

**MAGKS**



**Joint Discussion Paper  
Series in Economics**

by the Universities of  
**Aachen · Gießen · Göttingen  
Kassel · Marburg · Siegen**

ISSN 1867-3678

**No. 01-2024**

**Albina Latifi, David Lenz, and Peter Winker**

**Identification of innovation drivers based on technology-  
related news articles**

This paper can be downloaded from

[https://www.uni-marburg.de/en/fb02/research-  
groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics](https://www.uni-marburg.de/en/fb02/research-groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics)

Coordination: Bernd Hayo • Philipps-University Marburg  
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg  
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: [hayo@wiwi.uni-marburg.de](mailto:hayo@wiwi.uni-marburg.de)

# Identification of innovation drivers based on technology-related news articles\*

Albina Latifi      David Lenz      Peter Winker

January 17, 2024

## Abstract

Innovations contribute to economic growth. Hence, knowledge about drivers of innovation activities is a necessary input for economic policy making when it comes to implement targeted support measures. We focus on firms as potential drivers of innovation and use a novel data-driven approach to identify them. The approach is based on news articles from a technology-related newspaper for the period 1996–2021. In a first step, natural language processing (NLP) tools are used to identify latent topics in the text corpus. Expert knowledge is used to tag innovation-related topics. In a second step, a named entity recognition (NER) method is used to detect firm names in the news articles. Combining the information about innovation-related topics and firms mentioned in news articles linked to these topics provides a set of firms linked to each innovation-related topic. The results suggest that the approach helps identifying drivers of innovation activities going beyond the usual suspects. However, given that the rate of false alarms is not negligible, at the end also human judgement is needed when using this approach.

*Key Words: Innovation drivers, topic modeling, entity recognition*

*JEL classification: C49, C55, O30*

---

\*Financial support by the German Federal Ministry of Education and Research (16IFI001 and 16IFI002) is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

Affiliation: Department of Economics, Justus Liebig University Giessen, Email:  
albina.latifi@wirtschaft.uni-giessen.de,      david.lenz@wirtschaft.uni-giessen.de,  
peter.winker@wirtschaft.uni-giessen.de

# 1 Introduction

Innovative companies contribute to economic growth by introducing new products and services, opening up new markets and creating jobs (Carree and Thurik, 2010). Hence, economic policy making requires knowledge about drivers of innovation activities to implement targeted support measures. Existing methods for obtaining such information include surveys,<sup>1</sup> analysis of patents and aggregate measures on R&D expenditures. However, these methods exhibit substantial limitations as a base for economic policy making. These drawbacks include long publication lags for surveys and patent information, and a lack of information about the domain of innovation activities for surveys and aggregate data.

To overcome some of the shortcomings of established methods, we propose a data-driven approach to understand who the drivers of innovation are, i.e. which companies are engaged in particular fields of innovation. The approach has the advantage of providing timely information at the firm level. To this end, we proceed as follows: Our analysis is based on a large set of scraped technology related news articles.<sup>2</sup> These textual data are pre-processed so that we are able to apply methods from NLP (natural language processing) such as topic modelling and NER (named entity recognition).<sup>3</sup>

Topic modelling is a widely used method to uncover latent topics from text corpora. Recent applications in the context of innovation include, e.g., Mühlroth and Grottke (2020), who apply topic modelling on documents from the AAI to detect emerging technologies and innovations, the use of STM topic modelling by Dwivedi et al. (2023) to study the evolution of research output on artificial intelligence in the Journal of Technological Forecasting & Social Change, or the application of LDA topic modelling of patent text data by Savin et al. (2022) for tracing the evolution of service robotics.

NER stands for a class of methods allowing to identify proper names in documents, which has found a broad variety of applications in the literature. For example, in the chemical industry, chemical entities such as molecule

---

<sup>1</sup>The European Commission runs a biennial survey at the enterprise level in the EU, EFTA and the candidate countries. Details can be found at <https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey> for details.

<sup>2</sup>For the empirical application presented here, we use the period 1996 – 2021, which might be extended easily to include more recent news articles.

<sup>3</sup>A related approach focusing on potential innovators in the pharmaceutical industry has been proposed by Chen et al. (2021), who apply BERT for text classification and named entity recognition based on 1.9 million news articles between 2013 and 2018 in the domain of pharmaceutical industry. The authors use BERT to first classify newspaper articles that may mention innovation and then NER to identify company names in newspaper articles that may have mentioned innovation.

names are extracted (Eltyeb and Salim, 2014). Leitner et al. (2019) identify various entities from legal documents such as persons, regulations, and ordinances. Passonneau et al. (2015) use NER to recognise statements referring to specific companies in financial press releases. Latifi (2023) trains a NER-model with a customized entity to identify the beginning of speeches within the stenographic protocols of the German Bundestag.

The output of fitting a topic model are, on the one hand, topic-term-distributions, which can be interpreted as latent topics in a corpus, and, on the other hand, topic-document-weights, which indicate the importance of a specific topic within a document (news article). Often, the topic-term-distributions are presented in form of word clouds. We use this presentation to identify topics linked to fields of innovation with the help of experts.

For selected fields of innovation, we are interested in finding innovation drivers. Therefore, we link the topic-document-weights to the entities that appear in the news articles. Thereby, an entity stands for a proper name classified as an organization by a NER-model. Ideally, these entities all represent companies. All steps of the procedure are summarized in Figure 1.

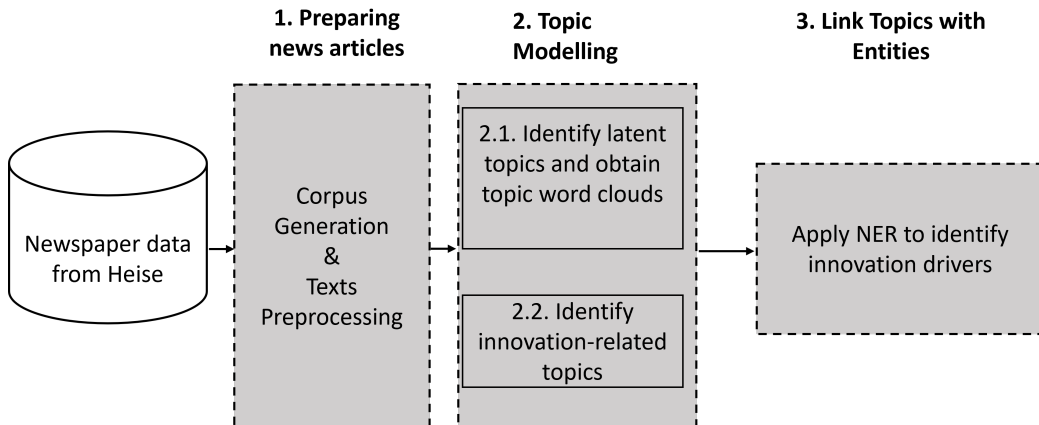


Figure 1: Depiction of the Pipeline for Identification of Innovation Drivers.

The remainder of this paper is structured as follows. Section 2 provides details on the data set used for the analysis and the pre-processing required for the application of topic modelling. Topic modelling methods and the identification of topics related to innovation are described in Section 3. Details of the approach used for linking innovation related topics and specific firms are presented in Section 4 as well as some examples. Section 5 summarizes the main findings and points to possible extensions and applications of the approach.

## 2 Corpus Generation and Data Set Description

In order to make statements about innovation drivers, we analyse newspaper articles from the archive of the German IT news publisher Heise Medien. We developed a web scraper to download automatically texts from heise online.<sup>4</sup> Thereby, we also scraped metadata, e.g., date of publication. When scraping the text data, we restricted ourselves to articles from the categories “MIT Technology Review”, “c’t Magazin” and “newsticker”, as we wanted to achieve as general a coverage of companies (and thus possible drivers of innovation) as possible. For example, including additionally the category “Mac & i”, would result in a too large impact of the technology company “Apple”. All articles in the above categories were scraped from the start of the archive in 1996 up to 18.10.2021, resulting in a text corpus of 190,722 articles. Of these, 5,991 articles fall into the category “MIT Technology Review”, 4,362 articles into the category “c’t Magazin” and 180,369 articles into the “newsticker”.

To clarify the terminology used in this paper, we refer to a (pre-processed) article as a document and to the set of all documents as corpus. Before the corpus can serve as input for topic modelling, a number of text pre-processing steps are carried out. First, we added the headline to the text of the newspaper article, since it can be assumed that the headline contains meaningful information about the content of the text. Second, in order to exclude encoding errors of the newspaper articles, we converted all German umlauts and the sharp S to “ae”, “oe”, “ue” and “ss”, respectively. Afterwards, we converted all letters to lower case and removed line breaks, numbers, spaces, punctuation marks and special characters. The next step consisted in tokenising the texts. Thereby, in our application, a token stands for a word, since the topic modelling is performed at the word-level. The set of unique tokens is called vocabulary. In addition, all tokens consisting of only one letter were excluded. Tokens with two letters were not removed because in this domain-specific corpus many meaningful tokens consist of two letters, such as: “ki”, “vr”, “it”, “tv”, “pc”.

After these pre-processing steps, a document contains on average 305.40 tokens and at the median 254 tokens, with the longest and shortest document containing 6,216 and 28 tokens, respectively. The standard deviation of document length is 204.11. Since a very large variation in document length may hinder the robust estimation of topic models, very short documents are excluded from the corpus. To this end, the 5%-percentile of the distribution

---

<sup>4</sup><https://www.heise.de/newsticker/archiv/>

of document lengths was used, which corresponds to 100 tokens, i.e. all documents containing up to 100 tokens were removed from the corpus (see also [Tang et al. \(2014\)](#)).

After removing these very short documents, the corpus comprises 181,402 documents. Next, German stop words are removed from the documents from a predefined list provided in the *nlTK*-module ([Bird and Klein, 2009](#)). The predefined list contains 232 stop words and can be found in Appendix [A](#). Additionally, a custom list of bigrams is constructed using word pair frequencies and the inverse document frequency measure (idf-value). Further details on how this list is constructed can be found in Appendix [B](#). These bigrams are added to the list of tokens<sup>5</sup>.

In the final step prior to the topic modelling, the vocabulary, i.e. the set of different words used, is further reduced using popularity-based pre-filtering analogous to the approach in [Lenz and Winker \(2020\)](#), which allows an automatic removal of domain-specific stopwords. In this step, tokens are removed that occur in less than 0.05% or in more than 65% of all documents. Eventually, the corpus used for the further analysis consists of 181,402 documents with 4,681 unique tokens.

### 3 Topic Modelling and Labelling

In order to identify latent topics from the technology-related corpus that represent an innovation field, topic models are estimated and the resulting word clouds are labelled as “innovation-related” or “non-innovation-related” with the help of experts. For this purpose one of the most popular topic modelling methods, Latent Dirichlet Allocation (LDA), is applied, which was first described in [Blei et al. \(2003\)](#).

#### 3.1 Latent Dirichlet Allocation

The LDA model belongs to the class of probabilistic generative topic models. In the model, it is assumed that all documents in a corpus were generated from a random mixture of different latent topics, where the number of different topics  $K$  is assumed to be known. A topic, in turn, is defined as a probability distribution over the set of tokens in the vocabulary. Each document  $d$  contains all  $K$  topics. However, the documents differ in their weights for each of the topics. Thus, a document  $d$  can be considered as a probability distribution ( $\theta_d$ ) over topics and a topic  $k$  as a probability distribution ( $\beta_k$ )

---

<sup>5</sup>For the specific corpus, it was found that all remaining bigrams were removed in the final step of popularity-based-pre-filtering.

over the vocabulary. Since the topics are not known in advance, the aim of estimating an LDA model is to learn the topic weights  $\theta_d$  for each document and the word weights  $\beta_k$  for each topic from the data (see [Blei and Lafferty \(2009\)](#)).

In order to infer the probability distributions  $\beta_k$  and  $\theta_d$  from the text data using LDA, the corpus resulting from the pre-processing described in Section 2 is transformed into a document-term-matrix. In our application, this is a very large sparse matrix with 181,402 rows (corresponding to the number of documents) and 4,681 columns (corresponding to the size of the vocabulary), in which each cell contains the frequency of each unique token per document.<sup>6</sup>

In addition to the document-term-matrix, a further central input for estimating the LDA model is the assumption about the number of latent topics  $K$ . Choosing the optimal number of topics is still a major challenge in the topic modelling literature (see, e.g., [Campagnolo et al. \(2022\)](#), [Sbalchiero and Eder \(2020\)](#), and [Bystrov et al. \(2022a\)](#)). Since, after estimating the LDA model, the topic-word-distributions are to be interpreted by humans in the subsequent step and assessed with regard to the mapping of an innovation-related topic field, it is particularly important that the resulting topic model delivers topics that are as accessible as possible for human interpretation. In a study by [Röder et al. \(2015\)](#), a large number of coherence metrics are compared with each other. The authors conclude that the  $c_V$  coherence metric performs best with regard to human interpretability. The `gensim`-module ([Řehůřek and Sojka, 2010](#)) provides an implementation of this metric, which we use. To this end, we estimated 13 topic models with different numbers of topics<sup>7</sup> and evaluated them with respect to the  $c_V$  score. According to this metric, a higher  $c_V$  score speaks for a higher coherence of the occurring words in a topic and thus for a better interpretability for humans. Figure 2 visualises the  $c_V$  values for the 13 estimated topic models.

The scores show a first peak for  $K = 60$  corresponding to the highest value of the  $c_v$  score among the 13 candidates. However, after additional human judgement of the resulting word clouds, we decided to choose the model corresponding to the second peak for  $K = 120$ . When considering this choice, one has to keep in mind that the procedure for estimating the LDA model contains a stochastic component. Consequently, the  $c_V$  scores are not deterministic either and should only be understood as providing a reference for model selection. In fact, by choosing 120 topics, we achieve a

---

<sup>6</sup>The sparsity of the resulting document-term-matrix amounts to 98.02%, i.e. 98.02% of all entries in the matrix are equal to zero.

<sup>7</sup>Topic models with  $K = 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150$  and 200 topics have been estimated.

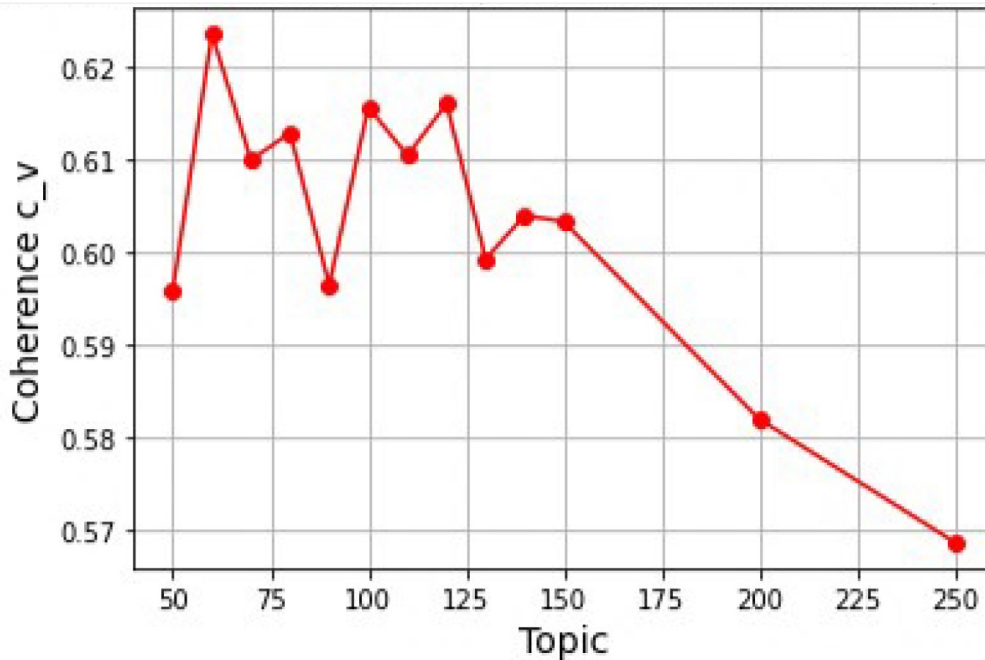


Figure 2:  $c_v$  Scores for various topic numbers

better differentiation of the topics as compared to the setting with  $K = 60$ , because we are more likely to obtain topics corresponding to a mix of different innovation themes with 60 topics than with 120 topics.

### 3.2 Robustness check with Paragraph Vector Topic Modelling

In order to check the robustness of the topics obtained when using the LDA model, we also estimate a Paragraph Vector Topic model (PVTM), which was already applied by [Lenz and Winker \(2020\)](#) on a similar corpus.

In contrast to LDA, PVTM first computes document vector representations for the documents resulting from the pre-processing steps described in Section [2](#) and embeds them in a lower dimensional vector space using Doc2Vec ([Le and Mikolov, 2014](#)). Then, the embedded documents are clustered into a pre-defined number of mixture components estimated using a Gaussian mixture model. For our application, the number of Gaussian mixture components is set to 120, which corresponds to the number of latent



topics previously determined for the LDA model. The dimensionality of the feature vectors (Doc2Vec), i.e. the dimension of the embedding space, is set to 100. The model is trained in 15 epochs. Apart from the topic number, we use all parameters for the PVTM as in [Lenz and Winker \(2020\)](#).

To compare the output of both models, we follow the procedure of [Bystrov et al. \(2022b\)](#), who compare the topic term distributions of LDA models trained on different corpora in order to identify similar topics. Obviously, we do not use this approach to find similar topics in different corpora, but to check the robustness of the learned topics on the same technology-specific corpus using different models. Thus, we compare the topic term distributions obtained from the LDA model ( $\beta_{LDA}$ ) with the one resulting from PVTM ( $\beta_{PVTM}$ ).

Since the original PVTM implementation by [Lenz and Winker \(2020\)](#) does not provide the topic term distributions  $\beta_{PVTM}$ , we first need to generate topic term probabilities for PVTM. This is done by multiplying the topic-document matrix, i.e.  $\theta_{PVTM}$ , by the document-term-matrix (which served as input for the LDA model estimated in Subsection [3.1](#)) and then normalising the resulting matrix so that all token probabilities within a topic sum up to one.

Finally, we compare the resulting topic-term-distributions across models using the cosine similarity measure. To this end, we assign to each LDA topic the PVTM topic with the highest cosine similarity, which is labeled by [Bystrov et al. \(2022b\)](#) as the “best-matching-approach”. The results show that similar topics to the one obtained by LDA can be found in the output of the PVTM model. Therefore, we assume that a robust estimation of the topics is achieved. Some examples of the matched topics can be found in Appendix [C](#).

### 3.3 Topic Labelling

Given our focus on innovation drivers, we want to focus on topics, which are related to innovations. In general, labelling of topics is a challenging task as it consists in assigning some meaning to high-dimensional vectors of term weights. For the present application, it is sufficient to identify innovation-related topics without having to assign specific labels to each of these topics. Our choice of the  $c_V$  coherence metric for selecting the number of topics was motivated by the expectation that the resulting topics allow for easy interpretation by humans. Furthermore, [Smith et al. \(2017\)](#) show that automatic topic labelling procedures are inferior to human evaluation. Therefore, we make use of human expertise for the task of labelling the 120 topics obtained by the LDA model as “innovation-related” / “non-innovation-related”.

The ultimate aim of conducting this assessment by experts is to identify innovation-related topics.

The labeling was done by eight experts with experience in topic modelling and innovation economics, who provided their assessment in two rounds of the DELPHI method (Dalkey and Helmer, 1963). We presented the 120 topics to the experts as word clouds generated from the LDA topic model. In addition, they received a labelling guideline of what should be understood by innovation, which is provided in Appendix D. This guideline is based on the OSLO manual (Commission et al., 2018), according to which innovation is defined as a “new or improved product or process” to the entity considered (e.g. company, organisation, person, etc.). However, since the word clouds do not allow following one to one such a classification, we use a broader definition of innovation to reflect product, process, platform, or technology innovation. Any topic corresponding to an innovation in the time interval under consideration should be classified as such.

Since there is no “ground-truth” that allows an unambiguous definition of what is specifically to be understood by the term “innovation”, it is to be expected that there will be divergent opinions when classifying the topics. For this reason, we evaluated the preliminary results after the first round of topic labelling by assigning a degree of innovation to each topic, which is given by the share of the eight experts labelling the topic as “innovation-related”. In the second round, we provided the experts again with the word clouds as well as with the preliminary degree of innovation of the individual topics, so that each expert could reconsider his or her opinion. For further details and results of the approach for determining the degree of innovation of a topic, see Appendix E. The results of this second round are again summarized by the share of experts labelling the topic as “innovation-related”. Finally, we consider topics as innovation-related for our application if this share is above 50%.<sup>8</sup>

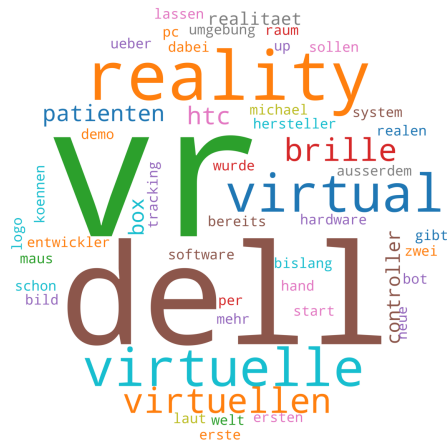
Figure 3 shows selected innovation-related topics. The selected topics are assigned a high degree of innovation according to the described approach of annotation by experts (100% or 87.5%, respectively, i.e. at least seven out of the eight experts labelled them as “innovation-related”). The figure is generated from the topic-term-distributions  $\beta_k$  of these topics. The word clouds contain the 50 most important tokens of a topic, where the font size reflects the probability that the respective token shows up in documents related to the topic. Thus, the larger the font size of the token is, the more

---

<sup>8</sup>An alternative approach would use weights corresponding to the share of experts for the further steps of the analysis. Preliminary analysis indicates that the results from using such a weighted measure would not differ qualitatively.

important the token is for the specific topic.

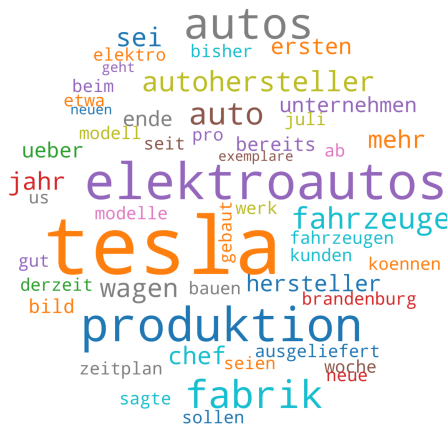
Topic #43



Topic #47



Topic #67



Topic #90

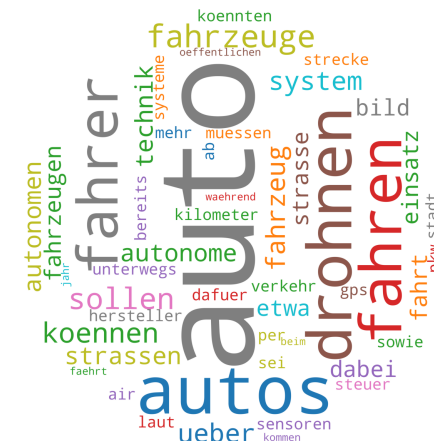


Figure 3: Word Clouds for some Topics related to Fields of Innovation: #43: “Virtual Reality”, #47: “Online Trading”, #67: “E-Mobility” and #90: “Autonomous Vehicles”

In the next step, we will focus on identifying innovation drivers in two of the four selected innovation fields visualized in Figure 3.

## 4 Link Topics to Entities

In this section, some innovation-related topics are examined in more detail with regard to the predominant entities, i.e. potential innovation drivers. As mentioned before, innovation-related topics are defined as those topics that the majority of the experts (i.e. at least 5 out of 8) have labelled as such in the second round of the DELPHI method.

Our approach comprises three steps. In the first step, we focus on identifying companies mentioned in the documents using a named entity recognition-model (NER) with a focus on company names. In a second step we use an embed-and-match approach to match the identified entities with company names from the ORBIS database. In the third step we link the companies to the topics and vice versa. Analyzing the companies linked to the topics provides information about the focus of innovation activities of those companies, while linking topics to companies provides a list of potential innovation drivers for selected innovation fields. The procedure will be explained in more detail below and is visualized in Figure 4.

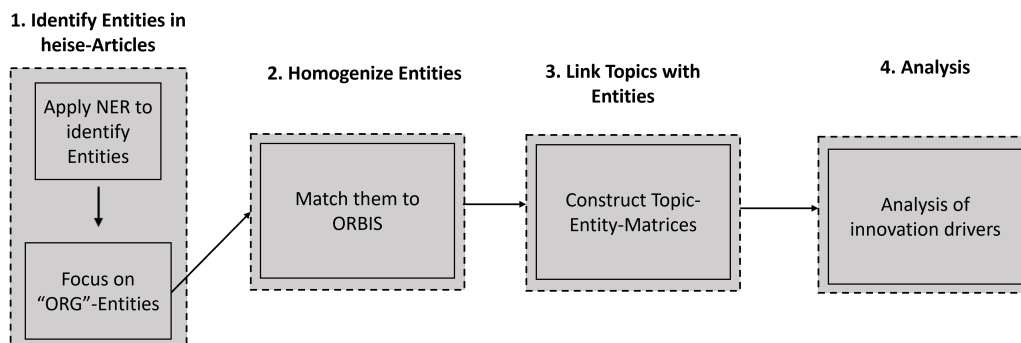


Figure 4: Visualization of the Entity Analysis Pipeline

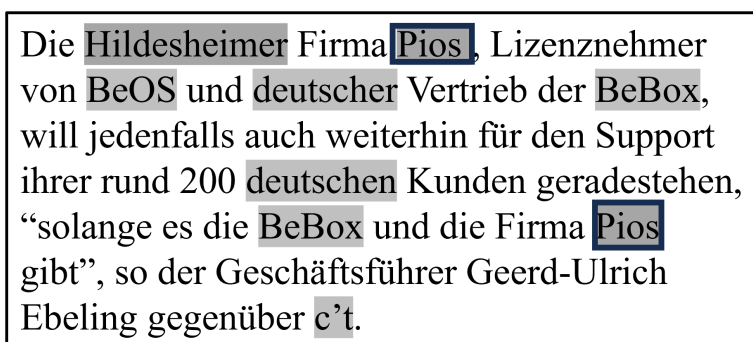
### 4.1 Identification and Homogenization of Entities

We use a standard NER model from the `spaCy`-toolkit (Montani et al., 2023), namely “`de_core_news_lg`” to identify company names in the corpus of 181,402 news articles. The model was selected as it was trained on data including, amongst others, German news corpora. Hence, it appears to be well suited for the present application. The NER model automatically identifies entities for the categories person names (“PERSON”), company names (“ORG”) and place names (“LOC”). Other proper names, which cannot

be assigned to any of the three categories, are labelled as miscellaneous (“MISC”).

Figure 5 provides an example of entity identification using a short news article from our corpus. The identified entities are highlighted. Apart from the company name “Pios” (correctly identified as “ORG”) or the place name “Hildesheimer” (correctly identified as “LOC”) further entities corresponding to product names (“BeBox”) or nationalities (“deutschen”) were assigned to the “MISC” category. For the further analysis, we focus solely on entities assigned to the category “ORG”, which are additionally framed in Figure 5.

Imposing the restriction to focus solely on “ORG” entities might lead to missing some companies, i.e. actual companies that the NER model falsely does not classify as “ORG”. We found a few examples of this case when manually checking a small random subset of companies. Furthermore, NER might classify some entities as “ORG”, which do not represent companies. Therefore, one has to keep in mind that this filtering might generate both false-positives and false-negatives. In the corpus, a total of 1,246,524 “ORG” entities were identified in 174,817 heise news articles (i.e. only 6,585 documents do not contain any “ORG” entity). The number of unique entities is 272,888.



Die Hildesheimer Firma **Pios**, Lizenznehmer von **BeOS** und **deutscher** Vertrieb der **BeBox**, will jedenfalls auch weiterhin für den Support ihrer rund 200 **deutschen** Kunden geradestehen, “solange es die **BeBox** und die Firma **Pios** gibt”, so der Geschäftsführer Geerd-Ulrich Ebeling gegenüber **c't**.

Figure 5: Example of identified entities (highlighted) in a short news article, where “ORG” identities are framed. For an English translation of the example see Figure 13 in the Appendix F.

In the next step, we link the identified entities to companies included in the ORBIS database. The most simplistic but also very restrictive approach would be to just allow perfect matches. However, we considered this approach to be too limited, since the entity names extracted from the news corpus are mostly not in the standardized format corresponding to the company names in the database. Consequently, we would lose a huge part of the observations.

Therefore, as a more robust alternative, we first create an embedding for each entity from our corpus and each company name included in the ORBIS-Database using the “all-mpnet-base-v2” model from the Sentence Transformer library, which has been trained on a large and diverse dataset of over 1 billion training pairs (Reimers and Gurevych, 2019, 2020). Thereby, we used the ORBIS database from June 2023<sup>9</sup>. Specifically we use a subset of ORBIS consisting of 1.786.878 economically active companies from Germany, Switzerland and Austria. While the database includes further more information on companies, e.g. age, address, size, we only employ the name of the companies for our analysis. Comparing each of the 1.786.878 ORBIS firms with all of the 272, 888 HEISE entities results in 487, 617, 563, 664 pairwise comparisons.

Ideally, the models in the Sentence Transformers library are applied on sentences or paragraph level length texts. However, preliminary tests showed that the model is still able to capture the meaning of the entities quite frequently, especially when the entities are rather descriptive: As an example consider “National Institute for Standards and Technology”. The model sometimes struggles with acronyms and abbreviations though, e.g. “NIST” might produce a rather insensible embedding for comparison with the embedding of “National Institute for Standards and Technology”. Overall, we found that the embedding model was mostly able to accurately match acronyms and abbreviations together anyway, displaying solid “knowledge” of the meaning of words beyond the mere characters forming the words.

After embedding the entities and the company names from the ORBIS database in a common vector space, we used the cosine similarity as similarity measure to calculate the pairwise similarity between the entities and the company names from the ORBIS database. Thereby, we consider those pairs with a similarity above 0.95 as matches. The 0.95 cut off value has been determined experimentally. Therefore, a match is only possible given at least a similarity of 0.95. If several company names from the ORBIS database fulfilled that criterion, the one with the highest similarity was utilized and represents the match for the entity identified from the news articles. Figure 6 shows the frequency distribution of pairwise cosine similarities. Here, for each Heise entity, the highest similarity to a ORBIS entity is considered.

A shortcoming of the approach described is that we may miss some actual matches because the similarity was below the cutoff value possibly due to noise in the extraction of entities during the NER step. Different news articles might refer to companies using different formats, i.e. SAP might sometimes show up as “SAP” and other times as “SAP AG”. Our goal was

---

<sup>9</sup><https://www.bvdinfo.com/en-gb/our-products/data/international/orbis>

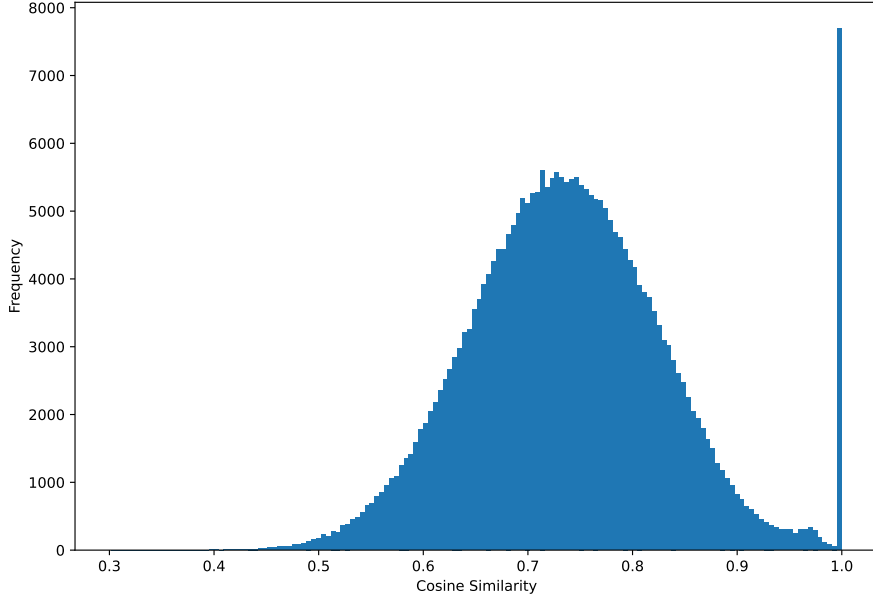


Figure 6: Histogram of pairwise cosine similarities for the highest cosine similarity for each Heise entity

therefore to find a common representation of each entity showing up in the news corpus, even if written in different ways. To this end, we developed an entity-clustering approach. Although this approach works well in some cases, it unfortunately also increases the noise in our data, so we prefer to avoid a larger entity data set and therefore stick to a dataset based on cosine similarities of the embedded ORBIS and Heise entities. The clustering approach is described in more detail in Appendix [G](#)

## 4.2 Descriptive Statistics of Matched Companies

Using the embed-and-match approach described in the previous sub-section, we were able to match 7,946 unique companies from the ORBIS database to entities extracted from the heise news corpus. In each document, we count the number of times one of these entities occurs resulting in a document-entity frequency matrix.

Table [1](#) provides the fifteen companies from the ORBIS database mentioned most frequently in the corpus. This list includes well-known companies such as NVIDIA, HP, or AMD with over 5,000 mentions. It should be noted

that these are primarily foreign parent companies, which are covered in the ORBIS database for the DACH region by their German subsidiaries.<sup>[10]</sup> The frequency of appearance of the 7,946 identified companies is 21.47 documents in mean with a median of 2 documents.

Furthermore, it has to be taken into account that also a non-negligible number of entities was assigned erroneously to a company in the ORBIS database. For example, the NER model recognised “SCO” as a proper name and classified the entity as “ORG”. Most likely, this entity in the news corpus refers to the “SCO Group”, which was active in the field of UNIX related software. However, in the entity homogenization process, the entity “SCO” was matched to the ORBIS entity “SCO GMBH”, a German facility management company.

Given the risk of incorrect matches, the proposed approach suggests interpreting entities from the perspective of technology-related entities as they would appear in a technology-focused news corpus. Furthermore, since we focus on innovation drivers and want to avoid the best possible noise in our data, we try to decrease the effect of incorrect matches by considering only those individual entities that are strongly associated with innovation-related topics. This step is described in the following sub-section.

### 4.3 Affinity of the ORBIS-entities to the innovation-related topics

For determining the affinity of the entities identified in the previous step (“ORBIS-entities”) to specific topics, we have to combine the information about the importance of topics in documents with the one about the documents in which the entities show up. To this end, we first construct a topic-entity-matrix by multiplying the topic-document weights  $\theta_d$  by the document-entity-frequency-matrix. All documents that do not contain any “ORG” entity are removed from the corpus for this analysis. The matrix multiplication thus results in a topic-entity-matrix with 120 rows and 7,946 columns. Each row corresponds to one topic and the entry in the columns to the ratio-scaled importance of the corresponding company in documents linked to this topic.

To determine the affinity to the innovation-related topics of the 7,946 ORBIS-entities, we divide the topic-entity-matrix for each  $entity_A$  by the

---

<sup>10</sup>“EBAY KLEINANZEIGEN” turned out to be the second most frequently mentioned entity. This is probably due to the advertising line of Ebay-Kleinanzeigen on the heise webpage, which was also scraped. Hence, for the following analyses, we exclude this entity, and also do not include it in Table [4](#)



“ORBIS-entity”	Frequency
NVIDIA	5550
HP	5215
AMD	5121
VODAFONE	4955
NSA	4951
T- MOBILE	3666
SAP	3434
SCO	3148
T-ONLINE INTERNATIONAL	2984
BNS	2504
NOVELL	2312
BSI	2105
FBI	2074
LG	1973
ADOBE SYSTEMS	1948

Table 1: Most frequent “ORBIS-entities”.

sum of the occurrence of  $entity_A$  across all topics. This generates a measure that expresses for each entity the distribution to which topics it is particularly associated. We call this association “topic affinity”.

To provide specific examples, we look at the topic affinity distribution of two “ORBIS”-entities: The first example is the Byton GmbH (“BYTON”), which is a now insolvent subsidiary of a Chinese start-up company that was active in the field of electrical cars. Consequently, one can assume that this company should have a pronounced affinity to some innovation related topics. The second example is the entity “SEC”, which can arguably be interpreted as the US Securities and Exchange Commission. Our intention is to exclude such entities that have obviously nothing to do with innovative activities by considering their topic affinity to innovation-related topics.<sup>11</sup>

Let us start with the topic-affinity-distribution for the entity “BYTON”, which is provided in Figure 7. The highest affinity of the entity “BYTON” is given with regard to topic 101 with 18.24%. This topic is not labelled as innovation related. However, “BYTON” also has a high affinity to two innovation-related topics, namely topic 67 (with 9.7%) and topic 90 (with

<sup>11</sup>It should also be noted that a match to an ORBIS entity should not actually occur. However, since we find company names with the letters “SEC” in our ORBIS database, we cannot exclude such matches beforehand. These false positives are the greatest obstacle in the homogenisation approach, but one that we approach with critical human judgement.

8.6%). Thus, we can classify “BYTON” as an entity with a relatively high affinity to innovation-related topics. The topics the entity is primarily associated with are shown in Figure 9. These are topic 101 (start-ups), topic 67 (e-mobility), topic 90 (autonomous vehicles) and topic 63 (company takeover). Considering that the entity represented a now insolvent startup in the field of electric cars, these empirical findings appear convincing.

In contrast, as shown in Figure 8, the entity “SEC” exhibits among the 10 topics with highest affinity only one innovation-related topic with a weight of just 2.04%. Thus, we might classify this entity as being not among the drivers of innovations, we are interested in. Additionally, we can confirm our suspicion that the Entity “SEC” represents the US Securities and Exchange Commission (SEC), as the word clouds for the topics with highest affinity shown in Figure 10 demonstrate, which are about the US stock market (topic 104), general reporting (topic 93), US enterprises (topic 50) and financial reporting (topic 6).

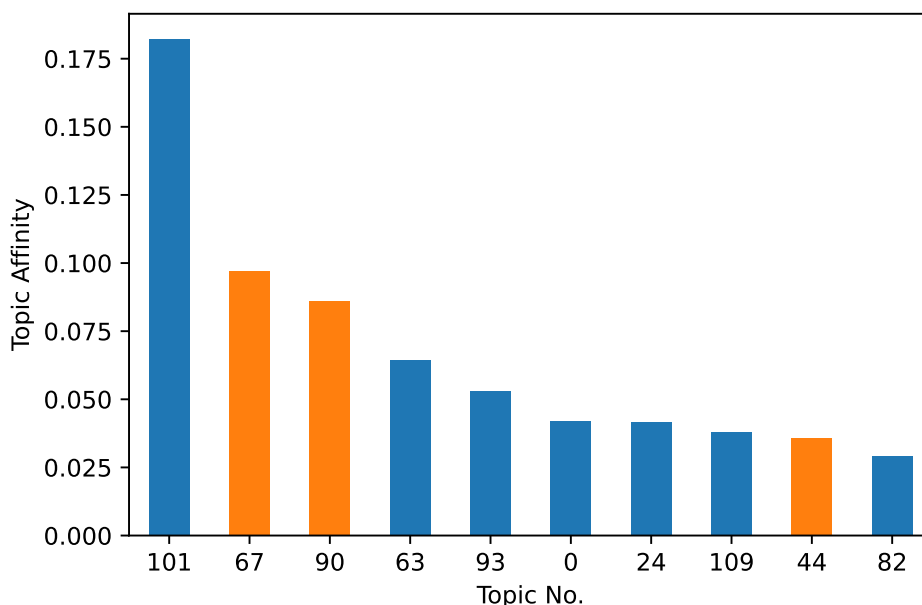


Figure 7: Topic-Affinity-Distribution for the Entity “BYTON”; orange bars stands for innovation related-topics

In the following, we would like to exclude entities that do not have a pronounced affinity to innovation-related topics which we have identified in Subsection 3.3. For this purpose, we look at the sum of affinity shares to the

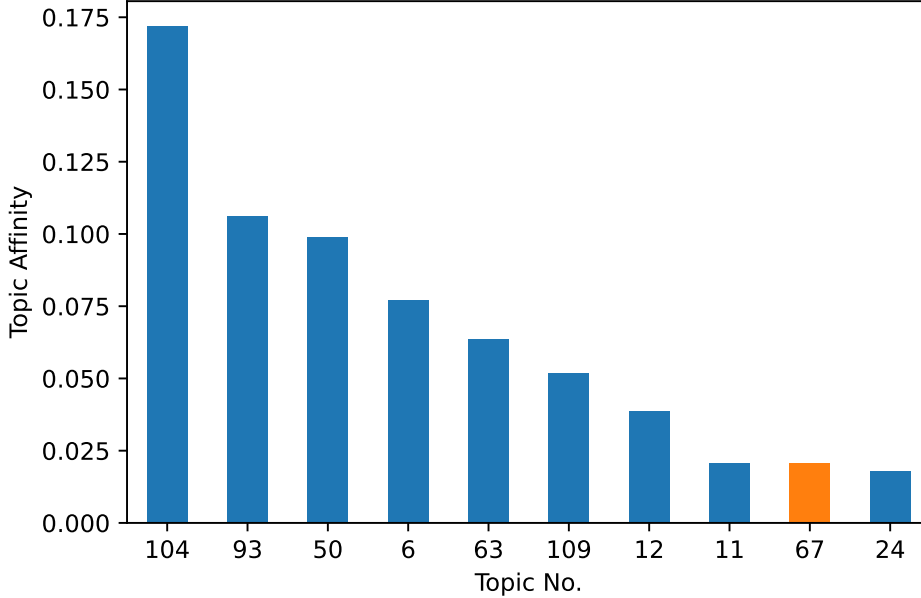


Figure 8: Topic-Affinity-Distribution for the Entity “SEC”; orange bars stands for innovation related-topics

42 innovation-related topics for each entity. The 50 entities with the highest affinity to innovation-related topics according to this measure are shown in Table 2.

The list in Table 2 contains several actual ORBIS-entities which are primary located in the DACH-region and have a high affinity for innovation. Examples are “NANOTRON TECHNOLOGIES”, which develops electronic components, “PIKKERTON”, which develops innovative sensors, and the former manufacturer of project management software (Scrum) “DANUBE”. Therefore, it can be concluded that the proposed approach is able to detect ORBIS-entities with a high affinity to innovation-related topics. Furthermore, we did a robustness check with InnoProb. InnoProb generates web-based innovation indicators that indicate the degree of innovation of a company (Kinne and Lenz, 2021). E.g., “NANOTRON TECHNOLOGIES” is among the most innovation-affine entities in our data set. This result is robust with the InnoProb value of 0.719, which is considered a very high value. These conclusions are also confirmed when looking at the topic affinity distribution in Figure 11. Three innovation-related topics in particular stand out there, representing the innovation fields “Wi-Fi components” (topic 7),

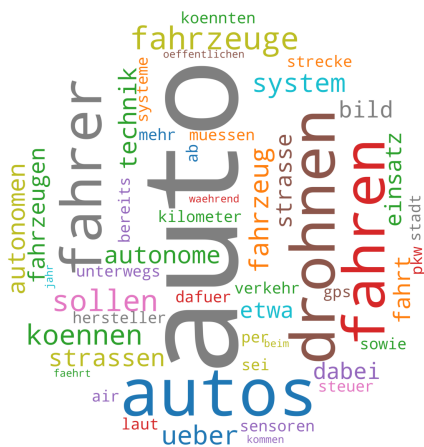
Topic #101



Topic #67



Topic #90



Topic #63



Figure 9: The Entity “BYTON” has the highest topic affinity to the following word clouds #101: “Start-ups”, #67: “E-Mobility” (innovation-related), #90: “Autonomous Vehicles” (innovation-related) and #63: “Company takeover”

“robotics” (topic 100) and “chip development” (topic 71).

Nevertheless, the listing should be used with necessary caution due to possible misclassifications, e.g. “ABOCOM” in the list does actually refer to the



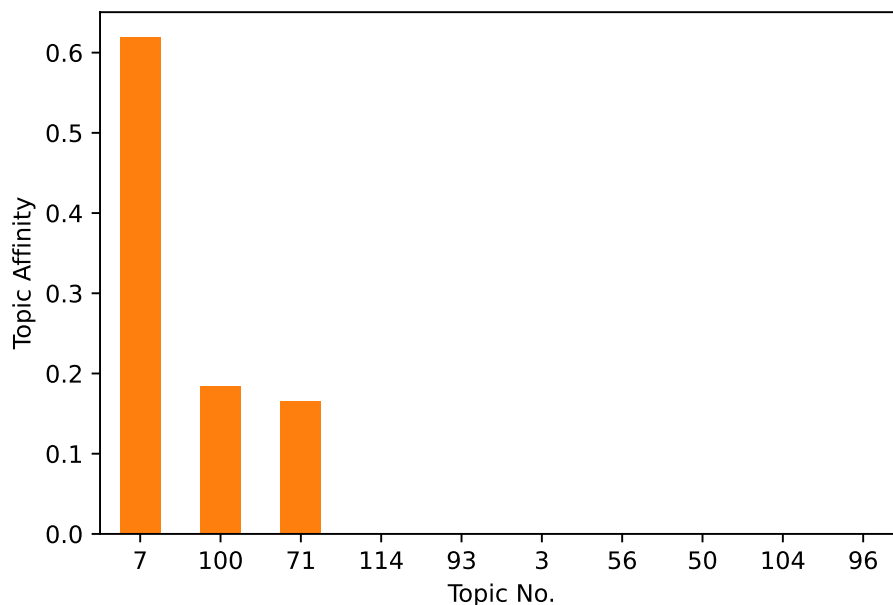


Figure 11: Topic-Affinity-Distribution for the Entity “NANOTRON TECHNOLOGIES”; orange bars stands for innovation related-topics

it might be interpreted as misclassification. However, making use of common sense or a company database with broader international covering, we could assign Abocom to the Taiwanese company, as would be correct according to our text corpus.

Other issues refer to entities which are mistakenly classified as “ORG” in the NER step. For example, “N5” is actually a manufacturing process for computer chips. Thus, “N5” has been incorrectly identified as an ORG-entity and subsequently matched to a company name in the ORBIS database that contains the characters “N5”. Similarly, “AIDE” refers to the Android IDE and “ABE” refers to the Apple Business Essentials. Finally, “VIGOR” refers to a router series, i.e. a product rather than a company. In general, companies with rather short names or even with acronyms are more difficult to match.

This approach may not be used to identify possible innovation drivers for different topics, as the risk of bias of the results is too high. It could happen that an entity that appears once in a document with a high proportion of innovation-related topics is immediately attributed an innovation affinity of 100%. If this entity appears only once in the entire corpus, this conclusion

would be very questionable. Thus, in order to identify innovation drivers of selected innovation fields, a different approach is taken further on.

Moreover, for the further analysis, we exclude all entities that have an innovation-affinity of less than 24.57%, which corresponds to the median of the innovation affinity distribution. This reduces the entity count to 3,973 unique entities.

“ORBIS”-Entity	innovation affinity
ABOCOM	0.997
PERLE	0.997
N5	0.996
AIDE	0.994
FCH	0.994
STR	0.993
AIB	0.992
ABE	0.989
ZERO	0.988
VIGOR	0.987
SMART NETWORK DEVICES	0.987
AMR	0.984
DANUBE	0.983
BSF	0.982
TERRASOFT	0.981
PIKKERTON	0.980
ROCKS	0.979
NANOTRON TECHNOLOGIES	0.976
TDA	0.977
FMD	0.976
ASIL	0.976
AMBA	0.974
EUPHORIA	0.974
NOBLE	0.971
ARCTURUS	0.971

Table 2: Twentyfive most innovation-affine “ORBIS-entities”

#### 4.4 Link Topics to ORBIS-Entities

Since we aim to link the ORBIS-entities to the topic-document-weights  $\theta_d$ , as mentioned previously, we multiply  $\theta_d$  by the document-entity-frequency-

matrix. Thereby, we only consider the document-entity-frequency-matrix with the 3,973 innovation-affine entities.

For the purpose of identifying the innovation drivers from the innovation-related topics, the ultimate measure obtained by the matrix multiplication is to be understood as a ratio-scaled metric. One can derive ratios between the entities, e.g. entity A is twice as important for topic X as entity B. Moreover, in order to be able to interpret the values of the resulting matrix multiplication as a percentage, each topic is divided by the sum of all values of the entities.

To give a concrete example, e.g. the innovation-related topics 67 & 90 are shown in Figure 3.

For the innovation related topics 67 (e-mobility) and 90 (autonomous vehicles), the top 15 possible innovation drivers are listed in table 3. Each ORBIS-entity can thus be interpreted as follows: For Topic 67, “NVIDIA” is the most important innovation driver with a share of 17.77% followed by “GM” with a share of 13.66% and “AMD” with a share of 4.18% respectively. In addition to the well-known companies, the ORBIS entity “BYTON” is also listed as an innovation driver with a share of 3.5%. “BYD”, which is actually matched to the ORBIS-entity BYD UG (haftungsbeschraenkt) should be interpreted simply as “BYD”. “BYD” is one of the largest manufacturers of electric and hybrid vehicles in the world, so this company is also listed as an innovation driver (due to the misclassification issue). We have to keep in mind, that the ORBIS-entities are restricted to companys based in the DACH region. Although the token ”tesla” is mapped very large in this innovation field (see wordcloud in figure 3), we would not expect to find the company “Tesla” in this list of innovation drivers. In topic 90 we also see that in addition to the established innovation drivers such as “GM” or “BOSCH”, smaller lesswell-known companies are also listed. For example, “MOIA” was founded in 2016 and offers mobility services or “RFID-Konsortium” provides AUTO-ID solutions, among other things, so that these companies are listed in the top 15 as plausible ORBIS-innovation drivers for the selected innovation fields. For the innovation-related topics, meaningful entities are listed from a qualitative point of view. Overall, it can be concluded that these results are suitable for identifying innovation drivers. Nevertheless, we have to keep in mind, that Non-DACH companies without a subsidiary in the DACH-region with similar company names in the ORBIS database also appear due to misclassifications in the listings. Accordingly, we need human assessment if we are interested exclusively in companies located in German-speaking countries.



“ORBIS”-Entity	67	“ORBIS-Entity”	90
NVIDIA	0.178	GM	0.057
GM	0.137	BOSCH	0.049
AMD	0.042	DFS	0.045
LG	0.040	NVIDIA	0.042
BYTON	0.035	DHL FREIGHT	0.037
BYD	0.019	MOIA	0.032
HP	0.019	KI	0.022
NPE	0.017	BBC	0.020
BOSCH	0.016	EAC	0.016
GROHMANN	0.015	ESP	0.011
GRUENE LIGA BRANDENBURG	0.013	BYTON	0.010
MUSIC UNLIMITED	0.012	RFID KONSORTIUM	0.010
RIM	0.012	PEGASUS	0.009
DHL FREIGHT	0.011	GARMIN	0.009
BBC	0.011	KIT	0.008

Table 3: Possible innovation drivers for topic 67 (“e-mobility”) and topic 90 (“autonomous vehicles”)

## 5 Discussion and Conclusion

We tried to identify drivers of innovation activities in different innovation fields based on technology-related newspaper articles. For this purpose, we first applied different topic models such as LDA to summarize the information contained in the textual corpus in form of latent topics. In the next step, these topics were classified by experts as “innovation-related” or “non-innovation-related”. From the subset of innovation-related topics, as examples, we focused on two innovation fields, namely “e-mobility” and “autonomous vehicles”. Therefore we examined these selected innovation-related topics in more detail with regard to the predominant entities linked to the documents with a high prevalence of these topics. This identification of companies was done using a NER-model, which helped to identify proper names belonging to an organization-entity (“ORG”). Since our focus is on the DACH-region innovation system, we were particularly interested in companies also listed in the ORBIS. We examined the affinity of the identified individual entities to the innovation-related topics and could list several innovation-affine ORBIS-entities. However, we have to point out that this listing should be used with common sense as it may include misclassifications. For two selected innovation related topics (#67: “e-mobility” and #90: “autonomous vehicles”), we presented potential ORBIS-innovation drivers.

From a qualitative point of view based on human judgement, the identified entities appear meaningful and do not only comprise the usual suspects for the particular innovation fields.

We also find that many of the entities identified as innovation drivers are DACH-located subsidiaries of large international corporations. For these companies it is not straightforward to decide whether the innovation contribution comes from the DACH-located subsidiary or from the international corporation as the matching of entities in the homogenization step as it is implemented does not allow for such a distinction. Future research will aim at a better differentiation in this respect, which might also be relevant for economic policy making.

Furthermore, we realize that there appear several misclassifications, which might be the result from the complex pre-processing steps. Consequently, some ORBIS-companies show up as innovation drivers which are in fact not, while we might miss some actual innovation drivers. These risks of errors of first and second type should always be kept in mind, when making use of the results, ideally combined with some critical expert judgment. Therefore, the lists generated with the proposed data-driven approach are intended to support human decision-making, but not to replace it. Future research will address the issues on how to reduce both type of errors mentioned above by improving both preprocessing and analysis of the obtained results. Proposed solutions include, for example, adapting the pre-trained NER model used specifically to the technology-related corpus through labeled entities. Thus, misidentifications of the entities or the entity span should be minimized. Furthermore, it would be ideal to perform the matching of the entities extracted with NER to a database with a broader coverage of entities. For this, an alternative entity linking to WikiData could be performed, which uses Wikipedia as knowledge base.<sup>12</sup>

Finally, we have to note that given the specific corpus of news articles, the approach is well suited for innovation drivers in the fields of technology and digitalisation, while other fields might be covered only to a limited extent. In this regard, future research might consider alternative text corpora for extending the scope of the analysis.

---

<sup>12</sup>[https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page)

## References

- Bird, Steven, E. L. and Klein, E. (2009). *Natural Language Processing with Python.*, O'Reilly Media Inc.
- Blei, D. M. and Lafferty, J. D. (2009). Topic models, *Text mining*, Chapman and Hall/CRC, pp. 101–124.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation, *J. Mach. Learn. Res.* **3**(null): 993–1022.
- Bystrov, V., Naboka, V., Staszewska-Bystrova, A. and Winker, P. (2022a). Choosing the number of topics in LDA models – a Monte Carlo comparison of selection criteria, *arXiv 2212.14074*, arXiv.
- Bystrov, V., Naboka, V., Staszewska-Bystrova, A. and Winker, P. (2022b). Cross-corpora comparisons of topics and topic trends, *Journal of Economics and Statistics* **242**(4): 433–469.
- Campagnolo, J. M., Duarte, D. and Bianco, G. D. (2022). Topic coherence metrics: How sensitive are they?, *J. Inf. Data Manag.* **13**.
- Carree, M. and Thurik, R. (2010). The impact of entrepreneurship on economic growth, in Z. J. Acs and D. B. Audretsch (eds), *Handbook of Entrepreneurship Research*, Vol. 1, Springer New York, NY, pp. 557–594.
- Chen, K., Cosgro, B., Domfeh, O., Stern, A., Korkmaz, G. and Kattampallil, N. A. (2021). Leveraging google bert to detect and measure innovation discussed in news articles, *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 1–6.
- Commission, E., Eurostat, for Economic Co-operation, O. and Development (2018). *Oslo manual 2018 : guidelines for collecting, reporting and using data on innovation*, Organisation for Economic Co-operation and Development.
- Dalkey, N. and Helmer, O. (1963). An experimental application of the DELPHI method to the use of experts, *Management Science* **9**: 458–467.
- Dwivedi, Y. K., Sharma, A., Rana, N. P., Giannakis, M., Goel, P. and Dutot, V. (2023). Evolution of artificial intelligence research in technological forecasting and social change: Research topics, trends, and future directions, *Technological Forecasting and Social Change* **192**: 122579.

- Eltyeb, S. and Salim, N. (2014). Chemical named entities recognition: a review on approaches and applications, *Journal of cheminformatics* **6**: 1–12.
- Ester, M., Kriegel, H. P., Sander, J. and Xiaowei, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise.
- Kinne, J. and Lenz, D. (2021). Predicting innovative firms using web mining and deep learning, *PLOS ONE* **16**(4): 1–18.
- Latifi, A. (2023). MaFiText-Bundestag speeches: Processing stenographic protocols of the German Bundestag, *Unpublished. Work in Progress* .
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents, *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14, JMLR.org*, p. II–1188–II–1196.
- Leitner, E., Rehm, G. and Moreno-Schneider, J. (2019). Fine-grained named entity recognition in legal documents, *Semantic Systems. The Power of AI and Knowledge Graphs: 15th International Conference, SEMANTiCS 2019, Karlsruhe, Germany, September 9–12, 2019, Proceedings*, Springer, pp. 272–287.
- Lenz, D. and Winker, P. (2020). Measuring the diffusion of innovations with paragraph vector topic models, *PLoS ONE* **15**.
- Montani, I., Honnibal, M., Honnibal, M., Landeghem, S. V., Boyd, A., Peters, H., McCann, P. O., jim geovedi, O'Regan, J., Samsonov, M., Orosz, G., de Kok, D., Altinok, D., Kristiansen, S. L., Kannan, M., Bournhonesque, R., Miranda, L., Baumgartner, P., Edward, Bot, E., Hudson, R., Mitsch, R., Roman, Fiedler, L., Daniels, R., Phatthiyaphaibun, W., Howard, G., Tamura, Y. and Bozek, S. (2023). explosion/spaCy: v3.5.0: New CLI commands, language updates, bug fixes and much more, *Technical report*, Zenodo.  
**URL:** <https://doi.org/10.5281/zenodo.7553910>
- Mühlroth, C. and Grottke, M. (2020). Artificial intelligence in innovation: How to spot emerging trends and technologies, *IEEE Transactions on Engineering Management* **69**: 493–510.
- Passonneau, R. J., Ramelson, T. and Xie, B. (2015). Named entity recognition from financial press releases, in A. Fred, J. L. G. Dietz, D. Aveiro, K. Liu and J. Filipe (eds), *Knowledge Discovery, Knowledge Engineering*

- and *Knowledge Management*, Springer International Publishing, Cham, pp. 240–254.
- Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.
- Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.
- Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the space of topic coherence measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15*, Association for Computing Machinery, New York, NY, USA, p. 399–408.
- Savin, I., Ott, I. and Konop, C. (2022). Tracing the evolution of service robotics: Insights from a topic modeling approach, *Technological Forecasting and Social Change* **174**: 121280.
- Sbalchiero, S. and Eder, M. (2020). Topic modeling, long texts and the best number of topics. some problems and solutions, *Quality & Quantity* **54**: 1095–1108.
- Schubert, E., Sander, J., Ester, M., Kriegel, H. P. and Xu, X. (2017). Dbscan revisited, revisited: Why and how you should (still) use dbscan, *ACM Trans. Database Syst.* **42**(3).
- Smith, A., Lee, T. Y., Poursabzi-Sangdeh, F., Boyd-Graber, J., Elmqvist, N. and Findlater, L. (2017). Evaluating visual representations for topic understanding and their effects on manually generated topic labels, *Transactions of the Association for Computational Linguistics* **5**: 1–16.
- Tang, J., Meng, Z., Nguyen, X., Mei, Q. and Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis, *International conference on machine learning*, PMLR, pp. 190–198.

# Appendix

## A Stopwords

We used the stop words from the `nltk`-package and preprocessed them analogously to our corpus preprocessing, so that the final list contains 232 stop words, which are listed below:

'aber', 'alle', 'allem', 'allen', 'aller', 'alles', 'als', 'also', 'am', 'an', 'ander', 'andere', 'anderem', 'anderen', 'anderer', 'anderes', 'anderm', 'andern', 'anderr', 'anders', 'auch', 'auf', 'aus', 'bei', 'bin', 'bis', 'bist', 'da', 'damit', 'dann', 'der', 'den', 'des', 'dem', 'die', 'das', 'dass', 'daß', 'derselbe', 'derselben', 'denselben', 'desselben', 'demselben', 'dieselbe', 'dieselben', 'dasselbe', 'dazu', 'dein', 'deine', 'deinem', 'deinen', 'deiner', 'deines', 'denn', 'derer', 'dessen', 'dich', 'dir', 'du', 'dies', 'diese', 'diesem', 'diesen', 'dieser', 'dieses', 'doch', 'dort', 'durch', 'ein', 'eine', 'einem', 'einen', 'einer', 'eines', 'einig', 'einige', 'einigem', 'einigen', 'einiger', 'einiges', 'einmal', 'er', 'ihn', 'ihm', 'es', 'etwas', 'euer', 'eure', 'eurem', 'euren', 'eurer', 'eures', 'für', 'gegen', 'gewesen', 'hab', 'habe', 'haben', 'hat', 'hatte', 'hatten', 'hier', 'hin', 'hinter', 'ich', 'mich', 'mir', 'ihr', 'ihre', 'ihrem', 'ihren', 'ihrer', 'ihres', 'euch', 'im', 'in', 'indem', 'ins', 'ist', 'jede', 'jedem', 'jeden', 'jeder', 'jedes', 'jene', 'jenem', 'jenen', 'jener', 'jenes', 'jetzt', 'kann', 'kein', 'keine', 'keinem', 'keinen', 'keiner', 'keines', 'können', 'könnte', 'machen', 'man', 'manche', 'manchem', 'manchen', 'mancher', 'manches', 'mein', 'meine', 'meinem', 'meinen', 'meiner', 'meines', 'mit', 'muss', 'musste', 'nach', 'nicht', 'nichts', 'noch', 'nun', 'nur', 'ob', 'oder', 'ohne', 'sehr', 'sein', 'seine', 'seinem', 'seinen', 'seiner', 'seines', 'selbst', 'sich', 'sie', 'ihnen', 'sind', 'so', 'solche', 'solchem', 'solchen', 'solcher', 'solches', 'soll', 'sollte', 'sondern', 'sonst', 'über', 'um', 'und', 'uns', 'unsere', 'unserem', 'unseren', 'unser', 'unseres', 'unter', 'viel', 'vom', 'von', 'vor', 'während', 'war', 'waren', 'warst', 'was', 'weg', 'weil', 'weiter', 'welche', 'welchem', 'welchen', 'welcher', 'welches', 'wenn', 'werde', 'werden', 'wie', 'wieder', 'will', 'wir', 'wird', 'wirst', 'wo', 'wollen', 'wollte', 'würde', 'würden', 'zu', 'zum', 'zur', 'zwar', 'zwischen'

## B Bigrams

When forming bigrams, two adjacent tokens in the corpus that are obviously dependent on each other are connected, e.g. by an underscore, to form a single token. An example of this corpus listed in [4](#) would be "light\_phone". The separate tokens "light" and "phone" would never suggest the meaning of the composite token (bigram) - a minimalist portable phone. A common

method for forming bigrams is implemented in the *Gensim.model.phrases*-module.

The bigram scoring function implemented in the module is calculated as follows:

$$(count(a, b) - min\_count) * \frac{N}{count(a) * count(b)} > Threshold \quad (1)$$

Thereby, *min\_count* is a parameter set to exclude all bigrams with a number of occurrences in the corpus lower than this value. *count(a,b)* stands for the total number of co-occurrences of tokens *a* and *b* as neighbours, while *count(a)* and *count(b)* stand for the total number of occurrences of tokens *a* and tokens *b* in the corpus, respectively. *N* denotes the size of the vocabulary. A bigram for *a* and *b* is created when the expression exceeds the a threshold parameter to be set. In order to avoid an ad hoc choice of this parameter, a data driven approach is followed. To this end, the selection of bigrams in our application was done as follows: First, all collocations occurring in the corpus were recorded. Then, for each collocation, the frequency of its occurrence in the corpus was determined and the corresponding idf-value calculated. The idf-value was calculated using the *scikit*-module and refers to the inverse document frequency. It is calculated as follows:

$$idf(collocation) = \log \left( \frac{D}{df(collocation)} \right) + 1 \quad (2)$$

It can be seen from equation (2) that the inverse document frequency is particularly small for collocations that occur in many documents. It can be assumed that collocations that occur in an excessive number of documents are not specific and thus should not be considered for forming bigrams. An example for such irrelevant collocations listed in Table 4 is "in\_der".

The frequency distribution of the idf-values is provided in Figure 12. This distribution shows that very few collocations occur in many documents. Therefore, all collocations for which the corresponding idf-value exceeds the 10 percent percentile – which corresponds to an idf-value of at least 11.5 – are considered as possible bigrams. Moreover, to be included in the list of bigrams, the collocation must occur at least five times in the corpus. According to the procedure described above, a list of 14,680 bigrams was created. Since the vocabulary was massively reduced to 4681 tokens in the last step of the corpus preprocessing, as described in Section 2, the bigrams do not play a relevant role among the most important tokens within a topic, as can be seen in the results shown in Section 3.

Collocation	Frequency	idf-value
in_der	141562	1.789840
fuer_die	119934	1.884407
in_den	106474	1.956937
...		
light_phone	22	11.722182
harmony_link	22	11.722182
pino_steps	22	12.415329
...		
strombedarfdecken	1	12.415329
gebe_konflikte	1	12.415329
energien_darueber	1	12.415329

Table 4: Some Examples of the collocations

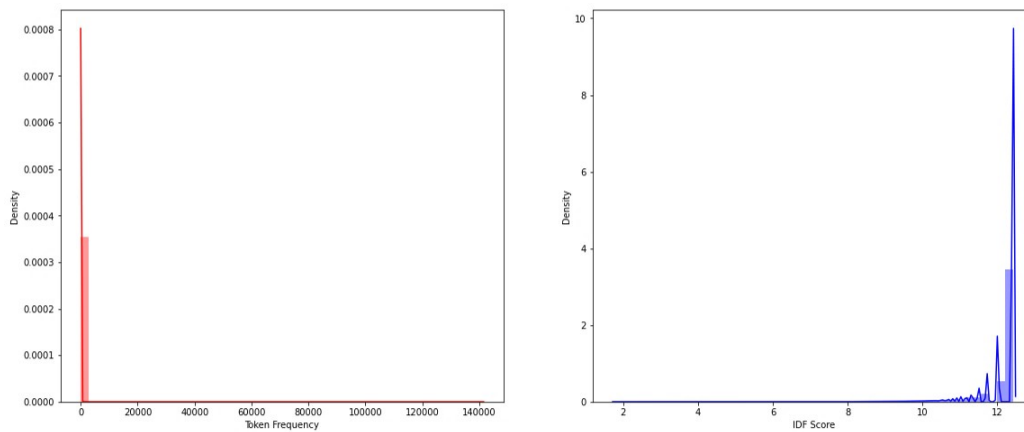


Figure 12: Histogram and Density Plot of the Descriptive Statistics of the Bigram Tokens



# C Matched Topics as Robustness check

LDA Topic: 43



PVTM Topic: 60  
Cosine similarity score:0.2728



LDA Topic: 47



PVTM Topic: 12  
Cosine similarity score:0.8597



LDA Topic: 67



PVTM Topic: 17  
Cosine similarity score:0.4587





## D Topic Labelling Instructions

Within the DynTOBI project we want to identify innovation-related topics resulting from a LDA model. We show the topics visualized as word clouds. The specific work assignment is as follows:

1. Read the definition of the term “innovation” provided below carefully.
2. Open both files in the folder.
3. In addition to the word clouds you can also see the diffusion curve of the topics, you can possibly use this information as an aid.
4. Then take small packages of topics at a time (no more than 20 in a work phase).
5. For each word cloud, consider whether or not that word cloud represents a topic to you that is related to innovation according to the definition provided.
6. If so, try to think of a name for this topic or at least try to narrow down the topic field somewhat in a few words. (If you find the topic innovation-related, but you can’t think of a name, then still list this topic).
7. Record the result in writing in this form:

E.g. innovative\_Topics\_LDA = 0:Tablets,22:Wikipedia,82:Smartphone etc..

Alternatively, use the attached excel file Topic-Labelling.xlsx, enter a 1 in the column ”LDA” if you consider the respective topic to be innovation-related, otherwise enter a 0. If you have entered a ”1”, then enter a name for the innovation-related topic in the second column “Naming of the innovation-related topic”.

According to the 2018 OSLO Manual, innovation is defined as follows: *“An innovation is a new or improved product or process (or combination thereof) that differs significantly from the unit’s previous products or processes and that has been made available to potential users (product) or brought into use by the unit (process).”*

Since this definition is to be interpreted from the perspective of entity considered (e.g. company, organization, person, etc.), we define innovation in a broader sense to reflect new products, new processes, new platforms, or new technologies in the time interval 1996 – 2021.

## E Results Topic Labelling

Topic Nr.	exp_1	exp_2	exp_3	exp_4	exp_5	exp_6	exp_7	exp_8	degree of innovation
0	0	0	0	0	0	0	0	0	0.000
1	0	0	0	0	0	0	0	0	0.000
2	0	0	0	0	0	0	0	0	0.000
3	0	0	0	0	0	0	0	0	0.000
4	0	0	0	0	0	0	0	0	0.000
5	1	1	1	1	1	1	1	0	0.875
6	0	0	0	0	0	0	0	0	0.000
7	1	1	1	1	1	1	1	1	1.000
8	1	1	1	0	0	1	1	0	0.625
9	0	1	0	0	0	0	0	0	0.125
10	1	0	1	1	1	1	1	1	0.875
11	0	0	0	0	0	0	0	0	0.000
12	0	0	0	0	0	0	0	0	0.000
13	0	0	0	0	0	0	0	0	0.000
14	0	1	1	0	0	1	1	0	0.500
15	0	0	0	0	0	0	0	0	0.000
16	1	1	1	1	1	1	1	1	1.000
17	1	1	1	1	1	1	1	1	1.000
18	1	1	1	1	1	1	1	1	1.000
19	0	0	0	0	0	0	0	0	0.000
20	0	0	0	0	0	0	0	0	0.000
21	0	1	0	0	0	0	0	0	0.125
22	0	0	0	0	0	0	0	0	0.000
23	1	1	1	0	1	1	1	0	0.750
24	0	0	0	0	0	0	0	0	0.000
25	1	1	1	1	1	1	1	1	1.000
26	1	1	1	1	1	1	1	1	1.000
27	0	0	1	0	0	0	0	0	0.125
28	1	1	1	1	1	0	1	1	0.875
29	0	0	0	0	0	0	0	0	0.000
30	0	0	0	0	0	0	0	0	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		

Topic Nr.	exp_1	exp_2	exp_3	exp_4	exp_5	exp_6	exp_7	exp_8	degree of innovation
31	0	0	0	0	0	0	0	0	0.000
32	1	1	1	1	1	1	1	1	1.000
33	0	0	0	0	0	0	0	0	0.000
34	1	1	1	1	1	1	1	1	1.000
35	0	0	1	1	0	1	1	1	0.625
36	0	0	0	0	0	0	0	0	0.000
37	0	0	0	0	0	0	0	0	0.000
38	0	0	0	0	0	0	0	0	0.000
39	0	0	0	0	0	0	0	0	0.000
40	0	0	0	0	0	0	0	0	0.000
41	0	1	1	1	0	1	1	0	0.625
42	0	0	0	0	0	0	0	1	0.125
43	1	1	1	1	1	0	1	1	0.875
44	1	1	1	1	1	1	1	1	1.000
45	1	1	1	1	1	1	1	1	1.000
46	1	0	0	0	0	0	0	0	0.125
47	1	1	1	1	1	0	1	1	0.875
48	0	0	1	0	0	0	0	0	0.125
49	0	0	1	0	0	0	0	0	0.125
50	0	0	0	0	0	0	0	0	0.000
51	1	1	1	1	1	0	1	1	0.875
52	1	1	1	1	1	0	1	1	0.875
53	1	1	1	1	1	1	1	1	1.000
54	0	0	1	0	1	0	0	1	0.375
55	1	1	1	1	1	1	1	1	1.000
56	0	0	0	0	0	0	0	0	0.000
57	0	0	0	0	0	0	0	0	0.000
58	0	0	0	0	0	0	0	0	0.000
59	0	0	0	0	0	0	0	0	0.000
60	0	0	0	0	0	0	0	0	0.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		

Topic Nr.	exp_1	exp_2	exp_3	exp_4	exp_5	exp_6	exp_7	exp_8	degree of innovation
61	1	1	1	1	1	1	1	1	1.000
62	0	0	0	0	0	0	0	0	0.000
63	0	0	0	0	0	0	0	0	0.000
64	1	0	1	0	1	1	0	0	0.500
65	1	1	1	1	1	1	1	1	1.000
66	0	0	1	0	0	0	1	0	0.250
67	1	1	1	1	1	0	1	1	0.875
68	1	1	1	0	0	0	0	0	0.375
69	0	0	1	0	0	0	0	0	0.125
70	0	0	0	0	0	0	0	0	0.000
71	1	1	1	1	0	1	1	0	0.750
72	1	1	1	1	0	0	0	1	0.625
73	0	0	0	0	0	0	0	0	0.000
74	0	0	0	0	0	0	0	0	0.000
75	1	1	1	0	0	0	1	0	0.500
76	1	0	1	1	1	1	1	0	0.750
77	0	0	1	0	1	0	1	0	0.375
78	1	1	1	1	1	1	1	1	1.000
79	1	1	1	1	1	1	1	1	1.000
80	0	0	1	0	0	0	1	1	0.375
81	0	0	0	0	1	0	0	1	0.250
82	0	0	0	0	0	0	0	0	0.000
83	0	0	0	0	0	0	0	0	0.000
84	0	0	0	0	0	0	0	0	0.000
85	0	0	1	0	1	0	1	0	0.375
86	1	1	1	1	1	0	1	1	0.875
87	0	0	1	0	0	0	0	0	0.125
88	0	0	1	0	0	0	0	1	0.250
89	0	0	1	0	0	0	0	0	0.125
90	1	1	1	1	1	1	1	1	1.000
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		

Topic Nr.	exp_1	exp_2	exp_3	exp_4	exp_5	exp_6	exp_7	exp_8	degree of innovation
91	1	0	1	0	0	0	0	0	0.250
92	0	0	1	0	0	0	1	1	0.375
93	0	0	0	0	0	0	0	0	0.000
94	0	0	0	0	0	0	0	1	0.125
95	1	1	1	1	1	1	1	1	1.000
96	0	0	0	0	0	0	0	0	0.000
97	0	0	0	0	0	0	0	0	0.000
98	1	1	1	1	1	1	1	1	1.000
99	1	1	1	1	1	1	1	1	1.000
100	1	0	1	0	1	1	1	0	0.625
101	0	1	1	0	0	1	1	0	0.500
102	1	1	1	1	1	0	1	1	0.875
103	0	0	0	0	0	0	0	0	0.000
104	0	0	0	0	0	0	0	0	0.000
105	0	0	1	0	1	0	1	1	0.500
106	0	0	0	0	0	0	0	0	0.000
107	0	0	1	0	0	0	1	1	0.375
108	0	0	1	0	0	0	0	0	0.125
109	0	0	0	0	0	0	0	0	0.000
110	0	1	0	0	0	0	0	0	0.125
111	1	1	1	1	1	1	1	1	1.000
112	0	1	1	0	0	1	1	0	0.500
113	1	1	1	1	1	1	1	1	1.000
114	0	0	0	0	0	0	0	0	0.000
115	1	0	1	0	0	0	0	0	0.250
116	0	0	0	0	0	0	0	0	0.000
117	0	0	1	0	0	0	0	0	0.125
118	1	0	1	1	1	1	1	1	0.875
119	1	1	1	1	1	1	1	1	1.000

## F Example Entity Extraction

The Hildesheim company Pios licensee of BeOS and German distributor of the BeBox, wants to continue to vouch for the support of their approximately 200 German customers, "as long as the BeBox and the company Pios exist", says the managing director Geerd-Ulrich Ebeling to c't.

Figure 13: Example of identified entities (highlighted) in a short news article, where “ORG” identities are framed.

## G Homogenization of Entities: Clustering approach

To include those entities that we just tightly missed due to the similarity threshold of 0.95, we homogenized the news article entities that had been matched to an ORBIS entity via clustering. To this end, we used the scikit-learn implementation of DBSCAN (Ester et al., 1996; Schubert et al., 2017).<sup>13</sup>

DBSCAN has two main parameters that control the cluster identification, epsilon (*eps*) and minimum samples (*minsamples*), which should be set according to the characteristics of the data at hand. Given a set of parameters, cluster centroids (core samples) are entities that satisfy the condition of having at least *minsamples* other entities within a proximity of “*eps*” in the vector space, thus qualifying them as neighbors of the core entity. This means that the cluster centroid exists within a region of high data density in the vector space. We set the epsilon to 0.15 and the minimum samples to 3 and used  $1 - \text{cosinesimilarity}$  as distance measure between samples.

For each cluster, the most frequent way of spelling the entity in the news articles has been chosen as the main representation of that entity. In cases where news article entities were not matched to an ORBIS entity (i.e., when

<sup>13</sup><https://scikit-learn.org/stable/modules/clustering.html#dbscan>



the highest values of similarity was smaller than 0.95) they are still assigned if another member of the same cluster could be matched to an ORBIS entity.

The details of the procedure are shown for the example “Microsoft” for clarification. The members of the corresponding cluster represent the different versions of the company name found in the news articles that have been clustered together. The cluster comprises the elements “microsoft”, “microsoft-DE”, “microsoft.”, and “microsoft::”. The corresponding entry in the ORBIS database is “Microsoft”.

The values of the similarity measure for the four elements in the cluster are 1, 0.9, 0.99, and 0.945, respectively. Thus, only the first and third element would lead to a match, while we would lose the news with entities “microsoft.” and “microsoft-DE”. However, since they belong to the same cluster as a news article entity that could be matched to an ORBIS entity, they are also assigned to the ORBIS entity “Microsoft”. Thus, the additional step of the cluster procedure helps identifying a larger share of all relevant entities in the news corpus.

Table G provides the results of the procedure for five further examples.

	IG Metall	XBit Labs	Bertelsmann AG	Schneider Electric	Barnes & Noble
0	IG Metall	XBit Labs	Bertelsmann AG	Schneider Electric	Barnes & Noble
1	IG Metall.	Xbit Laboratories	Bertelsmann-Stiftung	Schneider Technologies	Barnesandnoble.com
2	IG Metall,	Xbit Labs	Bertelsmann Stiftung	Schneider Electronics	Barnes & Nobles
3	IG-Metall	XBit Laboratories	Bertelsmann Stiftung.	Schneider Technologies AG	Barnes & Noble.
4	IG Metall Küste		Gütersloher Bertelsmann AG	Schneider Electronics AG	Barnes & Noble.com
5	IG Metall Bayern		Bertelsmann-Stiftung.	Schneider Electronics GmbH	Barnes and Noble
6	IG Metall-Chef Klaus Zwickel		Bertelsmann Verwaltungsgesellschaft	Schneider Electric,	Barnes & Nobles.
7	IG Metall und ver.di		Gütersloher Bertelsmann Stiftung	Schneider Electric, Selex	Barnes & Noble, Google
8	IG Metall Bocholt,		Bertelsmann Stiftung;		www.barnesandnoble.com
9	IG-Metaller		Bertelsmann Inc.		Barnesandnobles.com

Table 5: Examples of the results from the clustering step. Column header is the cluster name, i.e. the writing version found most often in news articles, and the rows represent the most frequent versions extracted from the news articles.

## H Possible “ORBIS”-Innovation drivers for Topic 43 & Topic 47

For the topics “virtual reality” and “online retailing” in the book industry, the potential ORBIS innovation drivers<sup>14</sup> are listed in 6.

<sup>14</sup>Again, only the innovation-affine ORBIS entities are considered.

43		47	
HTC	0.272380	FRANKFURTER BUCHMESSE	0.085808
VR PROJEKTE	0.240609	KINDLE	0.080105
HP	0.053828	MUSIC UNLIMITED	0.051733
NVIDIA	0.043867	ADOBE SYSTEMS	0.048073
AMD	0.025683	DHL FREIGHT	0.041520
IFA	0.018414	LIBRI	0.033058
LG	0.015165	AWS	0.030004
DAYDREAM	0.013823	ALEXA	0.027311
RED HAT	0.010971	LIDL	0.022572
UNITY	0.010200	PLASTIC LOGIC	0.020435
ADOBE SYSTEMS	0.008605	MVB	0.019963
MSI	0.007764	AUDIBLE	0.017372
EMC	0.007264	ABEBOOKS EUROPE	0.016736
CRYTEK	0.006829	IFA	0.016393
BBC	0.006460	HP	0.014269

Table 6: Possible innovation-drivers for topic 43 (“virtual reality”) and topic 47 (“digital books and online trading”)