Received: 20 November 2020 | Revised: 9 May 2022 | Accepted: 8 June 2022 | Published on: 11 July 2022

DOI: 10.1002/smj.3441

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents

Milan Miric[1] | **Nan Jia**[2] | **Kenneth G. Huang**[3]

[1]Department of Data Sciences & Operations, Marshall School of Business, University of Southern California, Los Angeles, California, USA

[2]Department of Management & Organization, Marshall School of Business, University of Southern California, Los Angeles, California, USA

[3]Department of Industrial Systems Engineering & Management, College of Design and Engineering & Department of Strategy & Policy, NUS Business School, National University of Singapore, Singapore, Singapore

Correspondence
Nan Jia, Department of Management & Organization, Marshall School of Business, University of Southern California, Los Angeles, CA 90007, USA.
Email: nan.jia@marshall.usc.edu

Funding information
Institute for Outlier Research (iORB)

Correction made on 21 July 2022 after first online publication: The affiliation of Kenneth G. Huang has been updated in this version.

Abstract

Research Summary: Researchers increasingly use unstructured text data to construct quantitative variables for analysis. This goal has traditionally been achieved using keyword-based approaches, which require researchers to specify a dictionary of keywords mapped to the theoretical concepts of interest. However, recent machine learning (ML) tools for text classification and natural language processing can be used to construct quantitative variables and to classify unstructured text documents. In this paper, we demonstrate how to employ ML tools for this purpose and discuss one application for identifying artificial intelligence (AI) technologies in patents. We compare and contrast various ML methods with the keyword-based approach, demonstrating the advantages of the ML approach. We also leverage the classification outcomes generated by ML models to demonstrate general patterns of AI technological innovation development.

Managerial Summary: Text-based documents offer a wealth of information for researchers and business analysts. However, researchers often need to find a way to classify these documents to use in subsequent research projects. In this paper, we demonstrate how supervised

ML methods can be used to automate the process of classifying textual documents into pre-defined categories or groups. We provide an overview of when such techniques may be used in comparison to other methods, and the considerations and tradeoffs associated with each method. We apply these methods to identify AI-based technologies from all patents in the United States, based on patent abstract text. This allows us to show interesting patterns of AI innovation development in the United States. We also provide the code and data used in this paper for future research.

KEYWORDS

artificial intelligence, keywords, machine learning, patent and innovation, text analysis

1 | INTRODUCTION

Text documents represent a rich and important source of data for social scientists (Gentzkow, Kelly, & Taddy, 2019). Researchers who need to construct empirical variables from the wealth of unstructured information embedded in text data—such as financial statements, patent texts, or company descriptions—for theoretically grounded constructs have traditionally relied on keyword dictionaries, which specify keywords that can be mapped with these concepts (e.g., Cho & Hambrick, 2006; Duriau, Reger, & Pfarrer, 2007; Kaplan, 2008). While this approach is effective when the theoretical constructs and language used to describe them are clear, well-established, and commonly accepted, in many other contexts it is challenging to construct and validate a dictionary of keywords. To address this issue, some strategy researchers have recently begun using machine learning (ML) techniques to construct theoretically grounded variables from text data (e.g., Choudhury & Kim, 2019; Leyden, 2018; Tidhar & Eisenhardt, 2020; Miric, Pagani, & El Sawy, 2021; Shrestha, He, Puranam and von Krogh, 2021).[1] Although these approaches are beginning to be implemented, best practices regarding the use of supervised ML tools to classify text data are not widely known among strategy researchers so those unfamiliar with these methods may find it challenging to apply them in their own research. This lack of clear guidance may deter many from adopting ML methods that would otherwise provide substantial value in strategy and management research.

We address these gaps by providing a walkthrough of the use of supervised ML methods in the large-scale classification of text documents and their application in the context of classifying patent abstracts based on their relationship to artificial intelligence (AI) or non-AI technologies.

[1]Researchers also use ML to analyze unstructured text data to (1) conduct topic modeling to identify the dominant patterns in text data for which no theoretical priors exist (Choudhury, Allen, and Endres, 2019; Hannigan et al., 2019; Kaplan & Vakili, 2015; Teodoridis, Lu, & Furman, 2020) and (2) identify predictors of performance outcomes in order to subsequently inform theory development; this approach is similar to an inductive research method (Shrestha, He, Puranam and von Krogh, 2021; Tidhar and Eisenhardt, 2020). We will provide a more detailed discussion in subsequent sections.

We first provide a conceptual distinction between supervised ML and other ML approaches to provide readers with a clear understanding of how different ML approaches may be used and for which research goals. We also discuss why researchers might use ML instead of a traditional keyword-based approach. Moreover, we identify critical choices arising from the implementation of supervised ML methods and demonstrate how to address them, including how to construct the training data, compare and contrast different classification approaches, and interpret the classification model (to alleviate concerns stemming from the common criticism of ML methods as being "black boxes"). Third, to further encourage the diffusion of these methods, we provide our code for identifying AI patents though both Git-Hub and Google Colab (Google Drive-based online workspace) so that researchers can adapt these tools easily to their own applications. The workspace includes free access to graphic processing unit (GPU) processors, which are necessary for more complex classification models. This can allow others to replicate our work and apply these methods in their research.

In addition to the in-depth discussion of the decisions that researchers must make when applying ML, we generate new insights into the empirical context in which we demonstrate these tools. Because "artificial intelligence" is an imprecise term referring to a great variety of statistical prediction techniques developed for different settings (Agrawal, Gans, & Goldfarb, 2018), our classification of AI patents provides insight into which technologies represent AI. This insight can, in turn, provide a theoretical understanding of which technologies actually fit the "AI" label. We report our insights in this paper and through an online interactive dashboard which, in addition to the online workspace containing data and code, we intend to keep updated providing up to date information about this phenomenon (http://www.aipatents.xyz/).

The contributions of this paper are twofold. First, its applied methodological contribution provides an in-depth exposition of supervised ML methods, showing their use in classifying pre-existing concepts from large-scale text data; this common goal in strategy research has thus far been achieved predominantly in rudimentary ways, such as by devising keyword-based dictionaries. Our exposition provides an in-depth comparison of the benefits and shortcomings of supervised ML and other approaches. We also provide clear guidance for researchers on whether to use supervised ML methods or the traditional, dictionary-based method. This comparative approach examines how ML methods can aid strategic management research and demonstrates best practices for their use in text analysis (Choudhury, Wang, Carlson, & Khanna, 2019; Tidhar & Eisenhardt, 2020; Furman & Teodoridis, 2020; Shrestha, He, Puranam and von Krogh, 2021). To enable researchers to easily apply these methods, we provide a link to our code in an online computational environment, where researchers can freely access our code and data and download these to their own Google drives.

The second contribution of this paper is insight into the context of AI technologies, a domain that is both theoretically meaningful and empirically important. The broad label "artificial intelligence" is used for a wide range of technologies. However, as many recent authors have argued (e.g., Agrawal et al., 2018), this label is imprecise because, in reality, the term covers a wide set of statistical prediction techniques applied in different settings. Therefore, we aim to demonstrate the extent to which different technologies are currently used in relation to AI, helping subsequent researchers study this phenomenon more precisely. Moreover, we intend to provide a large-scale, comprehensive, and evolving database of AI patents,[2] with the

[2]We intend to keep the database updated as new patents are granted and as new training data becomes available.

critical information needed to accurately understand the extent of patented innovations in this context.

This paper proceeds as follows. First, we review various ML tools commonly used by management researchers. We compare the supervised ML-based text classification tools described in this paper to other techniques, such as topic modeling, which are currently more common. This is not an exhaustive comparison of all possible ML applications but rather a framework for determining the most appropriate ML in a given situation. Second, we demonstrate the steps necessary to classify patents based on their text descriptions, the various critical decisions researchers must make throughout this process, and guidance regarding related best practices. Third, we present techniques for evaluating the effectiveness of ML classification methods and compare the ML approach with traditional keyword-based approaches. Finally, we provide a high-level discussion and recommendations for assessing when ML approaches are appropriate in research.

2 | OVERVIEW OF MACHINE LEARNING METHODS IN STRATEGIC MANAGEMENT RESEARCH

2.1 | Common applications of machine learning methods

ML refers to a set of statistical algorithms that "learns" or improves from previous iterations without explicit instructions. ML techniques are used increasingly by researchers in strategy and management, most commonly in one the following situations.

First, *supervised* ML algorithms have been used to classify observations into groups based on theoretical concepts, with the aim of creating categorical variables based on pre-specified theoretical constructs. For example, Miric et al. (2021) use ML techniques to distinguish platform companies from non-platform companies to investigate differences in their acquisition strategies. Choudhury and Kim (2019) use ML to construct categorical variables indicating ethnicity of inventors based on the names of those inventors.

Second, *unsupervised* ML approaches, such as topic modeling, have been used to identify dominant patterns or groups within text data or image data without the guidance of any pre-existing theoretical concepts (Choudhury et al., 2019; Kaplan & Vakili, 2015; Choi, Menon, Tabakovic, 2021; Teodoridis et al., 2020; for a recent review, see Hannigan et al., 2019).

Third, ML methods have been used to transform textual and image data into numerical representations. This approach allows the computation of distance metrics between text or image data, for instance finding patents that are related to similar topics (Choudhury, Starr, & Agarwal, 2020), distance between the areas of expertise of academic scientists (Furman & Teodoridis, 2020), novelty of products based on their textual descriptions (Miric, Ozalp, & Yilmaz, 2020) or overlap between the industry of companies based on the text in their financial disclosures (Hoberg & Phillips, 2016).

Fourth, ML algorithms have also been used to identify variables with considerable explanatory power to provide an input for theory-building. This approach involves using ML to identify relationships that emerge naturally in the data; then, researchers use this knowledge to develop theory as part of an inductive approach (Choudhury, Allen, & Endres, 2021; Tidhar & Eisenhardt, 2020; He, Puranam, Shrestha, & von Krogh, 2020; Shrestha, He, Puranam, and von Krogh, 2021). In this paper, we focus on *supervised* ML methods, with the specific application of constructing categorical variables theoretically defined from text data.

Researchers also use ML methods to augment existing econometric methods rather than to construct variables. For instance, variable selection algorithms (i.e., LASSO) have been used to automate the selection of control variables (Belloni, Chernozhukov, & Hansen, 2014; Guzman & Stern, 2020; Miric, Boudreau, & Jeppesen, 2019) and ML methods have been used to predict instrumental variables (Singh et al, 2020). Furthermore, researchers have used ML methods to derive causal estimates—either by using ML to construct a better-matched sample (Rathje & Katila, 2021) or by using prediction algorithms to explore heterogeneous treatment effects in experiments (Miric & Jeppesen, 2020; Wager & Athey, 2018).

Table 1 provides an overview of the common applications of ML observed in strategic management research and illustrates where the *supervised* text classification methods, which are the focus of this paper, are situated within the literature. It also highlights the differences and similarities of supervised methods compared with other approaches that use ML methods.

2.2 | The reason for using ML to construct categorical variables from text data

Grouping or classifying large-scale text data into theoretically grounded categories has typically been achieved by using what is referred to as "keyword-based" classification. The keyword-based classification of text documents is based on a simple programmatic conditional statement (i.e., if a text document contains the *KEYWORD*, then the text is coded as *TRUE*, which means it belongs to the category of interest). Researchers can specify the set of keywords—often referred to as a "dictionary"—required by an application.

An alternative approach to grouping data into theoretically grounded groups is to first construct a training dataset, and then use ML approaches to classify the full population of observations based on those observations. Supervised ML approaches are a data-driven way of constructing groupings of outcome variables from text data. By design, supervised ML approaches generate various metrics for assessing and comparing the classification performance. There are advantages and drawbacks to both the keyword-based and ML approaches, and either may be appropriate in certain situations. At times, a blended approach using both may be optimal. We discuss below the relative benefits of the ML approach, comparing it to the more commonly used keyword-based approach.

The keyword approach has several benefits. It is straightforward in terms of implementation when a dictionary is available. Second, these dictionaries are often transferable across contexts. For example, different types of texts, such as news articles, social media posts, and academic publications, may share common keywords. Third, keyword-based approaches are often easily interpretable because researchers can share a keyword dictionary.

The keyword approach also has specific limitations that ML methods may be able to overcome. First, implementing a keyword-based approach may be difficult if no dictionary exists. Alternatively, in such instances, text classification can still be performed using ML methods. For instance, Choudhury et al. (2019) studied CEO communication styles, a domain in which it may be difficult to ex ante specify the keywords associated with any particular style. Similarly, Miric et al. (2021) distinguish between platform companies and non-platform companies, which lack a common dictionary of keywords. In these situations, it is straightforward to manually classify a small number of observations to construct training data, as it may be easy for an individual to "know it when [they] see it" (Polanyi, 1966) but difficult for that same individual to

TABLE 1 Summary of common applications of machine learning methods in strategic management research

ML approach	Objective	Example of applications in strategic management research
Topic modeling (e.g. *LDA*, *PCA*, *NMF*)	Identifying prominent groups in the data	• Identify dominant technology groupings in patents (Teodoridis, Lu & Furman, 2020; Kaplan & Vakili, 2015) • Identify dominant business areas of companies to create more fine-grained measure of competitive overlap between companies (Hoberg & Phillips, 2016; Shi, Lee, and Whinston, 2016)
Exploratory data analysis *(*e.g. *Random Forest*, *LASSO*, *or other interpretable algorithms)*	Identifying the features (characteristics) associated with an outcome of interest (e.g. performance)	• Identify key business model features associated with performance (Choudhury et al., 2021; Shrestha, He, Puranam, & von Krogh, 2021; Tidhar & Eisenhardt, 2020)
Constructing categorical variables (e.g. *Naïve Bayes*, *Random Forest*, *SVM*, *Neural Networks)*	Constructing categorical variables in a large dataset where manual coding is infeasible	• Construct categorical variable indicating ethnicity based on individual names (Choudhury & Kim, 2019). • Construct categorical variable describing "Managerial Style" on the basis of video and transcripts (Choudhury et al., 2019)
Constructing distance measures between text or image documents	Constructing numerical representations of textual, image or voice data and computing distance metrics	• Measure distance between areas of scientific research based on text documents (Furman & Teodoridis, 2020) • Identify distance between scientific domain of companies (Hoberg & Philips, 2016) • Identify closest patents based on textual distance between patent claims (Choudhury et al., 2020) • Identify products which are novel based on their textual description (Miric et al., 2020). • Identify similarity between images (Gross, 2020) • Identify similarity between patents based on text embeddings (Hain, Jurowetzki, Buchmann & Wolf, 2022) • Identify similarity of firm strategies based on the text descriptions on their websites (Guzman & Li, 2022).

TABLE 1 (Continued)

ML approach	Objective	Example of applications in strategic management research
Automating human decisions in econometrics	Automating (removing human input) from econometric procedures (e.g., variable selection)	• Perform matching based method on text data using machine learning (Rathje & Katila, 2021) • Construct composite instrumental variables (Singh et al., 2020) • Explore heterogeneous effects following difference-in-differences regressions using random forest (Wager & Athey, 2018) • LASSO to select and include control variables (Miric et al., 2019; Guzman & Stern, 2020)

abstract intuitively to construct a list of keywords that reliably and comprehensively distinguishes the two groups.

A second limitation to the keyword-based approach is the common concern that dictionaries can be subjective—but, more importantly—that they are often difficult to validate.[3] Although supervised ML methods also require human coders to construct training data, and human coders are used to manually classify observations, concerns regarding subjectivity and error are reduced through the use of established techniques and practices—such as cross validation—that validate the accuracy of ML-based methods. By contrast, dictionaries required for keyword-based models are difficult to validate because researchers must manually code a validation sample; doing so may take the same effort as implementing an ML approach.[4]

A third limitation is the limited capacity of the keyword approach to capture the context surrounding the word. One example is the difficulty for simple keyword searches to reliably identify negations that contain a keyword (e.g., "this patent is not an AI patent"). In contrast, ML methods have the ability to capture more nuanced contextual information, including words that may be used in combination with others. In addition, recent developments in ML, such as text embeddings, can capture synonyms, antonyms, and other complex semantic relationships that humans can understand but which they have traditionally struggled to capture when using automated techniques.

Supervised ML-based text classification methods have specific advantages. First, because they "learn the importance of keywords," they are appropriate for settings in which a dictionary of keywords is not reliably defined. Second, recent developments in supervised ML-based methods allow for an integration of various other techniques, including leveraging the many benefits of *unsupervised* ML methods, as when creating a representation of concepts occurring

[3]Without a hand-coded sample, a dictionary approach is difficult to validate. However, if one has exerted the effort to create a hand-coded sample, then the ML method may be preferred, as it might help identify the relevant keywords from the dictionary.

[4]A keyword approach with a validation dataset is, in effect, a classification algorithm where the existence of a keyword is associated with an outcome of one. If this is the best possible classification approach, then the ML approach would reveal such a classifier. However, the ML approach may be able to identify a better classifier based on the same set of keywords, and this is validated through established validation approach.

TABLE 2 General comparison of keyword approach and ML methods in constructing categorical variables from texts

	Keyword-based approach	ML-based approach
Pros (benefits)	• Straightforward to implement if dictionary exists • Dictionary often transferable across contexts (text from news articles, patents, social media, etc. often contain the same keywords)	• Does not require researchers to specify a pre-existing dictionary, but instead need to only identify training data (ground truth) • Sophisticated ML techniques have the ability to capture contextual information • Established techniques, best practices and readily available software tools allow researchers to implement methods in a standardized, repeatable way
Cons (limitations)	• Accuracy is contingent on quality of dictionary, which is often subjective and difficult to validate • May require researchers to define dictionary, which may be costly and often ad-hoc • Context and nuances are difficult to capture (such as negation)	• Requires the construction of a training dataset (or ground truth/known values) which may require human coders
Example of when appropriate	• *Context with established dictionaries, vernacular*, etc. (e.g., *studying use of language in congressional debate*—Ridge et al., 2019)	• *Contexts where language for describing the concept is not well known or there may be inherent ambiguity* (e.g., *management communication style*—Choudhury et al., 2019)

in the textual data (e.g., Choi, Menon, & Tabakovic, 2021; Choudhury et al., 2019; Furman & Teodoridis, 2020; Kaplan & Vakili, 2015) and then classifying the data accordingly. Finally, supervised ML methods allow us to quantify the degree to which each method correctly classifies all observations; this enables us to compare the efficacy of the methods and, more importantly, to leverage additional statistical techniques to correct or account for any measurement errors. We elaborate on these features in subsequent sections of this paper.

We do not suggest that the ML approach is always superior to the keyword-based approach. Tradeoffs exist to using each approach. The keyword approach can be efficient (i.e., simple to implement and computationally fast) and interpretable (i.e., easy to understand why something belongs in a group). Well-established dictionaries of keywords can be defined to comprehensively represent a theoretically grounded concept and then reused across many studies. Additionally, researchers can construct a validation dataset to validate keyword-based approaches. In contrast, ML methods may be more appropriate when researchers lack access to a well-established dictionary of keywords and need to quantify the accuracy of their variable construction. However, implementing ML methods can be costly because they require large training datasets (much larger than what may be needed to validate the keyword-based approach). In Table 2, to further illustrate where ML methods may be useful to management researchers in text classification, we provide an overview of the relative advantages and limitations associated with each approach.

2.3 | Brief overview of key text classification procedures using supervised ML methods

In the context of ML methods, the term "supervised" refers to the use of training data, which are a set of observations for which the classification is already known. In practice, the term "training data" refers to a sample of observations in which researchers have already classified the text—often using human coders—into different groups.

There are four general steps in implementing the supervised ML approach in text classification. In Step 1, researchers construct the training data, a sample of classified data that will form the basis on which all observations will (eventually) be classified. In Step 2, the textual data are converted into numerical representations (i.e., features) that can be used by an ML algorithm; this can be achieved in multiple ways. In Step 3, the different classification approaches are tested using out-of-sample validation (e.g., a holdout sample) to identify the best-performing ML algorithm. Because it is difficult to know ex ante which type of algorithm may work best in a particular application, researchers may be best served by testing different ML algorithms and selecting the most accurate[5] in terms of out-of-sample (e.g., holdout sample) classification. Finally, in Step 4, the best-performing algorithm is applied to classify the remaining population of data. We will illustrate how we implemented these steps in Section 3.

3 | EMPIRICAL CONTEXT: STUDYING ARTIFICIAL INTELLIGENCE TECHNOLOGIES USING PATENT DATA

3.1 | Need for identifying artificial intelligence technologies in patents

Patents represent one of the largest repositories of technological knowledge in modern society over long periods of time, so they are often used as indicators of technological innovation. Strategy researchers have used patents extensively to study innovation outcomes at the individual, organizational, and societal levels (Fleming & Sorenson, 2004; Gambardella, Giuri, & Luzzi, 2007; Gans, Hsu, & Stern, 2008; Hall, Jaffe, & Trajtenberg, 2005; Hausman, Hall, & Griliches, 1984; Huang & Murray, 2009, 2010; Jain & Huang, 2022; Jia, Huang, & Zhang, 2019). The majority of patent-based metrics used by researchers are generated based on standardized fields in the patent documents, such as the patent technology classes, grant dates, and citations. In contrast, the bulk of patent documents, including the unstructured information in the main body of the text, is often underutilized because of challenges associated with making sense of unstructured data at the scale necessary for econometric analysis. However, information that would be useful for extraction exists in this area.

The study of patenting with regard to AI technologies provides a recent example of such issues. While patents constitute a useful indicator for these technologies (Goldbarb, Taska, and Teodoridis, 2019), those related to AI technologies are not easily identified. Existing patent classification systems typically contain some patent classes specifically denoting "AI technologies," but AI is often considered to be a general-purpose technology that cuts across many technological sectors and industries, thus encompassing many different patent classes (Goldfarb, Taska, & Teodoridis, 2019). Therefore, AI patents are likely distributed across many different technology

[5]This refers to the accuracy of the model in classifying observations that the algorithm has not previously encountered.

domains, including some general patent classes that are not specifically categorized as "AI technologies," making it difficult to examine a small selection of specific patent classes as a way to identify all patents on AI technologies.

Therefore, the use of supervised ML methods to classify patent documents based on the embedded text is an important practice. Additionally, identifying patents related to AI technologies aids researchers' understanding of AI technologies; this is another reason for our focus on this application as our empirical context.

3.2 | Importance of artificial intelligence for strategy

AI is poised to revolutionize many industrial processes, potentially reshaping value creation and value capture (Agrawal et al., 2018; Brynjolfsson & McAfee, 2014; Iansiti & Lakhani, 2020) and influencing both organizations and society in numerous ways. The importance of this phenomenon is reflected in the recent increase in research papers attempting to understand this and to use these tools (Goldfarb et al., 2019; Hartmann & Henkel, 2020; Shrestha et al., 2021; Tidhar and Eisenhardt, 2020; Tong, Jia, Luo, & Fang, 2021). However, while the broad label of "artificial intelligence" has proliferated (including its use in this paper as well as its title), it is imprecise and evokes a notion of either "the Terminator" or "the Jetsons."[6] In reality, AI—in practice—is a set of statistical prediction techniques used to simplify various processes. Agrawal et al. (2018) make this argument explicitly by referring to these AI technologies as "prediction machines." Many of the canonical papers that introduced the key concepts used in AI did not use the term "AI" explicitly but instead discussed specific ML techniques (Blei, Ng, & Jordan, 2003; LeCun, Bengio, & Hinton, 2015; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Therefore, it is valuable to unpack the wording actually used to refer to the concept of AI and, in particular, to determine the text features (e.g., words describing statistical techniques) that are used to refer to AI technologies. To achieve this goal, it is important to *let the data speak* and reveal patterns that researchers can then study.

As noted, patent data represent an important repository of applied knowledge that is continuously updated with the filing and granting of new patents. Therefore, in addition to classifying these patents for use in subsequent research, we can also use insights from the classification process to help us understand which technologies and concepts are associated with AI technologies, providing clarity for subsequent research.

4 | APPLICATION: IDENTIFYING PATENTS RELATED TO ARTIFICIAL INTELLIGENCE TECHNOLOGIES

In this section, we provide a walkthrough of how we applied supervised ML techniques to classify patent text documents. This is not intended to provide extensive background on ML techniques, as such information is widely available in numerous textbooks (e.g., James, Witten,

[6]"The Terminator or the Jetsons" was the title of a 2018 conference organized by the Technology Policy Institute regarding the future of AI technology. "The Jetsons" (from the TV show) evokes notions of AI and robotics that help simplify and automate everyday life. "The Terminator" (from the movie) evokes notions of an AI that disrupts or harms humanity. Conference information is accessible here: https://techpolicyinstitute.org/publications/economics-and-methods/terminator-or-the-jetsons-the-economics-and-policy-implications-of-artificial-intelligence/ (accessed October 12, 2021).

Hastie, & Tibshirani, 2013). Instead, we use this application of ML methods to: (1) demonstrate how we constructed the training data necessary for classification; (2) demonstrate how different supervised ML techniques can be used and compared to identify the best-performing model for a given study; (3) demonstrate how we can understand the performance of supervised ML methods by quantifying the out-of-sample classification performance; and (4) discuss the nuanced but critical judgment calls that researchers must frequently make when utilizing this approach in their own applications.

4.1 | Data

We sought to classify all patents issued by the United States Patent and Trademark Office (USPTO) from 1985 through 2018. *Each year* within this time frame, of all patent applications filed with the USPTO, approximately 650,000 were subsequently granted. Given the enormous size of the dataset, it would be cost-prohibitive to employ human coders to identify those patents related to AI technologies. However, patent texts can be quickly classified using a supervised ML method, provided that we construct a training dataset. We describe, in Section 4.2, the key procedures involved in implementing ML methods to identify patents.

4.2 | Implementation of ML methods to identify AI patents

4.2.1 | Step 1: Construction of training data: Diagnosis of size, balance, and representativeness

We constructed a training dataset of 4,000 USPTO patents. The training dataset was manually constructed by enlisting graduate student coders[7] to read through the patent abstracts and identify whether they were related to AI technologies. AI technologies were defined as those employing some form of "statistical learning" methodology. We were guided by the definition of AI used by Cockburn, Henderson, and Stern (2018), which provides the clearest empirical definition of the term and examples of the referenced technologies. We provided human coders with the textbook definitions of this paper, as well as a list of related technology terms (e.g. neural network, statistical learning, deep learning, etc.). Our approach to constructing training data, which has human coders read over text documents and identify whether techniques such as deep learning may be used but not mentioned by name, also augments the dictionary of Cockburn et al. (2018) by allowing human coders to identify observations that may use words such as "statistical learning," but are not terms such as "artificial intelligence." We provide more details in Online Appendix I.1.

AI technologies represent only a small subset of all patents. Therefore, it is important to select training observations appropriately. The first-order consideration is class balance. For an ML classification model to perform well (i.e., to classify out-of-sample observations correctly),

[7]We employed master's-level students in data science from the business school of a major US university under the supervision of professors in data science and management.

the training data should include a balance of both classes (AI and non-AI patents).[8] However, if we were to randomly sample from the entire dataset, AI patents are likely to account for only a small proportion of the training data, which may undermine our ability to achieve a good classification performance. We therefore adopted an "active learning" technique (Cohn, Ghahramani, & Jordan, 1996) to oversample the positive class (i.e., AI patents in our context).[9] We performed the classification with a small number of training observations, manually validated the predictions on a sample that oversampled the positive class, and then added these observations to the new training data. We repeated this process until we had training data of sufficient size (elaboration on this can be found in Online Appendix A). Our final dataset contained a sample of 20% AI-related patents out of all 4,000 patents.

We performed a set of validation diagnoses to ensure that this training data achieved (1) adequate size; (2) adequate level of balance, and (3) adequate representativeness. First, we traced the learning rate to demonstrate how the size of the training sample impacted the out-of-sample classification performance. As we increased our sample size from 500 to 3,000 observations, the out-of-sample classification performance (F1-score; we elaborate on F-1 scores in Section 5) improved from 0.83 to 0.925, which is sizable. However, increasing the training sample size from 3,000 to 4,000 observations increased the classification performance by only 0.005 to 0.93. We can connect these observations to create a curve and infer that, in adding a larger amount of data, we would gain only a miniscule benefit in classification performance. For example, doubling the number of observations of the training data could increase the F1-score by 0.03—from 0.93 to 0.96—but this requires considerable effort in terms of manually collecting 4,000 additional observations. A more detailed discussion on training data size and relevant results is reported in Online Appendix A.1.

As a second validation check, we estimated how classification performance changed with class balance. We observed that, with a low level of class balance, we generated high accuracy but a low recall metric—defined as the share of positive values in the training data correctly classified (see Online Appendix D)—as well as a high true-negative rate—defined as the share of correct predictions among observations that are classified as being negative (see Online Appendix D). Because of the small proportion of AI patents among all patents, if observations were overwhelmingly classified as non-AI-related, we could achieve high accuracy in identifying non-AI patents but at the expense of the accuracy of identifying AI patents (a high false-negative rate and thus a low recall metric). This is why we needed to increase the balance in constructing the training data. As we increased the balance, we observed an associated increase in the recall metric, indicating that we were more accurately classifying both classes (AI and non-AI patents). A more detailed discussion on class balance and the relevant results are reported in Online Appendix A.2.

As a final consideration, we examined the representativeness of our sample with respect to the population. Note that, because we oversampled the positive class, we did *not* want this

[8]Class imbalance (i.e., the unequal distribution of classes in the training data) presents a problem for supervised learning, because the algorithm will "learn" that it is optimal to classify all observations as belonging to the dominant class (i.e., an algorithm trained to identify a rare disease will learn that it can achieve an optimal classification accuracy if it identifies all patients as not having the disease). To correct for this, it is important to resample the data to ensure an equal distribution between the different classes. This was done in all of the cases.

[9]The intuition behind these methods is similar to the Rare Events Logit approach (King & Zeng, 2001), which may be more familiar to researchers.

training data to be mapped perfectly to the population. Instead, we wanted to ensure that the distribution of features (i.e., the keywords identified as important using the classification algorithm) were similarly distributed in the training and population data. We performed these checks to validate our sample and report the details in Online Appendix A.3.

4.2.2 | Step 2: Construction of numerical representation of text data

We adopted two methods for converting the text data into a numerical representation: (1) a "bag-of-words" approach and (2) an embedding-based approach.

The bag-of-words approach effectively involves counting the number of times each word occurs in each document, then using the count of these words (i.e., the term-document frequency) as a predictor in a classification model. The advantage of this approach is that it allows us to understand the impact of each word on the classification performance. The disadvantage is that the algorithm does not "understand" (i.e., capture) the context or word meaning. Therefore, two words with similar meanings will be treated as completely different if they contain distinct words. For example, "reinforcement learning" and "Q-learning" are typically used to refer to the same techniques but are semantically distinct, while two instances of the same word with different meanings are treated as the same (e.g., "online" can refer to either internet-based or interactive algorithms, which are distinct concepts).

The embedding-based approach involves first using an unsupervised method, which creates a textual representation of the data. For example, researchers can use the word2vec approach to create a representation of the words in the document (Mikolov et al., 2013). This algorithm predicts a specific word in relation to its neighboring words and provides a numerical representation of this word in the form of a vector of common neighboring words. Thus, words with similar meanings have geometrically close vectors, while words with different meanings have geometrically distant vectors, taking into account the context issues described above.

However, this approach requires an additional step: creating text embeddings. Text embeddings can be created by the research team—for example, in this case, by classifying the population of patent abstract documents. Alternatively, it is possible to download embeddings that have been trained on large corpuses (e.g., Global Vectors for Word Representation [GLoVe] embedding) and that are often freely released by large technology companies. These provide a more reliable representation of the general understanding of language because these companies have developed the representations based on very large samples of text, ensuring a large number of instances of words in a particular context. In some cases, we may want to access both a text classification model and the embeddings trained by a third-party with access to large datasets and computing power, then fine-tune this to our particular training data. This is the logic behind transformer models, which have recently become the leading tool for language understanding and text classification (Devlin, Chang, Lee, & Toutanova, 2018). The disadvantage of these embedding-based approaches is that they are less transparent in terms of how each keyword (feature) is mapped with each outcome. Additionally, any biases or issues in the embedding data might be projected into the final text classification.

We utilized both bag-of-words and embedding-based approaches and compared their effectiveness in terms of their out-of-sample classification performance (Step 3, discussed next).

4.2.3 | Step 3: Comparing classification models (k-fold cross-validation)

We trained each classification algorithm based on 80% of the training data (3,200 observations). We then evaluated the performance of each algorithm using a validation sample of 20% (800 observations). Following this, we used k-fold cross validation by repeating this process five times for each classification—that is, using k-fold cross validation, we randomly split the data into k (in our case, five) sub-samples and repeated the classification by alternating the validation sample. This approach ensured that our choice of holdout did not influence our model performance. K-fold cross validation is a well-established and more advanced way of constructing holdout samples because it is more robust than the conventional approach of using one holdout sample (e.g., James et al., 2013).

Specifically, 80% of the training data was used to train the classification model. Then the trained model was used to predict the outcomes for the remaining 20% of training data (as if we did not know the true values of the training data), the result of which was then used for comparison with the actual known classification of this remaining 20% of training data. We used multiple classification models; an overview is provided in Online Appendix C and we elaborate upon some in Section 5. Additionally, we discuss of how our selection of features affects the performance of classification in Online Appendix B.

Comparing the predicted and known values for the validation sample allows us to calculate how well a classification algorithm performs in terms of making out-of-sample predictions. The common metrics used to determine out-of-sample classification performance include accuracy, precision, recall, F1-score, and area under ROC curves (AUC), which we define in Online Appendix D (and also discuss in Section 5). The above classification procedure was repeated for different classification algorithms. Algorithms were compared based on the out-of-sample (20% validation sample in k-fold cross validation) classification accuracy and false-positive rate.

As an additional check, we used another holdout sample of 100 observations on which the model was never trained to validate classification performance. The classification results achieved on this additional holdout sample were comparable (results available upon request).

4.3 | Access to data and codes

In this paper, we focus specifically on a high-level rationale for our implementation approaches. In the accompanying code in the online workspace, we provide a walkthrough of the code we used; in the online appendices, we discuss a rationale for the different choices we made regarding key steps in the code. Notably, in our online workspace researchers can access all data, codes, and explanations of these codes and can use a pre-installed Python program running on advanced, high-performance GPU/TPU processors that may be necessary for models with particularly complex embeddings. These processors are made freely available by large technology companies for this purpose; otherwise, researchers need to acquire their own processors. Researchers can replicate our data and codes and/or can modify them with their own data and codes, either in the online workspace or on their own local drives. Finally, this online workspace allows us to provide more details and updates on the technical information behind these approaches, on how the code can be implemented, and on resources for the statistical foundation of these models.

5 | EVALUATING TEXT CLASSIFICATION APPROACHES

In this section, we discuss the performance of different classification methods based on two sets of key metrics. The first key metric is how closely the out-of-sample predicted values match the known values (classified by human coders) of the holdout training data, as described above. To compare the two datasets, we used several established classification metrics, including accuracy, precision, recall, F1-score, and AUC; the definitions of these are included in Online Appendix D. We also used the false-positive rate (i.e., misclassified negative values) and false-negatives rate (i.e., misclassified positive values). Finally, as a reference point, we demonstrate in Section 5.1 how keyword-based metrics compare as an alternative (e.g., to the keyword-based approach), which would be researchers' choice in the absence of ML methods.

5.1 | Comparison of model accuracy of ML classification approaches

In Table 3, we present a comparison of the different classification approaches. Group A (Row 1, highlighted) indicates the performance for manually classified training data. This information is included to demonstrate how metrics with zero misclassification should look. For instance, a perfect classification model would yield zero false positives as well as a value of one for true-negatives, true-positive, accuracy, precision, recall, F1-score, and AUC.

The performance of various ML classification models is reported in Group B (Rows 2–15) of Table 3. The models are listed in descending order based primarily on classification accuracy (share of correct predictions) and, secondarily, on the false-positive rate (share of non-AI patents wrongly predicted to be related to AI). Note that we only included representative models in Table 3; see Online Appendix E for a complete list with all possible models. The random forest model (Row 2, highlighted) was the best-performing ML classifier, with a classification accuracy of 96% (the highest of all ML methods) and a false-positive rate of 6% (the lowest of all classification methods). A more subjective metric was the share of AI patents among all patents. The random forest model calculated a proportion of AI patents (18%) that was comparable to the share of AI patents found in the training data (20%), thus helping to confirm its performance.

A comparison of the performance of the ML models following the bag-of-words approach and with those following the text-embedding approach (mentioned above) generated interesting results. Note that transformer-based models, the results of which are highlighted in Table 3, were among the best-performing models in general, whereas the random forest model was the only bag-of-words model among the best-performing models. It is encouraging that the most sophisticated models—such as BERT and SciBERT, found to be the best-performing in other settings—were also among the best-performing in our context. This comparison exemplifies the importance of testing and comparing a large number of different models, because it is difficult to ex ante predict which will perform the best.

5.2 | Comparison of performance of ML classification methods and keyword-based methods

In Group 3 of Table 3 (Rows 16–18), we show the performance of the keyword-based methods; based on this, we discuss how they compare with ML classification methods. Fortunately, for

TABLE 3 Comparison of classification results using different classification approaches

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
		Confusion matrix				Classification performance metrics				
	Share AI patents	*True-positives*	*True-negatives*	*False-positives*	*False-negatives*	*Accuracy*	*AUC*	*Precision*	*Recall*	*F1*
(A) Hand-coded data (ground truth)										
1. *Training sample*	0.20	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00
(B) ML-based classification										
2. *Random Forest*	0.18	0.94	0.97	0.06	0.04	0.96	0.92	0.94	0.86	0.90
3. *SciBERT*	0.20	0.88	0.97	0.12	0.03	0.95	0.93	0.88	0.88	0.88
4. *PatentSBERTa*	0.21	0.86	0.98	0.14	0.03	0.95	0.93	0.86	0.90	0.88
5. *BERT*	0.22	0.84	0.97	0.16	0.03	0.94	0.93	0.84	0.89	0.87
6. *XLNet*	0.22	0.83	0.97	0.17	0.03	0.94	0.93	0.83	0.90	0.87
7. *Support Vector Classifier w. RBF kernel*	0.17	0.91	0.95	0.09	0.05	0.94	0.89	0.91	0.79	0.85
8. *RoBERTa*	0.22	0.82	0.98	0.18	0.02	0.94	0.93	0.82	0.91	0.87
9. *CNN w. Doc2Vec Patent Embeddings*	0.20	0.84	0.96	0.16	0.04	0.94	0.91	0.84	0.85	0.85
10. *ELECTRA*	0.20	0.84	0.96	0.16	0.04	0.94	0.91	0.84	0.85	0.85
11. *Big-Bird*	0.23	0.80	0.97	0.20	0.03	0.94	0.92	0.80	0.90	0.85
12. *Logistic regression*	0.23	0.80	0.97	0.20	0.03	0.93	0.92	0.80	0.90	0.84
13. *CNN w. GloVe Embeddings (840B – 300)*	0.19	0.83	0.95	0.17	0.05	0.93	0.88	0.83	0.81	0.82
14. *Naïve Bayes*	0.34	0.55	0.98	0.45	0.02	0.83	0.88	0.55	0.95	0.70
15. *Nearest Neighbors*	0.35	0.48	0.95	0.52	0.05	0.79	0.80	0.48	0.83	0.61

TABLE 3 (Continued)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
		Confusion matrix				Classification performance metrics				
	Share AI patents	***True-positives***	***True-negatives***	***False-positives***	***False-negatives***	***Accuracy***	***AUC***	***Precision***	***Recall***	***F1***
(C) Keyword-based classification										
16. Cockburn et al. (2018)	0.09	0.97	0.87	0.03	0.13	0.88	0.71	0.97	0.43	0.59
17. *UK Govt AI Classification*	0.23	0.39	0.86	0.61	0.14	0.75	0.64	0.39	0.46	0.42
18. *WIPO AI Patent Classification*	0.08	0.66	0.84	0.34	0.16	0.82	0.62	0.66	0.27	0.38

Note: *All statistics based on out-of-sample (or holdout/validation sample) classification accuracy.* Out of sample prediction performed with fivefold cross-validation. Definitions of classification statistics such as precision and recall are reported in the Data S1. *Hand Coded Data (Ground Truth)* reported for illustrative purposes to provide reader with a clear indication of how a perfectly functioning classifier should perform. Classification performance is evaluated with respect to how well the model performs in achieving a classification rate comparable to these methods. *ML Classifiers*: Classification results for various ML classifiers are reported. Grid search used to optimize hyper-parameters for each model, then results reported for each type of classification model. *Results sorted (reported in order) by model with the highest prediction accuracy, then the model with the lowest false positive rate.* We believe that this is representative of how researchers care about selecting the models that they use (to maximize predictive accuracy and minimize the rate of false positives). Note that we calculated the metrics based on the share of values predicted to be positive/negative, instead of the share of correctly classified positive/negative observations as is customary, becaue this is again more aligned with what applied researchers care about (e.g., the share of incorrect values among those predicted). Regarding *Keyword-Based Classification*, please note the following. We constructed our sample based on active learning, identifying observations to balance the classes, but this process also enables us to identify observations that are meaningful. Therefore, our training data involved observations where the keyword-based classification would not perform well, as these observations were informative for the ML classifier. However, this means that the fact that the keyword-based classification performs poorly in this training sample is not a reflection of how well it would perform in the population. In a random sample of the population, the classification performance would be higher for the keyword-based methods, and likely also for the ML methods. We provide more discussion in Appendix F. *Gray highlights indicate the observations discussed in the paper*.

identifying AI patents, several comparable, authoritative dictionaries have been published. First, in academic publications we found keywords developed by senior scholars with topic expertise. Cockburn et al. (2018), in a chapter from the National Bureau of Economic Research publication *The Economics of Artificial Intelligence: An Agenda*, presented a dictionary for classifying AI patents. Second, in 2019 the UK Government Intellectual Property Office published a study of AI patents that included another dictionary for identifying such patents.[10] This agency has considerable expertise (equivalent to that of the USPTO in the United States) in this domain, so its dictionary may serve as a reasonable alternative set of keywords. Third, the World Intellectual Property Organization (WIPO), a specialized agency of the United Nations with the sole focus of promoting and protecting intellectual property rights—including patents—studied the growth of AI patents and published a dictionary for identifying AI-related patents (WIPO, 2019).[11]

Among these three keyword-based classification approaches, the keywords provided by Cockburn et al. (2018) yielded the highest classification accuracy (88%) as well as a low false-positive rate (3% of positive values). The WIPO and UK Government classifications yielded slightly lower rates. Because our training data were constructed based on active learning, we identified observations likely to be misclassified and included these in the training data so that the ML method would take them into account. Therefore, the keyword-based approach would likely perform better when tested on a sample that mirrored the population (as would the ML approach).

When we compared the keyword-based and ML classification metrics, we found that, on average, the ML methods achieved approximately 10% higher accuracy rates. In particular, on average, the ML approach yielded a much lower share of false positives, except in situations where the overall rate of classified values was very small (see Online Appendix F).

It is important to acknowledge that by performing this comparison we are not attempting to criticize or suggest that the text-based dictionaries are invalid. The validation was performed on the training sample which was constructed based on active learning and therefore contains many of the observations that were misclassified by the keyword-based approach. Therefore, the values in the training sample yield a much lower classification performance, particularly in terms of false positives, than what the classification of the keyword-based approach would be in the population. We want to stress, that this should not indicate that the keyword-based classification; Cockburn et al. (2018), WIPO and UK Govt classifications perform as poorly as these results suggest. Instead, these results represent a lower bound for the observations that are likely misclassified.

6 | EXAMINATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES BASED ON PATENT DATA

6.1 | Using ML model outputs to understand artificial intelligence

A common critique of ML approaches is that they seem like "black boxes" because of researchers' limited ability to interpret and understand the factors and underlying calculations

[10]Available at: https://www.gov.uk/government/publications/artificial-intelligence-a-worldwide-overview-of-ai-patents (accessed October 9, 2021).
[11]Available at: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf (accessed October 9, 2021).

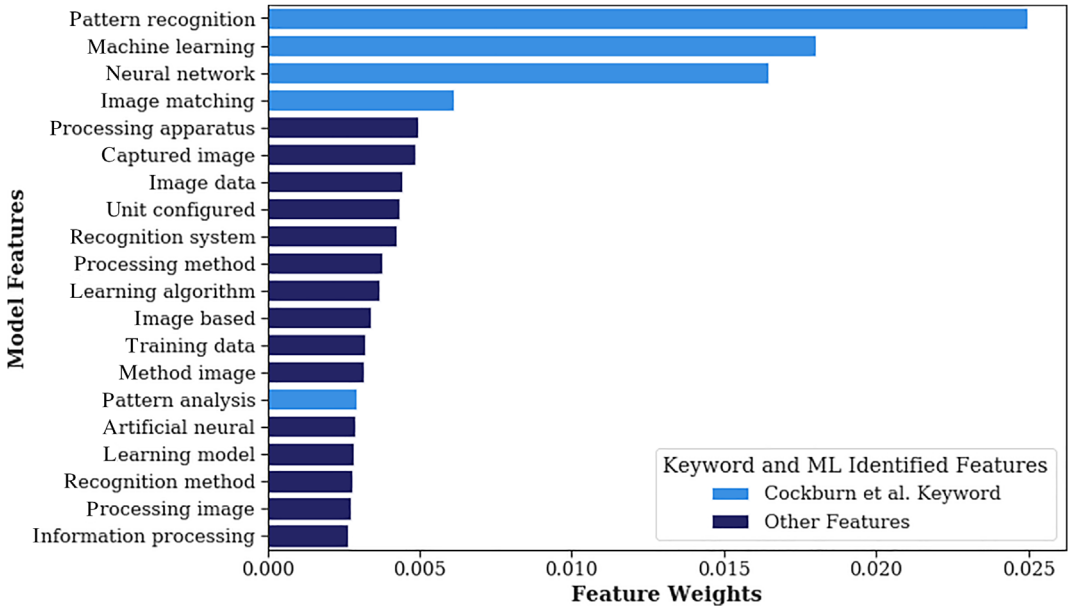

FIGURE 1 Feature weights generated by random forest classifier

that drive the approach. However, in reality, many ML algorithms provide "factor loadings," which explain the degree to which individual "features" are used to predict the outcomes. These loadings are similar to the estimated coefficients of variables in linear regression models in which researchers can compare different coefficients to understand the degree to which a given independent variable affects a particular dependent variable. This information is particularly accessible for bag-of-words models—such as the random forest classifier—because the inputs are the counts of the keywords.

In Figure 1, we show the top 20 features (bi-grams) for predicting AI patents using the random forest classifier, the best-performing ML classifier in our context. These features are ranked in descending order of "feature importance," which reflects the explanatory power of a given feature. We highlight (lighter color/shaded bars) features that overlap keywords identified by Cockburn et al. (2018). This comparison generated the following interesting insights.

First, the four features identified by the random forest classifier as the top predictors of AI patents—*pattern recognition*, *machine learning*, *neural networks*, and *image matching*—were also included in the dictionary provided by Cockburn et al. (2018). This suggests that the ML approach and the keyword approach overlap in identifying the most common and established word features that distinguish AI patents from non-AI patents. However, substantial divergence was found with keywords that occur less frequently than the above. Other predictors that were highly important according to the random forest approach—such as *image data*, *training data*, and *learning algorithm*—were not found in the keyword dictionary. Overall, of the top 20 features identified by the random forest approach for predicting AI patents, five (the four stated above and *pattern analysis*) were found in the keyword dictionary provided by Cockburn et al. (2018), while the other 15 were not. This outcome may have occurred because the keyword approach missed many patents not containing obvious keywords, likely because of the novelty and evolving language used to describe more nascent AI technologies.

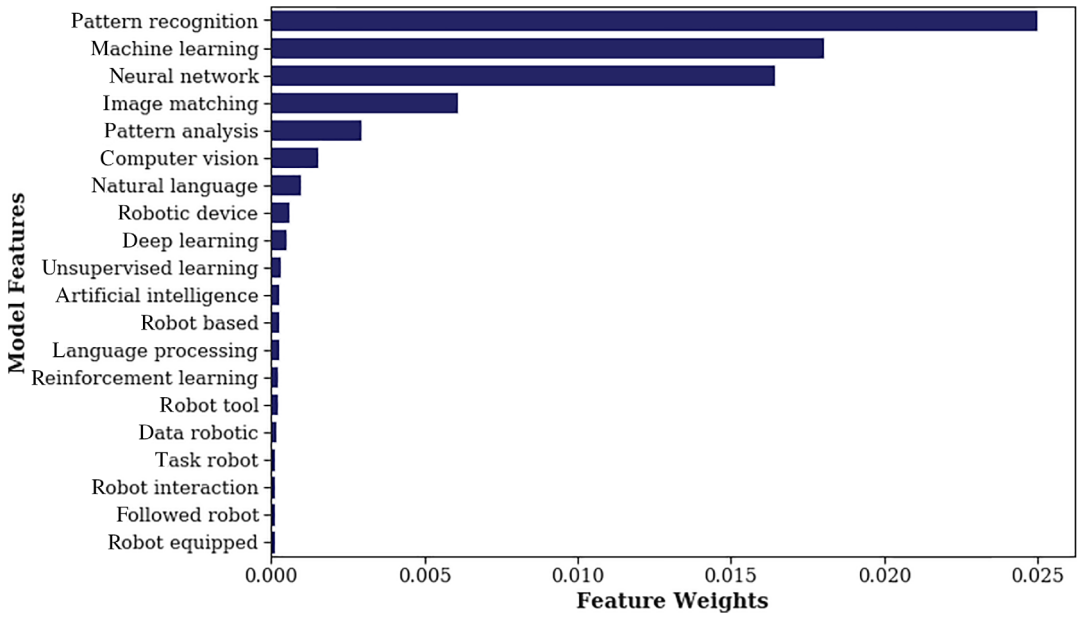

FIGURE 2 Feature weights generated by random forest classifier for keywords used by Cockburn et al. (2018)

To further probe the divergence of the results generated by the approaches, Figure 2 shows the "feature importance"—or predictive power—of the keywords provided by Cockburn et al. (2018) to identify AI patents; this feature importance was estimated by the random forest classifier. We ranked these keywords in descending order in terms of feature importance. The results suggested that, other than the top four or five keywords discussed above, the remaining keywords had very limited predictive power. For instance, interestingly, features such as "artificial intelligence" were very poor predictors of AI-related patents. This would have been difficult to confirm without the training data. As discussed above, studies developing AI technologies—including those that develop fundamental tools for AI technologies—frequently do not use this general label (Blei et al., 2003; LeCun et al., 2015; Mikolov et al., 2013). Instead, these studies tend to use specific terms that pertain to the technology being developed, such as "distributed representation" or "deep learning."

While researchers could debate the overlaps and differences in classification outcomes generated by the ML methods and keyword-based methods, this comparison highlights a general limitation of keyword-based approaches: it is difficult to quantitatively assess their validity without using training data.

6.2 | Patterns and AI patenting overview

The AI domain has attracted significant academic attention. Using our data and classification approach, we describe how this technology has developed over time.

First, we present the overall share and number of AI-related patent applications in the USPTO database from 1985 through 2018 in Figure 3. We can see that AI was not a particularly prominent technology prior to the mid-2000s but, in the past decade and a half, has grown to

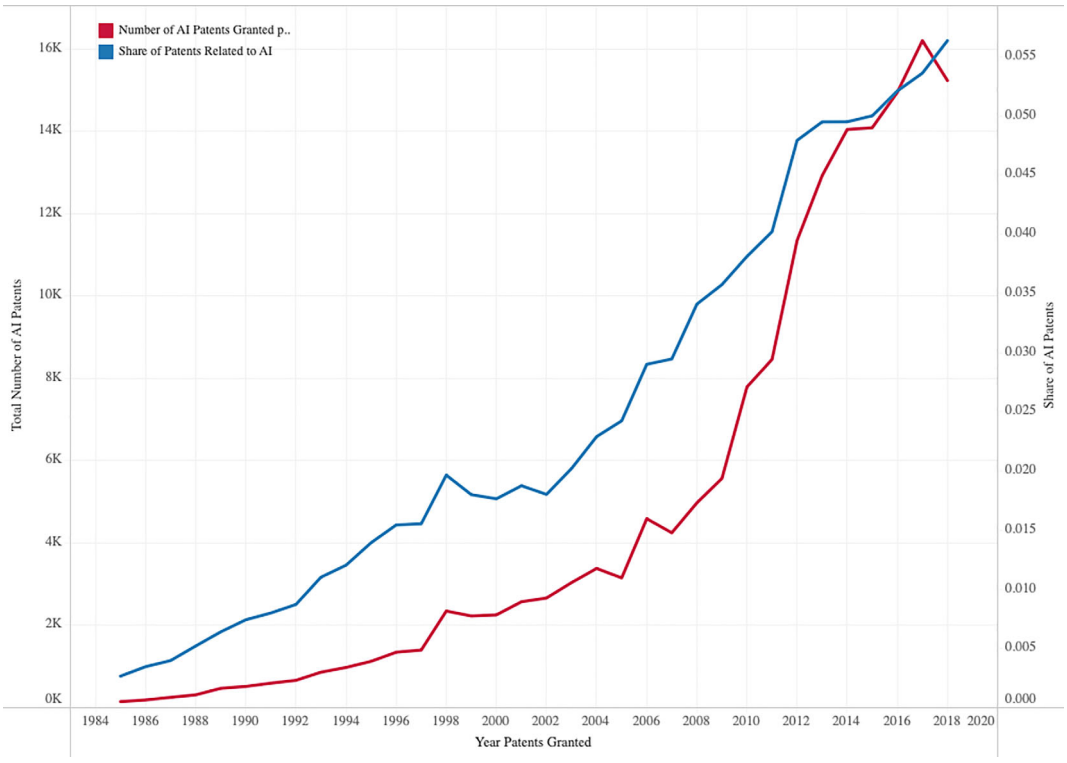

FIGURE 3 Number of AI patents and share of AI patents among all patents granted by year in USPTO database from 1985 to 2018

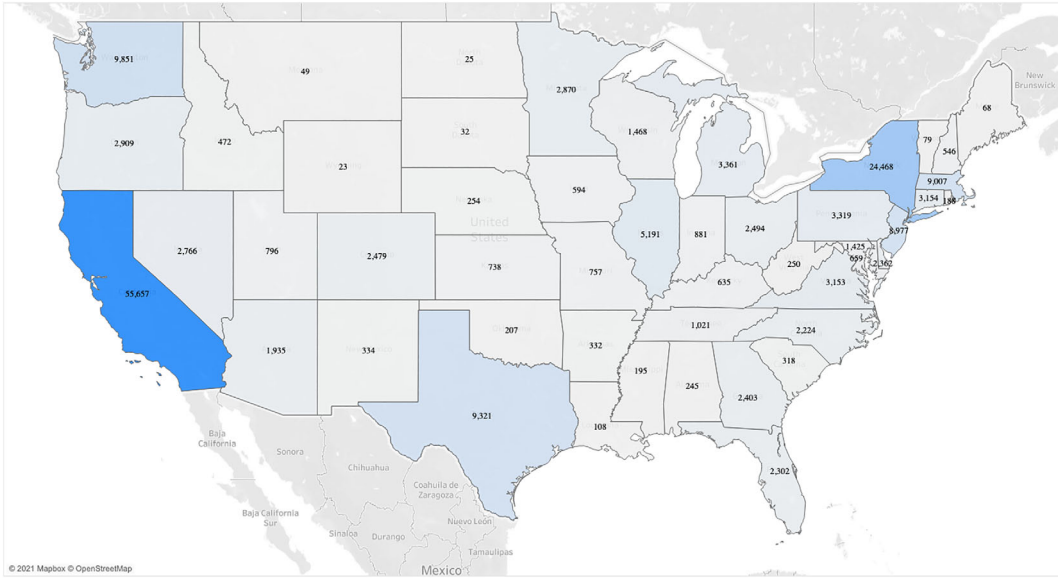

FIGURE 4 Number of AI patents granted by location of patent assignees (state level)

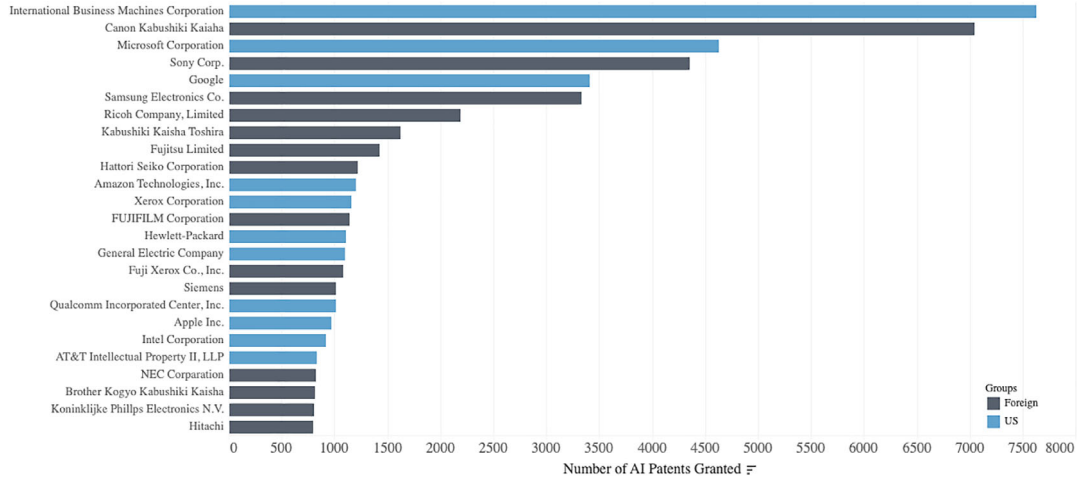

FIGURE 5 Top patent assignees by number of AI patents granted. We report the number of patents related to AI technologies that are granted from 2005 to 2018. We chose this period, as we are interested in the modern AI techniques (versus traditional techniques which did not really gain traction in practice). We distinguish US firms from foreign firms, as it is helpful to acknowledge that many of these technologies may be patented in the US but developed outside the country or the property rights owned by foreign entities. We see many of the well-known technology companies that are present among the top patent assignees, including Google, IBM, Microsoft, Apple, and Intel. We also see many prominent Japanese companies, particularly in the domain of imaging, including Canon, Ricoh, Sony, and Fujifilm.

represent more than 5.5% of all patents granted in the United States. This corresponds to approximately 15,000 AI-related patents granted in 2018.

Second, we present the data based on the geographic location of the patent assignees (Figure 4). "Patent assignee" refers to the entity that owns the intellectual property right of a patent. We observed that the largest share of AI patent-holders were located in the states of California, New York, and Washington, followed by Texas and Massachusetts. This pattern might be expected, as these states are home to many prominent technology companies and geographic hubs (e.g., Washington is home to Amazon and Microsoft). We provide a more detailed breakdown of the geographic locations in the online interactive dashboard, which includes visualizations of AI patents at the county level.

Third, we examine the top assignees in terms of the number of AI patents granted (Figure 5). We observe that large technology companies (e.g., Microsoft, Amazon, Google, and Apple) were featured prominently as holders of AI patents. However, many prominent technology companies (e.g., Canon, Sony, Ricoh, and Fujifilm) are in the imaging domain because AI technologies are often used to manipulate or analyze image data. For instance, in the radiology domain, where AI technology is often touted as a replacement for human medical professionals, the underlying technology imports, manipulates, and classifies images. Additionally, much innovation in digital imaging technologies—such as digital/phone cameras, text imaging software, and video manipulation software—is within the technological domain of these companies. It is worth noting that the vast amount of AI technologies is patented by IBM, which is often featured as a leader in this domain. For example, in its online profile, IBM states that, "in the field of AI, our researchers received more than 2,300 patents. To take two examples among

many: a novel way to search multilingual documents using natural language processing, and an ultra-efficient system for transferring image data taken by an on-vehicle camera."[12]

These patterns offer insight into the growing prominence of AI technologies and the dominant areas and technologies within which these tools have become commonplace. For further study of this data, we provide interactive visualizations in the online interactive dashboard (www.aipatents.xyz). We intend to keep this data updated.

7 | DISCUSSION: IMPLICATIONS FOR STRATEGY RESEARCH

This study demonstrates the use of supervised ML techniques to classify unstructured text data. This approach can generate categorical variables that researchers can then use in subsequent work, such as regression analysis. The traditional approach to classifying text data relies on predetermined dictionaries to identify texts containing specific keywords. Because AI technologies evolve rapidly, the traditional keyword-based approach faces an obvious challenge of not able to capture recent additions to AI technologies. Additionally, while "supervised methods" refers broadly to techniques that rely on training data to develop classifiers and validate the model, recently developed techniques can allow researchers to access a wide range of other resources—such as pre-trained transformer models and text embeddings—to overcome issues in a way not possible with alternative approaches.

This paper makes several contributions. First, we situate this paper in the burgeoning literature describing the use of ML methods to achieve a variety of goals in the research domain of strategy; we do so to help readers distinguish this study from and relate it to existing publications. Second, we provide a general overview of the research applications of supervised ML methods and their differences from traditional methods (i.e., the keyword-based approach). We also highlight the key procedures involved in implementing ML-based classification methods. We highlight the key benefits of supervised ML methods; in particular, they allow researchers to evaluate the performance of the classification algorithm and compare it with different methods. Moreover, we compare the classification results generated by bag-of-words representations, embedding-based representations, and transformer-based models, while also including comparisons with keyword-based methods as a reference point.

Finally, we demonstrate the core reason that supervised ML models perform better than keyword-based models. Specifically, we find that the top four "key text features" captured by random forest (the best-performing classifier in our context) were also included in the dictionary of Cockburn et al. (2018), demonstrating that both the ML and keyword-based methods can capture the most commonly used descriptions of AI technologies. However, that dictionary did not include many highly predictive "key features" identified by the ML method, and the remaining keywords in the dictionary had low predictive power in regard to the AI patents. Adopting supervised ML models allowed us to understand which text features were predictive of AI technologies. Concepts such as "neural networks" or "pattern recognition" were highly indicative of AI technologies; general terms such as "artificial intelligence" were not highly indicative of AI technologies. Therefore, future studies would benefit from a clearer conceptualization of AI and from a focus on this set of narrowly defined technologies, as strategy scholars have previously argued (Agrawal et al., 2018).

[12]Available at: https://www.ibm.com/blogs/research/2021/01/ibm-patent-leadership-2020/ (accessed October 9, 2020).

7.1 | Contribution and implications for research methods: Supervised and unsupervised ML methods

Empirical researchers increasingly use a wide range of quantitative tools, including sophisticated statistical estimation techniques, text analysis tools, ML, and estimation tools. As discussed in Section 2, most studies addressing text data have used *unsupervised* methods (Teodoridis et al., 2020; Kaplan & Vakili, 2015). We demonstrate how *supervised* ML tools can be applied to text analysis to classify text data and generate variables that can be used in subsequent analysis.

Thus far we have contrasted supervised ML methods with unsupervised ML methods (e.g., topic modeling). However, many supervised ML methods increasingly integrate both supervised and unsupervised approaches. Taking embedding-based methods as an example, in Section 4 we discuss their reliance on unsupervised methods (e.g., topic modeling and word2vec) to create a representation of textual data. This representation takes into account the context (e.g., different words with similar meanings and the same words with different meanings) and can be trained on a broad dataset (e.g., GLoVe embeddings are trained on a large corpus of news and Wikipedia articles) or on a highly specific dataset (e.g., training on Semantic Scholar, the USPTO database, or Pub Med). Transformer models take this step even further, as they are pre-trained on a very large dataset compiled by large technology firms that make them freely available for download by researchers, who can then fine-tune them for their specific datasets.

Therefore, we should not assume that supervised methods are unable to integrate the benefits of unsupervised methods, even though the two methods are appropriate for different situations. Specifically, supervised methods are designed for situations with a defined set of classes (e.g., AI vs. non-AI) and the resources to create a training dataset. However, researchers may still use unsupervised methods to create a representation of the text data, then use supervised methods to classify the data. Supervised methods enable researchers to construct metrics of classification performance, which allows us to understand how well these methods perform.

ML methods inherently introduces some degree of error; this can be seen as a feature, rather than as a limitation. Using supervised ML methods allows us to generate a quantifiable measure of the degree of error. Additionally, there are methods that take into account measurement error to quantify its impact on subsequent conclusions. Recent papers by Yang, Adomavicius, Burtch, and Ren (2018) and Qiao and Huang (2021) provide easily implementable methodologies that can be used to correct for classification bias in terms of using supervised ML-based variables, both as explanatory variables and as outcome variables.

7.2 | Contributions and implications for strategy research: ML methods and AI technologies

The insights generated in this study contribute to the growing conversation in strategy research regarding when and how to implement ML methods for research purposes (e.g., Choudhury et al., 2019; Menon et al., 2019; Shrestha et al., 2021). For researchers who aim to map unstructured text data with preexisting theoretical concepts, this paper demonstrates how to implement ML methods to achieve this goal. Moreover, as researchers want to know whether switching from the familiar keyword method to ML methods would be worthwhile, this paper provides

general qualitative guidance on the tradeoffs of each method and specific directions for quantitatively evaluating the performance of these methods.

The context of AI is also of interest to strategy and innovation researchers. The development of AI technologies is poised to transform many aspects of business and society (Agrawal et al., 2018); thus, a rapidly growing body of research in strategy and economics aims to understand the nature of technology and its consequences for production and business operations (e.g., Agrawal et al., 2018; Cockburn et al., 2018; Felten, Raj and Seamans, 2021; Goldfarb et al., 2019). Accurately identifying AI technologies is crucial to achieving these research objectives and paves the way for future research.

As argued above, the term "artificial intelligence" is broad and somewhat misleading. Our results shed light on the underlying technologies, labels, and keywords inherently associated with AI and provide greater precision and clarity. This outcome, of course, depends on our training data, and, as we—and other scholars using our online workspace—collect more data with the filing and granting of new AI patents, this classification can be repeated to continuously generate insights about how these concepts may change over time.

We do not claim our study is the first to apply ML methods in identifying AI patents; a recent report generated by the USPTO also adopts this approach.[13] Instead, our goals are to demonstrate the value of ML methods for strategy researchers and to show that AI patents constitute an appropriate empirical context for us to quantitatively assess ML methods and compare them with the keyword-based approach. Moreover, our online workspace illustrates the process of classifying AI patents, enabling easy replications and/or extensions of the data and results as new methods/techniques become available. We provide more detailed comparisons between our dataset and the USPTO classifications in Online Appendix G.

7.3 | Implications of studying artificial intelligence for theories of strategic management

As Felten et al. (2021) argue, AI is important because it may impact various fields of study. First, much AI literature emphasizes whether technology will assist or displace workers, or more poignantly, which workers and skills will be augmented, enhanced, or replaced (Acemoglu & Restrepo, 2019; Felten, Raj & Seamans, 2021; Goldfarb et al., 2019; Yilmaz, Naumovska, & Aggarwal, 2022). This discussion generates implications for the changing value of human capital, particularly as individuals shift their skills (Horton & Tambe, 2019) or exploit changes in the value of human capital to move between employers (Jain & Huang, 2022; Miric & Ozalp, 2022). Therefore, a comprehensive knowledge of the (changing) knowledge stock of AI, as captured in our data, may assist researchers of human capital in understanding shifts in human capital investment.

Moreover, AI technology enables the development of subsequent innovations (Cockburn et al., 2018). Griliches (1957, p. 502) describes it as "invention of a method of inventing." This means that AI technologies are likely to influence the development of a variety of other technologies. For example, AI has been introduced to distinct technological sectors such as autonomous vehicles, radiology, software services, and exploratory scientific research. As the technology within these industries is shaped by AI development, evolution of these industries

[13]"Inventing AI: Tracing the Diffusion of Artificial Intelligence with U.S. Patents," available at https://www.uspto.gov/sites/default/files/documents/OCE-DH-AI.pdf (accessed October 9, 2021).

could be impacted by AI. The use of patent data in AI-related technologies enables researchers to examine the evolution of these strategies (e.g., Goldfarb et al., 2019).

Further, patent data have traditionally been used to study knowledge flows, collaboration patterns, and knowledge recombination processes (e.g., Griliches, 1957). While these important issues have not yet been widely examined in the context of AI, recent evidence suggests there exist important knowledge flows and worker movements in AI-related technologies, particularly from universities to companies (Jurowetzki, Hain, Mateos-Garcia, & Stathoulopoulos, 2021). Datasets that provide information about AI-related patents will help researchers document the transformation of knowledge and the evolution of knowledge creation in AI-related technologies.

7.4 | Limitations and future research

Two limitations of this study yield potentially fruitful avenues for future research. First, while we examined patents with respect to AI technologies, we recognize that data on technological innovations may be stored in other locations or might not be patented. For example, Cockburn et al. (2018) also used the keyword approach to identify AI-related technologies in academic publications. To more comprehensively analyze AI technologies, additional study is needed that uses ML methods to analyze other repositories of data. Second, we have focused on the broad category of AI; others may be interested in a more narrowly defined definition of AI or in more recent patents. We share our classification algorithms and training data so others may apply this approach to different contexts or add new training data as more AI technologies are patented. We intend to update the online workspace that contains the repositories of this data and these insights so they remain relevant with the development of these technologies.

The AI patents classified here represent work from many years, wide geographic regions, and diverse technological classes. The richness of this dataset, which we only use to present basic patterns of AI technology development, will enable future scholars to examine a myriad of questions related to the development of AI technologies. Moreover, the online workspace through which we share data and codes will provide researchers with the computational capacity to extend and update the dataset of AI patents from the USPTO and to adapt ML methods to other research contexts and goals.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of the Institute for Outlier Research (iORB). The authors are grateful to the helpful comments provided by Vishal Gupta in earlier version of the paper. Excellent research assistance was provided by Jiatong Zhong, Jinhong Lu, Yutong (Eva) Tang and Rushan Zhang. The authors are grateful to the editor and two anonymous reviewers for their help and guidance throughout this process.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [insert provided URL from Open Research Disclosure Form].

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://www.aipatents.xyz.

ORCID

Milan Miric https://orcid.org/0000-0002-8318-9288
Kenneth G. Huang https://orcid.org/0000-0002-4811-5638

REFERENCES

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, *33*(2), 3–30.

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Boston, MA: Harvard Business Review Press.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, Progress, and prosperity in a time of brilliant technologies*. New York, NY: WW Norton & Company.

Cho, T. S., & Hambrick, D. C. (2006). Attention as the mediator between top management team characteristics and strategic change: The case of airline deregulation. *Organization Science*, *17*(4), 453–469.

Choi, J., Menon, A., & Tabakovic, H. (2021). Using machine learning to revisit the diversification-performance relationship. *Strategic Management Journal*, *42*(9), 1632–1661.

Choudhury, P., Allen, R., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57.

Choudhury, P., & Kim, D. Y. (2019). The ethnic migrant inventor effect: Codification and recombination of knowledge across borders. *Strategic Management Journal*, *40*(2), 203–229.

Choudhury, P., Starr, E., & Agarwal, R. (2020). Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, *41*(8), 1381–1411.

Choudhury, P., Wang, D., Carlson, N. A., & Khanna, T. (2019). Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles. *Strategic Management Journal*, *40*(11), 1705–1732.

Cockburn, I. M., Henderson, R., & Stern, S. (2018). The Impact of Artificial Intelligence on Innovation. *in* The Economics of Artificial Intelligence: *National Bureau of Economic Research Conference Report*.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Working Paper*. arXiv:1810.04805.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. (2007). A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods*, *10*(1), 5–34.

Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, *42*(12), 2195–2217.

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*(8), 909–928.

Furman, J., & Teodoridis, F. (2020). Automation, research technology, and Researchers' trajectories: Evidence from computer science and electrical engineering. *Organization Science*, *31*(2), 330–354.

Gambardella, A., Giuri, P., & Luzzi, A. (2007). The market for patents in Europe. *Research Policy*, *36*(8), 1163–1183.

Gans, J., Hsu, D., & Stern, S. (2008). The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays. *Management Science*, *54*, 982–997.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574.

Goldfarb, A., Taska, B., & Teodoridis, F. (2019). Could Machine Learning Be A General-Purpose Technology? Evidence from Online Job Postings. *Working paper*.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, *25*, 501–522.

Gross, D. P. (2020). Creativity under fire: The effects of competition on creative production. *Review of Economics and Statistics*, *102*(3), 583–599.

Guzman, J., & Stern, S. (2020). The state of American entrepreneurship: New estimates of the quantity and quality of entrepreneurship for 32 US States, 1988–2014. *American Economic Journal: Economic Policy*, *12*(4), 212–43.

Guzman, J., & Li, A. (2022). Measuring Founding Strategy. Management Science. *Articles in Advances*. https://doi.org/10.1287/mnsc.2022.4369

Hain, D. S., Jurowetzki, R., Buchmann, T., & Wolf, P. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, *177*, 121559

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, 36(1), 16–38.

Hannigan, T., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V., Wang, M., ... Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, *13*(2), 586–632.

Hartmann, P., & Henkel, J. (2020). The rise of corporate science in AI: Data as a strategic resource. *Academy of Management Discoveries*, *6*(3), 359–381.

Hausman, J. A., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica*, *52*(4), 909–938.

He, V. F., Puranam, P., Shrestha, Y. R., & von Krogh, G. (2020). Resolving governance disputes in communities: A study of software license decisions. *Strategic Management Journal*, *41*(10), 1837–1868.

Horton, J. J., & Tambe, P (2019). The death of a technical skill. Working Paper.

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

Huang, K. G., & Murray, F. E. (2009). Does patent strategy shape the long-run supply of public knowledge? Evidence from human genetics. *Academy of Management Journal*, *52*(6), 1193–1221.

Huang, K. G., & Murray, F. E. (2010). Entrepreneurial experiments in science policy: Analyzing the human genome project. *Research Policy*, *39*(5), 567–582.

Iansiti, M., & Lakhani, K. R. (2020). *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world*. Brighton, MA: Harvard Business Press.

Jain, A., & Huang, K. G. (2022). Learning from the past: How prior experience impacts the value of innovation after scientist relocation. *Journal of Management*, *48*(3), 571–604.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Jia, N., Huang, K. G., & Zhang, C. M. (2019). Public governance, corporate governance and firm innovation: An examination of state-owned enterprises. *Academy of Management Journal*, *62*(1), 220–247.

Jurowetzki, R., Hain, D., Mateos-Garcia, J., & Stathoulopoulos, K. (2021). The privatization of AI research (−ers): Causes and potential consequences – from university-industry interaction to public research brain-drain? Working Paper.

Kaplan, S. (2008). Cognition, capabilities, and incentives: Assessing firm response to the fiber-optic revolution. *Academy of Management Journal*, *51*(4), 672–695.

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, *36*(10), 1435–1457.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Leyden, B. T. (2018). There's an app (update) for that: Understanding product updating under digitization. Working Paper.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miric, M., Boudreau, K. J., & Jeppesen, L. B. (2019). Protecting their digital assets: The use of formal & informal appropriability strategies by App developers. *Research Policy*, *48*(8), 103738.

Miric, M., & Jeppesen, L. B. (2020). Does piracy Lead to product abandonment or stimulate new product development?: Evidence from Mobile platform-based developer firms. *Strategic Management Journal*, *41*(12), 2155–2184.
Miric, M, & Ozalp, H. (2022) Technological Standardization and The Generalizability of Human Capital: The Impact of Enabling Platform Technologies on Employee Mobility. Working Paper.
Miric, M., Ozalp, H., & Yilmaz, E. D. (2020). Trade-offs of using middleware: An innovation enabler and creativity constraint. Working Paper.
Miric, M., Pagani, M., & El Sawy, O. (2021). When and who do platform companies acquire? Understanding the role of acquisitions in the growth of platform companies. *MIS Quarterly*, *45*, 2159–2174.
Polanyi, M. (1966). The logic of tacit inference. *Philosophy*, *41*(155), 1–18.
Qiao, M., & Huang, K. W. (2021). Correcting misclassification bias in regression models with variables generated via data mining. *Information Systems Research*, *32*(2), 462–480.
Rathje, J. M., & Katila, R. (2021). Enabling technologies and the role of private firms: A machine learning matching analysis. *Strategy Science*, *6*(1), 5–21.
Ridge, J. W., Ingram, A., Abdurakhmonov, M., & Hasija, D., (2019). Market Reactions to Non-Market Strategy: Congressional Testimony as an Indicator of Firm Political Influence. *Strategic Management Journal*, *40*(10), 1644–1667.
Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity. *MIS quarterly*, *40*(4), 1035–1056.
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? *Organization Science*, *32*(3), 856–880.
Singh, A., Hosanagar, K., & Gandhi, A. (2020). Machine learning instrument variables for causal inference. In *Proceedings of the 21st ACM Conference on Economics and Computation* (pp. 835–836).
Teodoridis, F., Lu, J., & Furman, J. L. (2020). Measuring the direction of innovation: Frontier tools in unassisted machine learning. Working Paper.
Tidhar, R., & Eisenhardt, K. M. (2020). Get rich or die trying... finding revenue model fit using machine learning and multiple cases. *Strategic Management Journal*, *41*(7), 1245–1273.
Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*(9), 1600–1631.
Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.
WIPO (2019). *WIPO Technology Trends 2019: Artificial Intelligence*. Geneva: World Intellectual PropertyOrganization. Available at https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf (last accessed on July 5, 2022).
Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, *29*(1), 4–24.
Yilmaz E.D., Naumovska I., & Aggarwal V. (2022). Does AI replace labor (yet)? Evidence from Machine Translation. *Working Paper*.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Miric, M., Jia, N., & Huang, K. G. (2023). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, *44*(2), 491–519. https://doi.org/10.1002/smj.3441