International Journal of Data Science and Analytics (2025) 20:819–840
https://doi.org/10.1007/s41060-023-00483-9

REGULAR PAPER

Artificial intelligence trend analysis in German business and politics: a web mining approach

Philipp Dumbach[1] · Leo Schwinn[1] · Tim Löhr[1] · Tassilo Elsberger[1] · Bjoern M. Eskofier[1]

Received: 21 April 2023 / Accepted: 16 November 2023 / Published online: 20 December 2023
© The Author(s) 2023

Abstract
Current research on trend detection in artificial intelligence (AI) mainly concerns academic data sources and industrial applications of AI. However, we argue that industrial trends are influenced by public perception and political decisions (e.g., through industry subsidies and grants) and should be reflected in political data sources. To investigate this hypothesis, we examine the AI trend development in German business and politics from 1998 to 2020. Therefore, we propose a web mining approach to collect a novel data set consisting of business and political data sources combining 1.07 million articles and documents. We identify 246 AI-related buzzwords extracted from various glossaries. We use them to conduct an extensive trend detection and analysis study on the collected data using machine learning-based approaches. This study successfully detects an AI trend and follows its evolution in business and political data sources over the past two decades. Moreover, we find a faster adoption of AI in business than in politics, with a considerable increase in policy discourse in recent years. Finally, we show that the collected data can be used for trend detection besides AI-related topics using topic clustering and the COVID-19 pandemic as examples.

Keywords Artificial intelligence · Web mining · Trend detection · Topic clustering · Business · Politics

1 Introduction

As a result of the advances in computer science, research buzzwords like AI, machine learning (ML), and deep learning (DL) are receiving increased attention in academia [1]. This trend is underlined by the quadrupled amount of peer-reviewed AI-related journal publications during the last two decades [2]. Moreover, AI has a rising impact on our everyday life as well as on various industry sectors. As examples of the application of AI, Lee [3] names intelligent agents and software robots, which use a fuzzy-neural system to automatically collect and analyze weather information to support the weather forecast. Furthermore, in operation agent shoppers could be named that provide smart solutions in an online shop environment [3, 4]. Besides, AI is adopted in many business areas. This includes the field of predictive maintenance [5] when looking at the industrial environment, or the healthcare sector in form of robot-assisted surgery as well as in the field of automated image recognition [6].

In 2018, the German government published its AI strategy including three central goals: first to transform Europe and Germany into a leading location for AI research and its further application in various sectors, second the responsible and common good use of AI, and third to embed AI ethically, legally, culturally, and institutionally in society [7]. According to this German federal initiative, three billion euros will be invested until 2025 to reach these goals and to contribute to the development of a European AI ecosystem, and expand the quality seal *AI Made in Germany* [8].

These initiatives of the German government in AI are more recent examples that show the direct connectivity between the fields of politics and business [9, 10]. Already in the 1980s,

✉ Philipp Dumbach
philipp.dumbach@fau.de

Leo Schwinn
leo.schwinn@fau.de

Tim Löhr
tim.loehr@fau.de

Tassilo Elsberger
tassilo.elsberger@fau.de

Bjoern M. Eskofier
bjoern.eskofier@fau.de

[1] Machine Learning and Data Analytics Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universtität Erlangen-Nürnberg, Carl-Thiersch-Str. 2b, 91052 Erlangen, Germany

Yoffie [11] stated that no industry in the USA is resistant to the influence of political decisions. In political and economic research, the field of *political economy* evolved over decades and focuses on the effects of politics on the economy and investigates the influence of macroeconomic trends, but also of specific industries, on political decisions [12]. On the one hand, businesses can benefit from politics through value creation; on the other hand, the companies' success is dependent on national political nuances as well as on global politics due to market expansion [9].

After looking at the general (global) connectedness of business and politics, the focus is once again on specific activities in those sectors relating to the advancement of new technologies. It can be observed that a combination of scientific and industrial research contributes to the development process of new technologies also being dependent on the application field and sector [13]. The previous examples at the beginning of this section illustrate the impact of new technology like AI and show the importance of detecting and understanding trends in science, industry, and society to be able to respond to those trends from a business as well as a political perspective because they often evolve into new technologies in the future [14, 15].

Nevertheless, previous work focuses on the industrial-driven application of AI without considering political data sources and analyzing connections between them. Especially in the field of trend detection in AI, the current literature concerns mainly academic data sources. We argue that industrial trends are influenced by public perception and political decisions and should be reflected in political data sources. To fill this gap, we propose a web mining approach to create a novel data set including business and political data sources. In this work, we present a data-driven rather than a methodological approach. For our investigation, we extract data from five different business magazines, that are published especially in the DACH region (Germany, Austria, Switzerland) and show a high circulation, and additionally two types of political protocols from the German parliament *Bundestag*. Within this study, we are addressing the following research questions (RQ):

- *RQ1* Can we visualize the AI trend in German business and politics between 1998 and 2020 and examine its development beyond that?
- *RQ2* Can we identify specific and interpretable topics within the multiple political and business data sources using topic clustering?
- *RQ3* To what extent is political data a viable research medium for detecting trends in the field of AI and beyond?

In order to answer those research questions, we extract AI-related buzzwords from various glossaries and we use them to perform a trend detection and analysis study on the collected data using machine learning-based approaches. We use state-of-the-art algorithms based on deep learning to perform topic clustering. Furthermore, we fine-tuned a pre-trained transformer model based on the BERT architecture on political and economic data sources. The pipeline for topic clustering was built with *OCTIS* [16], an open-source technology, which builds upon the Hugging Face BERT models [17]. In the following, the main contributions of this study are summarized:

- We describe a web mining approach that was used to create a novel data set including German business and political data sources (in total 1.07 million articles and documents from 1998 until 2020).
- We identify 246 AI-related buzzwords and use them to successfully detect an AI trend and analyze the AI trend development over the past two decades.
- We investigate the adoption of AI trends in business and politics and observe a faster adoption of AI in business compared to politics with a considerable increase in political discourse in the past few years.
- We demonstrate that industrial AI trends are reflected in political data sources and identify unknown topics in business as well as in political data sources.
- We exemplify how the novel data set can be used for trend detection besides the field of AI and illustrate the transferability of the proposed approach for future research at the interface of politics and business.

Besides, we demonstrate that political data sources are an informative and from our perspective a very relevant research medium in the field of data mining, and here specifically in the field of web mining, even for non-political topics.

Reproducibility: The code for the crawlers and the data from political sources, that were used in the experiments (see Sect. 4.1), are publicly available in the GitHub repository at https://github.com/mad-lab-fau/trend-detection-data-set_politics (see Sect. 5).

2 Related work

We review the existing literature on web data mining and AI perception in general and include text mining in a political and economic context. We further outline current research in the field of buzzword identification and trend detection. In current literature, we find multiple examples of researchers scraping and mining web databases, such as the following study of Landers et al. [18], which proposes an approach to automatically extract big data from the web in psychological research. They discuss the notion of data source theories, which suggests starting with assumptions about a potential

data source in order to not only obtain a big data source but also to collect and analyze useful data. In their study, they scrape data from an online depression discussion board, which not only addresses psychologists as the main target group [18].

In addition, there are research projects that use this approach to examine large amounts of data in particular with regard to the public perception of AI. The *New York Times* articles from 1986 until 2016 were analyzed by Fast and Horvitz [14] to investigate the positive and negative change in public perception regarding the subject area of AI. In the absence of a professional, unified, and precise definition of AI, it has not been possible to perform automatic sentiment analysis. Therefore, they proceeded with a paragraph-wise manual annotation. Their results show an increase in news coverage related to AI, but also specific trends regarding the opportunities of AI. This includes healthcare and education, as well as negative attitudes related to ethical concerns or a loss of control. Over the three decades, they were able to demonstrate individual deviations from this development, such as the beginning of the *AI winter* in 1987 and finally in 1995, based on the lowest level of public discussions related to AI.

According to von Krogh [19], the adoption of AI by organizations is been driven by the following four advances: first the progress in underlying AI methods, second in capturing and storing task-related data, third reduced costs when it comes to computational power and finally the access to cloud-based services for all types of organizations. As a result, not only the public interest in the field of AI is increasing, but advances in computational power are also impacting other research areas like political sciences by making large amounts of public data, such as political minutes, freely and easily accessible.

Proksch and Slapin [20] looked at such a public database by analyzing the parties' election manifestos in Germany from 1969 until 2005 and explored a word-frequency-based position estimation technique. They applied the word scaling method *wordfish* to investigate the word frequency as they suspect a connection between the choice of words and the political ideology. It was feasible to identify the political orientation of the parties only by the frequency of certain words, even if the sentence structure itself was missing, and to create a word map including the top fifty words of both the left-wing and right-wing parties.

So far, the political environment was used not only to identify trends in the choice of words of politicians [20], but also to analyze the extent to which the official outcome of an election can be traced on the basis of social media data or public reporting [21]. The data linked to the federal election in Germany in 2013 was gained from 1.4 million Twitter messages from over 98,000 users. Relevant political terms within the analysis of Jungherr et al. [21] were party names or candidates. Their results show a deviation in Twitter-centered political coverage compared to the actual political coverage, for example, when small parties receive oversized attention on Twitter. Social media networks became powerful tools in the political investigation. Their research shows that inferences about the public opinion toward political parties based on social network data can lead to inaccurate conclusions. The previous studies show how relevant the recognition and the analysis of certain buzzwords as well as their changes over the years can be and investigate, for example, political developments and public opinion regarding a topic such as AI [14, 20].

According to the online *Cambridge Dictionary* [22], a buzzword is "a word [..] that is frequently used because of its importance or popularity especially among a certain age-group or profession." A buzzword develops from a term with a precise and specific meaning in the context of a specialist to a less precise and popular one among people outside the respective subject area [23].

Just as discussed in prior literature, Landers et al. [18] as well as Kleinberg [24] point out the challenge in the field of text mining to not only gather data, but instead extract the relevant information and structure from streams that arise again and again. Research paper archives and personal emails are examples for such document streams and are used as data source within his study. In text mining, the goal is to detect the emergence of a particular topic by measuring the so-called *bursts of activity*. Kleinberg [24] follows the methodological approach to utilize an infinite-state automaton to identify these bursts, which occur in state transitions. This study was conducted within the context of undergraduate courses and resulted in the observation of rising activity linked to assessment deadlines.

Researchers are aiming to detect a trend by the increased activity linked to a specific topic [24]. Another approach in trend analysis by Neuman et al. [35] looked at the appearance of buzzwords in comparison with well-established words or non-popular neologisms based on *blogosphere*. Blogosphere sums up all blogs and their connections and is described as a competitive field for the emergence of new buzzwords. The authors not only present an exponential growth rate but also a higher growth rate compared to the blogosphere itself. For this measurement, they used the number of appearances of *the* as the highest-ranked word in the English language. Furthermore, they contributed early warning signals to estimate the tipping point of decline of a buzzword by measuring the increasing autocorrelation between words and a unique variance pattern.

Based on the review of the diverse and wide-ranging literature from the fields of web scraping and web data mining, not only in AI trend analysis, but also related to policy and economics, and the still relatively young field of buzzword research, we did an extended literature review which is pre-

Table 1 Literature review on—web data mining, trend detection and analysis, and politics and public databases

Theme	Author(s)	Data	Methodology	Key findings
Web data mining	Yang et al. [25]	5 climatic sources, data of more than 10,000 weather stations since nineteenth century	Real-time web scraping, parsing, and filtering	Agricultural system consisting information from climatic, soil, and cropland databases
	Glez-Peña et al. [26]	3 antimicrobial peptide databases	Web scraping using the jARVEST framework	Review of scraping frameworks, web scraping as valuable tool in research projects
	Cha and Stow [27]	Twitter and Google Trends data of Toledo water incident in Aug 2014	Web data mining	Miningweb data as cost-effective, less time-consuming way to discover public perception
Trend detection and analysis	Cornwall [28]	Review collection in *The Development Dictionary*	Buzzword critique development discourse	Use of buzzwords to generalize and gain popularity, loss of subtle meaning
	Budak et al. [29]	Twitter data (41.7 million nodes, 417 million posts)	Diffusion model, structural trend detection, social network analysis	Two new structural trend definitions, Use of friendship information for topic identification
	Caled et al. [30]	DBLP database, 3 million bibliographic documents	Clustering, *k*-means, self-organizing maps	Characteristics for buzzwords, connection between buzzwords and innovation, identification of buzzwords with both approaches
	Holzinger et al. [31]	Tech sector, government funding, and AI conference attendance	Knowledge extraction, (statistical) ontologies	Mechanism in data-driven research, AI hype due to success in ML and knowledge extraction
Politics and public databases	Igan et al. [32]	Home Mortgage Conflict of interest Act, data from > 9000 lenders (1999–2007)	Quantitative analysis, data mining	Correlation between lobbying and legislation; closer monitoring of lobbying activities and reduction of political influence
	Mohr et al. [33]	11 US National Security Strategy (NSS) documents (1990–2010)	Automated text analysis, NLP, LDA	Burke's method of actors, acts and scenes; grammar of motives in NSS protocols, mapping actor relations, and political agenda
	Bunea and Ibenskas [34]	Position documents in European Commission consultations	Quantitative Analysis, Text Mining	Quantitative analysis using *wordfish* not a suitable solution for mining EU policy documents
Combined approach	Proposed work	1.07 million German business and political documents from 1998 until 2020	Web scraping, web data mining and analytics, knowledge extraction, topic clustering	Novel data set including scraping workflow, AI trend detection, faster AI adoption in business

sented in Table 1. In summary, we found multiple studies focusing on either buzzwords, the field of AI, or the political and economic domains, but there is a lack of research when it comes to addressing specific combinations of these areas of investigation. Especially in trend detection and trend analysis, social media data from networks like *Facebook* or *Twitter* is a common source. In return, business magazines or newspapers were identified as less used data sources within our extended literature review. There exist many research projects investigating the field of AI in general as well as its application in various industries [3–6]. However, when it comes to the political environment, fewer studies have been conducted investigating political trends especially linked to the field of AI. In addition, current studies have collected and analyzed data from a single type of data source, rather than including and comparing data from multiple source types and environments in the study.

3 Methodology

In this section, we motivate the adaption of our quantitative research approach and describe the selection process of the relevant data sources for this study. Furthermore, the analysis pipeline is discussed step-by-step including the phases of the data collection and the data processing as well as the buzzword identification and final creation of the AI-related buzzword list that are prerequisites for the subsequent data analysis. This section ends with the description of the procedure to perform topic clustering on the collected data.

3.1 Data collection and processing

Prior research showed that huge quantities of data lead to more generalizable results for quantitative data analysis [36, 37]. Therefore, for our analysis, we created an extensive data set from two different data sources. Namely, we extracted data from multiple different business magazines and parliamentary data over a period of more than two decades. The economic-related data was crawled from five different magazines, that are published and read especially in the DACH region (Germany, Austria, Switzerland), namely *Brand Eins (BE)* [38], *Focus Money (FOCM)* [39], *Handelsblatt (HB)* [40], *Manager Magazin (MM)* [41], and *Wirtschaftswoche (WW)* [42]. The decision was based on the circulation of the magazines, which is defined as sold editions per specified time. The German Audit Bureau of Circulation (IVW) reports the ten most published German business magazines [43]. According to the relevance and availability, four out of these ten magazines were selected as well as the *Handelsblatt* due to its volume and strong economic focus. The circulation of Austrian and Swiss business magazines was lower compared to the German ones and therefore not further considered. In summary, the articles, used as data set within this study, were published from January 1998 until December 2020 varying on a daily, weekly, or monthly basis.

Figure 1 illustrates the downloading process of the business magazines. First, the online archives and the respective overview pages of the individual years were accessed to iterate a year list. The same procedure was conducted for the volume and article overview pages before scraping the text blocks from the article pages. In the end, the raw text data for each individual article was saved in a separate file including the information about the buzzword occurrences, word counts, and dates.

The now further described processing steps were identical for all five business magazines to create a pandas [44] data frame for each business magazine. The individual files for each article were named to have a unique identifier for each article. Further columns were added to keep track of the affiliation of an article to the respective magazine, to represent the article's text, and to count the number of words in each article. The information was passed to a pandas data frame by iterative looping through all text files. As a result, a single data frame combining all information was created for each magazine and was modified according to the following step-by-step procedure:

1. The text was converted to lower case.
2. Special characters were replaced with blank space (exception for dashes ('–') having a specific meaning).
3. Multiple blank spaces were replaced by a single blank space.
4. The *Null* or *None* values were transformed to Zeros.
5. The blank space amount was counted to determine the number of words.
6. The text column was dropped after the relevant information was extracted.
7. The zero data columns and rows were removed to reduce the size of the data frame.

The parliament protocols are directly available on the web page of the German parliament *Bundestag*. Here, two types of political protocols are differed: the *Drucksachen (DRS)* [45] which summarize, for example, law drafts, inquiries, or surveys composed by individuals or by parties in the parliament, and the *Plenarprotokolle (PP)* [46], that are word-by-word transcripts of the individual plenary sessions. These documents exist since 1949, but according to the available business magazines, the protocols were used beginning with the election period 14 that started in 1998. The majority of the political data (election periods 14 to 18) was available in a condensed and ordered XML format and could therefore be manually downloaded. For election period 19 (which ended in 2021), the web page structure changed and the scraper was implemented similarly to the magazines' static

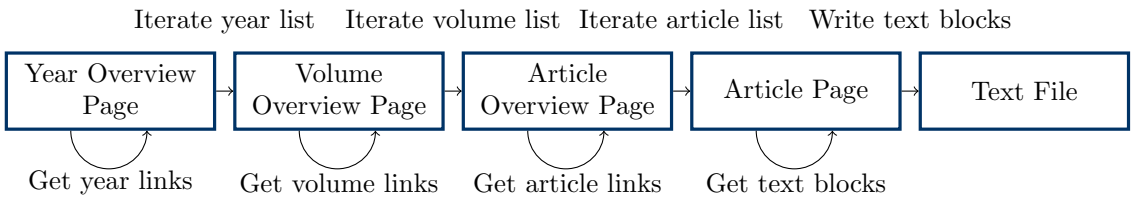


Fig. 1 Scraping and processing of the business magazines — step-by-step downloading workflow

HTML pages. For this last period, the processing procedure was comparable to the magazines. Therefore, the raw text files were saved in the same format as the business magazines to be compatible with the magazine processing scripts.

3.2 Buzzword identification and processing

The text data from the political protocols and business magazines is mostly written in German. Nevertheless, the primary language in the field of computer science as well as in related disciplines is English. Therefore, buzzword lists in both languages were generated. The following sources were used for the English glossary:

- Gartner Hype Cycle 2016–2020 [47–49],
- Microsoft Machine Learning Glossary [50],
- Accenture Applied Intelligence Glossary [51],
- Google Machine Learning Glossary [52],
- Stanford Machine Learning Glossary [2],
- Wtf Machine Learning Glossary [53],
- Oxford Dictionary for Computer Science [54].

The English buzzwords were manually taken from the given graphs, simply copied in the buzzword file, or extracted with the parsing library *BeautifulSoup* (BS4) [55] and afterward trimmed with the VIM text editor [56]. The procedure was similar for the German buzzwords as they were copied as well or extracted again by using *BS4*. The further sources for the German glossary were:

- PLS Glossary [57],
- Microsoft Maschinelles Lernen Glossary [58],
- Clickwork Glossary [59],
- Gulp Glossary [60].

After the extraction of the individual buzzword lists into respective files, a joined list of both languages was compiled. The created list was still in raw form and was further processed. For the English buzzword list in total 7249 words were collected, but 85.5% of the words were added from the *Oxford Dictionary of Computer Science* with a wide range of topics besides the field of AI. Therefore, the buzzword occurrences were counted. Only buzzwords that appeared in at least two different glossaries were further utilized and duplicates were removed. High-occurring English terms were translated and added to the German list. In the further analysis, words like *IoT* and *Internet of Things*, *ML* and *machine learning*, or *AI* and *Artificial Intelligence* (including the respective German translations and German abbreviations) were considered combined. Overall, the final list included in total 246 unique buzzwords in English and German (see "Appendix B"). This list of AI-related keywords is used for later analysis and helps in the process of defining the topic clusters.

As part of the experiments and analysis, we calculated the relative buzzword occurrence to investigate the total frequency of the buzzwords. In addition, we followed Cao et al. [61] and chose the term frequency–inverse document frequency (TF-IDF) as a metric to evaluate the relative importance of words. The TF-IDF value functions as an indicator to filter out common words [61].

3.3 Topic clustering

A topic cluster is a collection of content that revolves around a common theme [62]. Topic clustering refers to the automatic extraction of topic clusters from data using machine learning. One of the first approaches, called *latent Dirichlet allocation* (LDA), was introduced by Blei et al. [63] and is based on a Bayesian model to reduce the dimensionality of word embeddings, to group words into clusters as well as to differentiate between various topics.

As a supplement to the previous explanations, we would like to outline that among the political data sources, the *PP* covers various topics that have a strong link to the recent public interest, whereas the *DRS* contains documents with more specialized topics. For this analysis, we split the business magazines into the *HB* and the other magazines. In a preliminary experiment, we found that *HB*-based topics dominated the magazines otherwise.

Within this study, we use state-of-the art algorithms based on deep learning to perform topic clustering. Particularly, we fine-tuned a pretrained transformer model based on the BERT architecture on political and economic data sources. The pipeline for topic clustering was built with *OCTIS* [16], an

open-source technology provided on GitHub, which builds upon the Hugging Face BERT models [17].

Topic clustering requires that the number of topics is preset before the training phase is started. In accordance to prior work, we used the metrics *Coherence* and *NPMI* to estimate the ideal number of topics for each of the data sets [16]. The respective number of topics was used for the final training and inference process. Each extracted topic includes a list of ten words in string format. In order to name topics, we used the first five topics and connected them with underscores. A single string could now describe each topic. For the *PP*, the 20 topics are defined by the ten most contributed words to this topic, but not by a *generic term*, which needs to be self-defined (see "Appendix D"). The trained model can be used to extract an embedding vector $z \in [0, 1]^T$ for a given text, where the magnitude of every entry indicates the affiliation with one of the topics. In this context, $T \in \mathbb{N}^+$ is the preset number of topics. The embedding vectors can be used to calculate the similarity between two different chunks of text and extract the associated topics (see Sect. 4.4).

More information about the data source and data set building, the text cleaning pipeline as well as the training, optimization, and inference steps related to the topic clustering is given in more detail in "Appendix C".

4 Data set, results, and discussion

In this section, we present the novel data set and give a detailed view on the text distribution according to the seven different data sources as well as show the word count development from 1998 until 2020. Furthermore, we perform a baseline analysis of the proposed data set. For this purpose, we normalized the buzzword occurrences of the different data sources to the total number of words in each data source. This was done because the total number of buzzwords differs considerably between the different data sources (e.g., *PP* and *DRS*). In the following experiments, we always describe the relative total occurrence of the AI-related buzzwords.

4.1 Data set — overview

The distribution of the collected text data within this study is illustrated in Fig. 2 individually for the five business magazines (*FOCM*, *HB*, *BE*, *WW*, *MM*) as well as both types of German political protocols (*DRS*, *PP*).

The imbalance between the business and the political part of the data set is marginal. The political protocols, which combine *DRS* and *PP*, have a total of 52.6% compared to the business magazines with 47.4% of the overall collected text data. The generated data set consists of a bit more than 1.07 million articles and documents with a total of about 1.05 billion words. With this total amount of words, the data set

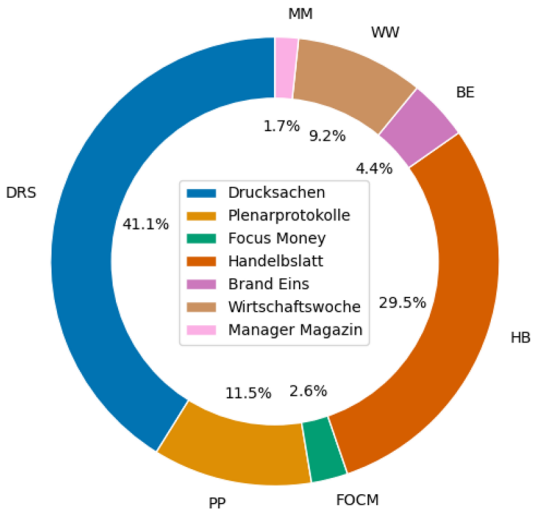


Fig. 2 Distribution of the total word count in the different data sets split into the political protocols and the business magazines

is large enough to overcome the statistical variance, which proves to be problematic, especially working with smaller data sets [64]. When looking at the average word distribution, around 6300 words were counted in average per political document compared to a bit more than 500 words in average per business article. Nevertheless, an imbalance between the data sources can be observed. When looking at the business data coming from the magazines, the *HB* contributes 29.5% of the overall text data compared to only 1.7% of the *MM*. Within the political protocols, the *DRS* provide 41.1% of the text in the data set compared to 11.5% of the *PP*.

Figure 3 gives a more detailed view on the yearly growth of the overall data set as well as the individual contribution of the seven data sources to the overall word count and its yearly development between January 1998 and December 2020. The business magazine *WW* had the latest first publication date, which was in January 2003, of all used data sources within this study.

The development of the total number of words in the area of business magazines shows a strong increase between 1998 and 2004, which, however, can only be partially attributed to the publication of new magazines in the early 2000s. In contrast, the annual development of the number of words in the two data sources from the policy context is initially relatively constant and a clear jump is then only discernible again from 2006 and onwards.

4.2 AI trend detection and development

In the context of RQ1, we wanted to detect and visualize the AI trend in German business and politics and then further track the AI trend development over the period of time between 1998 and 2020. As mentioned in the methodology

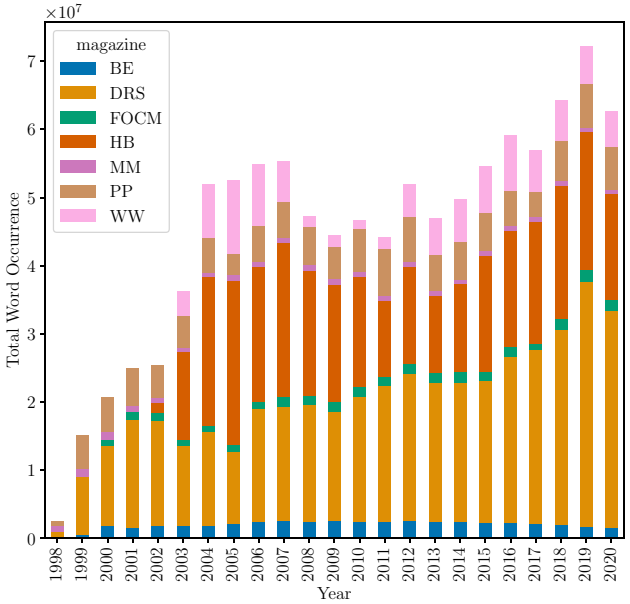


Fig. 3 Word count development for the individual sources and the complete data set from 1998 to 2020 (for abbreviations see Sect. 4.1)

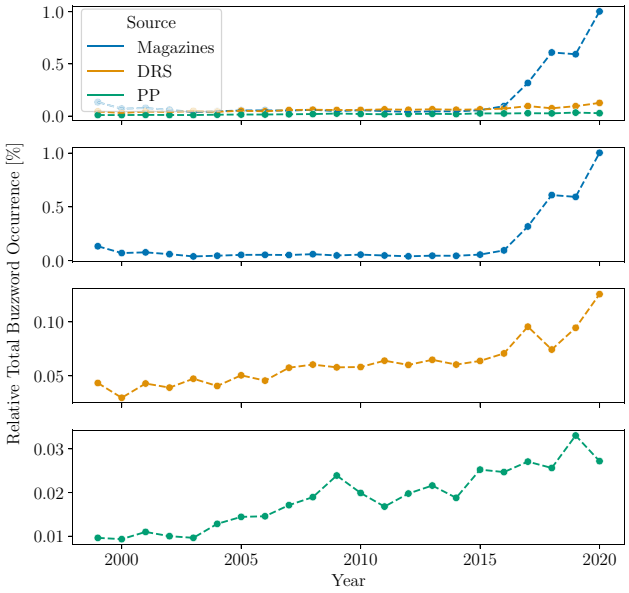


Fig. 4 Relative occurrence of combined AI-related buzzwords over the analyzed period

section, we identified 246 unique German and English AI-related buzzwords that amounted to the final buzzword list used in the data analysis step.

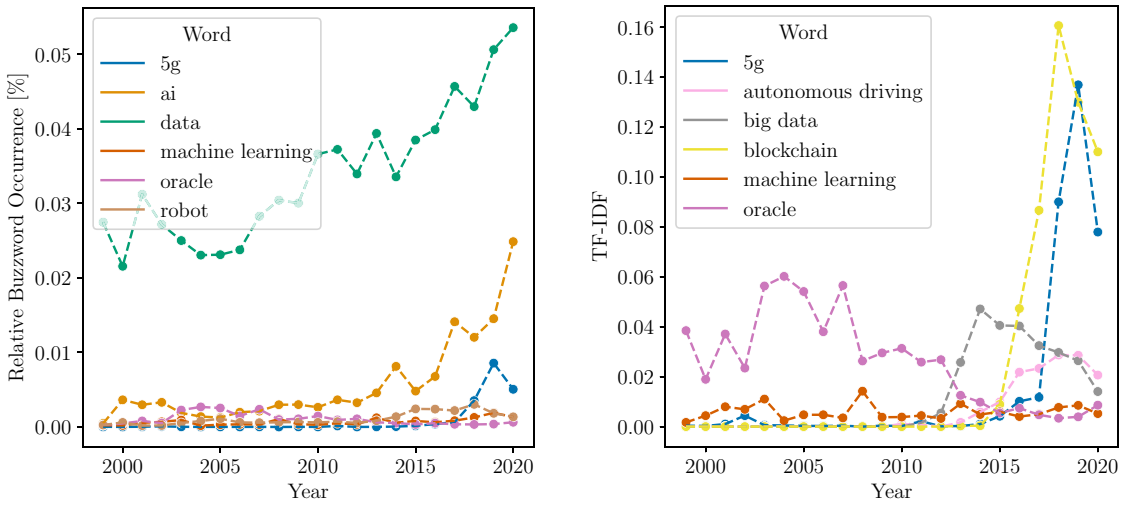
In this section, we are looking first at this combined list of all identified and finally selected AI-related buzzwords. Figure 4 displays the relative total buzzword occurrence for the different data sources: business magazines, *DRS*, and *PP*. The business magazines show a substantial rise in relative occurrence from 0.056 to 1.003% starting at 2015 until 2020 with a small drop between 2018 and 2019. The *DRS* demonstrates a more evenly distributed increase in the relative occurrence of AI-related buzzwords from 0.043% in 1999 to 0.126% in 2020. Similar to *DRS*, *PP* displays a relatively evenly distributed increase from 0.010 to 0.0272% from 1999 to 2020. Note that while *DRS* and *PP* show an earlier rise of relative occurrence, the relative occurrence of AI-related buzzwords is already 0.133% in 1999 for the business magazines, which is higher than the maximum of the other political data sources. Nevertheless, this may stem from a wider range of topics in the *DRS* and *PP* compared to the business-oriented magazines.

When looking at all three data sources in Fig. 4, the trend development based on AI-related buzzwords is clearly upwards from 2015 on. We are able to make the recent hype and increase of interest in the field of AI visible. There are multiple reasons that could have influenced this development. First of all, a general interest in AI and related technologies, for example, in healthcare and industry is observable [5, 6]. Furthermore, the *Industrie 4.0* initiative was introduced in Germany in 2011 with a strong focus on emerging technologies and the industrial transformation within this so-called fourth revolution [65]. Third, the German government passed the national AI strategy in 2018 with the goal to transform Europe and Germany into leading locations in the research and development of AI technology [7].

In addition to the trend of the total relative occurrence of all AI-related buzzwords, we analyzed the trend of the top six most abundant buzzwords in all data sources combined in the presented data set. Within this experiment, we looked at two metrics: the relative buzzword occurrence to investigate the total frequency of and the TF-IDF to evaluate the relative importance of the buzzwords. Figure 5 shows the comparison of the individual results of both metrics from 1998 to 2020. Figure 5a summarizes the relative occurrence of the buzzwords *5g*, *AI*, *machine learning*, *data*, *oracle*, and *robot*. Only the words *data* and *AI* show a rising trend over the whole analyzed period. The relative occurrence of *5g* increases from 2017 to 2019 but decreases from 2019 to 2020. The other words *machine learning*, *oracle*, and *robot* demonstrate no clear trend that could have been detected.

The word *data* is not only been used as an AI-related buzzword but is also used as a common term linked to various topics in business as well as politics. This explains the enormous gap between *data* and the second most frequently used key term *AI* in this evaluation.

In Fig. 5b, the relative importance based on the TF-IDF metric for the buzzwords *5g*, *autonomous driving*, *big data*, *blockchain*, *machine learning* and *oracle* was calculated. The results in the top six differ from those in Fig. 5a. In particular, the terms *blockchain* and *5g* are experiencing a tremendous increase from 2015 and 2017, respectively. They reached



(a) Top 6 - Relative Buzzword Occurrence (b) Top 6 - TF-IDF

Fig. 5 Top 6 — AI-related buzzwords (complete data set) by considering the metrics **a** relative buzzword occurrence and **b** TF-IDF

the highest TF-IDF value in 2018 and 2019. In addition, the words *big data* and *autonomous driving* do not show such strong increases in the development compared to the two terms just mentioned, but here, too, increases in the assessment of relative importance can be observed, particularly in the period from 2011 to 2014 concerning *big data* and from 2013 to 2019 concerning *autonomous driving*. These results show the existence of a trend linked to the field of AI, that we are able to detect. At first, we are not differentiating between the individual political and business sources, but are looking at the whole data set. Furthermore, we observe here as well a rising trend development over the more than two decades since 1998, especially when looking at the relative buzzword occurrence of the AI word group from 2015 and onwards. This development of the trend, partly already from 2013, very strongly visible from 2015, is again additionally underpinned by further observation based on the TF-IDF metric. These findings go along with the existing literature, which describes an increased impact of AI on our everyday life and its adoption and application in products as well as in industrial processes [3–6]. We investigated the relative occurrence and the TF-IDF as shown in Fig. 5 not only for the top six AI-related buzzwords but for the complete buzzword list (see "Appendix B"). At this point, a supplementary look at the top 15 AI-related buzzwords shows that their relative occurrence at the lower end of the list is already too low to be able to detect a trend or derive further development from this (see "Appendix A").

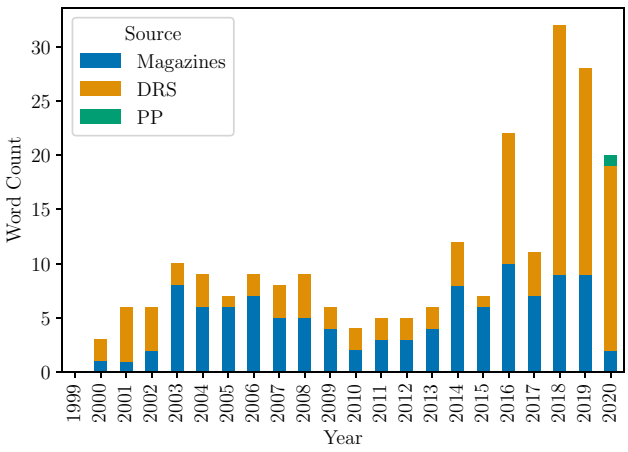


Fig. 6 AI-related keyword development (first 10% occurrence)

4.3 Comparison of the AI trend in business and politics

In the following subsection, we now concentrate on the comparison of the detected AI trend and its development by analyzing the data separately for business and politics to further address RQ1. We further explored at which point in time 10% of the total individual word occurrence was achieved for all different data sources (e.g., for a word with 1000 occur-

rences in all data sources in total, at what point did the word occur 100 times for one of the data sources?). This analysis indicates how fast the different data sources adapt the buzzwords and thus the AI trend. Figure 6 shows the number of words for each data source that reached 10% of the total word occurrence in the overall period between 1998 and 2020.

Generally, business magazines show a faster adoption of AI-related buzzwords from 2003 until 2015 compared to *DRS* and *PP*. Nevertheless, a decline in the development is observable in 2009, when looking at the business magazines. Slowly increasing again after 2010, the results show that the magazines reached again the fast adoption level as it was already identified for the first decade of the 2000s. Until 2015, the *DRS* show a relatively constant level regarding the adoption of AI-related buzzwords. In 2016, the *DRS* start to rapidly adapt buzzwords linked to the field of AI and until the end of our investigation period completely dominate the keyword development compared to the other data sources of the business magazines and also to the political documents summarized in *PP*. The *PP* only show one word in 2020 (*AI*), where the occurrence was a least 10% of the total occurrence. This can be explained by the overall lower word count for the PP (see again Fig. 3).

These results are also reflected in the relative occurrences of individual buzzwords in the different data sources that are displayed in Fig. 7a. Here, we have omitted the word *data* because it is more ambiguous than the other buzzwords in the previously presented list of the overall top six AI-related buzzwords, which was already discussed in the subsection before. For all sources, only the words *5g* and *AI* show an increase of the relative occurrence over time. The business magazines show a higher buzzword occurrence from the usage of the buzzwords between the year 2000 until 2012. In contrast, the *DRS* demonstrates a low relative buzzword occurrence until 2007 and before 2007 only show one burst of the word *AI* in 2002. For the *PP*, the relative occurrence of the top five most abundant buzzwords remains low until 2017 and increases until 2019 with a decline in 2020.

In the analysis of the different data sources regarding the AI trend, in addition to the relative buzzword occurrence just discussed, the evaluation of relative importance is performed using the TF-IDF metric. The results are shown in Fig. 7b. This evaluation was performed with respect to the same top six AI-related buzzwords as in the experiment presented in Fig. 5b. For all three sources, the words *5g* and *blockchain* show the clearest increase based on the TF-IDF value. Looking again first at the business magazines, the TF-IDF of the buzzwords is higher from the beginning of the 2000s to 2013, similar to the evaluation of Fig. 7a. The *DRS*, on the other hand, shows very low values from the beginning of the period under consideration until about 2017. However, this development is interrupted by three smaller peaks, which are visible in 2003, 2008 and 2013 following, for example, the word *machine learning* again in Fig. 7b. In the *PP*, TF-IDF values stay low for the top six buzzwords, with one exception *big data* in 2014, until 2017 and increase in the following years.

A connection between business and politics is visible when looking at the top AI-related buzzwords. Over the two decades, political documents and business magazines have followed the same patterns. In 2020, we see a drop in occurrences for the business magazines as well as in the *PP*. This development could be linked to the beginning of the global COVID-19 pandemic starting at the end of 2019 [66]. The magazines and the *PP* indicate here that they reflect the current public interest and a publicly visible debate to a much greater extent. In contrast, the *DRS*, which consist of draft legislation, inquiries, or surveys, arguably shows less dependence on the day-to-day business of society or, for example, the emergence of the COVID-19 pandemic. Therefore, the buzzword occurrences are further growing.

When looking again at Fig. 7b, differences in business and politics can be observed regarding the adoption of AI. The word *blockchain* showed an increasing development in all data sources, beginning in the business magazines in 2015, and in both political data sources with a delay in 2017. While the curve decreases, almost simultaneously, for the magazines and *DRS* until the end of the period under consideration, the *PP* still shows an ongoing upward development regarding this specific buzzword.

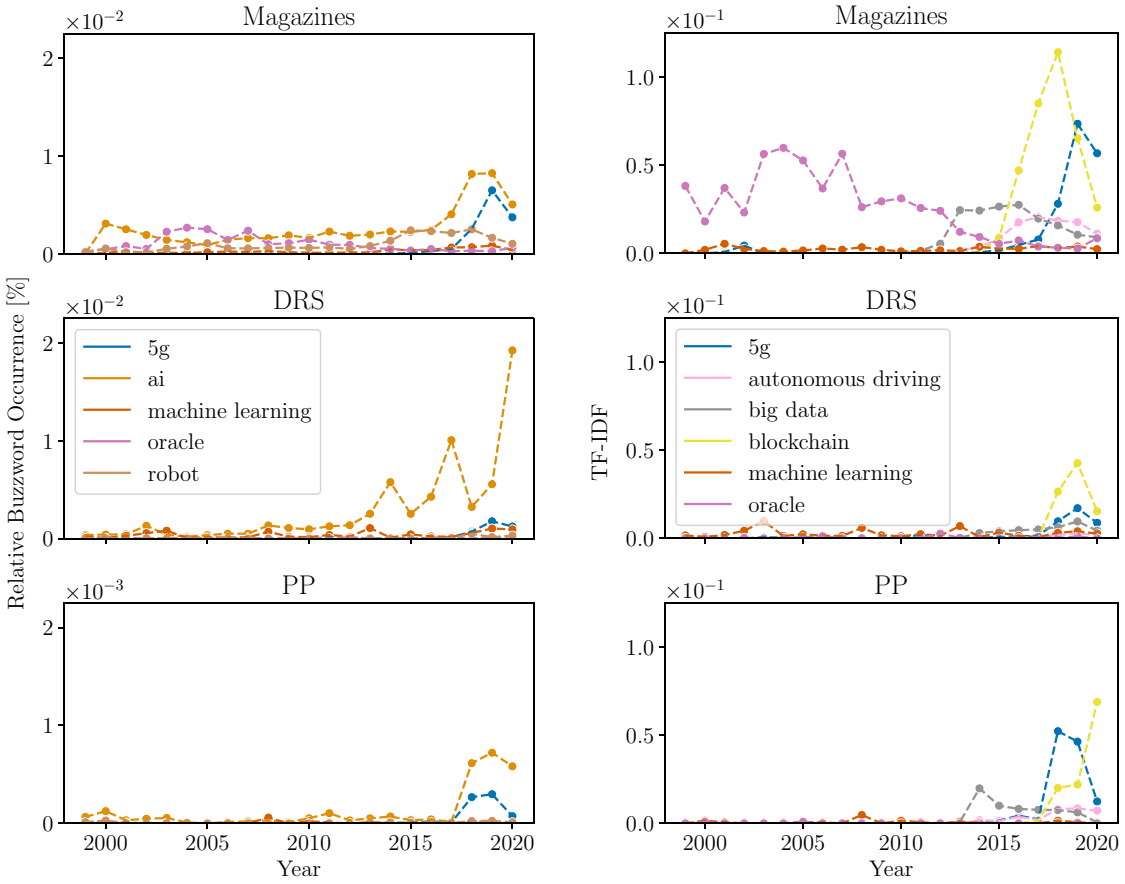
In Sect. 4.5, we will have a closer look at the potential influence of COVID-19 on the AI trend development in German business and politics.

4.4 Untargeted analysis using topic clustering

Toward investigating RQ2, we used topic clustering to identify specific and interpretable topics within the multiple political as well as business data sources. For this purpose, we calculated the cosine similarity between individual embedding vectors z of documents and articles to reveal which deal with similar topics. The visualization of all identified topics for these four separate data sets was not possible with in one plot for each single data set. Therefore, we present in Fig. 8 a selection of the most relevant topics for the individual data sets over the period from 1998 until 2020. The scale is a color bar with a range from 0 to 1. The closer a value comes to 1, the stronger the similarities are.

As illustrated in the methodology section, all of the topics for these four data sets are presented in the form of *generic terms*, which were self-defined based on the list of ten words. This list was, respectively, identified for each topic in string format. In "Appendix D," the identified topic strings as well as *generic terms* are summarized separately for the most important topics of the four data sets.

For the *DRS*, the most relevant topics related to this study are *AI*, *Cyber Security*, *Data Regulations*, and *Data Surveil-*



(a) Relative Buzzword Occurrence (Top 6 without *data*) (b) TF-IDF (Top 6)

Fig. 7 Comparison of **a** relative buzzword occurrence and **b** TF-IDF with the respective top six according to the used metrics in the individual sources *Magazines*, *DRS* and *PP*

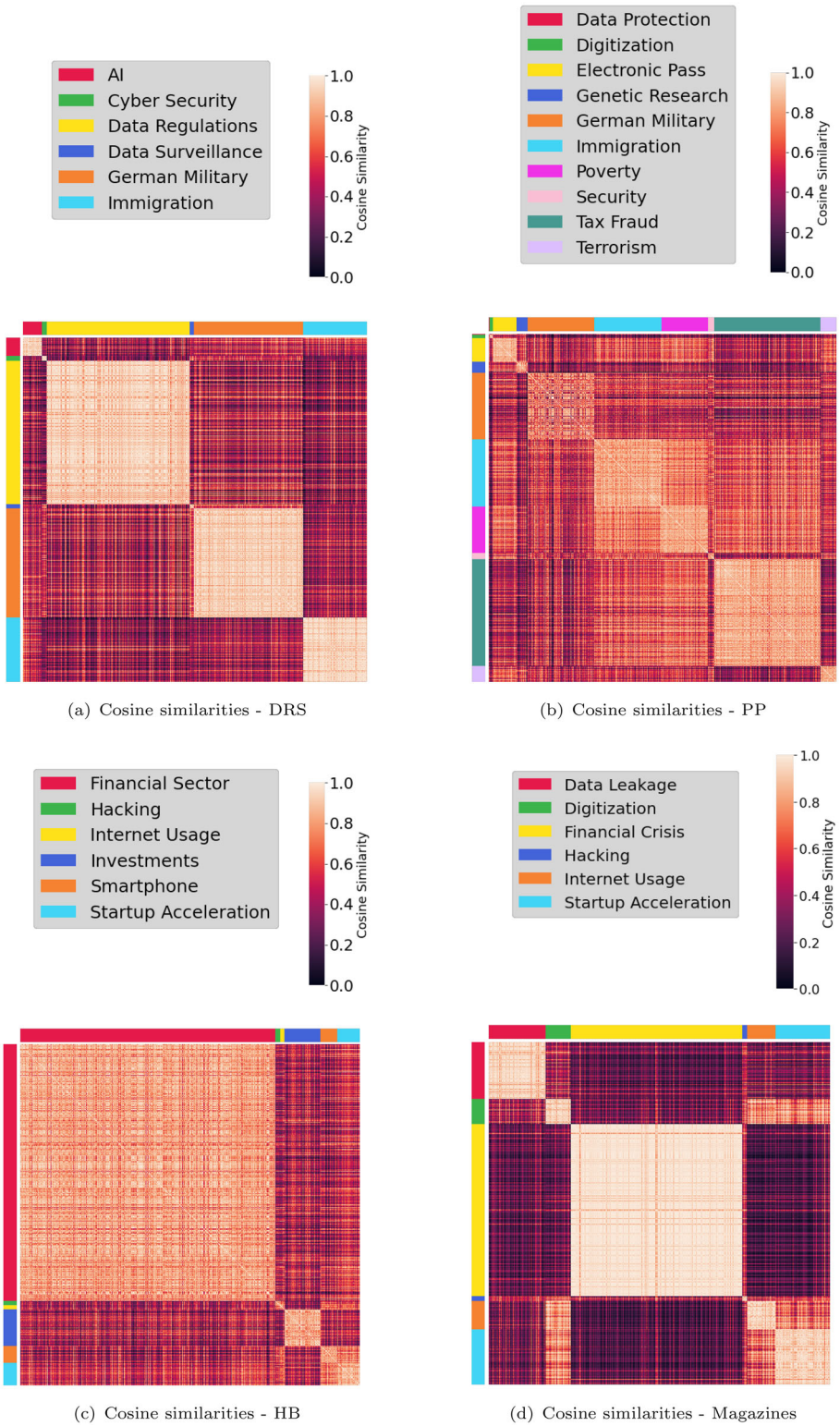
lance. All of these topics have a strong link to the research focus of this study in terms of trend detection in the field of AI. In addition, the topics *German Military* and *Immigration* were picked due to their relevance in the political debate and social discussion during the last decades [67, 68]. In Fig. 8a, these six topics are shown with color bars in the top row and most left column. When looking at the *DRS*, four distinct clusters were formed on the heat map. The two biggest clusters are located in the middle of the plot, which are representing *Data Regulations* and *German Military*. The two smaller clusters in the corner on the top left and bottom right corner, respectively, denote *AI* and *Immigration*. For the two further topics, the clusters are not clearly visible.

In addition, we observe in Fig. 8b, which shows the results for *PP*, that there are differences within both political data sources regarding the number of most relevant topics. Furthermore, we are able to identify a higher amount of clusters, but we notice equally the smaller size of the very same compared to the clusters in *DRS*. Besides, the identification of the biggest cluster, for example, in *PP* representing *Tax Fraud* or *German Military*, we see also strong similarities between differing topics like *Poverty* and *Immigration*, which indicates a strong connection in political debates. In the business data sources, similar results can be observed for the topics *Internet Usage* and *Startup Acceleration* in the magazines as shown in Fig. 8d.

Looking at the *PP*, we see that politics, and here in particular the German parliament, deals with a much wider range of issues and topics. However, these topics do not result in such large clusters, for example, with a focus on the financial sector, as we can observe in the business magazines and the *HB* (see Fig. 8c) in particular.

4.5 Influence of COVID-19 on AI trend and further exploration of the data set

In December 2019, the coronavirus 2 (SARS-CoV-2), which causes the respiratory disease COVID-19, was first identified in Wuhan, China. At the beginning of 2020, the first cases were also noted in Germany, like an outbreak in Bavaria, end of January, shows. A similar development was observed in various countries all over Europe and is seen as the beginning



(a) Cosine similarities - DRS

(b) Cosine similarities - PP

(c) Cosine similarities - HB

(d) Cosine similarities - Magazines

Fig. 8 Cosine similarities for the most relevant topics in all four data sets

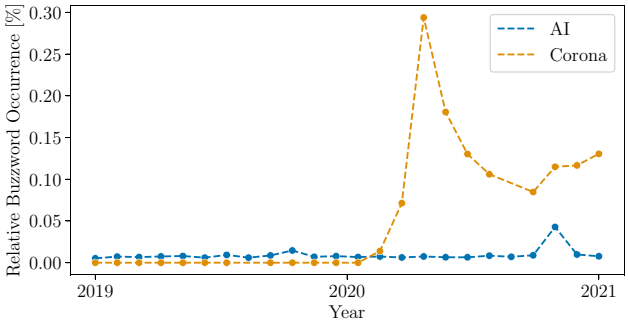


Fig. 9 Comparison of the AI and Corona trend (data *DRS*)

of the still ongoing worldwide pandemic [66]. On the 25th March 2020, the German parliament officially announced the epidemic situation of national importance [69].

Within RQ3, we explore if political data is a viable research medium for trend detection in the field of AI as well as in general. Therefore, we wanted to explore in an additional experiment if we on the one hand can investigate the influence of the COVID-19 pandemic on the trend development in the field of AI in German business and politics and on the other hand find further trends unrelated to AI in the proposed data set. Because of that, we analyzed the relative occurrence of the word *Corona* from the end of 2019 until the beginning of 2021. Figure 9 compares the trend development of the buzzwords *AI* and *Corona* during this 2-year period. Similar to the previous word groups that are described in the buzzword identification and processing section, the buzzword *Corona* combines here the terms *corona*, *covid*, *covid-19*, *pandemie*, *corona-pandemie*, *coronavirus*, and *covid-19-pandemie*.

Before the beginning of 2020, we were not able to detect a (rising) appearance of the *Corona* word group, which then shows an enormous increase by looking at the relative buzzword occurrence. After the first months in 2020, the *Corona* trend is declining and keeps a constant level from mid of year until the end of 2020. This observed increase in the corona trend development underscores the very high public profile of the COVID-19 pandemic and its dominance in both public and policy debates. Nevertheless, despite the particular global situation in the last 2 years and the Corona trend demonstrated here, we cannot detect any impact of the COVID-19 pandemic on the AI trend itself.

5 Conclusion, limitations, and outlook

In previous literature, AI trends and applications as well as the impact of AI are mainly considered in an industrial context. In addition, social media or academic data sources have been the primary means of identifying trends to date. We fill the gap in trend research and create a novel data set beyond these data sources. We publish code on GitHub to enable other researchers to recreate the used data set (see Sect. 5). We utilize the novel data set for a research study, where we investigate the hypothesis that industrial trends are influenced by public perception and political decisions and should therefore be reflected in political data sources. Therefore, we propose a web mining approach to collect and to examine the data from German business and politics between 1998 and 2020. We successfully detect an AI trend and demonstrate its progression over more than two decades. The results show a faster adoption of AI in business compared to politics. In contrast, we see an increase in the political environment during the last year. We further exemplify the usage of the collected data set for general trend analysis beyond the scope of AI and deliver insights regarding the suitability of political data sources in order to investigate trends by practitioners and decision makers. Here, we extract the most important subjects in each data source using a machine learning-based topic clustering approach. Additionally, we look at the corona development in recent years and cannot observe an influence of the COVID-19 pandemic on the AI trend in politics. Besides our focus on AI, this proposed approach is transferable for future research on any kind of topic, especially in the application field of trend detection and trend analysis, working with the same data set at the interface of politics and business.

One of the limitations of this study is linked to the data collection phase and needs to be discussed in terms of data availability. The collected data was often error-prone due to errors in the databases of the business magazines and of the German federal parliament as well as due to the scraping process. We addressed this challenge with time-consuming manual error finding and bug fixing. It was initially necessary to familiarize oneself with a completely new data set by combing through the data and generating basic statistics. This procedure simplified and accelerated subsequent processes. As described in the methodology section, we are following a quantitative research approach. Nevertheless, it is necessary to get familiar with language idiosyncrasies, especially when working with buzzword lists in German and English. Therefore, the literature review could be extended with a stronger focus on linguistics, politics, and political structures in Germany and beyond as complex research environments for data scientists.

In future work, we could imagine applying our approach to documents from the scientific community (e.g., paper submissions to research conferences or journals) as a third data source besides business magazines and political documents. The same analysis could be performed as within this study to further detect an AI trend development in additional sources and to observe potential connections between industry, politics, and academia. The proposed approach could be applied here alongside traditional approaches in trend analysis to fully consider trends. Furthermore, the data set could be increased not only by collecting the articles and protocols

that were published up to now but also by adding further business magazines from Germany, ranked by the *IVW* [43], as well as from the German-speaking countries Austria and Switzerland to represent the whole *DACH* region. We still see a research gap to investigate data from multiple sources. These sources could differ in form of the environmental context (e.g., industry, politics) or in form of textual and non-textual data (e.g., podcasts, briefings). After the transcription of media files with a speech-to-text engine, for example, the relative buzzword occurrences could be compared. In our work, we focused on the field of AI, but the proposed approach is not limited to this research field and could be used to analyze the same or an extended data set according to further topics of interest in the research community.

Acknowledgements Bjoern M. Eskofier gratefully acknowledges the support of the German Research Foundation (DFG) within the framework of the Heisenberg professorship program (Grant Number ES 434/8-1).

Author Contributions PD conceived the proposed project and method, conducted the experiments (partly in form of a master thesis supervised by the first author) and wrote the initial draft of the paper. LS, TL, and TE discussed the results and reviewed the paper. All authors analyzed the results and contributed to the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials The code for the crawlers and the data from political sources are included in the GitHub repository: https://github.com/mad-lab-fau/trend-detection-data-set_politics. For more information about the data crawling procedure regarding the business data sources please contact the corresponding author via philipp.dumbach@fau.de.

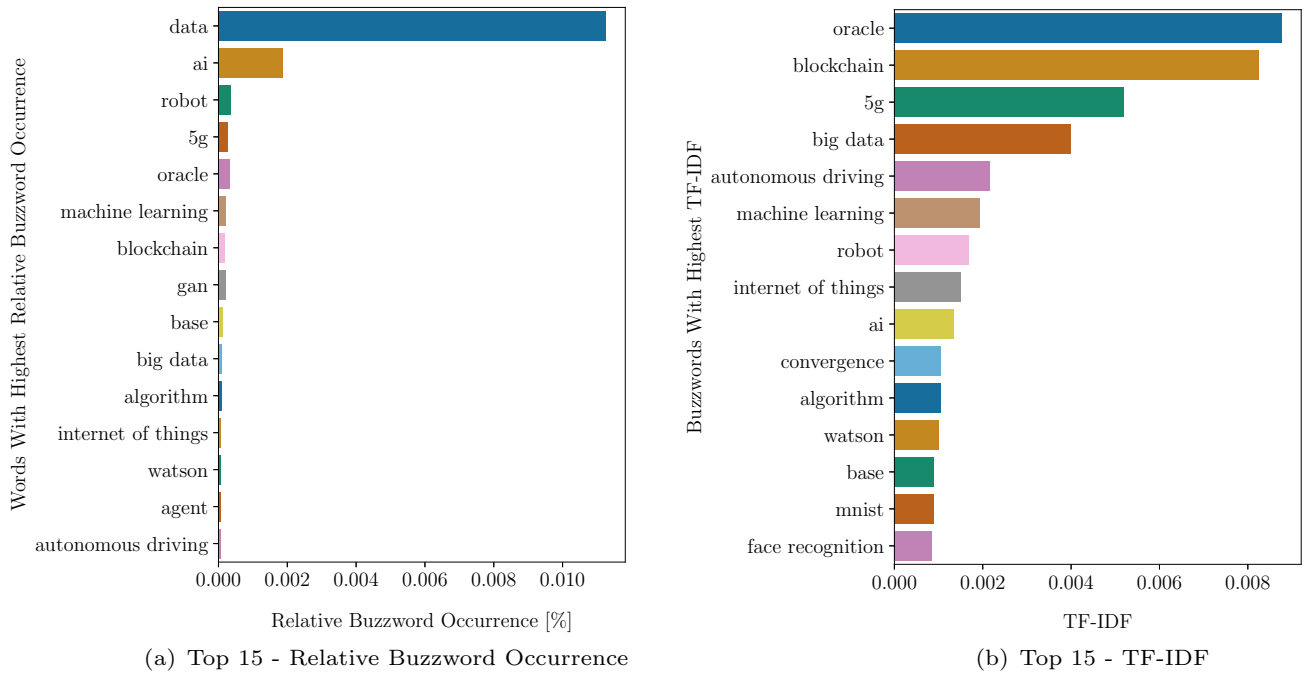
Declarations

Conflict of interest The authors have no competing interests to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Appendix A: TOP 15 — AI-related buzzwords

In Fig. 10, the relative occurrence and the TF-IDF of the top 15 AI-related buzzwords are summarized. In Fig. 10a, the occurrence of the terms *data* and *AI* is considerably higher than that of all other buzzwords. In Fig. 10b, the application of the TF-IDF resulted in a normalized evaluation of the top 15 buzzwords and shows *oracle*, *blockchain* or *5g* with the highest values in the ranking.



(a) Top 15 - Relative Buzzword Occurrence

(b) Top 15 - TF-IDF

Fig. 10 Top 15 — AI-related buzzwords (complete data set): comparison of the results by considering the metrics **a** relative buzzword occurrence and **b** TF-IDF

Appendix B: Buzzword list

In Tables 2 and 3, the English and German buzzword lists that were used for the data analysis are shown. Note that we summarized synonyms like *AI* and *Artificial Intelligence* and their respective German translations and German abbreviations into one word group.

Table 2 List of *English* buzzwords

English buzzwords	
Accuracy	Genetic programming
Activation function	Gradient
Adagrad	Gradient descent
Adaptive Ml	Graph
Agent	Ground truth
AI	Hashing
AI PaaS	Heuristic
AlexNet	Hinge loss
Artificial general intelligence	Human augmentation
Artificial intelligence	Hypergraph
Augmented reality	Hyperparameter
Autoencoder	Image recognition
Autonomous driving level 4	Inference
Autonomous driving level 5	Information retrieval
Autonomous vehicles	IoT platform
Backpropagation	Knowledge graphs
Base	Knowledge representation
Binary classification	Linear regression
Biochips	Long short-term memory
Biotech-cultured or artificial tissue	LSTM
Black box	Machine intelligence
Blockchain	Machine learning
Bounding box	Machine translation
Brain-computer interface	Mainframe
Case-based reasoning	Mean absolute error
Chomsky hierarchy	ML
Classifier	MNIST
Clustering	Model training
Co-adaptation	Nanotube electronics
Cognitive computing	Neuromorphic hardware
Cognitive expert advisors	N-Gram
Cognitive science	Normalization
Collaborative filtering	Object-oriented programming
Commercial Uavs	One-hot encoding
Computer vision	One-shot learning
Confusion matrix	Optical character recognition
Connected home	Oracle
Convergence	Overfitting
Conversational user interfaces	Perceptron

Table 2 continued

English buzzwords	
Convex hull	Perplexity
Convolution	Predicate calculus
Correlation	Principal component analysis
Covariance	Quantization
Cross-validation	Quantum computing
Cybernetics	Rectified linear unit (Relu)
Data	Recurrent neural network
Data analysis	Regression
Data augmentation	Reinforcement learning
Data mining	Reinforcement learning (Rl)
Data set	Responsible Ai
Data set	Root directory
Decision tree	Self-supervised learning
Deep learning	Semi-supervised learning
Digital twin	Signal processing
Dimension reduction	Smart dust
Drones	Smart robots
Dynamic programming	Smart workspace
Edge AI	Softmax
Enterprise taxonomy and ontology management	SQL
Epoch	Supervised learning
Error rate	Supervised machine learning
Explainable AI	Transfer learning
Feature	Turing test
Feature engineering	Unsupervised learning
Flying autonomous vehicles	Unsupervised machine learning
FPGA	Virtual assistants
Fuzzy logic	Virtual reality
Game theory	Volumetric displays
Gaussian distribution	Word2Vec
Generalized linear model	Zero-shot learning
Generative adversarial networks	4D printing
Genetic algorithm	5 G

Table 3 List of *German* buzzwords

German buzzwords	
Algorithmus	Multitask-Lernen
Assistenzsysteme	Natürliche Sprachverarbeitung
Automatisierte Systeme	Neuronale Netzwerke
Autonomes Fahren	Neuronales Netz
Autonome Systeme	N-gram
Backpropagation	Numerischer Funktionsvektor
Beaufsichtigtes Lernen	One-Shot-Lernen
Big Data	Online- und Datenstrom-Lernverfahren
Binäre Klassifizierung	OpenCL
Black Box-Modelle	Optische Charaktererkennung
Chatbot	Prädiktive Instandhaltung
Chat Bot	Predictive Maintenance
Computer Vision	Regression
Cyber Physical Systems	Rekurrente Neuronale Netze
Data Mining	Relationales Lernen
Data Mining	Relativer absoluter Fehler
Data Science	Relativer quadratischer Fehler
Daten	Roboter
Datenbereinigung	Robotik
Datenexploration	Schwarmintelligenz
Datenwissenschaft	Self-Supervised Deep Networks
Deep Learning	Sentiment-Analyse
DeepMind	Sentiment Detection
Eingebettetes System	Smart Data
Ende-zu-Ende-Lernen	Smart Factory
Entscheidungsmodell	Social Bots
Erklärbare KI	Soft Bots
Erklärbare Künstliche Intelligenz	Software Roboter
Face Recognition	sozialer Agent
Featureentwicklung	sozialer Roboter
GAN	Sprachassistenzsysteme
Gegnerisches Lernen	Spracherkennung
Generative Adversarial Networks	Starke KI
Gesichtserkennung	Starke Künstliche Intelligenz
Grey Box-Modelle	Subsymbolische Modelle
Hyperparameter	Symbolische Modelle
intelligente Fabrik	Technologische Singularität
Internet der Dinge	Tensorflow
KI	tiefes Lernen
KI-Chip	tiefgehendes Lernen
Kognitives Computing	Trolley-Problem
Kontextuelle Intelligenz	Turing-Test
Künstliche Intelligenz	Überwachtes Lernen
Künstliches neuronales Netz	Überwachtes maschinelles Lernen
Lernalgorithmus	Unsicherheit und Schlussfolgern
Lernen auf Datenströmen	Unüberwachtes Lernen

Table 3 continued

German buzzwords	
Lernen Systeme	Unüberwachtes maschinelles Lernen
M2M-Kommunikation	Verlustfunktion
Machine Learning	Verstärkendes Lernen
Machine-to-Machine Kommunikation	Verstehen und Generieren von natürlicher Sprache
Maschinelles Lernen	Verteiltes Rechnen
mehrschichtiges Lernen	Verteiltes System
Mensch-Maschine-Interaktion	Video-Kennzeichnung für Maschinelles Lernen
Mittlerer absoluter Fehler	Virtueller Assistent
ML	Vorhersagen-Modell
Monte-Carlo-Baumsuche	Watson
Monte-Carlo Tree Search	White Box-Modelle
Multiklassenklassifizierung	Zero-Shot-Lernen

Appendix C: Topic clustering pipeline

Data Source The scraped data sources for the purpose of this study were already introduced before (see again Sect. 3.1). As a supplement to the previous explanations, we would like to outline that among the political data sources the *PP* discusses various topics with a strong link to the recent public interest, whereas the *DRS* contains documents with a stronger specification on topics and are partly funneling into the *PP* when discussed in the parliament.

The business magazines were at this point separated as different data sets with on the one hand side the *HB* as by far largest data source and on the other side the following magazines (*BE*, *WW*, *MM*). *FOCM* was not included in this partial experiment as the combined amount of data could not be processed by the available systems. The following four data sets with their according size were created:

- *PP*: 1154
- *DRS*: 86,939
- Magazines (*BE*, *WW*, *MM*): 291,075
- *HB*: 188,659

Due to the limitation of 64GB RAM, only approximately eight million words were used as training data and a representative sample was generated from the four data sets.

Data Set Building This section describes all the processing steps on the raw transcribed text before it was used as input for the topic clustering algorithm. Existing literature about topic clustering performs model training on texts with a short total word length. An increase in corpus length for each training sample has adverse effects on finding the optimal number of topics [70]. Complying with existing research [14], using the full articles as a training sample has proven not to be ideal because each article has an approximate length of thousands of words. Therefore, a different strategy was chosen to divide the transcriptions into smaller chunks.

Text Cleaning Topic clustering uses the embeddings of words as input for their training and inference process. Words that are similar in meaning but differ slightly from each other, e.g., through the upper or lower case, past or present tense or simply by singular and plural, need to be uniformly processed so similar words are not counted twice. A text cleaning pipeline is a crucial step before applying topic clustering [71]. The step-by-step cleaning process (see Fig. 11) was performed for every article in all data sets, where each article transcription equals the *Raw Text*.

Keyword Data Set The buzzword list was used to find syntactical small but meaningful chunks. As already mentioned before, words and abbreviations with the same meaning were combined in the buzzword list. All articles were iteratively scanned for the listed buzzwords. From the raw text, we used each keyword occurrence and picked a fixed number of words post and prior to it, which will be called text chunk throughout this study.

The number of words, that were used to create the text chunks, was selected by comparing the development of two computed metrics. First, the relative proportion of overlapping text chunks reveals how significant the percentage of overlapping text chunks is within one article. The closer the number gets to 1, the worse the selected number k for the number of words prior and post to the keywords gets. The second metric is the relative proportion of used words. This metric tracks how many words of the total number of words per article are located in all text chunks for each article. This metric gets steadily closer to one when k is increasing because it approaches the length of the entire article text. For this reason, the desired number of k shall be as small as possible, but still preserves as many words as possible from the text. The ideal number of words post and prior to each buzzword, for example, for the *PP* is 200. As a result, the keyword-level chunk articles are of the following size:

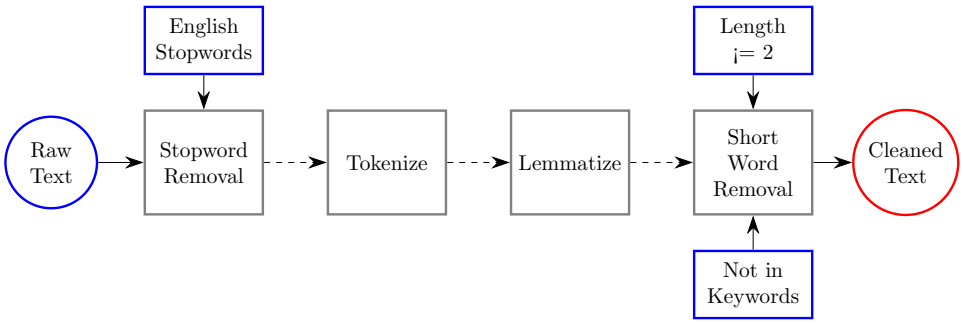


Fig. 11 Text cleaning

- *PP*: 8077
- *DRS*: 146,793
- Magazines (*BE*, *WW*, *MM*): 24,138
- *HB*: 46,394

It was observed that, for example, on average eight buzzwords were found in each *PP* article and four keywords in the *HB* articles. Due to the reason, that topic clustering on this keyword-level data sets would require more than 64GB RAM for more than 20,000 articles at once, three additional data sets were created to not exceed the length of 20,000 keyword-level chunks. The three new data sets are randomly chosen samples from the keyword-level data sets, to shorten them down to be usable for the training process. Since *PP* has only 8077 entries, no additional data set was required.

Training The BERT architecture was used for the model training. The pipeline for topic clustering was built with *OCTIS* [16], an open-source technology provided on GitHub, which builds upon the Hugging Face BERT models [17]. The transformer model *all-mpnet-base-v2* was used due to the fact that it provides the best performance for the language feature extraction [72]. The advantage of using *OCTIS* is that the metrics *Coherence* and *NPMI* can be calculated off-the-shelf. Topic clustering requires that the number of topics is preset before the training started.

Optimization The ideal number of topics is an important parameter to decide how many topics will be used for the training data. If the number is too low, the entire context of the training data could not be caught. Topics that are desired to be found, such as AI-related topics, could accidentally be omitted by the model. If the defined number of topics is too high, many overlapping topics will occur and the topics over time analysis could fail. Therefore, the metrics from *OCTIS* [16], namely *Coherence* and *NPMI*, were used as a basis estimation for the ideal number of topics. The metrics were iteratively calculated from the range [10, 40] with a step size of two. After shortening down the three data sets that exceeded the length of 20,000 keyword-level chunks, the following ideal number of topics for each of the four data sets was calculated:

- *PP*: 20
- *DRS*: 30
- Magazines (*BE*, *WW*, *MM*): 25
- *HB*: 35

This respective number of topics was used for the final training and inference process.

Inference After training the model with the hyperparameters from the optimization function and the ideal amount of clusters, the model training returned a large dictionary. For example, the dictionary from the *PP* contains the following entries after the training with the 8077 by 401 keyword-level chunks and 20 preset topics:

- *topics*: (20 × 10) the list of the most significant words for each topic (list of lists of strings),
- *topic-word-matrix*: (20 × 25,070) matrix of weights with 20 as the number of topics and 25,070 as the vocabulary length,
- *topic-document-matrix*: (20 × 8077) matrix of weights with 20 as the number of topics and 8077 as the number of documents in the corpus.

Each topic contains a list of ten words in string format. The naming convention was to use the first five topics and connect them with an underscore with each other. As a result, a single string could describe each topic. Each of the 20 topics is defined by the ten most contributed words to this topic. The 20 topics have words connected to them, but no *generic term*, which needs to be self-defined by hand for each topic (see again "Appendix D"). The embedding vectors are necessary to calculate the cosine similarities. The embedding data sets have the same length as the training data and the columns correspond to the number of topics. For example, the *PP* embedding, which is stored in the *topic-document-matrix* has a size of (8077 × 20). The columns accumulate to one for each row. This corresponds to how likely each topic is present in the current keyword chunk.

Appendix D: Topic strings and generic terms

Table 4 shows the mapping from the topic string (added in German and in English) to the self-defined generic term to describe at the end all selected topics from the four sub data sets of *DRS*, *PP*, *HB* and the combined other magazines.

Table 4 Mapping from topic string to generic term

Data set	Topic string (German)	Topic string (English)	Generic term
DRS	ki,einsatz,projektgruppe,kapitel,siehe	ai,operation,projectgroup,chapter,see	AI
	polizei,verbindungsdaten,datenbanken,speicherung,verkehrsdaten	police,connectiondata,databases,storage,trafficdata	Cyber Security
	daten,verwendet,person,dürfen,löschen	data,used,person,allowed,delete	Data Regulations
	daten,personenbezogene,dürfen,übermittelt,löschen	data,personal,may,transmitted,delete	Data Surveillance
	inwiefern,bundeswehr,antwort,eingestuft,europol	how,germanarmedforces,answer,classified,europol	German Military
	migrationshintergrund,jahre,jahr,quelle,tab	migration background,years,year,source,tab	Immigration
PP	reichenbach,gerold,datenschutzrecht,datenschutze,datenschutz	reichenbach,gerold,dataprotectionlaw,dataprotection,dataprotection	Data Protection
	elektronischen,elena,verfahren,elektronische,braucht	electronic,elena,procedure,electronic,needs	Digitization
	computer,pass,sicher,eröffnete,legislaturperiode	computer,passport,secure,opened,legislative,session	Electronic Pass
	forschung,biobanken,bereit,probanden,genetischen	research,biobanks,ready,test,subjects,genetic	Genetic Research
	soldaten,bundeswehr,militärische,sol,vereinten	soldiers,germanmilitaryforces,military,sol,united	German Military
	spd,integration,menschen,beifall,gute		Immigration
	kind,menschen,integration,geld,wenig	spd,integration,people,applause,good	Poverty
	binniger,datei,benneter,sicherheitsbehörden,ordneten	binniger,file,benneter,security,authorities,orders	Security
	erstattet,verheimlicht,steueranteil,deskanzleramtes,politikersatz	reimbursed,concealed,tax,chancellor's office,policy,substitute	Tax Fraud
	innern,staaten,vereinigten,zusammenarbeit,terrorismus	interior,states,united,cooperation,terrorism	Terrorism
HB	geld,banken,länder,land,klar	money,banks,countries,land,clear	Financial Sector
	hacker,phishingmail,viren,hackerangriff,email	hacker,phishingmail,virus,hackerattack,email	Hacking
	anbieter,internet,kunden,verhalten,verbraucher	provider,internet,customers,behavior,consumer	Internet Usage
	investoren,anleger,erwarten,anlegen,fond	investors,investor,expect,invest,fund	Investments
	smartphone,app,apple,iphone,server	smartphone,app,apple,iphone,server	Smartphone
	startup,big,datum,digitalisierung,digitalen	startup,big,date,digitization,digital	Startup Acceleration
Magazines	unterstützt,sicherheitslücke,right,twitteraccount,enthüllungen	supported,vulnerability,right,twitteraccount,revelations	Data Leakage
	chip,sensoren,auto,autonome,zukunft	chip,sensors,car,autonomous,future	Digitization
	banken,basis,situation,schuldner,lage	banks,basis,situation,debtor,situation	Financial Crisis
	hacker,angreifer,opfer,unbemerkt,nutzern	hacker,attacker,victim,unnoticed,users	Hacking
	iphone,smartphone,apple,musik,internet	iphone,smartphone,apple,music,internet	Internet Usage
	bald,startup,neuen,wozu,virtuelle	soon,startup,new,what,virtual	Startup Acceleration

References

1. Ongsulee, P.: Artificial intelligence, machine learning and deep learning. In: 15th International Conference on ICT and Knowledge Engineering (ICT&KE), pp. 1–6 (2017). https://doi.org/10.1109/ICTKE.2017.8259629
2. Perrault, R., et al.: The AI Index 2019 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA (2019). https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf
3. Lee, R.S.T.: Intelligent agents and software robots. In: Lee, R.S.T. (eds.) Artificial Intelligence in Daily Life. Springer, Singapore (2020). https://link.springer.com/chapter/10.1007/978-981-15-7695-9_8
4. Lee, R.S.: Fuzzy-Neuro Approach to Agent Applications: From the AI Perspective to Modern Ontology. Springer, Berlin (2006)
5. Keller, R., Stohr, A., Fridgen, G., Lockl, J., Rieger, A.: Affordance-experimentation-actualization theory in artificial intelligence research: a predictive maintenance story. In: ICIS 2019 Proceedings 1 (2019). https://aisel.aisnet.org/icis2019/is_development/is_development/1/
6. Kalis, B., Collier, M., Fu, R.: 10 promising AI applications in health care (2018). https://hbr.org/2018/05/10-promising-ai-applications-in-health-care
7. Presse- und Informationsamt der Bundesregierung. Ki als markenzeichen für deutschland: Strategie künstliche intelligenz (15.11.2018). https://www.bundesregierung.de/breg-de/themen/digitalisierung/ki-als-markenzeichen-fuer-deutschland-1549732
8. Bundesministerium für Bildung und Forschung. Nationale ki strategie für künstliche intelligenz - ai made in germany (2022). https://www.ki-strategie-deutschland.de/home.html
9. Boyles, M.: Understanding how politics can affect your business (2022). https://online.hbs.edu/blog/post/politics-and-business
10. Statista. Economy and politics (2023). https://www.statista.com/markets/2535/economy-politics/
11. Yoffie, D.B.: How an industry builds political advantage: competitive strategy. Havard Bus Rev **66**(3), 82–89 (1988)
12. Frieden, J.: The political economy of ecomomic policy: we should pay closer attention to the interactions between politics, economics and other realms. Finance Dev **57**, 4–9 (2020)
13. Fleming, L., Sorenson, O.: Science as a map in technological search. Strateg Manag J **25**(89), 909–928 (2004). https://doi.org/10.1002/smj.384
14. Fast, E., Horvitz, E.: Long-term trends in the public perception of artificial intelligence. Proc AAAI Conf Artif Intell **31**(1), 963–969 (2017)
15. Oliner, S.D., Sichel, D.E.: The resurgence of growth in the late 1990s: is information technology the story? J Econ Perspect **14**(4), 3–22 (2000). https://doi.org/10.1257/jep.14.4.3
16. Terragni, S., Fersini, E., Galuzzi, B.G., Tropeano, P., Candelieri, A.: Octis: comparing and optimizing topic models is simple!. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Systems Demonstrations, pp. 263–270 (2021)
17. Wolf, T., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6
18. Landers, R.N., Brusso, R.C., Cavanaugh, K.J., Collmus, A.B.: A primer on theory-driven web scraping: automatic extraction of big data from the internet for use in psychological research. Psychol Methods **21**(4), 475–492 (2016). https://doi.org/10.1037/met0000081
19. von Krogh, G.: Artificial intelligence in organizations: new opportunities for phenomenon-based theorizing. Acad Manag Discov **4**(4), 404–409 (2018). https://doi.org/10.5465/amd.2018.0084
20. Proksch, S.-O., Slapin, J.B.: How to avoid pitfalls in statistical analysis of political texts: the case of Germany. German Politics **18**(3), 323–344 (2009). https://doi.org/10.1080/09644000903055799
21. Jungherr, A., Schoen, H., Jürgens, P.: The mediation of politics through twitter: an analysis of messages posted during the campaign for the German federal election 2013. J Comput-Mediat Commun **21**(1), 50–68 (2016). https://doi.org/10.1111/jcc4.12143
22. Cambridge University Press. Buzzword: Cambridge dictionary (2022). https://dictionary.cambridge.org/dictionary/english-german/buzzword
23. Mjøs, O.J., Moe, H., Sundet, V.S.: The functions of buzzwords: a comparison of 'web 2.0' and 'telematics'. First Monday (2014). https://doi.org/10.5210/fm.v19i12.4896
24. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min Knowl Discov **7**(4), 373–397 (2003)
25. Yang, Y., Wilson, L.T., Wang, J.: Development of an automated climatic data scraping, filtering and display system. Comput Electron Agric **71**(1), 77–87 (2010). https://doi.org/10.1016/j.compag.2009.12.006
26. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., Fdez-Riverola, F.: Web scraping technologies in an API world. Brief Bioinform **15**(5), 788–797 (2014). https://doi.org/10.1093/bib/bbt026
27. Cha, Y., Stow, C.A.: Mining web-based data to assess public response to environmental events. Environ Pollut **198**, 97–99 (2015). https://doi.org/10.1016/j.envpol.2014.12.027
28. Cornwall, A.: Buzzwords and fuzzwords: deconstructing development discourse. Dev Pract **17**(4–5), 471–484 (2007). https://doi.org/10.1080/09614520701469302
29. Budak, C., Agrawal, D., El Abbadi, A.: Structural trend analysis for online social networks. Proc. VLDB Endow. **4**(10), 646–656 (2011). https://doi.org/10.14778/2021017.2021022
30. Caled, D., Beyssac, P., Xexéo, G., Zimbrão, G.: Buzzword detection in the scientific scenario. Pattern Recognit. Lett. **69**, 42–48 (2016)
31. Holzinger, A., Kieseberg, P., Weippl, E., Tjoa, A.M.: Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable AI. In: Holzinger, A., Kieseberg, P., Tjoa, A.M. & Weippl, E. (eds.) Machine Learning and Knowledge Extraction, Vol. 11015 of Lecture Notes in Computer Science, pp. 1–8. Springer, Cham (2018)
32. Igan, D., Mishra, P., Tressel, T.: A fistful of dollars: lobbying and the financial crisis. NBER Macroecon. Annu. **26**(1), 195–230 (2012). https://doi.org/10.1086/663992
33. Mohr, J.W., Wagner-Pacifici, R., Breiger, R.L., Bogdanov, P.: Graphing the grammar of motives in national security strategies: cultural interpretation, automated text analysis and the drama of global politics. Poetics **41**(6), 670–700 (2013). https://doi.org/10.1016/j.poetic.2013.08.003
34. Bunea, A., Ibenskas, R.: Quantitative text analysis and the study of EU lobbying and interest groups. Eur Union Politics **16**(3), 429–455 (2015). https://doi.org/10.1177/1465116515577821
35. Neuman, Y., Nave, O., Dolev, E.: Buzzwords on their way to a tipping-point: a view from the blogosphere. Complexity **16**(4), 58–68 (2011). https://doi.org/10.1002/cplx.20347
36. Grbich, C.: Qualitative Data Analysis—An Introduction, 1st edn. SAGE Publications, London (2007)
37. Winter, G.: A comparative discussion of the notion of 'validity' in qualitative and quantitative research. Qual. Rep. **4**(3), 1–14 (2000). https://doi.org/10.46743/2160-3715/2000.2078
38. brand eins. brand eins: Heft-archiv (2021). https://www.brandeins.de/magazine/brand-eins-wirtschaftsmagazin#anchor-1999
39. FOCUS MONEY Online. Money archiv (2021). https://www.focus.de/finanzen/money-magazin/archiv/

40. Handelsblatt GmbH. Handelsblatt archiv (2021). https://www.handelsblatt.com/archiv/
41. Manager magazin. Manager magazin archiv (2021). https://www.manager-magazin.de/print/mm/index-2022.html
42. Handelsblatt GmbH. Wirtschaftswoche archiv (2018). https://archiv.wiwo.de/?ticket=ST-5975351-1Gikv4ezUDuC7O5huPbw-ap5
43. IVW. Ivw (2021). https://www.ivw.eu/
44. The pandas development team. Pandas (2022). https://pandas.pydata.org/
45. Deutscher Bundestag. Drucksachen (2021). https://www.bundestag.de/drucksachen
46. Deutscher Bundestag. Plenarprotokolle (2021). https://www.bundestag.de/protokolle
47. Walker, M., Burton, B., Cantara, M.: Hype cycle for emerging technologies, 2016 (2016). https://www.gartner.com/en/documents/3383817/hype-cycle-for-emerging-technologies-2016
48. Walker, M.: Hype cycle for emerging technologies, 2017 (2017). https://www.gartner.com/en/documents/3768572-hype-cycle-for-emerging-technologies-2017
49. Walker, M.: Hype cycle for emerging technologies, 2017 (2018). https://www.gartner.com/en/documents/3885468-hype-cycle-for-emerging-technologies-2018
50. Microsoft Corporation. Machine learning glossary of important terms (2021). https://docs.microsoft.com/en-us/dotnet/machine-learning/resources/glossary
51. Accenture. The applied intelligence glossary (2021). https://www.accenture.com/gb-en/insights/applied-intelligence/artificial-intelligence-glossary
52. Google. Machine learning glossary (2021). https://developers.google.com/machine-learning/glossary
53. Mishra, J.: Machine learning glossary (2021). https://machinelearning.wtf/
54. Butterfield, A., Ngondi, G.E., Kerr, A.: A Dictionary of Computer Science, 7th edn. Oxford University Press, New York (2016)
55. Richardson, L.: Beautiful soup: beautiful soup documentation (2020). https://www.crummy.com/software/BeautifulSoup/bs4/doc/
56. Moolenaar, B.: Vi improved. GitHub (2019). https://www.vim.org/
57. Lernende Systeme - Die Plattform für Künstliche Intelligenz. Glossar (2020). https://www.plattform-lernende-systeme.de/glossar.html
58. Microsoft Corporation. Machine learning-glossar mit wichtigen begriffen (2021). https://docs.microsoft.com/de-de/dotnet/machine-learning/resources/glossary
59. clickworker GmbH. Ki glossar: Begriffserklärungen rund um lernende systeme und ki (künstliche intelligenz) (2021). https://www.clickworker.de/ki-glossar/
60. GULP Redaktion. Glossar künstliche intelligenz: Als unternehmer sollten sie diese ki begriffe kennen (2019). https://www.gulp.de/knowledge-base/19/iii/glossar-kunstliche-intelligenz.html
61. Cao, S., Lyu, H., Xu, X.: Insurtech development: evidence from Chinese media reports. Technol Forecast Soc Change **161**, 120277 (2020). https://doi.org/10.1016/j.techfore.2020.120277
62. Vayansky, I., Kumar, S.A.: A review of topic modeling methods. Inf Syst **94**, 101582 (2020). https://doi.org/10.1016/j.is.2020.101582
63. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J Mach Learn Res **3**(Jan), 993–1022 (2003)
64. Mitchell, T.M.: Machine Learning. McGraw Hill, New York (1997)
65. Kagermann, H., Lukas, W.-D., Wahlster, W.: Industrie 4.0: Mit dem internet der dinge auf dem weg zur 4. industriellen revolution. VDI nachrichten **13**, 2 (2011)
66. Böhmer, M.M., et al.: Investigation of a Covid-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. Lancet Infect Dis **20**(8), 920–928 (2020). https://doi.org/10.1016/S1473-3099(20)30314-5
67. Duffield, J.S.: Political culture and state behavior: why Germany confounds neorealism. Int Organ **53**(4), 765–803 (1999). https://doi.org/10.1162/002081899551066
68. Geddes, A., Scholten, P.: The Politics of Migration and Immigration in Europe Second edition/Andrew Geddes and Peter Scholten. SAGE, Los Angeles (2016)
69. Deutscher Bundestag. Epidemische lage von nationaler tragweite: Wd3 - 3000 - 141/20 (10.06.2020). https://www.bundestag.de/resource/blob/700666/da1b330c0d4f3ac6c8b14ed3844a4d38/WD-3-141-20-pdf-data.pdf
70. Sbalchiero, S., Eder, M.: Topic modeling, long texts and the best number of topics some problems and solutions. Qual Quant **54**(4), 1095–1108 (2020). https://doi.org/10.1007/s11135-020-00976-w
71. Schmiedel, T., Müller, O., vom Brocke, J.: Topic modeling as a strategy of inquiry in organizational research: a tutorial with an application example on organizational culture. Organ Res Methods **22**(4), 941–968 (2018). https://doi.org/10.1177/1094428118773858
72. Jayanthi, S.M., Embar, V., Raghunathan, K.: Evaluating pretrained transformer models for entity linking in task-oriented dialog (2021). https://arxiv.org/abs/2112.08327

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.