

Obtaining More Specific Topics and Detecting Weak Signals by Topic Word Selection



Laura Kölbl and Michael Grottke

Abstract With topic modeling methods, such as Latent Dirichlet Allocation (LDA), we can find topics in large text collections. To efficiently employ this information, there is a need for a method that automatically analyzes the topics with respect to their usefulness for applications like the detection of new innovations. This paper presents a novel method to automatically evaluate topics produced by LDA. The new approach puts the focus on finding topics with topic words that are not only coherent, but also specific. By using the documents associated with each word to calculate background topics, a baseline can be set for each topic word that helps assess whether its context fits the topic well. Experiments indicate that the resulting topics are more manageable in terms of their interpretability. Moreover, we show that the approach can be used to detect weak signals.

Keywords Text mining · Topic modeling · Weak signals · Topic coherence

1 Introduction

A major drawback in topic modeling is the (lack of) interpretability of topics. Especially for a large number of topics, human evaluation of models and topics is a time-consuming task. Since the results of such a human evaluation do not correlate well with automatic measures like perplexity and held-out likelihood [1], there is a need for more adequate metrics. State-of-the-art topic coherence measures perform fairly well [2], but they do not consider how specific and meaningful a word is.

Depending on the task it can be crucial that the topics are very specific. Chang et al. [1] found that topic coherence (as measured by humans) declines with an increasing

L. Kölbl (✉) · M. Grottke
Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberg, Germany
e-mail: laura.koelbl@fau.de

M. Grottke
e-mail: michael.grottke@fau.de; michael.grottke@gfk.com

M. Grottke
GfK SE, Global Data Science, Nürnberg, Germany

© Springer Nature Switzerland AG 2020

H. Pham (ed.), *Reliability and Statistical Computing*, Springer Series
in Reliability Engineering, https://doi.org/10.1007/978-3-030-43412-0_12

number of topics. However, fine-grained topics are important for application tasks such as emerging trend detection. Therefore, there is a need for an approach capable of identifying good topics among the many topics found by LDA.

In this paper, we assess the quality of a topic in terms of both coherence and specificity by using context information from the document corpus.

2 Related Work

2.1 Latent Dirichlet Allocation

LDA [3] is a generative probabilistic model of a document collection. It assumes that each document is a mixture of latent topics, which in turn are mixtures of words. The input is a corpus \mathcal{C} consisting of D documents. \mathcal{V} is the vocabulary containing all the different words present in this corpus. Document $d \in \{1, \dots, D\}$ consists of N_d words.

The topic-word distributions ϕ_1, \dots, ϕ_K of the K topics are drawn from a Dirichlet distribution with hyperparameter vector β , while the topic distributions $\theta_1, \dots, \theta_D$ of the D documents are sampled from a Dirichlet distribution with hyperparameter vector α . The documents are thus assumed to be generated using Algorithm 1.

Algorithm 1: Document generation assumed by LDA

```

1  $\phi_k \sim \text{Dir}(\beta), k \in 1, \dots, K;$ 
2 for the  $d^{\text{th}}$  document,  $d \in \{1, \dots, D\}$ , do
3    $\theta_d \sim \text{Dir}(\alpha);$ 
4   for the  $n^{\text{th}}$  word in document  $d$ ,  $n \in \{1, \dots, N_d\}$ , do
5     Choose a topic  $z_{d,n} \sim \text{Multinomial}(\theta_d);$ 
6     Choose a word  $t_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}});$ 
7   end
8 end

```

2.2 Coherence Measures

There are different approaches to the handling of topics modeled with LDA that are related to our work.

The paper by Lau et al. [4] about selecting topic words as well as subsequent papers on this subject matter [5] have aimed at finding the best label for a topic.

Recent work by Alokaili [6] is concerned with topic word reranking; the aim of this paper is to enhance the quality of the topics by reordering the topic words.

Although topic reranking seems closely related to our work, the methods used in the paper do not incorporate any semantic context information, but they rely for example on term frequencies or probabilities. We will use word selection in order to find specific topics with an approach that emphasizes context information; therefore, we will employ coherence measures as a benchmark in our work. Topic coherence captures the internal relatedness of the topic.

Research on topic coherence measures has mainly focused on topic words. In the literature, usually the top- W words (where $W = 5$ or $W = 10$), evaluated by their topic-word probabilities, are defined to form the set of topic words for topic k , $\mathcal{W}_k = \{w_{k,1}, \dots, w_{k,W}\}$. Word topic coherence can be determined by comparing pairs of these words [7], [8], or by comparing words with word subsets, as done by Rosner et al. [9] and Röder et al. [2], who considered one-all, one-any and one-many comparisons.

According to Röder et al. [2], the measure that showed the best results for topic coherence compared with human evaluation was the C_V measure, which is based on comparing each topic word with the entire topic word set. Calculating how $w_{k,i}$ is supported by \mathcal{W}_k results in a vector $\mathbf{v}_{k,i}$ whose j^{th} element represents the comparison of $w_{k,i}$ with $w_{k,j}$ using the normalized pointwise mutual information (NMPI):

$$v_{k,i,j} = \text{NMPI}(w_{k,i}, w_{k,j}) = \frac{\ln \frac{P(w_{k,i}, w_{k,j}) + \epsilon}{P(w_{k,i}) \cdot P(w_{k,j})}}{-\ln(P(w_{k,i}, w_{k,j}) + \epsilon)}.$$

The small constant ϵ is added in the logarithms to avoid problems with zero probabilities. These (joint) word probabilities are computed using a Boolean sliding window.

While $\mathbf{v}_{k,i}$ could be used as a direct confirmation measure for topic word $w_{k,i}$, the C_V metric is based on comparing this vector with vector \mathbf{v}_k^* , representing the support of word set \mathcal{W}_k with each word in \mathcal{W}_k . The j th element of this vector can easily be calculated as

$$v_{k,j}^* = \sum_{i=1}^W v_{k,i,j}.$$

An indirect confirmation measure for topic word $w_{k,i}$ is then calculated as the cosine vector similarity between $\mathbf{v}_{k,i}$ and \mathbf{v}_k^* ,

$$\psi_{k,i} = \text{Sim}_{\cos}(\mathbf{v}_{k,i}, \mathbf{v}_k^*),$$

and the C_V metric for topic k is the arithmetic average of the W measures $\psi_{k,1}, \dots, \psi_{k,W}$.

Our aim in this paper is to add information about the semantic context of the words to the coherence measure.

Recent work by Korencic et al. [10] has suggested to use document-based coherence measures for detecting concrete topics. The authors computed document-based coherence by selecting the top documents for a topic, and they evaluated their coher-

ence by calculating for example their cosine distance. The paper highlighted that some topics showed low word coherence although the documents were very similar. Document coherence approaches topic coherence from a very different and interesting angle.

In the following, we present a novel approach, which detects concrete topics by finding specific topic words.

3 Proposed Approach

AlSumait et al. [11] introduced the idea of comparing each topic with “junk topics”, which are word or document distributions that are inferred from the whole corpus. They used the similarity of a topic to these junk topics as a measure of topic quality.

Based on this idea, we also make use of background topics, but unlike in the approach by AlSumait et al. we define background topics that are word-specific in order to detect topic words that are tightly connected with a particular topic.

First, we run an LDA with K topics. We consider the top- W words of topic k to form its set of topic words \mathcal{W}_k . For each of these topic words $w_{k,i}$, we build the background corpus $\mathcal{C}_{k,i}$, i.e., the corpus consisting of all the documents that contain word $w_{k,i}$. From background corpus $\mathcal{C}_{k,i}$ a background topic $b_{k,i}$ is calculated by running an LDA with the topic number parameter set to one.

Using the Jensen-Shannon divergence (JSD), we can compute a similarity measure comparing the word distribution of background topic $b_{k,i}$ with the word distribution of any of the K topics from the original LDA:

$$\text{BGM}(k', b_{k,i}) = 1 - \text{JSD}(\phi_{k'}, \phi_{b_{k,i}}), \quad k' = 1, \dots, K.$$

If a word $w_{k,i}$ featuring in the topic list \mathcal{W}_k is specific to this topic, then comparing ϕ_k with $\phi_{b_{k,i}}$ should result in a relatively high value. On the contrary, the background topic of any word that is not specific to a topic will be unlikely to achieve the highest similarity with this one topic. Rather, if the context of the word is not closely related to the topic, then it might fit some other topic well.

Whether a topic word is to be considered specific also depends on the corpus studied. For example, a topic word like “health” could be deemed specific in a dataset about robotics but unspecific in a dataset about healthcare. We might thus be inclined to keep topic word $w_{k,i}$ in \mathcal{W}_k only if $\text{BGM}(k, b_{k,i})$ is the largest similarity value obtained.

However, even a topic-specific word might be relevant for more than one topic. In our algorithm, we therefore use the background indicator (BGI) parameter to decide whether or not a word is to be dropped from the topic list. Arranging the similarities $\text{BGM}(k', b_{k,i})$, $k' \in \{1, \dots, K\}$, in descending order, the BGI defines the rank position up to which the word is still accepted. The most restrictive choice $\text{BGI} = 1$ would lead us to the approach laid out in the previous paragraph, demanding

that the background topic $b_{k,i}$ attains the highest similarity when comparing it to topic k .

Algorithm 2: BGI Algorithm

Input: \mathcal{C} , K , BGI

```

1 Run LDA with  $K$  topics on document corpus  $\mathcal{C}$ ;
2 for the  $k^{\text{th}}$  topic,  $k \in \{1, \dots, K\}$ , do
3   for the  $i^{\text{th}}$  topic word in  $\mathcal{W}_k$ ,  $i \in \{1, \dots, W\}$ , do
4     Create background corpus  $\mathcal{C}_{k,i}$ ;
5     Obtain background topic  $b_{k,i}$  by running LDA with 1 topic on  $\mathcal{C}_{k,i}$ ;
6     Arrange similarities  $\text{BGM}(1, b_{k,i}), \dots, \text{BGM}(K, b_{k,i})$  in descending
       order;
7     if rank of  $\text{BGM}(k, b_{k,i}) > \text{BGI}$  then
8       | drop  $w_{k,i}$  from  $\mathcal{W}_k$ 
9     end
10  end
11 end

```

For a larger BGI value, such as 2, the algorithm drops fewer words from the list of topic words, because it also keeps those for which topic k only attains a lower (e.g., the second-largest) similarity value. After all, the ultimate aim of our approach is not to separate the contents of the various topics, but to assess the topic quality and to enhance the interpretability of the topics. The whole approach is depicted in Algorithm 2.

By selecting the topic words, our approach is implicitly selecting topics as well: if all of the topic words of a topic are dropped, then the topic can be considered unspecific.

Although this approach seems computationally intensive, there is no need to calculate all the similarities $\text{BGM}(1, b_{k,i}), \dots, \text{BGM}(K, b_{k,i})$ for each topic word: as soon as BGI topics have a better rating than topic k , the topic word can be dropped.

4 Topic Word Selection

4.1 Experimental Set-Up

In the following, we describe an experimental set-up to show the applicability of the approach as a topic word selection approach.

We are using two data sets to evaluate the performance of our approach: The first data set is a technology dataset publicly available at webhose.io. It comprises 22,292 news and blog texts from September 2015 on technology scraped from the web in

that month.¹ The second data set consists of 2669 abstracts of scientific articles on autonomous vehicles² from the years 2017 and 2018.

We execute LDA³ with the entries of the hyperparameter vectors α and β set to 0.5 and 0.01, respectively, the values used by Rosner et al. [9]. The topic parameter K is set to 50. We calculate the results for the BGI parameter set to 1 and 2. $W = 10$ topic words are considered for the BGI approach and the C_V measure; for the latter, we use $\epsilon = 10^{-12}$ as suggested by Röder et al. [2].

For $BGI = 1$, our approach selects 15 topics for the technology data set and 18 topics for the autonomous vehicles data set. In the following, we will present the top-10 results for each data set. The rank rk_B of each topic is determined by the number of words selected by the BGI algorithm; in case of ties we are using mid-ranks.

The C_V measure is the one most similar to our approach, because it also uses information from the documents associated with each topic word, and it accounts for indirect confirmation between the topic words. In our results tables, we therefore provide the values and ranks based on the C_V measure for comparison.

4.2 Topic Word Selection Results

Table 1 shows the results for the webhose technology data set. Words in bold (in italics) were chosen with the BGI parameter set to 1 (2).

As intended, our approach tends to pick words specific for one topic (such as “gene” and “dna” in the seventh topic) over more general terms (“paper”, “science”). Terms that are only slightly connected to a topic are rejected (e.g., “year” and “expect” for the eighth topic). This is not the case for all the 10 highest-ranking topics according to C_V . At the bottom of the table, we show the topics ranked 6–8 by C_V . While these topics have a relatively high word coherence, they comprise many general terms.

From the results it also becomes clear that named entities are often selected by the approach. For example, “Kumari” and “Biswas” are the last names of scientists who authored papers on secure authentication, which are cited in the news articles. Other examples include “medigus”, “nasa”, and “apple”. The named entities in the ninth topic point to the 2012 shooting of the imam Obid-kori Nazarov in Sweden. Obviously, this topic has been formed based on non-technology-related articles contained in the document corpus, and the BGI algorithm has correctly identified the topic words as topic-specific.

Table 2 shows the outcomes for the autonomous vehicles data set.

Again, these results confirm that the words are selected based on how specific they are for each topic, in relation to the other topics. The term “time” is rather

¹Publicly available at <https://www.webhose.io/datasets>; retrieved on August 27th, 2018.

²All the paper data sets were downloaded from Scopus API between January 4th, 2019 and February 4th, 2019 via <http://www.api.elsevier.com> and <http://www.scopus.com>.

³In all our experiments we used common preprocessing techniques such as stemming and custom stopword lists to enhance the topic quality.

Table 1 Topic selection results for the technology data set; the first column rk_B shows the ranking according to the BGI approach and the last column rk_{C_V} shows the ranking according to the C_V measure; topic words in bold (in italics) were selected with the BGI parameter set to 1 (2)

rk_B	Topic Words, BGI = 1 , <i>BGI = 2</i> , BGI > 2	C_V	rk_{C_V}
1	'cloud' , 'manag' , 'custom' , 'solut' , 'softwar' , 'platform' , 'servic' , 'data' , 'busi' , 'technolog'	0.647	10
2	'nasa' , 'earth' , 'ocean' , 'sea' , 'temperatur' , 'satellit' , 'water' , <i>'imag'</i> , <i>'space'</i> , <i>'smartwatch'</i>	0.602	13
3	'oneplu' , 'redmi' , 'processor' , 'batteri' , 'display' , 'camera' , 'samsung' , <i>'smartphon'</i> , <i>'note'</i> , <i>'phone'</i>	0.741	4
4.5	'kumari' , 'biswa' , 'cryptanalysisi' , 'telecar' , 'syst' , 'crossref' , <i>'authent'</i> , <i>'agreement'</i> , <i>'medicin'</i> , <i>'doi'</i>	0.926	1
4.5	'uzbek' , 'teacher' , 'indian' , 'friday' , 'student' , 'school' , <i>'distribut'</i> , <i>'technolog'</i> , <i>'industry'</i> , <i>'presid'</i>	0.660	9
6	'physician' , 'clinic' , 'patient' , 'surgic' , 'medic' , <i>'health'</i> , <i>'care'</i> , <i>'healthcar'</i> , <i>'procedur'</i> , <i>'announc'</i>	0.813	2
7	'gene' , 'dna' , 'scientist' , 'cell' , 'brain' , <i>'professor'</i> , <i>'paper'</i> , <i>'univers'</i> , <i>'science'</i> , <i>'human'</i>	0.754	3
8	'siri' , 'iphon' , 'appl' , 'pro' , <i>'box'</i> , <i>'watch'</i> , <i>'app'</i> , <i>'ipad'</i> , <i>'year'</i> , <i>'expect'</i>	0.695	5
9	'kupaisinov' , 'tashkent' , 'zhukovski' , 'nazarov' , <i>'amin'</i> , <i>'suspect'</i> , <i>'attack'</i> , <i>'imam'</i> , <i>'told'</i> , <i>'polic'</i>	0.504	33
10	'forwardlook' , 'medigu' , <i>'statement'</i> , <i>'futur'</i> , <i>'risk'</i> , <i>'includ'</i> , <i>'code'</i> , <i>'secur'</i> , <i>'releas'</i> , <i>'compani'</i>	0.504	32
–	<i>'famili'</i> , <i>'children'</i> , <i>'women'</i> , <i>'parent'</i> , <i>'age'</i> , <i>'child'</i> , <i>'tree'</i> , <i>'kid'</i> , <i>'young'</i> , <i>'men'</i>	.690	6
–	<i>'investor'</i> , <i>'bank'</i> , <i>'capit'</i> , <i>'compani'</i> , <i>'invest'</i> , <i>'busi'</i> , <i>'billion'</i> , <i>'fund'</i> , <i>'firm'</i> , <i>'financi'</i>	0.662	7
–	<i>'million'</i> , <i>'year'</i> , <i>'quarter'</i> , <i>'revenu'</i> , <i>'loss'</i> , <i>'end'</i> , <i>'total'</i> , <i>'compar'</i> , <i>'net'</i> , <i>'share'</i>	0.661	8

unspecific for the sixth topic about driver assistance systems or the ninth topic about lane changing; however, it is much more closely related to the third topic dealing with public transport (e.g., the departure times and the travel times experienced).

“Thing”, a highly unspecific term by definition, was nevertheless chosen for the second topic, because in this data set the semantic context of the word is deeply connected with the collective term “internet of things”.

Topic ten is concerned with traffic flow simulations. Microscopic traffic flow simulations are models used for simulating the traffic dynamics of a single vehicle.

Table 2 Topic selection results for the autonomous vehicles data set; the first column rk_B shows the ranking according to the BGI approach and the last column rk_{C_V} shows the ranking according to the C_V measure; topic words in bold (in italics) were selected with the BGI parameter set to 1 (2)

rk_B	Topic words, BGI = 1 , <i>BGI = 2</i> , <i>BGI > 2</i>	C_V	rk_{C_V}
2	'camera' , 'detect' , 'map' , 'imag' , 'vision' , 'video' , 'match' , 'visual' , 'point' , 'target'	0.666	6
2	'busi' , 'innov' , 'world' , 'internet' , 'iot' , 'digit' , 'technolog' , 'thing' , 'emerg' , 'creat'	0.725	3
2	'travel' , 'citi' , 'public' , 'transport' , 'servic' , 'urban' , 'user' , 'choic' , 'demand' , 'time'	0.695	4
4	'nonlinear' , 'program' , 'constraint' , 'solv' , 'comput' , 'dynam' , 'set' , 'uncertainty' , 'solut' , <i>'time'</i>	0.620	7
5	'energi' , 'emiss' , 'electr' , 'fuel' , 'batteri' , 'pollut' , 'power' , 'hybrid' , <i>'air'</i> , <i>'reduct'</i>	0.818	2
6	'drive' , 'autom' , 'driver' , 'task' , 'situat' , 'assist' , 'steer' , 'condit' , <i>'level'</i> , <i>'time'</i>	0.689	5
7.5	'neural' , 'deep' , 'recognit' , 'learn' , 'machin' , 'train' , 'artifici' , <i>'intellig'</i> , <i>'task'</i> , <i>'network'</i>	0.830	1
7.5	'navig' , 'filter' , 'posit' , 'error' , 'path' , 'updat' , 'plan' <i>'dynam'</i> , <i>'autonom'</i> , <i>'environ'</i>	0.506	13
9.5	'penetr' , 'lane' , 'connect' , 'rate' , 'vehicl' , <i>'autom'</i> , <i>'mix'</i> , <i>'time'</i> , <i>'level'</i> , <i>'commun'</i>	0.529	9
9.5	'flow' , 'microscop' , 'traffic' , 'fundament' , <i>'highway'</i> , <i>'congest'</i> , <i>'desir'</i> , <i>'condit'</i> , <i>'evid'</i> , <i>'assign'</i>	0.501	15
13	'stabil' , 'theoret' , 'paramet' , <i>'trajectori'</i> , <i>'scheme'</i> , <i>'platoon'</i> , <i>'refer'</i> , <i>'dynam'</i> , <i>'numer'</i> , <i>'robust'</i>	0.553	8
14	'fatal' , 'crash' , 'accid' , <i>'prevent'</i> , <i>'road'</i> , <i>'safeti'</i> , <i>'option'</i> , <i>'motor'</i> , <i>'rise'</i> , <i>'number'</i>	0.527	10

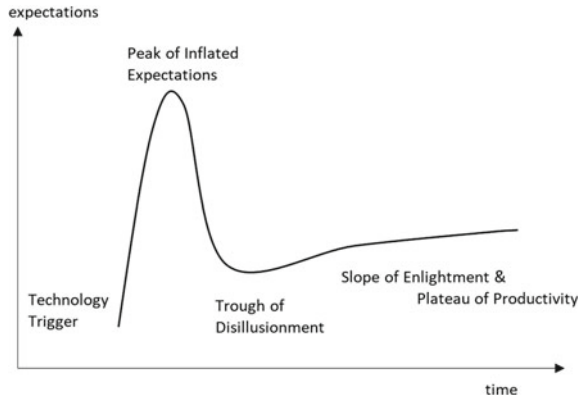
5 Weak Signal Detection

5.1 Terminology

In a second experiment, we show how to employ the proposed approach for finding weak signals. Weak signal detection is thereby located in the subject area of corporate foresight, where topic models have already been shown to be a useful tool to find new innovation opportunities [12].

A weak signal is defined as a new event with the capability of having an impact on future events [13]. This makes weak signals hard to pinpoint, because their status as a weak signal does not depend on whether or not they will *actually* have an impact later on, but rather on their *potential ability* to do so. In other words, a weak signal is an early indicator of change.

Fig. 1 Gartner hype cycle
[16]



This differentiates weak signals from trends; Saritas and Smith [14] define an (emerging) trend as follows: “Possible new trends grow from innovations, projects, beliefs or actions that have the potential to grow and eventually go mainstream in the future”. Therefore, a weak signal can be seen as the possible root cause for a potential upcoming trend. A summary of papers on weak signal and trend detection is available in the systematic literature review by Mühlroth and Grottko [15].

The Gartner hype cycle presents the maturity of technologies in a graph that displays the expectations over time (see Fig. 1), and divides the development of emerging technologies into five phases: The first stage is the “technology trigger”, where a technological breakthrough or a similar event draws a lot of attention. In the next phase, called the “peak of inflated expectations”, excessive expectations are created about the new technology, that it ultimately cannot meet. These high expectations therefore sooner or later result in a “trough of disillusionment”, which finally leads to phases of solid productivity and reasonable expectations (“slope of enlightenment”, “plateau of productivity”) [16].

A weak signal can be seen to occur at the earliest phase of the hype cycle, as being or announcing a technology trigger. Gartner hype cycles are published every year for technology trends in general but also for specific areas like 3D printing and artificial intelligence. Technology innovations at early stages, between the technology trigger event and the peak, are displayed as “on the rise”. Therefore, we consider a technology which appears on a hype cycle as “on the rise” and which was not contained in the hype cycle of the previous year as a sign of a newly-occurring weak signal.

5.2 Experimental Set-Up

In the experiment, we intend to retrace the weak signals on the gartner hype cycles on 3D printing from 1014, 2015 and 2016 using the BGI approach. To evaluate the

detection of weak signals, a data set on 3D printing and additive manufacturing was gathered.⁴ It consists of 15,403 documents from the years 2010 through 2018.

The idea is to find occurrences of new words in the data set; if a new word is also specific for a topic, then it might hint at a weak signal. Therefore we consider a word-topic combination of such kind as a potential weak signal.

For example, to detect small, new topics with our approach, the initial LDA can be run with the data from the current year while the word selection can be applied using the data from the previous three years. Hereby, we drop all those words and topics that the BGI approach does not pick for $BGI = 2$. Then $w_{k,min}$ is defined as the topic word associated with the smallest number of documents from the respective year.

In our experiment, we applied the algorithm using the documents available at the end of the years 2014, 2015, and 2016. For each year, the algorithm was run for $K = 100$ and $K = 150$. Similar to comparable literature, such as Chang et al. [1], we thus increased K in steps of 50; unlike Chang et al. we started with $K = 100$ (rather than $K = 50$), to catch smaller topics. Due to the stochastic nature of the LDA, we calculated each model twice; i.e., the algorithm was run four times for each year. The LDA parameter settings were the same as in Sect. 4.

The topics were sorted in descending order by the number of documents connected with the respective word $w_{k,min}$. The top-10 topics of each algorithm run were considered as potential weak signals, and we compared them to topics that are newly marked as “on the rise” in the Gartner hype cycles 2015 and 2016 on 3D printing compared with the first Gartner hype cycle on 3D printing from 2014.⁵

To consider a potential weak signal to match a topic from the hype cycle, it was required that both the topic words and the documents associated with the selected words are connected with the hype cycle topic. The top-5 topic words were chosen according to the topic-word probabilities from the model, and the documents were chosen by using the selected word as a query on the documents of that respective year; to discard documents that are only loosely connected with the words (e.g., as part of an enumeration or example), we only count documents that have at least two occurrences of the respective word as truly associated with the word and the topic.

Because of the topic word selection, small topics mainly consist of only a few but specific words. Thus sorting out the topics manually can be done much faster than with regular topics.

⁴All the paper data sets were downloaded from Scopus API between January 4th, 2019 and February 4th, 2019 via <http://www.api.elsevier.com> and <http://www.scopus.com>.

⁵www.gartner.com/doc/2803426/hype-cycle-d-printing; www.gartner.com/doc/3100228/hype-cycle-d-printing; www.gartner.com/doc/3383717/hype-cycle-d-printing; viewed on February 15th, 2019.

5.3 Weak Signal Detection Results

The first occurrences of the hype cycle topics that were found are shown in Table 3. We found traces of 9 of the 12 topics that newly appeared in the hype cycles of 2015 and 2016.

The approach detected the bioprinting of organs in 2014 and 3D-printing-aided hip and knee implants in 2015, and first traces on the topic “3D printed drugs” could already be found in 2014. Also the hype cycle topic on Workflow Software of 2016 was yet traced in 2014, with the abbreviation “CAD” translating to “computer-aided design”.

The topic about “wearables” was not detected per se, but it was possible to retrace early signs of research on the connection of 3D printing and biosensors in 2014.

The three Gartner hype cycle topics that have not been detected are nanoscale 3D printing (2016), 4D printing (2016), and sheet lamination (2016).

To show that the approach can detect the topics early in the data, the history of the words connected to the weak signals from Table 3 is provided in Table 4.

In 2015, when the topics “Consumable Products” and “Knee Implants” were detected by the approach, there were only 13 documents associated with the word “food” and merely 11 documents associated with the word “knee”.

Table 3 Results for the weak signal detection approach; the table shows when the weak signal was first found, and the respective topic words selected by the BGI approach as well as the top-5 topic words

Gartner hype cycle topic	First found	Selected words	Top-5 topic words
Consumable Products, 2015	2015	‘food’	‘food’, ‘product’, ‘manufactur’, ‘technolog’, ‘chain’
Knee Implants, 2015	2015	‘knee’	‘knee’, ‘nativ’, ‘prostesi’, ‘asc’ (stem cells), ‘nasal’
Hip Implants, 2015	2015	‘hip’	‘cartilag’, ‘graft’, ‘scaffold’, ‘week’, ‘rat’
BP for Life Science R&D, 2015	2015	‘market’	‘market’, ‘life’, ‘science’, ‘emerg’, ‘sector’
Organ Transplants, 2015	2014	‘organ’	‘organ’, ‘engin’, ‘tissu’, ‘scaffold’, ‘biolog’
3D-Printed Drugs, 2016	2014	‘tablet’	‘drug’, ‘release’, ‘innov’, ‘form’, ‘deliveri’
Powder Bed Fusion, 2016	2015	‘bed’	‘melt’, ‘thermal’, ‘life’, ‘cycl’, ‘compon’
Workflow Software, 2016	2015	‘software’, ‘cad’	‘comput’, ‘cad’, ‘aid’, ‘subject’, ‘geometr’
Wearables, 2016	2014	‘sensor’	‘vascular’, ‘sensor’, ‘descript’, ‘uniqu’, ‘hardwar’

Table 4 Document history for the selected words; the table shows the absolute frequencies of documents that contain the respective word at least twice (bold type indicates when the topic was first found); the last row shows the total sum of documents in the corpus for the respective year

Selected words	2010	2011	2012	2013	2014	2015	2016	2017	2018
Food	0	1	1	2	0	3	16	24	38
Knee	0	6	0	4	2	6	7	21	12
Hip	0	2	3	0	0	11	14	22	31
Market	1	0	1	6	6	15	19	14	23
Organ	3	9	3	12	18	31	51	65	90
Tablet	1	0	0	0	2	7	5	14	22
Drug	4	7	12	13	17	27	37	71	136
Bed	1	2	1	2	5	12	17	43	60
Cad	3	4	6	9	20	26	31	20	48
Sensor	4	2	3	5	10	41	61	88	128
Σ documents	208	299	350	535	913	1558	2474	3784	5282

We also added “drug” to the table as additional information, since “tablet” was the selected word for first occurrence of this weak signal in 2014 but it also appeared in 2015 as a weak signal with “drug” as a selected word. Furthermore, the word “drug” might be more representative for the topic “3D-Printed Drugs”.

The table also shows the total sum of documents in the corpus for each year. This number noticeably increases over the years, which is explicable due to the increasing interest in 3D printing in general but also with the general increase in published papers per year. This should be considered before interpreting increases in publications on a certain topic.

6 Conclusion

In this paper, a new approach was proposed for finding specific topics by using context information from the topic words.

The results for the topic word selection demonstrate that the topics become more readable and on point. Another advantage of our approach over the benchmark is that it automatically selects the topics: there is no need to set a topics threshold. Although it is still necessary to set the number of topic words as a parameter, the BGI approach is less prone to influences of topic cardinality on the topic quality estimation [17] due to the word selection.

We have also shown that we can find new topics at very early stages of their appearance in the corpus. The results indicate that the approach might be able to detect weak signals for corporate foresight very early when applied to more extensive data sets.

Acknowledgements This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the Program Concept Innovations for “Tomorrows Production, Services, and Work” (02K16C190) and managed by the Project Management Agency Forschungszentrum Karlsruhe (PTKA). The authors are responsible for the contents of this publication.

References

1. Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: *Advances in neural information processing systems*, vol 22. Vancouver, Canada, pp 288–296
2. Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on web search and data mining*. Shanghai, China, pp 399–408
3. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
4. Lau, JH, Newman D, Karimi S, Baldwin T (2010) Best Topic Word Selection for Topic Labelling. In: *Proceedings of the 23rd international conference on computational linguistics*. Beijing, China, pp 605–613
5. He D, Wang M, Khattak AM, Zhang L, Gao W (2019) Automatic labeling of topic models using graph-based ranking. *IEEE Access* 7:131593–131608
6. Alokaili A, Aletras N, Stevenson M (2019) Re-ranking words to improve interpretability of automatically generated topics. In: *Proceedings of the 13th international conference on computational semantics—long Papers*. Gothenburg, Sweden, pp 43–54
7. Aletras N, Stevenson M (2013) Evaluating topic coherence using distributional semantics. In: *Proceedings of the 10th international conference on computational semantics*. Potsdam, Germany, pp 13–22
8. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the conference on empirical methods in natural language processing*. Edinburgh, United Kingdom, pp 262–272
9. Rosner F, Hinneburg A, Röder M, Nettling M, Both A (2014) Evaluating topic coherence measures. In: *Computing research repository (CoRR)*, pp 1–4. [arXiv:1403.6397](https://arxiv.org/abs/1403.6397)
10. Korencic D, Ristov S, Snajder J (2018) Document-based topic coherence measures for news media text. *Expert Syst Appl* 114:357–373
11. AlSumait L, Barbara D, Gentle J, Domeniconi C (2009) Topic Significance Ranking of LDA Generative Models. In: *Proceedings of the European conference on machine learning and knowledge discovery in databases*. Bled, Slovenia, pp 67–82
12. Kölbl L, Mühlroth C, Wiser F, Grottk M, Durst C (2019) Big data im Innovationsmanagement: Wie machine learning die Suche nach Trends und Technologien revolutioniert. *HMD Praxis der Wirtschaftsinformatik* 56(5):900–913
13. Thorleuchter D, Scheja T, Van den Poel D (2014) Semantic weak signal tracing. In: *Expert systems with applications* 41(11):5009–5016
14. Saritas O, Smith JE (2011) The big picture-trends, drivers, wild cards, discontinuities and weak signals. *Futures* 43(3):292–312
15. Mühlroth C, Grottk M (2018) A systematic literature review of mining weak signals and trends for corporate foresight. *J Bus Econ* 88(5):643–687
16. Lajoie EW, Bridges L (2014) Innovation decisions: using the Gartner hype cycle. *Libr Leadersh Manage* 28(4)
17. Lau JH, Baldwin T (2016) The sensitivity of topic coherence evaluation to topic cardinality. In: *Proceedings of the North American chapter of the association for computational linguistics: human language technologies*. San Diego, USA, pp 483–487

Laura Kölbl has a M.Sc. in statistics from the Ludwig-Maximilians-University Munich and is currently working at the chair of statistics and economics at the Friedrich-Alexander University Erlangen-Nuremberg. As part of a research project funded by the German Federal Ministry of Education and Research (BMBF), her research focuses on NLP and machine learning methods with a special interest for possible applications in the automatic detection of trends.

Michael Grottke received an M.A. in economics from Wayne State University, USA, as well as a Diploma degree in business administration, a Ph.D., and a Habilitation degree from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany. He is now Principal Data Scientist at GfK SE, Nürnberg, Germany and Adjunct Professor at FAU. Before, he spent three years as a Research Associate and Assistant Research Professor in the Department of Electrical and Computer Engineering at Duke University, USA. His research focuses on stochastic modeling, statistical data analysis and machine learning for topics in information systems, computer science, and business administration, including crowdsourcing, social networks, software quality, and the automatic detection of upcoming technological trends. His research work has been funded by national and international organizations, such as the European Commission, the German Ministry of Education and Research, and NASA's Office of Safety and Mission Assurance. He has published more than 40 papers in international journals and peer-reviewed conference proceedings. He has acted as a Guest Editor for Special Editions of Performance Evaluation and the Journal of Systems and Software, and he is an Associate Editor of the International Journal of Performability Engineering and the IEEE Transactions on Reliability. He is a member of the IEEE, the German Statistical Society, and ESOMAR.