// NO.19-063 | 12/2019

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[†+*] & Patrick Breithaupt[†*]

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

[*] Correspondence: janna.axenbeck@zew.de; Phone: +49 621 1235 – 188, patrick.breithaupt@zew.de; Phone: +49 621 1235 – 217

December 31, 2019

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. In this study, we use 4,485 German firms from the Mannheim Innovation Panel (MIP) 2019 to analyze which website characteristics are related to innovation activities at the firm level. Website characteristics are measured by several text 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 the website's language, the number of subpages, and the total text length. Moreover, our website characteristics show a better performance for the prediction of product innovations and innovation expenditures than for the prediction of process innovations.

Keywords: Text as data, innovation indicators, machine learning
JEL Classification: C53, C81, C83, O30

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, Tobias Gließner and Steffen Mayer for valuable inputs.

1 Introduction

Innovation, defined as the implementation of either new or significantly improved products, services or processes, brings vast benefits to consumers and businesses. In some cases it may even lead to the creation of new markets. In other words, technological progress is considered as a main driver of economic growth (Solow 1957). It is, therefore, a matter of public interest to analyze and understand innovation dynamics.

A prerequisite for this is to correctly measure firm-level innovation activities within an STI (science, technology and innovation) system. Using valid innovation indicators is crucial when analyzing related economic questions. Traditionally, firm-level innovation indicators are constructed with data from large-scale questionnaire-based surveys like the biennial European Community Innovation Survey (CIS) or the annual German Mannheim Innovation Panel (MIP), which is also the German contribution to the CIS.[1] These innovation indicators suffer, however, from some major drawbacks (i.e. Axenbeck & Kinne 2018, Pukelis & Stanciauskas 2019). The MIP, for example, surveys around 18,000 firms every year. This corresponds to only a fractional share of the total stock of German firms and therefore lacks granularity and coverage. Additionally, questionnaire-based surveys – especially on a large scale – are costly and lack timeliness. Furthermore, most surveys require firm participation. As a consequence, surveys like the MIP suffer from low response rates. Besides, firm-level innovation can also be studied by patent or publication analysis. However, respective indicators cover only technological progress for which legal protection is sought (Archibugi & Planta 1996).

Some of these issues, however, could be solved by adding web-based information: Nowadays, nearly every firm is represented on the Internet. Firm websites can entail information about new products, key personnel decisions, strategies, and relationships with other firms. Those pieces of information might be directly or indirectly related to a firm's innovation status. Advances in computing power, methods for statistical learning as well as natural language processing tools allow extracting website information on a large scale. Therefore, e.g., Gök et al. (2015) propose to complement traditional innovation indicators with information from scraped firm websites. Extracting this information might allow constructing innovation indicators that enable an automatized, timely, and comprehensive analysis of firm-level innovation activities, carried out faster and in shorter intervals in comparison to traditional indicators.

This paper contributes to the question whether website-based innovation indicators are feasible: Our objective is to answer whether firm websites contain human-interpretable information to measure innovation activities. Additionally, we analyze which characteristics of a website relate most to a firm's innovation status.

In most industries, firms might have a greater incentive to inform customers about new products than to disclose new processes as the latter could provide an advantage for competitors. Therefore, it is assumed

[1] The innovation definition in both surveys is based on the Oslo Manual (OECD/Eurostat 2018).

that firms have a higher incentive to report product than process innovation on their websites. We test this by analyzing the prediction performance difference for different innovation indicators related either to product innovations, process innovations or innovation expenditures. First, we analyze the predictive performance of website characteristics for innovative firms. Then, we test whether our predictions improve or worsen when we predict only product or process innovations. Additionally, we compare our results to the prediction of whether a firm has innovation expenditures.

Data of 4,485 German firms from the Mannheim Innovation Panel (MIP) 2019 is used. We extract their website's text and hyperlink structure by applying the ARGUS web-scraper (Kinne 2018). Several methods including topic modelling and natural language processing tools are applied to generate features that potentially relate to the firm-level innovation status. Furthermore, we extract further information from the websites like how fast the website is responding and whether there is a version for mobile end user devices available. After extracting and calculating a wide variety of features, we divide them into three different groups: text-based features, meta information features, and link features. Based on these three 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 significantly better than a random prediction. Consequently, firm websites entail human-interpretable information that relate to a firm's innovation activity. Looking at the most important characteristics measured by the "mean decrease in impurity", 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 regardless of the innovation indicator. Moreover, there are characteristics that are highly important only for certain indicators, e.g., the German word "entwickeln" (develop) for innovation expenditures. We also find, as expected, that our website characteristics better predict firms with product innovations and innovation expenditures than with process innovations.

The remainder of this paper is organized as follows: At first, previous literature is reviewed in Section 2. In Section 3, we present our data. Section 4 describes the empirical approach and Section 5 shows the results, which are discussed in Section 6. The work ends with a conclusion in Section 7.

2 Literature Review

Previous literature has already shown that information produced online can be used to construct frequent real-time estimates ('nowcasting') (Genzkow, Kelly & Taddy 2017). Famous 'nowcasting' examples that utilize web information are Ginsberg et al. (2009), who use Google search queries to accurately predict influenza activities in the United States. Choi & Varian (2012) claim that search engine query indices are often correlated with economic activities and allow to generate frequent indicators. They show that forecasts about, e.g., 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 firm websites provide detailed information about the firm as well as its products, they appear to be suitable for measuring firm-level innovation activities (Gök et al. 2015). For an in-depth literature review on web-based innovation indicators, see Kinne & Axenbeck (2018).

Following the idea of web-based innovation indicators, Kinne & Lenz (2019) attempt to predict innovation at the firm level using textual information on websites and novel machine learning tools. They use traditional firm-level innovation indicators from the MIP 2017 to train an artificial neural network classification model on labelled (innovative/non-innovative) web texts.

Pukelis & Stanciauskas (2019) fit several machine learning models to develop a web-based innovation indicator. Their annotated data set is limited to 500 firms. One of the most important characteristics of their work is the individual analysis of firm website subpages instead of predicting the innovation status of an entire website, i.e., firm. Additionally, their text data was manually labelled as either innovation or non-innovation related messages instead of using survey or patent data. The best performance is achieved with an artificial neural network.

An 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. 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. We attempt to fill this gap and try to identify which website characteristics are linked to the innovation status of a firm in order to provide new and detailed insights on the question whether firm websites entail measurable information about firm-level innovation activities.

3 Data

Based on the Oslo Manual, we define an innovation 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)"* (OECD/Eurostat 2018, p. 20). Furthermore, we consider all expenditures that were spent for innovation purposes - independent of the magnitude - as innovation expenditures and summarize firm-level product or process innovation as well as innovation expenditures as innovation activity.

We use data from the Mannheim Innovation Panel (MIP) 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. We chose the MIP 2019[2] wave for our analysis as it is until today the only MIP wave that entails information about the actual firm-level innovation status in 2018, the year website texts were firstly scraped. In the MIP 2019, firms were asked whether they introduced a product or process

[2] Status: 2.10.2019

innovation between the years 2016 and 2018 and if they had *innovation expenditures* in 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. We use 13,747 firms from MIP 2019 as our initial sample. We merge these firms with the Mannheim Enterprise Panel (MUP), which consists of more than 3.2 million economically active firms, to receive information about the firms' website addresses (URLs).[3] Unfortunately, only 54 percent of our sample can be assigned to websites, as we limit ourselves to quality-assured web addresses. Hence, we end up with 6,368 firms with information on the website address and at least one innovation indicator.

We extracted website content by applying the ARGUS web-scraper, which allows us to collect texts as well as hyperlinks. The websites' texts were first scraped in September 2018 and then again in January 2019 because hyperlinks were added. We scraped again in October 2019 to add information about technical features, e.g., regarding the existence of firm websites for mobile end user devices.[4] The maximum limit of scraped subpages per website was set to 50.[5] Otherwise the size of the data would have been too large. Moreover, the scraping program was set to prefer subpages with shorter URLs, because we assume these subpages include more important and rather general information about the firm. Also, ARGUS was set to prefer websites in German language. Hence, when we calculate the share of different languages on a website we expect a small bias.[6]

While scraping the data, especially while collecting the meta information features, we received several error messages. Furthermore, we only use observations for which all features are non-missing. If, for example, the meta information is not available the observation will not be used for training or testing. Therefore, after the entire data collecting process, we end up with 4,485 firms in our sample (Table 1).

As we need to exclude a large share of our observations due to missing values, we cannot rule out a selection bias. Also, firms from certain industries or 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 the groups that are overrepresented in our sample and it could induce that an overrepresented class is predicted more often than expected. 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 Figure A.1 & A.2 in the Appendix A.4 as well as A.5). 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. Moreover, in the current

[3] The MUP serves as a sampling frame for surveys like the MIP.
[4] Unfortunately it is not possible to collect the data retrospectively. Therefore, we have to accept the time discrepancy.
[5] As the medium number of subpages is 15 (Kinne & Axenbeck 2018), we assume this a sufficient amount.
[6] However, only 1.5 percent of the firms in our subsample have 50 or more subpages. That is why we assume that this bias is unproblematic.

MIP 2019 population, 44 percent of firms are product innovative, 59 percent are process innovative, 68 percent are innovators and 37 percent have innovation expenditures. These values also correspond to those in our subsample (see Table 1).

Table 1- Summary statistics for innovators, product innovators, process innovators as well as innovation expenditures

Variable	Definition	N	Mean	SD	Min	Max
Innovators	1: If firm is a product or/and process innovator 0: Otherwise	4,485	0.63	0.48	0	1
Product innovators	1: If firm is a product innovator 0: Otherwise	4,387	0.40	0.49	0	1
Process innovators	1: If firm is a process innovator 0: Otherwise	4,346	0.54	0.50	0	1
Innovation expenditures	1: If innovation expenditures were reported 0: Otherwise	1,891	0.39	0.49	0	1

To analyze which website characteristics relate to firm-level innovation activities, we explore three different groups of information sources on firm websites: text-based, meta information, and link features. To capture website characteristics, we apply several methods like a keyword search and natural language processing as well as an analysis of hyperlinks. All collected features are described below. We use Python as programming language for the calculation of our features as well as for training our Random Forest models.

Text-based features:

First, information from website *texts (1)* is analyzed (Table 2). This might relate to a firm's innovation status for the following reasons: Presumably, most firms are using their websites to directly inform customers about new products or services. They might directly mention that their product is new or innovative. Moreover, a firm might report that some of the offered products or services include a recently *emerging technology (2)* like blockchain, 3D printing or augmented reality (see Appendix A.2). Hence, the firm's product or service is likely to be innovative, at least on an incremental level, as it makes use of technologies that are fairly new. Additionally, there might be *latent patterns (3)* on a website that reveal a firm's innovation status, i.e., a firm that uses a lot of outdated terminologies on its website might be less likely to be innovative than a firm which uses more contemporary words. Furthermore, the *languages (4)*, which are used, might relate to the export status of a firm and this might provide information about a firm's innovation status because the export status is linked to firm-level innovation (e.g., Kirbach & Schmiedeberg 2006). Also, a firm might highlight that it is innovative in order to emphasize that it provides something original, which distinguishes it from competitors. Especially in this case the innovation status reported on a website does not necessarily need to be true. This drawback cannot be overcome by web-based innovation indicators easily. Although, the MIP is self-reported as well, firms have no incentive to make false declarations in this survey as they do not gain any advantage. For this reason, we expect the MIP to reveal the true innovation

status. Therefore, using innovation indicators from the MIP as target variables may help to solve the issue of incorrect innovation information on firm websites.

Table 2 - Features related to text

Website texts	1) Texts	Term-document matrix with the 5000 most frequent words (TF-IDF applied).
	2) Emerging technologies	Discrete variable that counts how often technologies from *Wikipedia's list of emerging technologies* appear on a firm's website.
	3) Latent patterns	Topics generated by the latent Dirichlet allocation (LDA) based on our term-document matrix (no TF-IDF applied).
	4) International orientation	Percentage of subpages in English language, in German language as well as all other languages.[7]

Meta information features:

Second, meta information of firm websites (Table 3) might allow to distinguish innovative firms from non-innovative firms. For example, the *website size (5)* might help to predict a firm's innovation status. Large firms are more likely to be innovative (Rammer et al. 2019). As the number of subpages of a website correlates with the number of employees of a firm (Kinne & Axenbeck 2018), 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 (6)[8] of a website could be faster and a mobile version (7) might be more often available when firms are more technologically advanced. Another potentially relevant feature is the *age of a website* (8) 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.

[7] We use the share of English, German, and other languages as variables as the Random Forest classifier can deal with collinearity issues. However, collinearity can influence the feature relevance.
[8] However, there might be some noise because the *loading time* may also be short if the website is relatively simple.

Table 3 - Features related to meta information

Meta information	5) Website size	Number of subpages on a website, total amount of characters on a website.
	6) 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).
	7) Mobile version	Dummy variable that is one if a version for mobile end user devices exists and zero otherwise.
	8) Website age	The year of the first entry at web.archive.org.

Link features:

Third, links between websites (Table 4) might also help identify the firm-level innovation status. Firms that have more business relationships with other firms might be better informed and know earlier about new profitable applications. Moreover, innovation projects are often realized in cooperation with other firms (e.g., Becker & Dietz 2003). Hence, firms with more *relationships to other firms (9)* could be more likely to be innovative. Bertschek & Kesler (2017) show that a firm's use of the social networking site Facebook is linked to product innovations. Hence, the *use of social media (10)* might reveal information about a firm's innovation status.

Table 4 - Features related to links

Links	9) Relationships with other firms	The total number of incoming as well as the total number of outgoing hyperlinks.
	10) Social media	Number of hyperlinks to Facebook, Instagram, Twitter, YouTube, Kununu, LinkedIn, XING, GitHub, Flickr, and Vimeo.

A significant part of our study is the analysis of whether the three groups of features differ in their performance when predicting a firm's innovation status. A detailed description of the features can be found in Appendix A.1.

The descriptive statistics (Table 5) show differences between innovative and non-innovative firms for most of the features. Innovative firms are mentioning more often an emerging technology term. Furthermore,

relatively more German subpages on a website exist if a firm is not innovative whereas innovative firms have more subpages in English language. Subpages in other languages do not show any difference between both groups. Moreover, 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 seems to be faster for innovative firms according to all innovation activity indicators, except innovation expenditures. The website age of non-innovative firms is a little bit lower (the first occurrence on web-archive is later) and they have less often a version of their website for mobile end user devices. Differences also exist for outgoing and incoming hyperlinks as well as for hyperlinks to social media websites. Innovative firms have on average more links. Moreover, the difference is larger for incoming than for outgoing or social media hyperlinks. Also, innovative firms have more links to social media websites.

Table 5 – Descriptive statistics for selected variables

Feature	N	Mean	S.D.	Min	Max	Group specific means							
						Innovator	No innovator	Product innovator	No product innovator	Process innovator	No process innovator	Innovation expenditures	No innovation expenditures
Emerging technology term	4,485	0.17	0.51	0	8	0.21	0.1	0.25	0.11	0.21	0.12	0.24	0.1
Percentage of German	4,485	0.86	0.23	0	1	0.84	0.89	0.82	0.88	0.84	0.88	0.81	0.9
Percentage of English	4,485	0.12	0.21	0	1	0.14	0.08	0.16	0.09	0.14	0.1	0.17	0.08
Percentage of other lang.	4,485	0.02	0.08	0	1	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Website size: Length	4,485	63,974.26	68,269.68	42	1,411,200	70,598.44	52,548.95	74,968.4	56,580.67	71,444.78	55,234.18	75,296.56	52,876.32
Website size: Nr. of pages	4,485	26.91	14.73	1	50	28.87	23.55	30.34	24.57	28.77	24.58	31.24	23.69
Loading time	4,485	0.56	2.45	0.01	140.05	0.54	0.58	0.57	0.54	0.51	0.58	0.51	0.49
Mobile version	4,485	0.72	0.45	0	1	0.75	0.67	0.76	0.7	0.76	0.67	0.73	0.69
Age of a website	4,485	2,004.68	5.3	1996	2019	2,004.39	2,005.19	2,004.23	2,004.98	2,004.43	2,004.99	2,004.37	2,005.01
Outgoing hyperlinks	4,485	14.08	18.15	1	309	15.11	12.32	15.88	12.9	15.16	12.88	16.24	12.41
Incoming hyperlinks	4,485	8.93	76.44	0	3,433	11.79	3.98	14.78	5.14	13.23	4.19	12.07	3.7
Use of social media	4,485	1.25	1.9	0	17	1.45	0.92	1.61	1.02	1.51	0.98	1.62	0.91

4 Empirical Approach

The objective of our work is the identification of website characteristics that allow predicting firm-level innovation activities. For this purpose, we extract several features from the web. We integrate these features as explanatory variables in a Random Forest classification model (Friedman, Hastie & Tibshirani 2001).[9] To evaluate the performance of the collected website characteristics, we use as a baseline model a random coin toss model based on the sample distribution. A baseline model works as a benchmark to compare more complex solutions against and helps to analyze whether the trained prediction model performs better than a random prediction.

We use the metrics accuracy, precision, and recall (Fawcett 2006) to evaluate and compare our models. Accuracy measures the fraction of all correctly predicted firms. Precision measures, for example, the fraction of correctly classified innovative firms over all firms classified as innovative, while recall measures the fraction of innovative firms that have been identified as innovative over the total amount of innovative firms. The f1-score captures the harmonic mean between precision and recall. The same applies to non-innovative firms. Respective baseline outcomes of accuracy, precision, and recall as well as the f1-score for our different innovation activity indicators are presented in Section 5.

As we analyze four different innovation indicators (four different targets), the predictive power of three different groups of features as well as their joint predictive power (in total four different groups of features), we need to 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 model's performance with the observations that are already used to train the models: The data is split into a training sample (for fitting our model) and into a test sample (for evaluating the model). The training amount is 75 percent and the test amount consists of the remaining 25 percent. For supervised learning, 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 (Friedman, Hastie & Tibshirani 2001). We explore the parameter space for the number of trees (1000, 1250) and maximum tree depth (20, 70, 75, and 80). This leads to eight different hyperparameter combinations for every model. We select the combination with the best performance for evaluation.

[9] We use the Python package *scikit-learn* for estimating the Random Forest model.

5 Results

We predict the innovation status reported in the MIP for innovators, product innovators, process innovators, and innovation expenditures with four different sets of features using a Random Forest classification approach. Random Forest models have the property that the importance of the features can be easily measured, e.g., with "mean decrease in impurity" (Breiman & Friedman, 1984) applied in the study.

5.1 Innovators

Table 6 – Results for Random Forest classification models with innovators as target

	Label	Support	Precision	Recall	F1-Score	Accuracy
Baseline model	**Non-innovators**	420	0.37	0.37	0.37	0.53
	Innovators	702	0.63	0.63	0.63	
Text features	**Non-innovators**	420	0.6	0.18	0.28	0.65
	Innovators	702	0.65	0.93	0.77	
Meta information features	**Non-innovators**	420	0.47	0.29	0.35	0.61
	Innovators	702	0.65	0.80	0.72	
Link features	**Non-innovators**	420	0.39	0.25	0.31	0.57
	Innovators	702	0.63	0.76	0.69	
All features	**Non-innovators**	420	0.57	0.18	0.27	0.64
	Innovators	702	0.65	0.92	0.76	

Better than the baseline model Worse than or as good as the baseline model

Table 6 shows evaluation metrics for models with *innovator* as target variable. For innovative firms the baseline value for precision, recall, and f1-score is 0.63 and for non-innovative firms it is 0.37. Hence, the sample is slightly imbalanced. All four different feature combinations improve precision, recall, and f1-score for the innovative firms in comparison to the baseline model. For the non-innovative firms, however, only the precision improves. Recall and f1-score are worse than random. The baseline accuracy is 0.53. With an accuracy of 0.65, an improvement of 12 percentage points, the 'text features' model shows the best result. The 'all features' model performs just slightly worse. The 'link features' model shows the weakest performance. Looking at the feature importance values (Figure A.3), *length* has the highest value. It is followed by the features *English*, *nr_pages,* and *German*. Hence, the size of a website as well as the language seem to play a crucial role when predicting whether a firm is an innovator according to the features we use. Also, most LDA topics, especially the topics 11, 24, and 16, are important as well as *loadTime*. Topic 11 could be related to the self-description of a firm (Table A.2). Topic 24 might relate to opening times because the German word "uhr" (o'clock) and several days of the week belong to the most important words. Due to

the English language as well as the word "international", topic 16 seems to relate to the self-description of international firms.

5.2 Product innovators

Table 7 – Results for Random Forest classification models with product innovators as target

	Label	Support	Precision	Recall	F1-Score	Accuracy
Baseline model	**Non-product innovators**	670	0.61	0.61	0.61	0.52
	Product innovators	428	0.39	0.39	0.39	
Text features	**Non-product innovators**	670	0.71	0.83	0.76	0.69
	Product innovators	428	0.63	0.47	0.54	
Meta information features	**Non-product innovators**	670	0.64	0.72	0.68	0.59
	Product innovators	428	0.46	0.37	0.41	
Link features	**Non-product innovators**	670	0.64	0.75	0.69	0.59
	Product innovators	428	0.46	0.33	0.39	
All features	**Non-product innovators**	670	0.71	0.84	0.77	0.69
	Product innovators	428	0.64	0.46	0.54	

Better than the baseline model Worse than or as good as the baseline model

The baseline value for precision, recall, and f1-score for predicting whether a firm is a product innovator is 0.39 and for the non-product innovators it is 0.61 (Table 7). The sample is slightly imbalanced towards non-product innovators. Both, the 'text features' as well as the 'all features' model, show for all evaluation metrics a better performance than the random weighted coin toss. For example, the accuracy increases for both models by 17 percentage points. For the 'link features' model as well as the 'meta information features' model, the accuracy and precision improves for innovators and non-innovators, but the recall for non-innovators performs worse than random. Hence, both feature groups alone do not detect a sufficient amount of innovative firms. Figure A.4 shows the feature importance for the 'all features' model and product innovators as target. Topic 1 has the largest predictive power. Looking at the most important words (Table A.3), it entails "javascript" and "browser" as unigrams and as a bigram. Therefore, it might relate to the case when a website informs its visitors to enable JavaScript in their browser.[10] That is why we assume that this topic relates to more technical advanced websites. Moreover, the features *length*, *English,* and *nr_pages* are listed among the top characteristics. Similar to *innovators,* the website size and the language seem to play a decisive role. Additionally, the LDA topic 25, which probably relates to banking, has a high relevance. Also, the German term "produkte" (products) appears among the most relevant features.

[10] It is likely that the ARGUS web scraper always receives this message as it does not fully account for dynamically loaded web pages.

5.3 Process innovators

Table 8 – Results for Random Forest classification models with process innovators as target

	Label	Support	Precision	Recall	F1-Score	Accuracy
Baseline model	**Non-process innovators**	504	0.46	0.46	0.46	0.51
	Process innovators	583	0.54	0.54	0.54	
Text features	**Non-process innovators**	504	0.58	0.47	0.52	0.60
	Process innovators	583	0.61	0.71	0.65	
Meta information features	**Non-process innovators**	504	0.49	0.43	0.46	0.53
	Process innovators	583	0.56	0.62	0.59	
Link features	**Non-process innovators**	504	0.53	0.54	0.54	0.56
	Process innovators	583	0.6	0.58	0.59	
All features	**Non-process innovators**	504	0.6	0.5	0.55	0.61
	Process innovators	583	0.62	0.71	0.66	

Better than the baseline model Worse than or as good as the baseline model

Looking at process innovators, the baseline value of non-innovative firms for precision, recall, and f1-score is 0.46 and for the innovative firms it is 0.54 (Table 8). The baseline accuracy is 0.51. Hence, the sample is nearly balanced. With an accuracy of 0.61, the 'all features' model shows the best performance. This is 10 percentage points larger than the random prediction. However, the improvement is not as large as for product innovators. When comparing precision, recall, and f1-score for different models, the 'all features' model shows again the best performance, followed by the 'text features' model. Only the 'meta information features' model performs worse than random and does not distinguish sufficiently innovative firms from non-innovative firms. The feature importance for the Random Forest classification model with all features for process innovators is displayed in Figure A.5. *Length* is the feature with the highest importance, followed by the LDA topic 9. This topic seems to relate to the building sector as it entails the German words "architektur" (architecture), "haus" (house), and "bauen" (build) (Table A.4). Moreover, the features *nr_pages, German,* and *English* also belong to the five most important characteristics. Additionally, topic 16, which seems to relate to international business, is quite influential.

5.4 Innovation expenditures

In the following, results of Random Forest classification models trained to distinguish between firms with and without innovation expenditures are presented (Table 9). The baseline accuracy is 0.52. The baseline precision, recall, and f1-score for the firms with innovation expenditures is 0.4 and without innovation expenditures it is 0.6. The sample is imbalanced, but only very slightly.

Table 9 – Results for Random Forest classification models with innovation expenditures as target

	Label	Support	Precision	Recall	F1-Score	Accuracy
Baseline model	**No Innovation expenditures**	285	0.6	0.6	0.6	0.52
	Innovation expenditures	188	0.4	0.4	0.4	
Text features	**No Innovation expenditures**	285	0.7	0.88	0.78	0.7
	Innovation expenditures	188	0.7	0.43	0.53	
Meta information features	**No Innovation expenditures**	285	0.67	0.83	0.74	0.65
	Innovation expenditures	188	0.6	0.39	0.47	
Link features	**No Innovation expenditures**	285	0.63	0.78	0.7	0.59
	Innovation expenditures	188	0.48	0.31	0.38	
All features	**No Innovation expenditures**	285	0.71	0.88	0.79	0.71
	Innovation expenditures	188	0.71	0.45	0.55	

Better than the baseline model Worse than or as good as the baseline model

The evaluation metrics show that the 'all features' model performs best. In contrast to the baseline model, the accuracy is 19 percentage points larger. Comparing the models with only a single group of features, shows that most of the predictive performance is related to text features. Moreover, the 'meta information features' model as well as 'link features' model have a recall performing worse than random. Figure A.6 entails the list with the most important features. The share of English language has a very high relevance as well as the number of subpages of a website. Also the LDA topic 13 plays a crucial role and might relate to "digital solutions" indicated by the words "software", "solutions", "entwicklung" (development) & "technology" (Table A.5). The text length, the share of German subpages (similar to English subpages), LDA topic 11 as well as the number of incoming hyperlinks seem to be of high importance, too. Furthermore, words semantically related to the verb "entwickeln" (develop) seem influential. It is plausible that these words are important when predicting firms with innovation expenditures because innovation expenditures strongly relate to research and development (R&D).

6 Discussion

We show that website characteristics relate to firm-level innovation activity: For each innovation indicator, precision and accuracy of our models are in the majority of cases better than the random coin toss model. This property proves that our statistical models could actually learn from the data. However, the values leave room for improvement as we, for example, still misclassify the existence of innovation expenditures for 29 percent of the firms. Within our sample, there can be of course also innovative firms that do not mention their innovation activity (implicitly or explicitly) on their website.

Regarding the most dominant characteristics, we see a pattern in the models that is independent of different target variables. The features *German*, *English*, *length,* and *nr_pages* have a high relevance in all trained models. This indicates that these features are generally important for measuring innovation based on website information. It is noteworthy that these features are even more relevant than words such as "innovation".[11] One has to consider, however, that the relevance of these features is only compared to the relevance of single words. If one would add up the relevance of every word appearing in the term-document matrix, the aggregated relevance of the entire text corpus would probably be higher than the relevance of the most important features. In addition, it can be said that our models based on meta information features or link features never beat the text-based models. This illustrates the relevance of text in comparison to the additional collected data. Of course, we cannot exclude the possibility that this is due to the choice of our features. There may be other web-based features that are more meaningful and have not been considered in our analysis.

The descriptive statistics (Table 5) show that a low share of German subpages is associated with positive firm-level innovation activity. Therefore, it is plausible that our models capture a negative relationship between *German* and the probability that a firm is innovative.

The selected word-based features in the Random Forest models appear to be plausible. For example, technological words and terms like “entwickeln” (develop) and “product” are chosen. These words have a very strong and direct connection to innovation. Furthermore, no single word that does not belong to the innovation context appears in the top 10 features. This is a further sign that the models only use relevant information for classification. Within the learned LDA topics, however, there are also words that are not innovation specific. Our assumption is that this can be improved by a better text processing pipeline as explained below.

As expected, product innovations can be predicted comparatively well. This can easily be explained by the presumed property that firms have a larger incentive to present innovative products on their websites. This looks different for process innovations: Process innovations are often kept secret, which gives the firm an advantage over competitors and therefore disclosure is not sought. That is why a worse predictive performance for process innovators and innovators, as the latter one contains product and process innovators, is plausible. Comparatively good results for innovation expenditures could be explained by the fact that we have this data on an annual basis. The other indicators might include more noise as they cover a three year span. Also, firms might emphasize on their websites that they do R&D in order to be in favor of the government or other stakeholders. This might explain as well why predictions perform here relatively well.

[11] Nonetheless, the word "innovation" or respective lexical variations are always present in the top 100 most important features for text-based models (see Appendix A.3).

Moreover, text data is always very noisy and therefore a model with 100 percent accuracy almost never exists in approaches like ours. However, there are some parts of our study that can be improved. Natural language processing methods have not yet been exhausted. For example, a further study could consider to apply part of speech tagging, named entity recognition, and stemming. To be more precise, filler words (e.g., auxiliary verbs) can be deleted, words with the same lexical meaning but different endings (declensions) can be transformed into a common structure or whole word classes (e.g., adjectives) can be removed. These methods can be used to further reduce the noise in the data. Also, there are countless other features that can be used to improve performance, for example, server-hosting information (location and information whether the firm hosts the website itself), which is according to our expectations, a proxy for the firm's technical infrastructure.

A general issue of studies predicting MIP-based firm-level innovation activity by means of website data is a matching problem between the survey innovation indicator and website data. In a perfect world, we would have process and product innovation data for every year or even more frequently. In the survey, however, this information is only collected on an aggregate level for the last three years. The most inexact case is that a firm was innovative once three years ago and the associated survey variable is therefore positive in the current year. As websites can change a lot during this period and it is unclear to us whether it reflects the reported innovation status. Solving this matching problem seems to us a necessary step to improve predictions. Survey data on innovation expenditures might improve the model as it covers exactly one year. Matching in this case works better, according to our assumptions, and this is also reflected in the results. Even though our predictions are far better than the random coin toss model, our approach still leaves room for improvement. Further studies should consider using an innovation indicator which is measured more frequently as target.

The main criteria for choosing a Random Forest approach are the explainability of the results and the fact that nonlinear relationships can be learned. There is, of course, the possibility to use other statistical models that have these properties, for example, Gradient Boosted Trees. Neural networks unfortunately do not offer a direct possibility to visualize the 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 a LSTM neural network.

By complementing website data with other information sources, e.g., the MUP, the innovation status can be predicted more accurately. For example, including the number of employees, the annual turnover, and the NACE code improves predictions. The idea behind this is, for example, that firms in the IT sector are on average more innovative than, for example, firms in the construction industry. In this work, we have consciously decided against adding this data in our main analysis, since the websites information, which is freely accessible for everyone on the Internet and is up-to-date daily, should be in the foreground. As a brief digression: Adding firm turnover, number of employees, and NACE codes improves the results of the ‘all features’ models. The accuracy improves for product innovators by 2 percent, for process innovators by 5

percent, for innovators by 10 percent, and for innovation expenditures by 2 percent. It must be mentioned that the sample size for this calculations is about 3000 firms, as we do not have the mentioned data for our entire sample. A further study could explore in more detail the effects of adding additional non-web data. Other additional data sources might further improve the models. For example, data from social media platforms like XING could improve predictions as innovative firms could favor specific types of employees.

Our results for product innovators are in line with the results of Kinne and Lenz (2019). Their statistical model has reached a similar accuracy for product innovators only observed in one MIP wave. Here it would be interesting to know whether the combined features from this work have an additional predictive value on their proposed "undercomplete autoencoder-like neural network architecture".

The amount of observations for this work is unfortunately very limited. The analyses can therefore be repeated as soon as the data preparation of the MIP 2019 survey has been officially completed. We are expecting a few thousand new data points, which could improve the performance and generalization of the algorithms. This step requires the re-scraping of the meta information for the newly added firms and a retraining of all statistical models.

It is important to mention that patent data can also be used as an alternative target variable in a similar study. However, patent-based indicators suffer from large time lags and rather measure inventions than innovations. It is also worth mentioning that the websites are only a self-reported representation of a firm. It is of course possible that there are large differences to the actual innovativeness. Lastly, firm website-based innovation indicators can only be applied to firms that have a website.

7 Conclusion

Traditionally, firm-level innovation indicators are constructed with data from large-scale questionnaire-based surveys. According to Kinne and Axenbeck (2018), these indicators suffer from drawbacks that can be solved by complementing or substituting them with web-based information. By analyzing whether and which website characteristics identify if a firm is innovative or not, we contribute to this discussion.

We use data from 4,485 German firms from the Mannheim Innovation Panel (MIP) 2019, which stated whether they were innovative in the last three years with respect to product or process innovation or reported innovation expenditures in the last year. For these firms, we extract their website texts, additional meta information as well as the hyperlink structure and apply several methods like a keyword search, hyperlink classification, and unsupervised learning (LDA topic modelling) to generate website characteristics. By training different Random Forest models, we analyze which website characteristics improve predictions of firm-level innovation activity. The models are trained on four different innovation indicator variables. Additionally, each model is trained on different feature sets, e.g., only text-based features.

Our results show that human-interpretable website characteristics exist which relate to firm-level innovators, product innovations, process innovations, and innovation expenditures. This is shown by an increase in

accuracy of the prediction when adding website characteristics in comparison to our baseline models. Combining text-based information with meta data and hyperlinks further improves predictions. Moreover, it can be said that both product-innovative firms and firms with current data on innovation expenditures can be better predicted with the selected model and feature setup than process-innovative firms and innovators. Analyzing the most important features in our models shows that the percentage of subpages in English as well as in German language, the number of subpages, and the total amount of characters are always decisive for the prediction regardless of the innovation indicator. However, there are some features that are only highly important for the prediction of a specific innovation indicator like the German word “entwickeln” (develop) for innovation expenditures.

Our work and related studies show that state of the art web-based predictive modeling cannot fully replace traditional surveys. However, our models provide information about innovativeness that might be of interest for politicians and researchers as the results can be quickly updated, are on a very disaggregated level (firm-level), and less expensive than surveys. Trained models can also be used to make predictions about the total quantity of German firms.

References

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

Archibugi, D. & Planta, M. (1996), 'Measuring technological change through patents and innovation surveys', *Technovation* **16**(9), 451 – 519, URL**:** *http://www.sciencedirect.com/science/article/pii/0166497296000314.*

Arora, S. K., Youtie, J., Shapira, P., Gao, L. & Ma, T. (2013), 'Entry strategies in an emerging technology: a pilot web-based study of graphene firms', *Scientometrics,* **95**(3), 1189–1207.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999), *Modern information retrieval* (Vol. 463), New York: ACM press.

Beaudry, C., Heroux-Vaillancourt, M. & Rietsch, C. (2016), 'Validation of a web mining technique to measure innovation in high technology Canadian industries', OECD Blue Sky Forum on Science and Innovation Indicators, Ghent, Belgium.

Becker, W. & Dietz, J. (2003), 'R&D Cooperation and innovation activities of firms – evidence for the German manufacturing industry', *Research Policy*, 33, 209-223.

Bersch, J., Gottschalk, S., Müller, B. & Niefert, M. (2014), 'The Mannheim Enterprise Panel (MUP) and firm statistics for Germany', URL: *https://www.zew.de/en/publikationen/zew-discussion-papers/.*

Bertschek, I. & Kesler, R. (2017), 'Let the user speak: Is feedback on Facebook a source of firms' innovation?', *ZEW Discussion Paper No. 17-015*, Mannheim.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003), 'Latent Dirichlet allocation', *Journal of machine Learning research*, *3*(Jan), 993-1022.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984), 'Classification and regression trees', Belmont, CA: Wadsworth. *International Group*, *432*, 151-166.

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

Choi, H. & Varian, H. (2012), 'Predicting the present with Google trends', *Economic Record* **88**, 2–9.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of the American society for information science*, **41**(6), 391-407.

Fawcett, T. (2006), 'An introduction to ROC analysis', *Pattern recognition letters*, **27**(8), 861-874.

Friedman, J., Hastie, T., & Tibshirani, R. (2001), *The elements of statistical learning*, **1**(10), New York: Springer series in statistics.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012), 'A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches', *IEEE Transactions on Systems, Man, and Cybernetics*, Part C (Applications and Reviews), **42**(4), 463-484.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009), 'Detecting influenza epidemics using search engine query data', *Nature,* **457**(7232), 1012.

Gök, A., Waterworth, A. & Shapira, P. (2015), 'Use of web mining in studying innovation', *Scientometrics,* **102**(1), 653–671.

Hoberg, G. & Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of Political Economy,* **124**(5), 1423–1465.

Katz, J. S. & Cothey, V. (2006), 'Web indicators for complex innovation systems', *ReR search Evaluation,* **15**(2), 85–95.

Kinne J., Axenbeck J. (2018), 'Web mining of firm websites: A framework for web scraping and a pilot study for Germany', *ZEW Discussion Paper No. 18-033*, Mannheim.

Kinne J., Lenz D. (2019), 'Predicting innovative firms using web mining and deep learning', *ZEW Discussion Paper No. 19-001*, Mannheim.

Kirbach & Schmiedeberg (2006): 'Innovation and export performance: Adjustment and remaining differences in East and West German manufacturing', *Economics of Innovation and New Technology*, 17, S. 435-457.

Lenz, D. & Winker, P. (2018), 'Measuring the diffusion of innovations with paragraph vector topic models', *Technical report*.

Martin, J. H., & Jurafsky, D. (2009), *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition,* Upper Saddle River: Pearson/Prentice Hall.

Miner, G., Elder IV, J. & Hill, T. (2012), *Practical text mining and statistical analysis for non-structured text data applications*, Academic Press.

Nathan, M. & Rosso, A. (2017), 'Innovative events', Technical report*, Centro Studi Luca d'Agliano Development Studies Working Paper No. 429*.

OECD/Eurostat (2018), 'Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation', *The Measurement of Scientific, Technological and Innovation Activities*, 4th Edition, OECD Publishing, Paris/Eurostat, Luxembourg.

Peters, B. & Rammer, C. (2013), 'Innovation panel surveys in Germany', *Handbook of Innovation Indicators and Measurement*, Edward Elgar Publishing, chapter 6, pp. 135– 177, URL: *https://EconPapers.repec.org/RePEc:elg:eechap:144276.*

Pukelis, L., & Stanciauskas, V. (2019), 'Using Internet Data to Compliment Traditional Innovation Indicators', URL: https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf.

Shepherd, W. G. & Shepherd, J. M. (1979), *The economics of industrial organization*, Waveland Press.

Solow, R. M. (1957). 'Technical change and the aggregate production function', *The review of Economics and Statistics*, 312-320.

Sparck Jones, K. (1972). 'A statistical interpretation of term specificity and its application in retrieval', *Journal of documentation,* **28**(1), 11-21.

Rammer, C., Behrens, V., Doherr, T., Hud, M., Köhler, M., Krieger, B., ... & von der Burg, J. (2019), 'Innovationen in der deutschen Wirtschaft: Indikatorenbericht zur Innovationserhebung 2018', *ZEW Innovationserhebungen-Mannheimer Innovationspanel (MIP)*, Mannheim.

Wößmann, L. & Lachenmaier, S. (2006), 'Does innovation cause exports? Evidence from exogenous innovation impulses and obstacles using German micro data', *Oxford Economic Papers* **58**(2), 317–350, URL**:** *https://dx.doi.org/10.1093/oep/gpi043.*

Youtie, J., Hicks, D., Shapira, P. & Horsley, T. (2012), 'Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies', *Technology Analysis & Strategic Management,* **24**(10), 981–995.

Appendix

A.1 Detailed information on the calculation of our features

Text-based features:

1) Texts To identify the most relevant terms when predicting a firm's innovations status, we transformed the scraped texts into a format that allows to do mathematical operations: We converted the website texts into a term-document matrix (e.g., Baeza-Yates & Ribeiro-Neto 1999, Blei et al. 2003), 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. Every 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 5000 most frequent words in our entire training text corpus. To calculate the term-document matrix we need to do some preprocessing steps in advance. First, we merge all scraped subpages related to a single firm and we delete irrelevant subpages by using the gold standard approach (e.g., Kinne & Lenz 2019). The model distinguishes between subpages including relevant information like product descriptions and subpages with irrelevant information like the imprint, information about cookies or texts that are prescribed by law. It provides values regarding the relevance of subpages between zero and one. Values close to one indicate highly relevant subpages and values close to zero the opposite. We only keep subpages that have at least a regression value of 0.9. Also, every word is converted into lower case and punctuation is excluded. After calculating the term-document matrix, we manipulate the term-frequency counts by the TF-IDF scheme (Baeza-Yates, R., & Ribeiro-Neto 1999) as it usually improves predictions.

2) Emerging technology terms – To capture firms mentioning emerging technologies, we conduct a keyword search in which we count how often technologies from *Wikipedia's list of emerging technologies*[12] appear on a firm's website using all subpages and the entire vocabulary. A detailed list of all used keywords is provided in the Appendix A.2. The feature *SumTechWords* (only used for calculating the descriptive statistics) captures the total number of emerging technologies terms appearing on one 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) (Blei et al. 2003). The LDA algorithm assumes that a document consists of a fixed set of topics, while every topic is a distribution of words. By linking each word in the text corpus 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 LDA, the topic-document matrix shows how much every topic contributes to each document (website). This matrix is used for predicting the innovation status of a firm, i.e., the topic contribution to a document is used

[12] Retrieved from https://en.wikipedia.org/wiki/List_of_emerging_technologies (accessed on August 16, 2018). The article is updated several times a month.

as a feature (*lda_topics*). We use the term-document matrix of our training sample as a text corpus for fitting the LDA model. However, we do not manipulate the matrix by means of the TF-IDF scheme this time and use simple frequency counts instead.[13] To improve our model performance, we delete all words that appear less than 150 times in the text corpus as we only want to use words that appear often enough to identify topics that are not exclusively valid for our sample.[14] We use the Python package *scikit-learn* to train the LDA model. In the standard LDA approach the number of topics needs to be defined. For this purpose, we apply the grid-search technique.[15] It is evaluated which model parameter combination leads to the best result according to the log likelihood. To improve our LDA model, we conduct a grid-search over different values for the 'number of topics'-parameter. We try 15, 20, and 25 topics. For all models the optimal number of topics was 25.

4) 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.[16] Therefore, it is quite likely that firms with international customers choose to describe their products in English language. Therefore, we measure the share of subpages in English language, in German language as well as all other languages to approximate the export orientation of a firm (*English*, *German*, *other_lang*).[17]

Meta information features:

5) Page properties – Approximating firm size might help to predict a firm's innovation status. Kinne & Axenbeck (2018) show that the number of subpages correlates with firm size. Hence, we use the number of subpages as a feature to predict a firm's innovation status (*nr_pages*).[18] We additionally analyze to what extent the number of characters per website, which might also relate to firm size, informs about the firm's innovation status.

6) Loading time – The loading time (*loadTime*) of the web pages is determined by a simple http or https request. The time from sending the request until the arrival of the response is measured. Servers which are very far away or which only process the requests slowly (e.g., due to bad hardware or an overload) have a high loading time (in milliseconds). This feature serves as a proxy for a firm's IT hardware structure. However, it should be noted that the IT infrastructure can also be outsourced to professional hosting firms.

[13] We do not use the TF-IDF score in this case as it worsens the models' performance.
[14] We use *CountVectorizer* from *scikit-learn* to delete words.
[15] We use *GridSearchCV* from *scikit-learn* to conduct the grid-search.
[16] Retrieved from https://www.ethnologue.com/cloud/eng (accessed on April 1, 2019).
[17] To classify the language of a subpages, we apply the Python package *langdetec*t.
[18] One problem with this feature is that it is truncated at 50 subpages because this is the limit of the web-scraper. However, as we use a Random Forest model and it selects cut-off points for splitting, we can cope with truncated features.

7) Mobile version – For each website, it is recorded whether a version for mobile end user devices exists. A Google API [19] 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 Googles' mobile version passing score.

8) Website age – To determine the website age, we used web.archive.org.[20] 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 (*domainPurchaseYearProxy*). We collected this data in 2019. So if a firm was observed by web.archive.org for the first time in 2019, it is possible that firm age is declared as 2019, even though we scraped the firm website in 2018.

Link features:

9) Relationships with other firms – 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 and incoming hyperlinks to other firms is observed. We measure incoming hyperlinks by counting how often firms which are listed in the entire MUP refer to a particular firm (*outgoing_links*, *incoming_links*).

10) 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*).[21]

A.2 List of emerging technology terms used in the conducted keyword search

English terms: Agricultural robot, closed ecological systems, cultured meat, precision agriculture, vertical farming, micro air vehicle, neural-sensing headset, claytronics, four-dimensional printing, molecular assembler, utility fog, arcology, domed city, aerogel, amorphous metal, bioplastic, conductive polymers, cryogenic treatment, fullerene, graphene, lab-on-a-chip, high-temperature superconductivity, magnetorheological fluid, high-temperature superfluidity, metamaterials, metal foam, multi-function structures, nanomaterials, carbon nanotube, programmable matter, quantum dots, silicene, superalloy, synthetic diamond, time crystals, translucent concrete, 3D displays, ferroelectric liquid crystal display, field emission display, holography, interferometric modulator display, laser video displays, OLED displays, microLED displays, phased-array optics, screenless display, virtual retinal display, bionic contact lens, eyetap, 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, airborne wind turbine, artificial photosynthesis, concentrated solar power, electric double-layer capacitor, energy harvesting, flywheel energy storage, fusion power, generation iv reactor , grid energy storage, home

[19] https://www.googleapis.com/pagespeedonline/v3beta1/mobileReady
[20] We are aware that this feature can suffer from a lot of noise, as websites may not allow web-archive to crawl or websites may change their URL.
[21] Outgoing and incoming hyperlinks as well as hyperlinks to social media pages are counted by means of the *regex* package.

fuel cell, lithium-air battery, lithium iron phosphor battery, lithium-sulfur battery, magnesium battery, molten salt reactor, nanowire battery, nantenna, ocean thermal energy conversion, smart grid, space-based solar power, thorium fuel cycle, vortex engine, wireless energy transfer, zero-energy building, computer-generated imagery, immersive virtual reality, ultra-high-definition television, 5G cellular communications, ambient intelligence, artificial brain, artificial general intelligence, atomtronics, augmented reality, blockchain, carbon nanotube field-effect transistor, civic technology, cryptocurrency, dna digital data storage, 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, Li-Fi, 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, artificial uterus, body implants, prosthesis, cryonics, de-extinction, genetic engineering of organisms and viruses, suspended animation, hibernation, immunotherapy, immunoncology, life extension, nano medicines, nano sensors, oncolytic viruses, personalized medicine, whole genome sequencing, plantibody, regenerative medicine, robotic surgery, stem cell treatments, synthetic biology, synthetic genomics, tissue engineering, tricorder, virotherapy, vitrification, cryoprotectant, brain-computer interface, brain reading, neuro informatics, electro encephalography, head transplant, neuro prosthetics, caseless ammunition, cloaking device, directed energy weapon, electro laser, electromagnetic weapons, electrothermal-chemical technology, force field, green bullet, laser weapon, particle beam weapon, plasma weapon, pure fusion weapon, sonic weapon, stealth technology, vortex ring gun, wireless long-range electric shock weapon, anti-gravity, artificial gravity, asteroid mining, hyper telescope, stasis chamber, solar gravitational lens, inflatable space habitat, miniaturized satellite, android, gynoid, molecular nanotechnology, nanorobotics, powered exoskeleton, self-reconfiguring modular robot, swarm robotics, unmanned vehicle, airless tire, alternative fuel vehicle, beam-powered propulsion, electro hydrodynamic propulsion, flexible wings, fluidics, flying car, fusion rocket, hoverbike, high altitude platforms, jetpack, backpack helicopter, maglev train, vactrain, magnetic levitation, mass driver, float to orbit, nuclear photonic rocket, personal rapid transit, photon rocket, physical internet, scooter-sharing system, propellant depot, pulse detonation engine, reusable launch system, space elevator, spaceplane, super sonic transport, vehicular communication systems

German terms: Agrarroboter, Geschlossenes Ökosystem, Zuchtfleisch, Präzisions Landwirtschaft, Vertikale Landwirtschaft, Mikro-Luftfahrzeug, Neuronales Headset, Claytronics, Vierdimensionales Drucken, Molekularer Assembler, Versorgungsnebel, Arkologie, Kuppelstadt, Aerogel, amorphes Metall, Bio-Kunststoff, leitfähige Polymere, Kryogene Behandlung, Fulleren, Graphen, Labor auf einem Chip, Hochtemperatur-Supraleitung, Magnetorheologische Flüssigkeit, Hochtemperatur-Superfluidität, Meta-Materialien, Metall-Schaum, Multifunktions-Strukturen, Nano-Materialien, Kohlenstoffnanoröhre, programmierbare Materie, Quantum-Punkte, Silicen, Super-Legierung, Synthetischer Diamant, Zeit-Kristall, durchsichtiger Beton, 3D-Display, ferro-elektrische Flüssigkristallanzeige, Feld-Emissions-Anzeige, Holographie, interferometrische Modulatoranzeige, Laser-Video-Display, OLED Display, Mikro-LED Display, Gruppenstrahler-Optik, bildschirmlose Anzeige, virtuelle Netzhautanzeige, bionische Kontaktlinse, EyeTap, Teleskop-Pixelanzeige, zeitgemultiplexter optischer Verschluss, volumetrische Anzeige, Biometrie, digitale Duft-technologie, Elektronische Nase, E-Textil, Flexible Elektronik, Memoristor, Molekulare Elektronik, Nano-Elektro-Mechanisches-System, Spintronik, Thermo-Kupfer-Säulen-Stoß, dreidimensionale integrierte Schaltung, Luft-Wind-Kraftanlage, Künstliche Photosynthese, Konzentrierte Solarenergie, Elektrischer Doppelschicht-Kondensator, Energie-Ernte, Schwungrad-Energiespeicher, Fusionskraft, Reaktor der Generation IV, Netz-Energie-Speicher, Heim-Brennstoffzelle, Lithium-Luft-Batterie, Lithium-Eisen-Phosphor-Batterie, Lithium-Schwefel-Batterie, Magnesium-Batterie, Salz-Schmelz-Reaktor, Nano-Draht-Batterie, Nantenne, Ozean-thermische Energieumwandlung, Intelligentes Netz, weltraum-gestützte Solarenergie, Thorium-Brennstoff-Kreislauf, Vortex-Motor, drahtlose Energie-Übertragung, Null-Energie-Haus, computergeneriertes Bild, Immersive Virtualität, hochauflösendes Fernsehen, 5G zellulare Kommunikation, Umgebungs-Intelligenz, künstliches Gehirn, künstliche Intelligenz, Atomtronik, erweiterte Realität, Blockchain, Kohlenstoff-Nanoröhren-Feldeffekt-

Transistor, zivile Technik, Krypto-Währung, digitale DNA-Datenspeicherung, Exascale-Computing, Gestenerkennung, Internet der Dinge, Neue Speichertechnologie, Neue magnetische Speichertechnologie, optische Platten der vierten Generation, holografischer Speicher, allgemeines Rechnen auf Grafikprozessoren, Exokortex, Li-Fi, Maschinen-Übersetzung, maschinelles Sehen, mobile Zusammenarbeit, Nano-Funk, optische Datenverarbeitung, Quanten-Computer, Quantenkryptographie, Radiofrequenz-Identifikation, semantisches Web, intelligenter Lautsprecher, Software-definiertes Funkgerät, Spracherkennung, subvokale Erkennung, Hybrid-Forensik, künstliche Gebärmutter, Körperimplantat, Kryonik, Löschung, Gentechnik, Immun-Therapie, Immunkologie, Lebensverlängerung, Nanomedizin, Nanosensoren, onkolytische Viren, individualisierte Medizin, Pflanzenkörper, regenerative Medizin, Roboterchirurgie, Stammzellentherapie, synthetische Biologie, synthetische Genomik, Gewebezüchtung, Tricorder, Virus-Therapie, Verglasung, Kälteschutzmittel, Gehirn-Computer-Schnittstelle, Gehirn-Lesen, Neuroinformatik, Elektro-Enzephalographie, Kopftransplantation, Neuroprothetik, Hülsenlose Munition, Tarn-Gerät, gerichtete Energiewaffe, Elektro-Laser, elektromagnetische Waffen, elektrothermisch-chemische Technologie, Kraftfeld, grünes Geschoss, Laser-Waffe, Strahl-Waffe, Plasma-Waffe, Fusions-Waffe, Schall-Waffe, Tarn-Technologie, Wirbelringkanone, Elektroschock-Waffe, Anti-Schwerkraft, künstliche Schwerkraft, Asteroiden-Abbau, Hyper-Teleskop, Stase-Kammer, Sonnengravitationslinse, Aufblasbares Weltraum-Habitat, Miniatur-Satellit, Android, Molekulare Nanotechnologie, Nanorobotik, Exoskelett, selbstkonfigurierender Roboter, Schwarm-Robotik, unbemanntes Fahrzeug, luftlose Reifen, Fahrzeug mit alternativen Kraftstoffen, Strahl-Antrieb, elektro-hydrodynamischer Antrieb, flexible Flügel, Fluidik, fliegendes Auto, Fusionsrakete, Schwebefahrrad, Hochplattform, Jetpack, Rucksackhelikopter, Magnetschwebebahn, Vactrain, Magnetische Schwebetechnik, Massenantrieb, Orbit-Flug, photonische Kernrakete. Personenschnellverkehr, Photonenrakete, physisches Internet, Roller-Sharing-System, Treibstofflager, Pulsdetonationssystem, wiederverwendbares Startsystem, Raumaufzug, Raumflugzeug, Überschall-Transport, Fahrzeugkommunikationssystem

A.3 Top 100 most relevant features for each 'all features' model (Table A.1)

Model	**Top 100 most relevant features**
Innovators & all features	length, English, nr_pages, German, lda_topic_11, lda_topic_24, lda_topic_16, lda_topic_20, lda_topic_7, lda_topic_17, lda_topic_18, lda_topic_19, loadTime, lda_topic_1, lda_topic_6, domainPurchaseYearProxy, lda_topic_4, lda_topic_12, lda_topic_23, lda_topic_13, produkte, lda_topic_2, incoming_links, lda_topic_9, lda_topic_5, lda_topic_14, lda_topic_3, lda_topic_21, outgoing_links, lda_topic_8, lda_topic_25, lda_topic_22, lda_topic_10, lda_topic_15, weltweit, social_media, entwickelt, unserer, entwickeln, entwicklung, kunden, gmbh, wurde, sowie, einsatz, unternehmen, bieten, seit, bereich, erfolgreich, mehr, partner, software, produkten, wurden, jahre, finden, dabei, erhalten, dass, anforderungen, weitere, stehen, lassen, anwendungen, sowohl, seite, immer, ganz, jahren, bitte, stellen, ca, entsorgung, anwendung, innovativen, ab, team, arbeiten, bereits, uhr, neuen, zeit, leistungen, hohe, funktionen, systeme, neue, beim, other_lang, stadt, rund, erfahrung, mitarbeiter, gerne, art, produktion, wunsch, nutzen, neben
Product innovators & all features	lda_topic_1, nr_pages, English, length, lda_topic_25, incoming_links, software, German, lda_topic_8, lda_topic_14, lda_topic_16, domainPurchaseYearProxy, lda_topic_12, lda_topic_22, entwickelt, systeme, lda_topic_24, lda_topic_13, loadTime, produkte, lda_topic_7, weltweit, lda_topic_20, outgoing_links, lda_topic_9, lda_topic_21, system, lda_topic_10, lda_topic_4, lda_topic_23, lda_topic_18, lda_topic_6, lda_topic_11, lda_topic_19, lda_topic_3, lda_topic_2, lda_topic_5, lda_topic_15, lda_topic_17, social_media, innovative, produkten, gmbh, seit, innovativen, test, entwickeln, automatisch, anwendung, unserer, anwendungen, daten, weitere, sowie, entwicklung, bereich, jahre, online, beratung, wurde, mehr, bieten, integration, finden, kunden, softwareentwicklung, jahren, dass, stellt, informationen, display, bau, neuen, unternehmen, dabei, mm, schnittstellen, stehen, hardware, flexible, version, technologie, angezeigt, programm, mitarbeiter, prozesse, neben, neue, design, tool, einsatz, anbieten, technologies, abgestimmt, viele, data, produktion, germany, rund, basis
Process innovators & all features	length, lda_topic_9, nr_pages, German, English, lda_topic_16, lda_topic_22, lda_topic_4, lda_topic_3, social_media, lda_topic_1, lda_topic_23, loadTime, outgoing_links, lda_topic_14, lda_topic_5, produkte, lda_topic_12, lda_topic_15, lda_topic_8, informationen, lda_topic_11, lda_topic_25, lda_topic_7, lda_topic_2, lda_topic_24, lda_topic_6, lda_topic_13, weltweit, lda_topic_19, lda_topic_17, lda_topic_20, kunden, lda_topic_18, incoming_links, zeit, lda_topic_21, lda_topic_10, domainPurchaseYearProxy, management, gmbh, sowie, entwickelt, seit, unserer, internationalen, flexible, bietet, entwickeln, integration, optimieren, partner, technologien, wurde, bieten, mehr, erfolgreiche, neue, neuen, prozess, iso, bereich, basis, document, team, system, anforderungen, knowhow, anwendungen, arbeiten, ca, international, hohe, innovative, dass, einsatz, moved, beim, unternehmen, entwicklung, neben, baugruppen, finden, zukunft, jahren, dabei, anwendung, bereits, notwendig, stehen, prozesse, erfahrung, standards, wunsch, zusammenarbeit, engagement, immer, mobile_version, verschiedenen, wurden
Innovation expend. & all features	English, nr_pages, lda_topic_13, length, German, lda_topic_11, entwickelt, incoming_links, entwickeln, weltweit, entwicklung, lda_topic_5, domainPurchaseYearProxy, lda_topic_7, system, lda_topic_23, lda_topic_24, lda_topic_21, outgoing_links, lda_topic_9, lda_topic_4, innovativen, loadTime, integration, lda_topic_10, lda_topic_1, produkte, lda_topic_16, lda_topic_18,

	forschung, lda_topic_6, lda_topic_2, high, lda_topic_25, social_media, technologien, lda_topic_14, lda_topic_22, anwendung, lda_topic_17, innovationen, support, lda_topic_3, lda_topic_8, umsatz, gmbh, engineering, anwendungen, lda_topic_15, development, lda_topic_20, lda_topic_12, flexible, technologie, basis, innovative, automatisch, lda_topic_19, ag, steuerung, data, markt, ml, kunden, produkten, kompetenzen, einsatz, systems, sowie, wurde, germany, sowohl, mehr, systeme, dr, jahre, automatisierung, unserer, group, zusammenarbeit, komponenten, technology, ab, product, innovation, management, finden, position, solutions, daten, partnern, kompetenz, beratung, drei, research, zeit, bieten, unternehmen, dabei, seit

A.4 Sector and firm size representation

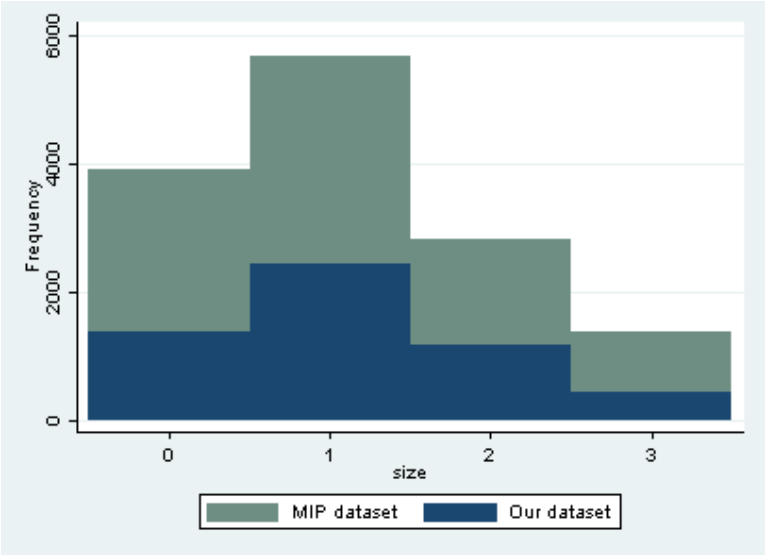

Figure A.1: The absolute frequency of data points (our dataset and MIP dataset) measured by the size of the firm. ("0": 0-9 employees, "1": 10-49 employees, "2": 50-249 employees, "3": 250+ employees)

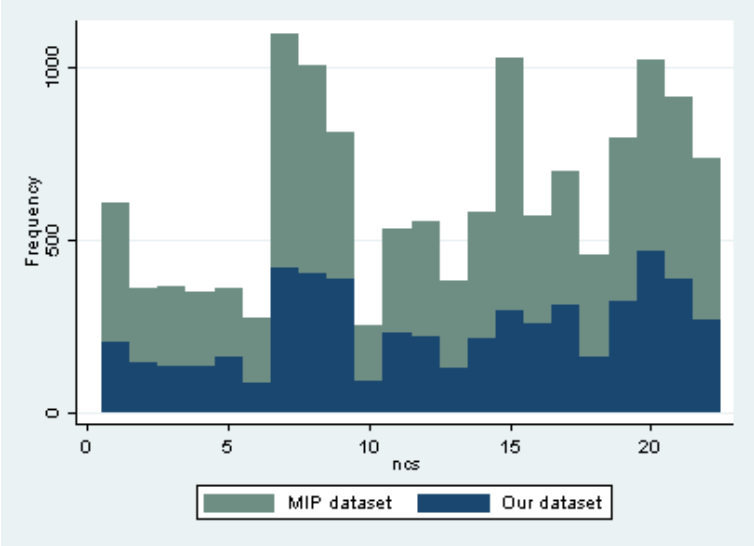

Figure A.2: The absolute frequency of data points (our dataset and MIP dataset) for different sectors (see Appendix A.5).

A.5 Sector definitions (abbreviation: ncs)

1 "Food, beverages, and tobacco products (10-12)"

2 "Textiles, wearing apparel, and leather products (13-15)"

3 "Wood and paper products (16-18)"

4 "Chemical and pharmaceutical products (20-21)"

5 "Rubber and plastic products (22)"

6 "Other non-metallic mineral products (23)"

7 "Basic metals and metal products (24-25)"

8 "Machinery (28)"

9 "Electronic and electrical products (26-27)"

10 "Vehicles and transport equipment (29-30)"

11 "Furniture and other manufacturing (31-33)"

12 "Water supply and waste management (36-39)"

13 "Energy, mining and oil refineries (5-9, 35)"

14 "Wholesale (46)"

15 "Transportation and post (49-53)"

16 "Media services, telecommunication (58-61)"

17 "Information and communication services (62-63)"

18 "Financial and insurance activities (64-66)"

19 "Consultancy and advertisement (69, 70.2, 73-74)"

20 "Technical services and scientific research (71-72)"

21 "Business services (78-82)"

22 "Missing"

A.6 Feature importance

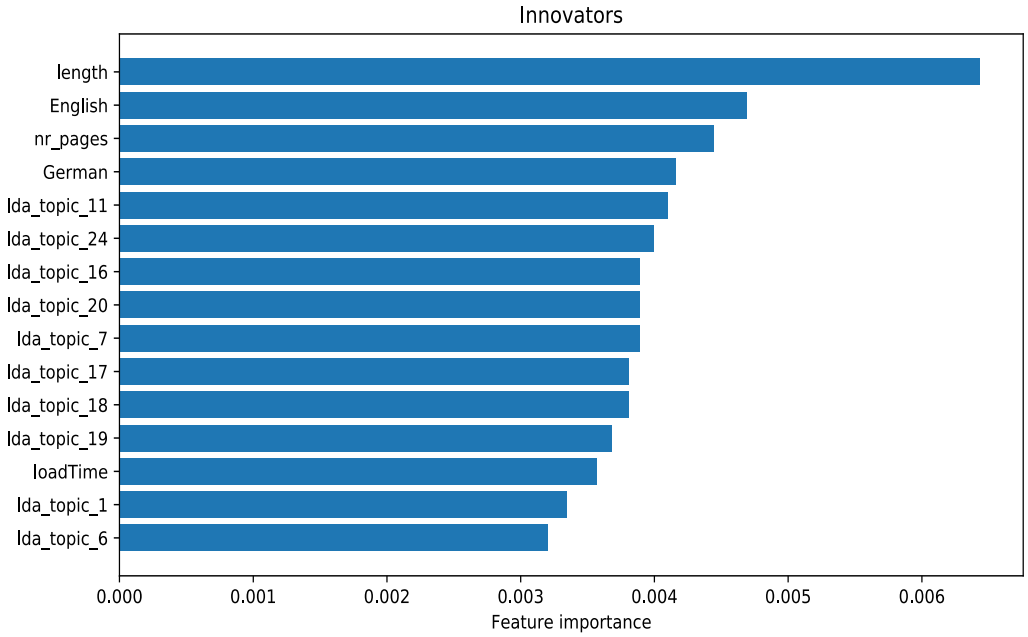

Figure A.3: Feature importance for the 'all features' model and innovators as target

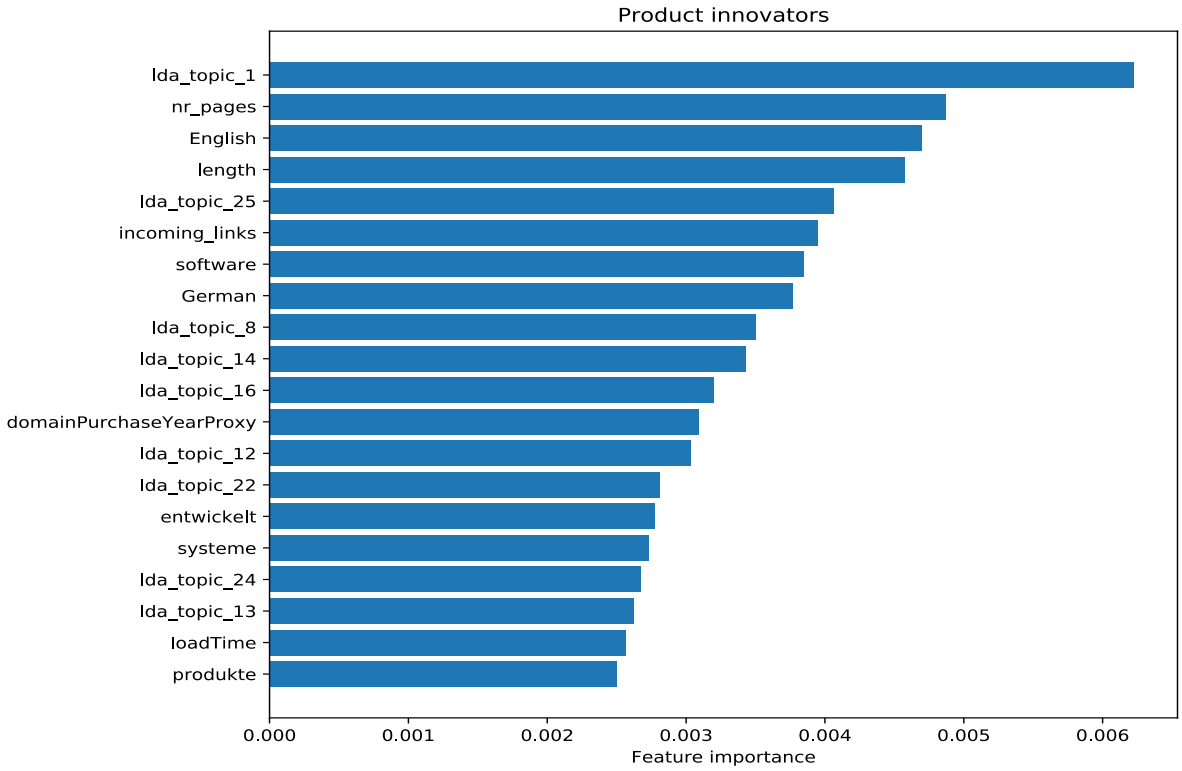

Figure A.4: Feature importance for the 'all features' model and product innovators as target

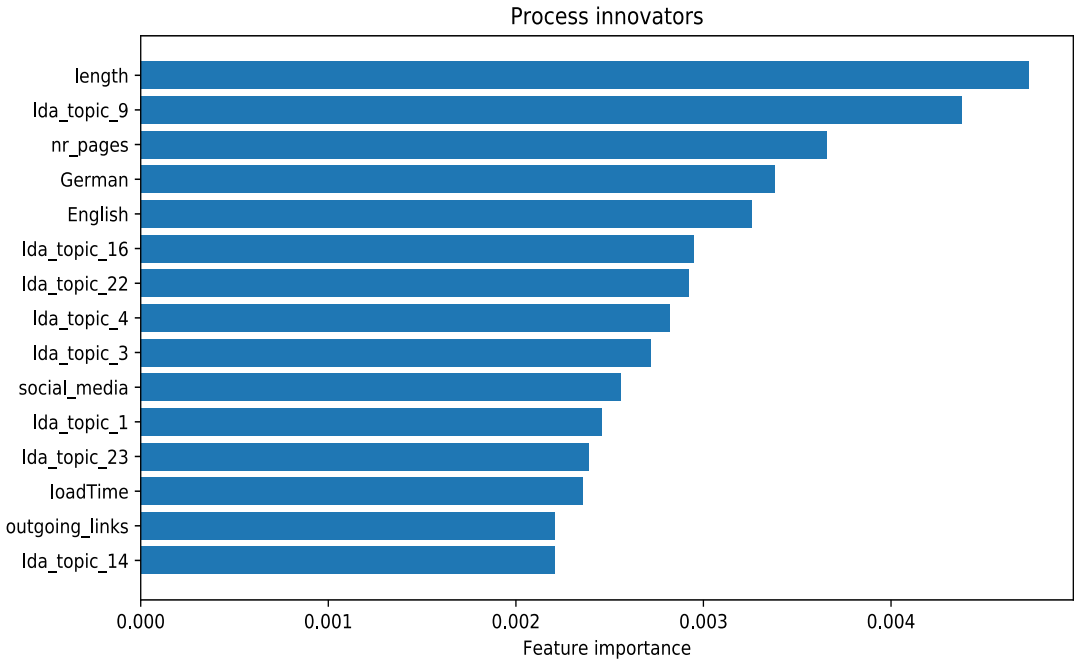

Figure A.5: Feature importance for the 'all features' model and process innovators as target

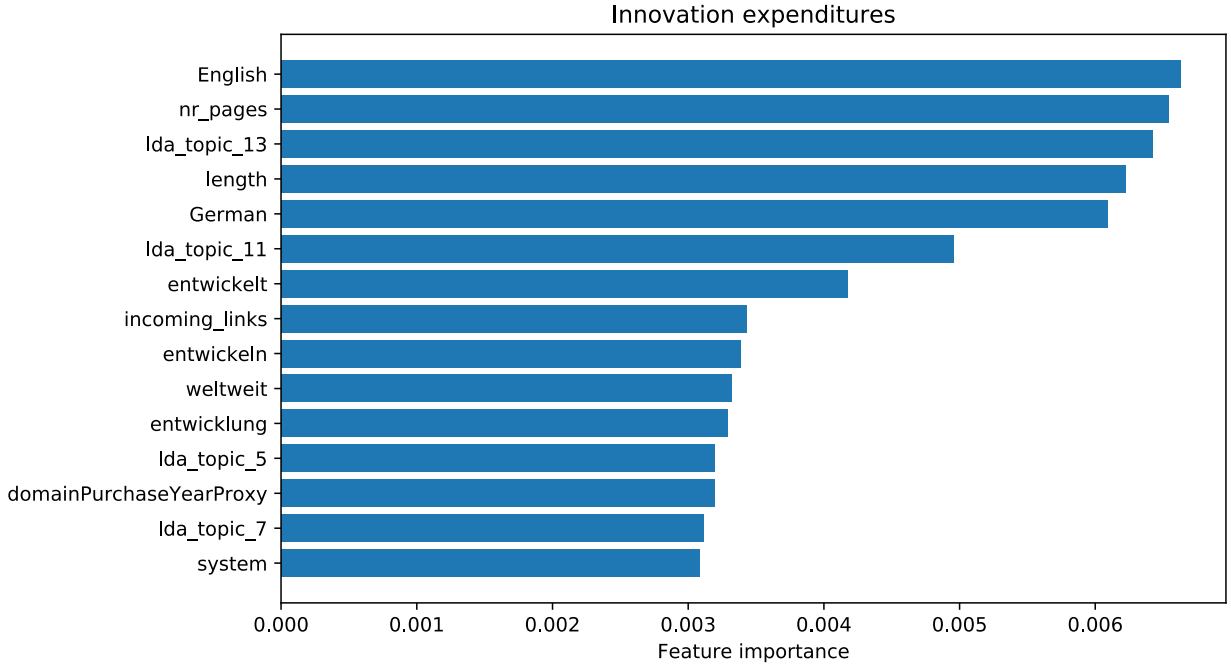

Figure A.6: Feature importance for the 'all features' model and innovation expenditures as target

A.7 Topic analysis

Table A.2 – Most important words for the three LDA topics with the highest feature relevance when predicting innovators

Topic identifier	**Most important words**
11	unserer, gmbh, gerne, sowie, kunden, finden, seit, produkte, bieten, holz, firma, bereich, beraten, fertigen, einfach
24	uhr, uhr uhr, montag, freitag, samstag, september, donnerstag, freitag uhr, euro, montag freitag, telefon, sonntag, samstag uhr, mittwoch, info
16	deutschland, group, data, services, business, international, germany, solutions, manager, hamburg, new, products, management, company, technology

Table A.3 – Most important words for the three LDA topics with the highest feature relevance when predicting product innovators

Topic identifier	**Most important words**
1	farben, javascript, design, browser, schwarz, produkt, material, nutzen, produkte, druck, verschiedenen, kategorien, informationen, newsletter, javascript browser
25	sparkasse, einfach, euro, bic, online, onlinebanking, app, geld, sparen, immer, nutzung, gemeinsam, sicher, finden, frei
8	gmbh, fertigung, produkte, sowie, maschinen, kunden, anlagen, unserer, montage, produktion, copyright, entwicklung, rights, bereich, herstellung

Table A.4 – Most important words for the three LDA topics with the highest feature relevance when predicting process innovators

Topic identifier	**Most important words**
9	wurde, bad, neubau, architekten ,planung, architektur, haus fenster, wurden, sowie, raum, sanierung, holz, bauen, bau
16	research, new, group, mio, porsche, update, services, business, international, germany, berlin, weltweit, management, company, solutions
22	news, halle, produkte, stand, auswahl, halle stand, anwendungen, informationen, messen, messe, test, entwicklung, newsletter, finden, artikel

Table A.5 – Most important words for the three LDA topics with the highest feature relevance when predicting innovation expenditures

Topic identifier	**Most important words**
13	systems, test, entwicklung, technology, services, products, gmbh, company, power, management, engineering, line, software, solutions, new
11	unternehmen, mehr, kunden, dass, mitarbeiter, neue, unserer, bietet, dabei, entwicklung, immer, partner, informationen, themen, data
5	gmbh, produkte, sowie, maschinen, unserer, kunden, hersteller, fertigung, einsatz, anlagen,produktion, technische, maschine, komponenten, bieten

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