// NO.19-063 | 11/2020

DISCUSSION PAPER

// JANNA AXENBECK AND PATRICK BREITHAUPT

Web-Based Innovation Indicators – Which Firm Website Characteristics Relate to Firm-Level Innovation Activity?

Leibniz Association

ZEW

Web-Based Innovation Indicators – Which Firm Website Characteristics Relate to Firm-Level Innovation Activity?

Janna Axenbeck[1,2,◑,*], Patrick Breithaupt[1,◑]

1 Department of Digital Economy, ZEW – Leibniz Centre for European Economic Research, L7 1, 68161 Mannheim, Germany
2 Justus-Liebig-University Giessen, Faculty of Economics, Licher Straße 64, 35394 Gießen, Germany

◑The authors contributed equally to this work.
* janna.axenbeck@zew.de

First version: 31.12.2019 This version: 24.11.2020

Abstract

Web-based innovation indicators may provide new insights into firm-level innovation activities. However, little is known yet about the accuracy and relevance of web-based information for measuring innovation. In this study, we use data on 4,487 firms from the Mannheim Innovation Panel (MIP) 2019, the German contribution to the European Community Innovation Survey (CIS), to analyze which website characteristics perform as predictors of innovation activity at the firm level. Website characteristics are measured by several data mining methods and are used as features in different Random Forest classification models that are compared against each other. Our results show that the most relevant website characteristics are textual content, the use of English language, the number of subpages and the amount of characters on a website. Furthermore, using several website characteristics jointly improves predictions of reported innovation activity up to 20 percentage points in comparison to our baseline model. Moreover, results also indicate a better performance for the prediction of product innovators and firms with innovation expenditures than for the prediction of process innovators.

Keywords: Text as data, innovation indicators, machine learning

JEL Classification: C53, C81, C83, O30

1 Introduction

Innovation, defined as the implementation of either new or significantly improved products or processes as well as combinations thereof [1], brings vast benefits to consumers and businesses. Moreover, technological progress is considered as a main driver of economic growth [2]. It is, therefore, a matter of public interest to analyze and understand innovation dynamics as it is conducted in several studies (e.g., [3–9]).

A prerequisite for the analysis of innovation-related questions is to correctly measure firm-level innovation activities. However, it should be noted that no universally accepted measurement approach exists. For example, firm-level innovation indicators are traditionally constructed with data from large-scale questionnaire-based surveys like the biennial European CIS or the annual MIP (see [10, 11]), which is also the German contribution to the CIS. However, these innovation indicators suffer from some major drawbacks (i.e., [12–14]). For instance, the MIP annually surveys around 18,000 firms. This only corresponds to a fractional share of the total stock of German firms and therefore lacks regional granularity and coverage. In addition to this, questionnaire-based surveys – especially on a large scale – have the added disadvantages of being costly and a lack of timeliness. Also, most surveys require firm participation and as a consequence, surveys such as the MIP suffer from low response rates [12]. Besides, firm-level innovation can also be studied by patent or publication analysis. However, respective indicators only cover technological progress for which legal protection is sought [15,16] and not every innovation can be patented. For example, due to the German regulatory framework it is quite difficult to patent software, i.e., digital innovations.

Issues, however, could be solved by adding web-based information: Advances in computing power, methods for statistical learning as well as natural language processing tools enable, e.g., researchers to extract website information on a large scale. This makes it technically possible to complement traditional innovation indicators with information from scraped firm websites. Nowadays, almost every firm has an online presence. Firm websites can entail information about new products, key personnel decisions, firm strategies, and relationships with other firms [17]. Those pieces of information might be directly or indirectly related to a firm's innovation status. By using this information, it is possible to conduct an automatic, timely and comprehensive analysis of firm-level innovation activities, as measurements can be carried out faster and in shorter intervals in comparison to traditional indicators.

The contribution of this paper to the question whether web-based innovation indicators are feasible is threefold. First, we analyze to what extent firm websites improve predictions of firm-level innovation activity. Second, we assess which characteristics of a website relate most to a firm's innovation status. Third, we examine which characteristics are appropriate for predicting different forms of innovation activity. We test the latter by additionally comparing the predictive power of different innovation indicators related either to product innovations, process innovations or innovation expenditures. We assume differences between indicators, for example, because firms with process innovations may have a smaller incentive to announce respective innovation activity. This may be due to the fact that new processes are less relevant for most website visitors.

For our analysis, data on 4,487 German firms from the MIP 2019 is used. We extract their websites' text and hyperlink structure by applying the ARGUS web-scraper [13]. Several methods including topic modelling and other natural language processing tools are applied to generate features that potentially relate to the firm-level innovation status. Furthermore, we extract information related to a website's technical maturity such as how fast it is responding and whether a version for mobile end user devices is available. After extracting and calculating a wide variety of features, we divide them into three different feature sets: I) text-based features including, e.g., words, document-topic probabilities derived from a topic modelling algorithm, and the share of English language, II) meta information features including, e.g., website size related features, availability of a mobile version and loading time, and III) network features including, e.g., hyperlinks to social networks as well as incoming and outgoing hyperlinks. Based on these three feature groups, we analyze which website characteristics best predict a firm's innovation status reported in the MIP 2019 by using a Random Forest classifier.

Our results show that predictions based on website characteristics perform unambiguously better than a random prediction. Consequently, firm websites entail information that relate to firm-level innovation activity. In addition, our website characteristics better predict firms with product innovations and innovation expenditures than with process innovations. Moreover, text features make the biggest contribution to our prediction performance.

Evaluating the predictive power of single variables across feature sets by means of the mean decrease in impurity (MDI), the language of a website and website size measured by the number of subpages as well as the total amount of characters are always relevant in the models with the highest predictive power for all considered innovation indicators. Moreover, there are characteristics that are highly important only for specific indicators, e.g., the verb "to develop" is more important for innovation expenditures and product innovators than for process innovators.

The remainder of this paper is structured as followed: Previous literature is reviewed in Section 2. In Section 3, we present our data and in Section 4 the descriptive statistics. Section 5 describes the methodology and Section 6 shows the results, which are discussed in Section 7. This paper concludes in Section 8.

2 Literature review

The usage of text data to generate innovation-related indicators has been tested in previous studies. For example, [18] show that the significance, i.e., relevance, of a patent is higher when its textual content is very distinct to previous patents but similar to subsequent ones. [19] generate innovation-related topics from 170,000 technology news articles using a Paragraph Vector Topic Model. They analyze the diffusion of the identified topics within the text corpus. Their results suggest that technology trends can be assessed by measuring the importance of topics over time. Using PATSTAT data, [20] show that context similarity of technological codes relates to innovative events. The likelihood that new combinations of technological codes appear in one patent can be predicted by their context similarity in patents where they have been used before.

Remarkable work is also conducted by [21]. In this study, a Latent Dirichlet Allocation (LDA) model is fitted with analyst reports of firms included in the S&P 500 index. The LDA topic that has the lowest Kullback-Leibler divergence to the wording of a mainstream economic textbook on innovation is chosen as innovation indicator. The authors show that firms have patents with greater impact (i.e., more citations per patent) if the innovation topic has a larger share in their analyst report. However, analyst (or also annual) reports are not available for every firm and smaller firms are particularly underrepresented. In contrast, firm websites are available for a large share of small and medium-sized firms.

Furthermore, previous literature shows that information produced online can be used to construct frequent real-time estimates [22]. Famous ‘now-casting’ examples that utilize web-based information are [23], who use Google search queries to accurately predict influenza activity in the United States. [24] claim that search engine query indices are also often correlated with economic activities and enable to generate frequent indicators. They show that forecasts concerning, for example, automobile sales and unemployment can be significantly improved by including search term indices in prediction models. Not only information from online searches but also firm website information can be used to generate economic indicators: As they provide detailed information about the firm as well as its products, they appear to be suitable for measuring firm-level innovation activities [17]. [13] summarize previous studies that analyze the possibility of firm website-based innovation indicators (e.g., [17], [25], [26], [27], [28] and [29]).

Most studies solely focus on the hyperlink structure of websites or only conduct a simple keyword search and are limited to small amounts of firms from a particular economic sector.

Firstly applying advances in statistical learning, [30] attempt to predict innovation at the firm level using textual information on websites and novel machine learning tools. They use a questionnaire-based firm-level product innovation indicator (innovative/ non-innovative) from the MIP years 2015-2017 as a target variable to train an artificial neural network classification model on website texts. The authors only consider stable product innovators in their main analysis. Firms that switch between innovation statuses, which is highly relevant in the field of innovation economics, are only observed in a secondary analysis. The average F1-score for the respective prediction is 0.68%. Additionally, [14] fit several machine learning models to develop a firm website-based innovation indicator, with their annotated data set being limited to 500 firms. One important characteristic of their work is the individual analysis of websites' subpages instead of predicting the innovation status of an entire website, i.e., firm. Additionally, their subpages are manually labelled as either innovation or non-innovation-related messages instead of using survey or patent data as target variables. The best performance is achieved with an artificial neural network. Even though the predictive performance is very high, the authors cannot show a credible external validity of their indicator.

Furthermore, another issue of both approaches is that neural networks do not reveal any decision rule that can be easily interpreted by humans, which is why they are often called black box models. It should also be noted that both studies only consider text. Nonetheless, previous results show that there must be distinct website characteristics that relate to a firm's innovation status, but the particular website characteristics are not identified yet.

[31] analyze whether firm's expenditures on innovation can be predicted by means of administrative records and balance sheet data. Using a Random Forest regression approach, the authors identified firm size, sectoral affiliation and investment in intangible assets as the most important predictors. Random Forests usually provide better predictive performance than linear methods while retaining the interpretability of feature relevance.

By applying a Random Forest approach to a large scale firm-level data set, we are able to analyze which website characteristics are linked to firms' innovation activity and are highly predictive. One further contribution of our paper is to address shortcomings of previous literature, as it provides new and detailed insights on the question whether firm websites entail measurable information on firm-level innovation activities.

3 Data

Based on the Oslo Manual, in our data set an innovation is defined as "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)" [1, p. 20]. Furthermore, we consider all expenditures spent for innovation purposes as innovation expenditures and summarize firm-level product or process innovation as well as innovation expenditures as innovation activity.

We use data from the MIP 2019 to classify firms as either innovative or non-innovative. The MIP is an annual survey conducted by the ZEW – Leibniz Centre for European Economic Research. The survey covers firms from manufacturing and service sectors and is conducted as a mail survey with the option to respond online.

In the MIP 2019, firms were asked whether they introduced a product or process innovation within the last three years (between 2016 and 2018) and for the total amount spent on innovation activities in the last year (2018). We consider a firm that stated, it introduced a product innovation within the considered time frame as a product innovator and a firm that stated that it introduced a process innovation within the considered time frame as a process innovator. A firm is an innovator if it introduced at least one of both. Every firm that spent financial resources on innovation - independent of the magnitude - is regarded as a firm with innovation expenditures. Our initial sample consists of 13,747 firms from the MIP 2019. We merge these firms with the Mannheim Enterprise Panel (MUP, see [32]), which consists of more than 3.2 million economically active firms, to receive information about the firms' website addresses. The MUP serves as a sampling frame for surveys like the MIP and, e.g., contains firm-level information on turnover, number of employees and sector affiliation. Only 54 percent of firms in our sample can be assigned to website addresses, as we limit ourselves to quality-assured observations. In total, we have 6,368 firms with information on the website address and at least one innovation indicator. We extract website content by applying the ARGUS web-scraper, which allows us to collect texts as well as hyperlinks to other websites. Firm websites were first scraped in September 2018 to collect texts, then again in January 2019 for adding hyperlinks. We scraped a third time in October 2019 to add information about technical features, e.g., capturing the existence of firm websites for mobile end user devices. The maximum limit of scraped subpages per website is set to 50, otherwise the amount of data would become too large. We consider this to be a sufficient number, as the median number of subpages in the MUP is 15 (see [13]) and only 1.5 percent of all firms in our subsample have 50 or more subpages. Moreover, the scraping program is set to prefer subpages with shorter website addresses because we assume these subpages include more important information about the firm. Also, ARGUS is set to prefer websites in German language. Hence, when we calculate the share of different languages on a website we expect a small bias. However, since only a few firms exceed the subpage limit, we assume this bias to be negligible. While scraping the data, especially while collecting meta information features, we received several error messages. Furthermore, we only use observations for which all features are non-missing. If, for example, a meta information feature is not available the observation will not be used for training or testing with other feature sets. Therefore, after the entire data collection process, we end up with 4,487 firms in our sample when predicting product innovators and innovators, 4,484 firms when predicting process innovators and 1,893 when predicting whether a firm has innovation expenditures (Table 1). There are three observations more for product innovators than for process innovators. Since these three observations are all product innovators, they are also in the innovator sample.

Table 1: Summary statistics for product innovators, process innovators, innovators as well as firms with innovation expenditures.

Variable	Definition	N	Mean	SD	Min	Max
Product innovators	1: If firm is a product innovator 0: Otherwise	4,487	0.39	0.49	0	1
Process innovators	1: If firm is a process innovator 0: Otherwise	4,484	0.52	0.50	0	1
Innovators	1: If firm is a product or / and process innovator 0: Otherwise	4,487	0.61	0.49	0	1
Innovation expenditures	1: If firm innovation expenditures were reported 0: Otherwise	1,893	0.39	0.49	0	1

Additionally, a random sample of 32,276 website addresses of firms not included in the MIP is drawn from the MUP and scraped with the ARGUS webscraper using the same settings as for the MIP sample. The sample is used for topic modelling. We train a topic model on a separate sample for two reasons. First, it allows to include more data points. Second, it ensures that no observation used for calibrating topics is considered for evaluating Random Forest models. Hence, it prevents data leakage. The sample is hereinafter referred to as the LDA sample.

As we need to exclude a large share of observations due to missing values in our MIP sample, we cannot rule out a selection bias. Also, firms from certain industries and smaller firms are less likely to have a website and may therefore be underrepresented. In machine learning, adverse selection might lead to two issues: It could cause that our model is better fitted for groups that are overrepresented in our sample and it could induce that the class correlated with the overrepresented group is predicted more often. To identify whether a potential selection bias exists, we analyze how the sample distribution changes with respect to the number of employees and industry sectors, when excluding observations with missing information (see S1 Appendix and S2 Appendix).

Except for “transportation and post” (sector 15), we do not see a notable change in the distribution of firms that could be linked to a severe selection bias.

To capture website characteristics, we apply several methods to generate features like a keyword search and natural language processing as well as an analysis of hyperlinks (network analysis methods). We use Python as programming language for calculating our features and for training our Random Forest models. For an overview of feature sets, see Table 2.

3.1 Text-based features

Information from website texts is analyzed, as it might be related to a firm's innovation status for the following reasons: Presumably, most firms are using their websites to inform customers about new products or services and might mention whether their product is new or innovative, i.e., it is likely that innovative firms use particular innovation-related words. Information about process innovations can also be detected and used if reported on the website. Moreover, a firm might report that it uses a recently emerging technology like blockchain, 3D printing or augmented reality (for an overview of recently emerging technologies, see S3 Appendix). Hence, an emerging technology term might appear on a firm's website and if so it is likely that the firm can be considered as innovative, at least on an incremental level, as it makes use of technologies that are fairly new. Additionally, there might be latent patterns on a website that reveal a firm's innovation status, these latent patterns can be captured by the LDA topic modelling approach as successfully shown in [21]. Furthermore, innovative firms might follow some general technological trends like the digital transformation. As these technological trends are quite general, LDA topics related to these trends might appear quite often on firm websites. To capture this, we construct a topic popularity index that indicates the distribution of popular and less popular topics on a website.

We additionally analyze the following text-based metrics: Languages that

Table 2: Features related to text, meta information and network measures.

Text-based features	
1) Textual content	Term-document matrix with the 5,000 most frequent words (TF-IDF applied).
2) Emerging technologies	Dummy variable that measures whether a technology of Wikipedia's list of emerging technologies appears on a firm's website.
3) Latent patterns	Topic-document probabilities of 150 topics generated by the LDA approach.
4) Topic popularity index	The sum of LDA topic probabilities per document. Each probability is weighted with the relative frequency of its appearance in the entire LDA sample.
5) International orientation	Share of subpages in English language and the share of all other non-German subpages in all subpages.
6) Share of numbers	The share of numbers in website text (characters).
7) Flesch-reading-ease score	Numerical metric assessing readability of texts.
Meta information features	
8) Website size	Number of subpages on a website, total amount of characters on a website.
9) Loading time	The time from sending a request (http/https) to a webserver (to get the start page of a website) until the arrival of the response (in ms).
10) Mobile version	Dummy variable that is one if a version for mobile end user devices exists and zero otherwise.
11) Domain purchase year	The year of the first entry at web.archive.org.
Network features	
12) Centrality	The total number of incoming, the total number of outgoing hyperlinks as well as the PageRank centrality.
13) Social media	Number of hyperlinks to Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub, Flickr, and Vimeo.
14) Bridges	Number of bridges a firm is part of in the hyperlink network.

appear on a website might relate to the export status of a firm and this could provide information about a firm's innovation status because the export status is linked to firm-level innovation (e.g., [33], [34], [35]). Also, we test whether the share of numbers in all string characters (text) as well as the text complexity measured by the Flesch-reading-ease score [36] differ between innovative and non-innovative firms.

3.2 Meta information features

Second, meta information of firm websites (see Table 2) might allow to distinguish innovative from non-innovative firms. For example, the website size might help to predict a firm's innovation status. Large firms are more likely to be innovative [10].

As the number of subpages of a website correlates with the number of employees of a firm [13], the size of a website might provide information about whether a firm introduced an innovation. Also, the technological properties of a website could be relevant. Innovative firms might have a better technical knowledge and are able to apply more technological advanced features on their websites. For example, the loading time of a website could be faster and a mobile version might be more often available when firms are more technologically advanced. However, there might be some noise because the loading time may also be short if the website is relatively simple.

Another potentially relevant feature is the age of a website, i.e., the domain purchase year, as it might relate to the actual firm age. One has to consider, however, that this relationship is unlikely to be linear. On the one hand, a website that is fairly new might indicate a start-up with an innovative idea. On the other hand, having a very old website means the firm has adopted this new technology very early. This could also relate to a more technological advanced, hence, innovative firm.

3.3 Network features

Third, hyperlinks between websites (see Table 2) might also help to identify the firm-level innovation status. Firms that have more business relationships with other firms or are more relevant according to centrality measures might be better informed and know earlier about new profitable applications. Hence, firms with more relationships to other firms could be more likely to be innovative. Moreover, innovation projects are often realized in cooperation with other firms (e.g., [37]). Thus, patterns in firm-level cooperation are expected to be of interest. A firm that connects (or bridges) different network parts is usually relevant and its removal will decompose the network.

Lastly, [38] show that a firm's use of the social networking site Facebook is linked to product innovations. Hence, the use of social media might reveal information about a firm's innovation status, as well. Our study analyzes whether the three groups of features differ in their performance when predicting a firm's innovation status. A more detailed description of the feature generation can be found in S4 Appendix.

4 Descriptive analysis

The descriptive statistics for our predictor variables are presented in this section. Table 3 shows mean values for innovative and non-innovative firms as well as p-values regarding the difference of both means for selected features.

Table 3: Descriptive statistics for selected variables.

	Group-specific means											
	Product innovator			Process innovator			Innovator			Innovation expend.		
Feature (Variable name)	Yes	No	P-val.	Yes	No	P-val.	Yes	No	P-val.	Yes	No	P-val
Text-based features												
Emerging technology term (*emerging_tech*)	0.18	0.07	0.00	0.15	0.07	0.00	0.15	0.05	0.00	0.19	0.06	0.00
Percentage of English language (*english_language*)	0.16	0.10	0.00	0.14	0.10	0.00	0.14	0.09	0.00	0.17	0.08	0.00
Percentage of other language (*other_lang*)	0.02	0.02	0.45	0.02	0.02	0.25	0.02	0.02	0.74	0.02	0.02	0.30
Topic popularity index (*pop_score*)	34.64	34.35	0.36	34.78	34.11	0.03	34.68	34.13	0.08	35.07	33.82	0.01
Share of numbers (*share_numbers*)	0.025	0.028	0.00	0.025	0.028	0.00	0.026	0.028	0.00	0.027	0.027	0.97
Flesch-reading-ease score (*flesch_score*)	40.09	41.22	0.01	40.54	41.03	0.26	40.47	41.26	0.09	39.28	41.28	0.01
Meta information features												
Website size: Length (*text_length*)	75269.35	56746.84	0.00	71629.95	55685.73	0.00	71193.63	52859.37	0.00	75334.75	52462.63	0.00
Website size: Nr. of pages (*nr_subpages*)	30.37	24.65	0.00	28.75	24.87	0.00	28.92	23.75	0.00	31.23	23.58	0.00
Loading time (*load_time*)	0.57	0.55	0.69	0.51	0.60	0.25	0.55	0.57	0.76	0.51	0.49	0.57
Mobile version (*mobile_version*)	0.76	0.70	0.00	0.76	0.68	0.00	0.75	0.67	0.00	0.73	0.69	0.06
Domain purchase year (*domain_purchase_year_proxy*)	2004.22	2004.98	0.00	2004.42	2004.96	0.00	2004.37	2005.17	0.00	2004.38	2005.01	0.01
Network features												
Outgoing hyperlinks (*outgoing_links*)	15.93	12.95	0.00	15.18	12.97	0.00	15.19	12.46	0.00	16.23	12.38	0.00
Incoming hyperlinks (*incoming_links*)	14.78	5.22	0.00	13.24	4.30	0.00	12.11	4.09	0.00	12.09	3.70	0.00
Use of social media (*social_media*)	1.62	1.02	0.00	1.51	0.98	0.00	1.47	0.92	0.00	1.62	0.91	0.00
PageRank centrality (*pagerank_index*)	$2 * 10^{-6}$	$1 * 10^{-6}$	0.00	$2 * 10^{-6}$	$1 * 10^{-6}$	0.00	$1 * 10^{-6}$	$1 * 10^{-6}$	0.00	$1 * 10^{-6}$	$1 * 10^{-6}$	0.01
Bridges (*bridge_index*)	0.43	0.26	0.01	0.38	0.28	0.05	0.37	0.27	0.04	0.31	0.27	0.35
Number of observations	4,487			4,484			4,487			1,893		

Source: MIP 2019 and web-scraped data; Own calculations. All variables were rounded to the second decimal place except PageRank centrality, which was rounded to the sixth decimal place and share of numbers which was rounded to the third decimal place.

Differences exist for most variables. Looking at 'text' features, innovative firms are more likely to mention an emerging technology term and have more subpages in English language. The share of subpages in other languages, however, does not show any significant difference between both groups. Differences are also small for the share of numbers, our topic popularity index and for the Flesch-reading-ease score, but the deviation is statistically significant for some forms of innovation activity.

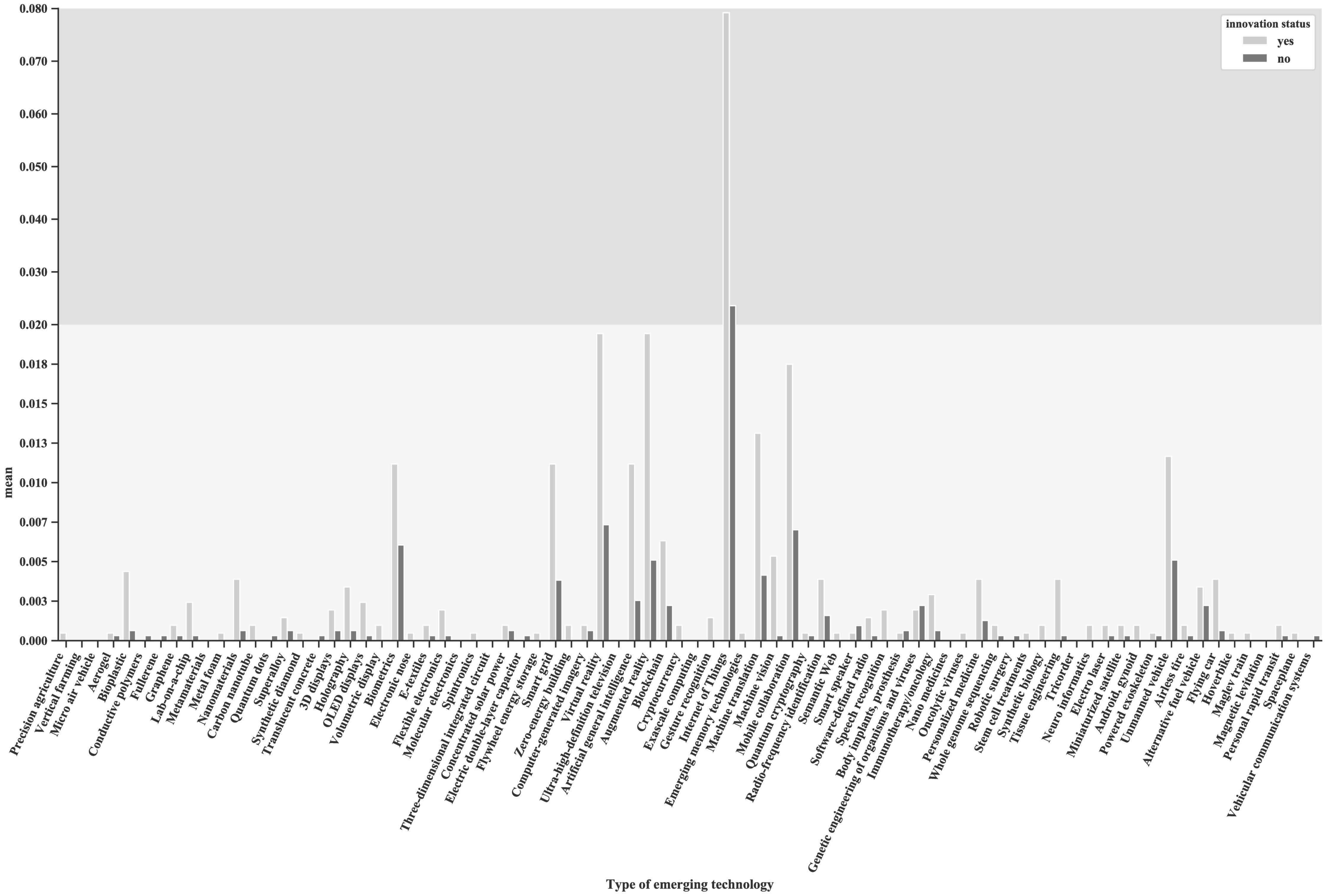
The descriptive statistics for 'meta' features show that innovative firms have larger websites with respect to the number of subpages as well as with respect to the number of characters. The loading time is slightly faster for process innovators and innovators, but not for product innovators and firms with innovation expenditures. However, differences are not statistically significant.

The first occurrence on web.archive.org is significantly later for non-innovative firms indicating their domain purchase year, i.e., website age is slightly lower. Additionally, non-innovative firms have less often a version of their website for mobile end user devices. Looking at 'network' features, significant differences also exist for outgoing and incoming hyperlinks as well as for hyperlinks to social media websites. Innovative firms have on average more hyperlinks. Moreover, the difference is larger for incoming than for outgoing or social media hyperlinks. Additionally, innovative firms also are significantly more important in firm networks looking at the PageRank centrality. The statistical significance of differences regarding the bridge index is, however, limited to the form of innovation activity. In summary, Table 3 confirms previous assumptions. Innovative firms seem more likely to apply emerging technologies, to have more technically advanced websites and to be better connected with each other according to most network indicators.

Fig 1 shows the average occurrence of different emerging technology terms on a firm website with respect to product innovation. The emerging technology terms differ strongly in their likelihood of occurrence. The emerging technology term *Internet of Things* is the most likely to occur. It appears on more than 8 percent of all product innovator websites and only on less than 2 percent of all non-product innovator websites. Also, terms relating to different machine learning applications, *biometrics*, *blockchain* technology and *mobile collaboration* appear relatively often. Moreover, for nearly every emerging technology term it is more likely to appear on a product innovator website than on a non-product innovator website. This result is the same for all innovation indicators.

Table 4 shows the ten most innovation-relevant LDA topics. The average highest absolute value of Pearson correlation coefficients between all four innovation indicators and the document-topic probabilities is used to identify the most relevant LDA topics. The topics are sorted in descending order. LDA topic 98,

Fig 1: Average occurrence of different emerging technology terms on firm websites with and without product innovations. Emerging technology terms not appearing on firm websites are not illustrated. The y-axis has a scale break at 0.02.



which relates according to its keywords to research & development, has a positive and by far the strongest relationship to innovation. Also, LDA topic 35, which relates to ICT infrastructure, has a comparatively strong positive correlation with our innovation indicators. Among the top 10, the LDA topics 20 (Tourism), 120 (Consulting & Costumer support) and 23 (Family business & craftsmanship) have the weakest correlation. Moreover, the correlation is negative.

Fig 2 also relates to the ten most innovation-relevant LDA topics. It shows for every topic the average share in a document for innovative and non-innovative firms. The figure reflects the results presented in Table 4. The selected topics considerably differ between innovative and non-innovative firms. Also, relation-

Table 4: Content of the LDA topics with the strongest relationship to MIP-based innovation indicators

Topic number	Content	Translated	Top words	Correlation*
LDA topic 98	Research & development	yes	'company' 'customer' 'development' 'to develop' 'department' 'employee' 'partner' 'project' 'successful'	positive (0.15)
LDA topic 35	ICT infrastructure	yes	'system' 'software' 'data centers' 'server' 'version' 'support' 'date' 'windows' 'automatic' 'document'	positive (0.10)
LDA topic 65	Construction	yes	'to build' 'project' 'new building' 'architect' 'planning' 'renovation' 'reconstruction' 'construction' 'to plan' 'architecture'	negative (-0.09)
LDA topic 134	Business software	no	'array' 'value' 'news' 'office' 'paket' 'error' 'data' 'page' 'SAP' 'search'	positive (0.08)
LDA topic 7	Product experience	no	'centro' 'company' 'best' 'use' 'experience' 'world' 'please' 'product' 'may' 'find'	positive (0.08)
LDA topic 41	Common terms	yes	'and' 'far' 'to take place' 'to put' 'frame' 'that' 'information' 'total' 'receive' 'department	negative (-0.07)
LDA topic 5	Carpentry	yes	'to tile' 'woods' 'to lay' 'laminate' 'tile' 'to put' 'material' 'stairs' 'floor' 'to glaze'	negative (-0.07)
LDA topic 20	Tourism	yes	'region' 'city' 'to be located' 'to offer' 'museum' 'old' 'historical' 'nature' 'tour' 'landscape'	negative (-0.06)
LDA topic 120	Consulting & costumer support	yes	'pleased' 'to offer' 'customer' 'to advise' 'individual' 'consulting' 'available' 'question' 'competent' 'to find'	negative (-0.06)
LDA topic 23	Family business & craftsmanship	yes	'company' 'to operate' 'visit' 'to stand' 'roofing' 'Michael' 'son' 'specialize' 'work'	negative (-0.06)

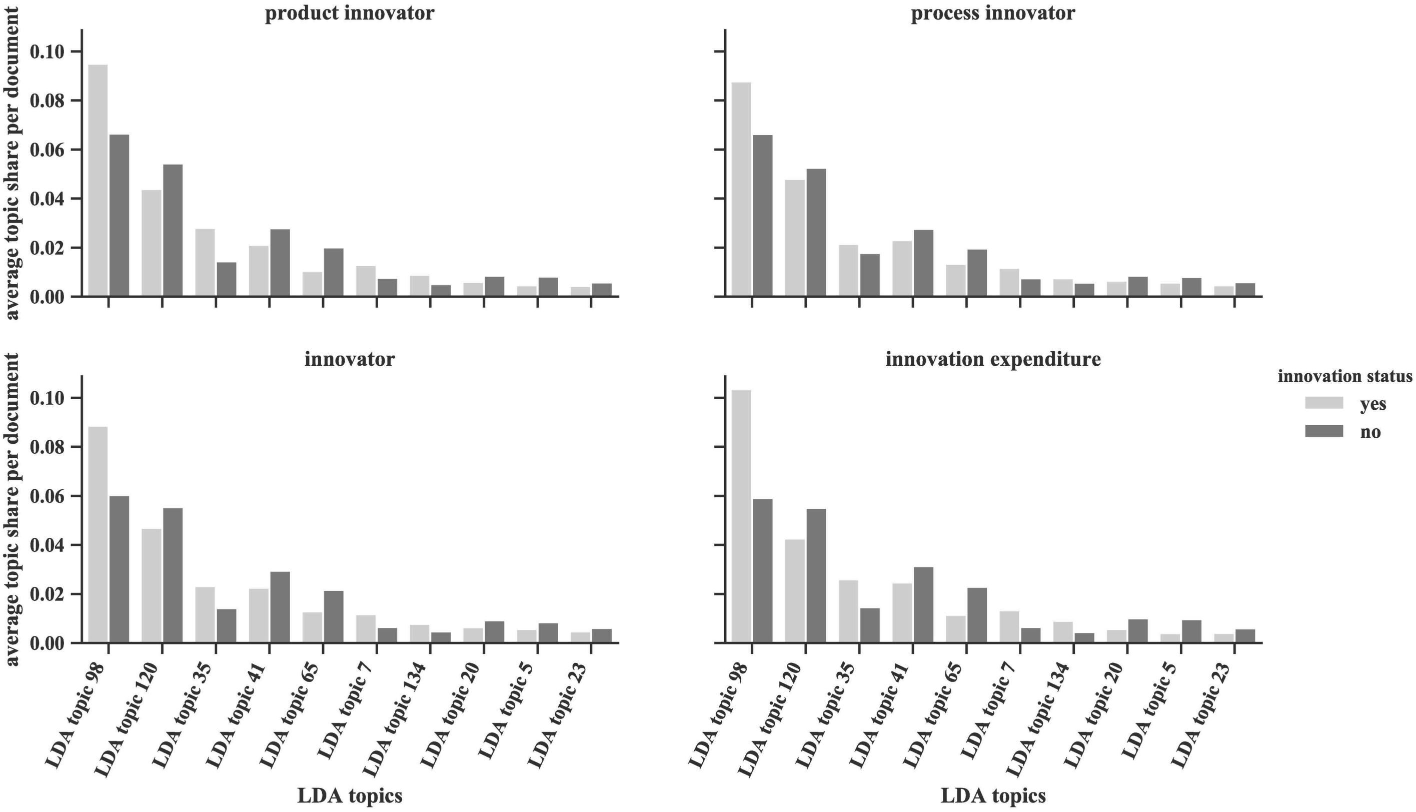
*Measured by the average of all Pearson correlation coefficients between the average topic share per document and each innovation indicator.

ships are constant, e.g., if a topic has a larger share on product innovator than on non-product innovator websites, it will also be relatively stronger represented on process innovator websites. Nonetheless, differences between innovation indicators exist. Average topic share differences diverge between indicators and are larger when considering firms' innovation expenditures than when taking product or process innovators into account.

5 Methodology

The objective of our work is the identification of website characteristics that allow predicting firm-level innovation activities. For this purpose, we integrate the described features as predictor variables in Random Forest classification models [39]. For each of our feature sets ('text', 'meta' and 'network' features) as well as for all features jointly a separate model is fitted. We use the Python package *scikit-learn* for the exercise. To evaluate the performance of the collected

Fig 2: Differences in the topic share of the top 10 topics with the strongest correlation with MIP-based innovation indicators on average.



website characteristics we use a baseline model. A random coin toss model based on the sample distribution is chosen. A baseline model works as a benchmark to assess the performance of more complex solutions, i.e., it helps to analyze whether a trained model performs better than a random prediction.

We use the metrics “area under the curve” (AUC), accuracy, improvement of accuracy in comparison to the baseline model, and the F1-score for positive as well as negative observations [40] to evaluate and compare models. AUC indicates the likelihood with which a model assigns a randomly selected innovative firm a higher probability of being innovative with a varying classification threshold. For the other metrics a classification threshold has to be set, which is for our models a probability of being innovative larger than 0.5. Based on this threshold, accuracy measures the fraction of all correctly predicted firms. The F1-score captures the harmonic mean between precision and recall for positive and negative observations respectively. In relation to a threshold as well, precision measures, for example, the share of correctly classified innovative firms in all firms classified as innovative, while recall measures the fraction of innovative firms that have been correctly identified as innovative. The same applies to non-innovative firms. Respective baseline outcomes of accuracy as well as F1-scores for our different innovation activity indicators are presented in Section 6. The random coin toss model assumes a fixed chance of being innovative (the sample mean). Hence, results do not change when varying the threshold and, therefore, the AUC value is not displayed for our baseline model.

Furthermore, we analyze four different innovation indicators (four different target variables), the predictive power of three different feature sets as well as their joint predictive power (in total four different groups of features). Accordingly, we train 16 Random Forest models. To analyze the performance of our out-of-sample prediction and to check for overfitting, we do not evaluate the models' performance with the observations that are already used for training: The data is split into a training sample (for fitting models) and into a test sample (for evaluating models). The training sample consists of 75 percent and the test sample consists of 25 percent of our observations. In the supervised learning context, this is a common partitioning method. It constitutes a trade-off between the generalization of the model and the validity of the evaluation. We also apply a gridsearch with 5-fold cross validation to tune the hyperparameters of all our models [39] for our training sample. We explore the parameter space for the number of trees (100, 500, 1,000, and 1,500), maximum tree depth (50, 100, 150, and 200), and minimum impurity decrease (0.01, 0.001). This leads to 32 different hyperparameter combinations for every model. We select the combination with the highest accuracy on the training sample for evaluation.

Random Forest models have the property that the feature importance can be easily measured, e.g., by the MDI [41], which is a split criterion to build single decision trees. For an overview of different split criterion measures, see [42]. The MDI is based on weighted impurity decreases evaluated at the node-level. For each node, it is calculated to what extent a variable will decrease the impurity of child nodes. The variable that leads to the best split, weighted by the cardinality of observations within each child node, will be selected. The average decrease for each variable, every time it is selected, is then calculated. This measure is known as the MDI. Feature importance is then derived by the respective MDI value divided by the sum of all MDI values. If multiple variables will lead to similar impurity decreases at one node, only one variable is selected for splitting. Hence, (multi-)collinearity of features can bias feature importance. This is obvious, for example, if we would include the same variable twice in a model. When choosing a variable for splitting, the model can randomly choose between the two and the feature relevance is thus divided between both variables.

6 Results

In this section, the predictions of MIP-based innovation indicators using a Random Forest classification approach are described. Table 5 shows evaluation metrics for all baseline as well as fitted models for different combinations of feature sets. Looking at product innovators, the highest AUC score (0.72) is realized with 'text' as well as 'all' features. The baseline accuracy is 0.53. The largest increase can be observed for the 'all' feature model (17 percentage points). Text-based features alone, however, lead to an increase of 16 percentage points. Moreover, 'network' and 'meta' features have a relatively weak impact. They just lead to improvements of 13 and 11 percentage points, respectively. This indicates that a large share of predictive power results from website text. The baseline F1-score for product innovators is 0.39 and for non-product innovators it is 0.61. Hence, the sample is slightly imbalanced towards non-product innovators and chances of randomly predicting this class correctly are higher. The F1-score shows a picture similar to other metrics. Only the 'text' feature model improves predictions notably and the F1-score for innovative firms is worse than a random prediction when only applying 'meta' or 'network' features. Also, the F1-score for non-innovative firms using 'all' features is lower than a random prediction. However, we find a very high performance (79 percent) for innovative firms (corresponding to an increase of 40 percentage points). This imbalanced result indicates that for this particular fitted model the classification threshold of 0.5 that a firm is innovative might be too low and should be adjusted when using this model for prediction. The 'text' feature model, by contrast, shows a more balanced increase in positive and negative F1-scores.

Our evaluation metrics for models predicting process innovators have predominantly a lower performance than those predicting the product innovator status. Nonetheless, models show better results as the process innovator baseline model. Hence, website characteristics still improve predictions. The best performance, in terms of accuracy, is reached with our 'all' feature model and leads to a performance increase of 10 percentage points. Moreover, 'meta' and 'network' features only perform slightly worse than 'text' features for all evaluation metrics.

The performance for innovators is slightly better than for process innovators in terms of AUC and accuracy. As the sample is slightly imbalanced towards innovators, this performance difference, however, is also partly related to different baseline values. Furthermore, similar to product innovators, we see a remarkably higher performance of models with 'text' features. Looking at F1-scores, predictions for the negative class always perform worse than a random prediction. In particular, predictions with only 'meta' or 'network' features lead to F1-scores of 0.0. This means that both models predict for every firm a likelihood that a firm

Table 5: Results for Random Forest classification models using different feature sets and target variables.

Feature sets					Accuracy		F1-Score		
Baseline	Text	Meta	Network	AUC	Value	Δ	Positive	Negative	Support
Product innovators									
x				-	0.53	-	0.39	0.61	1,122
	x			0.72	0.69	0.16	0.48	0.78	1,122
		x		0.66	0.64	0.11	0.37	0.75	1,122
			x	0.65	0.66	0.13	0.31	0.77	1,122
	x	x	x	0.72	0.70	0.17	0.79	0.51	1,122
Process innovators									
x				-	0.50	-	0.52	0.48	1,121
	x			0.62	0.59	0.09	0.63	0.54	1,121
		x		0.61	0.58	0.08	0.61	0.55	1,121
			x	0.59	0.58	0.08	0.62	0.53	1,121
	x	x	x	0.63	0.60	0.10	0.64	0.55	1,121
Innovators									
x				-	0.52	-	0.60	0.40	1,122
	x			0.67	0.63	0.11	0.75	0.31	1,122
		x		0.62	0.60	0.08	0.75	0.00	1,122
			x	0.62	0.60	0.08	0.75	0.00	1,122
	x	x	x	0.68	0.63	0.11	0.75	0.31	1,122
Innovation expenditures									
x				-	0.54	-	0.36	0.64	474
	x			0.74	0.72	0.18	0.55	0.80	474
		x		0.68	0.67	0.13	0.49	0.76	474
			x	0.64	0.65	0.11	0.17	0.78	474
	x	x	x	0.74	0.74	0.20	0.57	0.81	474

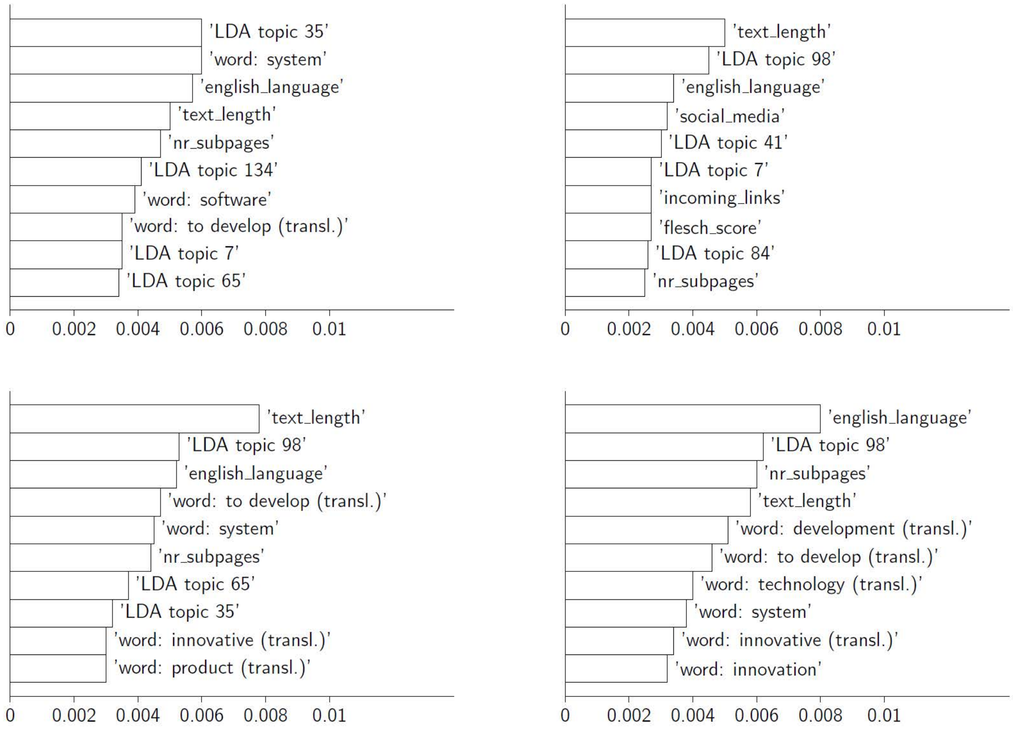
Source: MIP 2019 and web-scraped data; Own calculations. The results are rounded.

is innovative larger than 0.5 implying that all firms are classified as innovative, i.e., the model always predicts the majority class. This is known as zero rule prediction. For applying this rule, the information included in our baseline model is sufficient. In this regard, ‘meta’ and ‘network’ features do not provide information gains for innovators.

Even though the number of observations is the smallest, the predictive performance as well as the performance increase for firms with innovation expenditures is the highest in terms of AUC and accuracy. Looking at the ‘all’ feature model, firms with innovation expenditures can be predicted with an AUC and accuracy of 74 percent, which corresponds to a performance increase of 20 percentage points (with respect to accuracy). Similar to product innovators, the model solely based on ‘text’ features only performs slightly worse than the ‘all’ feature model. In addition, the F1-scores are always notably better than the baseline prediction, except for predicting the positive class with ‘network’ features.

In summary, it can be stated that the analyzed website characteristics show a better performance in the prediction of product innovators and firms with innovation expenditures than of process innovators. Moreover, text-based features show a greater relative relevance in comparison to other features for the first two indicators than for the latter. To compare the relevance of single features across feature sets, the ten most important predictor variables measured by the MDI are displayed in Fig 3 for each ‘all’ feature model respectively.

Fig 3: Feature importance values for ‘all’ feature models. Product innovators (top left), process innovators (top right), innovators (bottom left) and firms with innovation expenditures (bottom right) as target variable



Three features exist that always appear among the most relevant: The total number of characters, the number of subpages, and the share of English language. Looking in addition at the top 100 most relevant features (see S5 Appendix), reveals that further website characteristics exist with some general relevance. The words ‘worldwide’, ‘innovative’, ‘application’, ‘to develop’, ‘product’, ‘technology’ (all translated), the word ‘system’ as well as certain LDA topics, and the topic popularity index, incoming, outgoing as well as social media hyperlinks, the Flesch-reading-ease score, the loading time of a website, and the share of numbers are among the 100 most relevant features for every indicator. This shows that particular website characteristics exist, which have some relevance across indicators. In contrast, it is also noteworthy that features exist that show a large difference in the descriptive statistics but seem less important when predicting the innovation status. For example, neither the bridge index nor the emerging technology term dummy appear among the top 10 features for any indicator and are also not frequently observed among the top 100 features. Furthermore, some features are comparatively more relevant for certain innovation indicators. For instance, IT-related features seem to be highly relevant for product innovators. The IT-related LDA topics 35 ("ICT infrastructure") and 134 ("business software") as well as the words software and system are only among the top 10 features for this indicator.

On the contrary, the research & development related LDA topic 98 is more important when estimating process innovators and firms with innovation expenditure. Besides, the LDA topic 65 occurs in Fig 3 for product innovators and innovators, which should be related to a negative relationship to innovation activity, as the descriptive statistics show that this LDA topic is more likely to appear on websites of firms with no innovation activity. With respect to process innovators, it should be mentioned that no single word can be found in the 10 most important features and it is the only indicator that has 'network' features among its top 10. Furthermore, it is also interesting that the bottom left part of Fig 3, which relates to innovators, is at least for most features a combination of the most relevant features for product and process innovators. Last but not least, research & development related words are highly important in the prediction of firms with innovation expenditures.

7 Discussion

Descriptive statistics as well as our fitted Random Forest models show that website characteristics are relevant predictors for firm-level innovation activity. We see a significant difference in means between innovative and non-innovative firms for most of our features. For each innovation indicator, AUC and accuracy of the 'all' feature model are always higher than a weighted random coin toss. This property proves that our statistical models could actually learn from the data. Also, our results are in line with [30]. Their statistical model has reached a similar accuracy for product innovators only observed in one MIP wave.

Our exercise also reveals – especially when predicting product innovators and firms with innovation expenditures – that 'text' features are relatively more important than 'meta' and 'network' features. Besides, we see a pattern regarding the most important characteristics independent of different target variables: Across indicators, the total number of characters, the number of subpages and the share of English language belong always to the most relevant. It is also noteworthy that these features are more important than the word "innovative". This finding suggests that website size and language should be considered for different types of website-based innovation indicators, which has not been done in previous studies. Meeting expectations, features that show insignificant differences in Table 3 almost never belong to the top 10 most relevant features in Figure 3. An exception is the *flesch_score* in the case of process innovators. Furthermore, considering the poor performance of the 'meta' feature models and the result that 'text' is the most relevant feature set, the relevance of website size is quite counter-intuitive. One has to consider, however, that the importance of features is considered separately. The relevance of, e.g., the number of subpages is compared to the relevance of single words. If all words appearing in the term-document matrix would be considered jointly instead, their aggregated relative relevance would lie between 74 and 77 percent, depending on the indicator. This perspective illustrates why 'text' features and in particular textual content are still much more important for an accurate prediction. Nonetheless, as explained before, relative MDI importance should always be considered cautiously as it is affected by multicollinearity. Other web-based features may exist that possess predictive power and have not been considered in our analysis. These features would most likely change the result. Furthermore, it would also impact relative MDI importance, if this study's website data would be complemented with information from other sources, for example, non-web data from the MUP. In this case, innovation activity could potentially be predicted more accurately. However, we have deliberately decided against adding non-web data to our analysis, since this study focuses on the comparison of website information, which is up-to-date and freely accessible for everyone. Nonetheless, it would certainly be interesting to investigate in a further study the effect of adding additional non-web data. For potentially relevant features, see [31].

Another aspect that we want to emphasize is the fact that features which are only highly important for one indicator usually relate to its form of innovation activity. We see this as a strong indication that models use relevant information. Especially for firms with innovation expenditures, the selected word-based features appear particularly convincing. Terms like "to develop" (transl.) and "technology" (transl.) are highly ranked and have a very strong and direct connection to research & development expenditures. Another example is that the product experience LDA topic 7 has only a high importance for product innovators. Additionally, the 10 most relevant features of product innovators have a clear focus on information and communication (ICT) technologies, which is in line with the innovation spawning characteristic of ICT as well as the result of [43]. They find that ICT investment intensity is positively associated with innovation and stronger linked to product than to process innovation. Moreover, firms have a great incentive to present new products on their websites, but process innovations are often kept secret because this provides advantages over competitors. This might explain why results show a better predictive performance for product innovators than for process innovators and for innovators in general. In addition, no single word appears among the 10 most relevant features of process innovators and 'text' features alone do not lead to notably better predictions than 'meta' and 'network' features. This result supports the assumption that process innovations are often not mentioned explicitly but other patterns like the position of a firm within a hyperlink network or text complexity (*flesch_score*) include implicit information about the process innovation status. Regarding innovators, most of its top 10 features either appear in the product or process innovator ranking and the predictive performance of the 'all' feature model lies between both as well. This result meets our expectations as the innovator target variable is a combination of product and process innovators. Besides, 'meta' and 'network' features alone do not perform better than a weighted random coin toss for this indicator. This could be because these features relate differently to product and process innovators.

Interesting is also the fact that, contrary to our expectations, some features are not relevant. Even though the descriptive statistics show a large difference between innovative and non-innovative firms, the bridge index as well as the emerging technology dummy do not seem to be very decisive for predictions. Looking at the Pearson correlation coefficients between these terms and all other features reveals that both variables have a comparatively strong relationship with other features. Hence, their relative MDI importance is probably ranked lower due to multicollinearity. Besides, even though the descriptive statistics do not show a significant difference for every form of innovation activity, the Flesch-reading-ease score, the loading time of a website, and the share of numbers appear to be relevant for every indicator (according to the 100 most relevant features). These features, however, do not relate strongly to other features and might, therefore, provide some extra information. Hence, they are relatively relevant despite small differences.

Although we show a clear link between website characteristics and innovation status, the predictive performance of our models leaves room for improvement as we, for example, still misclassify the existence of innovation expenditures for 26 percent of the firms. Predictions might perform slightly better when applying neural networks. Our main criteria for choosing a Random Forest approach were the explainability of results and the fact that nonlinear relationships can be learned. Neural networks unfortunately do not offer a direct possibility to disclose decision processes. Hence, there is a trade-off, which often occurs in practice, between performance and explainability. If explainability is not necessary, predictive performance can most likely be improved by neural networks.

Within our sample, there can be of course also innovative firms that do not mention their innovation activity (implicitly or explicitly) on their website. In other words, some inaccuracy might relate to the nature of our data. In particular, product innovators, process innovators and innovators might suffer from noise as they cover a three year span. Websites can change a lot during this period. Comparatively good results for firms with innovation expenditures could be explained by the fact that this data is observed on an annual basis. Solving this matching problem seems to us a necessary step to improve predictions. Nonetheless, text data is always noisy and models with perfect accuracy are almost never identified. Furthermore, it could be criticized that website-based innovation indicators can only be applied to firms that have a website. Another point of criticism would be that it could cause noise if for marketing purposes firms falsely claim on their website that they are innovative. The MIP contains self-reported data as well, however, firms might not have the incentive to make false declarations as answers should not affect their public image. For this reason, we expect MIP data to reveal the actual innovation status and we consider the usage of MIP-based information as target variables as a solution to that problem. Besides, patent data could have also been used as an alternative target variable. However, patent-based indicators suffer from large time lags and rather measure inventions than innovations.

8 Conclusion

Firm-level innovation activity is often measured with data from large-scale questionnaire-based surveys. Resulting indicators, however, often lack, e.g., regional granularity and timeliness. Drawbacks could potentially be solved by adding web-based information. However, little is known yet about the accuracy and relevance of different website characteristics for measuring innovation. By exploiting a wide range of statistical learning and text mining techniques to predict firm-level innovation activities we contribute to the discussion on whether web-based innovation indicators are a feasible alternative to survey-based indicators.

In our analysis, data on 4,487 German firms surveyed in the MIP 2019 is used. We construct four different target variables from the MIP questionnaire to predict innovation activity: We use information on whether a firm has been a product innovator, process innovator or a combination thereof within the last three years and whether a firm reported innovation expenditures in the last year. We extract website texts, additional website-related meta information as well as hyperlinks of these firms. Several methods such as keyword search, network analysis tools and unsupervised learning techniques (LDA topic modelling) are applied to capture a wide range of different website characteristics. After generating a variety of features, we divide these into three different sets – ‘text’, ‘meta’ and ‘network’ features – and compare their performance with regard to every innovation indicator. A descriptive analysis already shows significant differences between innovative and non-innovative firms for most of the generated features. The Random Forest algorithm is chosen as it usually has a comparatively high predictive performance and the decision algorithm is comprehensive. Our results show that website characteristics unambiguously relate to MIP-based innovation indicators as the predictive performance notably improves in comparison to baseline values. The results also show that website characteristics increase performance for predicting product innovators and firms with innovation expenditures stronger than for process innovators. ‘Text’ features, in particular, appear to be decisive for the prediction of the first two indicators. Nonetheless, combining all feature sets shows the highest performance for every indicator. In addition to this, the comparison of single variables across feature sets indicates that for every indicator the language of a website as well as the website size belong to the most relevant features. Furthermore, there are some features that are only highly important for the prediction of a specific indicator that usually fits to its form of innovation activity, like the words “development” and “to develop” for innovation expenditures.

In summary, website characteristics seem more suitable to now-cast product innovators and firms with innovation expenditures than process innovators. This is most likely due to the fact that the latter group has a smaller incentive to announce innovation activity because new processes are less relevant for most website visitors. Hence, website characteristics seem to be rather suitable in measuring only certain aspects of innovation. In addition to this, the importance of certain website characteristics varies between indicators. Accordingly, different features should be taken into account depending on the kind of innovation activity that is analyzed.

Lastly, our work and related studies show that state of the art web-based predictive modeling cannot fully replace traditional surveys as error rates are still quite high. However, our models provide information about innovation activities that can be quickly updated, are on a very granular level (firm-level), and are less expensive than surveys. These results may be of particular interest for researchers as well as policy makers.

Acknowledgments

The authors would like to thank the German Federal Ministry of Education and Research for providing funding for the research project (TOBI - Text Data Based

Output Indicators as Base of a New Innovation Metric; funding ID: 16IFI001). We also thank Irene Bertschek, Reinhold Kesler, Christian Rammer, Bettina Schuck, Dominik Rehse, Thomas Niebel and Tobias Gließner for valuable inputs.

References

1. OECD/Eurostat. Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation, 4th ed. The Measurement of Scientific, Technological and Innovation Activities. Paris/Eurostat, Luxembourg; OECD Publishing. 2019.
2. Solow RM. Technical change and the aggregate production function. The Review of Economics and Statistics. 1957; 39(3): 312–320.
3. Hall BH, Jaffe A, Trajtenberg M. Market value and patent citations. The RAND Journal of Economics. 2005; 36(1): 16–38.
4. Crepon B, Duguet E, Mairesse J. Research, innovation and productivity: An econometric analysis at the firm level. Economics of Innovation and New Technology. 1998 Jan; 7(2): 115–158.
5. Kogan L, Papanikolaou D, Seru A, Stoffman N. Technological innovation, resource allocation, and growth. The Quarterly Journal of Economics. 2017 May; 132(2): 665–712.
6. Griffith R, Huergo E, Mairesse J, Peters B. Innovation and productivity across four European countries. Oxford Review of Economic Policy. 2006 Dec; 22(4): 483–498.
7. Belderbos R, Carree M, Lokshin B. Cooperative R&D and firm performance. Research Policy. 2004 Dec; 33(10): 1477–1492.
8. Klomp L, Van Leeuwen G. Linking innovation and firm performance: A new approach. International Journal of the Economics of Business. 2001; 8(3): 343–364.
9. Frenz M, Ietto-Gillies G. The impact on innovation performance of different sources of knowledge: Evidence from the UK Community Innovation Survey. Research Policy. 2009 Sep; 38(7): 1125–1135.
10. Rammer C, Behrens V, Doherr T, Krieger B, Peters B et al. Innovationen in der deutschen Wirtschaft: Indikatorenbericht zur Innovationserhebung 2019. ZEW Innovationserhebungen-Mannheimer Innovationspanel (MIP); 2019. Available from: http://ftp.zew.de/pub/zew-docs/mip/19/mip_2019.pdf.

11. Peters B, Rammer C. Innovation panel surveys in Germany. In: Fred Gault, editors. Handbook of Innovation Indicators and Measurement. Edward Elgar Publishing; 2013. pp.135-177.

12. Mairesse J, Mohnen P. Using innovations surveys for econometric analysis. In: Hall, B. H. and N. Rosenberg, editors. Handbook of the Economics of Innovation. vol. 2. Amsterdam and New York; Elsevier. 2010. pp. 1129-1155.

13. Kinne J, Axenbeck J. Web mining of firm websites: A framework for web scraping and a pilot study for Germany. Scientometrics. 2020: 1–31.

14. Pukelis L, Stanciauskas V. Using internet data to compliment traditional innovation indicators. [Preprint] 2019 [posted 2019 June; cited 2020 Oct 1]. Available from: `https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf`

15. Archibugi D, Planta M. Measuring technological change through patents and innovation surveys. Technovation. 1996 Sept; 16(9): 451–468, 519.

16. Arundel A, Kabla I. What percentage of innovations are patented? Empirical estimates for European firms. Research Policy. 1998 June; 27(2): 127–141.

17. Gök A, Waterworth A, Shapira P. Use of web mining in studying innovation. Scientometrics. 2015; 102(1): 653–671.

18. Kelly B, Papanikolaou D, Seru A, Taddy M. Measuring technological innovation over the long run. NBER Working Paper No. w25266. [Preprint] 2018 [posted 2018 Nov; revised 2020 Feb; cited 2020 Oct 1]. Available from: `https://www.nber.org/papers/w25266`.

19. Lenz D, Winker P. Measuring the diffusion of innovations with paragraph vector topic models. PLOS ONE. 2020 Jan; 15(1): e0226685.

20. Tacchella A, Napoletano A, Pietronero L. The language of innovation. PLOS ONE, 2020 Apr; 15(4): e0230107.

21. Bellstam G, Bhagat S, Cookson JA. A Text-Based Analysis of Corporate Innovation. Management Science; Forthcoming. doi: 10.1287/mnsc.2020.3682.

22. Gentzkow M, Kelly B, Taddy M. Text as data. Journal of Economic Literature. 2019; 57(3): 535–574.

23. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232): 1012–1014.

24. Choi H, Varian H. Predicting the present with Google Trends. Economic Record. 2012; 88: 2–9.

25. Katz JS, Cothey V. Web indicators for complex innovation systems. Research Evaluation. 2006; 15(2): 85–95.

26. Ackland R, Gibson R, Lusoli W, Ward S. Engaging with the public? Assessing the online presence and communication practices of the nanotechnology industry. Social Science Computer Review. 2010; 28(4): 443–465.

27. Arora SK, Youtie J, Shapira P, Gao L, Ma T. Entry strategies in an emerging technology: A pilot web-based study of graphene firms. Scientometrics. 2013; 95(3): 1189–1207.

28. Beaudry C, Héroux-Vaillancourt M, Rietsch C. Validation of a web mining technique to measure innovation in high technology Canadian industries. In: CARMA 2016–1st International Conference on Advanced Research Methods and Analytics. 2016. pp. 1–25.

29. Nathan M, Rosso A. Innovative events. Centro Studi Luca d'Agliano Development Studies Working Paper (N. 429). [Preprint] 2017 [posted 2017 Dec; cited 2020 Oct 1]. Available from: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3085935`

30. Kinne J, Lenz D. Predicting innovative firms using web mining and deep learning. ZEW Discussion Paper (19-001). [Preprint] 2019 [posted 2019 Jan; revised 2019 Dec; cited 2020 Oct 1]. Available from: `http://ftp.zew.de/pub/zew-docs/dp/dp19001.pdf`

31. Gandin I, Cozza C. Can we predict firms' innovativeness? The identification of innovation performers in an Italian region through a supervised learning approach. PLOS ONE. 2019 June; 14(6): e0218175.

32. Bersch J, Gottschalk S, Müller B, Niefert M. The Mannheim Enterprise Panel (MUP) and firm statistics for Germany. ZEW Discussion Paper. 2014; (14-104). Available from: `http://ftp.zew.de/pub/zew-docs/dp/dp14104.pdf`

33. Kirbach M, Schmiedeberg C. Innovation and export performance: Adjustment and remaining differences in East and West German manufacturing. Economics of Innovation and New Technology. 2008; 17(5): 435–457.

34. Cassiman B, Golovko E. Innovation and internationalization through exports. Journal of International Business Studies. 2011; 42(1): 56–75.

35. Lachenmaier S, Wößmann L. Does innovation cause exports? Evidence from exogenous innovation impulses and obstacles using German micro data. Oxford Economic Papers. 2006; 58(2): 317–350.

36. Flesch R. A new readability yardstick. Journal of Applied Psychology, 1948. 32(3): 221–233.

37. Becker W, Dietz J. R&D cooperation and innovation activities of firms-evidence for the German manufacturing industry. Research Policy. 2004; 33(2): 209–223.

38. Bertschek I, Kesler R. Let the user speak: Is feedback on Facebook a source of firms' innovation? ZEW Discussion Paper (17-015). [Preprint] 2017 [posted 2017 March; revised 2020 Aug; cited 2020 Oct 1]. Available from: http://ftp.zew.de/pub/zew-docs/dp/dp17015.pdf

39. Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning. 1st ed. New York; Springer Series in Statistics 2001.

40. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27(8): 861–874.

41. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont, CA: Wadsworth International Group. 1984; 432: 151–166.

42. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Advances in neural information processing systems 26 (NIPS 2013). 2013; pp. 431–439.

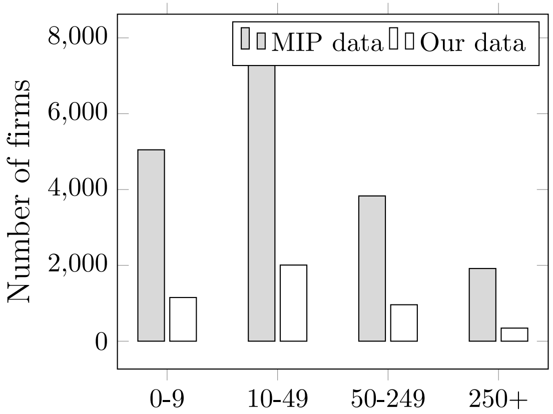
43. Hall BH, Lotti F, Mairesse J. Evidence on the impact of R&D and ICT investments on innovation and productivity in Italian firms. Economics of Innovation and New Technology. 2013; 22(3): 300–328.

44. Baeza-Yates R, Ribeiro-Neto B, et al. Modern Information Retrieval. vol. 463. New York: ACM press. 1999.

45. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning Research. 2003; 3(Jan): 993–1022.

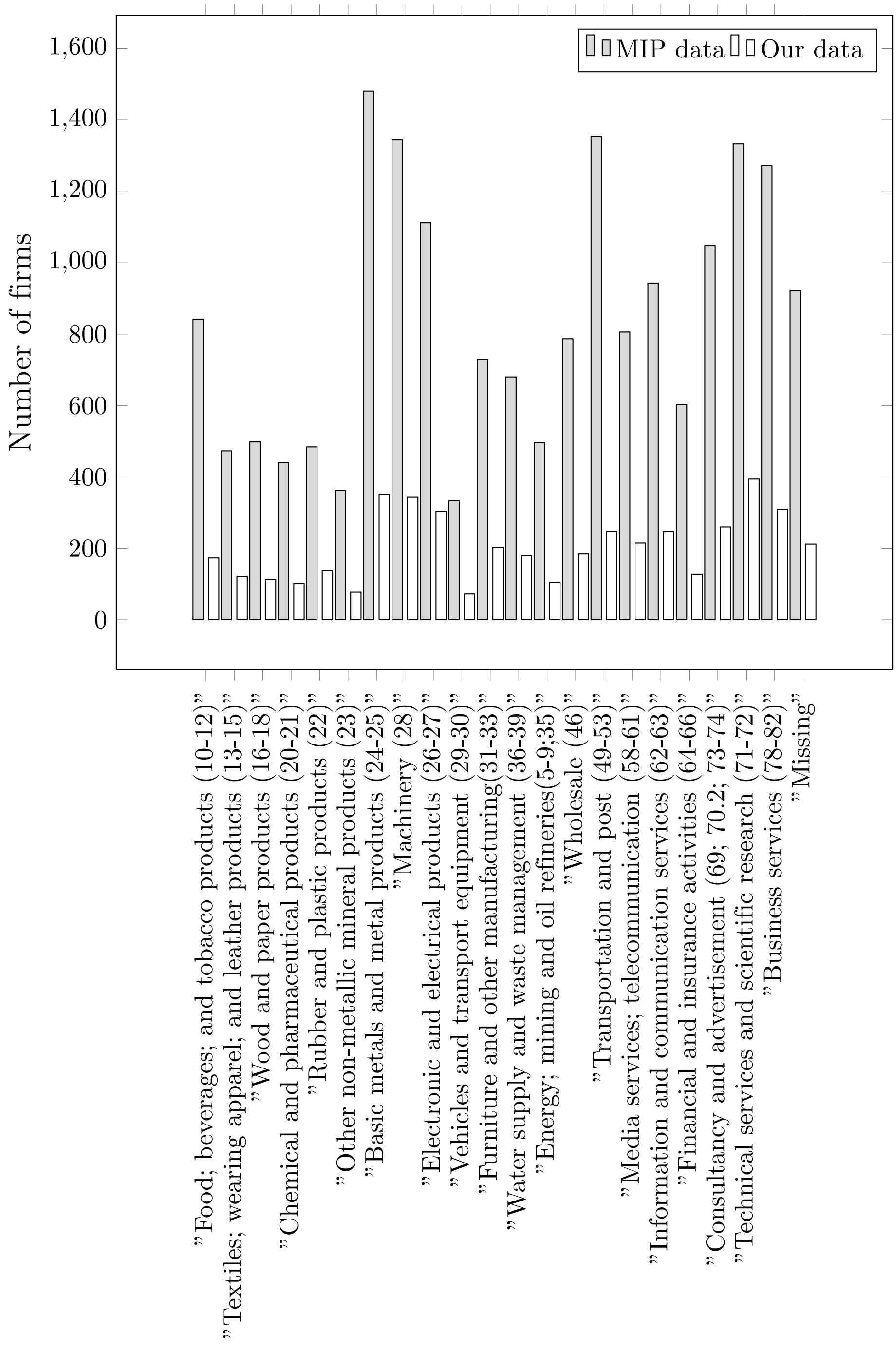
S1 Appendix

Fig 4: Firm distribution based on number of employees



S2 Appendix

Fig 5: Firm distribution for economic sectors based on 2 digit NACE codes.



S3 Appendix

English terms: Agricultural robot, closed ecological systems, cultured meat, precision agriculture, vertical farming, micro air vehicle, neural-sensing headset, four-dimensional printing, arcology, aerogel, bioplastic, conductive polymers, cryogenic treatment, fullerene, graphene, lab-on-a-chip, magnetorheological fluid, metamaterials, metal foam, multi-function structures, nanomaterials, carbon nanotube, quantum dots, superalloy, synthetic diamond, translucent concrete, 3D displays, ferroelectric liquid crystal display, holography, interferometric modulator display, laser video displays, OLED displays, micro LED displays, telescopic pixel display, time-multiplexed optical shutter, volumetric display, biometrics, digital scent technology, electronic nose, e-textiles, flexible electronics, memristor, molecular electronics, nano electro mechanical systems, spintronics, thermal copper pillar bump, three-dimensional integrated circuit, concentrated solar power, electric double-layer capacitor, flywheel energy storage, grid energy storage, home fuel cell, lithium iron phosphor battery, lithium-sulfur battery, magnesium battery, nanowire battery, ocean thermal energy conversion, smart grid, vortex engine, wireless energy transfer, zero-energy building, computer-generated imagery, virtual reality, ultra-high-definition television, 5G cellular communications, artificial general intelligence, augmented reality, blockchain, carbon nanotube field-effect transistor, civic technology, cryptocurrency, exascale computing, gesture recognition, internet of things, emerging memory technologies, emerging magnetic data storage technologies, fourth generation optical discs, holographic data storage, general purpose computing on graphics processing units, exocortex, machine translation, machine vision, mobile collaboration, nano radio, optical computing, quantum computing, quantum cryptography, radio-frequency identification, semantic web, smart speaker, software-defined radio, speech recognition, subvocal recognition, hybrid forensics, body implants, prosthesis, cryonics, de-extinction, genetic engineering of organisms and viruses, suspended animation, artificial hibernation, immunotherapy/oncology, nano medicines, nano sensors, oncolytic viruses, personalized medicine, whole genome sequencing, robotic surgery, stem cell treatments, synthetic biology, synthetic genomics, tissue engineering, tricorder, brain-computer interface, neuro informatics, electro encephalography, neuro prosthetics, caseless ammunition, directed energy weapon, electro laser, electromagnetic weapons, electrothermal-chemical technology, green bullet, laser weapon, particle beam weapon, sonic weapon, stealth technology, vortex ring gun, wireless long-range electric shock weapon, artificial gravity, stasis chamber, inflatable space habitat, miniaturized satellite, android, gynoid, nanorobotics, powered exoskeleton, self-reconfiguring modular robot, unmanned vehicle, airless tire, alternative fuel vehicle, electro hydrodynamic propulsion, flying car, fusion rocket, hoverbike, jetpack, backpack helicopter, maglev train, vactrain, magnetic levitation, mass driver, personal rapid transit, physical internet, scooter-sharing system, propellant depot, reusable launch system, space elevator, spaceplane, supersonic transport, vehicular communication systems

German terms: Agrarroboter, geschlossenes ökologisches System, Zuchtfleisch, Präzisionslandwirtschaft, vertikale Landwirtschaft, Mikro-Luftfahrzeug, neuronales Headset, vierdimensionales Drucken, Arkologie, Aerogel, Bio-Kunststoff, leitfähige Polymere, kryogene Behandlung, Fulleren, Graphen, Labor auf einem Chip, magnetorheologische Flüssigkeit, Metamaterialien, Metallschaum, Multifunktionsstrukturen, Nanomaterialien, Kohlenstoffnanoröhre, Quantenpunkte, Superlegierung, synthetischer Diamant, durchsichtiger Beton, 3D-Display, ferroelektrische Flüssigkristallanzeige, Holographie, interferometrische Modulatoranzeige, Laser-Video-Display, OLED Display, Mikro-LED Display, Teleskop-Pixelanzeige, zeitgemultiplexter optischer Verschluss, volumetrische Anzeige, Biometrie, digitale Dufttechnologie, elektronische Nase, E-Textil, flexible Elektronik, Memoristor, molekulare Elektronik, nanoelektromechanisches System, Spintronik, Thermo-Kupfer-Säulen-Stoß, dreidimensionale integrierte Schaltung, konzentrierte Solarenergie, elektrischer Doppelschicht-Kondensator, Schwungradspeicherung, Speicherung von Netzenergie, Heim-Brennstoffzelle, Lithium-Eisen-Phosphor-Batterie, Lithium-Schwefel-Batterie, Magnesium-Batterie, Nanodraht-Batterie, Ozean-Thermische Energieumwandlung, intelligentes Netz, Vortex-Motor, drahtlose Energie-Übertragung, Nullenergiehaus, computergeneriertes Bild, virtuelle Realität, hochauflösendes Fernsehen, 5G zellulare Kommunikation, künstliche Intelligenz, erweiterte Realität, Blockchain, Kohlenstoffnanoröhren-Feldeffekttransistor, zivile Technik, Kryptowährung, Exascale-Computing, Gestenerkennung, Internet der Dinge, neue Speichertechnologie, neue magnetische Speichertechnologie, optische Platten der vierten Generation, holografischer Speicher, allgemeines Rechnen auf Grafikprozessoren, Exokortex, maschinelle Übersetzung, maschinelles Sehen, mobile Zusammenarbeit, Nano-Funk, optische Datenverarbeitung, Quantencomputer, Quantenkryptographie, Radiofrequenz-Identifikation, semantisches Web, intelligenter Lautsprecher, Software-definiertes Radio, Spracherkennung, subvokale Erkennung, Hybrid-Forensik, Körper-implantat, Kryonik, Wiederbelebung ausgestorbener Tierarten, Gentechnik, verzögerte Reanimation, künstlicher Winterschlaf, Immuntherapie/-onkologie, Nanomedizin, Nanosensoren, onkolytische Viren, individualisierte Medizin, whole genome sequencing, Roboterchirurgie, Stammzellentherapie, synthetische Biologie, synthetische Genomik, Gewebezüchtung, Tricorder, Gehirn-Computer-Schnittstelle, Neuroinformatik, Elektroenzephalographie, Neuroprothetik, hülsenlose Munition, gerichtete Energiewaffe, Elektro-Laser, elektromagnetische Waffen, elektrothermisch-chemische Technologie, grünes Geschoss, Laser-Waffe, Strahlenwaffe, Schallwaffe, Tarntechnologie, Wirbelringkanone, Elektroschockwaffe, künstliche Schwerkraft, Stasiskammer, aufblasbares Weltraum-Habitat, Miniatursatellit, Android, Nanorobotik, Exoskelett, selbstkonfigurierender Roboter, unbemanntes Fahrzeug, luftlose Reifen, Fahrzeug mit alternativen Kraftstoffen, Elektrohydrodynamischer Antrieb, Fluidik, Fusionsrakete, Schwebefahrrad, Jetpack, Rucksackhelikopter, Magnetschwebebahn, Vactrain, magnetische Schwebetechnik, Massenantrieb, Personal Rapid Transit, physisches Internet, Roller-Sharing-System, fliegendes Treibstofflager, wiederverwendbares Startsystem, Raumaufzug, Raumflugzeug, Überschalltransport, Fahrzeugkommunikationssystem.

S4 Appendix

Text-based features:

1) **Texts** – To identify the most relevant terms when predicting a firm's innovation status, we transform the scraped texts into a format that allows to do mathematical operations: We convert the website texts into a term-document matrix (e.g., [44], [45]), which is a matrix that counts the frequency of terms that occur in a collection of documents (websites in this particular case). Every column represents a document and a row represents a word from a predefined vocabulary space. Accordingly, every cell counts how often a particular word appears in a particular document. We define our vocabulary space as the 5,000 most frequent words in our entire training text corpus. Before we calculate the term-document matrix, we conduct the following preprocessing steps. First, we merge all scraped subpages related to a single firm and delete irrelevant subpages (imprints, information about cookies or texts that are prescribed by law) by using the gold standard approach based on a supervised machine learning regression model, see [30]. Also, every word is converted into lower case and lemmatized by means of the Python package *spacy*. We exclude punctuation as well as English and German stop words (word lists are derived from the Python package *nltk*). Additionally, we manipulate the term-frequency counts by the TF-IDF scheme [44] as it usually improves predictions. Therefore, each document is tokenized and the term-document frequency is calculated by means of the *TfidfVectorizer* algorithm from *scikit-learn*.

2) **Emerging technology terms** – To capture firms that mention emerging technologies, we conduct a keyword search in which we calculate whether a technology from Wikipedia's list of emerging technologies (Retrieved from

https://en.wikipedia.org/ wiki/List_of_emerging_technologies accessed on August 16, 2018) appears on a firm's website using all subpages and the entire vocabulary as well as the Python package *regex*. We only search for a selection of technologies that are in a research, development, diffusion or commercialization stage, as it is a criterion for an innovation to be brought into use. A detailed list of all used keywords is provided in Appendix S2. The feature *emerging_tech* is a dummy variable that captures whether an emerging technology term appears on a firm website.

3) **Latent patterns** – Latent patterns on a website, which might reveal a firm's innovation status, are captured by the latent Dirichlet allocation model (LDA) (see [45]). The LDA algorithm assumes that a document consists of a set of topics, while every topic is a distribution of words. By linking each word in a document to a topic and iteratively improving assignments, the algorithm learns the distribution of topics in the text corpus as well as the distribution of words related to each topic. Moreover, after applying the LDA algorithm, the topic-document matrix shows how much every topic contributes to a document (website). We do not want our topic model to be exclusively valid for our sample. Hence, we calibrate our topics on a separate sample which consists of 32,276 websites of firms observed in the MUP 2019 but not in the MIP 2019. We apply the same text preprocessing to it as to our MIP sample, with two differences. First, we use a larger vocabulary space (15,000 most frequent words). Second, we do not manipulate word counts by means of the TF-IDF formula, but generate a TF-IDF stop word dictionary excluding words with a lower sum of TF-IDF scores within the LDA corpus than 3. The latter is applied to ensure that rather words that are characteristic for particular websites are considered. Also, to improve our model performance, we delete all words that appear less than 50 times and in more than 90 percent of all documents in the LDA corpus. We use the *TfidfVectorizer* to calculate the stop word dictionary. This dictionary as well as the *CountVectorizer* from *scikit-learn* is applied to generate a term-document matrix for our LDA sample. A term-document matrix for the MIP sample is calculated in the same manner. The Python package *scikit-learn* is used to train the LDA model. In the standard LDA approach, the number of topics needs to be defined. To solve this issue, we apply the grid-search technique to optimize the number of topics. For this, we use the *GridSearchCV* algorithm from *scikit-learn*. It is evaluated which model parameter combination leads to the best result according to the log likelihood. We conduct a grid-search over different values for the 'number of topics'-parameter as well as the document-topic prior. We try 200, 180, 150, 250 topics and values of 0.05, 0.1 for the document-topic prior. The optimal number of topics is 150, the highest log-likelihood is achieved with a document-topic prior of 0.1. After fitting the LDA model with the separate

sample, the topic distribution for each website in our MIP sample is predicted (*LDA topic*) and used in our Random Forest models, i.e., the predicted topic share in a document for each topic is used as a feature.

4) **Topic popularity index** – The topic popularity index is the sum of document-topic probabilities weighted by the relative frequency each topic appears in the entire text corpus (*pop_score*). A topic is considered to appear in a document if the document-topic probability is larger than 2%.

5) **Language classification** – The export orientation of a website might provide information about a firm's innovation status. English is worldwide the most widely spoken language by the total number of speakers. Therefore, it is quite likely that firms with international customers describe their products in English. We measure the share of subpages in English language, as well as all other languages except German to approximate the export orientation of a firm (*english_language*, *other_lang*). For the language classification of subpages, we apply the Python package *langdetect.*

6) **Share of numbers** – We also test whether the share of numbers in the total text length per document relates to the innovation status. The share was calculated by the ratio of digits within a string (document). For example, the text 'This book costs 500 dollars.' has a ratio of 3/28, i.e., 10.7 percent. The corresponding variable is named *share_numbers.*

7) **Flesch-reading-ease score** – The Flesch-Reading-Ease score is a metric used to assess the complexity of texts. The main idea for the index is that short words and short sentences are easier for readers to understand. The Python package *ReadabilityCalculator* (Retrieved from `https://pypi.org/project/ReadabilityCalculator/`) was used to calculate the score. The full definition can be found in [36] and the corresponding variable is named *flesch_score.*

Meta information features:

8) **Website size** – Approximating firm size might help to predict a firm's innovation status. For example, [13] show that the number of subpages correlates with firm size and larger firms tend to be more likly to implement an innovation. Hence, we use the number of subpages as a feature to predict a firm's innovation status (*nr_pages*). One problem related to this feature is that it is truncated at 50 subpages due to the scraping limit of the web-scraper. However, as only 1.5 percent of our observations exceed the scraping limit, we do not see a severe problem here. Moreover, we use a Random Forest model that selects cut-off points for splitting. Hence, it can cope with truncated features. We additionally analyze to what extent the number of characters per website (*text_length*), which might also relate to firm size, informs about the firm's innovation status.

9) **Loading time** – This feature serves as a proxy for a firm's hardware structure. A website's loading time (*load_time*) is determined by a http or https request. The time from sending the request until the arrival of the response is measured. Servers which are far away or which only process the requests slowly (e.g., due to bad hardware or an overload) have a higher loading time (in milliseconds). However, it should be noted that the IT infrastructure can also be outsourced to professional hosting firms. We retrieved the loading time by means of the the Python packages *requests* and *time*. The latter is a standard Python library.

10) **Mobile version** – For each website, it is retrieved whether a version for mobile end user devices exists. A Google API (`https://www.googleapis.com/pagespeedonline/v3beta1/mobileReady?url=http://`) is used to extract this information from the websites. The data is delivered as JSON object. Within the delivered data, the binary variable ”score” within the data structure “usability” is used (*mobile_version*). It indicates Google's mobile version passing score. The Python packages *json*, *mechanize*, *socket* and *urllib* are used for this exercise.

11) **Website age** – To determine the website age, we use web.archive.org. The website includes an Internet archive that allows to look at websites at earlier stages. We wrote a small program that automatically goes to web.archive.org and searches for the first entry of a particular website. This characteristic serves as a proxy for the digital age of a firm (*domain_purchase_year_proxy*). Our program uses the Python package *urllib*.

Network features:

12) **Centrality** – Relationships with other firms might also link to a firm's innovation status. If a firm is related to another firm, it is likely that the firm will refer on its website to it. Hence, to capture relationships with other firms, the sum of outgoing (*outgoing_links*) and incoming (*incoming_links*) hyperlinks to other firms is observed. Outgoing hyperlinks are measured by the number of external links on a firm website. We measure incoming hyperlinks by counting how often firms which are listed in the entire MUP refer to a particular firm. Additionally, a directed graph is constructed. Here, a vertex represents a firm and an edge a hyperlink from one firm to another. The Pagerank centrality measure is calculated with the Python package *igraph* (https://igraph.org) and the function ”pagerank”. The default parameters are used and the resulting variable is called *pagerank_index*.

13) **Social media** – The use of social media could also be correlated with the firm's innovation status. Therefore, the sum of hyperlinks to the websites Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub,

Flickr, and Vimeo is counted and used as another feature (*social_media*). This is calculated by means of *regex* again.

14) **Bridges** – An undirected graph is constructed, as well. A bridge is an edge of a graph whose removal increases the number of connected components. For each vertex, we count the number of times it is part of a bridge. The Python package *networkx* (https://networkx.github.io) and the function ”bridges” is used to calculate the bridges and the described measure. The resulting variable is named *bridge_index*.

S5 Appendix

Model	Top 100 most relevant features
Product innovators	’LDA topic 35’, ’word: system’, ’english_language’, ’text_length’, ’nr_subpages’, ’LDA topic 134’, ’word: software’, ’word: to develop (transl.)’, ’LDA topic 7’, ’LDA topic 65’, ’incoming_links’, ’LDA topic 105’, ’word: application (transl.)’, ’word: version’, ’word: test’, ’word: product (transl.)’, ’domain_purchase_year_proxy’, ’word: worldwide (transl.)’, ’LDA topic 98’, ’word: innovative (transl.)’, ’LDA topic 20’, ’word: innovative’, ’LDA topic 41’, ’social_media’, ’share_numbers’, ’word: automatically (transl.)’, ’word: technology’, ’flesch_score’, ’LDA topic 34’, ’outgoing_links’, ’word: technology (transl.)’, ’LDA topic 119’, ’LDA topic 127’, ’LDA topic 96’, ’emerging_tech’, ’LDA topic 97’, ’word: sensor’, ’pop_score’, ’word: development (transl.)’, ’LDA topic 78’, ’LDA topic 46’, ’word: support’, ’LDA topic 138’, ’LDA topic 75’, ’LDA topic 39’, ’LDA topic 38’, ’word: application (transl.)’, ’word: usage (transl.)’, ’LDA topic 60’, ’LDA topic 70’, ’LDA topic 103’, ’word: software development (transl.)’, ’LDA topic 101’, ’LDA topic 52’, ’load_time’, ’LDA topic 148’, ’word: to optimize (transl.)’, ’LDA topic 49’, ’LDA topic 117’, ’LDA topic 56’, ’LDA topic 128’, ’LDA topic 36’, ’LDA topic 53’, ’LDA topic 19’, ’word: digital’, ’word: interfaces (transl.)’, ’word: complex (transl.)’, ’LDA topic 113’, ’word: component (transl.)’, ’word: compact (transl.)’, ’word: user (transl.)’, ’LDA topic 0’, ’LDA topic 84’, ’LDA topic 143’, ’LDA topic 107’, ’LDA topic 5’, ’LDA topic 144’, ’LDA topic 104’, ’LDA topic 8’, ’LDA topic 51’, ’LDA topic 125’, ’word: production (transl.)’, ’LDA topic 114’, ’LDA topic 122’, ’LDA topic 26’, ’word: consulting (transl.)’, ’LDA topic 120’, ’LDA topic 32’, ’LDA topic 99’, ’LDA topic 16’, ’LDA topic 106’, ’word: product development (transl.)’, ’LDA topic 115’, ’LDA topic 140’, ’LDA topic 69’, ’LDA topic 15’, ’LDA topic 142’, ’LDA topic 13’, ’LDA topic 1’, ’LDA topic 63’
	transl: Translated from German to English language

Model	Top 100 most relevant features
Process innovators	’text_length’, ’LDA topic 98’, ’english_language’, ’social_media’, ’LDA topic 41’, ’LDA topic 7’, ’incoming_links’, ’flesch_score’, ’LDA topic 84’, ’nr_subpages’, ’outgoing_links’, ’LDA topic 75’, ’LDA topic 65’, ’word: product (transl.)’, ’word: to develop (transl.)’, ’word: technology (transl.)’, ’word: worldwide (transl.)’, ’word: system’, ’LDA topic 53’, ’LDA topic 106’, ’LDA topic 104’, ’LDA topic 127’, ’LDA topic 20’, ’LDA topic 57’, ’share_numbers’, ’LDA topic 140’, ’LDA topic 122’, ’LDA topic 103’, ’LDA topic 39’, ’word: innovative (transl.)’, ’LDA topic 6’, ’LDA topic 32’, ’LDA topic 12’, ’LDA topic 56’, ’load_time’, ’LDA topic 100’, ’LDA topic 120’, ’LDA topic 148’, ’LDA topic 31’, ’LDA topic 60’, ’LDA topic 36’, ’word: as well as (transl.)’, ’LDA topic 68’, ’LDA topic 35’, ’LDA topic 134’, ’LDA topic 22’, ’LDA topic 52’, ’LDA topic 64’, ’LDA topic 121’, ’LDA topic 133’, ’LDA topic 147’, ’pop_score’, ’LDA topic 99’, ’word: standard’, ’LDA topic 34’, ’word: successful (transl.)’, ’word: international’, ’LDA topic 19’, ’LDA topic 50’, ’LDA topic 101’, ’iso’, ’LDA topic 96’, ’LDA topic 23’, ’LDA topic 2’, ’LDA topic 1’, ’LDA topic 73’, ’word: challenge (transl.)’, ’LDA topic 5’, ’LDA topic 74’, ’LDA topic 43’, ’LDA topic 117’, ’LDA topic 114’, ’LDA topic 145’, ’LDA topic 89’, ’LDA topic 88’, ’LDA topic 93’, ’LDA topic 0’, ’LDA topic 109’, ’LDA topic 16’, ’LDA topic 83’, ’LDA topic 46’, ’LDA topic 79’, ’LDA topic 24’, ’LDA topic 146’, ’word: application (transl.)’, ’LDA topic 87’, ’LDA topic 82’, ’LDA topic 85’, ’word: process (transl.)’, ’LDA topic 81’, ’LDA topic 14’, ’LDA topic 61’, ’LDA topic 115’, ’LDA topic 125’, ’LDA topic 113’, ’LDA topic 129’, ’LDA topic 78’, ’LDA topic 105’, ’LDA topic 141’, ’LDA topic 90’
	transl: Translated from German to English language

Model	Top 100 most relevant features
Innovators	’text_length’, ’LDA topic 98’, ’english_language’, ’word: to develop (transl.)’, ’word: system’, ’nr_subpages’, ’LDA topic 65’, ’LDA topic 35’, ’word: innovative (transl.)’, ’word: product (transl.)’, ’LDA topic 84’, ’LDA topic 20’, ’word: worldwide (transl.)’, ’social_media’, ’LDA topic 41’, ’LDA topic 7’, ’LDA topic 31’, ’LDA topic 134’, ’flesch_score’, ’word: application (transl.)’, ’incoming_links’, ’word: development (transl.)’, ’domain_purchase_year_proxy’, ’LDA topic 75’, ’share_numbers’, ’LDA topic 127’, ’outgoing_links’, ’word: successful (transl.)’, ’LDA topic 94’, ’LDA topic 103’, ’LDA topic 100’, ’LDA topic 96’, ’LDA topic 106’, ’LDA topic 6’, ’LDA topic 50’, ’LDA topic 105’, ’LDA topic 102’, ’pop_score’, ’LDA topic 148’, ’LDA topic 2’, ’LDA topic 19’, ’LDA topic 140’, ’LDA topic 39’, ’LDA topic 71’, ’LDA topic 78’, ’LDA topic 53’, ’word: usage (transl.)’, ’word: to provide (transl.)’, ’LDA topic 43’, ’LDA topic 36’, ’load_time’, ’LDA topic 52’, ’LDA topic 147’, ’LDA topic 89’, ’LDA topic 144’, ’LDA topic 56’, ’LDA topic 82’, ’LDA topic 90’, ’word: innovative’, ’LDA topic 87’, ’LDA topic 138’, ’LDA topic 51’, ’LDA topic 122’, ’LDA topic 93’, ’LDA topic 85’, ’LDA topic 101’, ’LDA topic 11’, ’LDA topic 0’, ’LDA topic 23’, ’word: experience (transl.)’, ’LDA topic 120’, ’LDA topic 59’, ’LDA topic 5’, ’LDA topic 60’, ’LDA topic 27’, ’LDA topic 113’, ’LDA topic 13’, ’word: international’, ’LDA topic 12’, ’LDA topic 118’, ’LDA topic 69’, ’LDA topic 34’, ’LDA topic 61’, ’LDA topic 114’, ’word: software’, ’LDA topic 135’, ’emerging_tech’, ’LDA topic 28’, ’LDA topic 109’, ’LDA topic 133’, ’LDA topic 46’, ’word: as well as (transl.)’, ’LDA topic 67’, ’word: technology (transl.)’, ’LDA topic 123’, ’LDA topic 33’, ’word: complex (transl.)’, ’LDA topic 70’, ’LDA topic 24’, ’LDA topic 25’
	transl: Translated from German to English language

Model	Top 100 most relevant features
Innovation expend.	’english_language’, ’LDA topic 98’, ’nr_subpages’, ’text_length’, ’word: development (transl.)’, ’word: to develop (transl.)’, ’word: technology (transl.)’, ’word: system’, ’word: innovative (transl.)’, ’word: innovation’, ’word: international’, ’LDA topic 134’, ’incoming_links’, ’LDA topic 105’, ’word: application (transl.)’, ’LDA topic 36’, ’LDA topic 148’, ’word: worldwide (transl.)’, ’LDA topic 5’, ’word: product (transl.)’, ’outgoing_links’, ’LDA topic 100’, ’LDA topic 84’, ’LDA topic 7’, ’word: integration’, ’LDA topic 35’, ’LDA topic 20’, ’flesch_score’, ’LDA topic 106’, ’LDA topic 28’, ’load_time’, ’word: research (transl.)’, ’LDA topic 1’, ’LDA topic 65’, ’LDA topic 120’, ’domain_purchase_year_proxy’, ’LDA topic 104’, ’LDA topic 39’, ’LDA topic 125’, ’share_numbers’, ’social_media’, ’LDA topic 13’, ’LDA topic 109’, ’pop_score’, ’LDA topic 49’, ’word: process (transl.)’, ’LDA topic 6’, ’LDA topic 102’, ’word: support’, ’LDA topic 73’, ’LDA topic 140’, ’LDA topic 75’, ’LDA topic 41’, ’word: high (transl.)’, ’LDA topic 53’, ’LDA topic 67’, ’word: innovative’, ’LDA topic 57’, ’LDA topic 138’, ’LDA topic 94’, ’LDA topic 82’, ’LDA topic 56’, ’LDA topic 113’, ’LDA topic 88’, ’LDA topic 26’, ’LDA topic 59’, ’LDA topic 83’, ’LDA topic 14’, ’LDA topic 31’, ’word: high’, ’LDA topic 64’, ’LDA topic 81’, ’LDA topic 69’, ’LDA topic 130’, ’LDA topic 34’, ’word: complex (transl.)’, ’LDA topic 95’, ’LDA topic 128’, ’LDA topic 86’, ’word: automatically (transl.)’, ’word: to optimize (transl.)’, ’LDA topic 141’, ’LDA topic 80’, ’LDA topic 132’, ’word: workshop’, ’LDA topic 68’, ’LDA topic 16’, ’LDA topic 127’, ’LDA topic 101’, ’LDA topic 61’, ’word: new (transl.)’, ’word: process (transl.)’, ’LDA topic 24’, ’LDA topic 131’, ’LDA topic 96’, ’LDA topic 60’, ’word: management’, ’LDA topic 144’, ’LDA topic 22’, ’LDA topic 0’
	transl: Translated from German to English language

S6 Appendix

Table 6: Software packages. Details on used software packages.

Software	Version
python	3.7.3
igraph	0.8.2
json	2.0.9
langdetect	1.0.7
networkx	2.3
nltk	3.4.4
numpy	1.16.4
mechanize	0.4.3
pandas	0.24.2
ReadabilityCalculator	0.2.37
regex	2018.1.10
requests	2.22.0
sklean	0.21.2
spacy	2.2.4

Download ZEW Discussion Papers from our ftp server:

http://ftp.zew.de/pub/zew-docs/dp/

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

//
IMPRINT

ZEW – Leibniz-Zentrum für Europäische
Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European
Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de