsthema Digitalisierung zugeordneten Vorhaben in den Leistungsplanklassen aus Tabelle 1 nochmals überprüft, um sicherzugehen, dass durch die Anpassungen der Analyse keine bedeutende Anzahl an inkorrekten negativen Klassifikationen entstanden ist. Dieses Vorgehen wurde iterativ wiederholt, bis die Anzahl von False-Negatives und False-Positives sehr gering war.

Die semantische Analyse auf Basis des ursprünglich vom ZEW entwickelten Textanalyseprogramms zum Querschnittsthema Digitalisierung (Analyse der Geschäftstätigkeitsbeschreibungen von Startups) erwies sich als nicht geeignet für den Textkorpus der BMBF-geförderten Vorhaben. Viele Vorhaben der Leistungsplanklassen aus Tabelle 1 wurden nicht als "digital" erkannt. Dies lag daran, dass die ursprünglich verwendeten zehn Unterkategorien zur Beschreibung des Querschnittsthemas "Digitalisierung" zu eng abgegrenzt waren und wesentliche Digitalisierungsfelder nicht erfasst haben. Sie wurden daher innerhalb mehrerer Wiederholungsrunden in ihren Stichwortkombinationen erweitert und um 13 weitere Unterkategorien ergänzt.[5]

Die Unterkategorie zur Künstlichen Intelligenz umfasste beispielweise ursprünglich überwiegend verschiedene Schreibweisen des Begriffs "künstliche Intelligenz". Sie wurde unter anderem durch Begriffe zu maschinellem Lernen und neuronalen Netzten ergänzt. Hinzugefügte Unterkategorien zur Digitalisierung sind beispielsweise Cybersicherheit, Internet der Dinge sowie intelligente Produkte und Dienstleistungen. Als besondere Herausforderung des Querschnittsthemas Digitalisierung zeigte sich bereits hier seine große Themenbereite, welche ein besonders zeitaufwendiges Prüfen der als nicht-digital klassifizierten Vorhaben nötig machte, da jede Unterkategorie der Digitalisierung ihre eigenen Schlagwortkombinationen erfordert.

[5] Eine finale Liste der genutzten Unterkategorien der Digitalisierung findet sich in Tabelle 4 im Anhang.

**Tabelle 1: Leistungsplanklassen mit hohem Potenzial für das Querschnittsthema Digitalisierung**

| LP-ID | LP-Name |
|---|---|
| AA0240 | Computational Neuroscience |
| FA3061 | Datenmanagement |
| FA5060 | Instrumente, Methoden sowie Plattformen und Netzwerke |
| GA1010 | Entwicklung von Softwaremethoden und -Werkzeugen |
| GA1011 | Eingebettete Systeme |
| GA1012 | Integrierte Anwendungssysteme |
| GA1040 | Korrektheit und Redundanz bei Informationssystemen |
| GA1050 | Manipulationssicherheit von Informationssystemen |
| GA1060 | Sicherheit in DV-Netzen |
| GA1080 | Sonstiges im Rahmen der Softwaretechnologie |
| GA2010 | Parallelarchitekturen |
| GA2020 | Parallelsoftware |
| GA2030 | Mathematische Grundlagen der wissenschaftlichen Computeranwendungen |
| GA2040 | Modellierung / Simulation |
| GA2060 | GRID |
| GA2080 | Sonstiges im Rahmen des Höchstleistungsrechnens |
| GA4010 | Neuronale Netze und ihre Anwendungen |
| GA4030 | Erkennen und Verstehen von Schrift und Bildern |
| GA4040 | Wissensverarbeitung/Expertensysteme |
| GA4080 | Sonstiges im Rahmen der intelligenten Systeme |
| GA5010 | Erkennen, Verstehen und Übersetzen von Sprache |
| GA5020 | Intelligente Methoden der Mensch-Maschine-Kommunikation |
| GA5030 | Virtuelle Realität / Erweiterte Realität |
| GA5080 | Sprachtechnologie und Mensch-Maschine-Kommunikation |
| GA9010 | Analysen, Prognosen und Auswertungen Informatik |
| GA9020 | Internationale Zusammenarbeit im Rahmen der Informationsverarbeitung |
| GA9081 | DV-Systeme und -Technologien (abgeschl. DV-Progr.) |
| GA9099 | Sonstiges (auch Normung) im Rahmen der Informatik |
| GB1010 | Arbeitsprogramm IT-Sicherheit |
| GB1011 | Sicheres Cloud-Computing |
| GB1012 | IT-Sicherheit in Kritischen Infrastrukturen |
| GB1013 | Hightech für die IT-Sicherheit |
| GB1070 | Quanteninformationstechnologie |
| GB1080 | Privatheit in der digitalen Welt |
| GB1099 | Sonstiges im Rahmen der IT-Sicherheit |
| GB2010 | Netzbasierte Dienste in der Medizin |
| GB2011 | Netzbasierte Dienste im Verkehr |
| GB2099 | Sonstiges im Rahmen der Netzbasierten Dienste |
| GB8040 | Internettechnologien |
| GB9099 | Sonstiges im Rahmen der Kommunikationstechnologie (einschl. Querschnittsunters.) |
| GC2020 | Aufbau- u. Verbindungstechnik, 3 D - Integration |
| GC2025 | Chipbasierte Sicherheit für die Digitalisierung |

*Quelle: Zusammenstellung des ZEW*

Die Anwendung der auf diese Art erstellten Textfeldanalyse auf alle vom BMBF neu geförderten Vorhaben im Zeitraum von 2005 bis 2018 führte in ihren ersten Runde zu einer hohen Anzahl von inkorrekt als digital klassifizierten Vorhaben. Die Unterkategorie Künstliche Intelligenz klassifizierte beispielsweise fälschlicherweise Vorhaben im Bereich der Neurobiologie ohne Verknüpfung zur künstlichen Intelligenz als digital. Weitere Beispiele sind die Unterkategorien selbstfahrende Fahrzeuge, integrierten Systeme oder automatisches Erkennen von Bildern und Tönen. Hier lösten Wörter wie „Verfahren“ oder „Anerkennung“ fälschlich positive Klassifikationen aus, da sie die Wortbestandteile „fahren“ und „erkenn“ beinhalteten. Auch wurden Vorhaben noch anhand verschiedener Arten integrierter, aber nicht notwendigerweise digitaler, Systeme, wie Ökosystemen oder Wassernetzwerke, als digital eingestuft. Zur Vermeidung dieser Fehlklassifikationen wurden Nebenbedingungen zu den entsprechenden Unterkategorien hinzugefügt. Vereinfacht dargestellt mussten etwa für eine Klassifikation als digital in der Unterkategorie Künstliche Intelligenz neben den Worten „neuronal“ und „Netz“ auch weitere Worte wie „Algorithmus“, „selbstlernend“ oder „automatisiert“ innerhalb eines definierten Wortabstandes im Textfeld vorkommen.

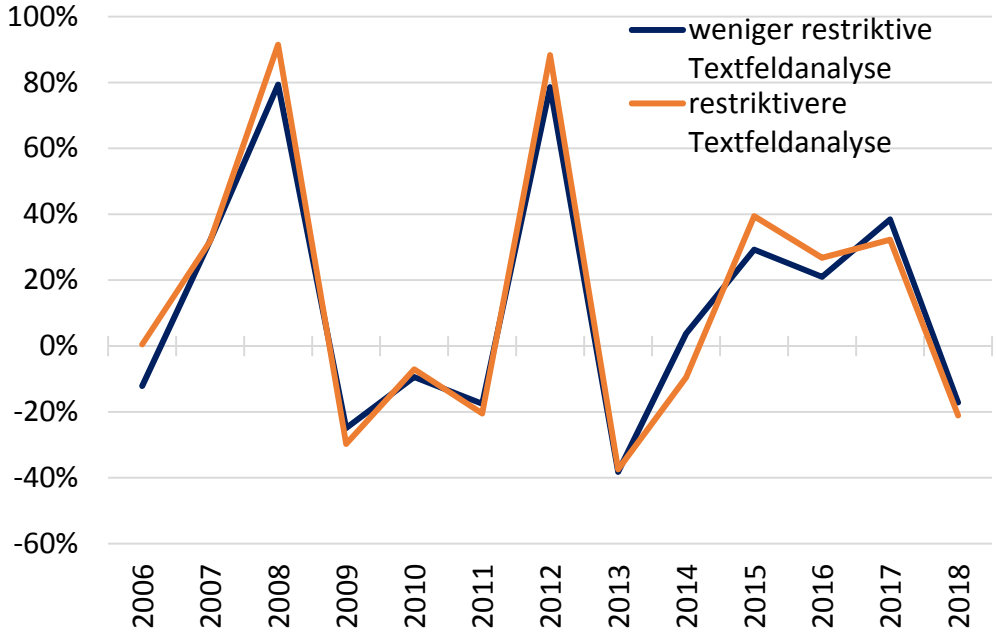
Das Ergebnis der wiederholten Anpassung der TexAn-Analyse sind zwei alternative Textfeldanalysen. Eine Textfeldanalyse ist restriktiver und verkleinert die Wahrscheinlichkeit von inkorrekt positiven Klassifikationen, die andere ist weniger restriktiv und verkleinert die Wahrscheinlichkeit inkorrekt negativer Klassifikationen. Hauptunterschied zwischen den beiden Varianten sind vier Unterkategorien, die nicht in der restriktiven, aber in der weniger restriktiven Analyse enthalten sind.[6] In der weniger restriktiven Variante sind z.B. simple Schlagwörter wie „digital“, „IKT“ und „Internet“ ohne weitere Restriktionen enthalten. Zum anderen umfasst die weniger restriktive Variante die Unterkategorie "integrierte Systeme", da diese bis zuletzt relativ fehleranfällig für False-Positives geblieben ist. Der Grund für das Erstellen von zwei unterschiedlichen Textfeldanalysen zur Digitalisierung ist erneut der Breite des Themas geschuldet. Diese erschwert eine eindeutige Abgrenzung von Vorhaben als "digital" und "nicht-digital". Deshalb empfehlen wir, zwei Varianten zu verwenden und die Ergebnisse als Ober- und Untergrenzen zu interpretieren.

Die restriktivere Analyse identifiziert 84 % der Vorhaben in den Leistungsplanklassen mit hohem Digitalisierungspotenzial als "digital". Angewendet auf alle Leistungsplanbereiche werden 17.923 Vorhaben des Bewilligungszeitraums 2005-2018 dem Querschnittsthema Digitalisierung zugeordnet. Die weniger restriktive Variante identifiziert 91 % der Vorhaben in den Leistungsplanklassen mit hohem Digitalisierungspotenzial als "digital". und weist insgesamt 23.388 Vorhaben dem Querschnittsthema zu. Dies ent-

[6] Die finale Liste der genutzten Unterkategorien in Tabelle 4 im Anhang zeigt, welche Unterkategorien in welcher Analyse genutzt wurden.

spricht einer Bewilligungssumme von 8,7 Mrd. EUR (restriktiv) und 11,3 Mrd. EUR (weniger restriktiv) im Querschnittsthema Digitalisierung und demonstriert einen merklichen Niveauunterschied zwischen den beiden Analysen. Abbildung 2 zeigt allerdings, dass das prozentuale Wachstum der jährlichen neuen Bewilligungssummen im Zeitraum von 2006 bis 2018 weitestgehend gleichverläuft. Demnach unterscheiden sich die Analysen zwar in ihren Niveaus, folgen aber der gleichen Entwicklung.

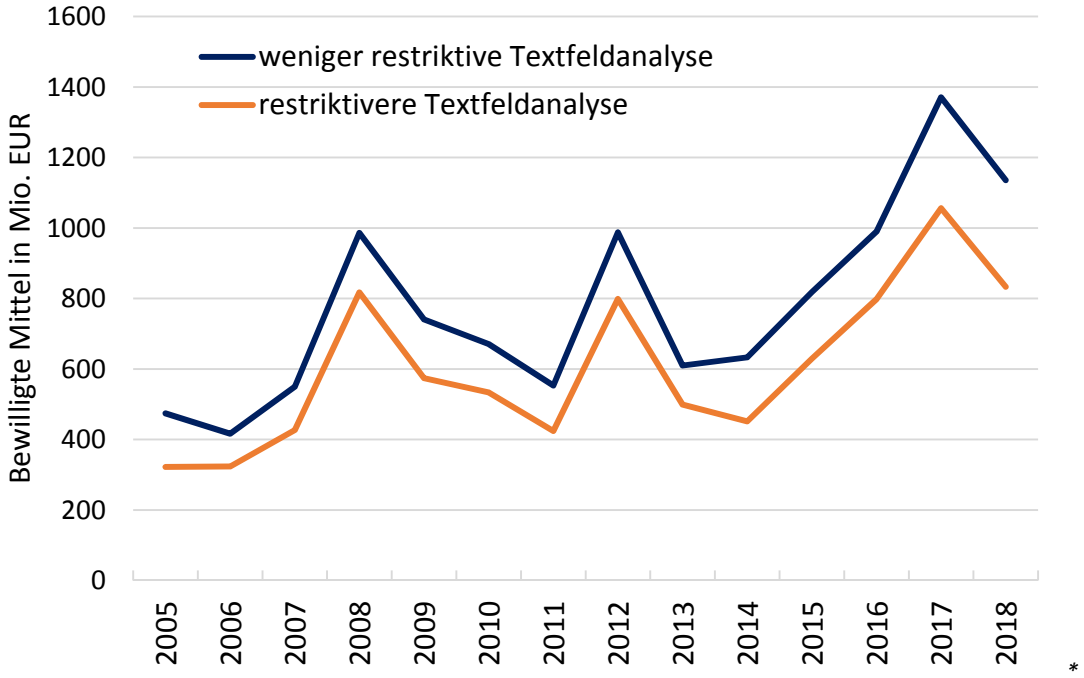
**Abbildung 2: Jährliches Wachstum der bewilligten Mittel von neu geförderten Vorhaben* zum Thema Digitalisierung 2006-2018**




** ohne Leistungsplanbereich T*
*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

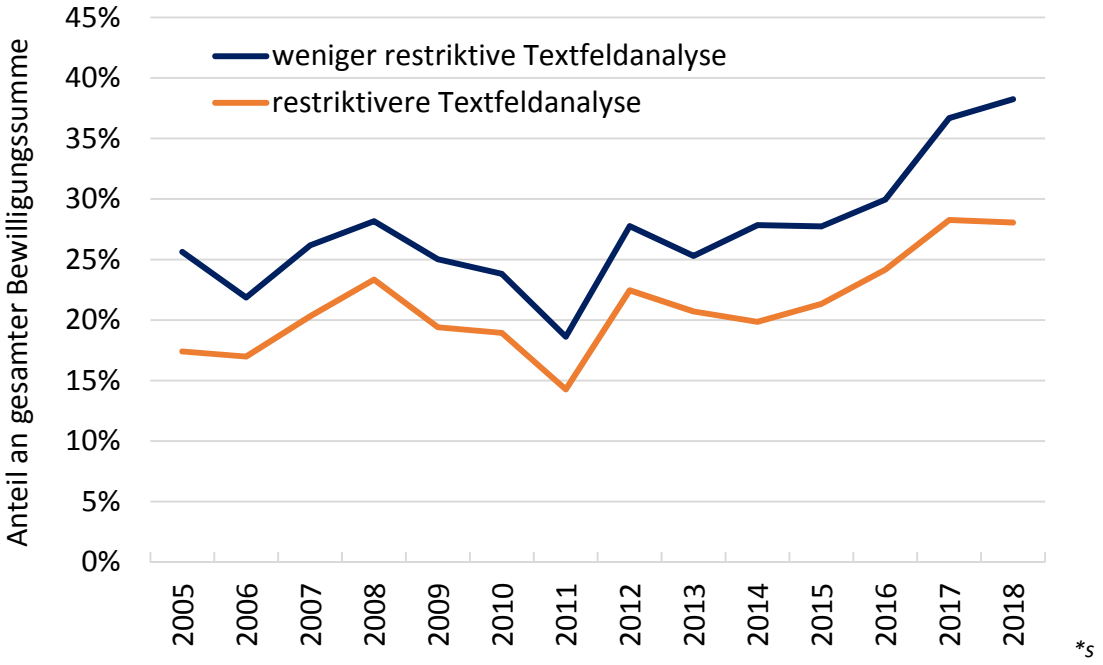
Abbildung 3 und Abbildung 4 zeigen die Höhe der bewilligten Mittel im Querschnittsthema Digitalisierung in den einzelnen Beobachtungsjahren sowie den Anteil dieser Mittel an der Gesamtbewilligungssumme des BMBF pro Jahr. Beide Grafiken zeigen einen ansteigenden Trend beider Werte und verlaufen zwar auf unterschiedlichen Niveaus, aber parallel. Im Durchschnitt identifiziert die weniger restriktive Analyse einen um 7 %-Punkte höheren Anteil (absolut: ca. 175 Mio. EUR mehr an Bewilligungen pro Jahr). Absolut betrachtet stiegt die jährliche Bewilligungssumme von Vorhaben im Bereich Digitalisierung zwischen 2005 und 2018 von 322 Mio. auf 1,06 Mrd. EUR (bei weniger restriktiver Abgrenzung: von 474 Mio. auf 1,14 Mrd. EUR). Der Anteil erhöhte sich von 17 % auf 28 %, beziehungsweise von 26 % auf 38 %.

**Abbildung 3: Bewilligte Mittel von neu geförderten Vorhaben* zum Querschnittsthema Digitalisierung 2005-2018**




*ohne Leistungsplanbereiche T*
*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

**Abbildung 4: Anteil bewilligter Mittel von neu geförderten Vorhaben* zum Querschnittsthema Digitalisierung am gesamten Bewilligungsvolumen des BMBF 2005-2018**




*ohne Leistungsplanbereiche T*
*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

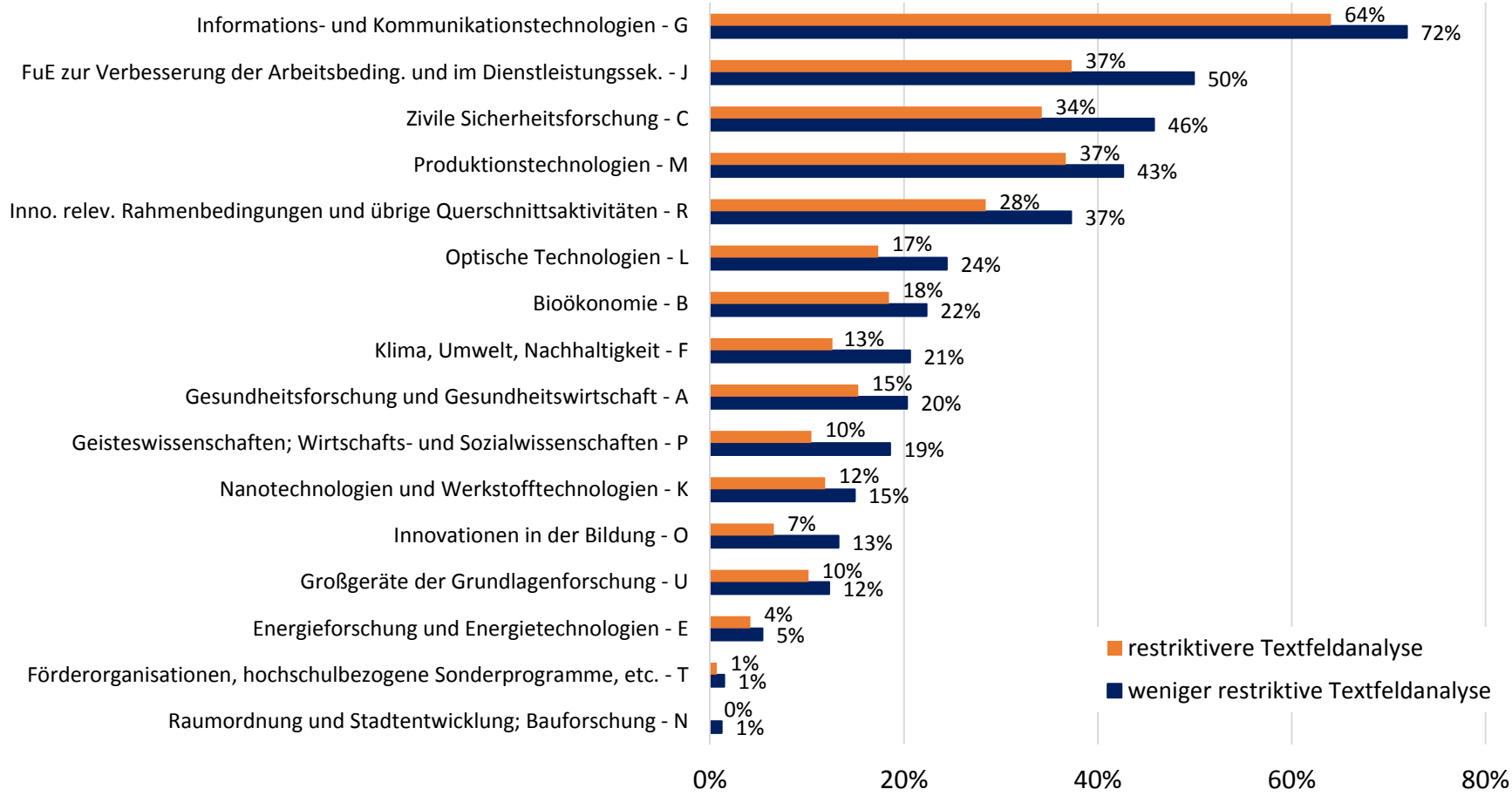
Abbildung 5 verdeutlicht außerdem eine ähnliche Rangordnung der Leistungsplanbereiche in ihrer Digitalisierungsintensität in beiden Analysen. Auch plausibilisiert die Digitalisierungsintensität, gemessen als Anteil der bewilligten Mittel im Themenbereich Digitalisierung innerhalb eines Leistungsplanbereichs, die Ergebnisse unsere TeXan-Analysen. Leistungsplanbereiche, in denen viele Vorhaben mit Digitalisierungsinhalten erwartet werden können, insbesondere „G – Informations- und Kommunikationstechnologien“, belegen die obersten Ränge der gezeigten Anordnung.

Abbildung 6 und Abbildung 7 zeigen die absoluten Bewilligungssummen sowie die Anzahl neu geförderter Vorhaben je Leistungsplanbereich. Hier ist zu erkennen, dass die meisten Vorhaben im Querschnittsthema Digitalisierung und die meisten bewilligten Mittel mit großem Abstand in den Leistungsplanbereichen „R – Innovationsrelevante Rahmenbedingungen und übrige Querschnittsaktivitäten“ und „G – Informations- und Kommunikationstechnologien“ anfallen. Die beiden nächstgrößeren Leistungsplanbereiche in Bezug auf die absolute Höhe der Förderung von Vorhaben im Querschnittsthema Digitalisierung sind „A – Gesundheitsforschung und Gesundheitswirtschaft“ und „F – Klima, Umwelt, Nachhaltigkeit“. Es ist nicht überraschend, dass die absoluten Zahlen der Bereiche R, A und F hoch sind, obwohl ihr Rang in Bezug auf den Anteil von Vorhaben, die dem Querschnittsthema Digitalisierung zugeordnet wurden, relativ niedrig ist. Der Grund hierfür liegt in den insgesamt hohen Fördersummen in diesen Leistungsplanbereichen.

Insgesamt hat die Zuordnung von Vorhaben des BMBF zum Querschnittsthema Digitalisierung einen nicht unerheblichen Arbeitseinsatz erfordert. Insgesamt war ein wissenschaftlicher Mitarbeiter über einen Zeitraum von vier Kalendermonaten im Umfang von ca. 150 Arbeitsstunden damit beschäftigt. Der Arbeitsaufwand entstand insbesondere durch die breite und schwierige Abgrenzbarkeit des Querschnittsthemas Digitalisierung und der vielen Iterationen (insgesamt acht Runden), um False-Negatives und False-Positives weitgehend auszuschließen. Eine kurzfristige Durchführung solch einer Analyse zu einem vergleichbar komplexen Querschnittsthemenfeld ist daher nicht machbar.

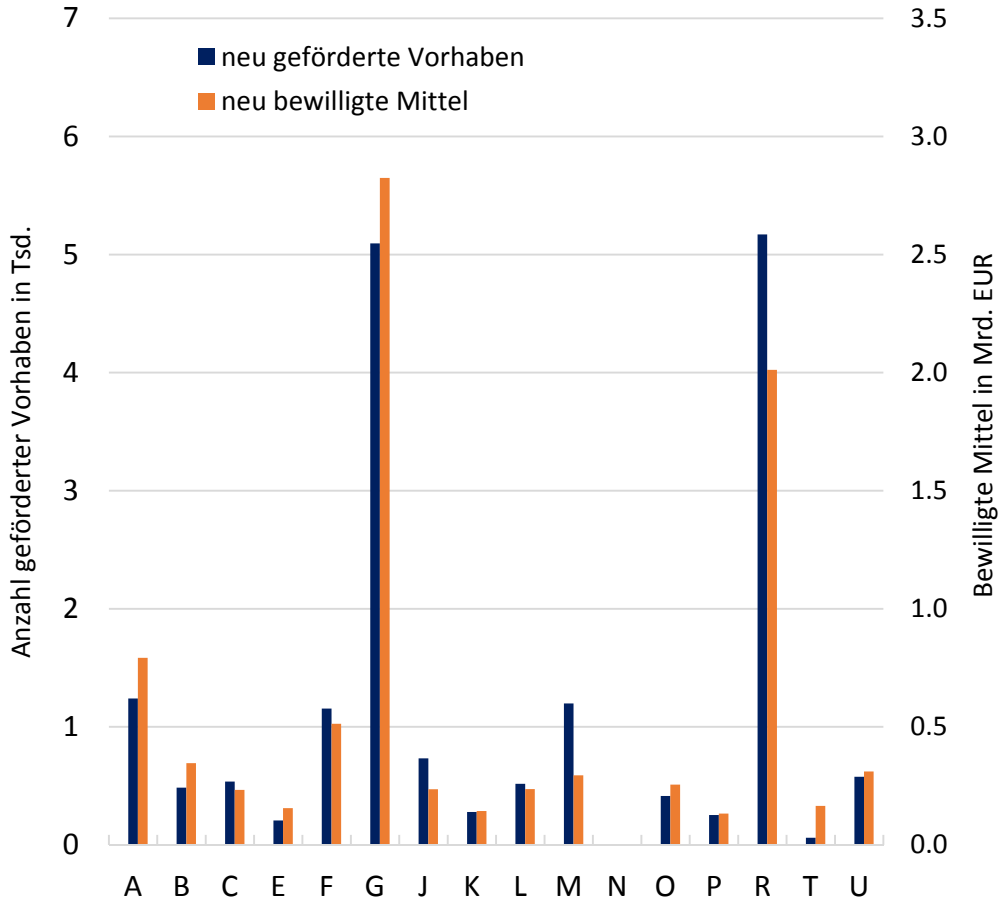
Des Weiteren ist zu beachten, dass in einem so dynamischen Themenfeld wie der Digitalisierung eine regelmäßige Anpassung und Prüfung der semantischen Textfeldanalyse notwendig ist. Die jetzt vorliegende Textfeldanalyse kann vermutlich noch für einige Jahre zuverlässige Ergebnisse bringen. In spätestens fünf Jahren ist nach unserer Einschätzung jedoch eine Überarbeitung notwendig, für die ein substanzieller Arbeitsaufwand (ca. die Hälfte des in dieser Studie benötigten Aufwands) zu veranschlagen ist.

**Abbildung 5: Anteil bewilligter Mittel von neu geförderten Vorhaben zum Querschnittsthema Digitalisierung am gesamten Bewilligungsvolumen nach Leistungsplanbereichen**




*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW*

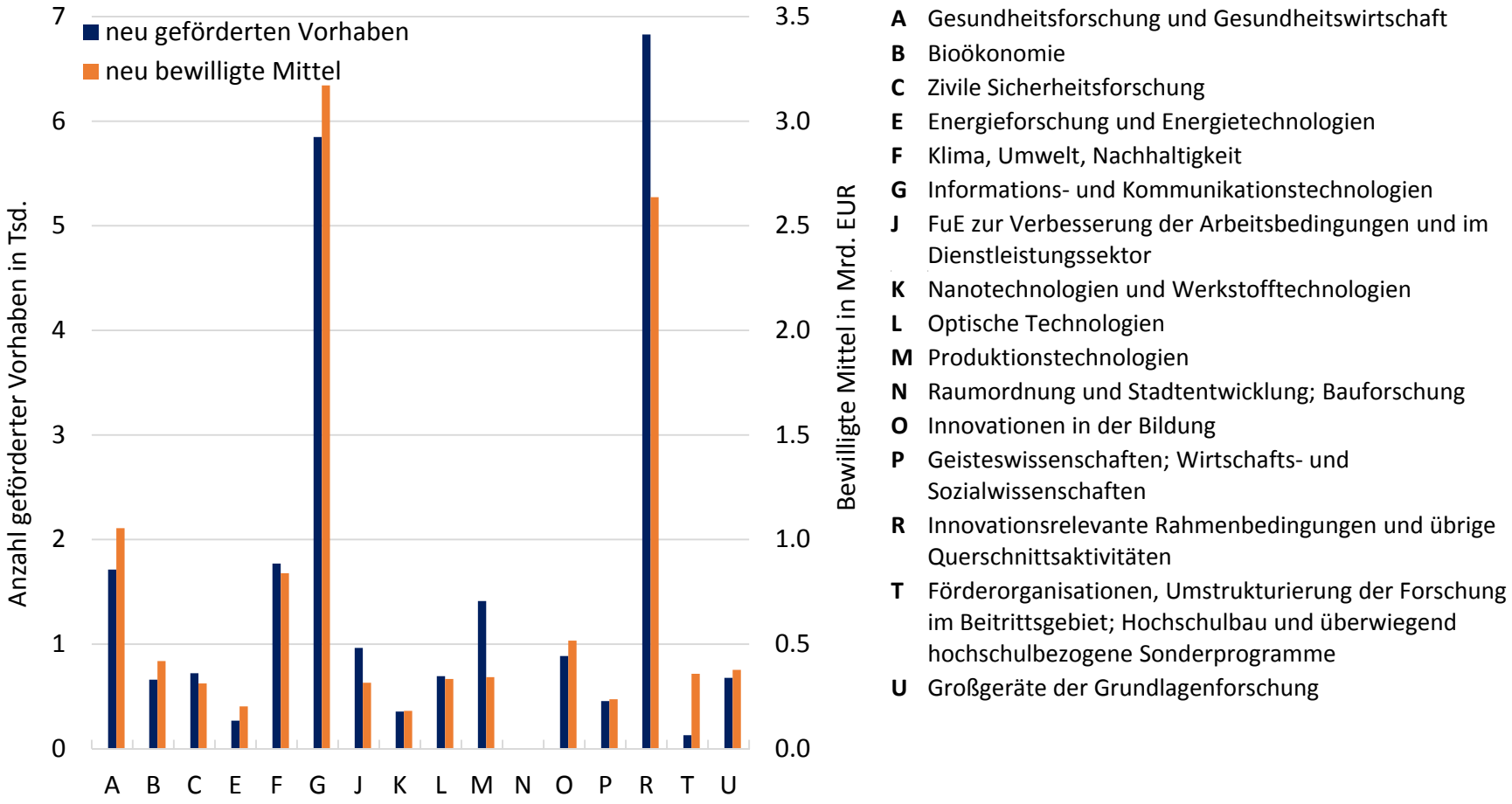
**Abbildung 6: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Digitalisierung nach Leistungsplanbereichen, restriktivere Abgrenzung**




**A** Gesundheitsforschung und Gesundheitswirtschaft
**B** Bioökonomie
**C** Zivile Sicherheitsforschung
**E** Energieforschung und Energietechnologien
**F** Klima, Umwelt, Nachhaltigkeit
**G** Informations- und Kommunikationstechnologien
**J** FuE zur Verbesserung der Arbeitsbedingungen und im Dienstleistungssektor
**K** Nanotechnologien und Werkstofftechnologien
**L** Optische Technologien
**M** Produktionstechnologien
**N** Raumordnung und Stadtentwicklung; Bauforschung
**O** Innovationen in der Bildung
**P** Geisteswissenschaften; Wirtschafts- und Sozialwissenschaften
**R** Innovationsrelevante Rahmenbedingungen und übrige Querschnittsaktivitäten
**T** Förderorganisationen, Umstrukturierung der Forschung im Beitrittsgebiet; Hochschulbau und überwiegend hochschulbezogene Sonderprogramme
**U** Großgeräte der Grundlagenforschung

*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

**Abbildung 7: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Digitalisierung nach Leistungsplanbereichen, weniger restriktive Abgrenzung**



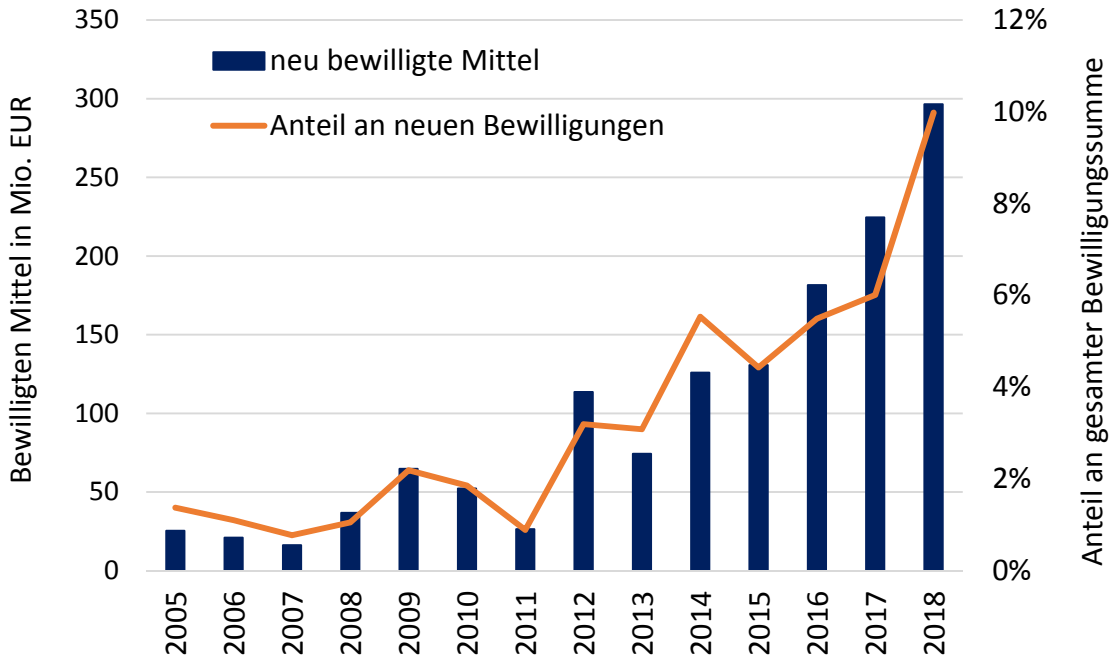

*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

## 3.2 TexAn-Analyse zum Themenfeld Künstliche Intelligenz

Die semantische Analyse zum Themenfeld Künstliche Intelligenz wurde als Teil der Analysen zum Querschnittsthema Digitalisierung durchgeführt. Die Künstliche Intelligenz bildete eine der Unterkategorien der Digitalisierung. Anders als für das Querschnittsthema insgesamt wird für das Themenfeld Künstliche Intelligenz nur eine Variante der semantischen Analyse vorgeschlagen ist, da hier eine eindeutige Zuordnung der Vorhaben besser möglich ist. Die Definition von Künstlicher Intelligenz in der semantischen Analyse umfasst vereinfacht zusammengefasst die Themen Maschinelles Lernen, Big Data Analysen, Mensch-Maschine/Roboter-Interaktionen sowie automatisiertes Erkennen von Bild und Ton.[7]

Aufgrund der hohen innovationspolitischen Bedeutung dieses Themenfelds wird hier eine separate Auswertung der Ergebnisse vorgenommen. Außerdem dienen diese Ergebnisse als Referenz für die Analysen im Abschnitt 4, wo mittels Methoden des maschinellen Lernens eine automatisierte Klassifikation von Vorhaben zum Themenfeld Künstliche Intelligenz erprobt wird.

**Abbildung 8: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Künstliche Intelligenz* 2005-2018**




** ohne Leistungsplanbereich T*
*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

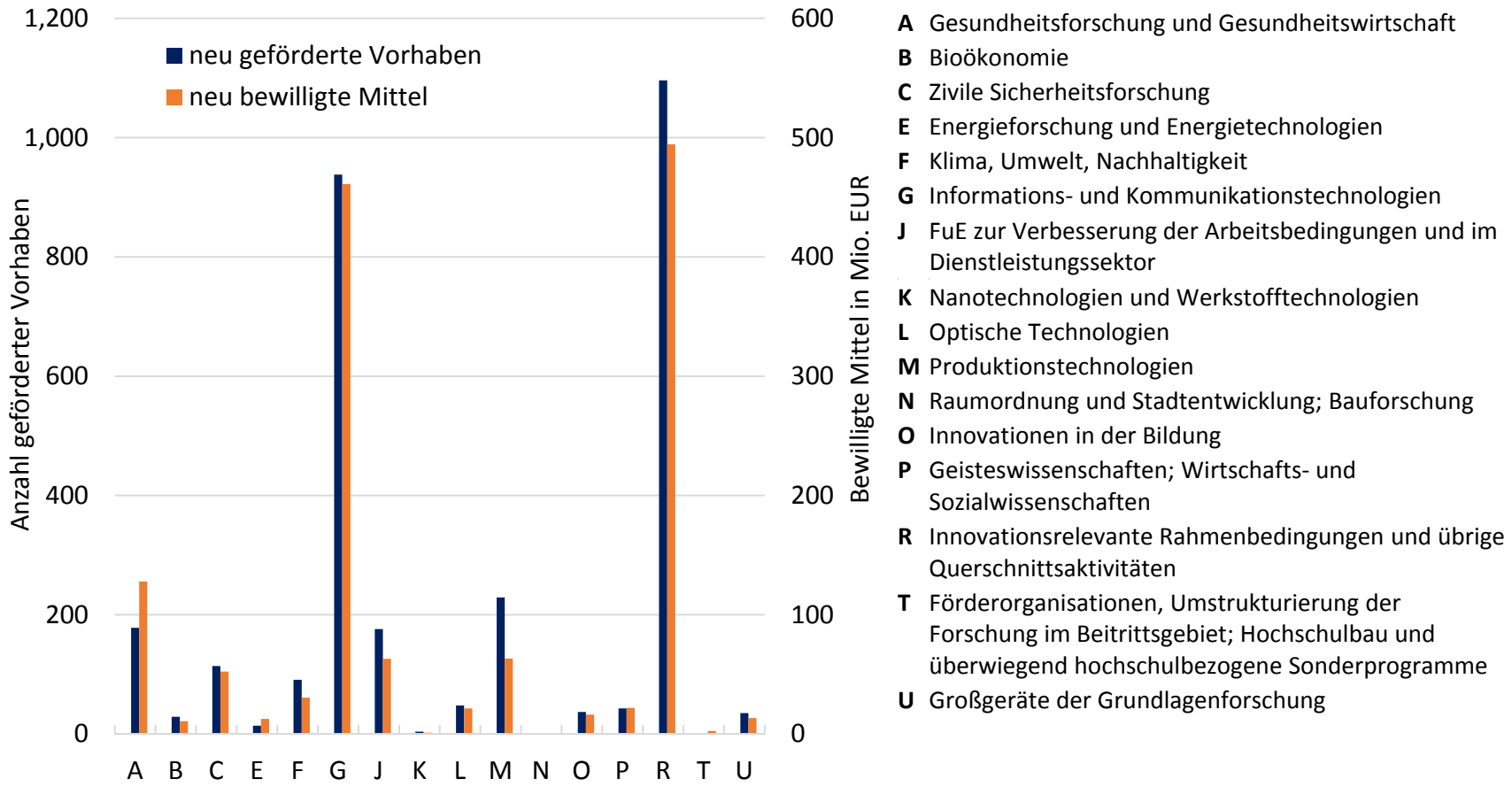
[7] Die finale Liste der genutzten Unterkategorien in Tabelle 4 im Anhang zeigt auch die Unterkategorien, die für die Analyse zur Künstlichen Intelligenz genutzt wurden.

Insgesamt wurden im Zeitraum von 2005 bis 2018 3.033 Projekte mit einem Bewilligungsvolumen von 1,4 Mrd. EUR dem Themenfeld Künstliche Intelligenz zugeordnet. Abbildung 8 zeigt die Entwicklung der bewilligten Mittel im Themenfeld Künstliche Intelligenz und ihren Anteil an den Gesamtbewilligungen des BMBF. Hier ist ein starker Anstieg der Mittel von 25 Mio. EUR in 2005 auf 225 Mio. EUR in 2018 zu erkennen. Auch nimmt die Bedeutung des Themenfelds Künstliche Intelligenz innerhalb der geförderten Vorhaben zu. Im Jahr 2005 lag ihr Anteil an der jährlichen gesamten Bewilligungssumme des BMBF bei 1,4 %, bis 2018 stieg dieser Wert auf 10,0 %.

Die Verteilung der Förderungen im Themenfeld Künstliche Intelligenz über die verschiedenen Leistungsplanbereiche ist in Abbildung 9 (Anzahl Vorhaben, bewilligte Mittel) und Abbildung 11 (Anteil an den gesamten Förderaktivitäten je Leistungsplanbereich) zu sehen. Das Muster der absoluten Bewilligungssumme je Leistungsplanbereich und der Anzahl der geförderten Vorhaben ähnelt dem Muster, das für das Querschnittsthema Digitalisierung insgesamt gefunden wurde. Insbesondere weisen die Leistungsplanbereiche „R – Innovationsrelevante Rahmenbedingungen und übrige Querschnittsaktivitäten“ und „G – Informations- und Kommunikationstechnologien“ mit großem Abstand erneut die höchsten Werte aus. Auch ist die Rangfolge des Anteils an den gesamten Förderaktivitäten in Abbildung 11 ähnlich zu denen in Abbildung 5, mit beispielsweise dem Bereich G an erster Stelle.

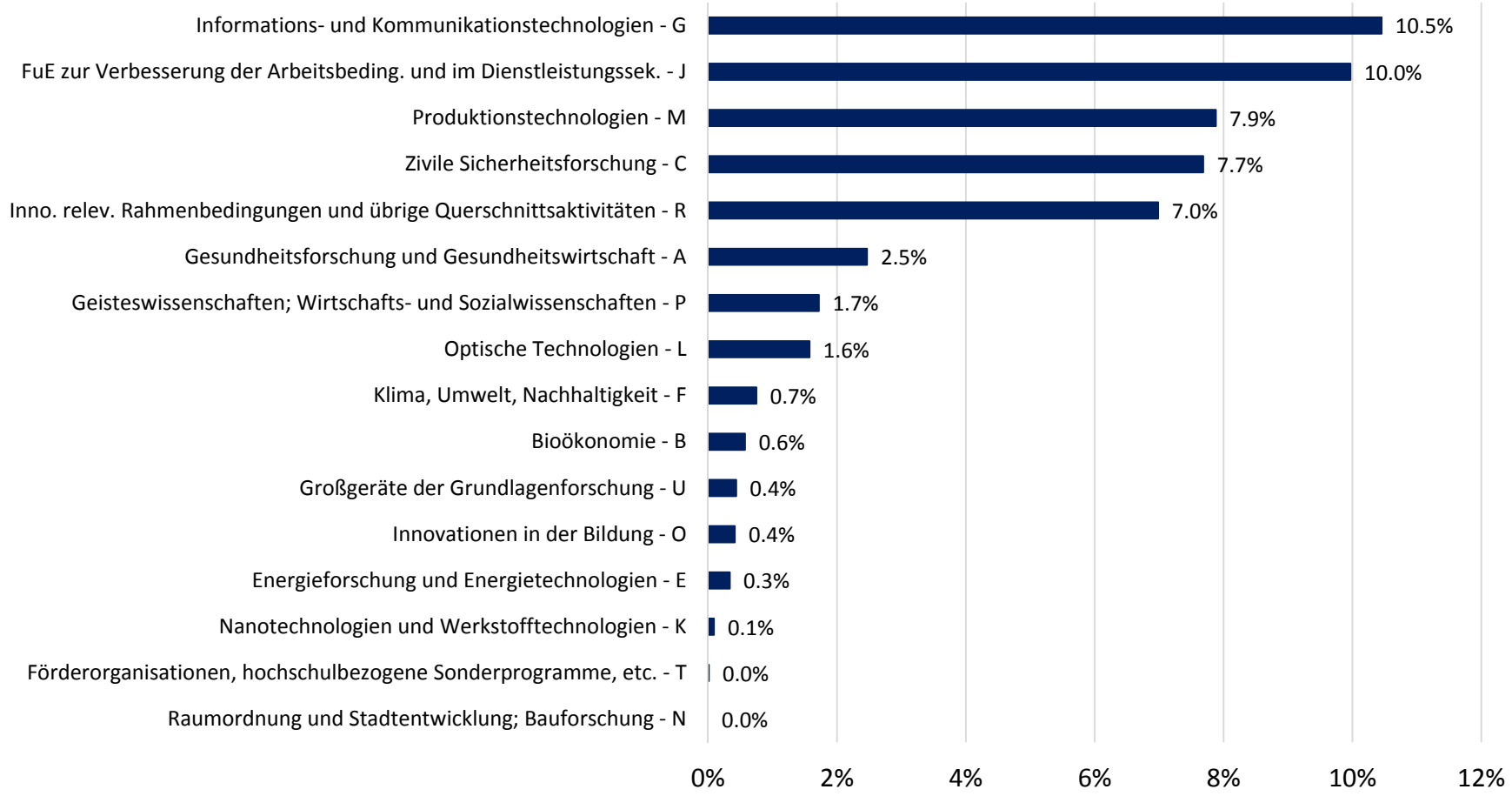
Der Bearbeitungsaufwand nur für das Themenfeld Künstliche Intelligenz war erheblich niedriger als für das Querschnittsthema Digitalisierung insgesamt, profitiert aber auch von "Spillovers", da die Prüfung von False-Negatives und False-Positives zu anderen Unterkategorien der Digitalisierung immer wieder Erkenntnisse für die verbesserte Erfassung des Themenfelds Künstliche Intelligenz ergaben. Isoliert betrachtet war für die Bearbeitung des Themenfelds Künstliche Intelligenz ein Aufwand von ca. 40 Arbeitsstunden eines wissenschaftlichen Mitarbeiters notwendig.

**Abbildung 9: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Künstliche Intelligenz nach Leistungsplanbereichen**




*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

**Abbildung 10: Anteil bewilligter Mittel von neu geförderten Vorhaben zum Querschnittsthema Künstliche Intelligenz am gesamten Bewilligungsvolumen nach Leistungsplanbereichen**




*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

### 3.3 TexAn-Analyse zum Querschnittsthema Soziale Innovationen

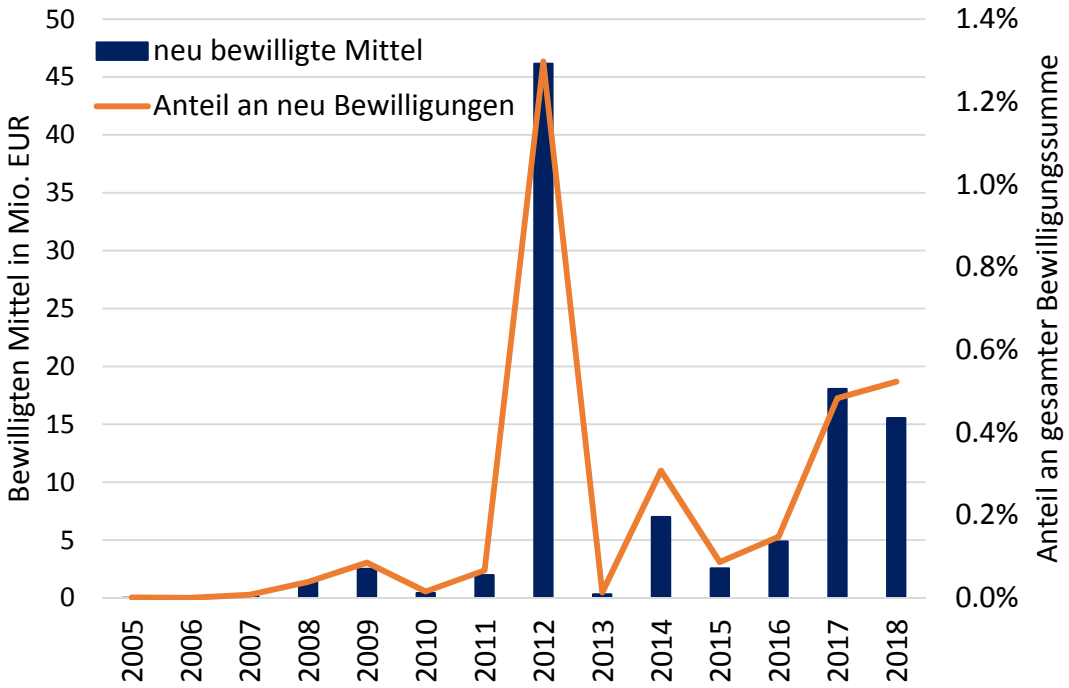
Als zweites Querschnittsthema wurde in dieser Machbarkeitsstudie eine Textfeldanalyse zu Sozialen Innovationen durchgeführt. Dieses Querschnittsthema unterscheidet sich deutlich von dem der Digitalisierung. Erstens geht es bei Sozialen Innovationen i.d.R. nicht um die Entwicklung neuer Technologien, sodass eine technologische Eingrenzung des Themas nicht zielführend ist. Zweitens bezieht sich das Thema auf die intendierte Wirkung von FuE-Vorhaben (nämlich die Entwicklung und/oder Einführung eines bestimmten Typs von Innovation). Drittens bezieht sich der Begriff Soziale Innovationen auf einen sehr breiten Anwendungsraum, ohne dass eine allgemein anerkannte, international standardisierte Definition vorliegen würde. Dadurch können potenziell fast alle Förderbereiche des BMBF hierzu Beiträge leisten, während es keine direkt ersichtlichen Leistungsplanklassen mit einem besonders hohen Potenzial für Soziale Innovationen gibt. Daher ist es auch nicht möglich, anders als Querschnittsthema Digitalisierung, die Güte des Ergebnisses einer semantischen Analyse an der "Trefferquote" im Bereich von vorab bestimmten Leistungsplanklassen zu bemessen.

Für dieses Querschnittsthema wurde daher eine andere Vorgehensweise gewählt. Zunächst wurde eine einfach strukturierte Textfeldanalyse (Suche nach den Wörtern sozial" und "Innovation"/"innovativ" sowie von häufig verwendeten Synonymen wie "Gesellschaft" und "Wandel") vorgenommen, die zum Ziel hatte möglichst wenige False-Negatives zu produzieren. Da das Suchergebnis eine relativ geringe Anzahl von Treffern ergab, konnten in einem zweiten Schritt alle klassifizierten Vorhaben manuell auf inkorrekt positive Klassifikationen geprüft und die Textfeldanalyse entsprechend angepasst werden. Die angepasste Textfeldanalyse wurde dann ausgeführt und erneut kontrolliert. Dieses Vorgehen wurde iterativ wiederholt bis die Zahl der False-Positives nahe Null war.

Die erste Textfeldanalyse mit einer hohen Wahrscheinlichkeit wenige False-Negatives aber vielen False-Positives ergab weniger als 500 geförderte Vorhaben zum Querschnittsthema Soziale Innovationen. Dabei zeigte sich, dass die Verwendung der Synonyme Gesellschaft und Wandel fast immer zu False-Positives führte, weshalb auf diese Suchwörter verzichtet wurde. Auch lieferte die Suche nach Wortkombinationen von "sozial" und "Innovation"/"innovativ" in den ersten Runden noch einige inkorrekt positiv klassifizierte Vorhaben, insbesondere aufgrund von Begriffen wie Sozialberuf, -einrichtung, -wesen oder -versicherung. Es wurde in die Textanalyse daher eingearbeitet, dass diese Wörter nicht in nächster Nähe zur Wortkombination sozial/innovation auftauchen dürfen.

Insgesamt wurden letztlich 127 Vorhaben mit einer Bewilligungssumme von 101 Mio. EUR dem Querschnittsbereich Soziale Innovation zugeordnet.[8] Abbildung 11 zeigt die Entwicklung der Förderung von Vorhaben zu Sozialen Innovationen. Das Großprojekt der VDI/VDE Innovation + Technik GmbH ist ein klarer Ausreiser in 2012. Ansonsten ist eine Zunahme der Bedeutung des Querschnittsthemas Soziale Innovationen zu erkennen. Die jährliche Summe an neubewilligten Mittel stieg von 20 Tsd. EUR in 2005 auf 16 Mio. EUR in 2018. Auch erhöhte sich ihr Anteil an der Gesamtbewilligungssumme im selben Zeitraum von quasi 0,0 % auf 0,5 %.

**Abbildung 11: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Soziale Innovationen 2005-2018**




** ohne Leistungsplanbereich T; zum Ausreißerwert im Jahr 2012 siehe Fußnote 8.*
*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

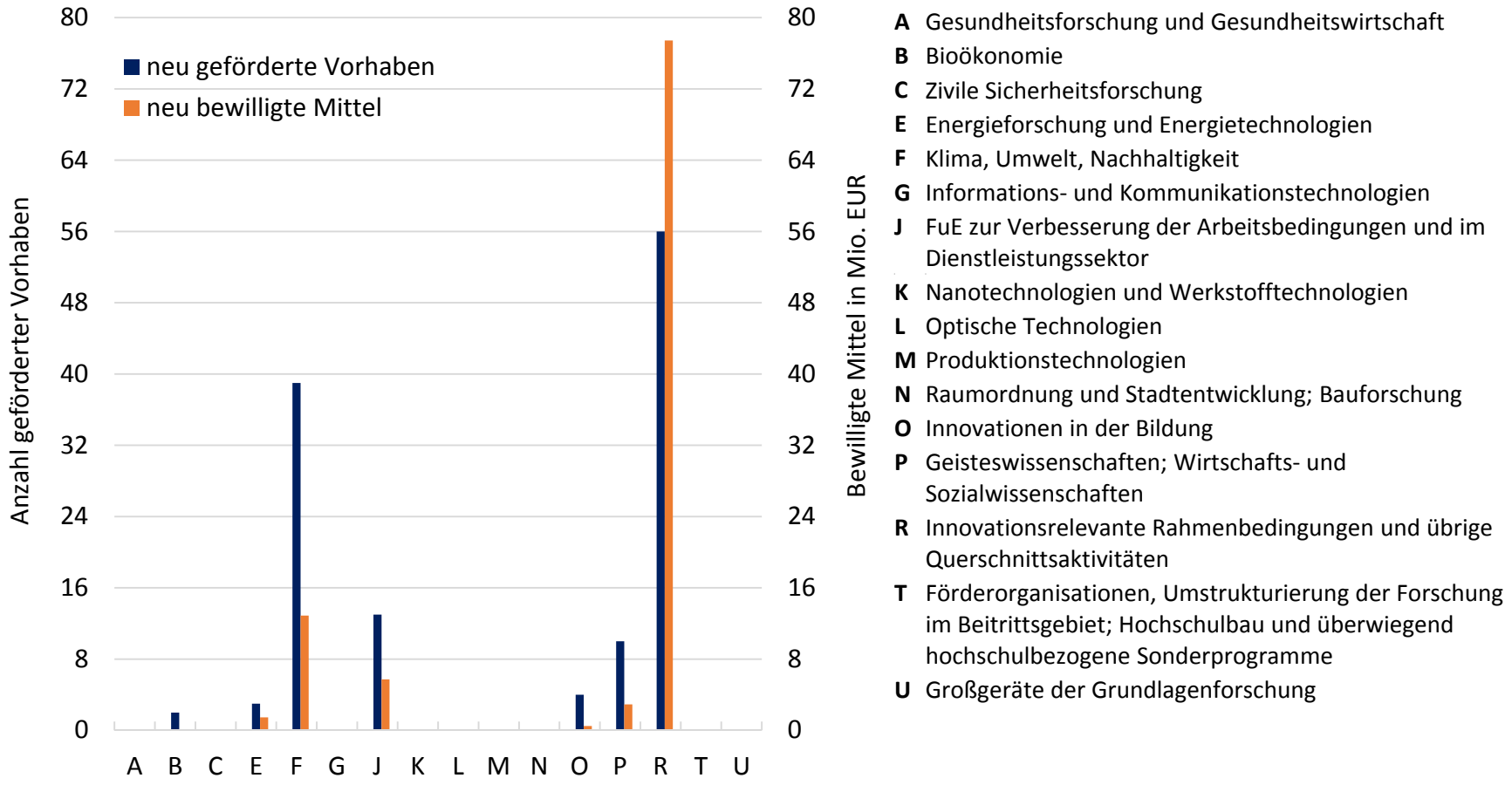
Abbildung 12 zeigt die absolute Höhe der bewilligten Mittel und die Anzahl der geförderten Vorhaben im Querschnittsthema Soziale Innovationen in den einzelnen Leistungsplanbereichen. Abbildung 13 zeigt den Anteil der Vorhaben zum Querschnittsthema Soziale Innovationen an dem Bewilligungsvolumen der einzelnen Leistungsplanbereiche. Die drei Leistungsplanbereiche mit den höchsten Werten in allen Kategorien

[8] Dabei ist anzumerken, dass 44 Mio. EUR auf ein einziges Projekt der VDI/VDE Innovation + Technik GmbH im Jahr 2012 entfallen, nämlich die „Projektträgerschaft Mensch-Technik-Interaktion 2012-2016 – Projektstabskosten“. Innerhalb dieses Vorhabens werden technische und soziale Innovationen im Themenfeld Demographischer Wandel und Mensch-Technik-Interaktion gefördert.

entsprechen sind R – Innovationsrelevante Rahmenbedingungen und übrige Querschnittsaktivitäten, J – FuE zur Verbesserung der Arbeitsbedingungen und im Dienstleistungssektor und F – Klima, Umwelt, Nachhaltigkeit. Wie bereits in den Analysen zu Digitalisierung können besonders viele Vorhaben im Bereich R identifiziert werden. Dies zeigt, dass dieser Bereich entsprechend seines Namens viele Vorhaben zu Querschnittsthemen umfasst.

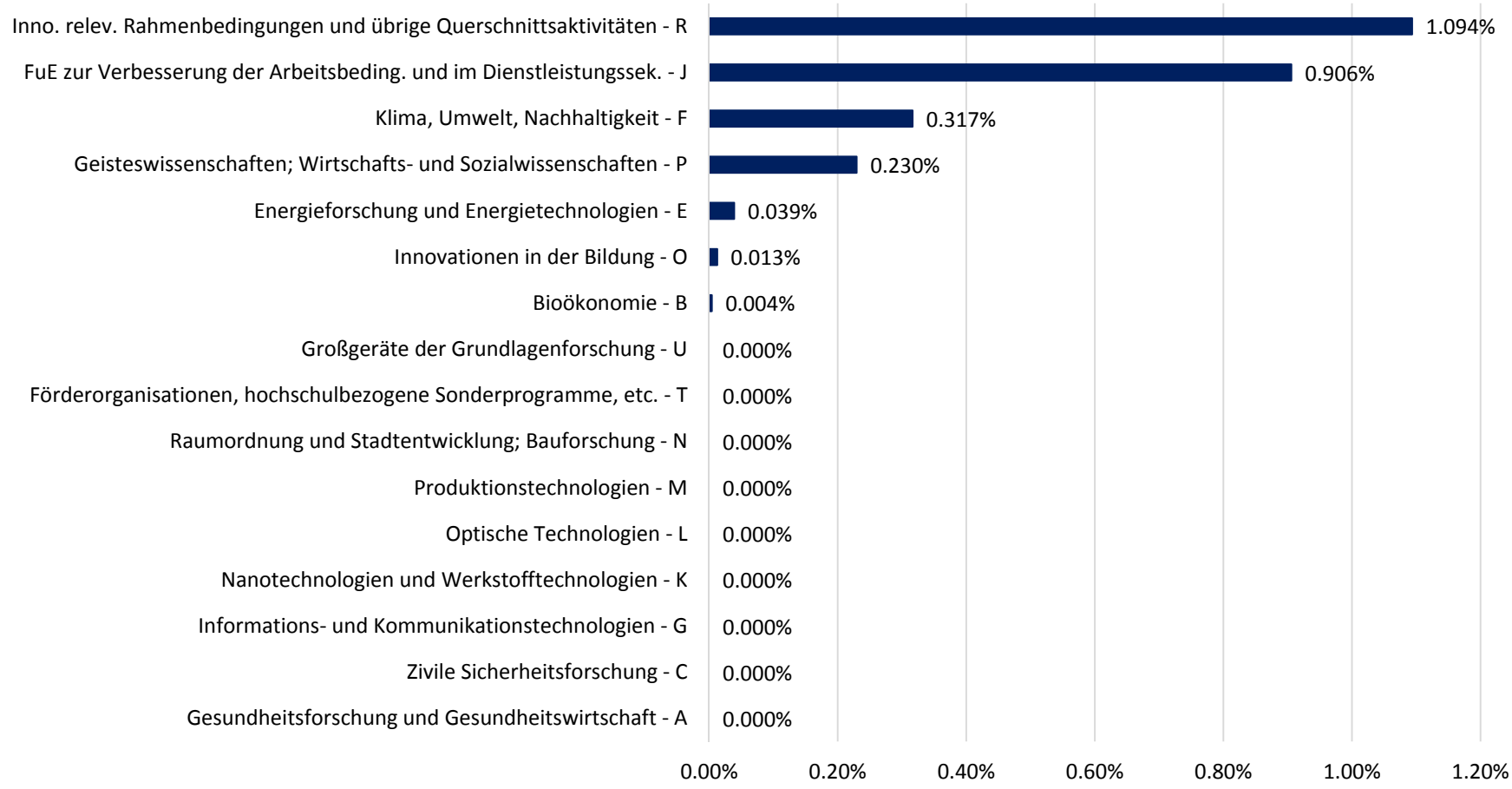
Die Zuordnung der neu geförderten Vorhaben des BMBF zum Querschnittsthema Soziale Innovationen konnte mit einem vergleichsweise geringen Aufwand realisiert werden. Die Zuordnung konnte innerhalb von zwei Kalenderwochen bei einem Einsatz von ca. 20 Stunden Arbeitszeit eines wissenschaftlichen Mitarbeiters umgesetzt werden. Die relativ kurze Arbeitszeit ergab sich insbesondere aufgrund der geringen Anzahl an Vorhaben, die selbst bei einer sehr einfachen und damit groben Suche nach Schlagwörtern zu Sozialen Innovationen gefunden wurden. Dadurch war der Aufwand für die manuellen Überprüfungen im Vergleich zum Querschnittsthema Digitalisierung erheblich geringer.

**Abbildung 12: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Soziale Innovationen nach Leistungsplanbereichen**




*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

**Abbildung 13: Anteil bewilligter Mittel von neu geförderten Vorhaben zum Querschnittsthema Soziale Innovationen am gesamten Bewilligungsvolumen nach Leistungsplanbereichen**




*Quelle: PROFI-Datenbank, Vorhabenbeschreibungen des BMBF, Berechnungen des ZEW.*

# 4 Analyse mittels maschinellem Lernen zum Themenfeld Künstliche Intelligenz

In diesem Abschnitt wird getestet, inwiefern ein trainierter, automatisierter Algorithmus geeignet ist, das Themenfeld Künstliche Intelligenz in den Abstracts von geförderten Vorhaben zu identifizieren. Hierzu wird auf die Ergebnisse aus Abschnitt 3.2 zur Künstlichen Intelligenz und auf einen Machine Learning Ansatz zurückgegriffen. Insgesamt besteht diese Analyse aus fünf Schritten: Textaufbereitung, KI-Indikator, Modellwahl, Training und Evaluation.

**Textaufbereitung.** Texte erfordern in der Regel eine Aufbereitung bevor diese in Machine Learning Modellen verwendet werden können, da diese in der Regel nur mit numerischen Größen in einer vordefinierten Datenstruktur arbeiten können. Tabelle 2 beschreibt die neun Schritte unserer Aufbereitung kurz.

**Tabelle 2: Textaufbereitung für Machine Learning Modelle**

| Pipeline | |
|---|---|
| 1. Datenselektion | Löschen von Datenpunkten ohne Text oder mit Textduplikaten. |
| 2. Tokenisierung | Die Texte werden in einzelne Wörter aufgeteilt. |
| 3. Stoppwort- Filter | Löschen von Wörtern auf Basis von mehreren Stoppwort-Listen. Die gelöschten Wörter sind üblicherweise nicht relevant für die Klassifikation, ein Beispiel ist das Wort „und“. |
| 4. Stemming | Verschiedene Wortvarianten werden auf ihre Grundform zurückgeführt. Zum Beispiel: Die Wörter „Wortes“ und „Wörter“ werden zu „Wort“. |
| 5. Löschen von kurzen Wörtern | Wörter mit einer Länge von Eins werden gelöscht. Dies sind in der Regel Satz- oder Sonderzeichen. |
| 6. Vereinheitlichen des Textes | Alle Großbuchstaben werden in Kleinbuchstaben überführt, um das Vokabular zu verkleinern. |
| 7. Löschen von Sonderzeichen | Die Sonderzeichen „!"#$ %&()*+,-./:;<=>?@[\]^_`{\|}~“ werden aus dem Text entfernt. |
| 8. Selektion von Wörtern | Nur die 50.000 häufigsten Wörter werden extrahiert. Die restlichen Wörter werden gelöscht. Dies verringert die Dimension der Daten und das “Rauschen“ im Text. |
| 9. Erstellen von Sequenzen | Die Wörter werden eindeutig durch Ganzzahlen ersetzt und in Sequenzen der maximalen Länge von 250 überführt. |

Die Abstracts der Vorhaben haben nach diesen Schritten eine Datenstruktur, die für Machine Learning Modelle als Eingabe geeignet ist. Nach der Durchführung aller neun Punkte verbleiben 68.191 nutzbare Vorhaben.

**KI-Indikator.** Für jedes der 68.191 verbliebenen Vorhaben existiert ein KI-Indikator der angibt, ob ein Vorhaben zum Themenfeld Künstliche Intelligenz gehört oder nicht. Dieser Indikator wurde in der zuvor beschrieben TexAn-Analyse ermittelt. Insgesamt gehören nur etwa 4 Prozent der Vorhaben zum Themenfeld Künstliche Intelligenz. Daraus ergeben sich 2.924 sogenannte positive und 65.267 negative Trainingspunkte.

**Modellwahl.** Für die Studie wurde ein „long short-term memory" neuronales Netzwerk[9] verwendet. Dieser Modelltyp ist auf dem neusten Stand der Technik bei der Klassifikation von Texten und wurde bereits in vielen Projekten eingesetzt. Zahlreiche Beispiele hierfür lassen sich unter anderem auf Kaggle[10] finden. Für die praktische Umsetzung wurde die Programmiersprache Python verwendet.

**Training.** Die 68.191 Vorhaben wurden in zwei Teile aufgeteilt. 80 % der 2.924 dem Themenfeld Künstliche Intelligenz zugeordneten Vorhaben sowie 80 % der nicht als KI klassifizierten Vorhaben wurden als Trainingsdaten verwendet. Diese dienen dazu, das oben genannte neuronale Netzwerk zu trainieren, sprich das neuronale Netzwerk optimiert seine Vorhersage über den KI-Indikator eines Vorhabens iterativ anhand der transformierten Abstract-Inhalte innerhalb dieses Datensatzes. Die restlichen 20 % der Vorhaben wurden als Evaluationsdatensatz zurückgehalten.

**Evaluation.** Für jedes Vorhaben im Evaluationsdatensatz wird mit Hilfe des trainierten neuronalen Netzes die Wahrscheinlichkeit berechnet, dass das Vorhaben Teil des Themenfelds Künstliche Intelligenz ist. Sofern die Wahrscheinlichkeit über 50 % betrug, wurde das Vorhaben dem Themenfeld zu geordnet, sonst nicht.

Die Güte des neuronalen Netzes bezüglich der Identifikation von Vorhaben zum Themenfeld Künstliche Intelligenz im Testdatensatz lässt sich mittels der Kennzahlen Precision, Recall und Accuracy[11] bewerten. Eine Übersicht der Kennzahlen ist in Tabelle 3 dargestellt. Es werden nur 68 % aller vom TexAn identifizierten Vorhaben zum Thema Künstliche Intelligenz durch das neuronale Netz erkannt *(68 % - Recall)*. Außerdem handelt es sich bei 22 % der durch das neuronale Netz identifizierten Vorhaben zum Themenfeld Künstliche Intelligenz um False-Positive *(78 % - Precision)*. Demnach besteht

[9] Sepp Hochreiter, Jürgen Schmidhuber (1997): Long short-term memory, *Neural Computation* 9(8), 1735–1780.

[10] https://www.kaggle.com/

[11] https://en.wikipedia.org/wiki/Precision_and_recall

eine relativ hohe Fehlerrate bei der Identifikation von Vorhaben im Themenfeld Künstliche Intelligenz.[12] Allerdings ist das LSTM Modell grundsätzlich in der Lage, aus den Daten zu lernen. Würden allen Vorhaben im Testdatensatz nicht dem Themenfeld Künstliche Intelligenz zugeordnet, wäre diese Klassifikation in 96 % aller Fälle Korrekt, da insgesamt lediglich 4 % der Vorhaben zu dem Themenfeld gehören. Das neuronale Netzwerk identifiziert dem gegenüber 98 % aller Vorhaben korrekt und ist demnach zuverlässiger *(98 % - Accuracy)*. Eine Steigerung der drei Gütekennzahlen wäre mit einer Vergrößerung des Testdatensatzes, insbesondere die Anzahl dem Themenfeld zugehöriger Vorhaben, wahrscheinlich.

**Tabelle 3: Zusammenfassung der Gütekennzahlen des Machine Learning Ansatzes für die Identifikation von Vorhaben im Themenfeld Künstliche Intelligenz**

| | |
|---|---|
| Anteil der mit dem Machine Learning Ansatz identifizierten KI-Vorhaben im Evaluationsdatensatz, die auch über die TexAn-Methode als KI-Vorhaben identifiziert wurden ("Precision") | 78 % |
| Anteil der über die TexAn-Methode identifizierten KI-Vorhaben im Evaluationsdatensatz, die auch mit dem Machine Learning Ansatz als KI-Vorhaben identifiziert wurden ("Recall") | 68 % |
| Anteil aller Vorhaben, die mit dem Machine Learning Ansatz korrekt als KI-Vorhaben oder Nicht-KI-Vorhaben identifiziert werden (Referenzwert ohne Einsatz von maschinellem Lernen: 96 %) | 98 % |

Aus diesen Ergebnissen lässt sich die Schlussfolgerung ziehen, dass der Machine Learning Ansatz für das Themenfeld Künstliche Intelligenz nicht geeignet ist, um mit hinreichender Genauigkeit KI-Vorhaben auf Basis der Abstract-Texte in der Gesamtheit aller vom BMBF geförderten Vorhaben zu identifizieren. Hierfür wären erheblich höhere Werte für Precision und Recall von zumindest 90 % notwendig. Dieses ungünstige Ergebnis liegt an der geringen Anzahl von Trainingsdatensätzen, was wiederum an der insgesamt niedrigen Anzahl von geförderten Vorhaben im Bereich KI liegt. Außerdem sind die zu jedem Vorhaben vorliegenden Texte relativ kurz und verhältnismäßig stark standardisiert. Für Themenfelder mit einer deutlich größeren Anzahl geförderter Vorhaben und damit einem umfangreicheren Trainingsdatensatz sowie auf Basis von umfangreicheren Texten (gesamte Vorhabenbeschreibung) könnte das Resultat deutlich anders

[12] Für Vorhaben, die nicht dem Themenfeld Künstliche Intelligenz zugeordnet wurden, sind Precision und Recall jeweils ca. 99 % und relativ nah am Optimalwert von 100 %. Dies liegt allerdings an dem sehr hohen Anteil von Nicht-KI-Vorhaben.

aussehen. Für Themenfelder, die noch seltener als das Themenfeld Künstliche Intelligenz vorkommen, (wie z.B. Soziale Innovationen) eignet sich ein Machine Learning Ansatz grundsätzlich nicht.

## Fazit

Die Machbarkeitsstudie hat untersucht, inwieweit Querschnittsthemen durch eine semantische Analyse von Vorhabenbeschreibungen mit hinreichender Genauigkeit identifiziert werden können. Innerhalb der Studie wurden die Querschnittsthemen Digitalisierung (inkl. des Teilgebiets Künstliche Intelligenz) und Soziale Innovationen mit einer vom ZEW entwickelten Software (TexAn – Textanalyser) untersucht. Des Weiteren wurde aufbauend auf den Ergebnissen der TexAn-Analysen zu Künstlicher Intelligenz ein Analysetool für eine automatisierte Zuordnung von Vorhaben zu Querschnittsthemen implementiert. Hierfür wurde auf Methoden des Natural Language Processings und des Maschinellen Lernens zurückgegriffen. Dabei wurde getestet, inwiefern ein trainierter, automatisierter Algorithmus geeignet ist, um das Querschnittsthema Künstliche Intelligenz in den Abstracts von geförderten Vorhaben zu identifizieren. Eine große Herausforderung dabei ist die geringe Anzahl von Vorhaben in dem Querschnittsthema, wodurch nur wenige Trainingsdatensätze vorliegen, an denen der automatisierter Algorithmus lernen kann.

Die Ergebnisse zur Machbarkeitsstudie sind gemischt. Die Klassifizierung des besonders breiten und schwer abgrenzbaren Querschnittsthemas Digitalisierung hat bei einem Einsatz eines wissenschaftlichen Mitarbeiters einen Arbeitsaufwand von ca. 150 Stunden verteilt über einen Zeitraum von vier Kalendermonaten in Anspruch genommen. Dies bedeutet, dass eine kurzfristige Klassifizierung ähnlicher komplexer Thematiken, etwa im Fall einer kleinen oder großen Anfrage des Bundestags oder kurzfristigen politischen Informationsbedarfs, nicht möglich ist. Auch konnte zum Querschnittsthema Digitalisierung keine Vorgehensweise gefunden werden, die eine eindeutige Klassifikation ermöglicht. Stattdessen wurden zwei Klassifizierungsvarianten (eine restriktive und eine weniger restriktive) entwickelt, die quasi einen oberen und unteren Bereich von Vorhaben zum Querschnittsthema Digitalisierung eingrenzen. Schließlich ist zu beachten, dass die hier entwickelte semantische Analyse von Vorhabenbeschreibungen auf dem aktuellen technischen Stand der Digitalisierung beruht. Da sich dieser rasch ändert, sind regelmäßig arbeitsaufwendige Anpassungen der semantischen Analyse notwendig.

Das Themenfeld Künstliche Intelligenz und das Querschnittsthema Soziale Innovationen konnten rascher bearbeitet werden, da sie zum einen besser einzugrenzen sind und zum anderen die Anzahl der relevanten Vorhaben deutlich geringer ist, was den Aufwand der manuellen Prüfung erheblich reduziert. Der benötigte Arbeitszeitraum zum Themenfeld Künstliche Intelligenz betrug ca. 40 Arbeitsstunden. Für das Querschnittsthema Soziale Innovationen fiel ein Aufwand von 20 Arbeitsstunden an. Die Klassifizierung des Querschnittsthemas Soziale Innovationen konnten innerhalb von 2 Arbeitswochen und damit recht kurzfristig umgesetzt werden.

Abschließend ist des Weiteren festzustellen, dass der getestete Machine Learning Ansatz nicht geeignet ist, um das Themenfeld Künstliche Intelligenz in den Abstracts von

geförderten Vorhaben mit hinreichender Genauigkeit zu identifizieren. Grund hierfür ist die geringe Anzahl an Vorhaben zu diesem Thema und der dadurch kleine Trainingsdatensatz sowie die geringe Länge der Texte und ihr recht hoher Grad an Standardisierung. Allerdings wäre eine Identifikation von Themen mit einem höheren Vorkommen sowie unter Nutzung der gesamten Vorhabenbeschreibungen mit Hilfe von Machine Learning Ansätzen grundsätzlich denkbar. Es ist allerdings auch hier anzumerken, dass diese Art von Analyse immer auf dem aktuellen technischen Stand beruht und regelmäßig angepasst werden muss.

# 5 Anhang

## 5.1 Unterkategorien der Digitalisierung

Tabelle 4 zeigt die verschiedenen Unterkategorien der TexAn-Analyse zur Digitalisierung. Die Spalte Unterkategorie umfasst die Namen der einzelnen Unterkategorien der Digitalisierung wie sie in der Programmierung der Analyse verwendet wurden. Die Themenüberschrift gibt einen kurzen Überblick, welche Themen die entsprechende Unterkategorie umfasst. Die weniger restriktive Analyse zur Digitalisierung umfasst alle gelistet Unterkategorien, die restriktivere Analyse abstrahiert von den blauhinterlegten. Die TexAn-Analyse zur Künstlichen Intelligenz nutzt die grünhinterlegten Unterkategorien. Die genannten Unterkategorien sind nicht vollkommen überschneidungsfrei in ihren Themen und genutzte Schlagwortkombinationen.

**Tabelle 4: Unterkategorien der TexAn-Analyse zur Digitalisierung**

| Unterkategorien | Themenüberschrift |
|---|---|
| DIGI_E | Digital - simple |
| INTSYS | Integrierten Systemen - simple |
| IKT_E | Informations- und Kommunikationstechnologien - simple |
| IT_E | Hochleistungsrechner, Softwarewerkzeuge, Soft- Middle- und Hardware |
| ITTECH | Chiptechnologie, EDV, Hochleistungsrechner |
| QUANT | Quantencomputer |
| KI | Maschinelles Lernen, Künstliche Intelligenz |
| MENSCHMA | Mensch-Maschine/Roboter-Interaktion |
| ERKENN | Automatisches Erkennen von Tönen, Bildern etc., Blicktracking |
| BIGDATA | Big Data |
| DRIV | Selbstfahrende Fahrzeuge |
| SMART | Intelligente Produkte/Dienstleistungen/Software/Fahrzeuge/Netzwerke |
| VIRTU | Virtual Reality, Augmented Reality, 3D- und 4D-Simulation |
| INTE_E | Internet - simple |
| IOF | Internet der Dinge |
| ECOOM | E-Commerce, Onlinehandel, E-Business |
| APPL | Appentwicklung |
| DEEP | Deepweb |
| CYBER | Cybercrime, Cyber-Technical-Systems |
| SECU | Cybersicherheit |
| CLOUD_E | Cloudlösungen - simple |
| CLOUD | Cloudcomputing , - technologien, -anwendungen, -dienste |
| DATENPLAT | Datenplattformen, Datennetzwerke |
| PLAT | Datenplattformen, Plattformservices |
| KRYPTO | Kryptowährung |
| DOKU | Elektronische Dokumentation |

## 5.2 Code zur TexAn-Analyse

Innerhalb der TexAn-Software werden für jede definierte Klasse, wie beispielsweise Soziale Innovationen oder eine der in Tabelle 4 genannten Unterkategorien, sogenannte `seeker` definiert. Diese beinhalten die Schlagwörter die zur Identifikation einer Klasse genutzt werden. Diese Schlagwörter werden entweder eins-zu-eins verwendet oder nochmals durch die Option `using standard` standardisiert. Diese Option wandelt die eingegebenen Schlagworte sowie die einzelnen Textfelder der Vorhaben um, sprich alle Buchstaben werden in Großbuchstaben geändert, die Umlaute werden zu z.B. "ä" zu "ae" umgewandelt, alle Sonderzeichen werden durch Leerzeichen ersetzt und mehrere Leerzeichen zusammengefasst.

Der Befehl `texan` führt die eigentliche Textfeldanalyse durch. Nach ihm wird definiert, ob die definierten `seeker` einzeln gesucht werden, Kombinationen von `seekern` in gewissen Wortabständen Auftauchen müssen, oder bestimmte `seeker` nicht auftauchen dürfen. Am häufigsten genutzt werden die `texan`-Optionen `max # words near` und `not`. In der ersten Option wird definiert wie viele Wörter (#) die Begriffe zweier `seeker` voneinander entfernt sein dürfen, damit eine Klasse in einem Textfeld vorliegt. In der zweiten Option werden Begriffe definiert, die nicht in einem Textfeld vorkommen dürfen, damit das Vorliegen einer Klasse identifiziert werden kann.

**Übersicht 1: Code der TexAn-Analyse zum Querschnittsthema Digitalisierung (inkl. Künstliche Intelligenz)**

```
seeker digi1_e seeks "digital" using standard
texan DIGI_E analyses digi1_e


seeker inte1 seeks "4.0"
seeker inte2 seeks "Industr" using standard
seeker inte3 seeks "integrier", "eingebette", "kooperierend", "automatis", "simulatio"
using standard
seeker inte4 seeks "system" using standard
seeker inte5 seeks "data", "daten", "algorythm", "algorith", "softwa", "selbstlern",
              "automatis", "computer", "digital", "smart" using standard
seeker inte6 seeks "Informationssystem", "Kommunikationssystem", "information system",
              "communication system"
seeker inte7 seeks "I4.0", "I 4.0"
texan INTESYS analyses inte1 max 3 words near inte2 or inte3 max 2 words near inte4 or
              inte5 max 5 words near inte4 or inte6 or inte7


seeker ikt1_e seeks " IKT ", " IKT-", " IKT,"
texan IKT_E analyses ikt1_e


seeker it1_e seeks "software", "hardware", "middleware", "informationstechn", "inter-
              nettechn" using standard
seeker it2_e seeks " IT ", " IT-", " IT,", " IT-,", "-IT-", " HPC ", " HPC-", " HPC"
seeker it3_e seeks "computing", "computer" using standard
```

```
seeker it4_e seeks "programmieren", "Programmierung", "Programmcode"

seeker it5_e seeks "compute", "softwa", "kommunik"

seeker it6_e seeks " Bits ", " Bit ", " Byte ", " Bytes ", " Kilobyte", " Megabyte", "
             Gigabyte", " Terabyte"

texan IT_E analyses it1_e or it2_e or it4_e max 5 words near it5_e or it6_e


seeker it1 seeks "Hochleistungsrech", "hochleistungsrech", "Quantencompu", "Quanten-
             rechn", "EDV-ger", "EDV-Ger", "Multicore", "Multi-core", "Multi-Core",
             "MultiCore", "Supercomput", "Mobilfunk", "Hochleistungsprozes"

seeker it2 seeks "IT-Werkzeug", "IT Werkzeug", "Software Werkzeug", "Software-Werkzeug",
             "Softwarewerkzeug", "Informationstechno", "Informatik"

seeker it3 seeks "quanten", "qubit", "hochleistu", "höchstleist", "high perfomance" u-
             sing standard

seeker it4 seeks "computer", "repeater", "rechner", "server", "prozessor", "EDV-ger",
             "datenverarbeitungsger", "hardware", "chiptech", "computing" using stan-
             dard

seeker it5 seeks "3d-druck", "chiptech" using standard

seeker it6 seeks " IT ", " IT-", " IT,", " IT-,", "-IT-", " HPC ", " HPC-", " HPC", "
Java "

seeker it7 seeks "technolo", "technik", "system", "netzwerk" using standard

seeker it8 seeks "software" using standard

seeker it9 seeks "code", "lösung", "system", "analys", "entwick" using standard

seeker it10 seeks "glasfaser" using standard

seeker it11 seeks "netz", "leitung" using standard

seeker it12 seeks "schnittstelle" using standard

seeker it13 seeks "optisch" using standard

seeker it14 seeks "drahtlos", "mobil" using standard

seeker it15 seeks "kommunikation", "kommunizier" using standard

seeker it16 seeks " IP ", " IP-"

seeker it17 seeks "adres", "addres" using standard

texan ITTECH analyses it1 or it2 or it3 max 3 words near it4 or it5 or it6 max 2 words
             near it7 or it8 max 3 words near it9 or it10 max 3 words near it11 or
             it12 max 3 words near it13 or it14 max 3 words near it15 or it16 max 2
             words near it17


seeker quant1 seeks "quanten" using standard

seeker quant2 seeks "plattform", "platform", "netzwerk" using standard

seeker quant3  seeks "kommunikat" using standard

seeker quant4  seeks "computer", "repeater", "rechner", "server", "prozessor", "EDV-
             ger", "datenverarbeitungsger", "hardware", "chiptech" using standard

seeker quant5 seeks "technolo", "technik", "system", "technisch", "netzwerk" using
             standard

texan QUANT    analyses quant1 max 2 words near quant2 or quant1 max 2 words near quant3
             or quant1 max 2 words near quant4 or quant1 max 2 words near quant5


seeker ki1 seeks "lern" using standard

seeker ki2 seeks "maschinel", "tiefes", "selbstständig" using standard

seeker ki3 seeks "netzwerk", "network", "netz" using standard

seeker ki4 seeks "neuronal", "neural" using standard

seeker ki5 seeks "intelli", "autonom", "vernetz" using standard

seeker ki6 seeks "service", "dienst", "maschin", "robot", "computer" using standard

seeker ki7 seeks "system" using standard
```

```
seeker ki9 seeks "data", "daten", "algorythm", "algorith", "softwa", "selbstlern", "au-
              tomatis", "computer", "digital", "smart" using standard

seeker ki10 seeks "Deep Learn", "deep learn", "deep-learn", "Deep-Learn", "künstliche
              Intellig", "künstlicher Intellig", "künstlichen Intellig", "künstliche
              intellig", "künstlicher intellig", "künstlichen intellig", "artificial
              intellig", "K.I.", "KI-Anwend", "KI Anwend", "KI-Appli", "KI Appli", "KI-
              Method", "KIMethod", "machine learning", "machine-learning", "Machine
              Learning", "Machine-Learning", "Machinelearning", "wearable computing",
              "Wearable-Computing", "Maschinenlern"

seeker ki11 seeks "maschi", "robot", "computer" using standard

seeker ki12 seeks "service", "autonom", "dienstleist" using standard

seeker noki4 seeks "brain", "gehirn", "hirn", "gesundheit", "health", "Kopf", "nerven",
              "synap", "schmerz", "Hippocamp", "gedächtnis", "Polyam", "anatomis",
              "psychisch", "psyche", "Krank" using standard

texan KI analyses ki10 or ki1 max 2 word near ki2 or ki3 max 2 words near ki4 max 10
              words near ki9 or ki5 max 4 words near ki6 max 20 words near ki9 or ki11
              max 2 words near ki12 max 20 words near ki5 or ki11 max 2 words near ki12
              max 20 words near ki9 or ki7 max 2 words near ki5 max 10 words near ki9
              or ki5 max 3 words near ki9


seeker mensch1 seeks "mensch", "user", "benutzer" using standard

seeker mensch2 seeks "maschi", "robot", "computer", "techni" using standard

seeker mensch3 seeks "interaktio", "interactio" using standard

texan MENSCHMA analyses mensch1 max 2 words near mensch2 or mensch1 max 5 words near
              mensch2 max 5 words near mensch3


seeker erkenn1 seeks "erkenn"

seeker erkenn2 seeks " Gesicht", " Bild", " Muster", " Ton", "Interface", " Sprach",
              "muster", "Schrift", "Person" using standard

seeker erkenn3 seeks "data", "daten", "algorythm", "algorith", "softwa", "selbstlern",
              "automatis", "computer", "digital", "smart" using standard

seeker erkenn4 seeks "eye" using standard

seeker erkenn5 seeks "tracking" using standard

texan ERKENN analyses erkenn1 max 3 words near erkenn2 max 50 words near erkenn3 or er-
              kenn4 max 2 words near erkenn5


seeker bigdata1 seeks "Big data", "big data", "bigdata", "Bigdata", "Big Data", "big
              Data"

texan BIGDATA analyses bigdata1


seeker driv1 seeks "fahren ", "fahrzeug", "Fahrzeug", "drive", "driving ", " Auto",
              "PKW", "LKW", "Lastfahr"

seeker driv2 seeks "vernetz", "autonom", "selbst", "intellige", "smart" using standard

seeker nodriv seeks "verfahr" using standard

texan DRIV analyses driv1 max 5 words near driv2 minimum 10 words near nodriv ignore


seeker smart1 seeks "smart ", "Smart ", "smart,", "smart.", "Smart", "Smartcard",
              "smartcard", "smart product", "Smart Product", "smart-product", "Smart-
              Product"

seeker smart2 seeks "computing", "service", "dienst", "lösung" using standard

seeker smart3  seeks "program", "software", "programm", "technolog", "informationstech"
              using standard

seeker smart4  seeks " Grid ", " grid ", "-Grid ", " car ", " cars ", "-Cars", "-Car ",
              " Cars ", " Cars", " Meter ", " Meters ", " meter ", "-Meter ", "-Meters
              "
```

```
seeker smart5 seeks "driving", "fahrzeug", "automobil", "lastkraft", "robot",
              "messstell"
texan SMART analyses smart1 max 3 words near smart2 or smart1 max 3 words near smart3 or
              smart1 max 2 words near smart4 or smart3 max 2 words near smart5


seeker virt1 seeks "virtuel", "virtual" using standard
seeker virt2 seeks "augment" using standard
seeker virt3 seeks "reality" using standard
seeker virt4 seeks "3D", "4D"
seeker virt5 seeks "druck", "system", "simulati", "visuali", "fertigung" using standard
seeker virt6 seeks "LCD"
seeker virt7 seeks "techno" using standard
texan VIRTU analyses virt1 or virt2 max 3 words near virt3 or virt4 max 2 words near
              virt5 or virt6 max 2 words near virt7


seeker inte1_e seeks "internet" using standard
seeker inte2_e seeks " Web ", " Web-", " Web,", "world wide web", "World Wide Web",
              "World-Wide-Web", "world-wide-web", "www", "WWW", "Web2.0", "Web 2.0",
              "Web3.0", "Web 3.0", "Web-2.0", "Web-3.0"
seeker inte3_e seeks "semantic", "semantisc" using standard
seeker inte4_e seeks "web" using standard
texan INTE_E analyses inte1_e or inte2_e or inte3_e max 3 words near inte4_e


seeker iof1 seeks "Internet der Dinge", "Internet-der-Dinge", "Internet Der Dinge", "In-
              ternet-Der-Dinge"
seeker iof2 seeks "internetderdinge" using standard
seeker iof3 seeks "Internet of Things", "Internet-of-Things", "internet of things", "in-
              ternet-of-things"
seeker iof4 seeks "internetofthings" using standard
texan IOF analyses iof1 or iof2 or iof3 or iof4


seeker ecomm1 seeks "E Commerc", " Ecommerc", " e commerc", " eCommerc", "e Commerc", "
              E-Commerc", " E-commerc", " e-business", " e-Business", " E-Business", "
              ebusiness", " Ebusiness", " EBusiness", " eBusiness"
seeker ecomm2 seeks"Internethandel", "Onlinehandel", "Online-Handel", "Online-Shop",
              "Onlineshop", "elektronischer Handel", "elektronisch handel", "elektroni-
              sche handel"
texan ECOMM analyses ecomm1 or ecomm2


seeker app1 seeks "App,", "App ", "Apps ", "App.", "Apps,", "usability", " App-"
texan APPL analyses app1


seeker deep1   seeks   "Deep Web", "Deep-Web", "Deepweb", "Hidden-Web", "Hidden Web",
              "invisible web", "invisible-web", "verstecktes web", "dark web", "dark
              net", "opaque web", "opaque-web", "proprietary web", "proprietary-web",
              "visible web", "visible-web", "clear web", "clear-web", "surface-web",
              "surface web" using standard
texan DEEP analyses deep1


seeker cyber1 seeks "cyber" using standard
texan CYBER analyses cyber1
```

```
seeker secur1 seeks "sicher", "secur" using standard
seeker secur2 seeks "cyber", " IT ", " IT-", " IT,", " IT-,", "internet", "Internet",
"Netz", "                    Netzwerk", "Website", "Webseite", " Web ", " Web-"
seeker secur3 seeks "trojan", "firewall", "honeypot", "schadsoftwar" using standard
seeker secur4 seeks "IPv6", " Bot ", " Bots ", " Botnetz "
seeker secur5 seeks "intrusion" using standard
seeker secur6 seeks "detection" using standard
texan SECU analyses secur3 or secur1 max 3 words near secur2 or secur5 max 5 words near
            secur6


seeker cloud1_e seeks "Cloud ", " Cloud,", " Cloud-"
texan CLOUD_E analyses cloud1_e


seeker cloud1 seeks "cloud" using standard
seeker cloud2 seeks "anwendung" using standard
seeker cloud4 seeks "computing", "service", "dienst", "lösung" using standard
seeker cloud3 seeks "technolo", "technik", "system", "netzwerk", "netz" using standard
texan CLOUD analyses cloud1 max 2 words near cloud2 or cloud1 max 2 words near cloud3 or
            cloud1 max 2 words near cloud4


seeker platt1 seeks "plattform", "platform", "netzwerk" using standard
seeker platt2 seeks "internet", "website", "webseite" using standard
seeker platt3 seeks "online ", "on-line", "online-", "Online"
seeker platt4 seeks "Daten ", "Daten.", "Daten,", "Datenanal", "Datenbas"
seeker platt5 seeks "program", "software", "programmier", "programment" using standard
seeker platt6 seeks "intellig", "simulatio", "service", "dienstleistu", "technolog"
            using standard
texan PLAT analyses platt1 max 5 words near platt2 or platt1 max 5 words near platt3 or
            platt1 max 3 words near platt4 or platt5 max 3 words near platt1 or
            platt1 max 3 words near platt6


seeker datenplat1 seeks "Daten ", "Daten.", "Daten,"
seeker datenplat2 seeks "plattform", "platform", "netzwerk", "austausch",        "kommu-
            nikati" using standard
texan DATENPLAT analyses datenplat1 max 5 words near datenplat2


seeker kryp1 seeks "krypto" using standard
texan KRYPTO analyses kryp1


seeker doku1 seeks "dokumentat" using standard
seeker doku2 seeks "digital" using standard
seeker doku3 seeks "elektro" using standard
texan DOKU analyses doku1 max 3 words near doku2 or doku1 max 2 words near doku3
```

## Übersicht 2: Code der TexAn-Analyse zum Querschnittsthema Soziale Innovationen

```
seeker soin_1 seeks "sozial" using standard

seeker soin_2 seeks "innovati", "neuerung", "neuheit", "erneuer", "neugestalt", "neu-
              ordn" using standard

seeker nosoin_1 seeks "sozialwissensch", "sozialberuf", "sozialwesen", "sozialversi-
              cher", "sozialhilf" using standard

seeker nosoin_2 seeks "beruf", "dienst", "betreuung", "einricht" using standard

texan SOZINN analyses soin_1 max 3 words near soin_2 and not nosoin_1 and not nosoin_2
max 1 words near soin_1
```

## Abbildungsverzeichnis

Abbildung 1: Anzahl und bewilligte Mittel vom BMBF geförderter Vorhaben* 2005-2018 ..... 10
Abbildung 2: Jährliches Wachstum der bewilligten Mittel von neu geförderten Vorhaben* zum Thema Digitalisierung 2006-2018 ..... 15
Abbildung 3: Bewilligte Mittel von neu geförderten Vorhaben* zum Querschnittsthema Digitalisierung 2005-2018 ..... 16
Abbildung 4: Anteil bewilligter Mittel von neu geförderten Vorhaben* zum Querschnittsthema Digitalisierung am gesamten Bewilligungsvolumen des BMBF 2005-2018 ..... 16
Abbildung 5: Anteil bewilligter Mittel von neu geförderten Vorhaben zum Querschnittsthema Digitalisierung am gesamten Bewilligungsvolumen nach Leistungsplanbereichen ..... 18
Abbildung 6: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Digitalisierung nach Leistungsplanbereichen, restriktivere Abgrenzung ..... 19
Abbildung 7: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Digitalisierung nach Leistungsplanbereichen, weniger restriktive Abgrenzung ..... 20
Abbildung 8: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Künstliche Intelligenz* 2005-2018 ..... 21
Abbildung 9: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Künstliche Intelligenz nach Leistungsplanbereichen ..... 23
Abbildung 10: Anteil bewilligter Mittel von neu geförderten Vorhaben zum Querschnittsthema Künstliche Intelligenz am gesamten Bewilligungsvolumen nach Leistungsplanbereichen ..... 24
Abbildung 11: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Soziale Innovationen 2005-2018 ..... 26
Abbildung 12: Anzahl und bewilligte Mittel von neu geförderten Vorhaben zum Querschnittsthema Soziale Innovationen nach Leistungsplanbereichen ..... 28
Abbildung 13: Anteil bewilligter Mittel von neu geförderten Vorhaben zum Querschnittsthema Soziale Innovationen am gesamten Bewilligungsvolumen nach Leistungsplanbereichen ..... 29

## Tabellenverzeichnis

Tabelle 1: Leistungsplanklassen mit hohem Potenzial für das Querschnittsthema Digitalisierung ..... 13
Tabelle 2: Textaufbereitung für Machine Learning Modelle ..... 30
Tabelle 3: Zusammenfassung der Gütekennzahlen des Machine Learning Ansatzes für die Identifikation von Vorhaben im Themenfeld Künstliche Intelligenz ..... 32
Tabelle 4: Unterkategorien der TexAn-Analyse zur Digitalisierung ..... 36

## Verzeichnis der Übersichten

Übersicht 1: Code der TexAn-Analyse zum Querschnittsthema Digitalisierung (inkl. Künstliche Intelligenz) ..... 37
Übersicht 2: Code der TexAn-Analyse zum Querschnittsthema Soziale Innovationen ..... 42

## Verzeichnis der Boxen

Box 1: PROFI-Datenbank ..... 9
Box 2: Leistungsplansystematik ..... 10