Artificial Intelligence in Innovation: How to Spot Emerging Trends and Technologies

Christian Mühlroth and Michael Grottke, *Member, IEEE*

***Abstract*—Firms apply strategic foresight in technology and innovation management to detect discontinuous changes early, to assess their expected consequences, and to develop a future course of action enabling superior company performance. For this purpose, an ever-increasing amount of data has to be collected, analyzed, and interpreted. Still, a major part of these activities is performed manually, which requires high investments in various resources. To support these processes more efficiently, this article presents an artificial-intelligence-based data mining model that helps firms spot emerging topics and trends at a higher level of automation than before. Its modular structure consists of components for query generation, data collection, data preprocessing, topic modeling, topic analysis, and visualization, combined in such a way that only a minimum amount of manual effort is required during its initial set up. The approach also incorporates self-adaptive capabilities, allowing the model to automatically update itself once new data has become available. The model parameterization is based on latest research in this area, and its threshold parameter is learnt during supervised training using a training data set. We have applied our model to an independent test data set to verify its effectiveness as an early warning system. By means of a retrospective analysis, we show in three case studies that our model is able to identify emerging technologies prior to their first publication in the Gartner Hype Cycle for Emerging Technologies. Based on our findings, we derive both theoretical and practical implications for the technology and innovation management of firms, and we suggest future research opportunities to further advance this field.**

***Index Terms*—Artificial intelligence (AI), computer-aided foresight, corporate foresight, innovation management, machine learning, strategic decision making, strategic foresight, technology management, trend detection.**

I. Introduction

THE accelerating pace of technological innovation and social change exposes firms to an increasingly complex space of strategic options. These dynamics in the corporate environment are further amplified by the decreasing predictability of future developments. In this context, firms are continuously confronted with new influence factors: Global trade encourages the market entrance of new or previously unknown competitors, risk capital and lean go-to-market strategies enable start-ups and new ventures to gain remarkable market shares in short time, technological breakthroughs increasingly accelerate technological change, and rapidly evolving customer needs lead to radically new products, services, and business models [1], [2]. Since today's cycle of innovation is characterized by high technological dynamisms, particular emphasis needs to be placed on the management of technology [1], [3].

In order not only to maintain competitive vitality in this highly competitive environment, but also to find new opportunities for competitive advantage, firms have a strong incentive to detect relevant emerging topics and trends at an early stage to develop adequate response strategies for the future [1], [4]. Still, these early signals are predominantly detected by chance rather than in a systematic manner [2]. Firms are thus facing the challenge to extract meaningful insights out of the plethora of information contained in big data sets, in order to effectively support the management of technology and innovation [5].

A recent literature review revealed that previous research had dealt with these challenges by applying various data mining techniques [6]. The goals are twofold: First, relevant topics have to be identified as early as possible [7]–[10]. Second, these topics have to be observed over a certain period of time to detect relevant trends [4], [11]–[13]. In addition, it was pointed out [6] that a stronger emphasis on improved search strategies, data quality, and automation is required to reduce the involvement of human experts and to thus decrease the likelihood of being influenced by the human actor bias. Existing approaches were also found to lack the ability to learn and accumulate knowledge over time; moreover, the need for incorporating multiple source types to provide a comprehensive basis for strategic decision making was identified [6].

This article does not intend to forecast any technological innovations or future trajectories. Rather, it aims at addressing the aforementioned challenges and demands by offering the following contributions for research and practice.

1) First, we propose an innovative artificial intelligence (AI)-enabled data mining model to detect relevant emerging and trending topics for technology and innovation management. We rigorously select, parameterize, and combine machine learning techniques in such a way that the model is able to continuously adapt to changes in the

Manuscript received April 30, 2019; revised November 8, 2019; accepted April 10, 2020. Date of publication May 27, 2020; date of current version January 7, 2022. This work was supported by the German Federal Ministry of Education and Research (BMBF) under Grant 02K16C191. Review of this manuscript was arranged by Department Editor D. Ehls. *(Corresponding author: Christian Mühlroth.)*

Christian Mühlroth is with the Department of Statistics and Econometrics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 90403 Nuremberg, Germany (e-mail: christian.muehlroth@fau.de).

Michael Grottke is with Global Data Science, GfK SE, 90443 Nuremberg, Germany, and also with the Department of Statistics and Econometrics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 90403 Nuremberg, Germany (e-mail: michael.grottke@gfk.com).

Digital Object Identifier 10.1109/TEM.2020.2989214

data, minimizing the manual effort required. In addition, we provide simple and effective recommendations for follow-up activities based on the model output.

2) Second, we use our model in a supervised manner with labeled training data in order to learn its threshold parameter in a first step; in a second step, we apply it in an unsupervised manner to independent testing data. Three exemplary case studies show, by means of a retrospective analysis, that our proposed model can detect technological innovations years before their first publication in the Gartner Hype Cycle for emerging technologies.
3) Finally, we derive implications and recommendations for detecting emerging topics and trends in strategic foresight processes in technology and innovation management, and we suggest future research opportunities to further advance this field.

The rest of this article is organized as follows. In Section II, we explain the theoretical framework that encompasses our article, with special regard to the role of technological change in innovation. The research method employed in this article is presented in Section III. Section IV describes our proposed AI-based model and its basic working principles. We present details on the implementation, the parameter settings and the results from the supervised learning of the similarity threshold parameter in Section V, and we then demonstrate the effectiveness of our unsupervised learning approach in three case studies in Section VI. In Section VII, we discuss the theoretical and practical implications for technology and innovation management, as well as limitations of our approach and future research opportunities. Finally, Section VIII concludes this article.

II. Theoretical Framework

The nature of innovation and the evolution of technology has been a central question in research for decades. Since first studies were published in the 1960s, different models to explain the driving forces behind this phenomenon emerged. The gradual development of these models can be categorized into five generations, ranging from science push (first generation), demand pull (second generation), the coupling (also called chain-linked) model (third generation), the integrated model (fourth generation), and the systems model (fifth generation) [14], [15]. While no direct hierarchy of these models can be established, adjacent models were influenced by each other, and various generations of models are still in application in today's firms [14].

Although these models differ in their explanation of how innovation arises, they are in agreement with their observations on where it happens: The key driving forces emerge from needs of humans, needs of other technologies, or both. Innovations therefore do not emerge from the void; instead, existing technologies and capabilities (such as the steam engine, wireless technology, or AI) serve as their building blocks [16]. The extensive exchange of information, coupled with the capability to combine and improve previous building blocks in such a way that additional value is produced, eventually leads to a new innovation. Traces of evidence for this can be found, for example, in patent data, where patents are more likely to cite (and therefore involve) recent patents rather than old ones [17]. The phenomenon of technological convergence provides further pieces of evidence: The first joint occurrence in a patent of previously unrelated technology areas (represented by their patent classification codes) signals the emergence of a technological innovation. Such merges have been found to be more frequent if the focal technology areas had already been closely related to each other (represented by their cross citations), and to often result from a collaboration between different firms [18]–[20]. A constantly growing number of technologies which can satisfy human and technological needs better or more economically than before thus produces even more building blocks on which new innovations can be built; as a consequence, technological progress keeps on accelerating [16]. For this reason, today's cycle of high-tech innovation is considered to be fundamentally different from earlier innovation cycles in other eras of science and economy [1].

However, the emergence of a technological innovation does not necessarily mean that it immediately replaces the preceding technology. Interactions between existing and emerging technologies are not unitary, but they can range from mutual benefit to mutual damage; examples are symbiotic interaction (the growth of an emerging technology stimulates the growth of existing ones), predator-prey interaction (either the emerging or the existing technology benefits while the other is disadvantaged), and purely competitive interaction (one technology replaces the other) [21], [22]. Furthermore, the mode of interaction can change over time; for example, the emergence of a new and disruptive technology may at first lead to a growing application of the previously existing one, only then to transition into a predator-prey interaction, before the existing technology is finally replaced [22], [23]. These dynamics bear a high risk: Firms (and in particular leading firms) may mistakenly hold on to their existing portfolio of technologies, products, and services for too long. Because they still observe growth from it, firms often focus investments in innovation on the incremental optimization of their existing portfolio, putting it at the core of their competitive vitality [1]. However, this strategy ignores (knowingly or unknowingly) that the growth experienced may also result from an imminent disruption of an emerging technology which may already have been on the rise [22]. With such a strong focus on incremental innovation, disruption in the form of technological discontinuity is not likely to come from leading firms, but rather from others [1]. The implications are twofold: First, leading firms are encouraged to behave ambidextrous, i.e., to maintain a diversified portfolio of both incremental and radical technologies and innovations in order to increase their likelihood of gaining (temporary) competitive advantage. Second, to lower the risk of getting surprised and disrupted, firms are required to detect discontinuous change as early as possible in order to develop adequate response strategies, helping them to maintain competitive vitality [3], [24]–[26].

For several decades, strategic foresight has been used to anticipate change in the corporate environment, interpret its consequences and develop future courses of action for responding to transformational change [27], [28]. Its outcomes can provide valuable inputs for related business processes, such

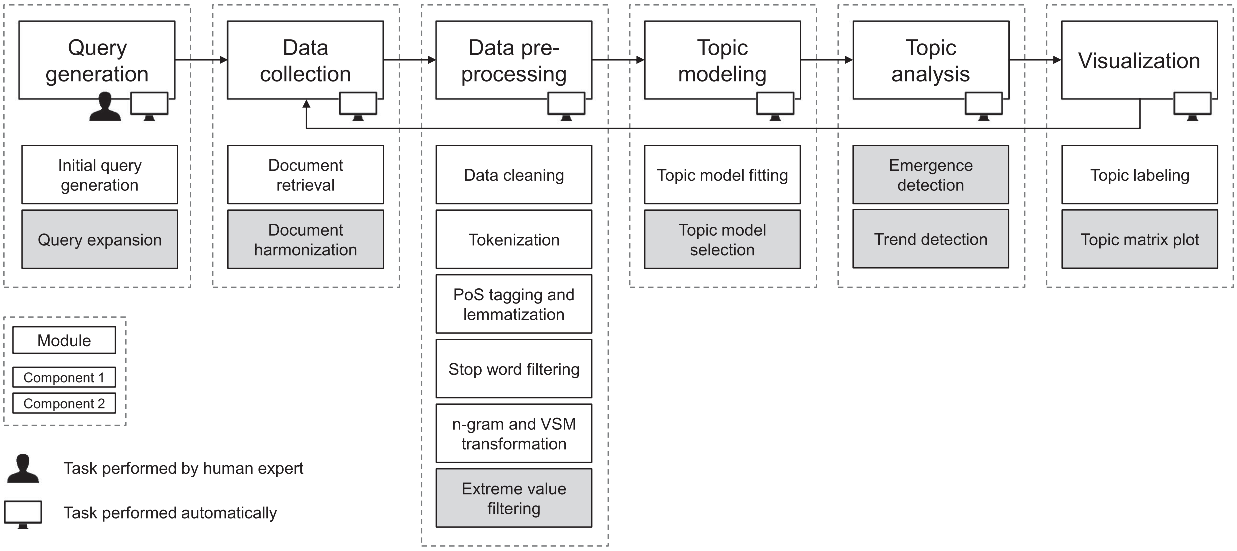


Fig. 1. Overview of the proposed model.

as directions for research and development (R&D) as well as opportunity spaces for growth at the early stages of innovation management [29], [30]. A recent study also confirmed that firms repeatedly applying strategic foresight practices are more likely to attain superior profitability and to outperform the industry by gaining better market capitalization growth [31]. An integral part of this process is the continuous perception of change, i.e., monitoring the corporate environment for strategically relevant signals, trends, and technologies. Still, necessary steps such as data collection and analysis are often performed manually, which requires a large amount of time, money, and human resources. With the ongoing advances in computing power, there is an increasing demand for computer-aided systems and AI techniques, in order to shift the work performed by human experts from tasks that can be automated (for example, data collection and data processing) to tasks that require creative skills (e.g., interpretation, decision making, and taking action) [5]. Against this backdrop, we propose an innovative approach that automates many of the tasks that are still performed manually today, in order to have more resources available for the tasks of tomorrow.

III. Research Method

Verifying a model which aims at discovering early signals for potential changes (e.g., upcoming trends or technological innovations) would require advance knowledge about the future. At the time when our model indicates a signal for a possible future technological innovation, we would already have to know whether this technology will successfully be established in the future. Since this is not possible, we have decided to employ the research method of testing our model by means of three retrospective case studies against a reference benchmark [32], [33].

As a widely recognized reference benchmark, we chose the Gartner Hype Cycle for Emerging Technologies [34]. It is created on a yearly basis by collecting, condensing, and communicating opinions of a large group of leading industry and domain experts at a global scale. Its emerging technologies are classified into five different phases: innovation trigger, peak of inflated expectations, trough of disillusionment, slope of enlightenment, and plateau of productivity, incorporating earlier research on the estimation of technological effects known as Amara’s law [35].

For our purpose, we solely focus on technologies in the first phase of the Hype Cycle (i.e., the innovation trigger). Therefore, we consider a retrospective case study to be successful if our proposed model is able to detect signs of a technological innovation prior to its first publication in the first phase of the Hype Cycle.

IV. Approach to Spotting Emerging Topics and Trends

Our proposed model is based on existing concepts from AI as well as existing approaches from advanced machine learning [5], [36], [37], and it both improves such components and combines them in a novel way, as depicted in Fig. 1. Components with essential changes and improvements are highlighted in gray.

The proposed model has a modular structure and consists of components that are connected to each other in such a way that the required manual effort is reduced to a minimum. A human expert is only required for the initial definition of the search field during query generation; all subsequent steps, including the visualization of the results, are then performed in an entirely automated way. Hereby, the parameterization of such an unsupervised learning approach is of major importance [38], [39]. Whenever possible, we therefore incorporate recent research findings (e.g., concerning the hyperparameterization in topic modeling [40], see Section IV-D); in the case of the initial threshold determination, the parameters are learnt in a supervised manner from an independent training data set (see Section V).

As the corporate environment is constantly changing, the model must allow for updates over time, too. We therefore incorporate self-adaptive capabilities [41] into our model, such that it collects new data on the search field at specified time intervals, automatically adapting itself to the data found. The iterative and AI character of the model makes it possible to continuously monitor the corporate environment; it thus allows not only for a retrospective analysis, but also for a system set up in preparation for the analysis of future data. The modules and components of our model are described in the following sections.

A. Query Generation

The *initial query generation* begins with the definition of the search field in which relevant emerging topics and trends are to be scouted and monitored, based on the firm's strategic foresight goals [5], [42]. A human expert defines the initial search terms, the data source types to be collected (e.g., scientific publications, patents, and web sources [6]), the relevant time frame (e.g., within the last five years), and the preferred update interval (e.g., quarterly). The initial search terms are then concatenated to a search query using Boolean operators in order to focus on subsets of data rather than simply collecting all available data.

So far, these queries have exclusively been created by human experts, whereas a structured and computer-aided query generation process is not used [6]. However, a search query that is too broad can lead to results that are too general, whereas a search query that is too narrow increases the risk of missing important data. Either way, results may be not specific enough, incomplete, or biased toward the human expert's field of expertise [8], [9], [43].

In order to improve the quality of the search and thus the quality of the data collected, we have added a component for *query expansion* to our model [44]. Hereby, the human-expert-created search query is extended by synonyms and semantically-related terms from the web-based semantic knowledge graph ConceptNet [45]. First, the initial search query is tokenized into various n-grams, and suggested synonyms and terms are received from ConceptNet for each of them. Next, the human expert decides which of these suggestions to include. Finally, the initial query is extended by concatenating the selected suggestions with the Boolean operator "OR." This improved search strategy makes it possible to include previously unknown or unconsidered search terms, and it thus increases the likelihood of relevant new discoveries that would otherwise go unnoticed. For example, based on the initial query *"autonomous driving,"* the following final query would be derived: *"autonomous driving" OR (("autonomous navigation" OR "self-directed" OR "self-driving" OR "self-governing" OR semiautonomous OR superautonomous) AND (automobile OR car OR drive OR driver OR "motor vehicle" OR vehicle)).*

B. Data Collection

For data collection, the final search query is automatically translated into the proprietary query language of each database for the data source types selected by the human expert. The queries are then sent to the connected databases, and the results are fetched during *document retrieval*. In the further course of our article, each resulting item (such as one scientific article or one patent) will be referred to as a "document," and the set of all documents for a given search query will be referred to as the "corpus."

Since the data formats of the documents vary due to their different sources, *document harmonization* is applied to all received documents using a unified document format. Each document thus consists of the following parts.

1) *Head:* Internal information required for data storage, such as a unique document identifier and a UNIX timestamp of when the document was retrieved.
2) *Meta:* Harmonized metadata of the document, such as its publication date, its authors, and the uniform resource identifier (URI) of the original source.
3) *Body:* Harmonized contents of the document, such as its title, abstract, and full-text content.
4) *Specifics:* Data that is available only for a particular document type, e.g., patent families.

The collected and harmonized data is then stored and passed on to data preprocessing.

C. Data Preprocessing

In text mining, the application of data preprocessing techniques is expected to increase the quality of the subsequent data analysis [46], [47]. Although a wide range of such techniques is applied in research, a universal standard has not yet been established [6]; in many cases, the intermediate results of this process are refined iteratively with the help of human experts until a sufficient result has been achieved (e.g., see [43], [48], and [49]). Since we intend to have a highly automated approach in our case, we have introduced the following data preprocessing pipeline.

At first, we apply *data cleaning* to the data collected. After removing duplicate documents, we remove all digit-only characters, special characters, and diacritics. We further remove terms that have less than 2 or more than 30 characters, as they are likely to be either leftovers, a URI, or a string of garbled text due to optical character recognition or document conversion [50]. Finally, we strip multiple whitespaces and transform the text to lowercase [47].

In the cleaned documents the words are still separated by whitespaces. During *tokenization*, the text is chunked by using the available whitespaces as delimiters. Each token now represents a word made up of alphanumeric characters.

We then apply *point of speech (PoS) tagging and lemmatization* to the generated tokens, thus reducing both inflectional and derivationally related forms of a word to a common base form. We prefer this method over stemming, as the latter one only cuts off common prefixes and suffixes of an inflected word, whereas lemmatization also considers each word's morphological information [51]. For example, the tokens "am," "are," and "is" are all reduced to the token "be," but the tokens "speaker" (tagged as a noun) and "speak" (tagged as a verb) are still kept as individual tokens, because they do not represent inflected word forms [52]. Moreover, we only keep proper nouns (such as "london" or "nato") and nouns (such as "car" or "machine"), verbs, adjectives, and numerals, while discarding all other tokens with different PoS tags.

After PoS tagging and lemmatization, *stop word filtering* removes commonly used words for which the information content is expected to be very low to none. Stop words are, for example, "about," "be," "have," and "particular." To remove as many of these words as possible from our list of lemmatized tokens, we have combined lists from different sources (i.e., from the Python NLTK package [51], the United States Patent and Trademark Office, and Oracle's MySQL open-source database), and have lemmatized them, too.

The remaining tokens are then vectorized during the *n-gram and vector space model (VSM) transformation*. An n-gram is an adjacent sequence of n tokens. We transform our tokens into bi-grams (tokens of length $n = 2$, such as "autonomous driving"), since they in particular have been found to improve both the quality and the coherence of data analysis results [53], [54]. In our case, a bi-gram needs to be observed in the document corpus at least once, otherwise it is discarded. Subsequently, all tokens are transformed into the VSM by creating a document-term matrix (DTM) using raw term-frequency counts.

As the final data preprocessing step, we apply *extreme value filtering* to the DTM. Tokens that appear in less than three documents are removed in accordance with the Zipf distribution [55], [56]. Tokens that appear in more than 90% of the documents are removed, too, because the subsequent topic modeling is otherwise suspected to overestimate their importance [57].

D. Topic Modeling

We apply latent Dirichlet allocation (LDA), an unsupervised machine learning technique for natural language processing. LDA is a generative and probabilistic topic model to discover latent topics in a collection of text documents [57], and it has shown superior performance in comparison to other text mining methods [58]. Its underlying assumption is that a topic captures word cooccurrences that are semantically related [40].

During *topic model fitting*, the tokens within the documents (referred to as "words") represent the only observed variables, whereas the topics and the per-document topic proportions are hidden variables of interest that need to be inferred. To do so, LDA works backwards and calculates which hidden structure (i.e., topic model) is most likely to have generated the observed words within the documents. This generative process is described in Algorithm 1.

The number of topics is denoted by K. Each topic k ($k = 1, \ldots, K$) is represented by a probability distribution $\boldsymbol{\beta}_k$ over all observed words (referred to as the vocabulary $\mathcal{V}$) and is drawn from a Dirichlet distribution, $\boldsymbol{\beta}_k \sim$ Dirichlet($\boldsymbol{\eta}$). The dth document from the collection of all D documents (referred to as corpus $\mathcal{D}$) is represented by a distribution over all K topics and is also drawn from a Dirichlet distribution, $\boldsymbol{\theta}_d \sim$ Dirichlet($\boldsymbol{\alpha}$). The joint distribution of all hidden and observed variables [57] can be expressed as

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y} | \alpha, \eta) = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k | \boldsymbol{\eta}) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d | \boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_{d,n} | \boldsymbol{\theta}_d) p(y_{d,n} | z_{d,n}, \boldsymbol{\beta}_{d,k}) \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$, $\boldsymbol{z} = (z_{1,1}, \ldots, z_{D,N_D})$, and $\boldsymbol{y} = (y_{1,1}, \ldots, y_{D,N_D})$.

Algorithm 1: Generative Process of LDA.

```
for topic $k \in \{1, \ldots, K\}$ do
    Choose distribution over vocabulary $\mathcal{V}$, $\boldsymbol{\beta}_k \sim$ Dirichlet($\boldsymbol{\eta}$)
for the $d$th document in corpus $\mathcal{D}$, $d \in \{1, \ldots, D\}$, do
    Choose distribution over topics, $\boldsymbol{\theta}_d \sim$ Dirichlet($\boldsymbol{\alpha}$)
    for the $n$th word in the $d$th document, $n \in \{1, \ldots, N_d\}$, do
        Choose topic assignment, $z_{d,n} \sim$ Multinomial($\boldsymbol{\theta}_d$), with $z_{d,n} \in \{1, \ldots, K\}$
        Choose word, $y_{d,n} \sim$ Multinomial($\boldsymbol{\beta}_{z_{d,n}}$), with $y_{d,n} \in \mathcal{V}$
```

Fig. 2 visualizes LDA as a graphical model, using plate notation.

The variables for the topics $\boldsymbol{\beta}$, the per-document topic distributions $\boldsymbol{\theta}$ and the per-word topic assignments $\boldsymbol{z}$ are not observed and would have to be conditioned on the only observable variables $\boldsymbol{y}$. However, the exact computation of the posterior is intractable due to its denominator [57]. In this article, we make use of variational Bayesian (VB) inference to compute from the documents available at time t the posterior per-document topic distributions in $\boldsymbol{\theta}^t = (\boldsymbol{\theta}_1^t, \ldots, \boldsymbol{\theta}_D^t)$ and the posterior per-topic word distributions in $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \ldots, \boldsymbol{\beta}_K^t)$. Specifically, we apply the multipass online learning variant of VB described in [59] to enable the self-adaptive capabilities of our process model as described in Section IV-G.

The parameterization of LDA has been subject to an ongoing discussion in research, where its Dirichlet concentration hyperparameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ as well as the parameter for the number of topics K are of special interest. These (hyper-)parameters were found to influence the performance of the algorithm to a large extent and should therefore be adjusted to the respective purpose of analysis [39], [40].

Hyperparamter vector $\boldsymbol{\alpha}$ influences the per-document topic density. If its entries consist of small values (e.g., 0.1) few components will have a positive probability, while most of them will have a probability close to zero; the distribution becomes more sparse. In contrast, higher values (e.g., 1.0) will lead to a more even distribution [57]. For our purpose of analysis lower values mean that the documents are more likely to contain a mixture of just a few of the topics (or even only one topic), whereas higher values mean that the documents tend to contain a mixture of most of the topics. In research, there have been various suggestions for choosing the entries of $\boldsymbol{\alpha}$ (e.g., $50/K$ [60], $1/K$ [61], and 0.1 [62]). Moreover, using an asymmetric parameter vector for $\boldsymbol{\alpha}$ was found to yield more interpretable results [40]. Fig. 3 illustrates 1000 random draws from three-dimensional (3-D) Dirichlet distributions, where two of the concentration parameter vectors $\boldsymbol{\alpha}$ are symmetric (namely, $[0.1, 0.1, 0.1]$ and $[1.0, 1.0, 1.0]$), while one is asymmetric ($[0.1, 0.05, 0.01]$).

Hyperparameter vector $\boldsymbol{\eta}$ influences the per-topic word density. Here, low values for the entries mean that the topics are more likely to contain a mixture of just a few of the observed words (i.e., they are more distinct), whereas high values lead to topics which contain a mixture of most of the words (i.e., they

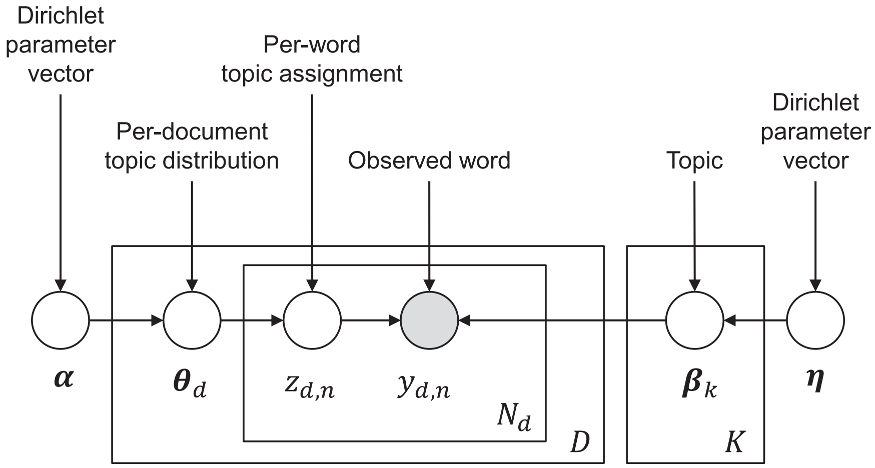


Fig. 2. LDA in plate notation.

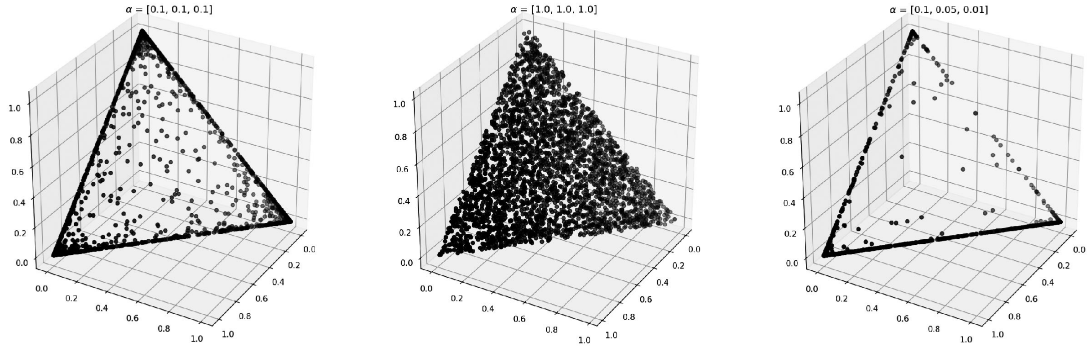


Fig. 3. Example of 1000 random draws each from the Dirichlet distributions with different concentration parameter vectors.

are more similar). Typically, the values for the entries of $\boldsymbol{\eta}$ are considerably lower than those for $\boldsymbol{\alpha}$ [40], [63]. High values are used when few and more general topics should be discovered, whereas values close to zero (e.g., 0.01 [62]) are employed when searching for a larger number of specific topics. Similar to documents, we want the topics to be as distinct as possible. However, in contrast to $\boldsymbol{\alpha}$, an asymmetric parameter vector for $\boldsymbol{\eta}$ has been found to yield topics that are more similar to each other [40].

The number of topics K is the only parameter that is mandatory to be specified before executing existing LDA implementations, whereas $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ are optional and can default to standard values defined in the respective software libraries. The selection of K was found to be crucial for the power, quality, and interpretability of the data analysis [38]–[40]. In general, if K is too low, the topics become too general, while too large a K leads to topics that are too granular; in both cases, they will be hard to interpret by human experts. In existing research, it is usually up to these experts to heuristically determine K [6].

However, manually tuning the number of topics and selecting the best resulting topic model requires expert knowledge and consumes valuable time which could better be used in other areas [5]. Since our goal is to reduce the involvement of human experts as much as possible, we propose a new procedure for a fully automated *topic model selection* as described in Algorithm 2.

Algorithm 2: Automated Topic Model Selection.

```
Input: $\mathcal{D}$, $\mathcal{V}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$
for number of topics $K$ in $\{\mathrm{round}(0.7 \cdot \sqrt{D}); \mathrm{round}(1.3 \cdot \sqrt{D})\}$ do
    Fit topic model LDA based on $\mathcal{D}$, $\mathcal{V}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, $K$
    Calculate topic coherence $C_V$
Select the topic model with max($C_V$)
```

At first, we apply an exhaustive grid search (also called "parameter sweep") [64], [65] over a dynamic range of K which depends on the number of documents in corpus $\mathcal{D}$. Hereby, we keep a low and asymmetric $\boldsymbol{\alpha}$ and a low and symmetric $\boldsymbol{\eta}$ fixed, as described above.

Each topic model is then evaluated for its quality. To this end, various metrics have been proposed in the literature, e.g., the predictive likelihood of held-out data [66], or density-based methods [67]. However, these earlier metrics were found to correlate negatively with human interpretability [68]. As a result, researchers have proposed new metrics based on topic coherence [69], [70]. In particular, the four-stage topic coherence metric C_V, as described in [71], was found to be closer than others to human ratings in terms of topic interpretability. First, a coherence score is calculated for each topic k by applying a four-step pipeline (segmentation, probability calculation, confirmation measure, and aggregation). Next, the coherence scores obtained for all K topics are averaged into the final score C_V. Finally, the one topic model from the exhaustive grid search attaining the highest C_V metric is selected.

E. Topic Analysis

By using the online variant of LDA, the model is updated at the selected frequency (e.g., quarterly) with new documents that were published within the latest time interval (e.g., quarter). Thus, a sequence of topic models is created over time, where each topic model (referred to as a topic model generation) represents a snapshot of the available topics at the respective point in time. The per-topic word distributions in $\boldsymbol{\beta}^t$ and the per-document topic distributions in $\boldsymbol{\theta}^t$ resulting from the tth topic model generation (i.e., the one obtained at current time t) are then used to analyze both the emergence and the trend of each topic. A topic can be either popular or emerging, and it can be either upward trending or downward trending.

For *emergence detection*, the word distribution of the kth topic in the tth topic model generation, $\boldsymbol{\beta}_k^t$, is compared to the word distributions of all topics in the $(t-1)$st topic model generation, $\boldsymbol{\beta}_1^{t-1}, \ldots, \boldsymbol{\beta}_{K^{t-1}}^{t-1}$, to assess their semantic similarity. To this end, we apply the well-known Jensen–Shannon divergence (JSD) [72] in a novel way, namely, to determine the semantic difference of two topics by measuring the divergence between their two probability vectors

$$\mathrm{JSD}_{k,l}^{t,t-1} = \frac{1}{2}\mathrm{KLD}\left(\boldsymbol{\beta}_k^t \parallel \bar{\boldsymbol{\beta}}_{k,l}^{t,t-1}\right) + \frac{1}{2}\mathrm{KLD}\left(\boldsymbol{\beta}_l^{t-1} \parallel \bar{\boldsymbol{\beta}}_{k,l}^{t,t-1}\right)$$
$$k = 1, \ldots, K^t;\ l = 1, \ldots, K^{t-1}. \tag{2}$$

Here,

$$\bar{\boldsymbol{\beta}}_{k,l}^{t,t-1} = \frac{1}{2}(\boldsymbol{\beta}_k^t + \boldsymbol{\beta}_l^{t-1}) \tag{3}$$

while JSD is the symmetrized and smoothed version of the Kullback–Leibler divergence (KLD), which quantifies the divergence between two probability mass functions $\boldsymbol{f}$ and $\boldsymbol{g}$ (represented as vectors on their joint domain $\mathcal{X}$)

$$\mathrm{KLD}(\boldsymbol{f} \parallel \boldsymbol{g}) = \sum_{x\in\mathcal{X}} f(x)\log_2\left(\frac{f(x)}{g(x)}\right). \tag{4}$$

A high $\mathrm{JSD}_{k,l}^{t,t-1}$ value indicates that topic k from topic model generation t and topic l from topic model generation $t-1$ have highly differing word distributions and hence a low semantic similarity. As the logarithm in (4) is taken to base two, all JSD values are in the range from zero to one. We can therefore easily transform the distance metric into a similarity metric

$$\mathrm{JSS}_{k,l}^{t,t-1} = 1 - \mathrm{JSD}_{k,l}^{t,t-1}, \quad k = 1, \ldots, K^t;\ l = 1, \ldots, K^{t-1}. \tag{5}$$

Since the topic model generations are calculated at different points in time and are thus based on different corpora, the observed vocabulary is expected to differ, too. For example, words that are observed in vocabulary $\mathcal{V}^t$ may not have been observed in vocabulary $\mathcal{V}^{t-1}$, because it is possible that they were not contained in any of the documents available at this earlier point in time. However, in the vocabulary of the previous generation the probabilities for these words are not zero, as one might expect, but the entries do not exist at all. As a consequence, the dimensions of the word distribution vectors from two consecutive topic model generations may differ, and the calculation of their similarity might thus not be possible. To cope with this issue, we insert the unobserved words into the respective probability vectors, and we apply Laplace smoothing [73] by assigning them a small value (namely, 10^{-12}), in order to shift as little probability mass as possible. After calculating the JSS values for all pairs of topics from generations t and $t-1$, respectively, the values are stored in a $(K^t \times K^{t-1})$-dimensional similarity matrix.

For topic k from the current model generation t, we define emergence as the maximum JSS value with any topic from model generation $t-1$

$$\mathrm{emergence}_k^t = \max_{l\in\{1,\ldots,K^{t-1}\}} \mathrm{JSS}_{k,l}^{t,t-1}. \tag{6}$$

To determine whether or not the topic is to be considered emerging, we apply a simple rule-based classification: If its emergence exceeds a certain threshold, it can be assumed that it shares its content with a topic from the previous model generation; the topic is thus not new and is classified as "popular." In contrast to this, if its emergence metric lies below the threshold, then it seems to feature content not seen before, and it is hence classified as "emerging."

The selection of an appropriate similarity threshold is crucial for the classification to be meaningful. Since the input data is different for each search field and each model generation, the corpora, the vocabularies, and the resulting probability vectors will differ; a fixed similarity threshold does thus not seem advisable.

While a dynamic threshold could be determined as an empirical quantile of all the JSS values in the similarity matrix, the automated topic model selection (Algorithm 2) might lead to a small number of topics in both generations, and a limited number of comparisons. Rather, with

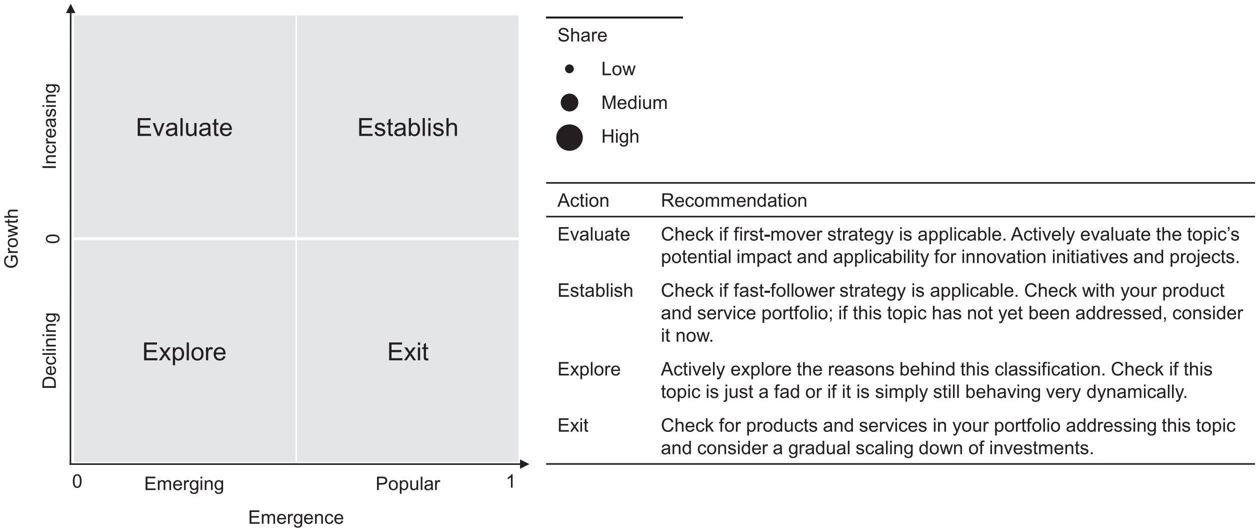
$$\mathrm{JSS}_{\min}^{t,t-1} = \min_{k\in\{1,\ldots,K^t\},l\in\{1,\ldots,K^{t-1}\}} \mathrm{JSS}_{k,l}^{t,t-1} \quad \text{and}$$
$$\mathrm{JSS}_{\max}^{t,t-1} = \max_{k\in\{1,\ldots,K^t\},l\in\{1,\ldots,K^{t-1}\}} \mathrm{JSS}_{k,l}^{t,t-1} \tag{7}$$

we calculate the threshold as the value located at a certain percentage π into the interval $[\mathrm{JSS}_{\min}^{t,t-1}; \mathrm{JSS}_{\max}^{t,t-1}]$; i.e., the threshold is given by

$$\mathrm{JSS}_{\min}^{t,t-1} + \pi \cdot (\mathrm{JSS}_{\max}^{t,t-1} - \mathrm{JSS}_{\min}^{t,t-1}). \tag{8}$$

The percentage value π to be used can be learnt during supervised training; in Section V, we describe this for a specific applied setting. With our approach, we ensure that the process model can autonomously adapt itself to new data without requiring a human expert to adjust the similarity threshold.

After all topics have been classified as either popular or emerging, the resulting per-document topic distributions in vector $\boldsymbol{\theta}^t = (\boldsymbol{\theta}_1^t, \ldots, \boldsymbol{\theta}_D^t)$ are analyzed for the temporal distribution of their associated documents during *trend detection*. First, we define the length the and number of time slices that are used to



Action	Recommendation
Evaluate	Check if first-mover strategy is applicable. Actively evaluate the topic's potential impact and applicability for innovation initiatives and projects.
Establish	Check if fast-follower strategy is applicable. Check with your product and service portfolio; if this topic has not yet been addressed, consider it now.
Explore	Actively explore the reasons behind this classification. Check if this topic is just a fad or if it is simply still behaving very dynamically.
Exit	Check for products and services in your portfolio addressing this topic and consider a gradual scaling down of investments.

Fig. 4. Schematic visualization of the topic matrix plot including recommendations for follow-up activities.

split $\boldsymbol{\theta}^t$ into parts of equal length, based on the publication dates of its associated documents. For example, setting the length of the time slices to one quarter (i.e., three months) and the number of time slices to four causes documents published within the last year to be used for trend analysis, whereas older documents are discarded. Next, we define a per-document topic probability threshold (denoted as δ) to indicate whether a particular topic is associated with a document. When using low entries of the LDA hyperparameter vector $\boldsymbol{\alpha}$, we expect documents to contain a mixture of just a few topics (see Section IV-D). Still, due to the probabilistic nature of LDA, a topic is sometimes assigned to a document with a relatively small probability value (e.g., 0.01). We thus introduce a threshold δ that needs to be exceeded, otherwise the association of a topic to a document is not taken into account. On this basis, we propose a new metric for the share of each topic, which is then calculated for each time slice s based on the set of all documents published in this time slice, $\mathcal{D}_s$

$$\text{share}_k^{t,s} = \frac{1}{\mid \mathcal{D}_s \mid} \sum_{d \in \mathcal{D}_s} I(\boldsymbol{\theta}_{d,k}^t \geq \delta) \tag{9}$$

where the indicator function $I(\cdot)$ equates to one if the condition in brackets is satisfied, and to zero otherwise. The share of a topic thus represents the fraction of documents published in a specific time slice that are associated with a particular topic. Note that the sum of shares of one generation may exceed the value one, due to multiple assignments caused by the probabilistic approach.

Depending on the number of time slices used, a sequence of share values is calculated for each topic k of topic model generation t, and is then analyzed using simple linear regression [74]. The topic share sequence values are plotted vertically against the indexes s of the time slices. Values for a and b are chosen such that the sum of the squared deviations from the linear trend line $a + bs$ is minimized

$$(\text{init}_k^t, \text{growth}_k^t) = \arg\min_{a,b} \sum_s \left(\text{share}_k^{t,s} - (a + bs)\right)^2. \tag{10}$$

We thus interpret the slope of the fitted least squares line as the (estimated) growth of topic k determined based on the current model generation. A positive (negative) value of growth$_k^t$ indicates an increasing (a declining) trend of topic k.

F. Visualization

To support the human expert in interpreting the analysis results, they are visualized, making use of a simple *topic labeling* technique. Hereby, each topic is labeled with its top-10 words according to the scheme word_probability*"word" (e.g., 0.367*"autonomous vehicle"), sorted by the word probabilities in descending order.

Various types of visualizations have been presented in research on mining weak signals and trends, with basic charts such as line charts, pie charts, or tables being used most frequently [6]. In this article, we propose a novel type of visualization, a *topic matrix plot*. Based on the previous topic analysis, each topic carries the information about its emergence and trend, which is depicted in the topic matrix plot as shown in Fig. 4. The emergence axis (abscissa) represents the normalized value range of the maximum semantic similarities of all topics of the current generation to the topics of the previous generation, whereas the trend axis (ordinate) shows the positive (trending) or negative (declining) growth of each topic. Additionally, the size of each element in the plot represents the relative topic share. Depending on the position of the topics in the quadrants of the matrix, recommendations for follow-up activities are proposed to help human experts decide which topics to focus on.

G. Self-Adaptive Capabilities

To cope with continuous change in the data available in the corporate environment, we incorporated self-adaptive capabilities [41] into our model. For this purpose, the process from data collection to visualization is triggered at regular intervals (e.g., quarterly, see Section IV-A), where the use of the online variant of LDA makes it possible to update the model with the

Algorithm 3: Iteration Step $t+1$ on Model Update.

```
Input: $\mathcal{D}^t, \boldsymbol{\alpha}, \boldsymbol{\eta}, K^t$
Collect new documents published since the previous generation $t$
Add new documents to $\mathcal{D}^t$, creating corpus $\mathcal{D}^{t+1}$
Pre-process documents and create new vocabulary $\mathcal{V}^{t+1}$
Perform online learning LDA based on $\mathcal{D}^{t+1}$, $\mathcal{V}^{t+1}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$, $K^t$
Calculate topic coherence $C_V^{t+1}$ of topic model from online learning LDA
if $C_V^{t+1} < C_V^t$ then
    Perform automated topic model selection based on $\mathcal{D}^{t+1}$, $\mathcal{V}^{t+1}$, $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$ (Algorithm 2)
    Set $C_V'$ to highest coherence value found in automated topic model selection
    if $C_V' > C_V^t$ then
        Use topic model resulting from automated selection as new generation $t+1$
    else
        Keep topic model from generation $t$ as topic model for generation $t+1$
    Set $K^{t+1}$ to the number of topics of topic model generation $t+1$
    Set $C_V^{t+1}$ to the coherence value of topic model generation $t+1$
else
    Use updated topic model from online learning LDA as new generation $t+1$
    Set $K^{t+1} = K^t$
Perform topic analysis using the resulting $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\theta}^{t+1}$, as well as $\boldsymbol{\beta}^t$, and $\boldsymbol{\theta}^t$
Visualize results for topic model generation $t+1$
```

newly collected data. After the update, the model performs a self-assessment based on its new topic coherence score C_V^{t+1}. If this value has declined compared to the topic coherence score of the preceding topic model generation C_V^t, the model tries to find a better state by performing automated topic model selection as described in Algorithm 2. If a topic model with a higher coherence score is found, it will be kept; otherwise, the model calculated by online LDA will be retained. Algorithm 3 describes this procedure in detail.

The iterative character of the model thus allows not only for a retrospective analysis, but it also prepares the model for an analysis of future data.

V. Implementation, Parameterization, and Supervised Learning

Our proposed model is designed to achieve the highest possible level of automation. To this end, we have combined different unsupervised machine learning techniques, and have parameterized them in a specific way. We have implemented our approach in Python, making use of the libraries NumPy, [75], Pandas [76], spaCy [51], scikit-learn [77], and gensim [61]. The parameters have been set as follows: As we want the topics to be as distinct as possible, we employ a low and asymmetric parameter vector for $\boldsymbol{\alpha}$ by using the input setting "asymmetric" as provided by gensim, and we choose the same low value (namely, 0.01) for all entries of $\boldsymbol{\eta}$. The document-to-topic probability threshold δ used in the topic analysis is 0.1, as smaller assignment probabilities can be considered the effect of noise [57].

While the settings for the necessary parameters can be derived from existing research, there is no default for the percentage value π used to calculate the interpolated value within the interval of similarity values (see Section IV-E). Although the actual similarity threshold value itself remains flexible by this approach, the percentage value must be defined. Instead of setting this value heuristically using trial and error, we decided to learn it from available data via supervised training.

For this purpose, we collected all conference proceedings available from the Association for the Advancement of Artificial Intelligence (AAAI) for five consecutive years (2018 and the four preceding years). The official conference tracks and their associated documents were used as the ground truth data. General tracks (e.g., "poster papers" or "student abstracts") and their documents were removed to focus on the main tracks, which describe more specific topics. Table I gives an overview of the data collected.

To learn the threshold, a human expert in the domain of AI and machine learning selected the following tracks: *Robotics* (2015), *AI and the Web* (2016), *Vision* (2017), and *Game Theory* (2018). For each track, its documents were kept for the respective year but removed for the previous year. For example, documents associated with the track on *Game Theory* were kept for 2018 but removed for 2017. Next, we ran our model on the data and let it autonomously select the best-fitting topic model by applying an exhaustive grid search (as described in Section IV-D) and threefold cross validation. Following the approach described in Section IV-E, we first calculated the similarity matrix, the corresponding interval, and the emergence of the topic that represents the track selected by the human expert. We then computed the percentage into the interval at which this specific emergence value is located. The arithmetic mean of all four percentage values calculated (one percentage value for each track selected by the human expert), 0.3625, represents the final percentage value π.

Table II summarizes the results from the supervised training. The topic terms (and their weights) indicating the AAAI track of interest are highlighted in bold and italics. The total training time for all four selected tracks amounted to approximately 1.5 h of parallelized runs using multiprocessing on 8 cores @ 2.60 GHz and 16 GB RAM.

VI. Case Studies

In order to test whether the proposed model can serve as an early warning system for detecting relevant emerging topics and trends, we tested its effectiveness by means of a retrospective analysis. To this end, we chose the Gartner Hype

TABLE I
Overview of the Supervised Training Data Set

	Total		Main tracks		General tracks	
Year	**# Tracks**	**# Documents**	**# Tracks**	**# Documents**	**# Tracks**	**# Documents**
2014	28	474	22	398	6	76
2015	32	673	23	537	9	136
2016	35	691	24	548	11	143
2017	36	786	24	639	12	147
2018	33	1102	23	937	10	165

TABLE II
Results From Supervised Training

	2018 vs. 2017	**2017 vs. 2016**	**2016 vs. 2015**	**2015 vs. 2014**
AAAI track	**Game Theory**	**Vision**	**AI and the Web**	**Robotics**
Topic detected	0.261*“human” + ***0.192*“game”*** + ***0.174*“visual”*** + 0.061*“response” + 0.055*“paper introduce” + 0.042*“draw” + ...	0.166*“dataset” + ***0.085*“image”*** + 0.079*“source” + ***0.060*“visual”*** + 0.056*“analysis” + ... + ***0.019*“computer vision”***	***0.069*“web”*** + ***0.062*“address”*** + 0.055*“real” + 0.053*“input” + 0.049*“improvement” + 0.048*“operator” + ...	0.076*“task” + ***0.033*“robot”*** + 0.029*“complete” + 0.028*“assume” + 0.027*“action” + 0.026*“learn” +...
Most similar previous topic	0.069*“human” + 0.066*“setting” + 0.062*“event” + 0.038*“knowledge base” + 0.035*“ontology” + 0.032*“structured” + ...	0.048*“learn” + 0.029*“base” + 0.026*“representation” + 0.025*“source” + 0.024*“exist” + 0.022*“show” + ...	0.116*“algorithm” + 0.041*“performance” + 0.038*“task” + 0.030*“real world” + 0.026*“develop” + 0.025*“show” + ...	0.056*“datum” + 0.045*“behavior” + 0.042*“approach” + 0.031*“structure” + 0.031*“learn” + 0.020*“propose” + ...
Similarity	0.09129212	0.10619812	0.09018212	0.12127712
Percentage value	0.34	0.39	0.32	0.40
Average value	0.3625			

Cycle for Emerging Technologies [34] as ground truth data, as explained in Section III. We consulted a human expert to select the technologies. Special care was given to ensure that the selections were as distinct as possible from each other, to prevent blurred data and results. For each selected technology, we applied our proposed model, starting from query generation and query expansion with the human expert. A data set containing papers from ScienceDirect (Elsevier) and patents from the European Patent Office was then collected over a period of five years, starting backwards from the initial year of publication of this technology in the hype cycle. Both data types were analyzed independently to identify if and where the technology could be found first. For the application of our model, we set the model parameters exactly as described in Section V. In the following sections, we present three case studies for *edge computing*, *bitcoin/blockchain*, and *3-D bioprinting*.

A. Case 1: Edge Computing

Edge computing enables the data processing infrastructure to exist closer to the sources of data than before. Instead of sending data across long routes to clouds or data centers, it is processed at the edge of a network, thus enabling to perform latency-sensitive data analyses in near real-time [78]. This technology was mentioned for the first time in the Gartner Hype Cycle for Emerging Technologies in 2017, and it was rated as transformational, yet still adolescent solution.

Since it is considered to be the next development step after cloud computing, which has already matured, the term *cloud computing* was used as initial query, and related terms such as *cloudification* and *cloudify* were added during query expansion to build the final search query of the retrospective analysis. We thus examined whether and when our proposed model would have indicated the new developments in edge computing if a firm had monitored ongoing developments in cloud computing. Within the time span of five years (2013 to 2017) 9623 papers and 9566 patents were collected.

The results of the unsupervised learning are summarized in Table III. The number of papers and patents already shows that there has been high activity in research and development since 2013. Our model signals an emerging topic in the papers data set, including the terms *iot* (an abbreviation for *Internet of Things*), *edge*, and *fog computing* during the year 2016. These terms are indicative of the topic edge computing, and they have a medium-sized share, a positive growth as well as a low maximum semantic similarity to topics from earlier topic model generations. Based on the calculated metrics, this topic is placed in the quadrant *evaluate* of the topic matrix plot, suggesting that the topic should be evaluated actively.

Also in 2016 an emerging topic is detected in the patent data, including similar terms such as *edge network* and *real time*. This topic has a relatively high share, most probably due to the included term *cloud computing* and its associated documents. Also, a positive trend and a low similarity to previous topics are found, leading to its classification into the quadrant *evaluate*.

It has thus been shown that the emerging technology edge computing manifested itself in 2016 as a newly emerging topic in our data sets, one year before its publication on the Gartner Hype Cycle in 2017. Presumably it were the relatively high weights of the abovementioned specific terms which led to the classification of the identified topic as being emerging, and the rather high share in patent documents seems to support the assumption that the edge computing topic emerged out of the previously existing generic topics related to cloud computing.

B. Case 2: Bitcoin/Blockchain

The blockchain technology can be described as a distributed, append-only, and time-stamped data structure. It allows to store any type of transaction sequentially into blocks, where each block is chained to the previous one, immutably stored across a peer-to-peer network and secured using cryptographic mechanisms. Due to its decentralized nature, its potential is seen to disrupt traditional business processes which rely on centralized architectures in health, education, and especially governance and the financial industry, as its popularity emerged mainly due to its associated use case bitcoin [79]. Mentioned for the first time in the Gartner Hype Cycle for Emerging Technologies in 2016, its terminology is still subject to ongoing discussions, and its business impact and use cases are yet to be proven.

In our retrospective analysis, we consider the topic from the perspective of the financial industry. Banks in particular have been intensively monitoring developments in the area of digital payment transactions. Accordingly, our data set was built using related search terms, such as *ledger* or *digital payment*. The analysis was to show whether and when companies observing this area would have been informed about developments concerning the blockchain technology by our proposed model. For the specified time period of five years, we collected 1893 papers and 1540 patents.

Based on scientific papers, a strong signal is already observed for the year 2014. The topic detected contains associated terms (such as *bitcoin*, *currency,* and *cryptocurrency*), and it is placed in the quadrant *evaluate* of the topic matrix plot.

A similar topic is identified in the patent data one year later. The topic terms include *block chain* and *digital currency*, and the topic has an almost zero yet negative growth value (-0.0019), which indicates that its development should be monitored and explored during the upcoming model updates. Table IV presents the results of the analysis.

In this case study, the emerging technology was found in the papers data set in 2014, and in the patents data set in 2015, both of which is earlier than its first appearance in the Gartner Hype Cycle in 2016. The topic detected in the research papers alludes to the first known use case of the blockchain technology (namely, bitcoin), whereas the topic found in the patent data is of a more general nature. In both cases, our proposed model was able to give an early signal concerning the technology.

C. Case 3: 3-D Bioprinting

3-D printing, as a technology for additive manufacturing, is widely employed in various industries and areas of application including engineering, manufacturing, and medicine. In recent years, the application of this technology has been found particularly promising for tissue engineering using cultured and artificial tissue and biocompatible scaffolds. Thus, the term *3-D bioprinting* emerged [80]. While 3-D printing in general has been represented on the Gartner Hype Cycle for quite some time, 3-D bioprinting was published for the first time in 2011 as a type of specialization. We deliberately chose this technology to test whether a newly emerging technology field nested inside a broader technology area (as 3-D bioprinting is to 3-D printing) can also be detected with our approach.

It would be reasonable to expect that if our proposed model had been used to monitor the general field of 3-D printing, signs of the development of its application in medicine should have been indicated early on. Therefore, the term *3-D printing* was selected for the initial query, and similar terms such as *three dimensional printing* and *additive manufacturing* were added during query expansion in order to build the data set. In total, 770 papers and 434 patents were collected during the five-year time period, which is considerably less than in the previous

TABLE III
RESULTS FOR THE EMERGING TECHNOLOGY EDGE COMPUTING (2017)

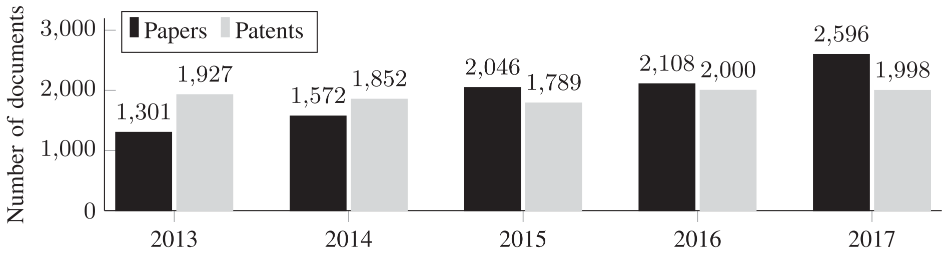
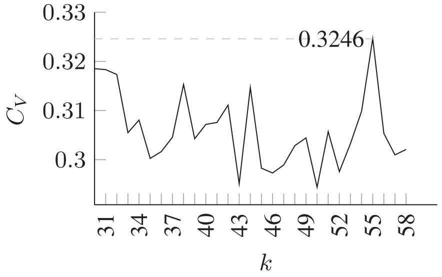
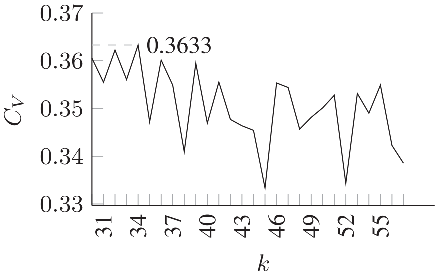
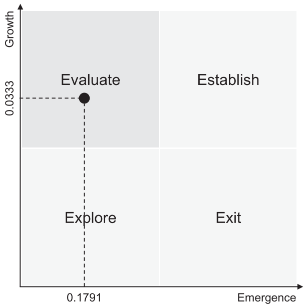
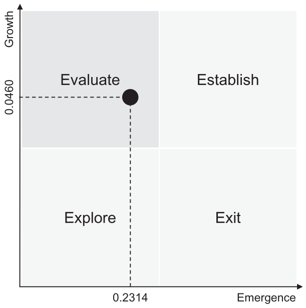
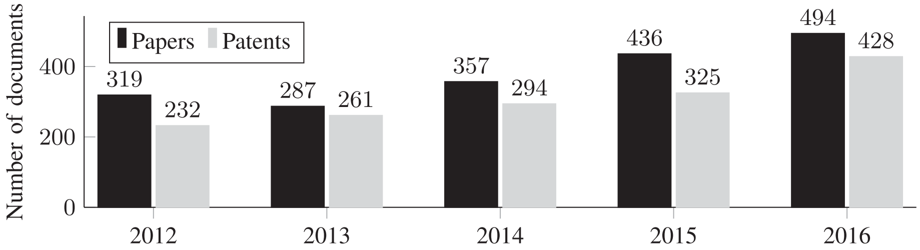
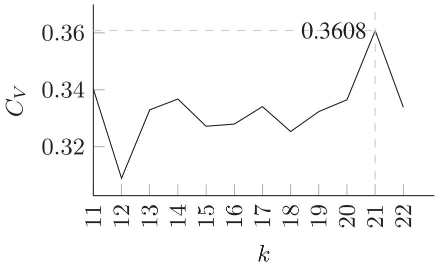
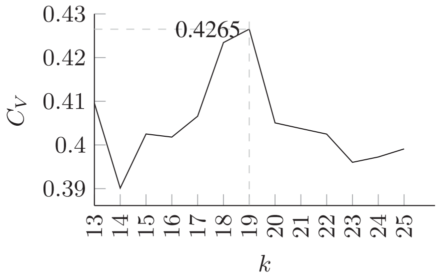
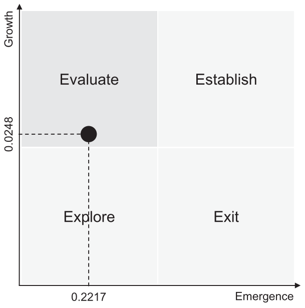
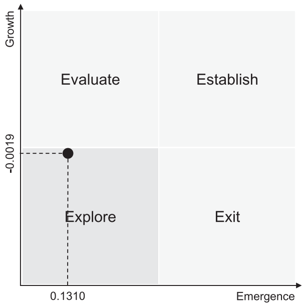
Query generated		
Initial query	"cloud computing"	
Query expansion	cloudification, cloudify, "utility computing", m2m, "machine-to-machine"	
Final query	"cloud comput*" OR cloudification OR cloudify OR "utility comput*" OR "m2m communicat*" OR "machine-to-machine communicat*"	
Data collected		
Documents collected		
Results achieved		
Data source	Papers	Patents
Grid search		
Year detected	2016	2016
Topic detected	***0.055*"iot" + 0.033*"edge"*** + 0.033*"production" + 0.033*"tool" + ***0.028*"sensor network"*** + ... + ***0.023*"fog computing"*** + ... + ...	***0.139*"edge network" + 0.102*"cloud computing"*** + 0.074*"sensor" + 0.070*"analysis" + ... + ***0.046*"real time"*** + ... + ***0.022*"internet thing"*** + ...
Share	0.0436	0.1073
Emergence	0.1791 (threshold: 0.3945)	0.2314 (threshold: 0.2914)
Growth	0.0333	0.046
Visualization		

TABLE IV
RESULTS FOR THE EMERGING TECHNOLOGY BITCOIN/BLOCKCHAIN (2016)

<table>
<tr><th colspan="3">Query generated</th></tr>
<tr><td><b>Initial query</b></td><td colspan="2">“alternative currency”, “digital currency”, “alternative payment”, “digital payment”, ledger</td></tr>
<tr><td><b>Query expansion</b></td><td colspan="2">“digital money”, “digital cash”, “electronic currency”, “electronic money”</td></tr>
<tr><td><b>Final query</b></td><td colspan="2">“alternat* currenc*” OR “digit* currenc*” OR “alternat* payment*” OR “digit* payment*” OR “digit* money” OR “digit* cash” OR “electronic* currenc*” OR “electronic* money” OR “ledger*”</td></tr>
<tr><th colspan="3">Data collected</th></tr>
<tr><td><b>Documents collected</b></td><td colspan="2">
</td></tr>
<tr><th colspan="3">Results achieved</th></tr>
<tr><td><b>Data source</b></td><td>Papers</td><td>Patents</td></tr>
<tr><td><b>Grid search</b></td><td>
</td><td>
</td></tr>
<tr><td><b>Year detected</b></td><td>2014</td><td>2015</td></tr>
<tr><td><b>Topic detected</b></td><td><b><i>0.087*“bitcoin”</i></b> + 0.052*“tool” + 0.048*“impact” + <b><i>0.040*“currency”</i></b> + ... + <b><i>0.030*“cryptocurrency”</i></b> + ... + 0.023*“technology” + ...</td><td><b><i>0.215*“block chain”</i></b> + <b><i>0.152*“digital currency”</i></b> + 0.098*“payment” + 0.084*“transfer” + ... <b><i>0.037*“alternative payment”</i></b> + ...</td></tr>
<tr><td><b>Share</b></td><td>0.0847</td><td>0.0526</td></tr>
<tr><td><b>Emergence</b></td><td>0.2217 (threshold: 0.447)</td><td>0.131 (threshold: 0.3946)</td></tr>
<tr><td><b>Growth</b></td><td>0.0248</td><td>-0.0019</td></tr>
<tr><td><b>Visualization</b></td><td>
</td><td>
</td></tr>
</table>

two case studies, probably due to the early starting date (2007). This case provides us with the opportunity to test whether our approach also yields effective results with a rather small number of documents.

The results of the analysis are shown in Table V. For paper data, an emerging topic containing terms such as *engineering* in combination with *biomaterial* and *tissue engineering* is already found in the year 2008, indicating a medium-sized share and a positive growth. It is classified approximately at the boundary between the quadrants *evaluate* and *establish*, presumably because of its semantic similarity to the field of generic 3-D printing.

The analysis of the patent data signals an emerging topic in 2009 about *print*, *tissue,* and *3d*, which is classified into the quadrant *evaluate* with a relatively high topic share.

Although the number of documents collected was relatively low, our proposed model has been able to identify emerging topics in the papers data set in 2008 and in the patents data set in 2009, years before the initial publication of this technology in the Gartner Hype Cycle in 2011. This suggests that even a small number of documents may be sufficient to observe a certain topic area. Particularly for newly emerging technologies this may often be the case, and our model has achieved a good result even under such circumstances.

D. Sensitivity Analysis

To check the robustness of our results, we now perform a sensitivity analysis testing whether a different parameterization would have changed the results. Remember that our case studies have been conducted using the gensim input setting "asymmetric" for hyperparameter $\boldsymbol{\alpha}$ and the symmetric vector $(0.01, \ldots, 0.01)$ for hyperparameter $\boldsymbol{\eta}$.

For the sensitivity analysis, we choose $\boldsymbol{\alpha}$ from the set $\mathcal{A} = \{(0.1, \ldots, 0.1), (0.2, \ldots, 0.2), \ldots, (1.0, \ldots, 1.0),$ "asymmetric"$\}$ and $\boldsymbol{\eta}$ from the set $\mathcal{E} = \{(0.01, \ldots, 0.01), (0.02, \ldots, 0.02), \ldots, (0.1, \ldots, 0.1)\}$, accounting for the fact that the entries of $\boldsymbol{\eta}$ are usually much smaller than those of $\boldsymbol{\alpha}$. More specifically, we consider the full grid resulting from the Cartesian product $\mathcal{A} \times \mathcal{E}$, which contains 110 different combinations of vectors for $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$. For each case study and data source (papers and patents), we run our approach for all 110 combinations, checking whether and when the respective technology is detected.

Table VI shows if the results obtained under the initial hyperparameter values and presented in Section VI-A to VI-C are reproduced under the 109 other vector combinations in the grid. Based on the patent data sets for the edge computing and 3-D bioprinting case studies, there are three and eight cases, respectively, for which the technology is detected one year later. Also, for three hyperparameter constellations the bitcoin/blockchain technology is found in the papers data set two years later as compared with the original settings. It should be noted that even in these less favorable cases the detection occurred at the latest in the year in which the technology was first listed in the Gartner Hype Cycle. However, the edge computing technology was not found at all in the papers and the patents data sets under 7 and 3 hyperparameter settings, respectively. Interestingly, such worse performance tends to occur more frequently for higher entries in the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$. This tendency supports our assumption that low values for the hyperparameter vectors of our model are beneficial for our purposes (see Section IV-D).

Nevertheless, on average over all case studies and data sources, the misses account for a mere 1.53% of the cases, and the late detections for 2.14%. In contrast, the huge majority of hyperparameter settings (96.33% on average) have led to the exact same results as reported before. These findings suggest a low sensitivity of our approach to the hyperparameters employed.

VII. Discussion

Results from our three case studies demonstrate that it is indeed possible to detect emerging technological innovations earlier than traditional approaches, given the Gartner Hype Cycle for Emerging Technologies as a reference benchmark. The effectiveness of our approach has a number of implications for the management of technology and innovation, which will be presented first in this section, while later this section will discuss some limitations of our method, together with opportunities for future research.

A. Implications for Technology and Innovation Management

Results from our three case studies indicate that early signals of technological innovations can be found at different points in time in different databases: In the bitcoin/blockchain case study, we detected the new topic first in the papers data set (2014), then in the patents data set (2015), prior to its first publication in the Gartner Hype Cycle for Emerging Technologies (2016). Similarly, signals for 3-D bioprinting were spotted by our approach in scientific papers first (2008), then in patent documents (2009), before the technology first appeared in the Gartner Hype Cycle (2011). For edge computing, the topic was detected in both papers and patent data in 2016, and its first mention on the Gartner Hype Cycle dates to the year 2017. These results have a defining implication on the management of technology and innovation: Monitoring the trajectory of the detected emerging technology at an early stage can help firms predict future change and its estimated time to market. Hereby, each data source type can be considered to reflect a different stage of the technological lifecycle. Managers might benefit from an improved understanding of this diffusion process when they analyze the proposed metrics share, emergence, and growth, with respect to each database. This understanding could help them to explore potential new opportunities ahead of the competition and to detect emerging risks at an early stage in order to develop mitigation strategies against them [22], [81], [82].

Our results further reveal that big data sets are not a mandatory prerequisite to detect early signals of technological innovations. In our third case study on 3-D bioprinting, only 108 additional papers from 2008 (and a total of 236 papers) as well as 64 additional patents from 2009 (and a total of 211 patents) were necessary to spot the emerging and trending topic in the respective data sets. Given that in most cases only few data points are likely to exist at the outset of a new technological development,

TABLE V
RESULTS FOR THE EMERGING TECHNOLOGY 3-D BIOPRINTING (2011)

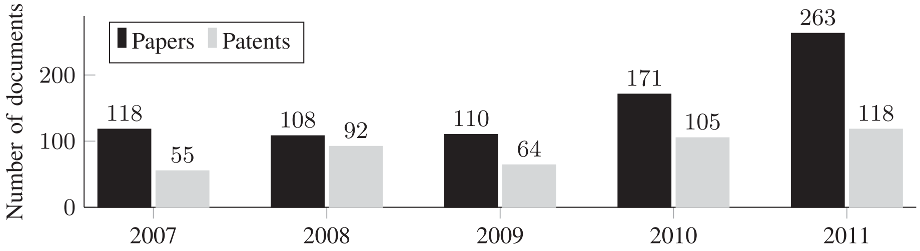
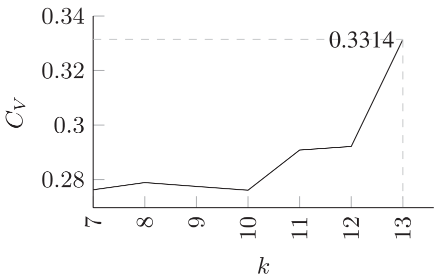
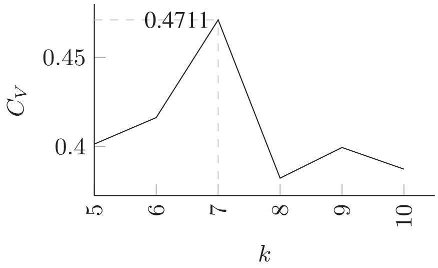
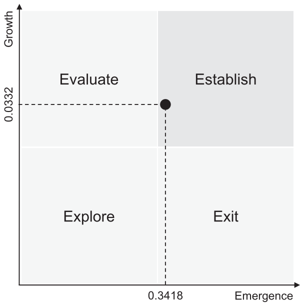
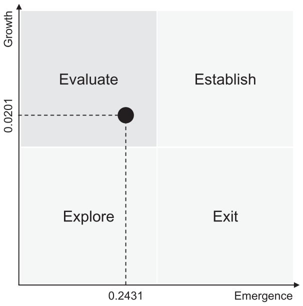
	Query generated	
Initial query	“3d printing”	
Query expansion	“three dimensional printing”, “three-D printing”, “3d printer”, “three dimensional printer”, “additive manufacturing”	
Final query	“3d print*” OR “three dimensional print*” OR “three-D print*” OR “3d print*” OR “three dimensional print*” OR “additive manufactur*”	
	Data collected	
Documents collected		
	Results achieved	
Data source	Papers	Patents
Grid search		
Year detected	2008	2009
Topic detected	***0.099*“engineering”*** + ***0.066*“biomaterial”*** + ***0.047*“cell”*** + ... + ***0.032*“tissue engineering”*** + 0.026*“challenge” + 0.026*“developed” + ...	***0.218*“print”*** + ***0.085*“tissue”*** + 0.071*“medium” + 0.063*“light” + 0.062*“liquid” + ***0.055*“3d”*** + ...
Share	0.092	0.15
Emergence	0.3418 (threshold: 0.3206)	0.2431 (threshold: 0.3141)
Growth	0.0332	0.0201
Visualization	 Growth Evaluate Establish 0.0332 Explore Exit 0.3418 Emergence	 Growth Evaluate Establish 0.0201 Explore Exit 0.2431 Emergence

TABLE VI
PERFORMANCE FOR THE 109 OTHER HYPERPARAMETER SETTINGS AS COMPARED WITH THE ORIGINAL PARAMETERIZATION

	Edge Computing		**Bitcoin/Blockchain**		**3-D Bioprinting**		**Average**
	Papers	**Patents**	**Papers**	**Patents**	**Papers**	**Patents**	**Percentage**
Identical result	102	103	106	109	109	101	96.33%
Detection delayed by 1 year	0	3	0	0	0	8	1.68%
Detection delayed by 2 years	0	0	3	0	0	0	0.46%
Technology not detected	7	3	0	0	0	0	1.53%

this property of our approach is making it practical for early warning in the real world.

Moreover, the effectiveness of the selected and combined machine learning techniques may provide both theoretical and practical guidance on which of these techniques to apply when mining emerging technological innovations, and which data sources to exploit at which stage of the technological lifecycle.

Due to the high level of automation in our approach, there are basically no limits to the number of monitored search fields once they have been set up. This provides firms with the opportunity to monitor many strategically relevant search fields in less time than required by human experts performing this task manually. We therefore recommend firms to consider this or a similar approach as a further value-adding tool alongside existing and proven methods in the toolbox of strategic foresight. It may allow humans to devote their valuable time to interpret the results and to make decisions, leaving them at a better starting position in the race after competitive vitality and competitive advantage [5], [22].

B. Limitations and Future Research Opportunities

Some limitations and future research opportunities derived from them are as follows.

First, since no commonly accepted queries and data sets for analyzing the emergence and trajectory of technologies and innovations have yet been established, every firm has to build its own queries and data sets. The query generation carried out to this end is often subjective and susceptible to the human actor bias [6], [8], [9], [43]. To counteract this limitation, we have proposed and applied a structured query expansion approach (see Section IV-A). Future efforts in this research field should focus on providing substantial amounts of standardized data for this purpose so that the risk of the human actor bias can be further reduced.

Second, the selection, parameterization and combination of the machine learning techniques applied, as well as the parameter thresholds learnt and used, can be subject to legitimate debate. We thoroughly developed our data preprocessing pipeline based on prior research (see Section IV-C), carefully selected and tuned the hyperparameters as well as the optimal number of topics for LDA (see Section IV-D), and learnt the parameter of the similarity threshold in a supervised manner (see Section V). Still, to a certain extent our approach incorporates our own beliefs about the data and the expected results. However, experiments by exhaustive grid search over the LDA hyperparameter vectors α and η (see Section VI-D) have found that our results are not highly sensitive to the hyperparameterization of our model.

Third, our article is also limited by the assumption that data about emerging technological innovations can be found in the databases used in this article, which might not always be the case (e.g., when companies conceal information about highly confidential R&D projects). Furthermore, our work assumes that the Gartner Hype Cycle for Emerging Technologies, which is considered to be a thought leader in this field, publishes the emergence of new technologies as early as possible.

Fourth, our method for detecting emergence is based solely on the semantics of the topical structure of the analysis result. Future studies could also take into account further potentially relevant information (such as the author or the firm associated with a newly detected topic).

Finally, our model is not capable of explaining the reasons behind a growing or declining trend of a particular topic. The interpretation for such a phenomenon detected (e.g., symbiotic, predator prey, or competitive interaction effects; see Section II) is left to the human experts.

VIII. CONCLUSION

This article was motivated by the lack of data-driven support and automation in identifying and analyzing relevant changes in the corporate environment of firms. We have therefore developed an AI-based model using unsupervised machine learning techniques to support the early stages of the strategic foresight process, which helps detect emerging technologies and innovations at an early stage.

In contrast to existing approaches in this field, which require frequent manual intervention and individual adjustments, our model has been designed to achieve the highest possible degree of automation. Its modules and components are based on the latest state of knowledge in this field, and they have been adapted, expanded, and combined in such a way that the model automatically adapts itself to given data. It already supports the human expert in the initial definition of the search field by suggesting synonyms and related terms. The built-in component for document harmonization allows any other machine-readable database to be connected and used. It automatically cleans and preprocesses the data collected at regular time intervals, it learns different topic models, it autonomously selects the best-fitting topic model, it analyzes the topics learnt, and it then visualizes them for interpretation and decision making by the human expert. To our knowledge, this is the first model to automate the entire data analysis process without requiring any human intervention [6]. As research advances in this field, its modular structure allows for adding new data sources as well as for extending and replacing individual components or entire modules.

We have trained the model on ground truth data in a supervised manner using documents from the AAAI, and have tested it in an unsupervised setting with the help of selected technologies of the Gartner Hype Cycle in three case studies. Our results indicate that it would have been possible to recognize the selected technologies before they were published on the Hype Cycle, if the model had been in place at that time. We recommend to permanently set up such a system for a variety of strategically relevant search fields in order increase the likelihood of detecting future technologies and innovation not yet known today.

We hope that this article will inspire future research and will help practitioners recognize relevant changes more quickly for making better-informed decisions benefiting our economy and society.

References

[1] J. M. Utterback, *Mastering the Dynamics of Innovation: How Companies Can Seize Opportunities in the Face of Technological Change*. Cambridge, MA, USA: Harvard Univ. Press, 1994.

[2] R. Rohrbeck and M. Bade, "Environmental scanning, futures research, strategic foresight and organizational future orientation: A review, integration, and future research directions," in *Proc. XXIII ISPIM Annu. Conf.*, 2012, pp. 1–14.

[3] M. Coccia, "Sources of technological innovation: Radical and incremental innovation problem-driven to support competitive advantage of firms," *Technol. Anal. Strategic Manage.*, vol. 29, no. 9, pp. 1048–1061, 2017.

[4] A. Gordon, *Future Savvy*. New York, NY, USA: Amer. Manage. Assoc., 2009.

[5] J. Keller and H. A. von der Gracht, "The influence of information and communication technology (ICT) on future foresight processes: Results from a Delphi survey," *Technol. Forecasting Social Change*, vol. 85, pp. 81–92, 2014.

[6] C. Mühlroth and M. Grottke, "A systematic literature review of mining weak signals and trends for corporate foresight," *J. Bus. Econ.*, vol. 88, no. 5, pp. 643–687, 2018.

[7] T. U. Daim, G. Rueda, H. Martin, and P. Gerdsri, "Forecasting emerging technologies: Use of bibliometrics and patent analysis," *Technol. Forecasting Social Change*, vol. 73, no. 8, pp. 981–1012, 2006.

[8] M. A. Palomino, A. Vincenti, and R. Owen, "Optimising web–based information retrieval methods for horizon scanning," *Foresight*, vol. 15, no. 3, pp. 159–176, 2013.

[9] D. H. Milanez, L. I. L. de Faria, R. M. do Amaral, D. R. Leiva, and J. A. R. Gregolin, "Patents in nanotechnology: An analysis using macro-indicators and forecasting curves," *Scientometrics*, vol. 101, no. 2, pp. 1097–1112, 2014.

[10] R. Rohrbeck, N. Thom, and H. M. Arnold, "IT tools for foresight: The integrated insight and response system of Deutsche Telekom innovation laboratories," *Technol. Forecasting Social Change*, vol. 97, no. 8, pp. 115–126, 2015.

[11] O. Saritas and J. E. Smith, "The big picture – trends, drivers, wild cards, discontinuities and weak signals," *Futures*, vol. 43, no. 3, pp. 292–312, 2011.

[12] J. Kim, M. Hwang, D.-H. Jeong, and H. Jung, "Technology trends analysis and forecasting application based on decision tree and statistical feature analysis," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12 618–12 625, 2012.

[13] J. Huang, M. Peng, H. Wang, J. Cao, W. Gao, and X. Zhang, "A probabilistic method for emerging topic tracking in microblog stream," *World Wide Web*, vol. 20, no. 2, pp. 325–350, 2017.

[14] A. D. Andersen and P. D. Andersen, "Innovation system foresight," *Technol. Forecasting Social Change*, vol. 88, pp. 276–286, 2014.

[15] J. C. Barbieri and A. C. T. Álvares, "Sixth generation innovation model: Description of a success model," *RAI Revista de Administração e Inovação*, vol. 13, no. 2, pp. 116–127, 2016.

[16] W. B. Arthur and W. Polak, "The evolution of technology within a simple computer model," *Complexity*, vol. 11, no. 5, pp. 23–31, 2006.

[17] S. Valverde, R. V. Solé, M. A. Bedau, and N. Packard, "Topology and evolution of technology innovation networks," *Phys. Rev. E, Statist., Nonlinear, Soft Matter Phys.*, vol. 76, no. 5, pp. 193–199, 2007.

[18] C. S. Curran and J. Leker, "Patent indicators for monitoring convergence—Examples from NFF and ICT," *Technol. Forecasting Social Change*, vol. 78, no. 2, pp. 256–273, 2011.

[19] B. Kim, G. Gazzola, J. M. Lee, D. Kim, K. Kim, and M. K. Jeong, "Inter-cluster connectivity analysis for technology opportunity discovery," *Scientometrics*, vol. 98, no. 3, pp. 1811–1825, 2014.

[20] F. Caviggioli, "Technology fusion: Identification and analysis of the drivers of technology convergence using patent data," *Technovation*, vol. 55–56, pp. 22–32, 2016.

[21] M. Coccia, "A theory of classification and evolution of technologies within a generalised darwinism," *Technol. Anal. Strategic Manage.*, vol. 31, no. 5, pp. 517–531, 2018.

[22] J. M. Utterback, C. Pistorius, and E. Yilmaz, "The dynamics of competition and of the diffusion of innovations," MIT Sloan School Working Paper, 5519-18, 2019.

[23] M. Coccia, "The theory of technological parasitism for the measurement of the evolution of technology and technological forecasting," *Technol. Forecasting Social Change*, vol. 141, pp. 289–304, 2019.

[24] C. A. O'Reilly and M. L. Tushman, "Ambidexterity as a dynamic capability: Resolving the innovator's dilemma," *Res. Org. Behav.*, vol. 28, pp. 185–206, 2008.

[25] K. T. Ulrich and S. D. Eppinger, *Product Design and Development*, 5th ed. New York, NY, USA: McGraw-Hill, 2012.

[26] T. N. Al-Geddawy and H. Elmaraghy, "Optimum granularity level of modular product design architecture," *CIRP Ann.—Manuf. Technol.*, vol. 62, no. 1, pp. 151–154, 2013.

[27] H. I. Ansoff, "Managing strategic surprise by response to weak signals," *California Manage. Rev.*, vol. 18, no. 2, pp. 21–33, 1975.

[28] R. Rohrbeck, C. Battistella, and E. Huizingh, "Corporate foresight: An emerging field with a rich tradition," *Technol. Forecasting Social Change*, vol. 101, pp. 1–9, 2015.

[29] H. A. von der Gracht, C. R. Vennemann, and I.-L. Darkow, "Corporate foresight and innovation management: A portfolio-approach in evaluating organizational development," *Learn. Future Faster*, vol. 42, no. 4, pp. 380–393, 2010.

[30] L. Kölbl, C. Mühlroth, F. Wiser, M. Grottke, and C. Durst, "Big data im innovationsmanagement: Wie machine learning die suche nach trends und technologien revolutioniert," *HMD Praxis der Wirtschaftsinformatik*, vol. 56, no. 5, pp. 900–913, 2019.

[31] R. Rohrbeck and M. E. Kum, "Corporate foresight and its impact on firm performance: A longitudinal analysis," *Technol. Forecasting Social Change*, vol. 129, pp. 105–116, 2018.

[32] A. J. Mills, E. Wiebe, and G. Durepos, "Retrospective case study," in *Encyclopedia of Case Study Research*, A. J. Mills, E. Wiebe, and G. Durepos, Eds. Newbury Park, CA, USA: Sage, 2010.

[33] R. K. Yin, *Case Study Research and Applications: Design and Methods*, 6th ed. Newbury Park, CA, USA: Sage, 2018.

[34] O. Dedehayir and M. Steinert, "The hype cycle model: A review and future directions," *Technol. Forecasting Social Change*, vol. 108, pp. 28–41, 2016.

[35] R. Amara and A. J. Lipinski, *Business Planning for an Uncertain Future: Scenarios & Strategies*. New York, NY, USA: Pergamon, 1983.

[36] R. Eckhoff, J. Frank, M. Markus, M. Lassnig, and S. Schön, "Detecting innovation signals with technology-enhanced social media analysis—experiences with a hybrid approach in three branches," *Int. J. Innov. Scientific Res.*, vol. 17, no. 1, pp. 120–130, 2015.

[37] O. Ena, N. Mikova, O. Saritas, and A. Sokolova, "A methodology for technology trend monitoring: The case of semantic technologies," *Scientometrics*, vol. 108, no. 3, pp. 1013–1041, 2016.

[38] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[39] A. Agrawal, W. Fu, and T. Menzies, "What is wrong with topic modeling? and how to fix it using search-based software engineering," *Inf. Softw. Technol.*, no. 98, pp. 74–88, 2018.

[40] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking LDA: Why priors matter," in *Proc. Neural Inf. Process. Syst. Conf.*, 2009, pp. 1973–1981.

[41] F. D. Macías-Escrivá, R. Haber, R. del Toro, and V. Hernandez, "Self-adaptive systems: A survey of current approaches, research challenges and applications," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7267–7279, 2013.

[42] R. Rohrbeck, "Trend scanning, scouting and foresight techniques," in *Management of the Fuzzy Front End of Innovation*, O. Gassmann and F. Schweitzer, Eds. Berlin, Germany: Springer, 2014, pp. 59–73.

[43] Y. Huang, Y. Zhang, J. Ma, A. Porter, and X. Wang, "Tracing technology evolution pathways by combining tech mining and patent citation analysis," in *Portland Int. Conf. Manage. Eng. Technol.*, 2015.

[44] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1698–1735, 2017.

[45] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, 2017, no. 31, pp. 4444–4451.

[46] L. M. Aiello *et al.*, "Sensing trending topics in Twitter," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.

[47] D. Thorleuchter and D. van den Poel, "Weak signal identification with semantic web mining," *Expert Syst. Appl.*, vol. 40, no. 12, pp. 4978–4985, 2013.

[48] H. Guo, S. Weingart, and K. Börner, "Mixed-indicators model for identifying emerging research areas," *Scientometrics*, vol. 89, no. 1, pp. 421–435, 2011.

[49] S. Kim *et al.*, "NEST: A quantitative model for detecting emerging trends using a global monitoring expert network and Bayesian network," *Futures*, vol. 52, pp. 59–73, 2013.

[50] A. Karl, J. Wisnowski, and W. H. Rushing, "A practical guide to text mining with topic extraction," *Wiley Interdisciplinary Rev.: Comput. Statist.*, vol. 7, pp. 326–340, 2015.

[51] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1373–1378.

[52] S. Bird, E. Klein, and E. Loper, *Natural Language Processing With Python* (ser. Safari Books Online). Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.

[53] C.-M. Tan, Y.-F. Wang, and C.-D. Lee, "The use of bigrams to enhance text categorization," *Inf. Process. Manage.*, vol. 38, no. 4, pp. 529–546, 2002.

[54] J. H. Lau, T. Baldwin, and D. Newman, "On collocations and topic models," *ACM Trans. Speech Lang. Process.*, vol. 10, no. 3, pp. 1–14, 2013.

[55] J. Zeng, J. Duan, W. Cao, and C. Wu, "Topics modeling based on selective Zipf distribution," *Expert Syst. Appl.*, vol. 39, no. 7, pp. 6541–6546, 2012.

[56] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic Bull. Rev.*, vol. 21, no. 5, pp. 1112–1130, 2014.

[57] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[58] S. Madnick, W. L. Woon, and A. Henschel, "Technology forecasting using data mining and semantics: Third & final annual report," MIT Sloan Working Paper CISL# 2011-01, 2011.

[59] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent Dirichlet allocation," in *Proc. Neural Inf. Process. Syst. Conf.*, 2010, pp. 856–864.

[60] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci.*, vol. 101, pp. 5228–5235, 2004.

[61] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, 2010, pp. 45–50.

[62] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 27–34.

[63] B. Grün and K. Hornik, "topicmodels: An R package for fitting topic models," *J. Statist. Softw.*, vol. 40, no. 1, pp. 1–30, 2011.

[64] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[65] M. Herlihy and N. Shavit, *The Art of Multiprocessor Programming*, rev. 1st ed. San Mateo, CA, USA: Morgan Kaufmann, 2012.

[66] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1105–1112.

[67] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive LDA model selection," *Neurocomputing*, vol. 72, no. 7–9, pp. 1775–1781, 2009.

[68] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. Neural Inf. Process. Syst. Conf.*, 2009, pp. 288–296.

[69] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.

[70] F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both, "Evaluating topic coherence measures," in *Proc. Topic Models: Comput., Appl. Eval., NIPS Workshop*, 2013, pp. 1–4.

[71] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 399–408 .

[72] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, Jul. 2003.

[73] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge Univ. Press, 2008.

[74] M. Bianchi, M. Boyle, and D. Hollingsworth, "A comparison of methods for trend estimation," *Appl. Econ. Lett.*, vol. 6, no. 2, pp. 103–109, 1999.

[75] T. Oliphant, *Guide to NumPy*. Austin, TX, USA: Trelgol Publishing, 2006.

[76] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 51–56.

[77] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[78] A. Yousefpour *et al.*, "All one needs to know about fog computing and related edge computing paradigms: A complete survey," *J. Syst. Archit.*, 2019.

[79] F. Casino, T. K. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," *Telematics Informat.*, vol. 36, pp. 55–81, 2019.

[80] E. S. Bishop *et al.*, "3-D bioprinting technologies in tissue engineering and regenerative medicine: Current and future trends," *Genes Diseases*, vol. 4, no. 4, pp. 185–195, 2017.

[81] R. M. Grant and K. E. Neupert, *Cases in Contemporary Strategy Analysis*, 3rd ed. Oxford, U.K.: Blackwell, 2003.

[82] M. Coccia, "Sources of disruptive technologies for industrial change," *L'Industria - Rivista di Economia e Politica Industriale*, vol. 38, no. 1, pp. 97–120, 2017.