European Journal of Futures Research

RESEARCH ARTICLE Open Access

Trend mining with Orange – using topic modeling in futures research with the example of urban mobility

Matthias Sonk[1*] and Dirk Tunger[2]

Abstract

Today, assumptions about probable future developments (at least as far as they make use of quantifiable scientific methods and are not pure speculation) are generally based on data from the past. An interesting way to analyze the future through this type of data is text mining or individual methods out of the spectrum of text mining, such as topic modeling. Topic Modeling itself is a combination of quantitative and qualitative methodology and is based on the full spectrum of social science methodology. Therefore, the method is an interesting way for futures research to analyze futures. This publication addresses the question of how a combination of different methods can contribute to trend monitoring or trend mining. For this purpose, a set of scientific publications was first generated with the help of a search query in the Web of Science (WoS), which is the basis for all evaluations and statements and topics. In essence, the method considered here should be more fully integrated into the scientific practice of futures research because it can make a valuable contribution to estimating future development based on past development.

Keywords Text mining, Topic modeling, Bibliometric analysis, Trend mining, Mixed methods, Urban mobility, Flexible mobility

Introduction

Futures research makes a strong promise by saying, that it can provide orientational knowledge for society, policy makers, and business. This promise is not easy to keep, and futures research has repeatedly developed new methodological approaches to meet this promise. Often, quantitative methods by means of statistical procedures are used for the analysis of the future but recently, more and more qualitative methods are being used in the analysis of possible, probable, or desirable futures as well. Especially the availability of more and more data about past events e.g., in the form of texts and articles, could be more strongly integrated into the methodological approach in futures research. An interesting way to analyze the future through this type of data is text mining or individual methods out of the spectrum of text mining, such as topic modeling. An important assumption here is that text data at a high-quality level can objectively represent the past.

There are some papers in which certain future-relevant aspects are investigated with the help of text mining, but usually it is about the method of keyword extraction and not about more complex possibilities of text mining [e.g.

*Correspondence:
Matthias Sonk
m.sonk@fu-berlin.de
[1]Freie Universität Berlin, Berlin, Germany
[2]Faculty of Information Science and Communication Studies, Institute of Information Management, TH Köln and Project Management Jülich, Center of Excellence "Analyses, Studies, Strategy", Forschungszentrum Jülich GmbH, Jülich, Germany

1–5]. Nevertheless, there are already some papers, which also apply topic modeling [e.g. 6–8].

Text mining is often applied to technological topics as the method has an affinity to more quantitative scientific work. But Topic Modeling itself is rather a combination of quantitative and qualitative methodology and is based on the full spectrum of social science methodology. Therefore, the method is an interesting way for futures research to analyze futures while using a "trend mining" method more frequently.

Method and scientific approach

Gathering data

The bibliometric analysis of publications and citations is based on university affiliations in the Web of Science (WoS). The "Science Citation Index" (SCI), which was first introduced by Eugene Garfield [9] and from which WoS was subsequently developed, is the most widely-used multidisciplinary publication and citations database in the academic community. The basic idea of Garfield was to select the journals covered in the database according to their significance for the respective field area: the most relevant journals from each scientific field were to be covered (core journals). This selection procedure of WoS, which is largely based on the Journal Impact Factor (JIF) [10], led to the creation of a database which can be used for bibliometric analyses of a variety of natural sciences disciplines.

For this paper on urban mobility, a topic search was performed using the Advanced Search of the Web of Science, which contains the terms.

"mobility" or "transportation"

In order to narrow down to urban mobility, a connection was made with.

"city" OR "cities" OR "town*" OR "urban" OR "rural"

A specific search was made for publications with a reference to the future, so that the following terms were also included in the search query:

"future*" OR "trend*" OR "scenario*" OR "transition*" OR "transformation*"

In order to obtain a dataset that contains publications on urban mobility, the search was limited to disciplines of the Web of Science Subject Categories that are related to urban mobility:

"TRANSPORTATION" OR "TRANSPORTATION SCIENCE TECHNOLOGY" OR "URBAN STUDIES" OR "REGIONAL URBAN PLANNING" OR "DEMOGRAPHY".

Overall, the search was carried out very openly in order to avoid losing publications as far as possible; the time period covers the years 1991–2021, i.e. 30 years. A long period was deliberately chosen in order to reflect the trend development as comprehensively as possible. For the further analysis, a data set was generated which, in addition to the titles and keywords, also contained the abstracts of the relevant publications (about 4400).

The overall search strategy chosen for Web of Science was.

(TS=("mobility" OR "transportation" AND ("city" OR "cities" OR "town*" OR "urban" OR "rural")) AND TS=("future*" OR "trend*" OR "scenario*" OR "transition*" OR "transformation*")) AND (TASCA==("TRANSPORTATION" OR "TRANSPORTATION SCIENCE TECHNOLOGY" OR "URBAN STUDIES" OR "REGIONAL URBAN PLANNING" OR "DEMOGRAPHY").

Stopwords

Using the final search strategy, a download with the corresponding publications was generated from the Web of Science, containing abstracts and keywords to the publications in addition to the bibliographic information (e.g., title, journal name, etc.). However, in order to generate quantitative evaluations from this, further preparation is required: The word frequency of individual terms cannot simply be determined from the text corpus, because this would mean that the words with the highest frequency and at the same time the lowest significance would be ranked highly. To prevent this, it is necessary to have a list of terms that are defined in advance as not carrying meaning and thus are not used further in the analysis. Such a list is called a "stopword list" and it contains, for example, all kinds of numbers and years, number words, calendar months, special characters, and publisher information from the abstracts. However, this list also contains terms that do not make sense in a quantitative analysis if they are taken out of context, e.g., words like "new", "although" or "search" (See Table 1).

Topic modeling

Orange: data mining

This study uses a topic modeling approach to discover abstract topics in a corpus based on clusters of words

Table 1 Examples out of the stopword list

Study	Survey	Paper	Research	Approach	Significant	Results	Evidence
conclusion	implications	findings	analysis	sample	factors	scopus	citation
IEEE	elsevier	sage	springer	wiley	methods	design	within
related	however	google	scholar	review	studies	used	although
certainty	search	terms	model	found	using	also	may

found in each document and their respective frequency. This analysis was performed using "Orange Data Mining". Orange is a machine learning and data mining suite for data analysis using Python scripting and visual programming [11]. Orange is used because it's a graphically programmable tool and coding experience is not necessary. This circumstance makes it possible that many people can replicate this method even if they have no programming experience (See Fig. 1).

There are different algorithms for topic modeling: e.g., the Latent Dirichlet Allocation (LDA), Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP). This study uses the Latent Dirichlet Allocation, which was first described by David Blei, Andrew Ng and Michael Jordan in 2003. LDA is a probabilistic model that is mainly used in the field of natural language processing. It helps to quickly determine the topic of a long text. LDA makes predictions about topics in texts based on the frequency of words that occur together. For example, a text about urban public mobility often contains the words "transit", "public", "rail", "service", "systems". Technically, LDA is a three-level hierarchical Bayesian model in which each element of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is in turn modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, topic probabilities are an explicit representation of a document. The topics are generated as a word lists, which are then each named by the authors in an interpretative and qualitative process [12–14]. Put simply, LDA attempts to find the most likely topics that can be generated in the given set of documents. This is done by iteratively mapping words to topics and adjusting the topic-word distribution until the best fit is found. These topic-word distributions are then labeled for further analysis.

The starting point for the analysis was a corpus of 4,415 scientific publications from the years 1991 to 2021. Based on scientific abstracts, topic trends were analyzed over time, and thus a general overview of the thematic structure of the corpus was generated. In relation to the evaluation over time, each topic trend was analyzed in relation to the time period of the publications used. Some topics already appeared in publications before 2000, while many other topics were only increasingly discussed in scientific discourse after 2000. This results in a LDA model, which includes 500 topics. In a first step of analysis, a linear regression was performed for each topic, examining the linear trend of the probability of occurrence over time of the topic in the documents. On this basis, 211 topics with a positive and 266 topics with a negative linear trend could be identified. 23 topics had no linear trend at all. In addition, the 84 topics with the strongest positive trend were analyzed and named. These can be seen as possible trend-setting topics for the future development of topic-specific research.

Models

The most interesting 9 models with a positive trend related to the period from 1991 to 2021 were selected and interpreted in a qualitative discussion. They are not the topics that had the strongest trend, but topics that had the highest thematic plausibility and therefore could be interpreted adequately (See Table 2).

Topics 98, 233 and 40 are focused on the development of e-mobility and it is no surprise that these topics are among those with the highest positive trends. Topic 98 is the further development of autonomous shared mobility, while topic 233 deals with even more flexible mobility and topic 40 describes the aspect of intelligent - AI-controlled - mobility. The topics 194, 464 and 249 are centered around the infrastructure regarding mobility. Topic 194 addresses the possibility of using e-mobiles as electricity storage while stationary, topic 464 addresses the infrastructure needed for shared mobility, and topic 249 addresses the connection between housing and mobility. These three themes are also to be expected in the analyzed corpus. Rather surprising are topics 306 and 364, which deal with the functionality of public transport and the further development of rail mobility as opposed to individual mobility by car. Finally, topic 215 is about environmental protection regarding mobility, which is also quite unexpectedly ranking high.

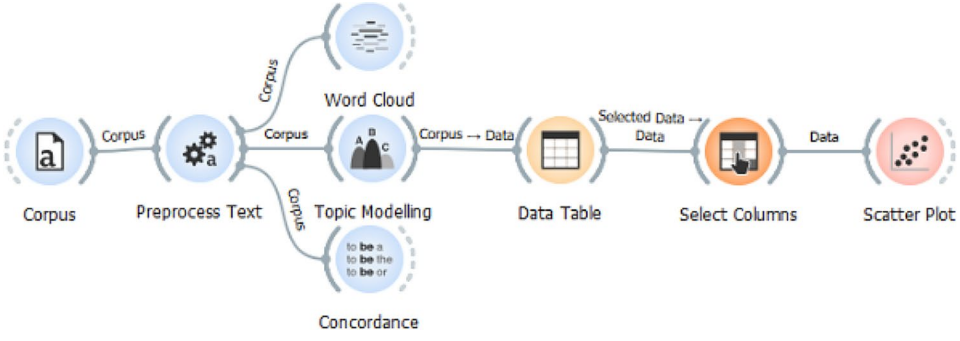

Fig. 1 Orange workflow for topic modeling

Table 2 Trend-setting topics

Topic Nr.	Topic	Word list
98	Autonomous shared mobility	Vehicles, autonomous, vehicle, automated, avs, av, shared, mobility, driving, technology
233	Flexible mobility	Sharing, shared, way, systems, users, services, floating, reservation, utilization, bike
194	e-mobility as an electricity storage system	Electric, vehicles, vehicle, battery, evs, conventional, electricity, energy, potential, charging
40	Intelligent autonomous driving	Based, trajectory, position, learning, positions, proposed, mobility, machine, trajectories, algorithm
464	Infrastructure deployment	Infrastructure, charging, stations, station, fast, public, vehicles, description, adequate, electric
249	Housing and mobility	Home, smart, concept, products, preserving, digital, engaged, drawn, natives, removing
306	Importance of public transport	Transit, public, rail, service, systems, oriented, low, area, station, influence
364	Further development of train mobility	Line, los, hsr, rail, upgrade, ratios, bands, evaluation, reflection, sight
215	Environmental protection regarding mobility	Epa, cycles, reasonable, deprivation, negotiating, control, introduced, area, agency, hood

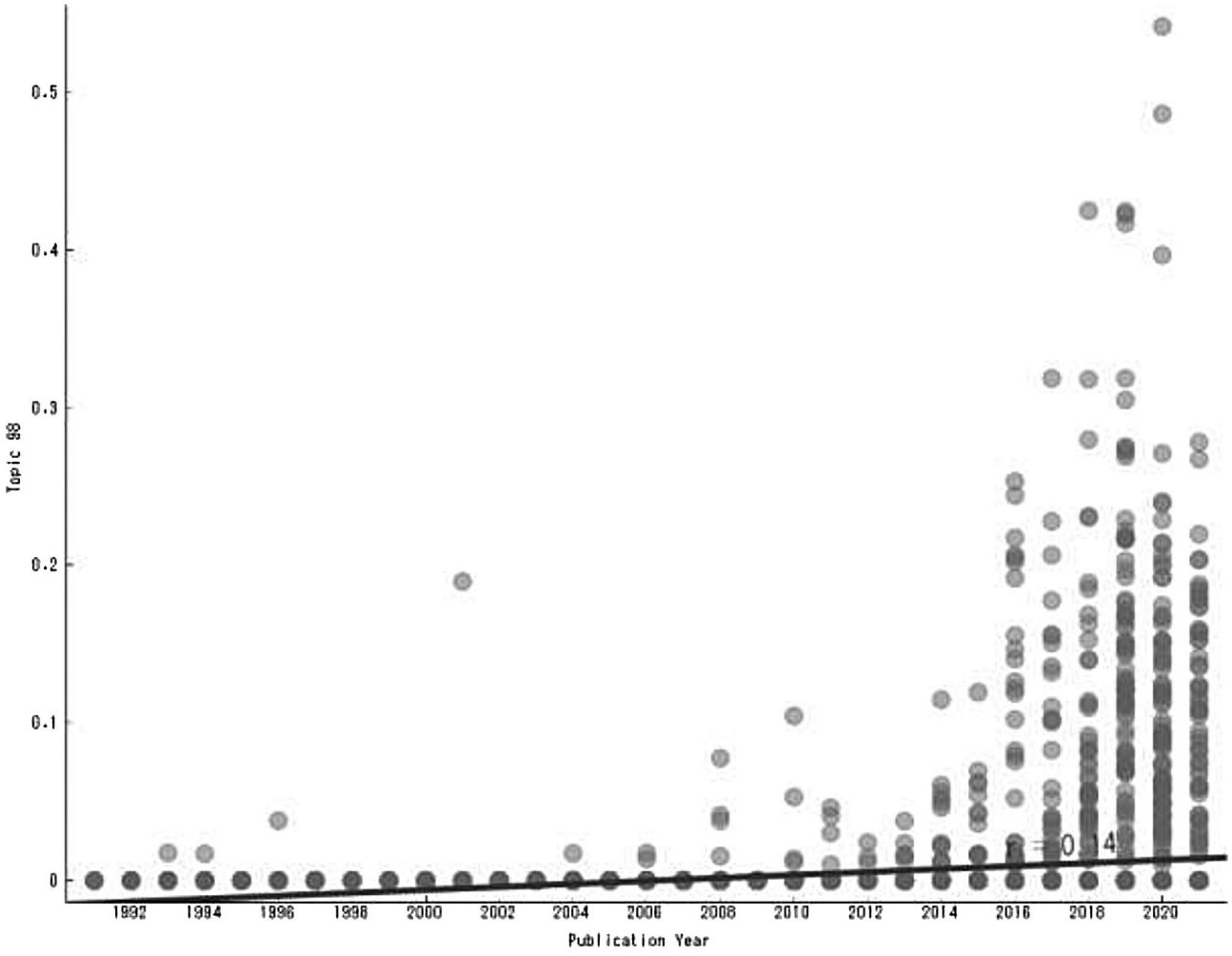

Fig. 2 Trend-setting topic: 98 autonomous shared mobility

Example figure of a topic

Figure 2 uses the example of topic 98 "autonomous shared mobility" to show how the individual publications and the corresponding probability of reference to the topic in the individual documents are distributed over time. From the entire corpus, 363 documents contained a reference to this topic with a topic probability ranging from 0.01 to 0.54. This means that these documents are more likely to contain words that belong to this topic than the rest of the corpus.

Results and discussion

Interpretation and description (model perspective)

Based on the 9 described models, a few interpretations can be derived. First, the models with the highest positive trends are at the same time very technology-centered, which is not uncommon in connection with innovation management as a very popular business-related specialization of futures research. For futures research, however, it is equally interesting to see to what extent trends have an impact on the political, social, or societal level, e.g., how the mobility of the future could change life in cities. Since technological developments are very dominant in the scientific literature, it is difficult to consider aspects that are not related to technology. This can be seen as a fundamental bias regarding text mining in futures research and makes it difficult to consider the social aspects of technology-dominated trends.

On the other hand, it is very interesting that topics can be derived from the models that take a broader perspective related to mobility. Thus, the topic of infrastructure, or charging infrastructure, is an important topic in the corpus. The infrastructure for e-mobility is an important driver of acceptance among the population and is necessary for a successful mobility transition. Furthermore, it is positive that the topic of mobility and housing has such a high priority in the studied literature, as e-mobility must adapt to current housing situations. Nevertheless, it is also important that current housing concepts adapt to the new mobility options to create sustainable climate-friendly mobility. It is a very positive sign that these topics also feature strongly in the literature, and this shows that future topics can be considered more broadly through the method of text mining.

Also interesting and important is the fact that local and long-distance public transport takes up quite a large part of the 9 urban mobility models. Public rail transport will continue to be a very important component of mobility in the city in the future and must adapt equally to the new requirements. Here, it can be seen that the method used allows topics to be examined from different angles. The topic of environmental protection in relation to mobility in the city also shows the direction that the discussion has taken in the scientific context. From this, too, it can be deduced that future topics can be found with the help of text mining.

Discussion (meta perspective)

The search strategy, as described in Chap. 2.1, is created as part of a process: each piece of this search strategy is first developed and tested individually before it is combined with the other parts. This makes it possible to assess whether the hits newly acquired through a search step fit the topic under investigation. In this way, a search strategy is created piece by piece, which in the end is also known to optimally represent the searched topic.

The stopword list has already been described in more detail in Chap. 2.2. Its function is to further narrow down the field of investigation with the help of meaning-bearing words. This is a subjective step because everyone sometimes sees other words as meaning bearing. Nevertheless, this is exactly what this method is also about because most of the words are probably indisputable. In the end, the stopword list works like a filter that further narrows down the field of investigation.

Today, assumptions about probable future developments (at least as far as they make use of quantifiable scientific methods and are not pure speculation) are generally based on data from the past, as collected in many kinds of statistics. The temporal development of such influencing factors is called a trend [15, 16].

Accordingly, a trend is a basic tendency that characterizes the direction in which a development is going in chart analysis, John Murphy for example, describes the direction of peaks and valleys in the graphical representation of data (e.g., stock market prices) as a trend [17, 18].

Trend developments often run with strong fluctuations and are often not linear. Every trend comes up against limits at which a maximum or minimum value can quickly be reached [15, 16].

Bibliometrics can be even further extended using data sources to provide support in recognizing trends (Ball & Tunger, 2006). The following example shows how the development of scientific topics can be analyzed with the aid of bibliometrics to provide information on future developments.

Is a trend just a chain of events consisting of coincidences, are they strategies or coincidences? Or can patterns be perceived? Trend research was introduced into classical economic theory by Igor Ansoff in 1975 and has become known as "weak signal research" [19]. This concept provides a fairly accurate description of what a trend is: a weak signal that must be identified in a large amount of data.

Three aspects will be taken into consideration when looking on trends in science [according to 20]:

a. The past is characterized by the development of the articles on the topic in question which can be found on the Web of Science literature databases. The development should be outlined over a sufficiently long period to draw the correct conclusions.
b. The present is represented by the citation behavior of the community in question. The response generated can be read off from the development of the citation curve over time.

c. The future can be derived from the convergence of the regions of the past (a) and present (b).

Although the methodology used in this publication is slightly different, a common method of futures research is to examine data from the past to generate statements about the future. Data from the Web of Science is thus a very valuable source for trends in science.

There are several aspects that speak for the method of text mining and its more frequent application in futures research. Basically, text mining as a semiquantitative method (e.g., topic modeling) is interesting for otherwise very qualitative futures research. For example, the consideration of large data sets leads to the inclusion of surprising aspects in the further process. Furthermore, the scalability of the method brings a high flexibility: Very focused, but also very broad text corpora can be considered, which leads to different trend observations.

The method also helps to give unknown topics a structure. This is also very interesting for futures research, since, for example, environment analyses are needed for the concretization of scenarios, and these can be supplemented by text mining. Of course, looking at big thematically focused data can also lead to novel results and entirely new pictures of the future can emerge.

Finally, the use of large amounts of text data is a way to look at the past of a thematically focused discussion. This is an important aspect and requirement for looking into the future because trends result from developments in the past and present.

Conclusions and outlook

This publication addresses the question of how a combination of different methods can contribute to trend monitoring. For this purpose, a set of scientific publications was first generated with the help of a search query in the Web of Science, which is the basis for all evaluations and statements. With the help of a stopword list, words without meaningfulness were removed before further evaluation, so that topic modeling is based solely on meaningful words. The most relevant topics were identified and named. As a result, we obtain a list that contains essential topics of urban mobility. The set of methods we used describes a possibility of foresight to use a combination of qualitative and quantitative methods, the results of which are as little subjective as possible. Of course, a search query contains subjective elements, as does a stopword list. With the help of the statistics contained in the topic modeling, however, this subjectivity is to be removed to some extent. A little subjectivity remains, as in every method of foresight.

The problem of biases in the study cannot be completely reduced, as the perspective of the researchers, the selection of stopwords, the database used, and the qualitative analysis cannot be free of external or personal influences. In the context of topic modeling, researchers must be aware of these external influences and biases. This also applies to all other social science methods.

In essence, the method considered here should be more fully integrated into the scientific practice of futures research because it can make a valuable contribution to estimating future development based on past development: this is an important source of data precisely because the knowledge contained in many individual publications can be regarded as the wisdom of the crowds, especially when it is considered cumulatively. And since Orange is a graphically programmable tool and coding experience is not necessary it is easy to apply for everybody. This circumstance makes it possible that many people in futures research can replicate this method even if they have no programming experience.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40309-024-00229-1.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
Both authors contributed equally.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Data availability
The data and materials are presented in additional supporting files.

Declarations

Competing interests
Not applicable.

Received: 4 July 2023 / Accepted: 13 February 2024
Published online: 11 March 2024

References

1. Lee M, Kim S, Kim H et al (2022) Technology Opportunity Discovery using deep learning-based text mining and a knowledge graph. Technol Forecast Soc Chang 180:121718. https://doi.org/10.1016/j.techfore.2022.121718
2. Lim C, Cho G-H, Kim J (2021) Understanding the linkages of smart-city technologies and applications: key lessons from a text mining approach and a call for future research. Technol Forecast Soc Chang 170:120893. https://doi.org/10.1016/j.techfore.2021.120893
3. Moro A, Joanny G, Moretti C (2020) Emerging technologies in the renewable energy sector: a comparison of expert review with a text mining software. Futures 117:102511. https://doi.org/10.1016/j.futures.2020.102511
4. Gokhberg L, Kuzminov I, Khabirova E et al (2020) Advanced text-mining for trend analysis of Russia's Extractive industries. Futures 115:102476. https://doi.org/10.1016/j.futures.2019.102476
5. Kayser V, Blind K (2017) Extending the knowledge base of foresight: the contribution of text mining. Technol Forecast Soc Chang 116:208–215. https://doi.org/10.1016/j.techfore.2016.10.017

6. Ma T, Zhou X, Liu J et al (2021) Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. Technol Forecast Soc Chang 173:121159. https://doi.org/10.1016/j.techfore.2021.121159
7. Rosa AB, Gudowsky N, Repo P (2021) Sensemaking and lens-shaping: identifying citizen contributions to foresight through comparative topic modelling. Futures 129:102733. https://doi.org/10.1016/j.futures.2021.102733
8. Erzurumlu SS, Pachamanova D (2020) Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. Technol Forecast Soc Chang 156:120041. https://doi.org/10.1016/j.techfore.2020.120041
9. Garfield E (1964) Science Citation Index-A New Dimension in indexing. Science 144:649–654. https://doi.org/10.1126/science.144.3619.649
10. Garfield E (1972) Citation analysis as a tool in journal evaluation. Science 178:471–479. https://doi.org/10.1126/science.178.4060.471
11. Demsar J, Curk T, Erjavec A et al (2013) Orange: Data Mining Toolbox in Python. J Mach Learn Res 14:2349–2353
12. Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. J Mach Learn Res 3:993–1022
13. Blei DM (2012) Probabilistic topic models. Commun ACM 55:77–84
14. Vayansky I, Kumar SA (2020) A review of topic modeling methods. Inform Syst 94:101582. https://doi.org/10.1016/j.is.2020.101582
15. Leutzbach W (2000) Das Problem mit der Zukunft: wie sicher sind Voraussagen? Alba, Düsseldorf
16. Dwivedi YK, Venkitachalam K, Sharif AM et al (2011) Research Trends in Knowledge Management: analyzing the Past and Predicting the Future. Inform Syst Manage 28:43–56. https://doi.org/10.1080/10580530.2011.536112
17. Murphy J (2006) Technische Analyse Der Finanzmärkte: Grundlagen, Strategien, Methoden, Anwendungen. FinanzBuch, München
18. Madsen DØ, Silva ES, Sohail SS (2023) 15 years of research on Google Trends. A bibliometric review and future research directions
19. Holopainen M, Toivonen M (2012) Weak signals: Ansoff today. Futures 44:198–205. https://doi.org/10.1016/j.futures.2011.10.002
20. Ball R, Tunger D (2006) Bibliometric analysis - a new business area for information professionals in libraries? Scientometrics 66:561–577. https://doi.org/10.1007/s11192-006-0041-0

Publisher's Note