// NO.19-001 | 01/2019

DISCUSSION PAPER

// JAN KINNE AND DAVID LENZ

Predicting Innovative Firms using Web Mining and Deep Learning

Leibniz Association

ZEW

Predicting Innovative Firms using Web Mining and Deep Learning

Jan Kinne[a,b,1,2] **and David Lenz**[c,1,2]

[a]Department of Economics of Innovation and Industrial Dynamics, ZEW Centre for European Economic Research, Mannheim, Germany; [b]Department of Geoinformatics Z_GIS, University of Salzburg, Salzburg, Salzburg, Austria; [c]Department of Econometrics and Statistics, Justus-Liebig-University, Gießen, Germany

This manuscript was compiled on January 21, 2019

Innovation is considered as a main driver of economic growth. Promoting the development of innovation through STI (science, technology and innovation) policies requires accurate indicators of innovation. Traditional indicators often lack coverage, granularity as well as timeliness and involve high data collection costs, especially when conducted at a large scale. In this paper, we propose a novel approach on how to create firm-level innovation indicators at the scale of millions of firms. We use traditional firm-level innovation indicators from the questionnaire-based Community Innovation Survey (CIS) survey to train an artificial neural network classification model on labelled (innovative/non-innovative) web texts of surveyed firms. Subsequently, we apply this classification model to the web texts of hundreds of thousands of firms in Germany to predict their innovation status. Our results show that this approach produces credible predictions and has the potential to be a valuable and highly cost-efficient addition to the existing set of innovation indicators, especially due to its coverage and regional granularity. The predicted firm-level probabilities can also directly be interpreted as a continuous measure of innovativeness, opening up additional advantages over traditional binary innovation indicators.

Web Mining | Web Scraping | R&D | R&I | STI | Innovation | Indicators | Text Mining | Natural Language Processing | NLP | Deep Learning

JEL Classification: O30, C81, C83

Innovations can disrupt individual industries with game-changing technology and the most radical innovations can even reshape whole economies. Despite having a destructive element, innovation is widely considered to be a main driver of long-term economic growth. Such growth may be kick-started by radical innovations or driven forward by a constant stream of so called incremental innovations which cause continuous change. Measuring and promoting innovation is the main objective of STI (science, technology and innovation) policy, which requires an accurate and timely picture of the current state of the STI system to implement policy measures in an evidence-based manner. However, traditional innovation indicators from questionnaire-based surveys or patent-based studies struggle to provide the full picture (1–3). (4) identified shortcomings of traditional innovation indicators concerning their coverage, granularity, timeliness, and cost. They proposed to use firm website content as a source of firm-level information, leveraging the fact that almost all relevant firms have websites nowadays. They argue that these websites are used as platforms to provide information on a firm's products and services, achievements, strategies, and relationships. All these aspects may be related to innovations developed by a firm. Innovation, in this context, is defined as the introduction of a new or significantly improved product or process (5). Most of the information on websites is codified as text and extracting innovation-related information from these web texts and transferring it into a credible firm-level innovation indicator is the aim of this study.

Text mining algorithms can be used to extract knowledge from large document collections and turn them into valuable economic information (6–10). As a result, text mining became one of the most promising approaches in economic analysis to provide novel tools and insights to economists. At the methodological level, great progress has been made in natural language processing (NLP), driven by the rapid increase in computational power and textual data availability (11). Especially neural networks have shown very promising results when used for the classification of text documents into certain categories (12, 13).

We utilize information from the Mannheim Innovation Panel (MIP), a questionnaire-based innovation survey of firms, to label the websites of surveyed firms as either innovative or non-innovative. This labelled data set is then used to train a deep neural network. The resulting classifier determines the innovativeness of firms based solely on their website text. Figure 1 outlines our proposed approach. The predicted innovation probabilities can directly be interpreted as a continuous firm-level indicator of innovation.

Two research questions are meant to help us assess the credibility of our proposed continuous innovation indicator:

- **Research Question 1**: Can deep neural networks be used to reliably classify innovative and non-innovative firms solely based on their website texts?
- **Research Question 2**: Does such a classification model produce a credible sectoral and regional pattern of innovative and non-innovative firms when applied to a large out-of-sample dataset of firm website texts?

The remainder of this paper is structured as follows. First, we present our data, followed by our methods. Our results section is twofold. In the first part, we present the results of our artificial neural network classification model. In the second part, we apply our classification model to a large dataset of German firms to predict their innovativeness. The results are discussed and summarized in the last two sections.

[2]To whom correspondence should be addressed. E-mail: jan.kinne@zew.de or david.lenz@wi.jlug.de

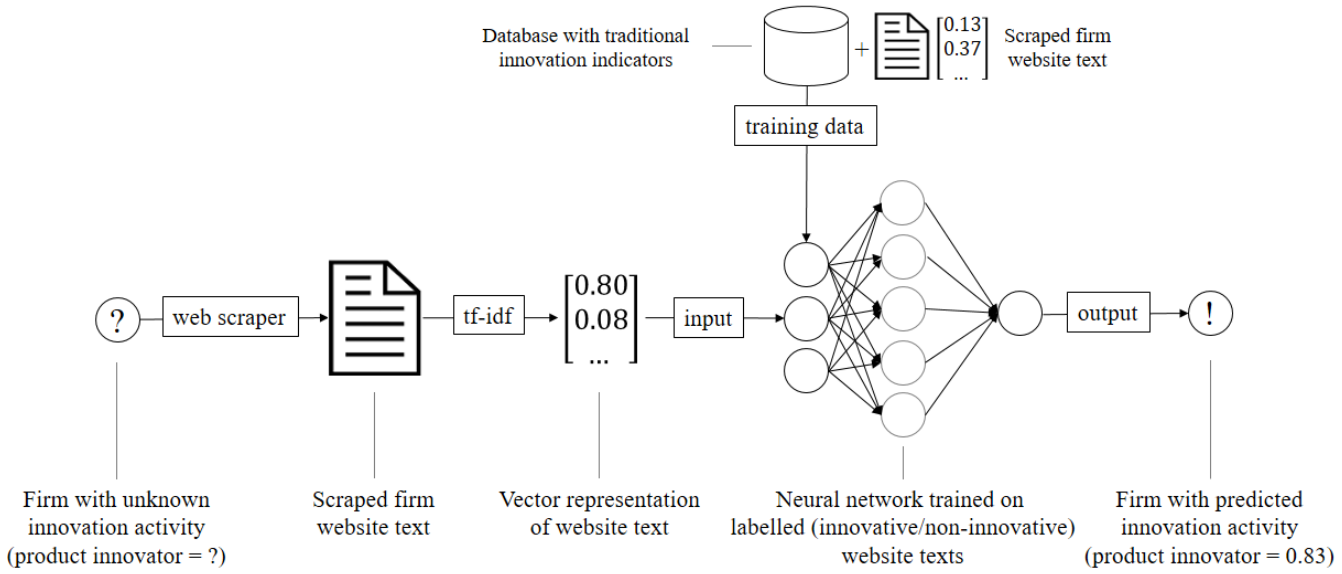

Fig. 1. Proposed firm innovation prediction model. Web scraped texts from firms with know innovation status are used to train a neural network to predict firms' innovation status from unlabelled website texts.

Data

In the following section we present our base datasets, the Mannheim Enterprise Panel (MUP), a firm database containing all economically active firms in Germany, and its derivative, the Mannheim Innovation Panel (MIP), which is a annual survey on innovation activities in firms sampled from the MUP database. The website texts of firms in both the MUP and MIP were scraped using ARGUS web scraper (14).

MUP firm panel dataset. The Mannheim Enterprise Panel (MUP) is a panel database which covers the total population of firms located in Germany. It contains more than three million firms which are updated on a semi-annual basis. The data covers firm characteristics such as the industry (NACE codes; a classification of economic activities in the European Union), postal addresses, number of employees, as well as the website address (URL) of the firm. For more information on the MUP see Bersch et al. (2014).

For our analysis, we use MUP data restricted to firms that were definitely economically active in early 2018. The resulting dataset contains 2.52 million firms and 1.15 million URLs (*URL coverage* of 46%). A prior analysis of this dataset by (4) showed that URL coverage differs systematically with firm characteristics. Only a fraction of very young (younger than two year) and very small firms (fewer than five employees) are covered by an URL after controlling for the search quality of the data provider. They also find sectoral and regional differences. Some regions (especially those with low broadband Internet availability) and some sectors, like agriculture, exhibit lower URL coverage. However, given that most of the innovative activity is conducted by middle-sized and larger firms (15), which are well covered in the dataset, they conclude that the data is suitable to analyze the German innovation system.

MIP innovation survey data. The Mannheim Innovation Panel (MIP) is an annual questionnaire-based innovation survey of firms sampled from the MUP database. The survey is designed as a panel survey, such that firms in the sample are surveyed every year. Firm closures and mergers are substituted by randomly sampled additional firms every two years. The MIP is the German contribution to the Community Innovation Survey (CIS), which is conducted every two years in the European Union. CIS data has been used as base data in an array of studies (16). The survey methodology and definition of innovation follows the "Oslo Manual" (5) and covers firms with five or more employees in manufacturing and business-oriented services. Each year, firms are asked whether they introduced new or significantly improved products ("product innovations") and/or implemented new methods of production or the delivery of products and services ("process innovations") during the three years prior to the survey. In our study, we use the firms' status as product innovators (yes/no) as the target variable, because we assume that firms are more likely to disclose product innovations than process innovations on their websites.

The sampling procedure of the survey (oversampling industries and size classes where innovation in more prevalent) results in an over-representation of innovative firms, such that 36% of the firms in our database are product innovators. Projected to the overall firm population in Germany, the share of product innovators can be expected to be in the range of 25% for the target population in the MIP (17)). This oversampling of innovative firms and a restriction to firms with at least five employees results in a dataset, in which firms are larger, in terms of their number of employees, than firms in the overall firm population in Germany. The median number of employees in our dataset is 20 and the mean is 246.8.

We use MIP survey data of 2017, 2016, and 2015. As each annual MIP survey, following the CIS, covers a three-year reference period (the three years preceding the year the survey is conducted), our data covers a time period from 2012 to 2016. Given that we use web texts scraped from the firms' websites in 2018 and match it to survey results covering the year 2016,

some firms may have changed their innovation status during this time lag. A relevant fraction of firms actually change their innovation status between years, such that they may be an innovator in 2015 but not in 2016 (18). We decided to cope with this issue by restricting our analysis to firms that had a "stable" innovation status in the surveys of 2015, 2016, and 2017. This means they were either product innovators in all of these years or none of these years. We assume that such firms are less likely to have changed their innovation status between the survey of 2017 and our test day in 2018. This approach reduces our sample of firms (and our training data of web texts of firms with a known innovation status) from 18,062 to 4,481. In the restricted sample, 32% of the firms are product innovators. The mean number of employees is 277.5 and the median is 23. Very young firms cannot be in our sample, as they have to have taken part in the survey at least three times. Nevertheless, we are able to show that this approach results in a higher quality of our training data (see Discussion section).

ARGUS web texts. We used ARGUS (4), a free web scraping tool based on the Scrapy Python framework, to scrape texts from the websites of both MUP and MIP firms. We used ARGUS simple language selection heuristic (set to German), which was shown to help limiting the downloaded texts to a certain language (4). According to (4), about 90% of the webpages downloaded this way can be expected to be in German, with some sectors, like the pharmaceuticals and mechanical engineering sector, exhibiting higher shares of non-German webpages (most of them in English).

Webpages are not downloaded randomly from the firms' websites. Instead, ARGUS starts at a firm's main page ("homepage") and continues downloading webpages with the shortest URL. The rationale is that more general information on the firm is available at the top level of the website (e.g. "firm-name.com/products", "firm-name.com/team"), which should be given priority over more specific information (e.g. "firm-name.com/news/2017/august/most_read"). This top-level approach is intended to generate firm-level business activity profiles instead of specific product-level descriptions found on low-level webpages. Even though the latter may inform about individual product innovations, we think that the top-level description of firms may allow the neural network to learn combinations of more general signal words that reliably predict innovative firms, regardless of what exact product innovations they implemented. Such a generalization would be desirable, especially as our training data consists of firms from both ends (consistently innovative or consistently non-innovative) of the overall firm distribution.

The number of downloaded webpages per firm website is defined by a limit parameter in ARGUS, which was set to 25 in this analysis. Hereby, we follow the finding by (4) that 50% of all firm websites can be scraped completely when this limit parameter is set to 15. Reaching 90% would require raising this threshold to 250, highlighting that web-based studies have to deal with outlier websites, as some firms (especially large ones) have massive websites with ten-thousands of webpages. (4) found that the amount of text per webpage is not statistically related to a firm's age, size, and sector. Following the suggested practice by (4), we excluded websites which redirect to a different domain when requesting their first webpage (i.e. homepage). A practice which should ensure that crawled websites belong to the corresponding firms. This was also shown to not result in any sectoral or firm age selection bias. Table 1 presents the resulting database after excluding such initial redirects and download errors caused by non-existing websites.

Training dataset of bloat webpages. During web scraping, all texts found on a firm's website are downloaded, regardless of their content and relevance to the study. Alongside valuable web texts describing the firm itself, products, employees, employed technologies, and the like, web texts from imprints, legal information, and HTTP cookie pop-ups are downloaded as well. We also face the problem that our textual data is highly ambiguous in the sense that many websites share common features, e.g. login pages or contact and legal sections. To filter webpages which contain text of mostly unwanted nature (*bloat webpages*), we created a dataset of webpages labeled as either bloat (containing unwanted information) or gold (containing relevant information) which can be used to train a bloat/gold classification model.

For this purpose, we sampled 10,000 firms from our MUP base dataset. Subsequently, we used ARGUS to scrape their websites with a limit of 100 webpages per website and German as the preferred language. We then kept only non-empty webpages written in German (as classified by Python's langdetect library; (19)). From this sample, we drew 10,000 webpages of which 8,080 could be unambiguously labeled as either gold or bloat by hand.

Methodology

In this section, we present how our web texts were preprocessed and transferred to term frequency–inverse document frequency (tf-idf) vectors. Tf-idf represents documents as a fixed size vector by counting words in each document (term frequency) and weighting each frequency by the inverse of the term's overall document frequency. We then describe the architecture of our deep neural network model for binary text classification, and how we evaluate the model's classification performance.

Web text preprocessing. We reduced the preprocessing of our texts to a minimum. The scraped web texts were standardized to lowercase and all characters not in the German alphabet were removed (keeping *Umlaut* special characters, whitespaces, and ampersands). Tests with word stemming procedures, which reduce words to their stem (e.g. "innovation" and "innovator" to "innovat"), did not increase our classification performance and we refrained from using it.

Web texts as numerical tf-idf vectors. We use the term frequency–inverse document frequency (tf-idf) scheme to represent the website texts as sparse vectors (see e.g. Manning et al. 2009). The tf-idf algorithm transfers each document to a fixed size sparse vector of size V, where V is the size of a dictionary composed of unique words. We restricted our dictionary to words with a minimum document frequency of 1.5% and a maximum document frequency of 65% (*popularity based filtering*), resulting in a dictionary size V of $6,144$ words. Each entry in the tf-idf vector of a document corresponds to one word in the dictionary, representing the relative importance of this word in the document. Words that do not appear in a given document are represented by a 0 value. Specifically, in a first step (the tf step) the number of appearances per

Table 1. Firms in datasets after filtering steps.

dataset	base data	unstable innovation	no URL	errors/redirects	final data
MUP	2,523,231	N/A	-1,374,383	-463,791	685,057
MIP	18,062	-13,587	-456	-893	3,126

word in a single document are counted. In a second step, the inverse document frequency (idf) is used as a weighting scheme to adjust the tf counts. Conceptually, the idf weights determine how much information is provided by a specific word by means of how frequently a word appears in the overall document collection. The intuition is that very frequent words that appear in a lot of documents, should be given less weight compared to less frequent words, as infrequent words are more useful as a distinguishing feature.

Web text classification with a deep neural network. Deep neural networks showed remarkable success when applied as text classification models (12, 13). Different deep neural network architectures were proposed and showed varying performance in NLP tasks. We tried different neural network architectures (convolutional neural networks, recurrent neural networks, both with long short-term memory and gated recurrent units) and also compared their performance in our specific classification task with more traditional models (naive Bayes classifier, logistic regression, decision trees). In this iterative process, an *undercomplete autoencoder-like neural network* architecture turned out to be the model with the best classification performance. Autoencoder-style neural networks (see e.g. (20)) impose a "bottleneck" (hidden layers with very few neurons) in the network architecture which are intended to force the learning of a highly compressed representation of the network's input. While the output of a standard autoencoder network has the same dimensionality as the network's input, the output of an undercomplete autoencoder network has a smaller dimension than the network's input.

Our network consists of four hidden layers with intermediary dropout layers, which are intended to improve the network's generalization by ignoring (*dropping*) neurons during the training phase (21). The network's first hidden layer consists of 250 neurons, the following two hidden layers consist of only five neurons each (the "bottleneck"), while the forth and last hidden layer contains 125 neurons. We used *scaled exponential linear units* (SELU, (22)) as activation functions in the hidden layers. The network's output layer consists of a single neuron with a *sigmoid* activation function, a common approach to receive an output between 0 and 1 from a neural network in binary classification tasks (see e.g. (23)). We used the common Adam optimizing algorithm (24) for the stochastic optimization of the network weights.

Results

We first present the results of our bloat webpage classification model. We then use the model to filter bloat webpages from our dataset of firms with available innovation indicators from the MIP innovation survey. Based on the filtered dataset we train our innovation prediction model and test the model's performance using a retained part of the training data. In the last part of this section, we apply the innovation prediction model to about 700,000 firms from the MUP to predict their innovation status and examine the resulting sectoral and regional patterns.

Bloat webpage filtering. Training our classification model with the bloat webpage training data and testing it with a retained part of the bloat webpage data (test set), resulted in a precision, recall, f1-score and support indicated in Table 2. The *precision* score of 0.81 indicates that the trained model is correct in 81% of cases if the predicted label is "bloat" (i.e. in 19 % of cases the prediction is bloat even though the webpage is *no bloat*). Out of all bloat webpages, we identify 48% of webpages correctly (*recall* of 0.48) as being bloat, but fail to detect 52% of bloat webpages. Combining precision and recall by applying a harmonic mean, results in a *f1-score* of 87%. *Support* indicates the respective number of cases. Thus, while having high precision, the recall of the bloat class leaves room for improvements. However, in our case we think it is reasonable to prefer a high precision over high recall, as we only want to dismiss webpages that are certainly not relevant.

Table 2. Bloat classification report for test set.

label	precision	recall	f1-score	support
bloat	0.81	0.48	0.61	368
no bloat	0.89	0.97	0.93	1652
avg / total	0.88	0.89	0.87	2020

Based on these findings, we decided to set the threshold of the classification model to 0.9, i.e. we only kept a webpage if the model was certain that the webpage is no bloat with high confidence (probability(no bloat) $>$ 0.9). This filtering step resulted in the exclusion of 309 firms because their websites consisted of bloat webpages only.

Innovation prediction model. After filtering bloat webpages from our MIP dataset, we aggregated all remaining webpages to the firm level, keeping only the first 5,000 words per firm. We randomly selected 75% of this data as training data for our innovation classification model and retained 25% as test data. Table 3 details precision, recall, f1-score, and support for the resulting classifier applied to the test set (classification threshold for the probabilities of 0.5). If the model classifies a firm as innovative, it is correct in roughly 4 out of 5 cases, as can be seen by the 81% precision for the innovative class. The model retrieves 64% of all innovative firms and 91% of all non-innovative firms in the test dataset (recall). The overall f1-score of the model is 80%.

Table 3. Innovation classification report for test set.

label	precision	recall	f1-score	support
non-innovative	0.81	0.91	0.86	429
innovative	0.81	0.64	0.71	255
avg / total	0.81	0.81	0.80	684

Investigating the performance of the model within sectors and size classes is difficult, as our test dataset is small and

some classes are not covered at all or contain only very few observations. Due to this limitation we decided to use the trained model to predict product innovator probabilities for both the training set and the test set to assess the sectoral and size pattern yielded by the model. This approach is not suitable to obtain an unbiased evaluation of the model because overfitting cannot be identified when using training data sets for prediction. We can, however, use the resulting predictions to examine the model's fit within sectors and size classes (an information which was available to the neural network only implicitly through interpretation of the web texts).

Figure 2 shows that our model generally underestimates the share of product innovators in most sectors except for wholesale and information, technology and communication (ICT) services. This underestimation is also reflected in a lower total share of product innovators (27.45%) compared to the share obtained from the true labels (33.29%).

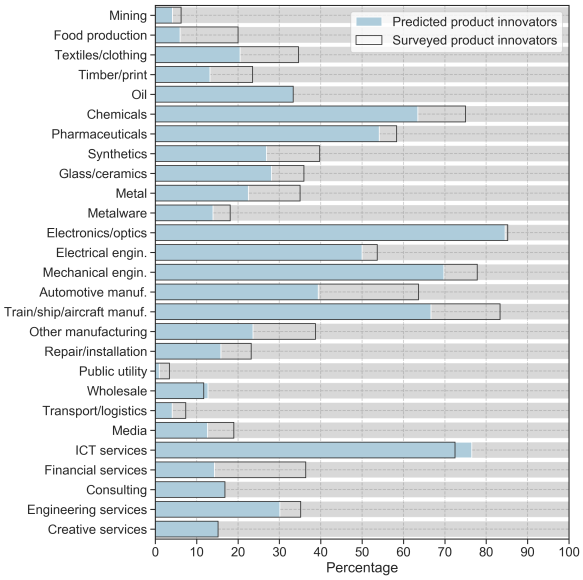

Fig. 2. Product innovator firms in training data by sector. Predicted (blue) and true (transparent) shares of product innovators.

The left panel of Figure 3 shows a non-linear relationship between firm size and the predicted product innovator probabilities, with mid-sized (around 1,000 employees) and very large firms (more than 3,000 employees) having the highest predicted product innovator probabilities. The same relationship cannot be seen when a binary classification is used (right panel). There, the above-mentioned underestimation of product innovators can be seen as well.

To assess the adequacy of our training data selection approach, we reran the entire procedure of web scraping, text preprocessing, model training and testing using three alternative datasets. First, we used only the product innovator variable from the 2017 survey (the same survey we used to create our main training dataset) instead of creating our "stable innovator" training dataset. Using this significantly larger dataset (11,506 firm websites) resulted in a f1-score of just 68% in the corresponding test set (see Table 4).

Second, we used the product innovator variable of the more recent MIP of 2018. The results in Table 5 show that this

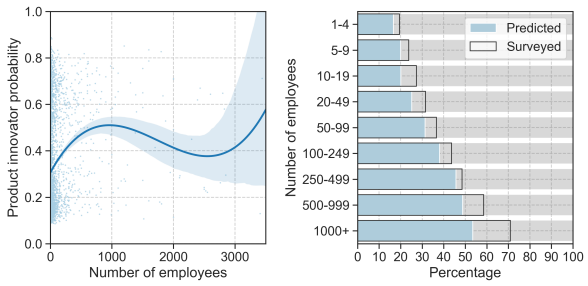

Fig. 3. Product innovator firms in training data by size. Left panel: Product innovator probabilities by firms size (number of employees) with fitted third order polynomial regression line with 95% confidence interval. Right panel: Share of product innovator firms by size classes (predictions in blue; true labels transparent).

Table 4. Innovation classification report for alternative data test set A.

label	precision	recall	f1-score	support
non-innovative	0.72	0.83	0.77	1,784
innovative	0.62	0.47	0.53	1,093
avg / total	0.68	0.69	0.68	2,877

convergence of survey data and web data results in a better f1-score, compared to the results using the same survey variable with a one year longer time lag (see Table 4).

Table 5. Innovation classification report for alternative data test set B.

label	precision	recall	f1-score	support
non-innovative	0.75	0.88	0.81	1,264
innovative	0.61	0.38	0.47	601
avg / total	0.70	0.72	0.70	1,865

Third, we used an alternative product innovator variable of the MIP 2018 which relates directly to the year of the survey. Instead of asking about product innovations introduced by the firm in the three consecutive years prior to the survey, product innovations in 2018 are surveyed. This survey data, which covers about the same time as the web data scraped in 2018, increases the predictive performance of the model in the corresponding test set even more (f1-score of 0.74; Table 6).

Table 6. Innovation classification report for alternative data test set C.

label	precision	recall	f1-score	support
non-innovative	0.79	0.95	0.86	751
innovative	0.63	0.26	0.37	258
avg / total	0.75	0.77	0.74	1,009

Out-of-sample innovation prediction. Figure 4 shows the distribution of product innovator probabilities for 685,057 MUP firms. The mean is 0.253 and the median is 0.203. The lowest predicted probability is 0.029 and the highest is 0.944.

Converting the firms into either innovators or non-innovators (to make them comparable to existing survey benchmark data), we are required to set a rather arbitrary classification threshold. Setting this threshold to 0.5 (the same that was

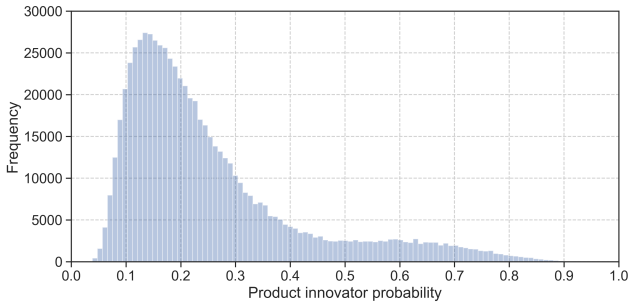

Fig. 4. Product innovator probabilities distribution. Histogram of predicted product innovator probabilities for 685,057 firms.

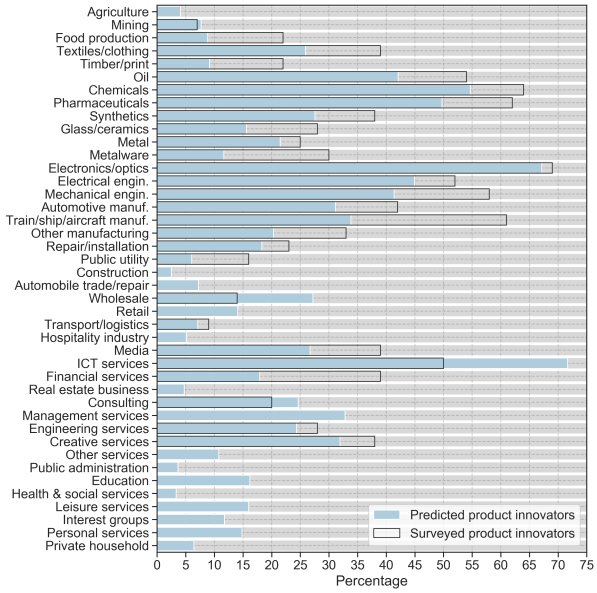

Fig. 5. Predicted product innovator firms by sector. Share of product innovator firms with five or more employees by sector. Blue bars indicate the predicted shares. Transparent bars indicate extrapolated shares from the MIP innovation survey.

used during the model's training) would result in 10.31% of the firms being classified as product innovators. As the MUP dataset covers the entire range of firm types in Germany, we do not only apply our innovation prediction model to classify firm types that are not covered in the training data (see Data section), it also means that there is no reference value for the total share of innovative firms in the overall German economy. We can, however, compare our prediction results to extrapolated MIP survey results in those sectors and firm size classes which are covered by the survey. The raw survey results are used to calculate weighted results at sector and size class level that are representative for the total firm population of the survey (firms with five or more employees in manufacturing and business-oriented sectors) (15).

According to the MIP extrapolations, the share of product innovators in manufacturing and business-oriented services is at 27%. 89,372 of the MUP firms have at least five employees and are from one of the covered sectors. 21% of those firms are predicted to be innovative using a classification threshold of 0.5. This result confirms the underestimation tendency of our model (see last section). To adjust for this discrepancy, we decided to calibrated our classification threshold to a value that produces the same number of innovative firms anticipated by the survey extrapolation benchmark of 27%. The calibrated classification threshold of 0.401 was subsequently used to label all 685,057 MUP firms as either innovative or non-innovative, resulting in an increase of the total share of product innovators from 10.31% to 15.12%.

Figure 5 presents the share of product innovator firms by sector after applying the calibrated classification threshold of 0.401. We also indicate the share of product innovator firms by sector as they are calculated from the MIP questionnaire-based survey (transparent bars) if available for the respective sector. Even though the overall trend and the proportions between sectors are similar to the survey benchmark, underestimation can be seen in all sectors except for wholesale, consulting, and especially ICT firms. For sectors without a survey benchmark, assessing the results is difficult, but overall these predictions look decent. Very low shares of product innovators in construction and agriculture, for example, and higher shares in management services is what was to be expected.

Figure 6 shows a breakdown of our predictions by firm size (number of employees). In the left panel, the number of employees is plotted against the predicted product innovator probabilities for all sectors (blue). ICT service (purple) and mechanical engineering firms (green) are also plotted as exemplary sectors. It can be seen that the fitted polynomial regression lines of third order indicate the same positive nonlinear relationship between the size of firms and their product innovator probabilities we saw in the training data already. The right panel of Figure 6 plots the share of product innovator firms by size groups for sectors covered in the MIP survey (blue) and the corresponding survey extrapolation benchmarks. It can be seen that our predictions match the survey benchmark very well, except for very large firms with more than 1,000 employees where we underestimate the share of product innovators by about 20 percentage points.

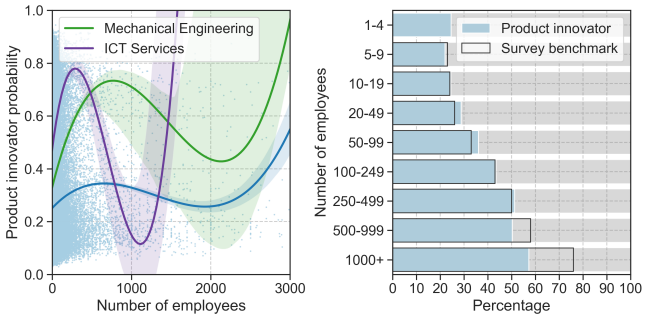

Fig. 6. Predicted product innovator firms by size. Left panel: Product innovator probabilities by firms size (number of employees) with fitted third order polynomial regression line with 95% confidence interval; all sectors (blue), ICT services (purple), mechanical engineering (green). Right panel: Share of product innovator firms by size classes (predictions in blue; survey extrapolation transparent).

Figure 7 maps the predicted ratio of product innovators to all firms by 432 German district. In general, city districts exhibit higher shares of innovative firms. This fact is confirmed by a high and significant Spearman's correlation coefficient of 0.61 between district population density (a proxy for urbanity)

and the ratio of innovative firms. It can also be seen that the vicinities of some major agglomeration areas in the South-West of Germany (Munich, Nuremberg, Stuttgart, Rhine-Neckar region around Mannheim, and Rhine-Main region around Frankfurt) exhibit high shares of innovative firms as well.

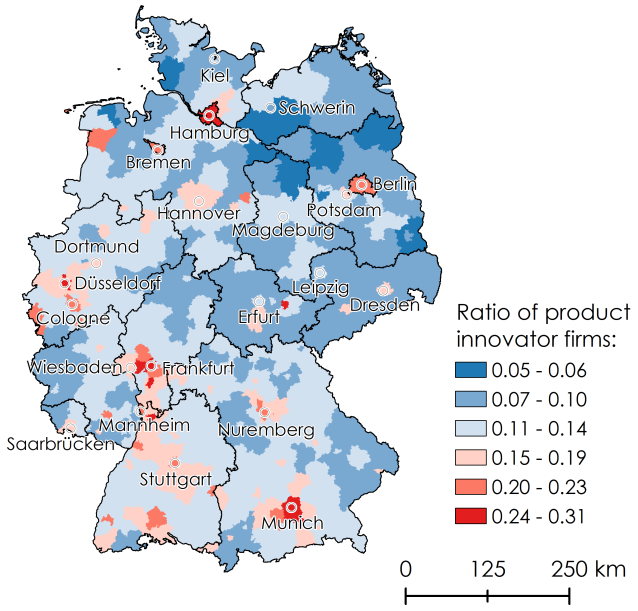

Fig. 7. Product innovator firms by district. Ratio of predicted product innovator firms to all firms by district.

The detailed address information in our firm database allows us to disaggregate the geographic pattern shown above and to analyze individual regions at a microgeographic level. For the German capital of Berlin, a special survey of the MIP is conducted every year (25), covering a high share of firms from manufacturing and business-oriented services in Berlin. This comprehensive dataset allows us to map the density of our predicted product innovators in Berlin against the observed density from the MIP special survey (see Figure 8).

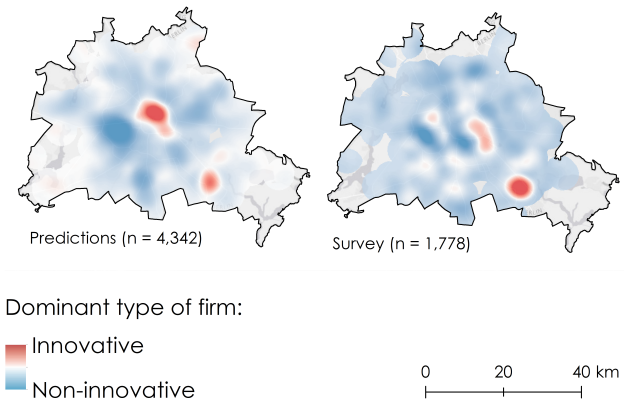

Fig. 8. Product innovator firms in Berlin. Predicted (left) and surveyed (right) densities of product innovator firms in Berlin.

For the density map in the left panel, we selected firms (4,342 of 35,998) that are from sectors and size classes covered in the survey. The right panel shows the same map for 1,778 firms that answered the MIP questionnaire. Both firm location patterns were used to calculate kernel density maps using the same set of parameters. It can be seen that the two densities resemble each other in their overall appearance, with major hotspots in the eastern city center of Berlin (city districts of Mitte, Prenzlauer Berg, and Friedrichshain-Kreuzberg), as well as the area around Adlershof (a major science and technology park) in the South-East.

Discussion

Based on a f1-score of 0.87, the overall performance of our bloat webpage classification model can be assessed as being good but not perfect. The high recall of 0.97 for non-bloat webpages ensures that only few valuable information is lost during the filtering process. The recall of 0.48 for bloat webpages is not very high, but at least half of the unwanted webpages are filtered at a classification threshold of 0.5. Given these results and the fact that we have quite a lot of webpages for each firm, we decided to set a rather high classification threshold of 0.9 to ensure that most unwanted webpages are filtered and only highly relevant ones remain in the dataset.

The innovation prediction model's f1-score in the test set turned out to be 0.80. Recall for non-innovative firms is substantially higher than for innovative ones, which suggest that a rather low classification threshold should be chosen if we want to recover most of the innovative firms in the overall firm population. Precision is balanced for both labels (0.81). Overall, we are satisfied with the performance of the model in the test dataset, especially given that our training (2,531 observations) and test sets (684 observations) are rather small. Our investigation of sectoral and firm size patterns based on the entire training dataset showed that the model generally underestimates the share of product innovators when using the training step classification threshold of 0.5. We also compared the predictive performance of our model using different training datasets to validate our original approach of using only "stable (non-)innovators" for training. The results confirmed that we were indeed able to extract firms with clearer business profiles in our original approach that eventually allowed our neural network to learn web text features to distinguish between innovators and non-innovators. However, using more recent survey data, we were also able to show that the model's predictive performance benefits from a smaller time lag between survey data and web data as well.

The classification of 685,057 out-of-sample MUP firms resulted in a predicted product innovators probability distribution where most firms have a probability between 10% and 40%. 10.31% of all firms would be classified as product innovators when using the training step classification threshold of 0.5. As there is no reference value for the overall population of firms in Germany, we decided to compare our prediction results within the same target population that is used in the MIP innovation survey and for which extrapolated reference numbers are available (15). In total, 89,372 MUP firms fall into this group (manufacturing and business-oriented services; five or more employees). Within this group, a classification threshold of 0.5 results in a 21% share of product innovators, just short of the surveyed 27%. We decided to calibrate the

classification threshold to 0.401 such that we classify the same share of firms as innovative. We also used this threshold to classify MUP firms from sectors and size classes not covered in the MIP, as there are no better reference values. This resulted in an increasing of the total share of innovative firms in the MUP dataset from 10.31% (0.5 threshold) to 15.12% (0.401 threshold).

The breakdown by sector allowed us to compare our model's results to the sector-level reference results from the extrapolated MIP survey. Overall, the sectoral pattern (proportions between sectors) is similar to the MIP benchmark. However, our model underestimates the share of product innovator firms in most sectors for which a reference value is available. Wholesale and ICT services, however, are exceptions, as our model overestimates the share of product innovators in these sectors. Concerning the wholesale sector, we assume that our model may not be able to distinguish between products produced or just sold by a firm. This may lead to a high share of assumed product innovators in the wholesale sector that are actually just presenting products of other firms on their website. One possible explanation for the overpredicted share of product innovators in the ICT service sector is that the "tech" sector is nowadays widely considered the sector with the most innovative and future-oriented technologies (with buzz words like Digitalization, Industry 4.0, Internet of Things, Artificial Intelligence and the like). Firms with an innovative agenda or self-concept may use these technologies (or at least the associated buzz words) and mention them on their websites. This could result in a bias in our classification model such that the neural network learns to over-relate these words to innovativeness. ICT firms, which by nature use tech vocabulary on their websites, may then be classified as innovative too often. Very preliminary analysis concerning the importance of word features using SHAP (26) points in exactly this direction and suggest that tech sector affiliated words ("software", "data", "cloud" etc.) may indeed play an important role during the classification.

The share of product innovator firms in sectors for which no survey benchmark is available can be considered to be reasonable and indicate that our model can be used for innovation prediction in sectors for which no training data is available. However, we assume that the retail sector, for example, may suffer from overestimated product innovator shares for the same reasons as the wholesale sector.

In conclusion, we suggest to use the raw (continuous) product innovator probabilities for future studies using our innovation prediction approach. If a binary indicator of innovation is needed, sector-level classification thresholds should be used to cope with the bias we found to be present in our predictions. Given that survey data is available at the sector level, researchers may want to select different classification threshold for each sector such that the predicted share of product innovators matches the shares from the extrapolated survey.

The breakdown by firm size revealed an interesting, nonlinear relationship between firm size and predicted innovativeness. In the aggregate, product innovator probability peaks at 500 to 1,000 employees. This effect is even more distinct for the exemplary sectors of ICT services and mechanical engineering. ICT firms were predicted to be the most innovative at a rather small size of about 300 employees. Mechanical engineering firms plateau between 600 and 1,000 employees. The well-known German *Mittelstand* (the bulk of mid-sized and highly innovative German firms) seems to be identified here, emphasizing the power of our model. Compared to the MIP survey benchmark, our model almost perfectly predicts the share of product innovators for all size classes. Only for very large firms with 1,000 and more employees, our predictions are clearly below the MIP benchmark. This has to be examined in follow-up studies.

The regional patterns of our predictions turned out to be highly credible. The observed East-West, North-South, and urban-rural trends are well documented in the literature (27). The innovation density map of Berlin highlights two things. First, our model's results compare very well to the benchmark data from the MIP special survey of Berlin and similar hotspots of innovation were identified in both patters (city center East and technology park Adlershof in the South-East). Again, we think that our model's bias towards the ICT sector may be the cause for a more pronounced innovation hotspot in the eastern city center, an area with exceptionally high shares of firms from this sector (28). Second, our predicted continuous indicator of innovation can be used to conduct large-scale analysis of regions in any desired geographical resolution, from individual firm locations to aggregated geographical units. The latter can be considered an important contribution because it allows scientist to analyze innovation policies with unprecedented regional and sectoral granularity.

Conclusion

In this paper, we presented a novel approach on how to predict a continuous and highly granular firm-level indicator of innovation using deep learning and web mining. We motivated our approach with the need to provide innovation policy making with an innovation indicator that overcomes some of the limitations of traditional indicators from questionnaire-based surveys or patents. Using web texts of firms surveyed in a traditional innovation survey as training data, we created a neural network classification model which predicts the innovation probability of firms using only their website texts. Our choice of training data, as well as our web text selection and preprocessing procedure were intended to allow the neural network to learn firms' business activity profiles. Eventually, we do not identify distinct product innovations, but firms with business activity profile that make product innovations very likely. This likelihood (i.e. probability) can be interpreted as a continuous firm-level innovation indicator. The following two research questions were intended to answer the question on the credibility of this novel innovation indicator.

RS1: Innovation prediction model performance. We concluded that our innovation prediction model offered a good performance within the test dataset of MIP surveyed firms. Looking at the predictions for both the test and the training data, we found that our model tends to underestimate the share of product innovators firms, which is reflected in a rather low recall concerning innovative firms. We also found first evidence that our model may be positively biased towards firms from the ICT sector.

RS2: Patterns from out-of-sample innovation prediction. The prediction of our proposed continuous innovation indicator for 685,057 out-of-sample firms resulted in credible

sectoral, size, and regional patterns. We concentrated mainly on comparing a subset of firms for which extrapolated survey reference values are available. We also used these survey extrapolations to calibrate the classification threshold which was applied to transfer our continuous firm-level product innovator probabilities to a binary variable (innovator/non-innovator). The resulting sectoral pattern followed the same trend anticipated by the survey extrapolation benchmarks with a positive bias towards ICT firms, resulting in an highly overestimated share of product innovators in this sector and underestimated shares in most other sectors. We recommended to calibrate individual classification thresholds at the sector level for future research to cope with this bias, if suitable survey data is available. Concerning the relation between firm size and our predicted product innovator probabilities, we found an interesting, non-linear relationship that seems to identify innovative and mid-sized German *Mittelstand* firms. Aggregated to size groups, our predictions almost perfectly match the extrapolated survey benchmarks. Finally, we were also able to show that our novel indicator exhibits very similar micro-geographical patterns compared to the MIP benchmark data, making us confident that we created a highly valuable tool for scientist to analyze innovation at any geographical and sectoral scale.

Future research. Future research should concentrate on both the methodological development and the application of our approach. Methodologically, it would be interesting to further investigate which words and word combinations have the biggest on the neural network's prediction outcome. Additional development on the network's architecture and additional training data, as well as a different preprocessing of the training data, could lead to better prediction performance. Our proposed approach could also be applied to other target variables from surveys in economics (e.g. process innovators) or other fields of social science. Empirical follow-up studies could apply our proposed approach to a wide array of research questions, from innovation policy evaluations to the analysis of knowledge spillovers and technology diffusion. Frequent crawling of firm websites would allow us to build up a panel database of web-based innovation indicators suitable for time-series analysis.

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) of which this study is a part. We also want to thank Martin Hud, Christian Rammer, Georg Licht and Peter Winker for their valuable input.

1. Nagaoka S, Motohashi K, Goto A (2010) Patent Statistics as an Innovation Indicator in *Handbook of Economics of Innovation*, eds. Hall BH, Rosenberg N. Vol. 2 edition, pp. 1083–1127.
2. Squicciarini M, Criscuolo C (2013) Measuring Patent Quality.
3. OECD (2009) *OECD Patent Statistics Manual.* (OECD, Paris), p. 162.
4. Kinne J, Axenbeck J (2018) Web Mining of Firm Websites : A Framework for Web Scraping and a Pilot Study for Germany.
5. Eurostat, OECD (2005) *Oslo Manual: Guidlines for Collecting and Interpreting Innovation Data.* (OECD) Vol. Third edit, p. 162.
6. Levenberg A, Pulman S, Moilanen K, Simpson E, Roberts S (2014) Predicting Economic Indicators from Web Text Using Sentiment Composition. *International Journal of Computer and Communication Engineering* 3(2):109–115.
7. Larsen VH, Thorsrud LA (2015) The Value of News, (Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School), Technical Report 6.
8. Lüdering J, Winker P (2016) *Forward or Backward Looking ? The Economic Discourse and the Observed Reality.* (MAKGS Joint Discussion Paper Series in Economics).
9. Grentzkow M, Kelly BT, Taddy M (2017) Text as Data.
10. Rönnqvist S, Sarlin P (2017) Bank distress in the news: Describing events through deep learning. *Neurocomputing* 264:57–70.
11. Schmidhuber J (2015) Deep learning – An overview. *Neural Networks* 61:85–117.
12. Kim Y (2014) Convolutional Neural Networks for Sentence Classification in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* (Association for Computational Linguistics, Doha, Qatar), p. 1746–1751.
13. Yang Z, et al. (2016) Hierarchical Attention Networks for Document Classification.
14. Kinne J (2018) ARGUS - An Automated Robot for Generic Universal Scraping.
15. Rammer C, Aschhoff B, Doherr T, Peters B, Schmidt T (2017) Innovationsverhalten der deutschen Wirtschaft, (Centre for European Economic Research (ZEW), Mannheim), Technical report.
16. Gault F, Aho E, Alkio M, Arundel A, Bloch C (2013) *Handbook of Innovation Indicators and Measurement* ed. Gault F. (Edward Elgar Publishing Ltd, Glos, UK), p. 486.
17. Rammer C, et al. (2019) Innovationen in der deutschen Wirtschaft, (ZEW Centre for European Economic Research, Mannheim), Technical report.
18. Peters B (2009) Persistence of innovation: Stylised facts and panel data evidence. *Journal of Technology Transfer* 34(2):226–243.
19. Danilak M (2015) langdetect.
20. Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning.* (MIT Press, Cambridge, Massachusetts).
21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15:1929–1958.
22. Klambauer G, Unterthiner T, Mayr A, Hochreiter S (2017) Self-Normalizing Neural Networks. *Advances in Neural Information Processing Systems* 30(Nips):99–112.
23. Manning CD, Raghavan P, Schutze H (2009) *An Introduction to Information Retrieval.* (Cambridge University Press, Cambridge, England), Online edi edition, p. 569.
24. Kingma DP, Ba JL (2015) ADAM: A Method for Stochastic Optimization in *ICLR Conference Paper.* p. 15.
25. Feser D (2018) Innovationserhebung Berlin 2017, (Technologiestiftung Berlin, Berlin), Technical report.
26. Lundberg S, Lee SI (2017) A Unified Approach to Interpreting Model Predictions. 16(3):426–430.
27. European Commission (2017) Regional Innovation Scoreboard 2017, (Bruxelles), Technical report.
28. Rammer C, Kinne J, Blind K (2019) Knowledge Proximity and Firm Innovation: A Microgeographic Analysis for Berlin. *Urban Studies* forthcomin.

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
Phone +49 621 1235-01
info@zew.de · zew.de