// NO.22-065 | 12/2022

DISCUSSION PAPER

// JANNA AXENBECK AND PATRICK BREITHAUPT

Measuring the Digitalisation of Firms – A Novel Text Mining Approach

Leibniz Association

ZEW

Measuring the Digitalisation of Firms – A Novel Text Mining Approach

Janna Axenbeck[*†], Patrick Breithaupt[*†§]
December 21, 2022

Abstract

Due to the omnipresence of digital technologies in the economy, measuring firm digitalisation is of high importance. However, current indicators show several shortcomings, e.g., they lack timeliness and regional granularity. In this study, we show that advances in text mining and comprehensive firm website content can be leveraged to generate real-time and large-scale estimates of firm digitalisation. We use a transfer learning approach to capture the latent definition of digitalisation. For this purpose, we train a random forest regression model on labeled German newspaper articles and apply it on firm's website content. The predictions are used as a continuous indicator for firm digitalisation. Plausibility checks confirm the link to established digitalisation indicators at the firm and sectoral level as well as for firm size classes and regions. Lastly, we illustrate the indicator's potential for giving timely answers to pressing economic issues by analysing the link between digitalisation and firm resilience during the Covid-19 shock.

JEL Classification: C53, C81, O30
Keywords: web-mining, text as data, machine learning, digitalisation

* Digital Economy Department, 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
§ Corresponding author: patrick.breithaupt@zew.de

The paper has been written as part of the project „Taxation in the Era of Digital Transformation" which received funding from the Leibniz Association. The authors would like to thank all participants of the ZEW Digital Economy Seminar, MAGKS Doctoral Colloquium (2020 & 2022), 15th RGS Doctoral Conference, and 31st ITS European Conference for valuable inputs. Special thanks are owed to Irene Bertschek, Thomas Niebel, Christian Rammer, Peter Winker, Reinhold Kesler, Sandra Gottschalk and Manuel Lauer. All remaining errors are ours alone.

1 Introduction

During the last decades, digital technologies have led to fundamental changes in the economy. For example, the spread of information and communication technologies (ICT) positively affected the efficiency of firm-level production processes (Cardona et al. 2013), created completely new markets, such as the app market, and reduced economic costs, such as search and transaction costs (Goldfarb & Tucker 2019). As many digital technologies are constantly evolving and new ones are emerging, it is assumed that the digital transformation will cause many disruptive changes in the future as well. It is therefore of particular importance to continuously analyse the process and the effects of digitalisation. A prerequisite for this analysis is to collect reliable data and create measurement tools that provide useful insights into the ongoing and accelerating digital transformation.[1]

Measuring digitalisation poses some crucial challenges as the term *digitalisation* itself does not relate to a single technology, but to a group of established as well as emerging digital technologies. Hence, either the impact of a single digital technology, e.g., cloud capacity, or composite indicators[2], such as the *Digital Economy and Society Index (DESI)*[3], can be analysed. Composite indicators, however, require human assessment of how to measure and weigh different aspects related to digitalisation, and it is possible that different digitalisation indicators give diverging measurements of the level of digitalisation. Measuring digitalisation at the firm level also poses problems with respect to the collection of data. Traditionally, firm-level information is collected by using questionnaire-based surveys, but these often have major drawbacks. They are cost-intensive, lack timeliness as well as regional coverage, and require firm participation. In contrast, nowadays, a large share of firms has a website that provides a wide range of information. Firm websites often include online shops, information about digital products, job postings or applied technologies as well as links to social media websites.[4] Hence, information on a firm website can relate to the firm's use of digital technologies. As advances in computing power and natural language processing enable the collection and transformation of unstructured data into (semi-)structured data, texts on firm website can be used in real time and at a large scale. However, website data has also some drawbacks, e.g., firms without a website

[1] The OECD provides a roadmap for the measurement of the digital transformation in `https://doi.org/10.1787/9789264311992-en`.

[2] Handbook on Constructing Composite Indicators (OECD): `https://www.oecd.org/sdd/42495745.pdf`

[3] The DESI of the European Commission is described at `https://ec.europa.eu/newsroom/dae/redirection/document/88764`.

[4] The information is based on a (presumably positive) self-representation and may therefore be biased. Our focus is therefore rather on the relative instead of the absolute indicator value.

might have a low degree of digitalisation and the existence of a potential bias towards firms that market their products and services to many clients.[5]

In a dictionary-based approach, the terms describing digitalisation are to some extent generic and not limited to well-defined technologies, i.e., the words *apple* and *bit* can only be properly understood with contextual words. These examples illustrate that measuring digitalisation with a simple keyword search might not be sufficient. For this reason, a methodology is needed that captures a more complex and broad definition of digitalisation. We, therefore, propose to leverage the potential of firm websites in combination with classified news articles and advances in machine learning to create an indicator for firm digitalisation. Previous work illustrates that websites can contain useful information for measuring economic outcomes, e.g., innovation (Axenbeck & Breithaupt 2021, Kinne & Lenz 2021). However, measuring firm digitalisation faces a further challenge because no ground truth exists, i.e., a suitable target variable for a supervised learning approach is not available. To solve this problem, we use a transfer learning approach, in which a new task is solved through the transfer of knowledge from a related task (Torrey & Shavlik 2010). Accordingly, we use texts which are already labeled as either being or not being about digitalisation. For this purpose, we use newspaper articles, as they often appear within predefined sections or are labeled with keywords. The New York Times website[6], for example, has a section entitled *Tech*. To ensure compatibility with our German firm data set, we use four different German newspaper outlets. Two news outlets cover daily news and the other two have a more technical focus. As a result, we cover a broad definition of the term digitalisation. The labeled newspaper data can thus be viewed as an expert system on the subject of digitalisation. Using articles of these newspapers, we fit a multi-language random forest regression model (Hastie et al. 2009) that predicts the likelihood of a newspaper article dealing with the subject of digitalisation. With an accuracy of 97 percent on the test sample, our model can separate well news articles on digitalisation from all other topics. Furthermore, our approach has a decisive advantage over simple keyword-based searches, i.e., we are capable of automatically extracting a time-varying definition of digitalisation. We use our machine learning model that is trained on news articles on the website texts of German firms. This constitutes the transfer learning step of our approach. Thereby, we predict the likelihood that a firm's website text includes the subject of digitalisation. The predictions are used as a continuous indicator for firm digitalisation. Consequently, our digitalisation indicator varies from zero to one.

[5] Issues with web-based data sources are discussed by Rammer & Es-Sadki (2022) in more detail.
[6] `https://www.nytimes.com/`

Our results show that larger firms seem to be more digitalised than smaller firms. Firms in the sectors „computer programming, consultancy, related (service) activities", „telecommunications", and „publishing activities, video & television, music publishing, etc." are highly digitalised in 2018, and firms in the sectors „accommodation and food and beverage service activities", „beverages, food and tobacco", and „construction" have a low digitalisation score in 2018. Moreover, firms in Western Germany appear to be more digitalised than firms in Eastern Germany and firms in bigger cities are on average more digitalised than firms in more rural areas. For additional plausibility checks, we use Eurostat data at the aggregated level (firm size and sectors based on 2-digit NACE codes)[7], digitalisation intensities at the regional level by Prognos AG[8], and Mannheim Innovation Panel (MIP) firm survey data (Rammer et al. 2021). Comparing our web-based digitalisation indicator with already established indicators shows that our digitalisation indicator yields comparable results. We contribute to the literature by generating a large-scale and quality-tested indicator of the digitalisation of firms that covers all firm size classes, regions, and economic sectors in Germany.

Lastly, we illustrate the indicator's potential for giving timely answers to pressing economic issues by analysing the link between digitalisation and firm resilience during the Covid-19 shock. For this purpose, we interpret changes in credit ratings as a proxy for firm resilience to the exogenous Covid-19 shock. Our statistical analyses indicate the following: First, an increase in digitalisation between 2018 and 2020 is associated with an improvement in credit scores between 2019 and 2021. Second, a high pre-crisis level of digitalisation in the year 2018 is linked to an improvement in credit scores between 2019 and 2021. The results are consistent with the related literature.

The remainder of this paper is structured as follows. In Section 2, the related literature is summarised. Section 3 presents our analytical framework based on machine learning. Section 4 illustrates the plausibility of the results. In Section 5, the indicator is applied to a use case. In Section 6, we discuss the results and our contribution. Section 7 offers a conclusion.

[7] isoc_e data: `https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm`
[8] `https://www.prognos.com/de/projekt/digitalisierungskompass-2018`

2 Related Literature

The following section gives an overview of related literature that covers digitalisation indicators (Section 2.1), text mining, and statistical learning methods (Section 2.2).

2.1 Measuring Digitalisation

Digitalisation is broadly defined as the „mass adoption of digital technologies that generate, process and transfer information" (Katz & Koutroumpis 2013, p.314).[9] It is a general concept without a clear-cut definition and is often differentiated between being internal and external. Internal digitalisation refers, for example, to the digitalisation of firm-level processes and products. External digitalisation includes, e.g., the broadband availability in a region. Our web-based firm-level indicator falls, therefore, into the area of internal digitalisation. Internal digitalisation is already measured with different approaches, e.g., based on firm surveys or the analysis of public (web) data.

Governmental and affiliated institutions already provide a large variety of digitalisation indicators. Every year, the Federal Statistical Office of Germany (Destatis) collects data for Germany on the degree of firm digitalisation with the „ICT in enterprises" survey.[10] The data includes, for example, information on *enterprises with a website* and *enterprises using social media*. The digitalisation indicator of the German „Federal Ministry for Economic Affairs and Climate Action" (formerly known as „Federal Ministry of Economic Affairs and Energy") is measured at the firm level and describes the level of digitalisation with respect to sectors, regions, and firm size classes (Büchel et al. 2020). The indicator captures data on firm-level properties such as digitalisation of processes and products, as well as external factors like the availability of technical infrastructure. The Mannheim Innovation Panel includes survey questions about *digital technologies* used in firms (Rammer et al. 2021). The *OECD Going Digital index*[11] consists of the seven sub indicators *access*, *use*, *innovation*, *jobs*, *society*, *trust*, and *market openness*. The *Digital Economy and Society Index* (DESI) tracks digital performance indicators of European countries and makes

[9] In contrast, digitisation describes „the process of converting something to digital form" (Source: `https://www.merriam-webster.com/dictionary/digitization`).

[10] Description of the data: `https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Enterprises/ICT-Enterprises-ICT-Sector/_node.html`.
Access to the data on an aggregated level: `https://www-genesis.destatis.de/genesis/online?language=en&sequenz=statistikTabellen&selectionname=52911#abreadcrumb`

[11] The website `https://goingdigital.oecd.org/en` provides an interactive view on the OECD toolkit and `https://doi.org/10.1787/9789264312012-en` supplies the documentation.

their progress comparable. The dimensions are *human capital, connectivity, integration of digital technology*, and *digital public services*.[12] However, each of these indicators has its drawbacks. Especially, official statistics often cannot react to recent technological developments. For example, until today the European *„ICT usage in enterprises (isoc_e)"* data includes no information about quantum computing. An overview of further country-level indicators is provided by Kotarba (2017). These composite indicators are predominantly used for comparisons between countries.

In addition, there are already some approaches that utilise big data and online data to measure the degree of firm digitalisation and related metrics. For example, Bertenrath et al. (2017) capture on a large scale firm website metrics on used technologies, mobile maturity, amount of traffic, search rankings, social media usage, keywords and quality measures. Ashouri et al. (2021) create a firm digitalisation score for products. They classify the products by analysing the NACE code of the firm, its link to field of study (FOS) codes and the web-scraped website content.

Related academic literature makes use of a large set of proxies to measure firm digitalisation and its economic effects. However, our interest in this study is on the measurement of digitalisation: Hall et al. (2013) analyse the relationship between *research and development* (R&D) expenditures and monetary ICT investments, on the one hand, and innovation, and productivity at the firm level on the other hand. Similarly, Dhyne et al. (2020) derive firm-level ICT capital stocks and find that high ICT capital is linked to an increase of the value added of a firm. Cardona et al. (2013) and Schweikl & Obermaier (2020) provide an overview of empirical literature on ICT and productivity. Furthermore, the capital stock of computers or survey-based spending data are used as indicators in a multitude of studies (Brynjolfsson & Hitt 2003, 1998, 1995, Brynjolfsson et al. 2002). (Greenan et al. 2001, p. 1) make use of five different R&D and IT indicators, e.g., „the ratio of the gross book value of office and computing equipment to the gross book value of total physical assets". Other publications look at regional differentiation: Billon et al. (2010) analyse ICT adoption at the country level. They make use, for example, of the number of broadband subscribers and secure internet servers. Bloom et al. (2012) analyse the effect of IT on productivity for (non-) multinational firms by using UK Census Bureau data on IT expenditures. Niebel (2018) measures the impact of ICT on economic growth in developing, emerging and developed countries. He uses the Conference Board Total Economy Database with information on ICT capital services. Lastly, Forman et al. (2009)

[12] The methodological documentation of the DESI is provided by the European Commission at `https://ec.europa.eu/newsroom/dae/redirection/document/88557`.

analyse the effect of the internet diffusion on regional wage inequality. Similarly, the effects of mobile internet use on productivity (Bertschek & Niebel 2016) and the economic impacts of broadband internet (Bertschek et al. 2015) are investigated.

These approaches have several disadvantages. For example, firm-level investment and capital stock data are often only available for few firms or as an estimate. Furthermore, the proxies are usually based on specific technologies such as broadband or cloud. As a result, our web-based approach can make a valuable contribution to the literature as it covers many firms and technologies, and does not require weights.

2.2 Text Mining & Statistical Learning

Our digitalisation indicator is based on web-based text data and it relies on statistical forecasts. It, thus, relates to the machine learning, text mining, and web data literature.

Previous studies use web data to construct frequent real-time estimates, also known as nowcasting. Some relevant publications are: Ginsberg et al. (2009) utilising Google search queries to detect influenza epidemics in the United States and Choi & Varian (2012) showing that search engine data often correlates with economic activities such as automobile sales and unemployment claims. Several studies show that firm websites and related web data sources are suitable to generate firm-level indicators, e.g., in the innovation literature (Axenbeck & Breithaupt 2021, Kinne & Lenz 2021, Pukelis & Stanciauskas 2019, Gök et al. 2015). Lenz & Winker (2020) show that news articles are suitable to measure the diffusion of new technologies (innovations) by applying a Paragraph Vector Topic Model. Related to our approach, Larsen & Thorsrud (2019) decompose articles of major business newspapers by means of a Latent Dirichlet Allocation topic modelling and show the predictive power for variables such as asset prices.[13] Lastly, newspaper articles are already used to explain a wide range of economic outcomes, e.g., Groseclose & Milyo (2005), Tetlock (2007), Engelberg & Parsons (2011), and might therefore also be suitable for investigating the topic of firm digitalisation.

Unfortunately, we do not have a ground truth for firm digitalisation. Missing training data is a major obstacle in statistical learning. In these situations, *transfer learning* is often used. First, information from a large labeled data set is extracted and a supervised learning model is trained. In a second step, the model is applied and optionally fine-tuned

[13] For a further literature review, see Gentzkow et al. (2019).

to a related problem where a limited amount of labeled data is available. These models are usually called *pre-trained*. The transfer learning method has already been applied to several image recognition tasks, e.g., Lima et al. (2017). As another example, Xie et al. (2016) use a transfer learning approach to measure the poverty rates by taking advantage of the availability of nighttime light intensity rates.

3 Framework

This section describes the measurement of digitalisation using web data. First, we introduce a machine learning-based model (Section 3.1). Second, we describe the newspaper and firm website data as well as the data processing (Section 3.2). Third, we present the results (Section 3.3). Fourth, the performance metrics are presented (Section 3.4).

3.1 Model

We use a transfer learning approach to create a firm digitalisation indicator. The steps of our approach are illustrated and explained in Figure 1. First, the model is trained on news article text data. We use a random forest for the model training (Hastie et al. 2009, Pedregosa et al. 2011). The multi-language machine learning model supports German and English texts, learns on a binary outcome (digital vs. non-digital)[14] and can predict

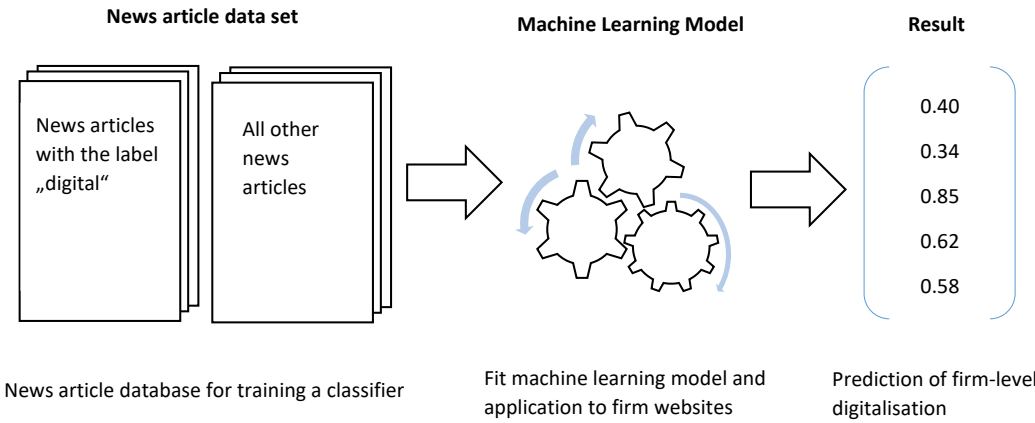


Figure 1: Empirical approach for the creation of the firm digitalisation indicator. News article data with binary labels (left), machine learning model (middle), and continuous firm digitalisation scores based on scraped websites (right). Own illustration.

the label of new or unseen news articles. This model type can also generate continuous regression forecasts between zero and one, indicating the probability of a text being about the topic *digitalisation*.[15] Second, it is applied to firm website text data to receive the continuous firm-level predictions that we interpret as digitalisation scores.

3.2 Data

The following subsection describes the newspaper articles (Section 3.2.1) and firm websites (Section 3.2.2) as well as their transformation into a matrix structure (Section 3.2.3).

3.2.1 News Articles

We scrape newspaper data from four major German online news providers by using the Python packages *Selenium*[16] and *newspaper*[17]. The news providers are not mentioned by their names, instead we give them the numbers (1), (2), (3), and (4) for referencing purposes. A mix of technical and daily news is chosen to cover a wide spectrum of content.

In a first step, we capture the URLs on the news article websites and store them in a unified list. In a following step, each URL is called and the HTML code is saved. Hereby, we scrape and process about 158K news article pages.[18] Errors occur as soon as the website structure can not be parsed. For example, we remove an observation if a relevant data field can not be extracted. The data fields are extracted by analysing the HTML code. Each article has the data fields URL, an abstract if applicable, the text content, the newspaper section (optional), a publication date, and further meta information, e.g., search engine optimisation keywords. The publication date is explicitly considered as the news articles reach many years into the past and might contain outdated information. The data extraction from news articles is exemplified in Figure 2. We restrict the news article data set to recent articles, as the topic *digitalisation* is learned on them and its definition is constantly changing. All articles before 2017 are therefore removed from the data set. News articles after 2019 are removed, because topics on digitalisation are often

[14] Examples: News articles on the German government's digital summit are a prime example of positive data points. On the other hand, news articles on welfare reforms are labeled as non-digital.

[15] Any machine learning model can be used, as long as it supports the output of regression forecasts.

[16] We choose the Selenium package because the websites are often dynamically loaded (`https://pypi.org/project/selenium/`).

[17] `https://pypi.org/project/newspaper/`

[18] We use the abbreviations „K" for thousand and „M" for million, e.g., 80K instead of 80,000.

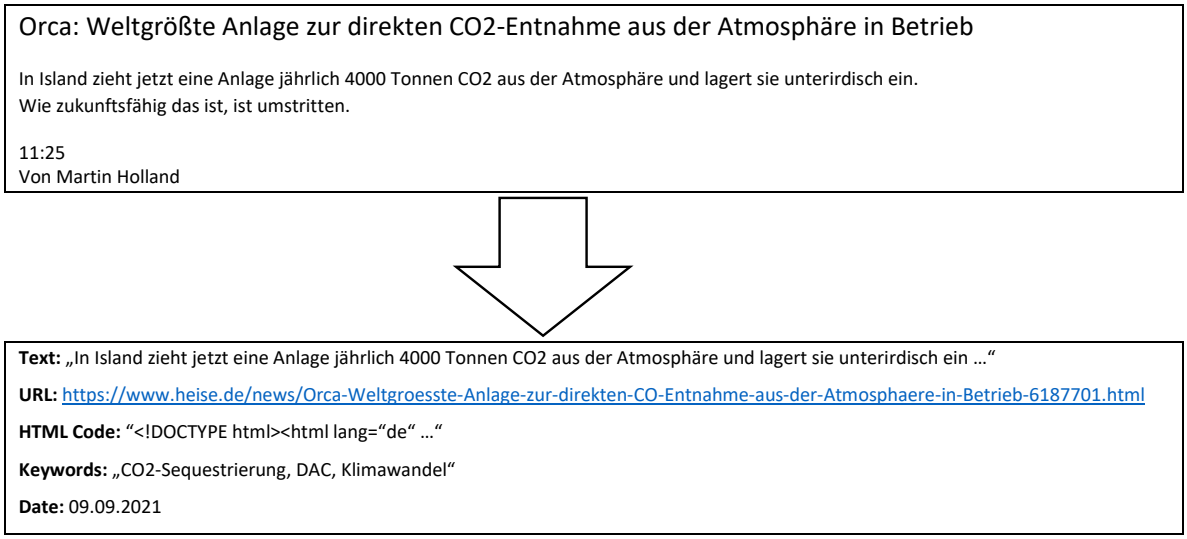


Figure 2: Example of a news article (top) and an excerpt of the extracted data (bottom). The illustrated news article is not in the data set of scraped articles. News article source: `https://www.heise.de/news/Orca-Weltgroesste-Anlage-zur-direkten-CO-Entnahme-aus-der-Atmosphaere-in-Betrieb-6187701.html`. Own illustration.

directly related to Covid-19. This extration process results in a data set consisting of 70,2K news articles. We ensure that no news texts are exact duplicates, i.e., the same news article from different news providers. Only news articles with a minimum length of 1,000 characters (including spaces) are used as input for the machine learning model. This results in a smaller data set consisting of 68,5K news articles without short texts.

A subset consisting of 25K newspaper articles is labeled with a binary variable. The variable is one if the article covers the topic *digitalisation* and zero otherwise. For the daily news sources (1) and (2), the label is set to one if the search engine optimisation (SEO) keywords[19], text or URL contain the word *digital*. For the two more technical news sources (3) and (4), two research assistants independently labeled news articles depending on whether or not they cover the topic *digitalisation*. Observations are only kept if the research assistants unanimously agreed on the labels. Table 1 shows the summary statistics for the four news services. There are positive and negative training examples for each news source and the percentage of news articles on the topic *digitalisation* is significantly higher for the news sources (3) and (4), because they report more frequently on technical topics.

[19] The correlation between the appearance of the word *digital* in the text and the search engine optimisation keyword „digital” is very high. We assume that some of the news providers automatically create the SEO keyword *digital* as soon as this word occurs in the text.

In a next step, we automatically translate the news articles into English by using the Python package *Deep-Translator*[20], because the machine learning model that is trained on the news articles has to classify firm website texts written in both languages. Table

News source	Number of articles	Articles about topic *digitalisation*
(1)	9,048	1,014
(2)	15,325	1,243
(3)	651	451
(4)	297	245
Total	25,321	2,953

Table 1: Label statistics for 25K German news articles. The articles are scraped from four different data sources. The translated English news articles have the same labels.

A.2 (Appendix) shows an overview of the text body size of the news articles. For example, German news articles contain on average approximately 500 words and consist of about 3,600 characters. The metrics for English texts differ for technical reasons.[21]

3.2.2 Website Data

The Mannheim Enterprise Panel (MUP) comprises almost all economically active firms in Germany. For example, by the end of 2013 it consisted of approximately 3.2M German economically active firms (Bersch et al. 2014). The MUP is fed with data from Creditreform[22], one of the largest credit rating agencies in Germany, and is updated every six months. From this data set, samples are taken for various studies, e.g., for the Mannheim Innovation Panel (Rammer et al. 2021). For part of these firms, a web address to the firm website is available.[23] The website URLs are the starting point for the web scraping approach. For our study, we use a snapshot of the MUP data for the years 2018 and 2020. The ARGUS Web Scraping Tool (Kinne & Axenbeck 2020) is then used to capture website content in both years. The program uses the Python package *Scrapy*[24] and has several settings. A firm website consists of at least one web page, i.e., a document that can be viewed in a web browser. In our case, we have limited the scraping to the 50 web pages with the shortest URL on a website. We assume that shorter URLs rather refer to

[20] `https://pypi.org/project/deep-translator/`. We use the „google translator" function.

[21] The software package has a limit for the text length (5,000 characters) and for the number of requests per day. As a result, less than 20% of texts are not completely translated. Optimisation potential: Separate long texts into smaller texts and carry out the translation over a long time period.

[22] The Creditreform Group operates as a credit reporting agency and debt collection service provider. Further information is available at `https://www.creditreform.de/`.

[23] In 2018, around 1.15M economically active firms have an URL (Kinne & Axenbeck 2020).

[24] `https://scrapy.org/`

general content, as the web page is usually fewer clicks away from the main page. Only internal web pages are considered, i.e., we exclude links to other sites such as Google or cooperation partners. Furthermore, we favour, based on a heuristic, German web page texts. This is important for international firms with multiple versions of their website. Therefore, there are mainly German and to a smaller extent English texts in our data set. In the next step, we remove irrelevant web pages with an approach similar to the machine learning-based *gold-bloat* method proposed by Kinne & Lenz (2021). For this, we train a model that predicts the probability of a text being relevant content. We use for the model training web page texts that are divided into the classes *relevant* and *non-relevant*. The model is then applied to the complete web page data set to remove login, document (file), contact and legal content and, thereby, reduces the noise. Lastly, the data is aggregated by combining the web pages of a website into one large text.

The schematic procedure of the website data preparation is shown in Figure 3. The total number of scraped websites is 738K in 2018 and 1.1M in 2020. We were able to successfully scrape, process, and select 663K websites in 2018 and 894K in 2020. We remove observations, e.g., when more than one observational unit is likely to represent the same MUP firm. The number of scraped firm websites between both years differs due to the following reasons: the available URL data set for 2020 is larger than in 2018; websites are accessible for the first time in 2020 or have been switched off since 2018; and the scraping tool ARGUS improved between the data collection time points.

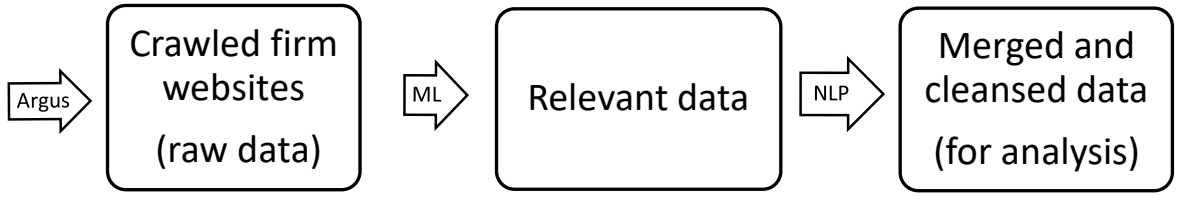


Figure 3: Firm website data crawling and text processing pipeline. Steps: [1] Data crawling using the 'ARGUS web scraper', [2] extraction of relevant data with a *gold-bloat* method, and [3] further cleansing and merging the web page text data. Own illustration.

Furthermore, the scraped websites often contain errors. Reasons for this are, for example, dynamic websites that load parts of the texts only after user interaction.[25] The

[25] The dynamically loaded content could also be scraped, e.g., with Selenium. However, this is not feasible for us, as we cannot implement tailored solutions for the large number of websites covered in this paper. Furthermore, this would also increase the amount of time it takes to scrape a website. There is a potential selection bias as we expect larger and IT-related firms to have dynamic websites.

scraped text content depends on the year in which the data was scraped, as firms might modify their website over time. Repeating the scraping process at a later point in time will most likely yield different results.[26] This gives us the possibility to create a time-varying panel to track the website content. Furthermore, the website can change, because of the characteristics of the visiting users, e.g., via cookies or the browser.

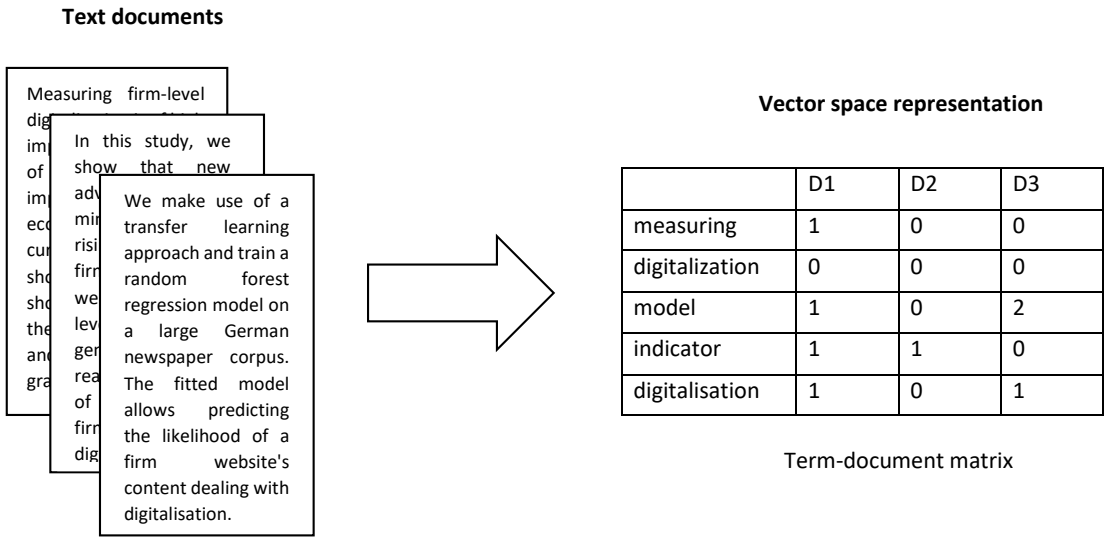
3.2.3 Transforming Texts to Matrices

The news and firm website texts are processed with the pipeline described in Table A.1 (Appendix). The steps are based on established methods from computational linguistics and data science. They consist of [1] data filtering, [2] text tokenisation, [3] stopword filtering, [4] stemming of words, [5] short word removal, [6] unification of words, [7] special character removal, and [8] selection of words, e.g., based on tf-idf.[27] As a result, the text is decomposed into the most important standardised words. The news article and website text data processing needs to be consistent, i.e., the vocabularies must be comparable. For example, the vocabularies might be not comparable if stemming is only performed for one of the two data sets. We expect that this will mitigate the problem that the news article and firm website text types are different to some extent.

Common methods of statistical learning cannot be directly applied to text data, i.e., the texts need to be transformed into a matrix structure (Gentzkow et al. 2019). News article and firm website texts are each transferred to a term-document matrix M. The columns of M represent the documents D, which can either be firm websites or news articles. The rows represent the words (or terms) of the vocabulary, e.g., the most important standardised words in the news article corpus. Figure 4 illustrates that the matrix entry at the position $m_{i,j}$ corresponds to the frequency of the word i in document j.

[26] The used newspaper and firm website data sets are archived for replication purposes.

[27] The term frequency–inverse document frequency (tf-idf) reflects the importance of a word in a text relative to a corpus of texts (Salton & Buckley 1988).



	D1	D2	D3
measuring	1	0	0
digitalization	0	0	0
model	1	0	2
indicator	1	1	0
digitalisation	1	0	1

Figure 4: Transformation of a text data set (left) to a term-document matrix M (right). Matrix columns refer to text documents and rows refer to words. Own illustration.

3.3 Model Training & Results

We use 75 percent of the news article data as training data and 25 percent as test data. Furthermore, a gridsearch optimisation is used for the selection of hyperparameters[28] and a five-fold cross-validation is performed, e.g., to identify overfitting. The term overfitting refers to a model that fits the training data too well and then performs significantly worse on the test data (Hastie et al. 2009).[29] The trained model is then applied to the texts of firm websites. To keep the model simpler, the definition of digitalisation is kept constant on purpose. The underlying machine learning model is, therefore, not changed or updated for the different firm website data sets, e.g., when we use more recent news articles.

The results are roughly 663K continuous predictions of firm digitalisation in 2018. For 2020, we have 894K firm-level predictions in our data set. The intersection between 2018 and 2020 is a set containing about 437K firms. All analyses hereafter refer to this subset unless stated otherwise. The predicted digitalisation score is between zero and one.

[28] We consider the following hyperparameters: number of features, minimum/maximum document frequency of words, stop words, number of estimators, and tree depth.

[29] We use the scikit-learn Python package (`https://scikit-learn.org/stable/`) and fix the random seed to make the results reproducible.

3.4 Performance on Newspapers

Powers (2011) and Fawcett (2004) provide an overview of the evaluation metrics' precision, recall, f1-measure, receiver operating characteristic (ROC) curve, and the area under curve (AUC) value (see Equation 1).[30] The *receiver operating characteristic* (ROC) curve plots the TPR against the FPR for different classification thresholds. The threshold is a value that transforms the continuous predictions of a model into class assignments. For example, assign class A if the continuous prediction is smaller than 0.5; otherwise assign class B. Lastly, the *area under curve* (AUC) value is defined as the area under the ROC curve.

$$\begin{aligned} &\text{Precision} = \frac{TP}{TP+FP}; && \text{Recall} = \tfrac{TP}{TP+FN}; \\ &\text{F-1 measure} = 2 * \frac{Precision * Recall}{Precision + Recall}; && \text{Accuracy} = \tfrac{TP+TN}{TP+TN+FP+FN}; \\ &\text{TPR} = \frac{TP}{TP+FN}; && \text{FPR} = \tfrac{FP}{FP+TN} \end{aligned} \tag{1}$$

The evaluation metrics confirm that the model learns from news articles and creates highly accurate predictions on unseen observations, see Table 2. Good *out-of-sample performance* is a requirement for a statistical model to be used later in a transfer learning approach. The performance of the model (precision, accuracy) is good for the two classes *digitalisation* and *non-digitalisation*, but the recall for the class *digitalisation* is slightly lower. The ROC curve is shown in Figure A.1 (Appendix). The AUC value, which is based on unseen data, is 98 percent. For skewed class distributions, one can use a baseline model that always predicts the majority class. The percentage of news articles about the topic digitalisation is around 12 percent. The accuracy of the baseline model, which always assigns the majority class label, is, therefore, 88 percent. Thus, the machine learning model performs better than the simple baseline model and is able to learn from the data. In summary, our model provides good predictions on unseen data and is capable to generalise. The *Mean Decrease in Impurity* (MDI) feature importance (Breiman et al. 1984) is calculated using random forests. The most important words and their feature importance values are presented in Table A.3 (Appendix). The list of words shows a clear connection to the topic of digitalisation. However, this list may also contain words that are negatively correlated with digitalisation, e.g., *analog*. Furthermore, some words

[30] The predictions fall into the classes true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Furthermore, there are the key figures true positive rate (TPR) and false positive rate (FPR), and the total number of predictions is denoted by N. The formulas for precision and recall are shown, as examples, for the positive class.

Metrics on test sample					
classes	precision	recall	f-1 measure	support	accuracy
non-digitalisation	0.98	0.99	0.98	11,164	
digitalisation	0.93	0.83	0.87	1,498	
				12,662	0.97

Table 2: Evaluation metrics on the news data test sample. *Precision, recall, f1-measure,* and *support* are reported separately for the two classes, and *accuracy* is reported for both together. The class 1 is equivalent to the label *digitalisation* and 0 to *non-digitalisation*.

are only relevant in combination with other words. For example, the word *apple* might refer in English texts to the firm or the fruit. Its meaning becomes clear if the context is considered, i.e., if we take multiple decision layers of the random forest trees into account. To sum up, the random forest model is able to sort unseen news article texts into the correct class and guarantees, to some extent, interpretability of the decision rules.

4 Plausibility of Digitalisation Scores

In the following, we check the plausibility of the digitalisation indicator based on firm websites by providing evidence that it measures firm digitalisation. Unfortunately, the NACE codes[31] (<1%), employee counts (29%), and information on the German federal state (<1%) are not available for all MUP firms. The missing rates are listed in brackets for the year 2018. Therefore, we use subsets of the firms for which the relevant data is available to validate the predictions by the machine learning model. Firstly, we examine whether the web-based digitalisation indicator yields plausible results with respect to, e.g., sectoral and regional differences. Secondly, the results are compared to independent data sources such as the Mannheim Innovation Panel (Rammer et al. 2021) and the aggregated Eurostat *„ICT usage in enterprises (isoc_e)"* data.[32] We show that our model produces meaningful results that are consistent with existing measures.

[31] The NACE codes are the „Statistical Classification of Economic Activities in the European Community". For a definition, see `https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF`.

[32] The *„ICT usage in enterprises (isoc_e)"* data is based on the annual surveys on *„ICT usage and e-commerce in enterprises"* by the National Statistical Institutes or Ministries in Europe. For more information, see `https://ec.europa.eu/eurostat/cache/metadata/en/isoc_e_esms.htm`.

Firstly, Figure 5 shows the continuous firm digitalisation scores. The calculations are carried out for about 437K firms scraped in 2018 and 2020. The distribution shows that many firms have a low digitalisation score and the existence of a long-tail on the right side. Furthermore, there is a small mass point in the range of about 0.7. It is not

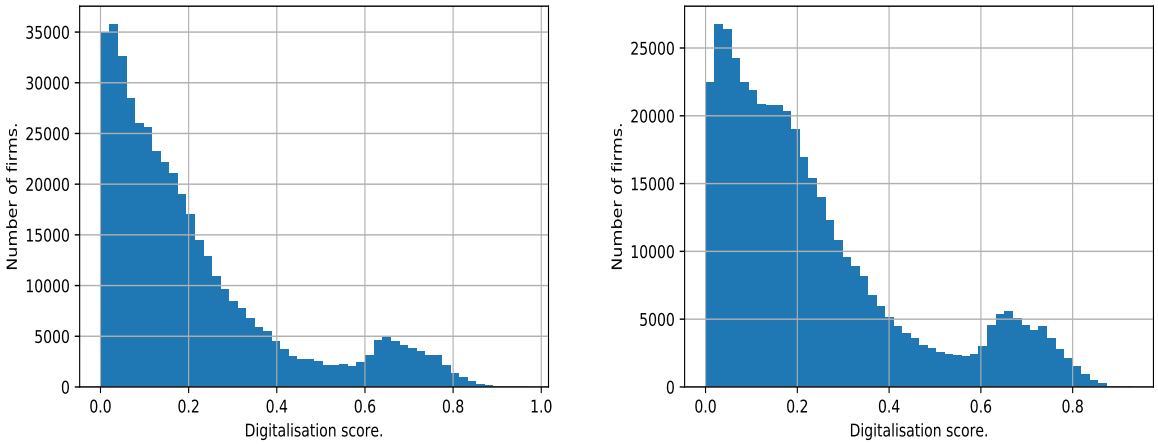


Figure 5: Number of firms in binned digitalisation score intervals. The predictions are based on the websites in 2018 (left figure) and 2020 (right figure). Own illustrations.

possible to identify a simple explanation for that as the used model is non-linear and has several complex decision levels. Therefore, the mass point cannot be simply associated with a single word of the vocabulary.[33] There are no strong outliers, i.e., large jumps, in the distribution. The average firm digitalisation is 0.21 in 2018 and 0.24 in 2020 (see Table 3), i.e., firms became more digitalised. However, there are also firms for which we observe a declining digitalisation score. The websites of these firms have often changed fundamentally, which in turn changed the digitalisation score. In Figure A.2 (Appendix), we can see that the digitalisation score has increased for the majority of the firms.

Year	∩	#obs.	Mean	Median	Std.	Min.	Max.
2018	No	663K	0.2050	0.1412	0.1979	0.0011	0.9680
2020	No	894K	0.2376	0.1783	0.2049	0.0008	0.9740
2018	Yes	437K	0.2112	0.1471	0.2007	0.0011	0.9680
2020	Yes	437K	0.2398	0.1792	0.2066	0.0008	0.9296

Table 3: Summary statistics for the digitalisation scores in 2018 and 2020, respectively. Each prediction is based on the textual content of a firm website. The first two rows show all firms per year and the last two rows present firms available in both years.

[33] One potential explanation for this finding might be a bimodal distribution, i.e., that there are two groups (non-digital and digital firms) in our data set. Furthermore, the predictions for the news article test data show a comparable distribution. The majority of the articles have a low predicted value and the distribution shows a small mass point at around 0.7.

Secondly, the average digitalisation scores per industry are shown in Figure 6. The data is aggregated to 2-digit NACE codes. For example, the ICT industry (NACE codes *26, 61–63*) is very digitalised and, for instance, the „food and beverage service activities" industry (NACE code *56*) has a low average degree of digitalisation. Furthermore, the number of observations per industry is crucial, as this could lead to outlier-sensitive results. In the context of this study, the number of observations in each industry is at

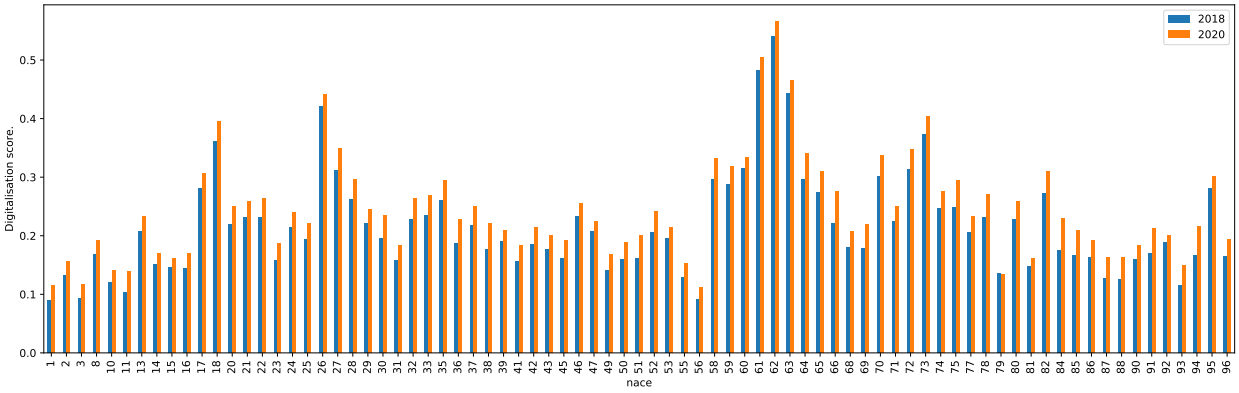


Figure 6: Average digitalisation scores per industry in 2018 (blue color) and 2020 (red color). The industries are defined on the 2-digit NACE codes. Groups with less than fifty observations are not shown. Own illustration.

least 50. Industries with a coverage below 50 firms are removed from the visualisation to ensure anonymity of the firms and to reduce instabilities of the predictions due to low observation numbers.[34] For the next analysis, we use the sector definition from the Eurostat ICT survey that is also based on aggregated 2-digit NACE codes. We use the data on „buy cloud computing services used over the internet" (2018), „enterprises who have ERP software package to share information between different functional areas" (2017), „enterprises' total turnover from e-commerce" (2018), „enterprises analysing big data from any data source" (2018), and „use two or more social media (as of 2014)" (2017) for the following analyses. The data is available aggregated by sector or firm-size.[35] These variables are selected because they are available at a fine granular level and are part of the „integration of digital technology" metrics of the DESI.[36] We calculate the unweighted average of the five variables and use it as a composite indicator to evaluate our approach.

[34] Note, only a few 2-digit industries are dropped due to low numbers, other missing 2-digits are just not defined in the NACE classification. The respective results are therefore not shown, e.g., *04, 05, 06,* and *07*.

[35] For sectors: without data from the financial sector; restricted to firms with at least 10 employees. For firm sizes: restricted to firms with at least 10 employees.

[36] Unfortunately, the DESI is only available at a highly aggregated level and thus not directly comparable with our results.

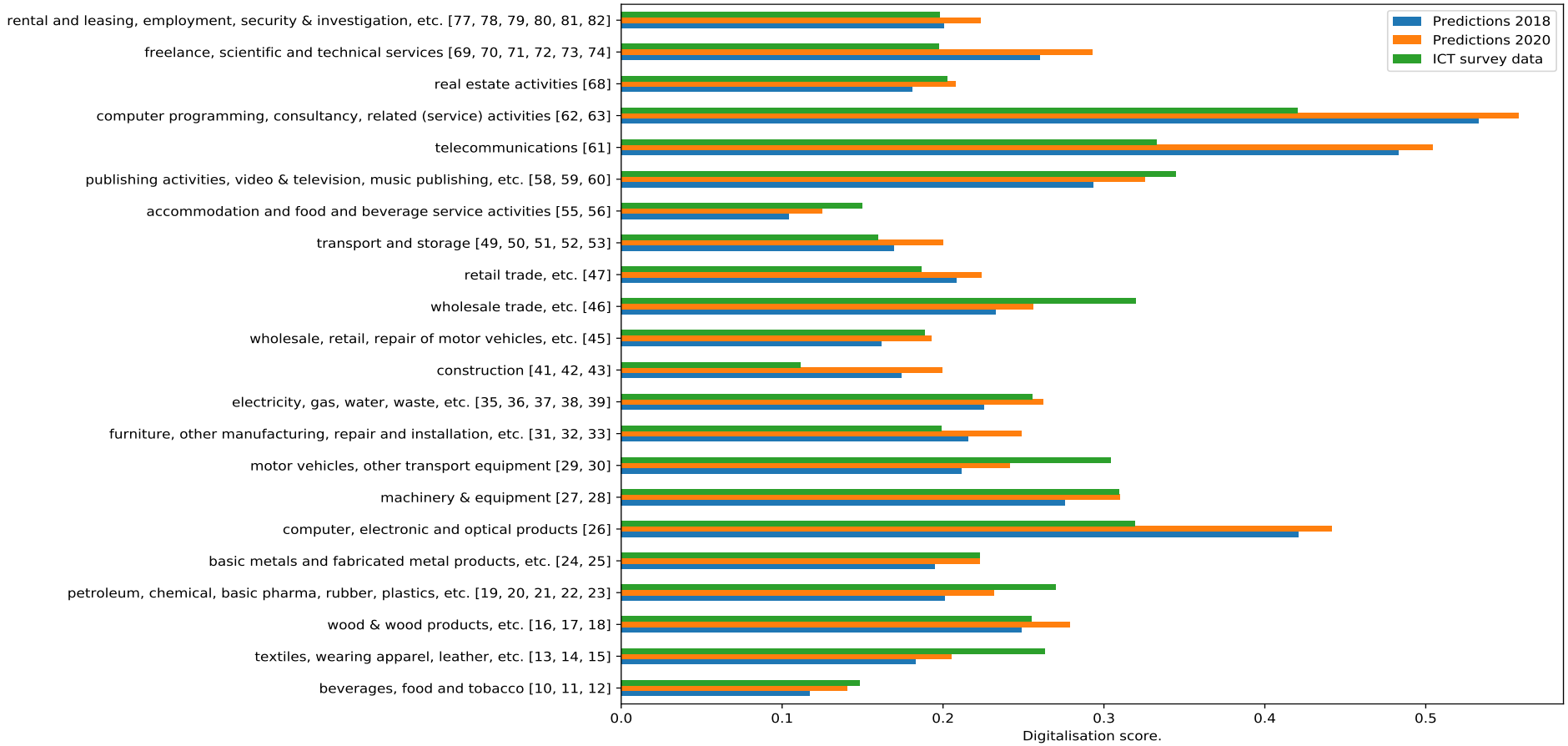


Figure 7: Average digitalisation per sector in 2018 and 2020, respectively, based on MUP, Eurostat ICT survey (2017/2018), and web data. Groups with at least fifty observations are shown. The sectors are defined on the 2-digit NACE codes (see brackets). Own illustration.

Figure 7 shows the average web-based digitalisation score per sector and the Eurostat composite indicator. Both data sets seem to be quite similar on the sectoral level and our indicator does not produce counter-intuitive results.[37] For example, the sectors *computer programming, consultancy, related (service) activities (62-63)*, and *telecommunication (61)* have a high score, but *beverages, food and tobacco (10-12)* has a low score.

Thirdly, Figure 8 shows the data for different firm size classes, i.e., the larger the firms, the higher the firm digitalisation. The difference becomes evident across indicators, e.g., when comparing large and small firms.[38] Lastly, our results also suggest validity with respect to regional differences. Figure 9 shows the distribution of firm digitalisation for

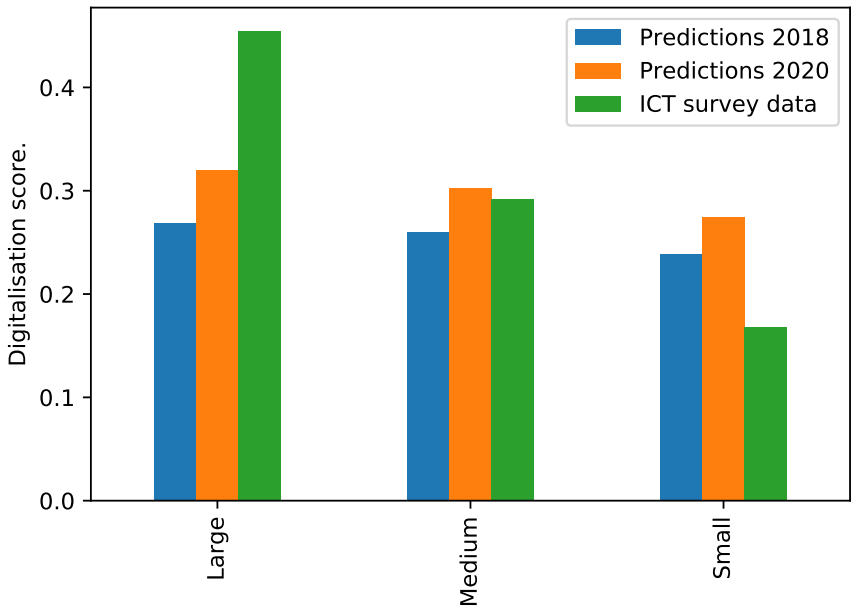


Figure 8: Average digitalisation per firm size in 2018 and 2020, respectively, that is based on MUP, Eurostat ICT survey, and web-data. The data is grouped by the number of employees (*10-49 = Small, 50-249 = Medium, and 250+ = Large*). Own illustration.

German districts. The illustrations on the left and in the middle show the web indicator for 2018 and 2020, respectively. The illustration on the right is based on the *Digitalisation Compass 2018* by Prognos AG[39] and uses a different scale. A dark red color illustrates a high average level of digitalisation within a German district. The pattern in all illustrations is similar, e.g., the eastern part of Germany is less digitalised and big cities such as Berlin and Munich are more digitalised than rural areas.

[37] For the sector data, we calculate the Pearson correlation coefficient, i.e., corr(web-based indicator 2018, Eurostat ICT survey) = 0.79 and corr(web-based indicator 2020, Eurostat ICT survey) = 0.80. The district data is not available for the „Digitalisierungskompass 2018" and the number of firm size classes is too small to reliably calculate a correlation coefficient.

[38] Our results match with findings from other studies, e.g., Büchel et al. (2020).

[39] `https://www.prognos.com/de/projekt/digitalisierungskompass-2018`

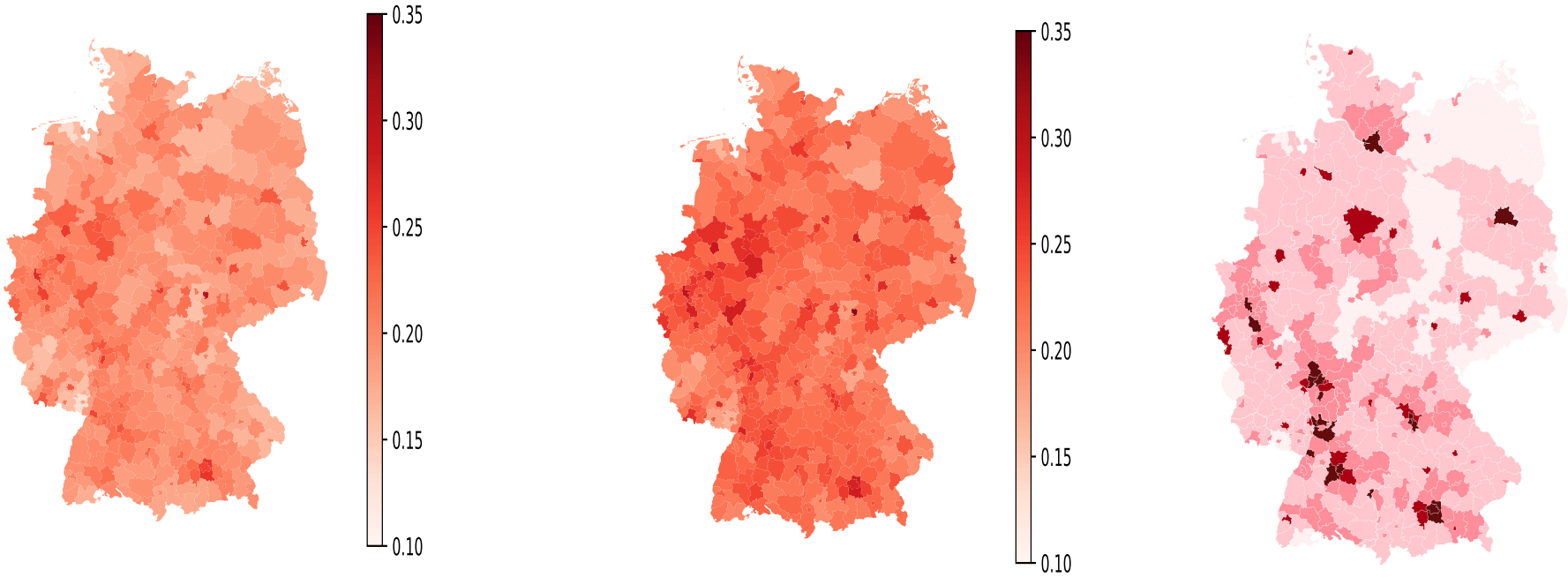


Figure 9: Digitalisation at the district level. Average digitalisation scores per German district in 2018 (left) and 2020 (middle) that are based on MUP and web-data. Right: Digitalisation Compass 2018 by the Prognos AG & index Gruppe. The left and middle figures are our own illustrations. The source of the right figure is `https://www.handelsblatt.com/politik/deutschland/digitalisierungskompass/` and its description is available at `https://www.prognos.com/de/projekt/digitalisierungskompass-2018`.

Fourthly, we use the MIP 2020 survey to create a digitalisation indicator at the firm level for comparison. Table A.4 (Appendix) provides the list of survey questions with a focus on digitalisation. The possible answers are *none, low, medium,* and *high*. The answer *none* is re-coded to 0, *low* to 1, *medium* to 2, and *high* to 3. Our MIP indicator is based on the unweighted average of the answers and consists of 894 observations. Figure 10 illustrates the MIP indicator and its link to the web-based digitalisation indicator. There is a positive link between the indicators, i.e., the web-based digitalisation score increases with the MIP score as indicated with the dotted regression line. However, the relationship does not become clear on the scatter plot as the points are too spread out.

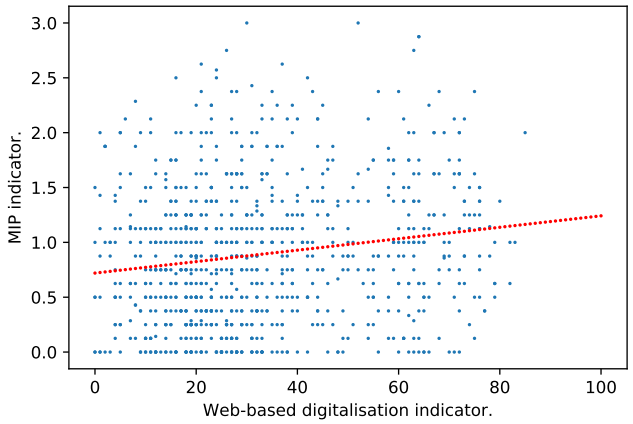


Figure 10: Comparison of the web-based digitalisation indicator (2020) and the MIP digitalisation indicator (2020). The blue dots represent the individual observations and the dotted red line is an estimate based on a linear regression. Own illustration.

In summary, our indicator seems to be plausible with respect to firm size classes as well as at the geographical and sectoral level. We believe that the reason for this lies in our data sets as well as in the machine learning model and its capability to capture the topic digitalisation. Lastly, evidence for model validity is given at the firm level.

5 Use Case: Firm Resilience

Finally, we check whether the web-based digitalisation indicator is related to other firm characteristics in the way one would expect. For this purpose, we use the established relationship between firm digitalisation and resilience in the literature. Firm resilience is defined as „the ability of a firm to persist in the face of substantial changes in the business and economic environment and/or the ability to withstand disruptions and catastrophic events" (Acquaah et al. 2011, p. 5528) and is thus an attribute of a firm. Our work is

consistent with findings that more digitalised firms have operational advantages during a health crisis, e.g., business activity has to be restricted less if the employees work from home.[40]

Economic literature already gives substantial evidence on the positive link between firm resilience and digitalisation. For example, Conz & Magnani (2019) and Saad et al. (2021) conduct a systematic literature review on firm resilience. Bertschek et al. (2019) show that ICT-intensive firms were hit less hard (productivity) during the crisis in 2008 & 2009 by analysing data from twelve countries and seven industries. Elgazzar et al. (2022) investigate the link between digital transformation and firm resilience during the Covid-19 crisis. Bianchini & Kwon (2021) investigate the role of digitalisation programmes by the government in strengthening SMEs' resilience in Korea during the Covid-19 shock. Fischer et al. (2022) explore the resilience of the public service sector during the Covid-19 pandemic. Abidi et al. (2022) investigate whether digitally-enabled firms in the Middle East and Central Asia reduced economic losses better during the Covid-19 crisis.

We use the MUP data (Bersch et al. 2014) and our web-based digitalisation indicator to examine the relationship between digitalisation and resilience of firms during the exogenous Covid-19 shock. For this purpose, we assess the solvency of firms by their credit ratings.[41] We assume that resilient firms, for example, do not or to a smaller extent experience a decline in solvency during the Covid-19 crisis. We, therefore, measure resilience as the change in firms' credit ratings based on the rating before and after the shock. The MUP data is a panel and consists of semi-annual data points.[42] For data preparation, we delete observations as soon as a variable is missing, so that all regression analyses are conducted on the same data set. In the original definition, the credit rating is in the range of 100 and 600, where 100 is the best rating. Our credit rating data preparation is similar to the data processing in Dörr et al. (2021) and Dörr et al. (2022).[43] For a simpler interpretation, we modify the credit rating using the function $(6 - (\textit{credit rating}/100))$. Thus, the credit rating range is between 1 and 5, and a high number indicates a good rating. We delete observations marked as duplicates or faulty, firms founded after Jan-

[40] Alternative: Official firm exit data cannot be analysed during this period, because there was no obligation to file for insolvency in Germany (`https://www.gesetze-im-internet.de/covinsag/`).

[41] Some factors of the credit rating: credit assessments, annual financial statement data, industry risk, turnover, legal form, firm age, regional risk, order situation, capital, management experience, number of employees, turnover / employees, capital / turnover. See also: `https://www.creditreform.de/aktuelles-wissen/praxisratgeber/wie-sie-ihren-bonitaetsindex-verbessern`.

[42] We use the waves delivered in the middle of the years for our analysis.

[43] Firms with a credit rating of less than 100 are deleted as this indicates that too little information is available about a firm, e.g., a start-up. In addition, the credit rating of firms with values above 500 are truncated to 500, e.g., insolvent firms.

uary 2018, and with an exit year before 2018. Table A.5 (Appendix) shows the summary statistics of the estimation sample. The average credit rating worsened between 2019 and 2021 by 0.09 points. The average digitalisation in our sample increased by 0.03 between 2018 and 2020 and firms have gained on average 0.71 additional employees. We control for firm properties at the regional level (sixteen German states) and the sector level (21 groups), as well as for the founding period (4 groups) and the legal form (14 groups). Their statistics are shown in Table A.6 (Appendix).

$$\Delta rating_{t+1,i} = \beta_1 \Delta digitalisation_{t,i} + \beta_2 digitalisation_{t,i} + ... + u_{t,i} \tag{2}$$

Equation 2 shows the model specification for our firm resilience analysis. In addition, we control for the sector, location, legal form, employee count, and founding period of the firm. The variable $\Delta rating_{t+1,i}$ refers to the change in the credit rating for firm i between 2019 and 2021. The main regressors $\Delta digitalisation_{t,i}$ and $digitalisation_{t,i}$ are the change of firm digitalisation between 2018 and 2020 and the degree of firm digitalisation in 2018. $\Delta employees_{t,i}$ is the change of employee counts between 2018 and 2020. Table 4 shows the

	(1)		(2)		(3)	
	Δ rating 2021 - 2019		Δ rating 2021 - 2019		Δ rating 2021 - 2019	
Δ digitalisation 2020 - 2018	0.0202***	(0.000)			0.0296***	(0.000)
digitalisation 2018			0.0794***	(0.000)	0.0361***	(0.000)
Δ employees 2020 - 2018					0.00453	(0.551)
employees 2018					-0.000870	(0.662)
Constant	-0.0906***	(0.000)	-0.108***	(0.000)	-0.358***	(0.000)
Founding Period Dummies	No		No		Yes	
Legal Form Dummies	No		No		Yes	
Sector Dummies	No		No		Yes	
Location Dummies	No		No		Yes	
Observations	176,902		176,902		176,902	
R^2	0.000		0.002		0.035	

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Lagged cross-sectional regression results for the firm resilience use case. The columns 1 and 2 show the baseline specifications without control variables. The column 3 includes additional control variables. The number of employees is stated in thousands.

empirical results. Columns (1) - (2) show a simple time-lagged cross sectional estimation without control variables, i.e., we expect a temporal delay between an increase in firm digitalisation and the corresponding change in the firm resilience. For example, the use of a new technology might require a learning phase. The results show that digitalised firms (pre-crisis) or firms that increased their digitalisation between 2018 and 2020 are more resilient with respect to the Covid-19 crisis, i.e., the coefficients are positive and significant. Column (3) shows the estimation including the control variables. A highly

significant positive relationship is found. If a firm increases its level of digitalisation from zero to one in the time period between 2018 and 2020, then, on average, the firm has a credit rating increase of 0.0296 for the years 2019 to 2021. Similarly, the pre-crisis level of digitalisation is positively and significantly linked to firm resilience. However, the coefficients are small with respect to the credit rating range and the R^2 value is only 0.035.[44] The results of the use case illustrate the indicator's potential for giving timely and plausible answers to pressing economic issues.

6 Discussion

The proposed approach is, to the best of our knowledge, the first purely web and text-based method to measure the digitalisation of German firms on a regular basis and on a large scale. The method is not based on survey data that requires an extensive and time-consuming data preparation. However, our method presupposes that the performance of the statistical model and the plausibility checks are convincing to the reader.

The definition of digitalisation changes over time. For example, the word *fax* was indicative of firm digitalisation a couple of decades ago. Today, it means quite the opposite because there are more advanced alternatives like *emails* to send documents. The vague definition of the term digitalisation is also a big challenge and illustrates the issue that we cannot evaluate against a ground truth. Therefore, an important part of this work is to provide convincing arguments that the derived web-based indicator provides plausible results. Moreover, a definition would necessarily also change over time. The problem can be solved with an extension of the model, e.g., by ignoring older news articles or reducing their weight over time. The use of further news article outlets might provide an even broader and perhaps better definition of digitalisation. So far, news articles from four providers are used, but this does not necessarily reflect all aspects of digital technologies. Furthermore, the issue may be raised whether about 3K news articles about digitalisation are sufficient to recognise all available technologies. If not, this may create a bias in the data and, as a result, also in the model. Subsequently, the question arises whether news articles are a good choice to learn the ground truth. In addition, not all available news articles were fully translated from German to English so far. Information like the firms' sector or size could also be used as input to the statistical model. For example, an IT

[44] Adding the firm-level credit rating for 2018 to the model, preserves the sign and significance levels of digitalisation 2018 and Δ digitalisation 2020 - 2018 ($R^2 = 0.100$). However, the sign of the constant becomes positive and the credit rating in 2018 has a negative and significant coefficient.

firm with little information on its website could be classified as more digitalised than based on the firm website content alone. However, this requires the sector information for every firm in Germany. Other machine learning methods than the selected random forest model might also prove to be more suitable for the transfer learning task, e.g., deep neural networks. Furthermore, it is necessary to highlight the challenges associated with the transfer of a model. The news article and firm website texts can be very different, e.g., in length and words used, so that the vocabularies are not comparable. Lastly, natural language processing methods have not yet been fully exhausted, e.g., word embeddings.

Websites only represent the public image of the firm and the digitalisation score derived from them could therefore be biased. However, our indicator is, on an aggregated level, in line with other digitalisation indicators. Some firms might talk too little or too much about digitalisation on their websites, which can result in noise that might be systematic. There may also be differences with respect to firm sizes and sectors. For instance, large firms in the chemical industry might use their websites to report about digitalisation differently than a software development firm. Some firms might even list buzzwords on websites to increase the attractiveness and competitiveness of the firm or for search engine optimisation. Vos (2009) discusses a similar deception in the context of climate protection, i.e., greenwashing. We cannot detect the accuracy of the websites' content and some dynamically loaded content may be missing. The website crawling can be further improved as dynamic websites are only crawled to a limited extent, but a tailor-made solution for all firm websites is not possible.[45] The content of firm websites could vary in timeliness. Some firms update their websites much less frequently than others. We cannot determine the age of firm website content, but we can calculate an approximation if the website data will be regularly collected in the future. The data processing is optimised for texts in German or English, but we cannot give a reliable estimate of how many firm website texts are neither German nor English. However, Axenbeck & Breithaupt (2021) showed in a comparable study that only about two percent of web pages fall in this category.

Lastly, we illustrate the indicator's potentials for timely answers to empirical questions of high economic and policy relevance. We show in a Covid-19 related use case that our indicator provides results that are consistent with findings of related studies: digitalised firms are more resilient to exogenous shocks. The research question could, depending on the data source, only be answered under restrictions such as time delays.

[45] Future studies could use other software packages such as *Selenium* instead of *Scrapy*.

7 Conclusion

In this paper, we introduce a web-based digitalisation indicator to address the problems of traditional indicators. For this purpose, we use 25K unique news articles to train a random forest model that is able to predict whether a text is about digitalisation. Using a transfer learning approach, we apply the fitted model to German firm websites to create a web-based digitalisation indicator. We predict digitalisation scores for 663K German firms in 2018 and 894K in 2020. The scores are available for 437K firms in both 2018 and 2020. Comparisons with established indicators show that our approach provides plausible results at the firm, regional, and sectoral level as well as for different firm size classes.

Our web-based indicator is a cost-effective way to measure firm digitalisation, it can be updated quickly and covers many firms. Thus, it does not have the usual disadvantages of questionnaire-based measures. In addition, the indicator is quality-checked and covers all firm size classes, regions, and economic sectors in Germany. The firm-level indicator of digitalisation constitutes our contribution to the literature.

Lastly, we illustrate the indicator's potential for giving timely answers to pressing economic issues by analysing the link between digitalisation and firm resilience during the Covid-19 shock. We find results that are consistent with the related literature, demonstrating the successful application in a use case. However, the analysis is only one example of a wide range of applications. Besides research, it is also applicable for economic policy advice and consulting, e.g., analyse regional differences in digitalisation.

References

1. Abidi, N., El Herradi, M. & Sakha, S. (2022), *Digitalization and Resilience: Firm-level Evidence During the COVID-19 Pandemic [Preprint]*, International Monetary Fund. Working Paper No. 2022/034. [Online; accessed 01.12.2022].
   URL: *https://www.imf.org/-/media/Files/Publications/WP/2022/English/wpiea2022034-print-pdf.ashx*

2. Acquaah, M., Amoako-Gyampah, K. & Jayaram, J. (2011), 'Resilience in family and nonfamily firms: an examination of the relationships between manufacturing strategy, competitive strategy and firm performance', *International Journal of Production Research* **49**(18), 5527–5544.

3. Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C. & Cunningham, S. (2021), 'Data in Brief: Indicators on Firm Level Innovation Activities from Web Scraped Data [Preprint]'. [Online; accessed 29.11.2022].
   URL: *https://doi.org/10.34894/W3W2JQ*

4. Axenbeck, J. & Breithaupt, P. (2021), 'Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity?', *PLOS ONE* **16**(4), e0249583.

5. Bersch, J., Gottschalk, S., Müller, B. & Niefert, M. (2014), 'The Mannheim Enterprise Panel (MUP) and Firm Statistics for Germany', *ZEW – Leibniz - Centre for European Economic Research Discussion Paper* (14-104). [Online; accessed 14.11.2022].
   URL: *https://ftp.zew.de/pub/zew-docs/dp/dp14104.pdf*

6. Bertenrath, R., Fritsch, M., Lichtblau, Karl & Schleiermacher, T. (2017), 'Digitale Wirtschaft in Nordrhein-Westfalen', *Studie im Auftrag der Initiative Digitale Wirtschaft NRW des Ministeriums für Wirtschaft, Energie, Industrie, Mittelstand und Handwerk des Landes Nordrhein-Westfalen, Köln* . [Online; accessed 29.11.2022].
   URL: *https://www.iwkoeln.de/fileadmin/publikationen/2017/334156/IW-Gutachten_Digitale_Wirtschaft_NRW_Endbericht.pdf*

7. Bertschek, I., Briglauer, W., Hüschelrath, K., Kauf, B. & Niebel, T. (2015), 'The Economic Impacts of Broadband Internet: A Survey', *Review of Network Economics* **14**(4), 201–227.

8. Bertschek, I. & Niebel, T. (2016), 'Mobile and more productive? Firm-level evidence on the productivity effects of mobile internet use', *Telecommunications Policy* **40**(9), 888–898.

9. Bertschek, I., Polder, M. & Schulte, P. (2019), 'ICT and resilience in times of crisis: evidence from cross-country micro moments data', *Economics of Innovation and New Technology* **28**(8), 759–774.

10. Bianchini, M. & Kwon, I. (2021), 'Enhancing SMEs' resilience through digitalisation: The case of Korea', *OECD SME and Entrepreneurship Papers* **27**.
URL: *https://www.oecd-ilibrary.org/economics/enhancing-smes-resilience-through-digitalisation_23bd7a26-en*

11. Billon, M., Lera-Lopez, F. & Marco, R. (2010), 'Differences in digitalization levels: a multivariate analysis studying the global digital divide', *Review of World Economics* **146**(1), 39–73.

12. Bloom, N., Sadun, R. & Van Reenen, J. (2012), 'Americans do IT better: US multinationals and the productivity miracle', *American Economic Review* **102**(1), 167–201.

13. Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), 'Classification and Regression Trees (1st Edition)', *Routledge, New York* .
URL: *https://doi.org/10.1201/9781315139470*

14. Brynjolfsson, E. & Hitt, L. (1995), 'Information technology as a factor of production: The role of differences among firms', *Economics of Innovation and New technology* **3**(3-4), 183–200.

15. Brynjolfsson, E. & Hitt, L. M. (1998), 'Beyond the Productivity Paradox', *Communications of the Association for Computing Machinery (ACM)* **41**(8), 49–55.

16. Brynjolfsson, E. & Hitt, L. M. (2003), 'Computing productivity: Firm-level evidence', *The Review of Economics and Statistics* **85**(4), 793–808.

17. Brynjolfsson, E., Hitt, L. M. & Yang, S. (2002), 'Intangible assets: Computers and organizational capital', *Brookings papers on economic activity* **2002**(1), 137–181.

18. Büchel, J., Demary, V., Goecke, H., Rusche, C., Burstedde, A., Engels, B., Koppel, O., Mertens, A., Scheufen, M., Wendt, J., Ewald, J., Hünnemeyer, V., Kempermann, H., Lichtblau, K., Schmitz, E., Bertschek, I., Niebel, T., Rammer, C.,

Schuck, B., Birtel, F., Harland, T., Hicking, J. & Wenger, L. (2020), 'Digitalisierungsindex 2020 - Langfassung Ergebnispapier'. [Online; accessed 14.11.2022].
URL: *https://www.de.digital/DIGITAL/Redaktion/DE/Digitalisierungsindex/Publikationen/publikation-download-Langfassung-digitalisierungsindex-2020.pdf*

19. Cardona, M., Kretschmer, T. & Strobel, T. (2013), 'ICT and productivity: conclusions from the empirical literature', *Information Economics and Policy* **25**(3), 109–125.

20. Choi, H. & Varian, H. (2012), 'Predicting the present with Google Trends', *Economic Record* **88**(s1), 2–9.

21. Conz, E. & Magnani, G. (2019), 'A Dynamic Perspective on the Resilience of Firms: A Systematic Literature Review and a Framework for Future Research', *European Management Journal* **38**(3), 400–412.

22. Dhyne, E., Konings, J., Van den Bosch, J. & Vanormelingen, S. (2020), 'The Return on Information Technology: Who Benefits Most?', *Information Systems Research* **32**(1), 194–211.

23. Dörr, J. O., Licht, G. & Murmann, S. (2021), 'Small firms and the COVID-19 insolvency gap', *Small Business Economics* **58**(2), 887–917.

24. Dörr, J. O., Kinne, J., Lenz, D., Licht, G. & Winker, P. (2022), 'An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers', *PLOS ONE* **17**(2), e0263898.

25. Elgazzar, Y., El-Shahawy, R. & Senousy, Y. (2022), The Role of Digital Transformation in Enhancing Business Resilience with Pandemic of COVID-19, *in* 'Digital Transformation Technology', Springer, Singapore, pp. 323–333.

26. Engelberg, J. E. & Parsons, C. A. (2011), 'The causal impact of media in financial markets', *The Journal of Finance* **66**(1), 67–97.

27. Fawcett, T. (2004), 'ROC graphs: Notes and practical considerations for researchers', *Pattern Recognition Letters* **31**(8), 1–38.

28. Fischer, C., Siegel, J., Proeller, I. & Drathschmidt, N. (2022), 'Resilience through digitalisation: How individual and organisational resources affect public employees working from home during the COVID-19 pandemic', *Public Management Review* pp. 1–28.
URL: *https://doi.org/10.1080/14719037.2022.2037014*

29. Forman, C., Goldfarb, A. & Greenstein, S. (2009), 'The Internet and Local Wages: Convergence or Divergence? [Preprint]', *National Bureau of Economic Research Working Paper (14750)* . [Online; accessed 14.11.2022].
URL: *https://www.nber.org/system/files/working_papers/w14750/w14750.pdf*

30. Gentzkow, M., Kelly, B. & Taddy, M. (2019), 'Text as data', *Journal of Economic Literature* **57**(3), 535–74.

31. 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), 1012–1014.

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

33. Goldfarb, A. & Tucker, C. (2019), 'Digital Economics', *Journal of Economic Literature* **57**(1), 3–43.

34. Greenan, N., Mairesse, J. & Topiol-Bensaid, A. (2001), 'Information Technology and Research and Developement Impacts on Productivity and Skills: Looking for Correlations on French Firm-Level Data [Preprint]', *National Bureau of Economic Research Working Paper (8075)* . [Online; accessed 14.11.2022].
URL: *https://www.nber.org/papers/w8075*

35. Groseclose, T. & Milyo, J. (2005), 'A measure of media bias', *The Quarterly Journal of Economics* **120**(4), 1191–1237.

36. Hall, B. H., Lotti, F. & Mairesse, J. (2013), 'Evidence on the impact of R&D and ICT investments on innovation and productivity in Italian firms', *Economics of Innovation and New Technology* **22**(3), 300–328.

37. Hastie, T., Tibshirani, R. & Friedman, J. (2009), 'The Elements of Statistical Learning: Data Mining, Inference and Prediction (Second Edition)', *Springer, New York* .
URL: *https://hastie.su.domains/Papers/ESLII.pdf*

38. Katz, R. L. & Koutroumpis, P. (2013), 'Measuring digitization: A growth and welfare multiplier', *Technovation* **33**(10-11), 314–319.

39. Kinne, J. & Axenbeck, J. (2020), 'Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study', *Scientometrics* **125**, 2011–2041.

40. Kinne, J. & Lenz, D. (2021), 'Predicting innovative firms using web mining and deep learning', *PLOS ONE* **16**(4), e0249071.

41. Kotarba, M. (2017), 'Measuring digitalization: Key metrics', *Foundations of Management* **9**(1), 123–138.

42. Larsen, V. H. & Thorsrud, L. A. (2019), 'The value of news for economic developments', *Journal of Econometrics* **210**(1), 203–218.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407618302148*

43. Lenz, D. & Winker, P. (2020), 'Measuring the diffusion of innovations with paragraph vector topic models', *PLOS ONE* **15**(1), e0226685.

44. Lima, E., Sun, X., Dong, J., Wang, H., Yang, Y. & Liu, L. (2017), 'Learning and Transferring Convolutional Neural Network Knowledge to Ocean Front Recognition', *IEEE Geoscience and Remote Sensing Letters* **14**(3), 354–358.

45. Loper, E. & Bird, S. (2002), 'NLTK: The Natural Language Toolkit [Preprint]'. [Online; accessed 14.11.2022].
URL: *https://arxiv.org/pdf/cs/0205028.pdf*

46. Niebel, T. (2018), 'ICT and economic growth – Comparing developing, emerging and developed countries', *World Development* **104**, 197–211.

47. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**(Oct), 2825–2830.

48. Powers, D. M. (2011), 'Evaluation: from precision, recall and F-measure to ROC, Informedness, Markedness and Correlation [Preprint]', *arXiv preprint* . [Online; accessed 14.11.2022].
URL: *https://arxiv.org/ftp/arxiv/papers/2010/2010.16061.pdf*

49. Pukelis, L. & Stanciauskas, V. (2019), 'Using internet data to compliment traditional innovation indicators. [Preprint]', *International Society of Scientometrics and Infometrics (ISSI) 2019* . [Online; accessed 14.11.2022].
URL: *https://www.ippapublicpolicy.org/file/paper/5d073ea805eb6.pdf*

50. Rammer, C., Doherr, T., Krieger, B., Marks, H., Niggemann, H., Peters, B., Schubert, T., Trunschke, M. & von der Burg, J. (2021), 'Indikatorenbericht zur Innovationserhebung 2020'. [Online; accessed 14.11.2022].
URL: *https://ftp.zew.de/pub/zew-docs/mip/20/mip_2020.pdf*

51. Rammer, C. & Es-Sadki, N. (2022), 'Using Big Data for Generating Firm-Level Innovation Indicators – A Literature Review [Preprint]', *ZEW – Leibniz - Centre for European Economic Research Discussion Paper* (22-007). [Online; accessed 28. Nov. 2022].
URL: *https://ftp.zew.de/pub/zew-docs/dp/dp22007.pdf*

52. Saad, M. H., Hagelaar, G., van der Velde, G. & Omta, S. (2021), 'Conceptualization of SMEs' business resilience: A systematic literature review', *Cogent Business & Management* **8**(1), 1938347.

53. Salton, G. & Buckley, C. (1988), 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management* **24**(5), 513–523.

54. Schweikl, S. & Obermaier, R. (2020), 'Lessons from three decades of IT productivity research: towards a better understanding of IT-induced productivity effects', *Management Review Quarterly* **70**, 461–507.

55. Tetlock, P. C. (2007), 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of Finance* **62**(3), 1139–1168.

56. Torrey, L. & Shavlik, J. (2010), 'Transfer Learning', *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques, IGI Global, Hershey,* pp. 242–264.
URL: *https://doi.org/10.4018/978-1-60566-766-9.ch011*

57. Vos, J. (2009), 'Actions speak louder than words: Greenwashing in corporate America', *Notre Dame Journal of Law, Ethics & Public Policy* **23**(2), 673.

58. Xie, M., Jean, N., Burke, M., Lobell, D. & Ermon, S. (2016), 'Transfer learning from deep features for remote sensing and poverty mapping', *Thirtieth AAAI Conference on Artificial Intelligence* . [Online; accessed 14.11.2022].
URL: *https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12196/12181*

Appendices

A Text Processing Pipeline

Step	Description
1. Data filtering	Delete data points without text or with text duplicates.
2. Tokenisation	The texts are split into single words using the Python *Natural Language ToolKit (NLTK)* software package (Loper & Bird 2002).
3. Stopword filter	Delete words based on multiple stop word lists. The deleted words are usually not relevant for classification. An example of a stop word is „and".
4. Stemming	Different word variants are reduced to their base form. For example: The words „tree" and „trees" become „tree". The *Snowball Stemmer* (NLTK package) is used for this.
5. Short word removal	Words with a length of one or two are deleted. These are usually punctuation marks or special characters.
6. Unification of words	All capital letters are converted to lower case to reduce the vocabulary size.
7. Special character removal	Special characters are removed from the text.
8. Word selection	The 10,000 words with the highest TF-IDF score are extracted. The remaining words are deleted to reduce the dimension of the data and the „noise" in the text.

Table A.1: The text data processing pipeline for news article and firm website data. The pipeline filters irrelevant data, extracts words from a text, and performs a standardisation.

B News Data

Metric	Language	mean	median	std. dev.	min	max
Number of characters	German	3,652	3,450	1,820	1,000	33,004
Number of words	German	515	485	261	115	4,563
Number of characters	English	3,125	3,188	1,180	795	5,196
Number of words	English	525	535	200	130	899

Table A.2: Descriptive statistics for 25K news (per language). The statistics are reported before the text processing of the data with the natural language processing pipeline.

C Model Performance

word	importance	word	importance
digital	0.13423	onlin	0.00578
digitalis („translated”)	0.01936	internet	0.00549
digitization	0.01674	app	0.00529
googl	0.01244	apps	0.00452
smartphon	0.01144	pixel	0.00436
softwar	0.01003	gmbh	0.00434
appl	0.00846	android	0.00421
user	0.00818	analog	0.00403
comput	0.00675	samsung	0.00401
facebook	0.00662	algorithm	0.00401
job market („translated”)	0.00656	data	0.00391
bitcoin	0.00592	iphon	0.00374
use („translated”)	0.00590	virtual	0.00364

Table A.3: Most important words in the random forest model for the classification of news. The illustrated feature importance is the „Mean Decrease in Impurity” (MDI) measure. Some of the listed words are manually translated from German to English.

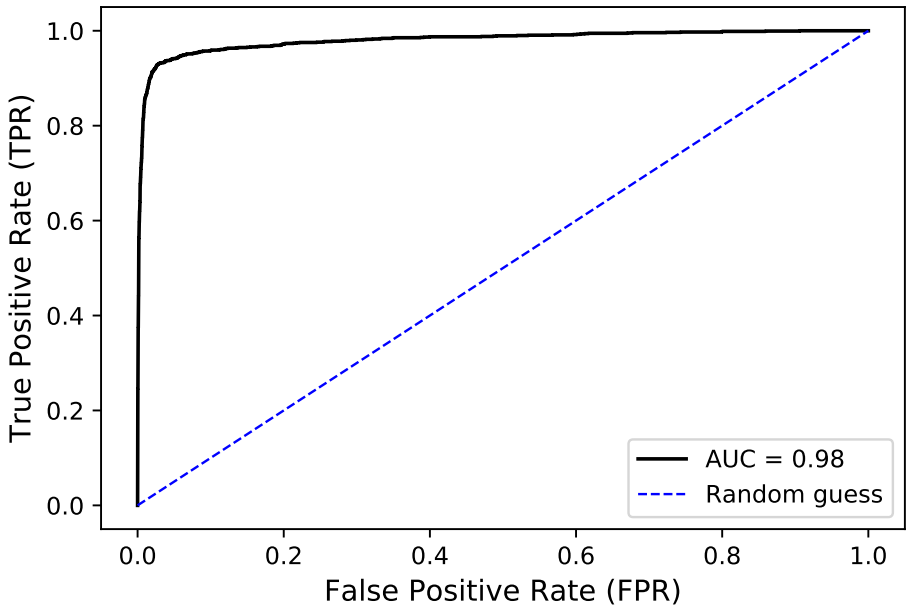


Figure A.1: ROC plot and AUC value (bottom right) for the regression model trained on the news articles. The baseline values are shown on the diagonal. Own illustration.

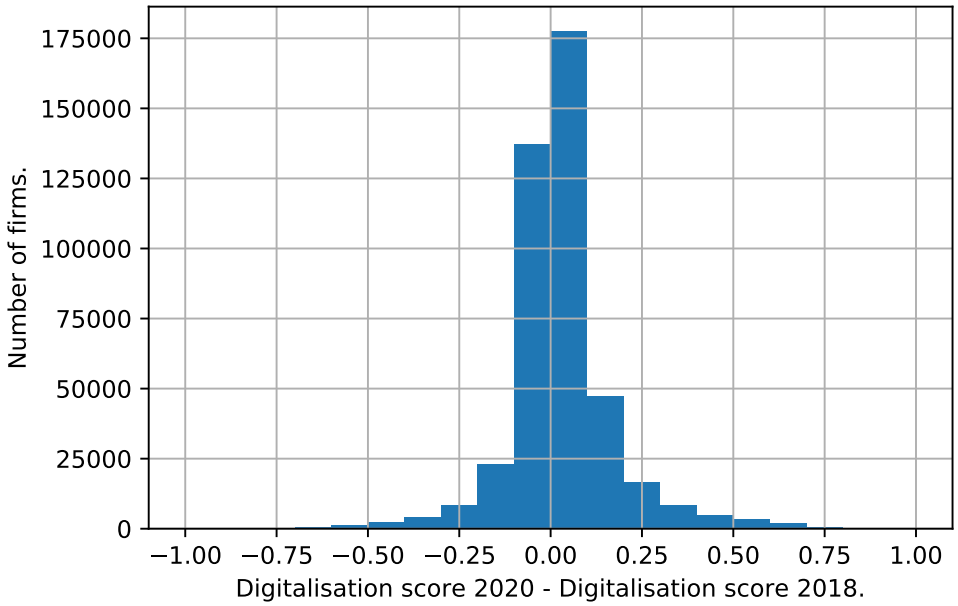


Figure A.2: Histogram indicating the change in the firm digitalisation score. The calculations are based on the data from the years 2018 and 2020. Own illustration.

D MIP Data

Number	Question	Answers
1	Use of digital platforms for delivering products or services	None, Low, Medium, High
2	Use of social networks to contact customers and obtain new customers	None, Low, Medium, High
3	Customisation of products through digital channels	None, Low, Medium, High
4	Methods of digital price differentiation	None, Low, Medium, High
5	Use of digital sources to collect data	None, Low, Medium, High
6	Digital integration of suppliers, business and other cooperation partners	None, Low, Medium, High
7	Use of digital media/tools for crowd sourcing of innovative ideas	None, Low, Medium, High
8	Use of machine learning or artificial intelligence	None, Low, Medium, High

Table A.4: List of digitalisation questions in the Mannheim Innovation Panel 2020 (MIP 2020) with the potential answers. The question block has the following title: „How important are the following digital elements for the current business model of your enterprise?".

E Firm Resilience

	mean	sd	min	max	count
digitalisation 2020	0.258363	0.2091660	0.001176	0.9266	176,902
digitalisation 2018	0.227544	0.2041340	0.001122	0.9165	176,902
Δ digitalisation 2020-2018	0.030818	0.1455610	-0.817273	0.8684	176,902
rating 2021	3.350102	0.5029563	1	5	176,902
rating 2020	3.379925	0.4906057	1	5	176,902
rating 2019	3.440050	0.4767542	1	5	176,902
rating 2018	3.458093	0.4701994	1	5	176,902
Δ rating 2021-2019	-0.089947	0.3549380	-3.29	3	176,902
employees 2020	0.033506	0.3395655	0.001	64.5	176,902
employees 2018	0.032793	0.3173355	0.001	64.5	176,902
Δ employees 2020_2018	0.000713	0.1276257	-36.77	29.3	176,902

Table A.5: Summary statistics for the „firm resilience" estimation sample that is based on MUP and web data. The reported number of employees per firm is stated in thousands. The statistics for the categorical data are shown in a separate table.

Location (German State)	Frequencies
Baden-Württemberg	27,556
Berlin	10,292
Brandenburg	3,951
Bremen	2,250
Hamburg	6,531
Bavaria	31,544
Saxony	4,059
Thuringia	1,641
Hesse	14,982
Mecklenburg-Vorpommern	1,268
Lower Saxony	19,034
North Rhine-Westphalia	35,117
Rhineland-Palatinate	8,059
Saarland	2,130
Saxony-Anhalt	1,269
Schleswig-Holstein	7,219

Sector	Frequencies
Agriculture, forestry and fishing (1-3)	1,647
Mining and quarrying (5-9)	200
Manufacturing industry (10-33)	20,579
Energy supply (35-35)	939
Water supply; sewage and waste disposal and pollution clean-up (36-39)	1,040
Construction (41-43)	19,597
Wholesale and retail trade; repair of motor vehicles (45-47)	35,602
Transport and storage (49-53)	4,108
Accommodation and food service activities (55-56)	6,910
Information and communication (58-63)	8,743
Provision of financial and insurance services (64-66)	6,133
Real estate and housing (68-68)	6,849
Provision of professional, scientific and technical services (69-75)	26,745
Administrative and support service activities (77-82)	11,440
Public administration, defense; social security (84-84)	824
Education and teaching (85)	3,460
Health and social services (86-88)	11,197
Art, entertainment and recreation (90-93)	3,210
Provision of other services (94-96)	7,658

Legal form	Frequencies
Self-employed profession („Freie Berufe")	9,715
Commercial operation („Gewerbebetrieb")	38,634
BGB association („BGB-Gesellschaft")	9,592
Single firm („Einzelfirma")	8,542
GmbH & Co. KG	10,231
OHG	1,388
KG	1,187
GmbH	89,512
AG	2,081
eG	907
eV	4,697
UG	415

Founding period	Frequencies
< 1990	49,571
1990 - 1999	44,932
2000 - 2009	49,723
> 2010	32,676

Table A.6: The number of firms for the different categorical and processed MUP firm properties *location, sector*, *legal form*, and *founding period*. Groups with less than fifty observations are not shown in the table. The labels of the legal forms have been translated from German into English where possible (see brackets). The sector definition is based on the 2-digit NACE code and is shown in the brackets.

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

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

Discussion Papers are intended to make results of ZEW research promptly available to other economists in order to encourage discussion and suggestions for revisions. The authors are solely responsible for the contents which do not necessarily represent the opinion of the ZEW.