Scaling-up of an Index of I4.0-Readiness of German Companies - applying ML models based on website data

Rainer Frietsch (rainer.frietsch@isi.fraunhofer.de), Fraunhofer ISI
Denilton Darold (denilton.luiz.darold@isi.fraunhofer.de), Fraunhofer ISI

Short abstract

Worldwide industrial production undergoes a transformation. Challenges like the reduction of environmental impact (including energy efficiency) as well as changes in supply and value chains are addressed in this context. The most prominent transformation of industry is the digitalization of production processes. This, however, is partly a mean to achieve the above mentioned goals, but also a core challenge in itself as it is supposed to offer efficiency gains as well as new business models. The digital transformation of production - in particular in Germany - has become known under the notion of Industry 4.0 (I4.0).

Empirical analyses have shown that the engagement and use of I4.0-like technologies are strongly biased toward large enterprises and that the progress and the adoption of these technologies vary greatly between companies. Measures have been created to assess the progress and diffusion, mainly based on surveys or on case studies. One of these surveys is the German Manufacturing Survey which contains questions on the digitalization of industrial production in the manufacturing sector in Germany since 2012. Based on this survey, indices - the I4.0 as well as the AI-Readiness-Index - have been created. As the main aim of this paper, we intend to upscale the coverage of these indices beyond the participants in the different surveys since 2012 and use website data to train and test a machine-learning classifier to automatically assign I4.0 indices to companies according to their website content.

Literature

The Industry 4.0 concept[1] represents the emergence of a complex transformation that holds great innovation potential and competitive advantages for manufacturing firms. At the core of this transformation are Cyber-Physical systems (CPS), a term that refers to the software-intensive embedded mechatronic production based on modern information and communication technologies (Bildstein, A. (2014); Lerch et. al (2019)).

CPS and related technologies associated with industry 4.0 are often evolutionary steps to overcome modern challenges. Companies are often unprepared for change due to

[1] https://www.plattform-i40.de/IP/Navigation/EN/Home/home.html

day-to-day demands hindering the capacity to foresight and prepare for the future. To better understand digital transformation, it is necessary to investigate what drives these changes, how transformation takes place, and how prepared companies are for change (Brisco, 2022).

In order to identify and consequently measure innovation capacities regarding Industry 4.0, Lerch et. al (2016) proposed the I4.0 Readiness Index. This index is based on the use and diffusion of I4.0 enabling technologies, aiming to quantify the maturity and intensity of use of such technologies. The I4.0 Readiness index addresses three central technologies that are represented in the following table.

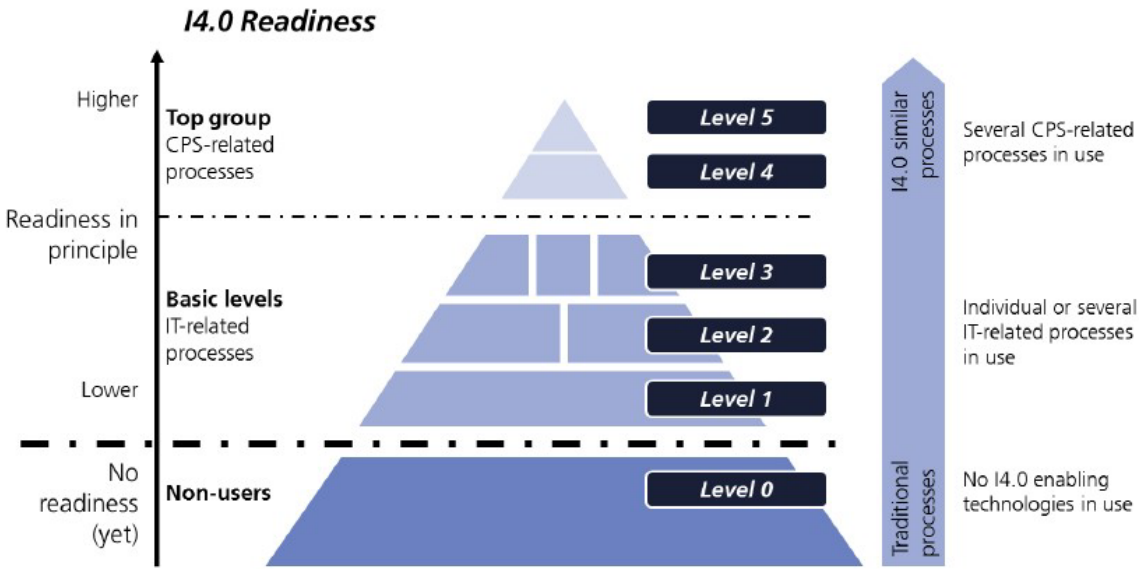
Table 1: Technology fields considered by the I4.0 Readiness Index

Technology field	Technologies	Level of application
A) Digital management systems	• software systems for production planning and steering • product lifecycle management systems	These applications are classified as basic technologies of IT and digitalisation and are thus assigned to IT-related processes.
B) Wireless human-machine communication	• digital visualisation at shop-floor-level • wireless mobile devices for programming machines and systems	This field is also assigned to the I4.0 basic technologies and thus to IT-related processes.
C) Cyber physical system (CPS) related processes	• near real-time production control system • automating of interal logistics • digital data exchange with suppliers and customers	These technologies already feature production elements of cyber-physical systems and are therefore considered advanced I4.0 technologies.

Source: Lerch et al. 2016.

By grouping these technology fields under different combinations, five levels emerged. The combinations have an important weight, as the relevance of these technology fields varies substantially. The fields (A) and (B), although crucial for digital integration, are still far from the CPS (C), which contains elements of networked production and, therefore, closer to the I4.0 than the previous ones. The result is an index in which companies can be classified closer to I4.0 implementations, based on technology fields and the use of several CPS-related processes in their production, as shown in Figure 1.

Figure 1: Industry 4.0 Readiness levels



Source: Lerch and Jäger 2019.

The index that was created by Lerch and Jäger (2019) is based on a number of items in the survey of about 1,300 German manufacturing firms. It has the following six levels and three aggregated stage-types:

Non-users do not show any I4.0-related activities:

Basic levels, have only minor/few I4.0-related acivities implemented in their production processes:

- Level 1 (beginners): companies that are active in only one out of three technology fields (see Table1).
- Level 2 (advanced beginners): active in two out of three technology fields (see Table1).
- Level 3 (advanced users): active in all three technology fields (see Table1).

Top group, as pioneers on the road to I4.0, with slightly higher readiness:

- Level 4: active in all three technology fields and use at least two CPS-related technologies (see Table 1, second column).
- Level 5: Companies that are active in all three fields of technology and use three CPS-related technologies (see Table 1, second column).

This index was included in the European Manufacturing Survey (EMS) 2018. The results, which are at the firm level, serve as the annotation to train a multi-class text classifier based on web scraped data. The details are described in the following methodological section.

Method

The approach to developing the hereby proposed classifier, which we name as I4.0 Readiness classifier, is to train a multi-class text classifier, where each one of the six levels (0-5) of the aforementioned I4.0 Readiness Index is one class. The input for the classifier is the data collected from the German part of the *European Manufacturing Survey* (EMS) 2018, combined with the text collected via web scraping from surveyed manufacturing companies' websites. In other words, we combined a semi-structured data source (survey) with an unstructured data source (web data) in order to create a machine learning model that automatically assigns classes, i.e., I4.0 Readiness levels, to companies based on their website content. The procedural steps regarding the data collection, cleaning, and processing are described in the following subsections.

Data collection

EMS Survey

The European Manufacturing Survey (EMS) is a survey of the diffusion of advanced production technologies and organizational concepts in the European manufacturing industry. It is organized by a Consortium of 16 European research institutions and has taken place every three years since 2001.

For this study, we use the data from the 2012, 2015, and 2018 surveys. This survey contains a comprehensive list of innovation-related data from which the I4.0 Readiness index was derived. The companies which contain the index are included in the classifier development. Of the 3,782 firms in the survey, 2,077 have the index and are therefore added to the training dataset.

Web scraping

In order to obtain text information from the firms, we scraped the manufacturing companies' websites. Of 3,782 firms in the survey, 3,501 had text collected via web scraping. The collection was performed using a python library named *scrawler*[2], which crawls the websites' domain identifying subpages and subsequently scrapes them, collecting the data according to the parameters provided, the main ones explained in Table 2:

Table 2: Parameter for the web scraper

[2] https://github.com/dglttr/scrawler

Parameter	Value Chosen	Comment
Max. distance to URL	4	Given an input URL, the tool collects the content recursively, crawling and scraping inner webpages to a distance of four subdomains/directories. Through that, we seek to collect not only the institutional and commercial text on the initial page, but also details of products and services in their specific subdomains.
Max. N of sites to be scraped	500	This choice seeks to balance comprehensive data collection and data volume within reasonable limits. Nevertheless, in most cases, there are less than 500 pages considering the distance to the input URL of 3.

Source: own compilation.

The resulting files are aggregated at the end of the web scraping process, splitting the data collected into metadata and text. The latter, our target, passes through a cleaning and basic transformation process. This process goes from removing special characters to splitting sentences, resulting in a readable, ready-to-use body of text.

Finally, the data sources are integrated via internal IDs and are ready to feed the machine learning model.

Machine Learning Model – Multi-text-classifier

Once the data is merged into one single dataset containing the index level and the text, we start the pipeline for the model training, which can be summarized in the following steps:

- Text normalization, removing punctuation, numbers, accentuation, especially the umlauts.
- Removal of stop words, i.e., semantically irrelevant words, like "the", "about", "such". On top of the standard list provided by the *nltk*[3] library, we added the list

[3] https://www.nltk.org/ - Natural Language Toolkit

from Sarica and Luo (2021), which contains a tailored list for engineering-related texts built upon the USPTO patent database.

- Stemming process, reducing the words to their root forms;
- Split the dataset into a training dataset and a testing dataset. The ratio adopted is 0.7, i.e., 70% for training and 30% for testing.
- Vectorization of text via TF-IDF[4], creating a numerical representation of the text.
- Model training using the algorithms Multinomial Naive Bayes, Multinomial Logistic Regression, Linear Support Vector Classifier (SVC) and Random Forests.
- Model evaluation using accuracy score through precision, recall, and FMeasure. The algorithm with the better indicators is then selected.
- A final evaluation is conducted via a confusion matrix, where it is possible to visualize the classifier's performance, having the number of actual and the predicted entries in a matrix. Additionally, a detailed F-Score table containing the accuracy scores of each class is generated.

Target Dataset

With the classifier already trained, we put it into action against a dataset of unknown registers, i.e., firms not present in the survey, to predict their I4.0 Readiness level based on their website text. The data source we used is the ORBIS[5] database which provides administrative records of firms worldwide. From ORBIS, we collect the firms from the manufacturing sector in Germany, using as criteria the two-digit NACE codes ranging from 21 to 33, encompassing firms framed as high-technology, medium-high-technology, medium-low-technology in the Eurostat[6] classification. The low-technology category in this classification refers to the food industry (Nace code 10-18) and is purposefully not included in our analysis. The other criteria applied are country (Germany), size (at least 20 employees), and a valid URL. These criteria are consistent with EMS selection with the exception of the sector, where we are stricter, as the EMS uses NACE codes ranging from 10 to 33 (total manufacturing sector). The resulting dataset contains 31.564 firms, which will be the target of the I4.0 Readiness index prediction.

---

[4] TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents

[5] ORBIS. Bureau Van Dijk. https://www.bvdinfo.com/en-gb/our-products/data/international/orbis

[6] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Hightech_classification_of_manufacturing_industries#:~:text=In%20order%20to%20compile%20statistics,%2Dtechnology%20and%20low%2Dtechnology.

Expected outcome and final remarks

The ultimate outcome is an I4.0 Readiness classifier, which can classify firms according to their website content. Such capability could drastically expand the reach of data collection and, consequently, the richness of empirical analysis regarding Industry 4.0.

We had scraped a number of websites from German manufacturing companies to which we intend to apply the I4.0 Readiness classifier, thereby extending the list of covered companies in the German manufacturing sector considerably. A limited set of about 20 of these assignments will be evaluated by experts in the field and by qualitative means.

Bibliography

Bildstein, A. (2014): Industrie 4.0-Readiness: Migration zur Industrie 4.0-Fertigung, In: Bauernhansl, T./ten Hompel, M./Vogel-Heuser, B. (Hrsg.): Industrie 4.0 in Produktion, Automatisierung und Logistik. Wiesbaden: Springer, 581-597.

Brisco, R., 2022. Understanding Industry 4.0 Digital Transformation. Proc. Des. Soc. 2, 2423–2432. https://doi.org/10.1017/PDS.2022.245

Lerch, C.; Jäger, A. (2019): Readiness for Industry I4.0 – Insights into the Upper-Rhine region. Karlsruhe: Fraunhofer ISI (Published online at: https://upperrhine40.eu/data/uploads/2020/02/Readiness_for_I4.0.pdf).

Lerch, C.; Jäger, A.; Meyer, N. (2016): I4.0-Readiness – Baden-Württemberg auf dem Weg zur Industrie 4.0? Karlsruhe: Fraunhofer ISI (Published online at: https://www.i40-bw.de/wpcontent/uploads/2020/08/Studie_I4.0_Readiness_Fh_ISI_publiziert-2_2016.pdf).

Sarica, S.; Luo, J. (2021). Stopwords in technical language processing. PloS one, 16(8), e0254937. https://doi.org/10.1371/journal.pone.0254937

Som, O. (2012): Innovation without R&D – Heterogeneous Innovation Patterns of NonR&D-Performing Firms in the German Manufacturing Industry. Wiesbaden: SpringerGabler.