Technological Forecasting & Social Change 207 (2024) 123627

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Unlocking machine learning for social sciences: The case for identifying Industry 4.0 adoption across business restructuring events

Fabio Lamperti

Department of Economic Policy, Università Cattolica del Sacro Cuore, Via Lodovico Necchi, 5 – 20123, Milano (IT), Italy

ARTICLE INFO

JEL classification:
C53
C61
O33

Keywords:
Machine learning
Natural language processing
Industry 4,0
Technology adoption
Restructuring events

ABSTRACT

In recent years, advancements in machine learning (ML) have facilitated the utilisation of big data across various academic disciplines. Nonetheless, these techniques still require a high-level of programming and data science expertise, making them inaccessible to many researchers and hindering the potential for knowledge advancements. This paper presents a framework for identifying the adoption of Industry 4.0 (I4.0) technologies among European firms that have undergone restructuring events. Existing studies on I4.0 adoption rely on diverse data sources at different levels of aggregation (e.g., countries, sectors, firms), spanning various time periods and technological domains. While this diversity often complicates result comparison, it also drives researchers and institutions to explore new data sources to assess technology adoption. Our identification methodology is based on the implementation of ML techniques using STATA, a well-established and user-friendly statistical software. We offer a step-by-step guide based on recently developed commands, allowing for comparison of model performance and analysis of model features. Our findings underscore the potential of ML algorithms as a robust tool for collecting new firm-level data on I4.0 adoption. Specifically, we observe that business restructuring events predicted as I4.0-related conform to adoption patterns identified in prior studies, across countries, sectors and over time.

1. Introduction

Nowadays, unstructured information embedded within texts presents a valuable complement to the structured and encoded data typically used in empirical research (Chen and Schintler, 2023; Miric et al., 2022). Consequently, social scientists are increasingly turning to textual documents in their research (Gentzkow et al., 2019). Traditionally, economists and other researchers in social sciences have relied on keyword-based approaches when needing to construct empirical variables that capture theoretical constructs from textual data, such as social media, firm reports, financial news, patent texts, advertisements, speech from politicians. Such approaches usually require researchers to create keyword dictionaries to define these constructs, followed by the searching of text documents for potential matches (e.g., Castellani et al., 2022; Cockburn et al., 2019; Felice et al., 2022; Martinelli et al., 2021; Van Roy et al., 2020). As Miric et al. (2022, pp. 2) argue, this approach proves effective in contexts where "*the theoretical constructs and language used to describe them are clear, well-established, and commonly accepted*". However, in many other contexts where theoretical constructs are rather vague or subject to individual interpretation – such as, the emotional value a customer attributes to a product or service – applying such methodology becomes more challenging due to the difficulties associated with defining and validating a dictionary of keywords.[1] Therefore, over the past decade, an increasing number of researchers in economics, business and management sciences have turned to machine learning (ML) algorithms to analyse large-scale text data. This allows them to define both theoretical constructs and build empirical variables directly from the data (e.g., Bhandari et al., 2023; Fantechi and Modica, 2022; Herrera et al., 2022; Kim et al., 2023; Mikko et al., 2022; Tidhar and Eisenhardt, 2020; Saura et al., 2023; Shrestha et al., 2021; Veiga et al., 2000; Vuorio and Torkkeli, 2023).

The recent excitement on the use of big data, ML, and other data science techniques has triggered a lively debate on where and how these methods can be applied in research (e.g., Agarwal and Dhar, 2014; George et al., 2016). More specifically, some studies have explored the conceptual distinction between supervised and unsupervised ML

E-mail address: fabio.lamperti@unicatt.it.

[1] The conversation on pros and cons of keyword-based and ML approaches has been extensively covered in the literature, highlighting that the choice of which is the best methodology to apply largely depends on the context of application, the complexity of the textual data to classify, and the research design. For a detailed discussion see, among the others, Miric et al. (2022).

https://doi.org/10.1016/j.techfore.2024.123627
Received 12 August 2023; Received in revised form 16 April 2024; Accepted 25 July 2024

approaches, and more traditional keyword-based approaches, along with their empirical implication (Miric et al., 2022). Other works have analysed how different ML techniques may be used for alternative research purposes in economics and social sciences (Gentzkow et al., 2019), or focused on the methodological and statistical details of ML techniques in general, and related methods (Jurafsky and Martin, 2009; Murphy, 2012).

Despite the latest surge in studies applying ML methods (e.g., topic modelling (TM) to uncover the underlying topics in large textual data, natural language processing (NLP) techniques to classify text data, latent semantic analysis (LSA) to understand hidden semantic relationships between words) in economic-related fields, most works present applications rooted in prior basic knowledge of data science methods and programming skills using specific software, mostly Python. Consequently, one critical drawback of this scenario is that many academics with traditional background in economics, business and management may hesitate to use such techniques due to the substantial investment of resources, notably time, required to grasp the fundamentals of these methodologies, i.e. learning the programming languages, debugging and optimising code. Therefore, most researchers in such situation will either refrain from venturing into the "*ML world*" or resort to hiring external experts to address the related challenges.

We address this gap by exploring the application of supervised ML models using a pool of recently developed commands in STATA, a well-known and user-friendly statistical software package largely used by researchers across all social sciences. We provide a walkthrough guide of the use of such methods in the classification of large unstructured text documents, starting from data preparation steps and techniques to the analysis of the features of ML models, and concluding with the evaluation and the comparison of the performance from different models. Specifically, we apply a range of ML algorithms – namely, support vector machine (SVM), k-nearest neighbour ($k-$NN), random forest (RF), decision tree (DT), gradient tree boosting (GTB) and regularised multinomial logistic regression (RM).

Our empirical application revolves around the classification of business restructuring events sourced from Eurofound's European Restructuring Monitor (ERM) database. We analyse these events in relation to mentions of the adoption of automation and Industry 4.0 (I4.0, or fourth industrial revolution) technologies in textual descriptions. Therefore, the methodology outlined in this study offers researchers a novel tool to identify the adoption of I4.0 technologies at the firm level. Furthermore, the ML approach we implement is not only scalable but also adaptable to alternative data sources (e.g., firm reports). This, in turn, may lead to increased availability of data concerning the adoption of emerging I4.0 technologies. Since large-scale and freely accessible data sources, especially at the firm-level, remain scarce – given the ongoing nature of the present digital transformation (Castellani et al., 2022) – it becomes crucial to explore new data sources. Through our research, we also show fresh insights into the characteristics and potential determinants of business restructuring events, providing a valuable instrument for stimulating future analyses on the international transformation of firms related to I4.0 (Luo and Zahra, 2023). We further delve into descriptive evidence regarding the distribution of identified I4.0-related restructuring events across various dimensions, including time, countries, industries, and event types. For simplicity, throughout the paper we will refer to such pool of technologies (e.g., Mariani and Borghi, 2019) using I4.0 and automation alternatively.

In summary, this study makes a threefold contribution. Firstly, we add to the expanding literature demonstrating how recent advancements in ML can enhance the quality of empirical research. For example, supervised ML tools can effectively learn and predict relevant textual classifications (e.g., patents) from vast data samples, tasks that would be significantly beyond human capacity to read, analyse, interpret, and code. Secondly, we contribute to practice in the field of social sciences broadly defined. Our methodology based on the application of new user-friendly commands in STATA is easily applicable for most researchers in social sciences who are interested in utilising ML techniques in their empirical investigations, as it does not require specific knowledge of programming languages (e.g., Python) or a comprehensive understanding of ML algorithms and data science techniques. Furthermore, we offer a detailed exposition and clear guidance on the preliminary steps researchers should take to prepare and explore their data before implementing supervised ML methods on large-scale text databases. Thirdly, we stress the need for new firm-level data sources on automation technologies, and we provide fresh insight on I4.0 adoption and its relationship with business restructuring events of varying nature (e.g., job-creating vs job-displacing). Our exploration of this relationship using ML algorithms represents just one of several interesting phenomena that embody the key characteristics (e.g., type of activity involved, type of worker involved) and determinants (e.g., cost-related factors, demand dynamics, profitability aims, changes in business model/strategy, adoption of new technologies) behind business restructuring events. We argue that these latter two contributions are both theoretically meaningful and empirically relevant for academics in economics, business, and management sciences.

The reminder of the paper is organised as follows. In Section 2, we briefly review the main studies applying ML techniques in economic-related fields, as well as the current state of data availability on I4.0 adoption. Section 3 presents the methodology implemented and the data used, providing step-by-step guidance. Section 4 first discusses the preliminary steps needed to analyse the features of ML models in order to better understand the data and then discuss the performance of different models. In Section 5, we discuss the insight from the predictions obtained from our best-performing classification model. Section 6 concludes presenting the implications for researchers, limitations, and future research directions.

2. Background literature

ML methods, alongside other statistical, big data, and data analytics tools broadly defined, relate to the interdisciplinary field of data science (George et al., 2016). In its essence, data science focuses on the collection, preparation, and analysis of large-scale data, emphasising their role in inference (Dhar, 2013). The array of tools encompassed within the discipline empowers researchers and practitioners to develop models able to capture, visualise and analyse the underlying trends and patterns within the data, thereby providing new insights. Specifically, McAfee and Brynjolfsson (2012) underscore that contemporary data science applications allow to tackle the three key features characterising big data: (i) *volume*, indicating the size of datasets nowadays available, which keeps increasing in terms of both available variables and number of observed elements; (ii) *velocity*, referring to the growing rate at which these data becomes available for collection and subsequent analysis, and; (iii) *variety*, denoting the plurality and the underlying differences of different types of data (e.g., structured vs unstructured, numerical vs textual).

In today's data-driven world, data science techniques play a pivotal role, offering both immense opportunities and challenges across various sectors. While the debate of new opportunities created for practitioners, such as more informed decision-making, has been ongoing for decades (e.g., McAfee and Brynjolfsson, 2012), the academic world has begun to explore big data and related opportunities on a wider basis only in recent years (e.g., Agarwal and Dhar, 2014; George et al., 2016). Consequently, the academic debate has primarily revolved around two main directions, which will be ever more relevant in the future of academic research. With the growth in both volume and variety of data over time, researchers can not only provide better and more accurate answers to existing research questions (Chen and Schintler, 2023; Lindner et al., 2022), but also pioneer the development of new theories and unexplored questions (Shrestha et al., 2021). Lately, these two main advantages have catalysed the application of data science techniques, such as ML,

across virtually every academic discipline.

Our work aligns with the first research direction devised. Specifically, as the digital transformation brought by I4.0 and automation technologies become ever more pervasive, we address the urgent need to explore new data sources to define reliable adoption measures. This remains paramount given the scarcity of firm-level sources. Therefore, in the subsequent sections, we review the recent, pivotal literature applying ML techniques, followed by a discussion on critical contemporary challenges concerning data availability on I4.0 adoption.

Table 1
Key recent contributions to the literature applying ML techniques in economics and related fields.

Authors	ML technique/s used	General objective	Empirical application	Field
Chung and Sohn (2020)	Deep learning model based on convolutional neural network and bidirectional long short-term memory.	Identify the semantic features of patent texts (abstracts and claims).	Classify semiconductor-related patents granted at the USPTO from 2000 to 2015 in three levels of importance based on annual forward citations.	Innovation studies
Bas et al. (2021)	Various supervised ML methods for classification: support vector machines, artificial neural networks, deep neural networks, gradient boosting models, distributed random forests, and extremely randomized forests.	Identify key elements driving customers' adoption of new technologies, so to provide additional ground for strategic decision making.	Use US customer data on self-reported perception of/inclination towards electric vehicles and the role played by their social structure in the attitude towards this type of vehicle.	Innovation studies
Montobbio et al. (2022)	Unsupervised ML model based on latent Dirichlet allocation	Identify the semantic features of patent texts.	Classify all patent applications at the USPTO filed between 2009 and 2018 based on their reference to labour-saving technologies or the application of the same.	Innovation studies
Kim et al. (2023)	Different predictive models: logistic regressions, decision trees, random forests, gradient boosting, support vector machines, and naïve Bayes.	Obtain future predictions using past data, evaluate the relative importance of different features in determining predictions.	Predict chances of market success for more than 200,000 start-ups using Crunchbase data between 2011 and 2021 and exploiting both firm-level and industry-level information.	Innovation studies
Saura et al. (2023)	Both supervised and unsupervised ML algorithms: support vector classifier, multinomial naïve Bayes, logistic regression, random forest classifier, and topic modelling based on latent Dirichlet allocation algorithm.	Identify the semantic features of user-generated contents on social media for sentiment analysis.	Explore the boundaries of open innovation using 586,348 user-generated contents on Twitter by first classifying tweets based on sentiment analysis, and then identifying the main topics.	Innovation studies
Veiga et al. (2000)	Supervised neural network analysis.	Demonstrate the usefulness of applying ML techniques in social sciences to uncover underlying patterns in textual information.	Recognise cultural patterns among top executives in French and British firms based on survey data.	International business
Garbe and Richter (2009)	Neural network analysis.	Identify causal relationships between constructs with an established theoretical base, but not properly tested on empirical ground.	Identify the appropriate level of internationalisation and test its causal effect on firm performance using a sample from UNCTAD's 100 most internationalised corporations.	International business
Bhandari et al. (2023)	Computer-aided text analysis based on a set of predefined keywords.	Identify causal relationships between constructs with an established theoretical base, but not properly tested on empirical ground.	Define a measure of firm-level digitalisation level starting from textual information in firm annual reports, using a sample of 571 US manufacturing firms. Such measure is then used to uncover the causal effect of digitalisation on firm performance and assess the role of internationalisation as moderator in such relationship.	International business
Vuorio and Torkkeli (2023)	Artificial neural network analysis.	Identify causal relationships between constructs with an established theoretical base, but not properly tested on empirical ground.	Explain the complex and multifaceted relationship between multiple, distinct aspects of dynamic managerial capabilities and firm's internationalisation, on a sample of 144 Finnish SMEs.	International business
Fantechi and Modica (2022)	Various supervised ML algorithms: logistic regression classifier, k-nearest neighbours, decision tree classifier, naïve Bayes Gaussian classifier, linear support vector machine, and simple neural network.	Obtain future predictions using past data, evaluate the relative importance of different features in determining predictions.	Predict different scenarios of recovery and resilience for 133 municipalities located in Italian regions hit by the 2016's earthquake, using socio-economic and natural characteristics.	Economic geography/ Regional studies
Mikko et al. (2022)	Unsupervised ML technique based on topic modelling.	Identify the semantic features of research papers, based on similarities in the vocabulary.	Identify the level of smart specialisation of regions based on common research domains across universities located in Arctic Scandinavia.	Economic geography/ Regional studies
Tidhar and Eisenhardt (2020)	Various ML algorithms: decision trees, random forests, and regularised multinomial logistic regression.	Apply ML techniques to support theory building.	Define optimal choices of revenue model combining ML and multi-case theory building, using data from 66,652 products on Apple's App Store.	Management studies
Savin et al. (2022)	Supervised topic modelling for classification.	Perform unbiased classifications based on relevant features without resorting to experts.	Classify more than 250,000 start-ups in the Crunchbase dataset between 2009 and 2019, using textual descriptions to model features predicting future success.	Management studies
Miric et al. (2022)	Compare hand-coded data with various supervised ML algorithms for classification (e.g., random forests, support vector machines, neural networks, logistic regression, naïve Bayes, and k-nearest neighbours).	Identify the semantic features of patent texts (abstracts and claims), present a guide to the use of ML techniques.	Classify about 650,000 patents granted at the USPTO from 1985 to 2018 as artificial intelligence-related or not.	Management studies

2.1. ML applications in economics and related fields

ML techniques are rapidly gaining popularity among researchers across various fields in economics. One significant reason for this increasing adoption is the recognition by many academics of the potential for gaining insights from these research techniques and their complementary use alongside traditional methodologies. As argued by Lindner et al. (2022), ML applications such as cross-validation, classification, and regression algorithms can expand the toolkit available to researchers, allowing them to produce more robust and generalizable results. This expansion is further facilitated by examining research practices in adjacent disciplines. Consequently, an increasing number of studies have recently begun to incorporate ML techniques into economics and related fields such as international business, economic geography, management, and innovation studies (see Table 1 for a summary).

The proliferation of research applying various ML techniques in innovation studies has been remarkable in recent years, resulting in a growing number of interesting empirical applications. For instance, a notable recent contribution from Chung and Sohn (2020) applies deep learning techniques based on different neural network algorithms for the early identification of valuable patents in the field of semiconductors, utilising both patent-specific and contextual information. Similarly, Bas et al. (2021) implement several supervised ML techniques (e.g., SVM, GTB, RF, neural networks) to categorize potential electric vehicle purchasers, thus predicting future adoption trends in the market, using data from a tailored survey of US residents. Montobbio et al. (2022) use probabilistic TM techniques, such as latent Dirichlet allocation, to analyse patent applications at the USPTO and identify those describing labour-saving technologies or applications of the same. Kim et al. (2023) focus on the critical tusk of predicting the future success of a new business venture by implementing several ML models and considering both internal, business-specific information of new startups and contextual, industry-level factors for a large sample of more than 200,000 firms. In addition to these practical applications, ML techniques have recently been used to push the boundaries of existing theories and to provide a "*reality check*" by extracting insights from unstructured text data. In this vein, Saura et al. (2023) employ a range of ML techniques, including NLP, TM, and computer-aided text analysis (CATA) techniques, to analyse more than half a million tweets, thus deriving current constraints and future research directions in the realm of open innovation.

Among international business studies, a significant contribution emerges from Veiga et al. (2000), who were pioneers in applying ML techniques in the field and, more broadly, among the earliest to do so in economic studies. They utilize neural networks to uncover underlying firm-specific patterns associated with national culture. Building on this groundwork, Garbe and Richter (2009) apply the same ML technique to investigate how decisions regarding the international configuration of firms impact their performance. More recently, ML methods have garnered increased attention in the international business literature. Bhandari et al. (2023) employ a CATA technique and an NLP algorithm to construct a measure of digitalization for US manufacturing firms based on their annual reports. This measure is then used by the authors to examine the moderating role of firms' internationalisation (e.g., FDI) on the relationship between digitalization and firm performance. Additionally, Vuorio and Torkkeli (2023) apply neural network analysis to explore the dynamic managerial capabilities of firms that are associated with early internationalisation decisions.

Driven by an increasing consensus regarding the potential of ML techniques (Chen and Schintler, 2023), a few studies have commenced applying such methods also in the field of economic geography and regional studies. For instance, Fantechi and Modica (2022) employ various supervised ML algorithms (e.g., SVM, k-NN, DT) to forecast the recovery capacity of local labour markets in Italy following three distinct earthquakes. Meanwhile, Mikko et al. (2022) utilize unsupervised TM tools to uncover common research themes among the vocabulary used by researchers across cities in the Arctic Scandinavian region.

Alike other research fields, also management studies have recently witnessed a notable increase in the application of various ML algorithms and methods. For instance, Tidhar and Eisenhardt (2020) leverage insights from multi-case studies and three ML techniques (i.e., DT, RF and RM) to support a novel theory-building approach for optimising firms' choice of their optimal revenue model. Similarly, Savin et al. (2022) use TM to classify start-up companies based on the relevance of specific industry-related topics, starting from text documents from Crunchbase dataset. Moreover, Miric et al. (2022) contribute to this trend by experimenting with several supervised ML classification techniques to identify patents related to artificial intelligence (AI) at the USPTO. They also offer guidance on employing different approaches (e.g., supervised vs unsupervised methods, bag-of-words vs embedding-based approaches) and interpreting results, making their work very close to the one we present here.

While our aim does not extend to a comprehensive review of the recent yet burgeoning literature applying ML across all economic-related fields, the aforementioned studies clearly indicate that ML applications are swiftly advancing research practices. They provide powerful new tools not only to corroborate and reinforce traditional research methodologies based on hypothesis-testing (Chen and Schintler, 2023; Lindner et al., 2022) but also to enhance theory building (Shrestha et al., 2021).

2.2. Data sources on Industry 4.0 adoption

Research on I4.0 technologies has significantly advanced over the past decade, encompassing various domains such as productivity, labour markets, business models, and international strategies. Correspondingly, empirical evidence surrounding these areas has steadily expanded. However, many empirical studies in this field offer insights that are challenging to compare on a quantitative ground, primarily due to the limited availability of measurement tools capturing technology adoption across diverse firms, countries, sectors, and time frames. While there has been a rise in cross-country survey data on technology uptake in recent years, they still provide only partial insights due to their cross-sectional nature. Additionally, existing studies often focus on a select few technologies due to the lack of comprehensive information sources (Brynjolfsson et al., 2019; Cockburn et al., 2019).

A significant portion of quantitative studies on automation concentrates on the adoption of individual technologies, particularly robots, leveraging cross-country and sectoral data provided by the International Federation of Robotics (e.g., Acemoglu and Restrepo, 2020; Anzolin et al., 2022; Cette et al., 2021; Du and Lin, 2022; Graetz and Michaels, 2018). Alternatively, some studies assess technology adoption within specific countries using administrative data on machinery imports (e.g., Acemoglu et al., 2020; Ballestar et al., 2020; Domini et al., 2021, 2022; Lamperti et al., 2023). Other works rely upon ad hoc private surveys – like, for instance, the European Manufacturing Survey – collecting detailed information on firm-level adoption across different European countries (e.g., Dachs et al., 2019; Jäger et al., 2015; Kinkel, 2020), or on single countries leveraging on either responses from dedicated questionnaires or survey data from national statistical institutes (e.g., Dalenogare et al., 2018; Frank et al., 2019; Marcucci et al., 2021; Pedota et al., 2023). Finally, several contributions have addressed the implications of adopting I4.0 technologies through case studies based on specific sectors or a small number of firms (e.g., Cugno et al., 2021; Müller et al., 2018).

On the one hand, data collected and freely available from international institutions and/or statistical offices, such as the European

Investment Bank (EIB, 2019, 2021) and Eurostat (Eurostat, 2023), now provide cross-country insights from a representative sample of firms adopting various I4.0 technologies at different unit of analysis (i.e., country level, sectoral level, firm-size groups). However, these sources are limited in scope, as they only provide cross-sectional evidence with limited comparability across technologies. On the other hand, the above-mentioned data sources on I4.0 technologies all present various limitations. Most notably, either the limited country coverage, the narrow focus on one specific technology, or the varying level of aggregation for adoption. Consequently, this scenario calls for new, updated, and complementary sources of information on the adoption of automation technologies at the firm level, spanning countries and time frames, and encompassing a much broader and comprehensive technological domain. Such data would enable researchers to deepen their understanding of the micro-level implications of digitalization across nations.

3. Methodology

The methodology framework devised in this paper is grounded in the application of machine learning (ML) techniques based on natural language processing (NLP) for classification. These include support vector machine (SVM), k-nearest neighbour (k-NN), random forest (RF), decision tree (DT), gradient tree boosting (GTB), and regularised multinomial logistic regression (RM). We implement such data science tools using a set of recent commands for STATA, a statistical software package widely used in social sciences. Hereafter, we present a step-by-step guide for applying such methods in the classification of large unstructured text documents. Specifically, we use this application of ML methods to: (i) show how to construct the training dataset for classification purposes and analyse input data in ML models; (ii) discuss how different supervised ML algorithms can be used and compared to identify the best-performing model; (iii) demonstrate how we can evaluate and understand the performance of different ML methods. Since the quest for a more objective science is rapidly surging (Cerulli, 2021), we believe that unlocking the potential of ML and other techniques used to work with big data for the broader public of social scientists with no robust skills in data science will help improve empirical evidence and spur new theoretical understanding of real-world problems.

We first present additional insights on data preparation steps and model features, based on recent STATA commands developed by Escobar (2015) and Williams and Williams (2014). Then, we discuss and show how to apply different ML commands in STATA by Cerulli (2022), Schonlau and Zou (2020), and Guenther and Schonlau (2016). These commands are based on the STATA/Python integration (available from STATA version 16 onward) and on STATA's Java implementation. Clearly, we note that none of these models or commands presented has been developed here. However, we integrate the application of the commands proposed by Schonlau and Zou (2020) and Guenther and Schonlau (2016) to implement optimal model tuning via k-fold cross-validation, which is already embedded in Cerulli's (2022) command.[2]

Since our scope lies in showing how these new commands allow overcoming existing barriers to a wider use of data science techniques, we do not focus on presenting the background on any of the models used. Therefore, we redirect researchers to the appropriate and established resources (e.g., textbooks, publications) where they can learn more about the foundations of the models used and the STATA commands implemented.

The implementation of all algorithms and graphing have been coded in STATA 17. One technical caveat to be acknowledged is that the commands we use to implement ML algorithms require the installation of specific plug-ins, libraries, or additional programs. These include the STATA/Python integrated interface available from STATA 16 onward, Python 3.7 or later versions, as well as specific Python libraries for Cerulli's (2022) command. We invite readers to refer to each command's description to check the specific requirements, as well as to replicate the implementation based on comparable commands using alternative statistical software (e.g., R or Python).

3.1. Data

The empirical application we present aims at classifying business restructuring events from Eurofound's European Restructuring Monitor (ERM) database[3] as automation/I4.0-related or not. Researchers in economic fields like international business, economic geography and management studies have made abundant use of data on firms' cross-border investments (i.e., greenfield investments) and other types of business restructurings (e.g., offshoring, reshoring). Most studies employ data from private sources like the fDi Intelligence's fDi Markets database[4] (e.g., Anderson and Sutherland, 2015; Andreu and Lavoratori, 2022; Castellani and Lavoratori, 2020), due to the richness of the information. Alternatively, some studies employ other sources like Eurofound's ERM database (e.g., Barbieri et al., 2019, 2022), the UniClub More Reshoring dataset[5] (e.g., Ancarani et al., 2019), or small-size dataset built upon dedicated surveys (e.g., Coucke et al., 2007; Coucke and Sleuwaegen, 2008). A common thread of all the large-scale data sources just cited is the availability of additional text information describing related events, which represents a rich source of information in addition to readily available and codified variables.

Eurofound's ERM database includes 23,618 restructuring events, recorded in from May 2002 to January 2023. The bulk of detailed ERM micro data is not freely available from Eurofound website. However, access is provided upon official request, subject to the approval of the research project for which data will be used. We gained access to the complete and current ERM database in March 2023. This database provides a rich set of unstructured textual information – in the form of detailed event descriptions, description of the context and motivations – about business restructuring events involving firms operating within the European Union (EU) 27 countries, the United Kingdom (UK) (until the end of 2019) and Norway. The information on restructuring events is collected from daily newspapers and business press and is integrated with online resources like company websites (Eurofound, 2023). To grant quality and consistency, the information in the ERM database undergoes continuous monitoring and cross-checking. Additionally, data in the ERM adhere to stringent criteria: it exclusively covers significant, large-scale restructuring events occurring across the EU, encompassing either the elimination or creation of a minimum of 100 jobs or, alternatively, at least 10 % of the workforce at facilities with over 250 employees (Eurofound, 2023). Beyond text information on the restructuring event, the ERM database includes specific information on the firm involved, the sector in which it operates, the announced number of jobs lost or created and location of the plant/firm (both current and planned, in the case of business expansions or offshoring).

Textual data describing business restructuring events represents a rich source of information in order to understand both characteristics (e.g., type of activity involved – such as, manufacturing, R&D, marketing, logistics – or type of worker involved – for instance, blue collar vs white collars) and determinants (e.g., cost-related factors, demand dynamics, profitability aims, change in business model and/or strategy, adoption of new technologies) of each event. Given the size and the continuously growing nature of the dataset, it would be cost-prohibitive to employ human coders to classify such information. Therefore, hereafter we present an application of different supervised ML techniques, specifically NLP ones, to classify restructuring events, specifically looking at

[2] All STATA coding used is available upon request.

[3] See: https://www.eurofound.europa.eu/observatories/emcc/european-restructuring-monitor.

[4] See: https://www.fdiintelligence.com/fdi-markets.

[5] See: https://reshoring.eurofound.europa.eu/.

mentions to the adoption of new automation/I4.0 technologies.

3.2. Implementing ML algorithm to classify automation-related business restructuring events

The overall process followed is described by the sequence of steps shown in Fig. 1.

Step 1: The initial data management and preparation involve performing a preliminary cleaning of the full dataset. Specifically, we dropped observations for restructuring events with no or limited (e.g., only few words) text information available. This preliminary procedure resulted in a reduction of the number of restructuring events form the original 24,842 to the final size of 23,618 (i.e., we discarded 1224 events). Subsequently, starting from the full set of 23,618 observations, we drew a sub-sample consisting of 7199 restructuring events (i.e., about 30 %), on which we performed a manual *ex-ante* validation by reading through the text information (i.e., the event description) associated to each event, and scouting mentions to the adoption of new digital technologies of the I4.0 (i.e., automation technologies, advanced manufacturing technologies, AI, etc.).

In supervised ML applications, the data validation procedure aims to define an unbiased set of observations truly identifying the phenomenon of interest, resulting in the definition of a training dataset. Such training dataset is later used for training ML models and optimising them by identifying the optimal parameters.[6] More generally, data validation procedures focus on ensuring the quality, completeness, and reliability of the input data by following both general and specific criteria. General criteria usually consist in excluding missing data, not complete information, and duplicates which would bias the model training. Conversely, specific criteria usually depend on the researcher's objective. In our case, the main criterion followed to validate the data is to identify as I4.0-related only those business restructuring events whose description specifically states that new automation technologies are implemented (see the examples reported in Table 2).

Automation/I4.0 technologies were identified starting from the definition of Martinelli et al. (2021) and Cockburn et al. (2019) and from the definition of advanced manufacturing technologies of Castellani et al. (2022). We employed the definitions and the related set of keywords (i.e., dictionaries) used in these studies to read over text information and identify whether restructuring events involved the adoption of I4.0 technologies. We also augmented our dictionary by including combinations of the retrieved keywords with more common, and generally recurring, terms frequently used to address the adoption of new automation technologies (e.g., new manufacturing technologies, advanced industrial systems, smart factory, etc.).

Business restructuring events associated with the adoption of automation technologies – either pushed by or making specific mention to it – are likely to represent only a small sub-set of all restructuring events. Therefore, it is important to appropriately select training observations. If we were to randomly sample from the entire ERM database, I4.0-related restructuring events are likely to account for a very small proportion of the training data, this potentially undermining our classification performance. In order to tackle this issue, we sample observations for the validated dataset, and for subsequently training our models, only between 2009 and 2020 as this period corresponds to the recovering of business activities after the 2008 global financial crisis, when the rise of I4.0 initiative and the subsequent uptake of related technologies across EU countries took place (Castellani et al., 2022; De Backer et al., 2018; Kagermann et al., 2013; Mariani and Borghi, 2019; Teixeira and Tavares-Lehmann, 2022), up to the outbreak of the global Covid-19 pandemic, which implied a shock for business activities and likely induced many firms worldwide to rethink the organisation of their activities (Antràs, 2020; Blit, 2020; Di Stefano et al., 2022). At the same time, given that the definition followed to identify automation technologies predominantly denote industrial/manufacturing applications, we sampled the 7199 observations mostly from firms operating in manufacturing industries (70 % manufacturing, 30 % services and other sectors).

Our final dataset for model training featured about 2 % of I4.0-related events. Given the nature of the data used in this paper and the purpose of our work, we believe our approach grants sufficient balance and alternative approach (e.g., active learning techniques; Miric et al., 2022) would not necessarily return better classification performance. Finally, to provide further robustness and consistency to our validation procedure, our manual classification of restructuring events as automation-related or not was *ex-post* validated by two independent researchers blindly reading a random sample of 700 descriptions (i.e., about 10 %) of training observations and independently classifying events. According to the results of this second, independent assessment, our approach turned out to be effective in identifying automation-related restructuring events accurately. Notably, the comparison of *ex-ante* and *ex-post* validation results (i.e., discussing feedback from the two experts) highlighted that the main criteria applied during the *ex-ante* validation produced a more restrictive validation in terms of requirements for identifying (i.e., in terms of vocabulary suggesting a semantic link to technology adoption), although more comprehensive in terms of the technological domains.

Step 2: Before building our NLP classification model, we implemented different techniques for pre-processing the input data later used for both training and testing purposes, and final out-of-sample prediction (i.e., the full, original dataset of 23,618 events). First, we removed punctuation, double-spacing, extra white spaces, and special characters. Second, we used a stop-word list to delete unnecessary words from the text data (e.g., articles, conjunctions, pronouns, adverbs and all the most common prepositions in the English language). Since our text data pertain the description of business restructuring events taken from news articles and alike, hence not featuring any particular or technical terminology, there is no specific need to use a dedicated library of stop-words as it would be necessary in the case of patents pertaining a specific sub-set of technologies (Chung and Sohn, 2020). Finally, we implemented Porter's (1980) stemmer algorithm in order to remove the commoner morphological and inflexional endings from words in our text data. This step is frequently performed as part of the term normalisation process that is usually done when setting up ML models. These pre-processing steps are necessary to clean row text data and reduce its complexity and variety.

Step 3: We adopted the bag-of-words (BoW) method to convert our text data into a numerical representation.[7] BoW embody the frequency of occurrence of each word in the text body and are frequently used as features for training NLP classifiers. However, directly using BoW vectors in our model poses the risk of overweighting words which may be naturally more common in the text, thus potentially leading to less accurate predictions. Therefore, we computed term frequency-inverse document frequency (TF-IDF) vectors, which are word frequency

[6] It's worth emphasising that the concepts of "*validated dataset*", "*training dataset*" and "*validated model*" are frequently confused, despite being distinct. Put simply, a validated dataset serves as the foundation for defining a training dataset, by splitting it into training and test datasets. Some works in the literature also refer to test dataset as "*validation dataset*", meaning that the data is used to test the performance of a model already trained using the training dataset. Conversely, model validation involves assessing the performance of a model in terms of accuracy, precision, and other metrics either during or after the process of defining the optimal model parameters.

[7] All pre-processing procedures in step 2 and the creation of BoW were implemented in STATA using `txttool` command (Williams and Williams, 2014).

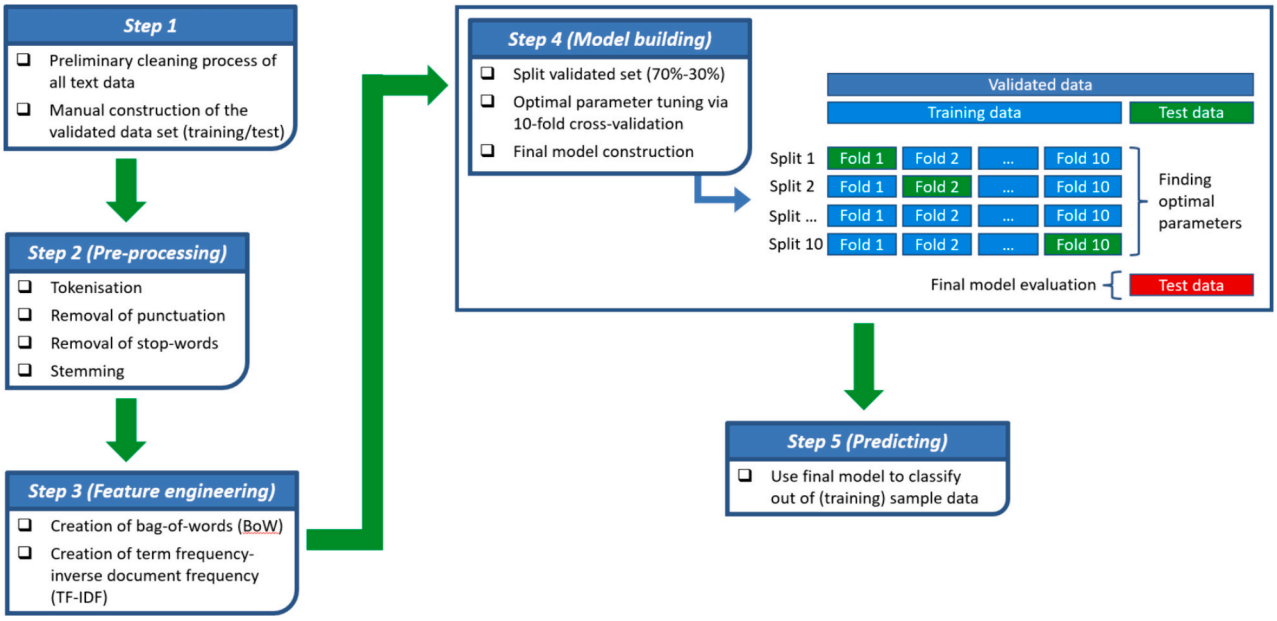

Fig. 1. Flow scheme of the proposed model.
Notes: Authors' own elaborations.

Table 2
Examples of business restructuring events identified as I4.0/automation-related during initial (ex-ante) dataset validation.

Firm	Country	Sector	Event type	Description
Progress-Werk Oberkirch	Germany	29.3 - Manufacture of parts and accessories for motor vehicles	Internal restructuring	[...] In order to remain competitive, the plant will also focus on ***digitalisation, big data analysis and machine learning*** in the future. […]
Alstom	France	30.20 - Manufacture of railway locomotives and rolling stock	Business expansion	[…] Bombardier wants to increase its capacities by creating new production lines, but also to ***develop digital technology and to install robots*** to reduce the arduousness of the work. […]
Thermo Fisher Scientific	Italy	26.60 - Manufacture of irradiation, electromedical and electrotherapeutic equipment	Business expansion	[...] The site will also house a new 2600 square meters development building and will include a ***flexible, highly automated, line*** for small-scale commercial development and production projects […]

scores that highlight the importance of words (i.e., words that are more interesting, hence frequent) in a text document, but not across all texts in the corpus (i.e., the set of all text information on all restructuring events, in our case). As discussed by Miric et al. (2022), this approach offers the advantage of allowing to understand the impact of each word on the classification performance. In turn, this approach offers higher potential for further optimisation of the model by dropping not relevant parameters which may reduce the model performance. At the same time, the drawback of this method is that it does not enable to capture the context or the word meaning. Hence, since two distinct words with similar meanings will be treated as completely different, it is crucial to properly perform and supervise pre-processing steps (i.e., stemming) in order to reduce similar words to their most common component.

Step 4: We randomly split the validated dataset in training and test sets on a 7:3 ratio base, balancing the two samples by ensuring there was a sufficient number of I4.0-related business restructuring events in both samples, in order to allow the ML algorithms to exploit the best information available in both sets of data to correctly classify events. Furthermore, we balanced the two sets of data exploiting additional information on the type of restructuring event, the country where it took place, the sector of the restructuring firm and the year of restructuring. This is crucial since I4.0-related restructuring events may be more concentrated in specific years, as well as in firms operating in specific countries and sectors.

We trained each NLP classification algorithm using the 70 % of the validated observations in the training data (5033 restructuring events) and tested a range of parameters specific to each model in order to find the optimal values. For each model considered in this study, Table 3 reports the set of parameters used for model optimisation, together with a brief description.

Subsequently, a conventional approach would be to use the remaining 30 % of the observations in the test dataset (2166 restructuring events) to evaluate the performance of each NLP classifier. This approach – known as "*hold-out*" cross-validation – is the simplest and most common technique to define optimal model parameters. However, this technique implies that each model is trained only once for each combination of the tested parameters, therefore potentially leading to a less accurate identification of the optimal parameters and lower model performance. To address this issue, we implemented k-fold cross-validation on the 70 % training dataset by repeating this process 10 times for each classification algorithm and combination of parameters: as shown in Fig. 1, we randomly split the training data into k (in our case, 10) sub-samples and repeated the classification by alternating the sub-testing sample. This approach ensures that the hold-out choice does not influence the identification of the optimal model parameters, hence its performance (Cerulli, 2021). K-fold cross-validation is a well-established and more advanced method of evaluating the model performance and parameter tuning thanks to its robustness compared to the traditional approach of using one hold-out sample (e.g., Bas et al., 2021; Cerulli, 2021; Miric et al., 2022).

The performance evaluation process was carried out in two stages:

Table 3
ML algorithms used, specific parameters, and related description.

ML algorithm	Parameters	Description	STATA command
Decision tree	No. of leaves	Maximum number of leaf nodes the decision tree can have	Cerulli (2022)
Random forest	No. of splitting features	Maximum number of features that are allowed to fit in each individual tree	
	No. of bootstraps	Number of trees built before identifying averages of predictions (higher values make predictions more reliable)	
	Tree depth	Longest path between the root node and the leaf node in each individual tree (smaller values avoid the risk of overfitting)	
Gradient tree boosting	Learning rate	Sets the weight of each individual tree on the final predictions (lower values make the model robust to tree-specific characteristics, i.e. increase generalisability)	
	No. of bootstraps	Number of sequential trees to be modelled (higher values make predictions more reliable, but should be tuned depending on the learning rate)	
	Tree depth	Longest path between the root node and the leaf node in each individual tree (smaller values avoid the risk of overfitting)	
k-Nearest neighbour	No. of neighbours	Number of nearest neighbours (observations featuring the lowest distance)	
Support vector machine	C	Regularisation parameter of the error term, i.e. the degree of correct classification that the algorithm has to meet (large C values imply moor data points used as support vector)	
	Gamma	Sets the curvature of the support vector (higher values correspond to more curvature)	
Regularised multinomial logistic regression	Penalisation	Sets the penalisation of the model by adding a loss function which regularises learned estimates towards zero, i.e. reduce overfitting	
	Elastic parameter	Defines the curvature of loss function (min 0 = Lasso; max 1 = Ridge)	
Support vector machine	C	Regularisation parameter of the error term, i.e. the degree of correct classification that the algorithm has to meet (large C values imply moor data points used as support vector)	Guenther and Schonlau (2016)
	Gamma	Sets the curvature of the support vector (higher values correspond to more curvature)	
	Kernel	Defines the shape of the support vector (i.e., linear, radial base function (RBF), sigmoid, polynomial)	
Random forest	No. of splitting features	Maximum number of features that are allowed to fit in each individual tree	Schonlau and Zou (2020)
	No. of bootstraps	Number of trees built before identifying averages of predictions (higher values make predictions more reliable)	
	Tree depth	Longest path between the root node and the leaf node in each individual tree (smaller values avoid the risk of overfitting)	

firstly, we compared the cross-validation performance of each ML model by evaluating train and test accuracy rates and by identifying the optimal parameter values for each model via "*grid search*". Secondly, we re-estimated each model using the full 70 % training set based on the identified optimal parameter values and tested its performance over the 30 % test set. We then evaluated overall model performance by comparing the predicted and known values of I4.0-related restructuring events for the whole validated sample, computing a set of commonly used metrics including accuracy, precision, recall and F1-score (e.g., Hassan et al., 2022) as defined in Table 4.

Step 5: In the final stage of our methodology, we utilised the estimated NLP classification model with the highest performance to classify automation-related restructuring events beyond the validated dataset used for training and testing. This enabled us to achieve an out-of-sample classification of I4.0-related events for the complete set of 23,618 observations in our original ERM dataset.

4. Results

Below, we first present the preliminary insights from the steps followed to estimate our NLP classification algorithms using text information in ERM data. Subsequently, evaluate and provide commentary on the performance of the various classification methods we tested, focusing on key metrics. Lastly, delve into the patterns observed in I4.0-related business restructuring events.

4.1. Analysis of ML model features

Miric et al. (2022) argue that a common critique of ML techniques is their perception as "*black boxes*" due to researchers' limited ability to interpret and comprehend the factors, parameters (or hyperparameters) and underlying calculations. While our work is not intended to provide or advance theoretical understanding of these methods, our aim is to demonstrate how researchers with limited knowledge of these methods can grasp the fundamental features driving ML model results and understand the extent to which individual features are utilised to classify and predict outcomes.

Table 4
Performance evaluation metrics, formulas, and definitions.

Measure	Formula	Definition
Train Accuracy (cross-validation)	$Accuracy = \frac{TP+TN}{P+N}$	Defines how much accurate the classifier classifies the data (calculated on the cross-validation train sample)
Test Accuracy (cross-validation)	$= (1 - MSE)$	Defines how much accurate the classifier classifies the data (calculated on the cross-validation test sample)
Accuracy (overall)		Defines how much accurate the classifier classifies the data (calculated on the whole validated sample)
Precision (overall)	$Precision = \frac{TP}{TP+FP}$	Defines how exact the model is, i.e. how many positive identified classes are correct (calculated on the whole validated sample)
Recall (overall)	$Recall = \frac{TP}{TP+FN}$	Defines how complete the model is, i.e. how many actual positives are identified correctly (calculated on the whole validated sample)
F1-score (overall)	$F1 - score = \frac{2*(Precision*Recall)}{Precision+Recall}$	It is the harmonic mean of precision and recall (calculated on the whole validated sample)

Incidence and coincidence analysis: The initial step researchers can follow to understand the potential importance of various features in the classification model – akin to variables in traditional regression models – involves analysing the frequency of occurrence of specific features (in our case, represented by BoW derived from the original text data) and examining how their occurrences intersect across different features. This process is conceptually similar to assessing correlations between variables, facilitating the study of their relationships.[8] Specifically, co-occurrence analysis and its closely related yet more fundamental counterpart, coincidence analysis, are foundational applications of many unsupervised learning techniques (Escobar, 2015). Given the aim of our preliminary analysis, we delve into incidence and coincidence analyses, where we consider as features the set of BoW matching the keywords identified from Martinelli et al. (2021), Cockburn et al. (2019) and Castellani et al. (2022) and discussed in Section 3.2.

Fig. 2 illustrates the incidence (%) of each of the I4.0-matching BoW, ordered from the most to the least frequent. Most matched BoW exhibit relatively low incidence rates, occurring for up to about 2 % of the overall validation sample. Among these, we observe features matching with various keywords capturing specific I4.0 technologies, such as AI (e.g., machine learning, big data analytics), robotics (e.g., autonomous, advanced, industrial), additive manufacturing/3D printing, internet of things (e.g., automation technologies, communication networks, sensors), and cybersecurity. Overall, the incidence of these BoW is consistent with the rate of business restructuring events identified and validated as automation related. Additionally, we identify two sub-groups of features that appear more frequently in our text data. The first sub-group exhibits incidence rates ranging between 2 % and about 7 %, encompassing relatively more common (i.e., general and potentially less specific) features related to either specific technologies (i.e., "addit") or more general terms included in our validation criteria beyond specific keywords found in the literature (i.e., "technolog", "system", "40"). The second sub-group displays significantly higher incidence rates, ranging from above 15 % to about 44 %, and includes features embedding more general terms used in combination with other, more specific terminology (i.e., "manufactur", "new", "factory", "industry"). Although these features may indeed appear in textual information regarding I4.0-related restructuring events, their exceedingly high incidence rates suggest that they are also very common terms throughout the entire corpus of our text data. This prevalence could potentially undermine the performance of our classification models by increasing the number of false positive events classified as I4.0-related. Hence, this insight validates our decision not to directly employ BoW as model features in the NLP classifiers, opting instead to utilize TF-IDF, which naturally attenuates scores for more frequently occurring features.

To provide further insight, Fig. 3 illustrates conditional coincidence plots for the four most frequent features identified in Fig. 2. These plots depict the actual coincidence (%) of each feature with each of the four most frequent BoW (indicated by the dark blue bars) and juxtapose them with the probability of their occurrence given the presence of the more frequent BoW (Escobar, 2015). Consequently, a green bar on the left of the blue one indicates that the feature is highly coincidental with the more frequent one (as its expected incidence value was lower in the case of independence of between features). Conversely, a red bar on the right of the blue one means that the feature is less likely to co-occur with the more frequent one. Overall, the four plots reported in Fig. 3 emphasise that most of the least recurring BoW are highly coincidental with the four most recurring BoW, with a few exceptions, such as the combination of "manufactur + technolog", "new + print", "factory + technolog", and "industry + print". These exceptions are consistently identified as not aligning with any combination of predefined keywords or specific terminology and thus, they lack relevance for our classification objectives.

Hierarchical clustering: Researchers might also wish to investigate the degree of similarity or dissimilarity between various features within the model. Typically, this involves employing hierarchical clustering algorithms, which group features together or separate them based on the distance measure between sets of observations. The outcomes of these analyses are often presented using dendrograms.[9] Dendrograms serve as practical tool to understand the distance and similarity between features: similar BoW are group together in the same cluster, while the length of the branches indicates the distance between clusters. In Fig. 4, we present four different dendrogram, all generated using the hierarchical clustering algorithm proposed by Ward (1963). This method employs an agglomerative hierarchical clustering procedure, where the criterion for merging pairs of features (and clusters, in the following iterations) is based on minimising the variance within each cluster. The resulting dendrograms are based on four frequently used measures of distance in the literature (e.g., Choi et al., 2010; Finch, 2005): Jaccard, Russel, Sneath and Rogers. All four dendrograms distinctly illustrate the presence of two main clusters of features: one comprising the more general and frequently occurring BoW highlighted in Figs. 2 and 3, and the other containing the remaining, less frequent, BoW. This distinction is particularly evident when using the Jaccard, Russell, and Rogers measures of distance. In these cases, the red vertical lines, indicating the cut-off point defining a two-cluster division, fall at relatively high distance values (1.38, 1.12 and 1.12, respectively). In contrast, the dendrogram based on Sneath distance reveals additional heterogeneity among the four most occurring BoW (i.e., "factory" and "industry" vs "manufactur" and "new"), with the resulting cut-off separating these four features from all others resting at 0.12 (the green vertical line), thereby identifying five distinct clusters. Importantly, all less recurring BoW matching automation-related keywords used in the data validation process show high similarity, suggesting their mutual relatedness.

All these insights offer a promising foundation for assessing the relevance of the examined I4.0-related features in the classification algorithms we estimate in the following section. They may prove useful in guiding researchers through the initial steps necessary to comprehend text data and its utilisation in ML methods.

4.2. Model optimisation and performance evaluation

Our objective is to classify business restructuring events as automation-related or not, based on their textual descriptions. Therefore, our target dependent variable is represented as a binary dummy variable indicating whether each event is automation-related (1 or 2, depending on the command requirements) or not (0 or 1). To manage computational complexity and streamline the implementation of NLP classification algorithms across various machine learning techniques tested, we opt to utilize only the TF-IDF representations of BoW matching the I4.0-related keywords analysed in the preceding section. Consequently, all our models include forty-four features. It's worth noting that a more comprehensive model could potentially incorporate a larger set of features, such as the TF-IDF of all features with incidence rates above a meaningful threshold.

Optimal parameter identification: We tested a wide range of values for each model-specific tuning parameter, as described in Table 5. This resulted in a large number of combinations tested to find the optimal values, depending on the number of parameters available in each model, ranging from less than 50 in the case of DT to more than 3000 in the case of GTB.

Figs. 5, 6 and 7 present (mean) train and (mean) test accuracy as a function of each combination of parameters tested for algorithms

[8] All incidence, coincidence and related analyses were implemented in STATA using `coin` command (Escobar, 2015).

[9] We do not aim at digging into the technicalities of such techniques, but we invite the reader to explore the relevant literature (e.g., Nielsen, 2016).

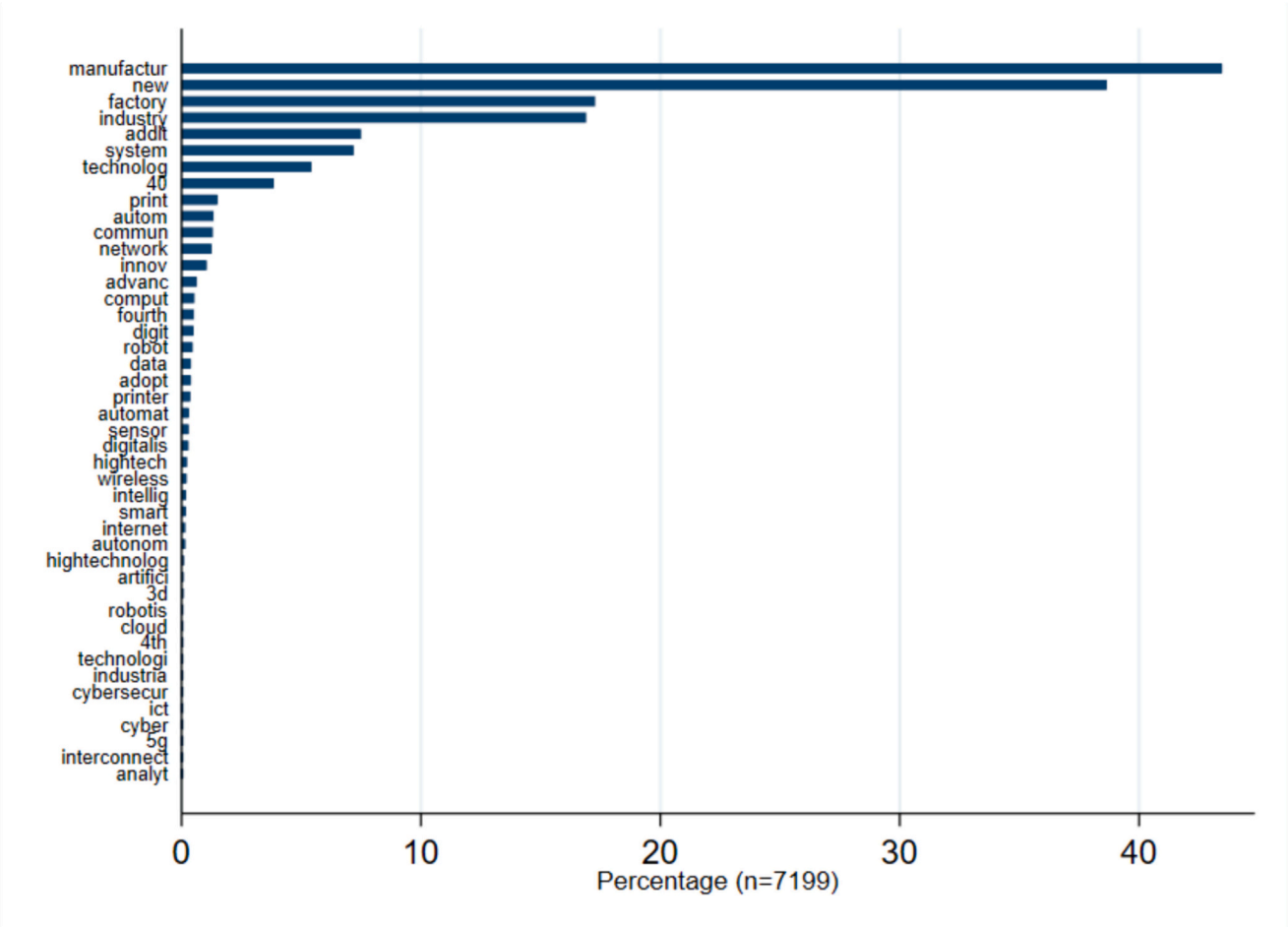

Fig. 2. Plot of incidence for BoW matching I4.0-related keywords.
Notes: Authors' own elaborations based on ERM data.

implemented using Cerulli (2022), Guenther and Schonlau (2016), and Schonlau and Zou (2020), respectively. For each algorithm, the optimal combination of parameters is that maximising (mean) test accuracy, obtained from a 10-fold cross-validation. For example, from the first graph in Fig. 5 presenting cross-validation results for the DT classification algorithm we clearly observe that the optimal number of leaves in the classification tree is 2, occurring at the combination indexed as 1 where the (mean) test accuracy is maximised. We can also observe that increasing the number of leaves results in a monotonic increase in the (mean) train accuracy, i.e. overfitting. Similarly, we can observe the overfitting appearing in Fig. 7 – looking at cross-validation results from the RF classification algorithm implemented through Schonlau and Zou's (2020) command – as the index of the tested combinations of parameters surpasses 400.

Comparison of performance: As highlighted by Cerulli (2021), the mean test accuracy computed via cross-validation serves as a reliable proxy for the out-of-sample predictive power of the chose ML technique. Nevertheless, as discussed in Section 3.2, to evaluate the performance of each model, we additionally computed the overall accuracy and other performance metrics using the predicted values of the NLP classification algorithms on the held out 30 % test dataset from the originally validated sample (as shown in Fig. 1). These predictions were based on the optimal parameters identified during the cross-validation process.

In Table 6, we provide a summary of the performance of the various NLP classification approaches we tested to identify I4.0-related business restructuring events. Upon reviewing the cross-validation outcomes (columns 1 to 3), we observe that all ML algorithms perform well in terms of both train and test accuracy. Particularly noteworthy is that algorithms implemented using Cerulli's (2022) STATA command all exhibit quite comparable performance. The SVM classifier results as the top-performing algorithm among the six tested using the Python-based implementation, achieving the higher train accuracy (0.9880), the second-best test accuracy (0.9865), and the lowest standard deviation (SD) of test accuracy (0.0006). Looking at the results obtained from the SVM classification algorithm implemented using Guenther and Schonlau's (2016) command, we observe that testing different kernel types has minimal impact on the performance of the algorithm (at least in our specific case). While the train and test accuracy measures align with those derived from the SVM model implemented using Cerulli's (2022) command, the SDs of the test accuracy are notably larger in the case of Guenther and Schonlau's (2016) implementation, indicating an overall lower test accuracy. This discrepancy is particularly pronounced in the case of the SVM classifier utilising the polynomial kernel type, suggesting that setting a more complex shape for the support vector does not perform well with the tested data. Lastly, the cross-validation results for the RF algorithm implemented using Schonlau and Zou's (2020) command show the best performance among the algorithms and commands tested. Specifically, it exhibits the lowest SD of test accuracy, implying the highest precision overall.

Focusing on the final evaluation results (columns 4 to 7 in Table 6), the RF classification algorithm based on Schonlau and Zou's (2020) command clearly emerges as the top performer. While all ML models achieve high accuracy and good precision rates, a comparison of their performance across recall and F1-score metrics reveals significant variations. Specifically, most other ML techniques fall short in fully identifying restructuring events manually validated as automation-related, as

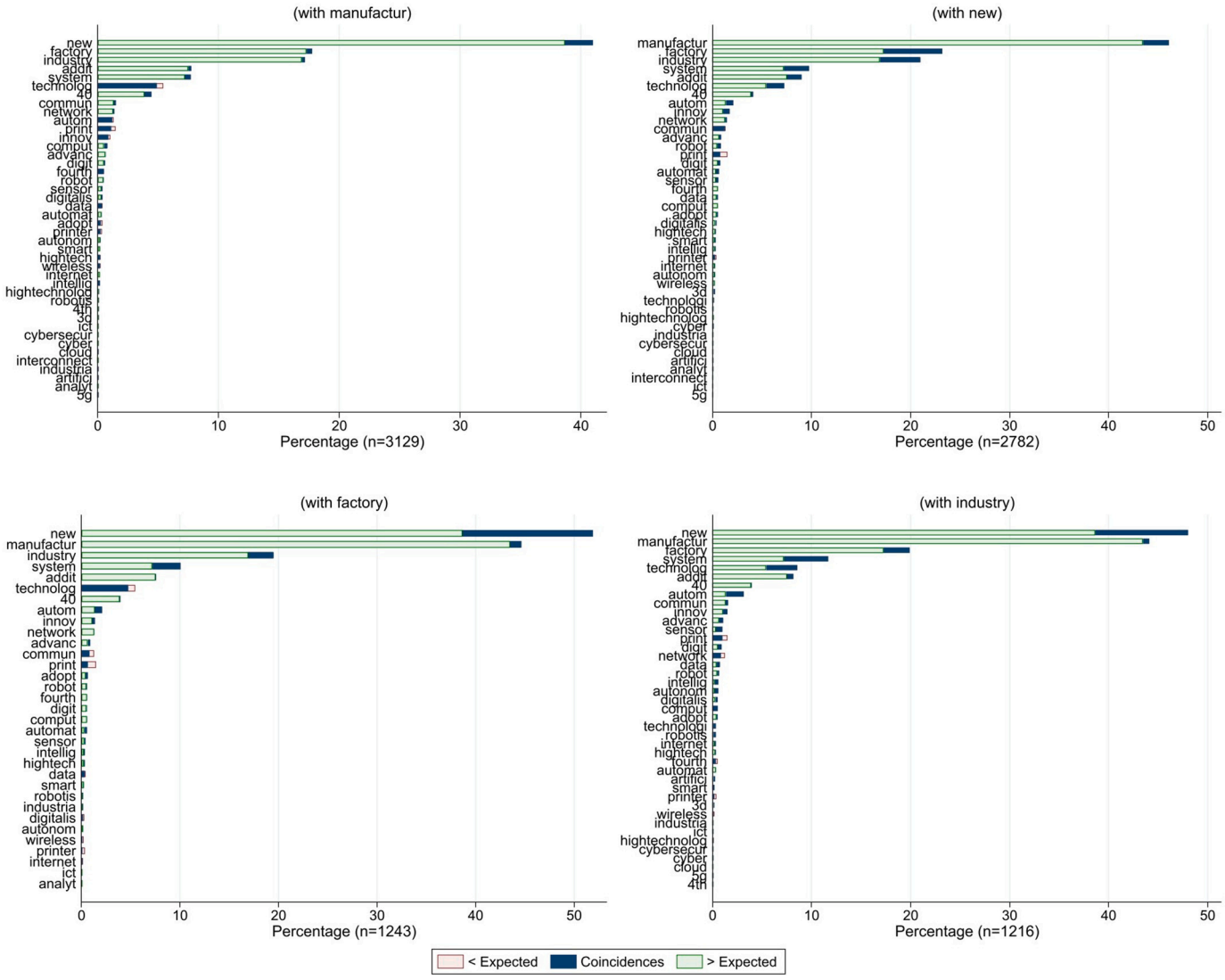

Fig. 3. Plot of conditional coincidence for BoW matching I4.0-related keywords with highest incidence rates.
Notes: Authors' own elaborations based on ERM data.

evidenced by their low recall metrics, which range between 0.0288 and 0.2885. in Contrast, Schonlau and Zou's (2020) implementation of the RF algorithm yields satisfying values for both recall and F1-score (0.7596 and 0.8272, respectively).

It is important to recognise that in presenting this comparison, we do not aim to criticise neither the underlying ML models nor their implementation using the STATA commands discussed above. Our results are contingent upon the specific text data used in this analysis, on the subset of features incorporated in our models, and on the data preparation process undertaken. Therefore, we emphasise that employing the same commands on a different set of features (for instance, to classify restructuring events as involving the hiring/displacement of blue or white collars) or on an entirely different text dataset may yield divergent outcomes regarding the optimal NLP classification algorithm. As such, we encourage researchers interested in applying such techniques to explore various ML algorithms and implementations thereof. This approach allows for the identification of the most suitable option aligned with the specific research objectives and contextual factors at play.

5. Analysis of model predictions

5.1. Evaluation of importance scores

Having compared different ML methods and identified the best-performing algorithm, we can now delve into another aspect of ML techniques, shedding light on what occurs within the often-cited "*black box*". In addition to exploring the characteristics of different features of the model as discussed in Section 4.1, we can analyse how these features impact the model's outcome computations. Many ML algorithms, particularly those leveraging on BoW or TF-IDF features, often offer measures explaining the extent to which each individual feature influence outcome predictions, known as "*importance scores*". These scores are conceptually akin to the coefficients of variables estimated in linear regression models (Miric et al., 2022).

In Fig. 8, we present the top 25 features (TF-IDF) ranked by importance score for classifying I4.0-related business restructuring events using the RF classification algorithm by Schonlau and Zou (2020). The TF-IDF representations of the features are arranged in descending order, providing insight into the relative explanatory power of each feature. Features embedding more general BoW like "analyt", "commun", "ict" or "automat" ranking highest, followed by BoW referring to AI,

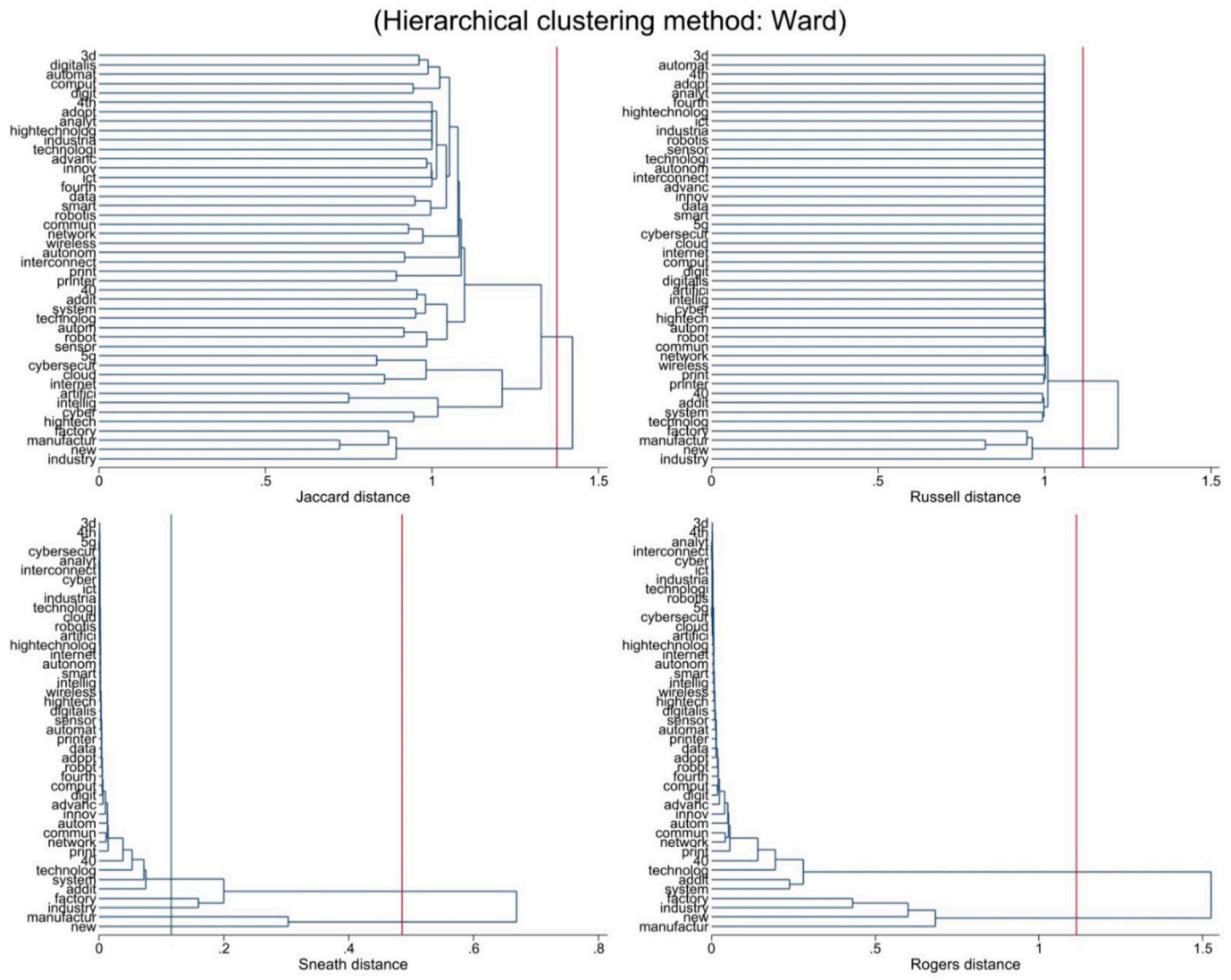

Fig. 4. Dendrograms of BoW matching I4.0-related keywords using hierarchical clustering.
Notes: Authors' own elaborations based on ERM data.

cybersecurity, and advanced automation. Meanwhile, features directly referring to Industry 4.0 or advanced manufacturing technologies like robots, additive manufacturing and internet of things (as defined by Castellani et al., 2022) exhibit lower, although still notable, importance scores. Notably, the TF-IDF representation of the BoW "analyt" emerges as significantly more crucial in terms of explanatory power within the RF classifier, showing at least double the explanatory power of all other model features. This finding is quite surprising considering that, as depicted in Fig. 2, the BoW "analyt" demonstrates the lowest percentage incidence in text descriptions across all restructuring events. However, this result underscores that despite its very limited presence the feature is tightly linked to the (validated) classification of an event as automation-related. This is indeed much higher than the importance scores computed for features like "manufactur" or "new", despite being more prevalent even in non-automation-related restructuring events. Similarly, the predictive power of AI-related features is found to be quite substantial as compared to their incidence across textual descriptions.

While we acknowledge that ML techniques are not inherently straightforward in terms of understanding the elements (i.e., features) upon which calculations are grounded and how they operate, the insights gleaned from this section (as well as those elucidated in previous one) should equip researchers interested in applying these techniques with a diverse array of tools, enabling them to experiment with ML-based methodologies using STATA software.

5.2. *Exploring automation-related restructuring events*

As a next phase in our explorative analysis of business restructuring events, we used the best-performing RF classification algorithm just discussed to categorize events outside the validated sample as either related to I4.0 technologies or not. This process yielded 371 restructuring events out of the total sample of 23,618 events being classified as automation-related (about 1.6 %). To provide deeper insights into the nature of these restructuring events, Tables 7, 8, and 9 showcase their distribution — as determined by the RF algorithm — over time, across various countries, and within different sectors, respectively.

Examining the yearly distribution of events in the ERM database (Table 7), we observe that automation-related restructuring events are primarily concentrated in the years just before the 2008 global financial crisis and from 2016 onward. Conversely, these events are less frequent during the years between these two periods, especially in relative terms, considering that the overall yearly number of restructuring events peaks in 2009 and then remains high in subsequent years. Notably, the second surge in automation-related events from 2016 onward occurs both in

Table 5
ML algorithms, specific parameters, and range of value tested.

ML algorithm	Parameters	Value tested
Decision tree	No. of leaves	All values from 1 to 44 (= N of features)
Random forest	No. of splitting features	All values from 1 to 44 (= N of features)
	No. of bootstraps	Values from 5 to 100, incrementing by 5 (e.g., 5, 10, 15, etc.)
	Tree depth	1 2 3
Gradient tree boosting	Learning rate	0.0001 0.00025 0.0005 0.00075 0.001 0.0025 0.005 0.0075 0.01 0.025 0.05 0.075 0.1 0.25 0.5 0.75
	No. of bootstraps	Values from 5 to 100, incrementing by 5 (e.g., 5, 10, 15, etc.)
	Tree depth	All values from 1 to 10
k-Nearest neighbour	No. of neighbours	All values from 1 to 30
Support vector machine	C	0.001 0.005 0.01 0.05 0.1 0.5 1 2 4 8 16 32 64 128 256 512 1024 2048 4096 8192
	Gamma	0.0001 0.001 0.01 0.1 1 10 100 1000
Regularised multinomial logistic regression	Penalisation	0.5 0.6 0.7 0.8 0.9 1 2 4 8 16 32 64 128 256 512 1024 2048
	Elastic parameter	Values from 0 to 1, incrementing by 0.05 (e.g., 0, 0.05, 0.1, etc.)
Support vector machine	C	0.001 0.005 0.01 0.05 0.1 0.5 1 2 4 8 16 32 64 128 256 512 1024 2048 4096 8192
	Gamma	0.0001 0.001 0.01 0.1 1 10 100 1000
	Kernel	Radial basis function (RBF), linear, sigmoid, polynomial
Random forest	No. of splitting features	All values from 1 to 44 (= N of features)
	No. of bootstraps	Values from 5 to 100, incrementing by 5 (e.g., 5, 10, 15, etc.)
	Tree depth	Unlimited (i.e., different trees might have different length, with no specified max depth)

absolute and relative terms, with these events accounting for approximately 5.8 % of all restructuring events in 2021, coinciding with the second lockdown phase of the Covid-19 pandemic in most, if not all, European countries. Furthermore, this increase in the share of automation-related restructuring events reflects the growing adoption of I4.0 and associated new digital technologies across European economies (Castellani et al., 2022), spurred by the implementation of I4.0-related policy initiatives (Mariani and Borghi, 2019; Teixeira and Tavares-Lehmann, 2022).

Turning to the cross-country distribution of business restructuring events in Table 8, we observe that they are spread across most European countries, albeit unevenly. Specifically, we witness a consistent share of I4.0-related events located across Western-European (and to a lesser extent, Northern-European) countries, accounting for around 2 % of total events per country (notably, France, Germany, Ireland, followed by Italy, Spain and the UK). This finding is unsurprising given that these economies are more advanced in terms of both technology adoption in manufacturing and the proportion of service activities. Moreover, these results align with previous research on the geographic distribution of adopters (Castellani et al., 2022) and developers (Balland and Boschma, 2021; Ciffolilli and Muscio, 2018; Corradini et al., 2021) of I4.0 technologies. Notwithstanding, we also note that certain Eastern-European countries stand out for their high number of automation-related restructuring events: Poland, Hungary, Romania, Czech Republic, and Slovakia show shares comparable to those found in more advanced Western-European economies. As discussed by Castellani et al. (2022) this is the result of substantial investments in advanced manufacturing technologies, resulting in increased adoption of I4.0 technologies.

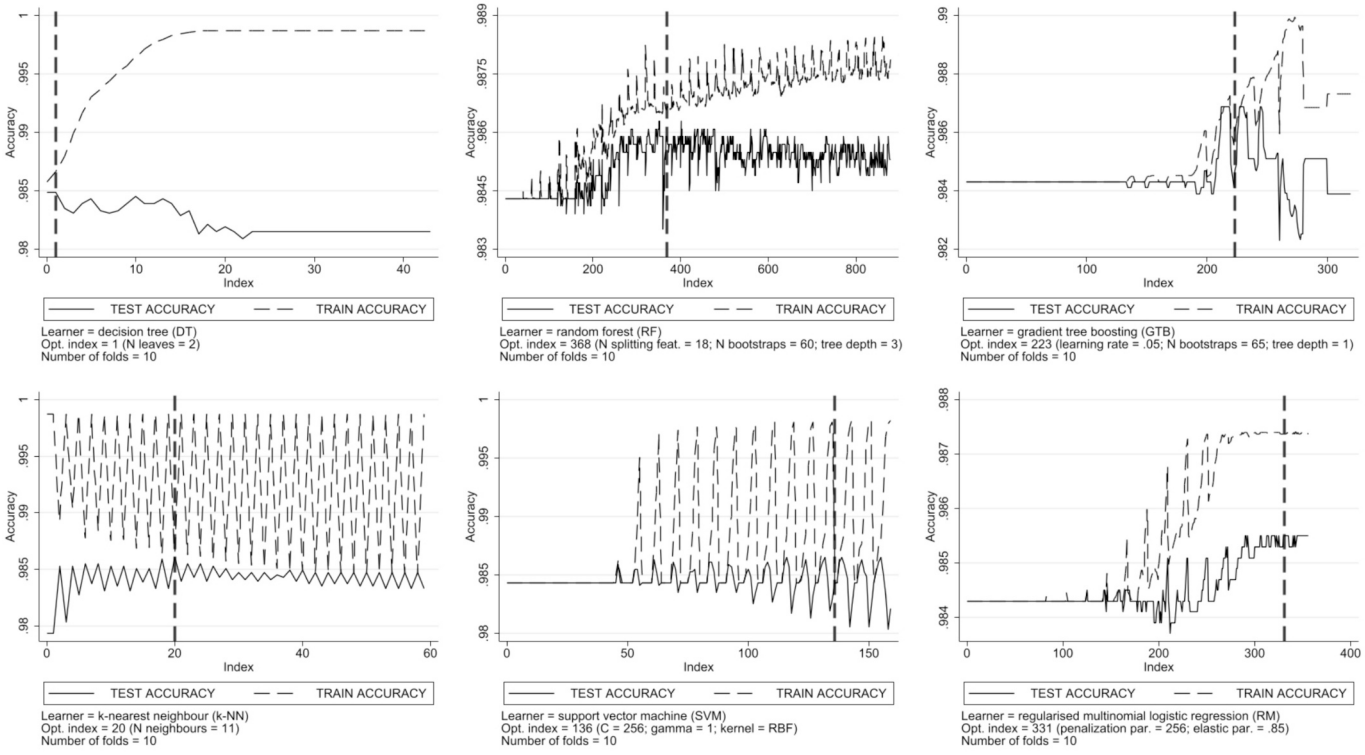

Fig. 5. Test and training accuracy as a function of tuning parameters for different NLP classification algorithms supported by Cerulli's (2022) STATA command.
Notes: Authors' own elaborations based on ERM data. For graphical representation reasons, graphs for RF and GTB algorithms only reports test and train accuracy and the optimal parameter combination holding "tree depth" at its optimal value.

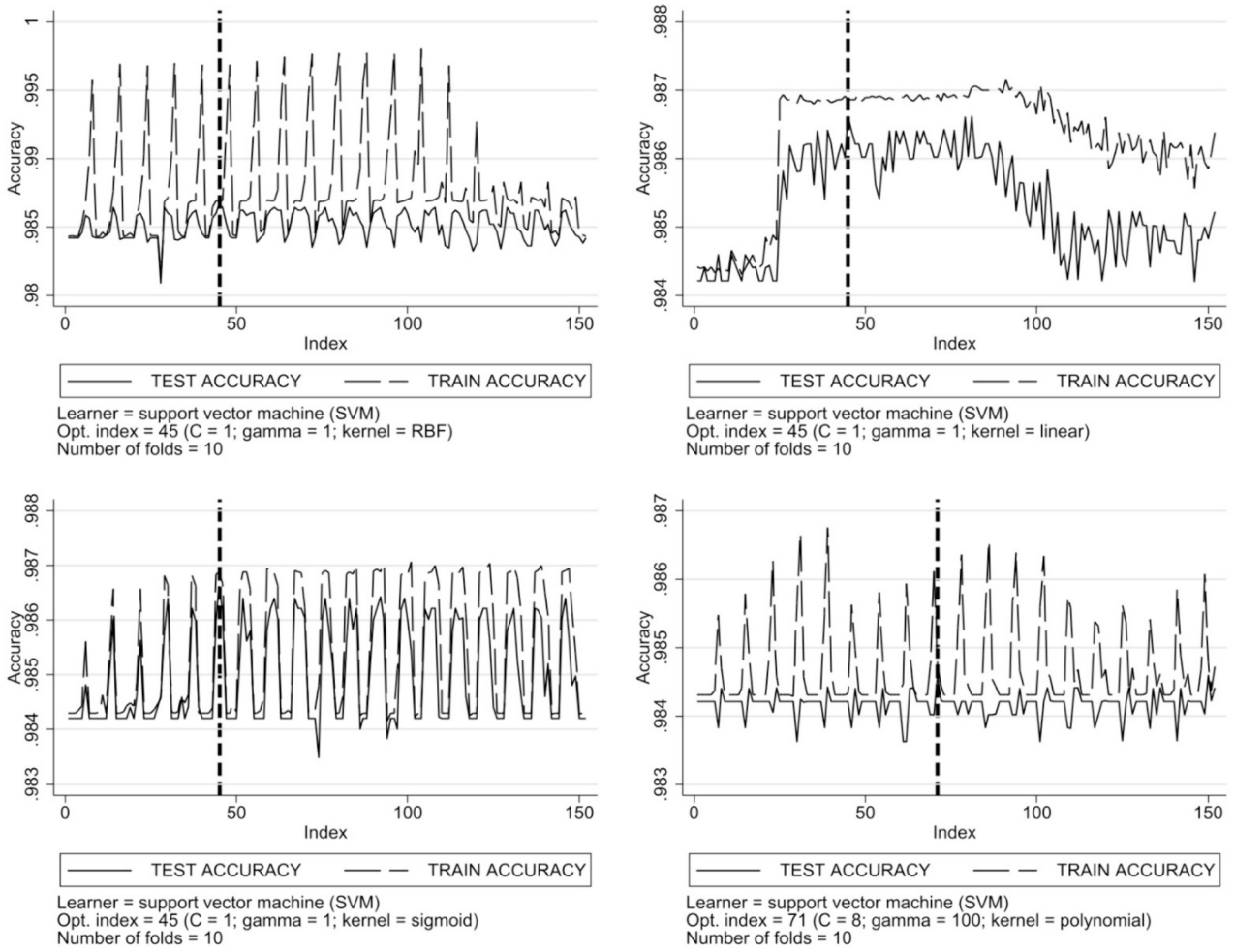

Fig. 6. Test and training accuracy as a function of tuning parameters for SVM classification algorithm supported by Guenther and Schonlau's (2016) STATA command.
Notes: Authors' own elaborations based on ERM data.

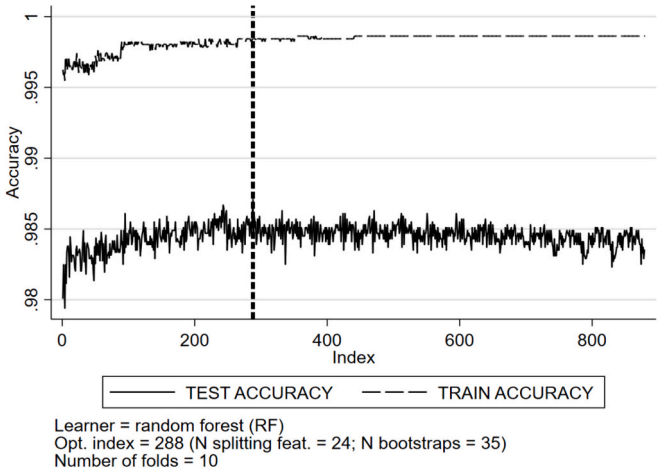

Fig. 7. Test and training accuracy as a function of tuning parameters for RF classification algorithm supported by Schonlau and Zou's (2020) STATA command.
Notes: Authors' own elaborations based on ERM data.

Beyond mere technology adoption, on the one hand this phenomenon should be viewed in the context of effective industrial policies aimed at technological upgrading and leapfrogging in these economies (e.g., Ministry of Industry and Trade of the Czech Republic, 2019). On the other hand, Eastern- and Western-European countries have strengthened their business and industrial ties over the past decade through participation in global value chain. Specifically, the accession to the European Single Market as a result of the 2004 enlargement of the EU (Cséfalvay, 2020), which largely benefitted Eastern-European countries, combined with the availability of relatively skilled workers, low labour costs, as well as a stable political and institutional setting (Carstensen and Toubal, 2004), resulted in rising foreign direct investments (FDI) from Western-European multinationals towards these countries (ECB, 2013, 2020). This has positioned them as preferred locations compared to Asian countries (Pavlínek, 2018), owing to their geographical proximity and desirability.

Looking at the sectoral distribution of events in Table 9, we witness that the majority of I4.0-related events are concentrated in manufacturing industries (132), although they do not account for the larger sectoral share (1.2 %). Interestingly, machine learning-classified restructuring events related to automation technologies are predominantly found in firms operating within the information and communication industries (3.6 %), financial and insurance activities (3 %), and professional, scientific, and technical services (4.3 %). These findings align with our results regarding the relative importance of different features in our NLP classification model: while features capturing terms associated with advanced manufacturing technologies had limited explanatory power, those related to artificial intelligence and services

Table 6
Comparison of performance using different NLP classification models and STATA commands.

	Cross-validation results			Final evaluation results			
ML algorithm	(Mean) train accuracy	(Mean) test accuracy	(SD) test accuracy	(Overall) accuracy	(Overall) precision	(Overall) recall	(Overall) F1-score
Decision tree	0.9866	0.9849	0.0022	0.9869	0.8125	0.1250	0.2167
Random forest	0.9868	0.9863	0.0024	0.9869	0.7778	0.1346	0.2295
Gradient tree boosting	0.9870	0.9869	0.0051	0.9871	0.7037	0.1827	0.2901
k-Nearest neighbour	0.9865	0.9861	0.0020	0.9871	0.7619	0.1538	0.2560
Support vector machine (SVM)	0.9880	0.9865	0.0006	0.9882	0.7714	0.2596	0.3885
Regularised multinomial logistic regression	0.9874	0.9855	0.0022	0.9872	0.6250	0.2885	0.3947
SVM (RBF kernel)	0.9869	0.9866	0.0047	0.9872	0.7500	0.1731	0.2813
SVM (Linear kernel)	0.9869	0.9866	0.0047	0.9872	0.7500	0.1731	0.2813
SVM (Sigmoid kernel)	0.9869	0.9866	0.0047	0.9872	0.7500	0.1731	0.2813
SVM (Polynomial kernel)	0.9847	0.9846	0.0074	0.9856	0.5000	0.0288	0.0545
Random forest	0.9984	0.9871	0.0001	0.9954	0.9080	0.7596	0.8272

Notes: Authors' own elaboration based on ERM data. A perfect classification model would yield value 1 for all accuracy measures, value 0 for SD of test accuracy, and value 1 for precision, recall, F1-score measures.

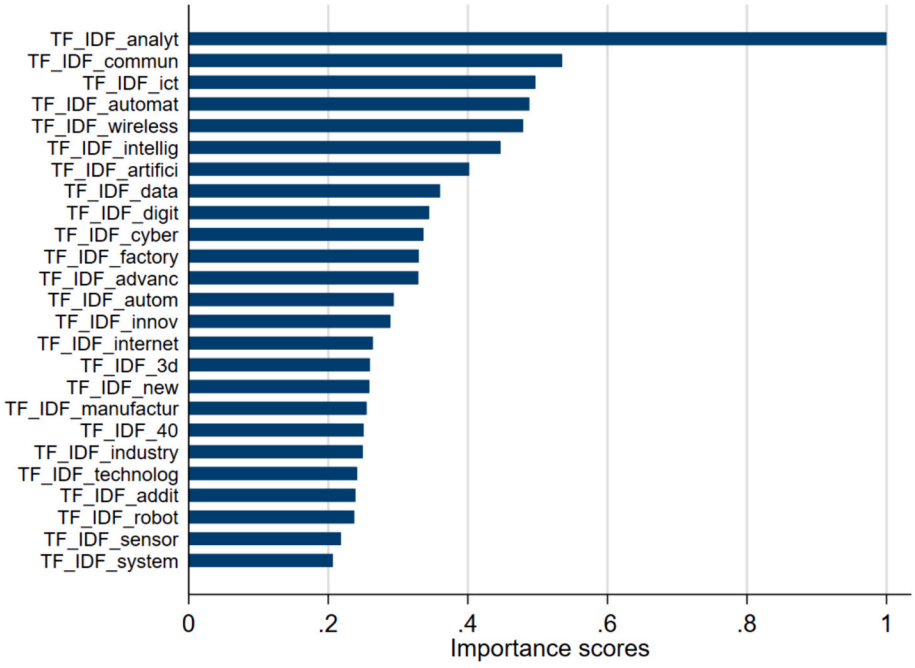

Fig. 8. Importance scores for top 25 features in best-performing NLP classification algorithm (RF).
Notes: Authors' own elaborations based on ERM data.

(such as "analyt", "data" and "cyber") were found to be way more relevant, having larger predictive power. Furthermore, our findings regarding the sectoral pattern of I4.0-related business restructuring events corroborate existing evidence indicating that AI-intensive industries are also knowledge-intensive sectors rather than more high-tech manufacturing industries (Igna and Venturini, 2023; Martinelli et al., 2021).[10]

Finally, we delve into the relative distribution of I4.0-related restructuring events, distinguishing between types of restructuring events. To accomplish this, we leverage a key piece of information provided by the ERM data: the number of jobs lost and/or gained as a result of these events. Specifically, we distinguish between job-creating – as identified by business expansion events in the ERM database – and job-displacing events – identified by all other types of restructuring events (i.e., bankruptcy/closure, internal restructuring, merger/acquisition, offshoring/delocalisation, outsourcing/relocation). We acknowledge that this is categorisation may oversimplify the job-related implications of restructuring events, since many of them may imply both the hiring and the displacement of workers, therefore we computed the net job loss/gain associated with each event.

Table 10 presents the frequency and relative share of restructuring events, distinguishing between I4.0-related and not, and between job-creating and job-displacing events. Additionally, we provide the means and SD of net job loss/gain by type of event: job-creating events constitute approximately 43 % of all events in the ERM database, of which about 2.5 % are relates to automation technologies. However, automation-related job-creating events feature lower mean number of jobs gained (about 187 per event, as compared to about 221), with the difference being statistically significant at the 1 % level as highlighted

[10] Indeed, the sectoral distribution of our I.40-related restructuring events as predicted by our RF classifier largely coincides with that of AI-related patents presented in Fig. 1 of Igna and Venturini (2023, pp. 7).

Table 7
Yearly distribution of business restructuring events.

	Not I4.0-related		I4.0-related	
Year	Frequency	%	Frequency	%
2002	42	100.00	0	0.00
2003	82	100.00	0	0.00
2004	171	100.00	0	0.00
2005	294	99.66	1	0.34
2006	1295	99.16	11	0.84
2007	1390	99.00	14	1.00
2008	1563	99.24	12	0.76
2009	1997	99.75	5	0.25
2010	1220	99.35	8	0.65
2011	1329	99.63	5	0.37
2012	1511	99.47	8	0.53
2013	1407	99.43	8	0.57
2014	1481	99.46	8	0.54
2015	1251	99.36	8	0.64
2016	1430	98.48	22	1.52
2017	1456	96.49	53	3.51
2018	1221	97.14	36	2.86
2019	1092	95.87	47	4.13
2020	1369	96.75	46	3.25
2021	834	94.24	51	5.76
2022	794	96.59	28	3.41
2023	18	100.00	0	0.00
Total	23,247	98.43	371	1.57

Notes: Authors' own elaboration based on ERM data.

by the *p*-value of the two-sample *t*-test (0.001). Conversely, I4.0-related events represent a minor relative share of all job-displacing events (only 0.86 %), which feature a lower mean number of jobs lost as compared to not I4.0-related events (about 425 vs around 459). However, this difference is not statistically significant (the two-sample t-test *p*-value is 0.616).

Table 8
Cross-country distribution of business restructuring events.

	Not I4.0-related		I4.0-related	
Country	Frequency	%	Frequency	%
Austria	434	99.31	3	0.69
Belgium	665	99.40	4	0.60
Bulgaria	196	97.51	5	2.49
Croatia	174	98.31	3	1.69
Cyprus	31	100.00	0	0.00
Czech Republic	992	98.32	17	1.68
Denmark	344	99.14	3	0.86
Estonia	194	98.98	2	1.02
Finland	603	98.53	9	1.47
France	2813	97.91	60	2.09
Germany	1843	98.03	37	1.97
Greece	210	99.53	1	0.47
Hungary	647	97.44	17	2.56
Ireland	993	98.03	20	1.97
Italy	939	98.63	13	1.37
Latvia	91	97.85	2	2.15
Lithuania	443	98.01	9	1.99
Luxembourg	40	90.91	4	9.09
Malta	55	96.49	2	3.51
Netherlands	674	98.54	10	1.46
Norway	289	98.30	5	1.70
Poland	2824	98.53	42	1.47
Portugal	435	98.19	8	1.81
Romania	1107	98.23	20	1.77
Slovakia	554	98.05	11	1.95
Slovenia	417	100.00	0	0.00
Spain	917	98.39	15	1.61
Sweden	886	98.55	13	1.45
United Kingdom	3437	98.96	36	1.04
Total	23,247	98.43	371	1.57

Notes: Authors' own elaboration based on ERM data.

6. Conclusion

6.1. Summary and contribution

Most research fields in social sciences are currently experiencing an evolution driven by the increasing size, variety, and complexity of available data, as well as by a growing consensus on the necessity of expanding the range of methodological tools employed. With more researchers utilising big data, including unstructured textual information, new ML-based methods such as NLP classifiers and TM techniques are gaining traction in economic-related disciplines spanning from international business to economic geography, management, and innovation studies. However, in many cases the will to adopt these new methodologies is not matched by researchers' familiarity with the tools and specific skills required for their implementation.

This study addresses the recent call made by academics in economics, business, and management field (Chen and Schintler, 2023; Lindner et al., 2022) to explore the use of ML methods to tackle the new challenges posed by the ever-expanding volume of unstructured information (e.g., Agarwal and Dhar, 2014; George et al., 2016). Furthermore, ML techniques hold the potential to deliver more robust and generalizable empirical results in economics and related domains by drawing on techniques, variables, and theoretical constructs from adjacent fields (Lindner et al., 2022; Shrestha et al., 2021). From this perspective, we contribute to the literature by leveraging the new opportunities of data science techniques and addressing the call for collecting new data on the adoption of I4.0 technologies at the firm level – still scarce, as of today – in a reliable and scalable way. In turn, this data could prove useful in providing further evidence on the economic and social implications coming from automation technologies, as well as defining new research directions.

Our contribution also holds practical significance. We offer a step-by-step guide to applying a suite of ML techniques for classification purposes, data preparation, and explorative data analysis, utilising a set of recently developed commands for STATA. Thus, our approach does not require extensive prior knowledge of the theoretical underpinnings of these methodologies or advanced programming and data science expertise. Our empirical test is based on textual information regarding business restructuring events sourced from Eurofound's ERM database. Specifically, we demonstrate the application of ML techniques to classify restructuring events occurring across European countries as I4.0-related or not, while also exploring the features driving such categorisation, their relative importance, as well as the distribution of classified events over time and across countries, sectors, and type of restructuring.

Our findings contribute to a deeper understanding of the defining characteristics of I4.0-related restructuring events, as discerned from their descriptions. Specifically, we observe clear patterns of incidence and coincidence between specific terms indicative of technology adoption by restructuring firms, alongside a notable differentiation in terms of similarity/distance between groups thereof. Analysing these patterns can serve as a preliminary ground to grasp the potential relevance of specific terminologies (i.e., features) in facilitating accurate predictions from ML algorithms, thus informing researchers' choices on model set-up and optimisation decisions. However, we underscore that only the ultimate analysis of the features' importance scores, resulting from the best-performing model, can provide an precise assessment of the relative importance each specific term truly has in influencing the outcome of the classification model.

The descriptive analysis of the predictions generated by our best-performing model highlights patterns of I4.0 technology adoption that align closely with prior evidence documented in the literature. These patterns are evident not only in terms of adoption rates but also in their distribution across countries, sectors, and over time. Remarkably, these patterns are coherent with trends in the organisation of value chain activities across European countries, largely driven by Western-European multinationals. Moreover, our analysis reveals a strong

Table 9
Sectoral distribution of business restructuring events.

	Not I4.0-related		I4.0-related	
Sector	Frequency	%	Frequency	%
A - Agriculture, forestry and fishing (01−03)	69	100.00	0	0.00
B - Mining and quarrying (05–09)	335	99.70	1	0.30
C - Manufacturing (10−33)	10,786	98.79	132	1.21
D - Electricity, gas, steam and air conditioning supply (35)	339	99.12	3	0.88
E - Water supply; sewerage, waste management and remediation activities (36–39)	73	100.00	0	0.00
F - Construction (41–43)	559	100.00	0	0.00
G - Wholesale and retail trade; repair of motor vehicles and motorcycles (45–47)	2294	98.54	34	1.46
H - Transportation and storage (49–53)	1719	98.91	19	1.09
I - Accommodation and food service activities (55–56)	488	100.00	0	0.00
J - Information and communication (58–63)	2283	96.45	84	3.55
K - Financial and insurance activities (64–66)	1500	97.02	46	2.98
L - Real estate activities (68)	65	97.01	2	2.99
M - Professional, scientific and technical activities (69–75)	651	95.74	29	4.26
N - Administrative and support service activities (77–82)	888	98.56	13	1.44
O - Public administration and defence; compulsory social security (84)	585	99.49	3	0.51
P - Education (85)	111	99.11	1	0.89
Q - Human health and social work activities (86–88)	330	99.70	1	0.30
R - Arts, entertainment and recreation (90–93)	124	98.41	2	1.59
S_U - Other service activities (94–99)	48	97.96	1	2.04
Total	23,247	98.43	371	1.57

Notes: Authors' own elaboration based on ERM data.

correlation with the sectoral intensity of innovation activities in automation technologies, indicating consistent adoption rates not only in manufacturing, but also in knowledge-intensive and creative industries.

An additional, interesting finding is that the predicted I4.0 adoption predominantly characterises job-creating restructuring events, such as business expansions, rather than the job-displacing outcomes typically emphasized in the automation literature. This result prompts new questions, such as: What are the critical factors driving firms' technological choices when expanding their activities? How do productivity-enhancing I4.0 technologies contribute to firms' competitiveness during periods of business growth? What are the implications of integrating I4.0 technologies into expansion strategies for workforce dynamics and organizational structure? Consequently, further research is necessary to investigate these lesser-explored benign effects of effective and highly productive automation technologies, capable of offsetting the negative job-displacing effect (Acemoglu and Restrepo, 2019).

6.2. Limitations and future research

Despite its rigor, our study is not without limitations inherent to the choices made at various stages of implementation. Firstly, we acknowledge that the results of our empirical exercise are conditional on the initial keywords' selection, as well as on the combination of parameters tested for each ML algorithm. This limitation can only be addressed by conducting further empirical testing. Therefore, we encourage researchers to explore new applications of the same algorithms using the STATA commands employed in this study. Alternatively, researchers could experiment with similar applications using alternative open-source statistical software, thus expanding further the set of practical applications for those who are approaching the world of data science. Secondly, regarding the data used in our empirical exercise, we recognise that restructuring data obtained from the ERM database comes with certain limitations. While it covers a remarkably extensive sample of restructuring events, encompassing various types of business restructuring across most European countries, it may not necessarily fully represent the entire population. It is noteworthy that ERM data primarily focuses on large-scale events involving either substantial layoffs or hiring of workers, thus potentially overlooking smaller organizational restructurings (Eurofound, 2023). Nevertheless, we underscore that despite these limitations, this dataset provides a rich source of information that remains largely untapped, holding potential for further interesting investigation.

In conclusion, given the scarcity of firm-level information on the adoption of I4.0/automation technologies – primarily available from private dedicated surveys or proxied by administrative data (e.g., imports of automation-related capital goods) – the classification exercise we undertake responds to a further call regarding the importance of studying I4.0 (Luo and Zahra, 2023). From this standpoint, while our work is methodological and empirical in nature – and therefore not situated within a specific theoretical framework – it offers a tool for examining how I4.0 technologies influence the international reorganisation of business operations. As discussed by Luo and Zahra (2023), the variable we construct using ML may facilitate investigations into how I4.0 technologies are reshaping global and regional (e.g., European) value chains, how they are shaping multinational corporations' decisions regarding the organisation and oversight of their resources, and how new automation technologies help multinationals in developing new firm-specific advantages and exploit existing country-specific advantages. Therefore, future research could investigate alternative classifications of the ERM data to dissect the types of value chain activities involved in restructuring events, the categories of workers affected by layoffs and hirings, as well as different economic and social

Table 10
Summary statistics and two-samples *t*-test on business restructuring events.

Type of restructuring	Job-creating	Job-displacing
Not I4.0-related		
Frequency	9927	13,320
%	97.49	99.14
Mean (net job gain/loss)	220.969	−458.618
SD (net job gain/loss)	197.170	1222.613
I4.0-related		
Frequency	256	115
%	2.51	0.86
Mean (net job gain/loss)	187.973	−425.330
SD (net job gain/loss)	157.996	699.922
Two-sample t-test with unequal variances (*p*-value)	0.001	0.616

Notes: Authors' own elaboration based on ERM data. Job-creating events: business expansion; job-displacing events: bankruptcy/closure, downsizing (internal restructuring), merger/acquisition, offshoring/delocalisation, outsourcing/relocation.

determinants behind these events.

CRediT authorship contribution statement

Fabio Lamperti: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The author has no conflicts of interest, no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are available from the author upon reasonable request.

Acknowledgements

I thank the Editor and three anonymous reviewers for their most constructive and helpful suggestions.

References

Acemoglu, D., Restrepo, P., 2019. Automation and new tasks: how technology displaces and reinstates labor. J. Econ. Perspect. 33 (2), 3–30. https://doi.org/10.1257/jep.33.2.3.

Acemoglu, D., Restrepo, P., 2020. Robots and jobs: Evidence from US labor markets. J. Polit. Econ. 128 (6), 2188–2244. https://doi.org/10.1086/705716.

Acemoglu, D., Lelarge, C., Restrepo, P., 2020. Competing with robots: firm-level evidence from France. AEA Papers and Proceedings 110, 383–388. https://doi.org/10.1257/pandp.20201003.

Agarwal, R., Dhar, V., 2014. Big data, data science, and analytics: the opportunity and challenge for IS research. Inf. Syst. Res. 25 (3), 443–448. https://doi.org/10.1287/isre.2014.0546.

Ancarani, A., di Mauro, C., Mascali, F., 2019. Backshoring strategy and the adoption of industry 4.0: evidence from Europe. J. World Bus. 54 (4), 360–371. https://doi.org/10.1016/j.jwb.2019.04.003.

Anderson, J., Sutherland, D., 2015. Developed economy investment promotion agencies and emerging market foreign direct investment: the case of Chinese FDI in Canada. Journal of World Business 50 (4), 815–825. https://doi.org/10.1016/j.jwb.2015.04.005.

Andreu, A.B., Lavoratori, K., 2022. History matters: colonial-based connectivity and foreign headquarter location choice. Manag. Int. Rev. https://doi.org/10.1007/s11575-022-00481-2.

Antràs, P. (2020). De-globalisation? Global value chains in the post-COVID-19 age. SELL J. (Vol. 5, issue 1). https://doi.org/10.3386/w28115.

Anzolin, G., Andreoni, A., Zanfei, A., 2022. What is driving robotisation in the automotive value chain? Empirical evidence on the role of FDIs and domestic capabilities in technology adoption. Technovation 115 (December 2020), 102476. https://doi.org/10.1016/j.technovation.2022.102476.

Balland, P.-A., Boschma, R., 2021. Mapping the potentials of regions in Europe to contribute to new knowledge production in Industry 4.0 technologies. Reg. Stud. 1–15. https://doi.org/10.1080/00343404.2021.1900557.

Ballestar, M.T., Díaz-Chao, Á., Sainz, J., Torrent-Sellens, J., 2020. Knowledge, robots and productivity in SMEs: explaining the second digital wave. J. Bus. Res. 108, 119–131. https://doi.org/10.1016/j.jbusres.2019.11.017.

Barbieri, P., Elia, S., Fratocchi, L., Golini, R., 2019. Relocation of second degree: moving towards a new place or returning home? J. Purch. Supply Manag. 25 (3), 100525 https://doi.org/10.1016/j.pursup.2018.12.003.

Barbieri, P., Boffelli, A., Elia, S., Fratocchi, L., Kalchschmidt, M., 2022. How does industry 4.0 affect international exposure? The interplay between firm innovation and home-country policies in post-offshoring relocation decisions. Int. Bus. Rev. 101992 https://doi.org/10.1016/j.ibusrev.2022.101992.

Bas, J., Cirillo, C., Cherchi, E., 2021. Classification of potential electric vehicle purchasers: a machine learning approach. Technol. Forecast. Soc. Chang. 168 https://doi.org/10.1016/j.techfore.2021.120759.

Bhandari, K.R., Zámborský, P., Ranta, M., Salo, J., 2023. Digitalization, internationalization, and firm performance: a resource-orchestration perspective on new OLI advantages. International Business Review 32 (4), 102135. https://doi.org/10.1016/j.ibusrev.2023.102135.

Blit, J., 2020. Automation and reallocation: will COVID-19 usher in the future of work? Can. Public Policy 46 (S2), S192–S202. https://doi.org/10.3138/cpp.2020-065.

Brynjolfsson, E., Rock, D., Syverson, C., 2019. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In: Agrawal, A., Gans, J., Goldfarb, A. (Eds.), The Economics of Artificial Intelligence. University of Chicago Press, pp. 23–60. https://doi.org/10.3386/w24001.

Carstensen, K., Toubal, F., 2004. Foreign direct investment in central and eastern European countries: a dynamic panel analysis. J. Comp. Econ. 32 (1), 3–22. https://doi.org/10.1016/j.jce.2003.11.001.

Castellani, D., Lavoratori, K., 2020. The lab and the plant: Offshore R&D and co-location with production activities. Journal of International Business Studies 51 (1), 121–137. https://doi.org/10.1057/s41267-019-00255-3.

Castellani, D., Lamperti, F., Lavoratori, K., 2022. Measuring adoption of industry 4.0 technologies via international trade data: insights from European countries. Journal of industrial and business. Economics 49 (1), 51–93. https://doi.org/10.1007/s40812-021-00204-y.

Cerulli, G., 2021. Improving econometric prediction by machine learning. Appl. Econ. Lett. 28 (16), 1419–1425. https://doi.org/10.1080/13504851.2020.1820939.

Cerulli, G., 2022. Machine learning using Stata/Python. The Stata Journal: Promoting Communications on Statistics and Stata 22 (4), 772–810. https://doi.org/10.1177/1536867X221140944.

Cette, G., Devillard, A., Spiezia, V., 2021. The contribution of robots to productivity growth in 30 OECD countries over 1975-2019 ☆. Econ. Lett. 200, 109762 https://doi.org/10.1016/j.econlet.2021.109762.

Chen, Z., Schintler, L.A., 2023. Rediscovering regional science: positioning the field's evolving location in science and society. J. Reg. Sci. 63 (3), 617–642. https://doi.org/10.1111/jors.12634.

Choi, S., Cha, S., Tappert, C.C., 2010. A survey of binary similarity and distance measures. Journal on Systemics, Cybernetics and Informatics 8, 43–48.

Chung, P., Sohn, S.Y., 2020. Early detection of valuable patents using a deep learning model: case of semiconductor industry. Technological Forecasting and Social Change 158, 120146. https://doi.org/10.1016/j.techfore.2020.120146.

Ciffolilli, A., Muscio, A., 2018. Industry 4.0: national and regional comparative advantages in key enabling technologies. European Planning Studies 26 (12), 2323–2343. https://doi.org/10.1080/09654313.2018.1529145.

Cockburn, I.M., Henderson, R., Stern, S., 2019. The impact of artificial intelligence on innovation: An exploratory analysis. In: Agrawal, A., Gans, J., Goldfarb, A. (Eds.), The Economics of Artificial Intelligence: An Agenda. University of Chicago Press, pp. 115–148. https://doi.org/10.3386/w24449.

Corradini, C., Santini, E., Vecciolini, C., 2021. The geography of industry 4.0 technologies across European regions. Regional Studies 1–14. https://doi.org/10.1080/00343404.2021.1884216.

Coucke, K., Sleuwaegen, L., 2008. Offshoring as a survival strategy: evidence from manufacturing firms in Belgium. Journal of International Business Studies 39 (8), 1261–1277. https://doi.org/10.1057/PALGRAVE.JIBS.8400403.

Coucke, K., Pennings, E., Sleuwaegen, L., 2007. Employee layoff under different modes of restructuring: exit, downsizing or relocation. Ind. Corp. Chang. 16 (2), 161–182. https://doi.org/10.1093/ICC/DTM002.

Cséfalvay, Z., 2020. Robotization in central and Eastern Europe: catching up or dependence? Eur. Plan. Stud. 28 (8), 1534–1553. https://doi.org/10.1080/09654313.2019.1694647.

Cugno, M., Castagnoli, R., Büchi, G., 2021. Openness to industry 4.0 and performance: the impact of barriers and incentives. Technological Forecasting and Social Change 168 (March), 120756. https://doi.org/10.1016/j.techfore.2021.120756.

Dachs, B., Kinkel, S., Jäger, A., 2019. Bringing it all back home? Backshoring of manufacturing activities and the adoption of industry 4.0 technologies. Journal of World Business 54 (6), 101017. https://doi.org/10.1016/j.jwb.2019.101017.

Dalenogare, L.S., Benitez, G.B., Ayala, N.F., Frank, A.G., 2018. The expected contribution of industry 4.0 technologies for industrial performance. International Journal of Production Economics 204, 383–394. https://doi.org/10.1016/J.IJPE.2018.08.019.

De Backer, K., De Stefano, T., Menon, C., Ran Suh, J., 2018. Industrial Robotics and the Global Organisation of Production. In: OECD Science, Technology and Industry Working Papers 2018/03. OECD Publishing, Paris. https://doi.org/10.1787/dd98ff58-en.

Dhar, V., 2013. Data science and prediction. Commun. ACM 56 (12), 64–73. https://doi.org/10.1145/2500499.

Di Stefano, E., Giovannetti, G., Mancini, M., Marvasi, E., Vannelli, G., 2022. Reshoring and plant closures in Covid-19 times: evidence from Italian MNEs. International Economics 172, 255–277. https://doi.org/10.1016/j.inteco.2022.09.009.

Domini, G., Grazzi, M., Moschella, D., Treibich, T., 2021. Threats and opportunities in the digital era: automation spikes and employment dynamics. Research Policy 50 (7), 104137. https://doi.org/10.1016/j.respol.2020.104137.

Domini, G., Grazzi, M., Moschella, D., Treibich, T., 2022. For whom the bell tolls: the firm-level effects of automation on wage and gender inequality. Research Policy 51 (7), 104533. https://doi.org/10.1016/J.RESPOL.2022.104533.

Du, L., Lin, W., 2022. Does the application of industrial robots overcome the Solow paradox? Evidence from China. Technology in Society 68, 101932. https://doi.org/10.1016/j.techsoc.2022.101932.

Escobar, M., 2015. Studying coincidences with network analysis and other multivariate tools. In The Stata Journal: Promoting Communications on Statistics and Stata 15 (4), 1118–1156.

Eurofound, 2023. About the European Restructuring Monitor. https://www.eurofound.europa.eu/observatories/emcc/erm/about-the-european-restructuring-monitor.

European Central Bank (ECB). (2013). Monthly Bulletin June 2013. European Central Bank, Germany. Retrieved from https://www.ecb.europa.eu/pub/pdf/mobu/mb201306en.pdf.

European Central Bank (ECB). (2020). ECB Economic Bulletin, Issue 1/2020. European Central Bank, Germany. Retrieved from https://www.ecb.europa.eu/pub/economic-bulletin/html/eb202001.en.html.

European Investment Bank, 2019. Investment report 2019/2020: accelerating Europe's transformation. https://www.eib.org/attachments/efs/eibis_2019_european_union_en.pdf.

European Investment Bank, 2021. EIB investment survey 2021: European Union overview. https://www.eib.org/attachments/publications/eibis_2021_european_union_en.pdf.
Eurostat, 2023. ICT usage in enterprises. Retrieved from Eurostat, the statistical office of the European Union: https://ec.europa.eu/eurostat/databrowser/explore/all/science?lang=en&subtheme=isoc.isoc_e.isoc_eb&display=list&sort=category.
Fantechi, F., Modica, M., 2022. Learning from the past: a machine-learning approach for predicting the resilience of locked-in regions after a natural shock. Regional Studies 1–14. https://doi.org/10.1080/00343404.2022.2089644.
Felice, G., Lamperti, F., Piscitello, L., 2022. The employment implications of additive manufacturing. Ind. Innov. 29 (3), 333–366. https://doi.org/10.1080/13662716.2021.1967730.
Finch, H., 2005. Comparison of distance measures in cluster analysis with dichotomous data. Journal of Data Science 3 (1), 85–100. https://doi.org/10.6339/JDS.2005.03(1).192.
Frank, A.G., Dalenogare, L.S., Ayala, N.F., 2019. Industry 4.0 technologies: implementation patterns in manufacturing companies. Int. J. Prod. Econ. 210, 15–26. https://doi.org/10.1016/j.ijpe.2019.01.004.
Garbe, J.-N., Richter, N.F., 2009. Causal analysis of the internationalization and performance relationship based on neural networks — advocating the transnational structure. J. Int. Manag. 15 (4), 413–431. https://doi.org/10.1016/j.intman.2008.10.002.
Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as data. J. Econ. Lit. 57 (3), 535–574. https://doi.org/10.1257/jel.20181020.
George, G., Osinga, E.C., Lavie, D., Scott, B.A., 2016. Big data and data science methods for management research. Acad. Manage. J. 59 (5), 1493–1507. https://doi.org/10.5465/amj.2016.4005.
Graetz, G., Michaels, G., 2018. Robots at work. Rev. Econ. Stat. 100 (5), 753–768. https://doi.org/10.1162/rest_a_00754.
Guenther, N., Schonlau, M., 2016. Support vector machines. The Stata Journal: Promoting Communications on Statistics and Stata 16 (4), 917–937.
Hassan, S.U., Ahamed, J., Ahmad, K., 2022. Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers 3, 238–248. https://doi.org/10.1016/J.SUSOC.2022.03.001.
Herrera, R., Climent, F., Carmona, P., Momparler, A., 2022. The manipulation of Euribor: an analysis with machine learning classification techniques. Technological Forecasting and Social Change 176, 121466. https://doi.org/10.1016/j.techfore.2021.121466.
Igna, I., Venturini, F., 2023. The determinants of AI innovation across European firms. Res. Policy 52 (2), 104661. https://doi.org/10.1016/j.respol.2022.104661.
Jäger, A., Moll, C., Som, O., Zanker, C., 2015. Analysis of the impact of robotic systems on employment in the European Union. In European Commission. https://doi.org/10.2759/516348.
Jurafsky, D., Martin, J.H., 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd ed. Pearson, London.
Kagermann, H., Wahlster, W., Helbig, J., 2013. Recommendations for implementing the strategic initiative INDUSTRIE 4.0. http://alvarestech.com/temp/RoboAseaIRB6S2-Fiat/CyberPhysicalSystems-Industrial4-0.pdf.
Kim, J., Kim, H., Geum, Y., 2023. How to succeed in the market? Predicting startup success using a machine learning approach. Technological Forecasting and Social Change 193. https://doi.org/10.1016/j.techfore.2023.122614.
Kinkel, S., 2020. Industry 4.0 and reshoring. In: de Propris, L., Bailey, D. (Eds.), Industry 4.0 and Regional Transformations. Routledge, pp. 195–213 (www.routledge.com/).
Lamperti, F., Lavoratori, K., Castellani, D., 2023. The unequal implications of industry 4.0 adoption: evidence on productivity growth and convergence across Europe. Econ. Innov. New Technol. 1–25. https://doi.org/10.1080/10438599.2023.2269089.
Lindner, T., Puck, J., Verbeke, A., 2022. Beyond addressing multicollinearity: robust quantitative analysis and machine learning in international business research. In: Journal of International Business Studies, Vol. 53. Palgrave Macmillan, pp. 1307–1314. https://doi.org/10.1057/s41267-022-00549-z (Issue 7).
Luo, Y., Zahra, S.A., 2023. Industry 4.0 in international business research. J. Int. Bus. Stud. 54 (3), 403–417. https://doi.org/10.1057/s41267-022-00577-9.
Marcucci, G., Antomarioni, S., Ciarapica, F.E., Bevilacqua, M., 2021. The impact of operations and IT-related industry 4.0 key technologies on organizational resilience. Prod. Plan. Control 1–15. https://doi.org/10.1080/09537287.2021.1874702.
Mariani, M., Borghi, M., 2019. Industry 4.0: a bibliometric review of its managerial intellectual structure and potential evolution in the service industries. Technological Forecasting and Social Change 149 (September), 119752. https://doi.org/10.1016/j.techfore.2019.119752.
Martinelli, A., Mina, A., Moggi, M., 2021. The enabling technologies of industry 4.0: examining the seeds of the fourth industrial revolution. Industrial and Corporate Change 30 (1), 161–188. https://doi.org/10.1093/icc/dtaa060.
McAfee, A., Brynjolfsson, E., 2012. Big data: the management revolution. Harv. Bus. Rev. 90 (10), 60–68.
Mikko, M., Stein, Ø., Jaakko, S., 2022. Machine learning and the identification of smart specialisation thematic networks in Arctic Scandinavia. Regional Studies 56 (9), 1429–1441. https://doi.org/10.1080/00343404.2021.1925237.
Ministry of Industry and Trade of the Czech Republic, 2019. National Artificial Intelligence Strategy of the Czech Republic. Retrieved from. https://www.mpo.cz/assets/en/guidepost/for-the-media/press-releases/2019/5/NAIS_eng_web.pdf.
Miric, M., Jia, N., Huang, K.G., 2022. Using supervised machine learning for large-scale classification in management research: the case for identifying artificial intelligence patents. Strateg. Manag. J. https://doi.org/10.1002/smj.3441.
Montobbio, F., Staccioli, J., Virgillito, M.E., Vivarelli, M., 2022. Robots and the origin of their labour-saving impact. Technological Forecasting and Social Change 174 (October 2021), 121122. https://doi.org/10.1016/j.techfore.2021.121122.
Müller, J.M., Buliga, O., Voigt, K.-I., 2018. Fortune favors the prepared: how SMEs approach business model innovations in industry 4.0. Technological Forecasting and Social Change 132 (December 2017), 2–17. https://doi.org/10.1016/j.techfore.2017.12.019.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge.
Nielsen, F., 2016. Hierarchical clustering. In: Introduction to HPC With MPI for Data Science. Springer, Cham, pp. 195–211. https://doi.org/10.1007/978-3-319-21903-5_8.
Pavlínek, P., 2018. Global production networks, foreign direct investment, and supplier linkages in the integrated peripheries of the automotive industry. Econ. Geogr 94 (2), 141–165. https://doi.org/10.1080/00130095.2017.1393313.
Pedota, M., Grilli, L., Piscitello, L., 2023. Technology adoption and upskilling in the wake of industry 4.0. Technological Forecasting and Social Change 187, 122085. https://doi.org/10.1016/j.techfore.2022.122085.
Porter, M.F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.
Saura, J.R., Palacios-Marqués, D., Ribeiro-Soriano, D., 2023. Exploring the boundaries of open innovation: evidence from social media mining. Technovation 119. https://doi.org/10.1016/j.technovation.2021.102447.
Savin, I., Chukavina, K., Pushkarev, A., 2022. Topic-based classification and identification of global trends for startup companies. Small Bus. Econ. https://doi.org/10.1007/s11187-022-00609-6.
Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. The Stata Journal: Promoting Communications on Statistics and Stata 20 (1), 3–29. https://doi.org/10.1177/1536867X20909688.
Shrestha, Y.R., He, V.F., Puranam, P., von Krogh, G., 2021. Algorithm supported induction for building theory: how can we use prediction models to theorize? Organ. Sci. 32 (3), 856–880. https://doi.org/10.1287/orsc.2020.1382.
Teixeira, J.E., Tavares-Lehmann, A.T.C.P., 2022. Industry 4.0 in the European union: policies and national strategies. Technological Forecasting and Social Change 180, 121664. https://doi.org/10.1016/j.techfore.2022.121664.
Tidhar, R., Eisenhardt, K.M., 2020. Get rich or die trying… finding revenue model fit using machine learning and multiple cases. Strateg. Manag. J. 41 (7), 1245–1273. https://doi.org/10.1002/smj.3142.
Van Roy, V., Vertesy, D., Damioli, G., 2020. AI and robotics innovation. In: Zimmermann, K. (Ed.), Handbook of Labor, Human Resources and Population Economics. Springer, Cham.
Veiga, J.F., Lubatkin, M., Calori, R., Very, P., Tung, Y.A., 2000. Using neural network analysis to uncover the trace effects of National Culture. Journal of International Business Studies 31 (2), 223–238. https://doi.org/10.1057/palgrave.jibs.8490903.
Vuorio, A., Torkkeli, L., 2023. Dynamic managerial capability portfolios in early internationalising firms. International Business Review 32 (1), 102049. https://doi.org/10.1016/j.ibusrev.2022.102049.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.
Williams, U., Williams, S.P., 2014. txttool: utilities for text analysis in Stata. The Stata Journal: Promoting Communications on Statistics and Stata 14 (4), 817–829.

Fabio Lamperti is Assistant Professor (RTD-A) in Economic Policy at Università Cattolica del Sacro Cuore of Milan (Italy) and an Affiliated Member at CIRCLE, Lund University (Sweden). He hold a PhD in Management – International Business and Strategy – from Henley Business School, University of Reading (UK) and a MSc in Management Engineering from Polytechnic University of Milan (Italy). Before joining UniCatt, Fabio has been research fellow in Economics at University of Perugia (Italy), and teaching/research assistant both at Henley Business School, University of Reading, and at the Department of Management Engineering, Polytechnic University of Milan. His main research interests focus on the technological change related to the Industry 4.0 transformation and its impact on employment, economic growth and productivity, and firm's restructuring decisions. He further investigates the multifaceted process of technology adoption, its determinants and dynamics. His publications have appeared in leading journals such as *Industry and Innovation*, *Economic of Innovation and New Technology* and *Research Policy*.