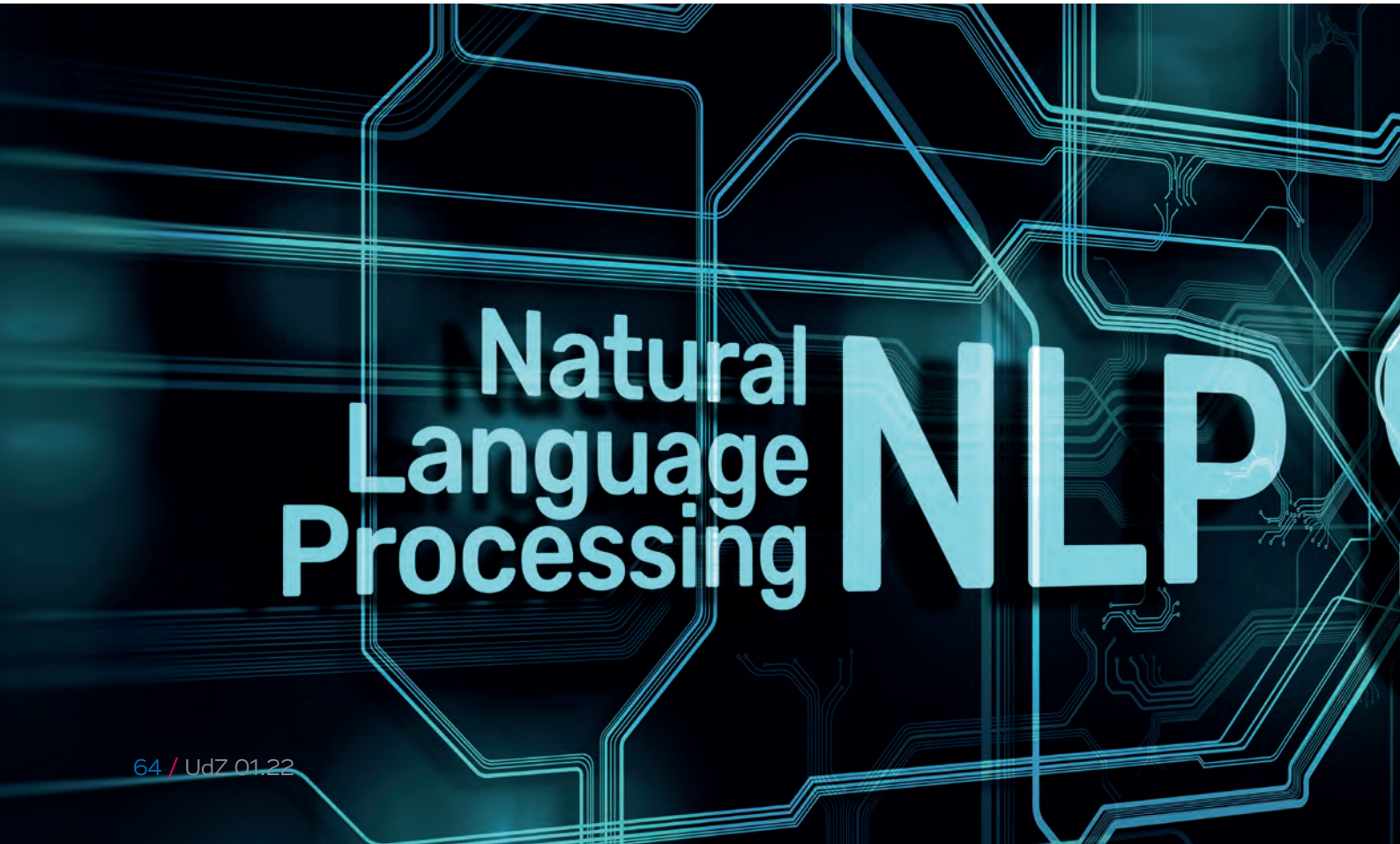
TechRad:

Unterstützung des Technologiemanagements in Unternehmen durch Natural-Language-Processing (NLP)

Technologiemanagement – die Basis für die Entscheidung über Einsatz, Entwicklung oder Beschaffung sowie die Verwertung von Technologien – kann strategische Entscheidungen eines Unternehmens maßgeblich beeinflussen und damit über dessen Erfolg oder Misserfolg entscheiden. Grundlegende Vorlage für das Technologiemanagement sind Technologieradare, inklusive der Bestimmung des (TRL), um die Reife neu eingesetzter Technologien (z. B. Newcomer vs. Etablierte) bewerten zu können. Sowohl Technologieradare als auch der TRL werden in zeitaufwendigen, manuellen Recherchen von Fachleuten ermittelt. Dieser Prozess wird aufgrund der Weiter- und Neuentwicklung von Technologien häufig wiederholt, sodass die notwendige Recherche als Daueraufgabe bestehen bleibt. Das Forschungsprojekt ‚TechRad' (Laufzeit: 01.06.2019 – 31.05.2022) zielt deshalb darauf ab, die Identifikation des TRLs sowie den Aufbau der Technologie-Radare mittels Webcrawling und Natural-Language-Processing (NLP) zu automatisieren. Im Artikel werden die Erkenntnisse aus der Entwicklung in Form eines generischen Leitfadens zur Entwicklung autonomer Technologieradare zusammengefasst. >



TechRad:

Supporting Enterprise Technology Management with Natural Language Processing (NLP)

Technology management, which provides the basis for decisions on the use, development, and procurement as well as the utilization of technologies, may have a significant impact on corporate strategic decisions and thus determine the success or failure of a company. Technology management in turn draws on so-called technology radars, which identify emerging technologies and assist in evaluating the technology readiness level (TRL) or maturity of newly deployed technologies (e.g. newcomers vs. established technologies). Both technology radars and TRLs are usually determined by experts in time-consuming, manual research. This type of research must be done over and over again as technologies are constantly being developed – it is an ongoing process. The TechRad research project (Project period June 1, 2019 to May 31, 2022) therefore aims to automate the creation of technology radars and the identification the TRLs of technologies using web crawling and natural language processing (NLP). The present article summarizes the insights from a development project in the form of a generic guide for the development of autonomous technology radars. >

Das Management von Technologien und Innovationen ist eine entscheidende Komponente für den unternehmerischen Erfolg, da es die Marktposition eines Unternehmens sichert[1]. Dem Markt immer einen Schritt voraus zu sein und die schiere Anzahl der verfügbaren Technologien zu managen ist sowohl für große als auch für kleine Unternehmen immer mehr zu einer Herausforderung geworden. Denn grundlegende Aufgaben des Technologiemanagements sind die Identifikation und Zusammenfassung von Technologien in Technologie-Radaren sowie die Bestimmung des Technical-Readiness-Levels (TRL). Dazu sind zeitaufwendige, manuelle Recherchen von Expert:innen notwendig, welche aufgrund der Weiter- und Neuentwicklung von Technologien häufig wiederholt werden müssen. Das ‚TechRad'-Forschungsprojekt zielt deshalb darauf ab, die Entwicklung von Technologieradaren sowie Bestimmung der TRLs mittels Webcrawling und Natural-Language-Processing (NLP) zu automatisieren. Im Folgenden werden die Erkenntnisse aus der Entwicklungsphase des Projekts in Form eines generischen Leitfadens zur Entwicklung autonomer Technologieradare zusammengefasst; der Leitfaden besteht aus sechs Phasen (s. Figure 1).

Zunächst muss in Phase 1 das Anwendungsfeld, für dessen Umsetzung Technologien gesucht werden sollen, mithilfe von Schlagwörtern beschrieben werden. Dieses Anwendungsfeld könnte beispielsweise Natural-Language-Processing sein. Passende Schlagwörter wären hier beispielsweise *translation* oder *sentiment analysis.*

Aus den Schlagwörtern entsteht eine Datenbasis, für deren Erstellung zwei mögliche Wege existieren: Zum einen können

[1] s. Klappert et al. 2011, S. 5

The management of technologies and innovations is a crucial component for business success, as it secures a company's market position[1]. Staying ahead of the market and managing the sheer number of available technologies has increasingly become a challenge for both large and small companies. This is because the basic tasks of technology management are to identify and summarize technologies in technology radars and to determine their technical readiness level (TRL). This requires time-consuming, manual research by experts, which must be performed again and again due to the new and further development of technologies. The TechRad research project therefore aims to automate the development of technology radars and the determination of TRLs by means of web crawling and natural language processing (NLP). In what follows, the findings from the development phase of the project are summarized in the form of a generic guide for the development of automated technology radars. This guide describes the six phases of the development process (see Figure 1).

First, in phase 1, the field of application for whose implementation suitable technologies are to be identified must be described using keywords. This application area could be, for example, natural language processing. Suitable keywords here would be 'translation' or 'sentiment analysis', for example.

The keywords are used to create a database, which can be developed in two possible ways: First, free or paid APIs such as arXiv, Springer Link, or Science Direct can be used to obtain text documents. Second, web scrapers can be

[1] Klappert et al. 2011, p. 5

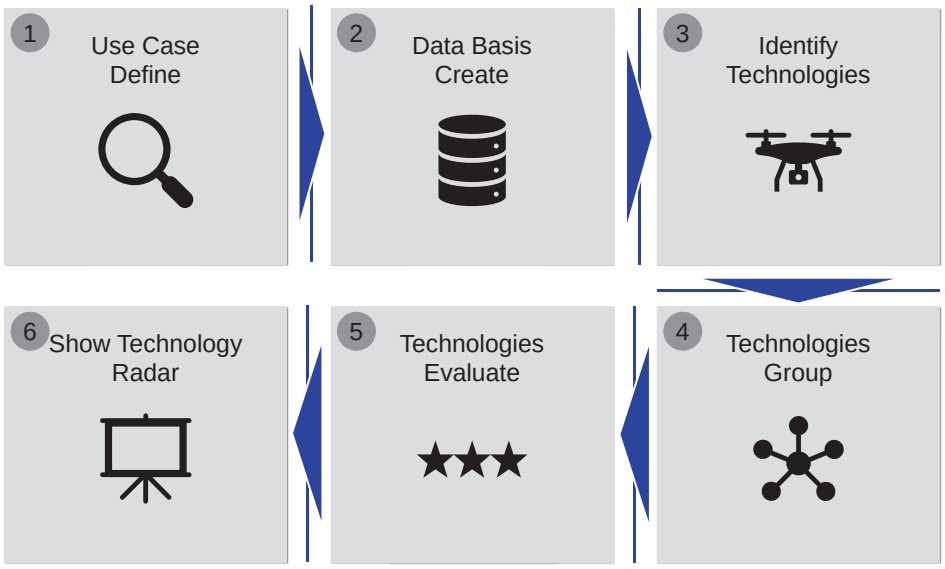


Figure 1: Phases in the development of autonomous technology radars (own illustration)

kostenfreie oder kostenpflichtige APIs verwendet werden, um Textdokumente zu beschaffen. Beispiele dafür sind *arXiv*, *Springer Link* oder *Science Direct*. Zum anderen eignet sich die Nutzung eines Webscrapers, der nach Internetdokumenten und Webseiten sucht und mit den Schlagwörtern verbundene Dokumente abspeichert. Beschreibende Schlagwörter können hierbei beispielsweise in Abstracts oder Keyword-Listen von Artikeln gesucht werden. Für die Programmierung von Webscrapern eignen sich Code-Bibliotheken der Pro-grammiersprache *Python* wie beispielsweise *Scrapy*, *Selenium* und *BeautifulSoup*.

Ist die Datenbasis geschaffen, folgt in Phase 3 die **Identifizierung** der relevanten **Technologien** für den beschriebenen Anwendungsfall. Dafür werden Methoden des Topic-Modelings eingesetzt. Dabei ist die Auswahl passender Algorithmen des Natural-Language-Processings maßgeblich für den Erfolg der Identifikationsaufgabe. Besonders gut eignen sich Transformer-Modelle (z. B. *SciBert*[2]), die mithilfe vortrainierter Sprachenmodelle komplexe, schriftliche Zusammenhänge (z. B. biologischer oder technologischer Art) in Texten erkennen können[3]. Die vorläufig identifizierten Technologien müssen abschließend als solche bestätigt werden, weshalb ein Abgleich derer mit bestehenden Technologiedatenbanken erfolgen muss. Sofern keine Datenbank verfügbar ist, müssen die Ergebnisse manuell evaluiert werden. Entsprechen die Ergebnisse der Identifikation nach einem Durchlauf der Phasen 2 und 3 nicht dem Gewünschten, sollten die Phasen beliebig oft wiederholt werden, z. B. durch die Anbindung neuer APIs (siehe Phase 2).

Nachdem ausreichend viele Technologien gefunden wurden, werden in Phase 4 die identifizierten Technologien inklusive der dazugehörigen Textdokumente nach den Schlagwörtern gruppiert, um sie im Technologieradar für den Nutzer als Sektoren anzeigen zu können. Technologien können dabei in mehreren Sektoren eines Radars erscheinen.

In Phase 5 wird für jede identifizierte Technologie das **TRL bestimmt**. Die Lösung dieser Problemstellung ist dem alten politischen Prinzip *Divide et impera* entlehnt (dt. Teile und herrsche, engl. *divide and conquer*). Hier bedeutet dies, dass eine komplexe Problemstellung in beherrschbare, lösbare Teilprobleme unterteilt wird. Die Bewertung einer Technologie basiert hier auf mehreren Textdokumenten aus der gesammelten Datenbasis. Im Umkehrschluss wird nicht direkt eine umfassende Bewertung für eine Technologie angestrebt, sondern mehrere kleine Bewertungen, die daraufhin durch eine Heuristik zusammengefasst werden. Die kleinste zu betrachtende Ebene ist nun nicht nur ein Textdokument, vielmehr sind es einzelne Abschnitte innerhalb eines Textdokuments, z. B. Paragrafen. Diese Abschnitte werden zur Bewertung mittels eines überwachten Lernansatzes (Sprachen-

used to search the Internet for documents and web pages and store documents containing the keywords. The search for descriptive keywords may focus on the abstracts or keyword sections of journal articles, for example. Python code libraries such as Scrapy, Selenium, and BeautifulSoup provide suitable tools for the programming of web scrapers.

Once the database has been created, phase 3 involves identifying the relevant technologies for the use case in question. To this end, topic modeling methods are being used. The selection of suitable natural language processing algorithms is crucial for the success of the identification task. Transformer models (e.g. SciBert), which can recognize complex (e.g. biological or technological) relationships in written texts with the help of pre-trained language models, are particularly suitable for this task. The preliminarily identified technologies must be subsequently confirmed as such – for this reason, the results must be compared with existing technology databases. If no database is available, the results must be manually evaluated. If the results of the identification process are not satisfactory after completion of phases 2 and 3, these phases should be repeated as often as desired, e.g., by connecting new APIs (see phase 2).

After a sufficient number of technologies has been identified, in phase 4 the identified technologies and the associated text documents are grouped according to the keywords in order to be able to show them as sectors in the technology radar. Technologies can appear in several radar sectors.

In phase 5, the TRL is determined for each of the identified technologies. The solution to this problem takes its cue from the ancient political principle of divide and conquer. In the present context, this means that a complex problem is divided into manageable, solvable sub-problems. The evaluation of a technology is based on several text documents from the created database. Conversely, the aim is not directly to produce a comprehensive evaluation for a technology, but rather to produce several small evaluations, which are then summarized using a heuristic. The smallest level to be considered is not a text document, but rather individual sections within a text document, e.g. paragraphs. These sections are classified for evaluation using a supervised learning approach (language model + transformer). As an example, the classes of the official TRL of NASA from 1 to 9 or the categories 'low', 'medium' and 'high' can be used. It is essential that an additional category is added, which labels the paragraph to be 'devoid of information'. The final step is to derive an overall score from this, using a heuristic over all the sub-scores of a technology. In the simplest case, using the mean is possible. More advanced approaches include

[2] s. Beltagy et al. 2019, S. 1 ff.

[3] s. Wolf et al. 2019, S. 1 ff.

[2] Beltagy et al. 2019, p. 1 et seqq.

[3] Wolf et al. 2019, p. 1 et seqq.

modell + Transformer) klassifiziert. Exemplarisch können die Klassen des offiziellen TRL der NASA von 1 bis 9 oder die Abstufung *niedrig, mittel* und *hoch* verwendet werden. Essenziell ist hierbei, dass eine zusätzliche Klasse hinzugefügt wird, welche den Paragrafen als *informationslos* beschreibt. Im letzten Schritt wird daraus eine Gesamtbewertung abgeleitet, wobei eine Heuristik über alle Teilbewertungen einer Technologie angewendet wird. Im einfachsten Fall ist das Verwenden des Mittelwertes möglich. Fortgeschrittenere Ansätze umfassen die Gewichtung der untersuchten Abschnitte – dies bedeutet, dass eine zusätzliche Klassifikation eingeführt wird, um die Gewichtung zum Einfluss der Bewertung darzustellen, etwa *niedrig* und *hoch*. Für die Trainingsphase ist es wichtig, dass beliebige Textdokumente verwendet werden, um die Abstraktion und Robustheit des Modells sicherzustellen. Die Textdokumente sollen dabei ähnliche Technologiebewertungen enthalten und mit den Klassen *informationslos* und *beliebige Klassen zur Reife* gelabelt werden.

Abschließend wird der Technologieradar in Phase 6 für den Anwender visualisiert. Dabei dienen die Schlagwörter als Sektoren des Radars, in denen die damit verknüpften Technologieradare angezeigt werden. Das TRL der einzelnen Technologien wird in Form eines Steckbriefs angezeigt. Ferner ist die Anzeige der relevanten Dokumente der Technologien inklusive Links zur Ursprungsquelle möglich. Zur Visualisierung der Ergebnisse eignen sich bekannte BI-Tools wie beispielsweise *Qlik-Sense, Power-BI* oder *Spotfire*.

cm · lc

weighting of the examined sections, i.e. an additional classification is introduced to represent the weight of a section in contributing to the evaluation, such as 'low' and 'high'. For the training phase, it is important that random text documents are used to ensure the abstraction and robustness of the model. The text documents should contain similar technology ratings and be labeled according to the 'devoid of information' category and the maturity categories.

Finally, the technology radar is visualized for the user in Phase 6. Here, the keywords serve as radar sectors in which the associated technology radars are displayed. The TRL of the individual technologies is displayed in the form of a fact file. Furthermore, it is possible also to display the relevant documents on the technologies including links to the original sources. Well-known BI tools such as Qlik-Sense, Power-BI, or Spotfire are suitable for visualizing the results.

cm · lc

References

BELTAGY, I.; LO, K.; COHAN, A.: SciBERT: A Pretrained Language Model for Scientific Text. In: EMNLP (2019). https://arxiv.org/pdf/1903.10676 (Link zuletzt geprüft: 23.02.2022)

KLAPPERT, S.; SCHUH, G.; AGHASSI, S.: Einleitung und Abgrenzung. In: Technologiemanagement. Reihe Handbuch Produktion und Management; Bd. 2. Hrsg.: G. Schuh; S. Klappert. Springer, Berlin [u. a.] 2011, S. 5 – 10.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A. ET AL.: Hugging Face's Transformers: State-of-the-art Natural Language Processing. 5. Version. 14.07.2020. https://arxiv.org/pdf/1910.03771 (Link zuletzt geprüft: 23.02.2022)

If you have any questions, please do not hesitate to contact the author.

Project Title: TechRad

Funding/Promoters: Europäische Union (EU); LeitmarktAgentur.NRW – Projektträger Jülich Forschungszentrum Jülich GmbH

Funding no.: EFRE-0801386 / IT-2-1-025

Project Partner: Deloitte Legal Rechtsanwaltsgesellschaft mbH; i2solutions GmbH; KEX Knowledge Exchange AG; RapidMiner GmbH

Website: techrad.fir.de

The project is funded by the European Union (EU).

Florian Clemens, M.Sc.
Project Manager
FIR e. V. at RWTH Aachen University
Phone: +49 241 47705-507
Email: Florian.Clemens@fir.rwth-aachen.de