Received: 29 January 2020 | Revised: 29 November 2022 | Accepted: 2 December 2022 | Published on: 27 December 2022

DOI: 10.1002/smj.3480

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

Using supervised machine learning to scale human-coded data: A method and dataset in the board leadership context

Joseph S. Harrison[1] | **Matthew A. Josefy**[2] | **Matias Kalm**[3] | **Ryan Krause**[1]

[1]Department of Management and Leadership, Neeley School of Business, Texas Christian University, Fort Worth, Texas, USA

[2]Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, Bloomington, Indiana, USA

[3]Department of Management, Tilburg School of Economics and Management, Tilburg University, Tilburg, Netherlands

Correspondence
Joseph S. Harrison, Department of Management and Leadership, Neeley School of Business, Texas Christian University, 2900 Lubbock Ave, Fort Worth, TX 76109, USA.
Email: j.s.harrison@tcu.edu

Abstract

Research Summary: Human coding of unstructured text can enable scholars to measure complex latent constructs for use in empirical analysis, but also requires substantial time and resources that limit the number and sample sizes of studies using this approach. We demonstrate how supervised machine learning (ML) can overcome these constraints by allowing scholars to scale human-coded data. Using board leadership as an illustrative context, we apply this method to create a large-scale dataset (N = 22,388) from smaller scale human codings of CEO duality and board chair orientations from company proxy statements. We further demonstrate the potential value of this approach by using the resulting dataset to examine the relationships among board leadership, firm performance, and CEO dismissal. The ML code and dataset are available at 10.5281/zenodo.7304697.

Managerial Summary: Manually converting unstructured text into usable data requires considerable time and resources. This article outlines a replicable process for applying supervised machine learning (ML) to overcome these constraints by scaling manually coded data.

The authors contributed equally and are listed in alphabetical order.

While ML is often used to identify patterns or predict relationships within a given dataset, we show how scholars and practitioners can build valuable custom algorithms at an earlier stage in the process—when first building a dataset. We illustrate this approach by training ML algorithms to replicate human codings of CEO duality and board chair control and collaboration orientations from over 22,000 company filings. We then show how this approach can support new knowledge development by using these data to explore the relationships among board leadership, company performance, and CEO dismissal.

KEYWORDS

board leadership, CEO duality, corporate governance, machine learning, measurement

1 | INTRODUCTION

Content analysis has become popular in the organizational sciences as a method to measure complex theoretical constructs and deepen our understanding of organizational phenomena (Duriau, Reger, & Pfarrer, 2007). In the strategic management domain, scholars have increasingly relied on unstructured text data from publicly available sources like earnings calls, shareholder letters, and social media posts to capture constructs that would otherwise be difficult or impossible to obtain, such as CEO cognitive traits (Gamache, McNamara, Mannor, & Johnson, 2015), impression management (Boudt & Thewissen, 2019), and strategic breadth (Eklund & Mannor, 2021). To assess these constructs, scholars have applied computer-aided text analysis, often using predefined dictionaries of words or phrases that are conceptually related to the construct of interest. Unfortunately, many constructs that might be gleaned from publicly available text data are too complex or context dependent to be measured through predetermined words or phrases. In such cases, scholars have relied on human coders to follow instructions and qualitatively evaluate text passages according to construct definitions (e.g., Nadkarni, Chen, & Chen, 2016; Wowak, Mannor, Arrfelt, & McNamara, 2016).

Human coders can identify constructs from text that may be too nuanced to capture using predefined dictionaries. Yet human coding also requires considerable resource investments (e.g., time, money, and training) that limit the size of the datasets that can reasonably be built using this technique, and in turn, the statistical power, accuracy, generalizability, and causal inferences that can be made from statistical analyses based on those data (Button et al., 2013; Colquhoun, 2014). Scholars also face a tradeoff between coding the text themselves, which can be time prohibitive, and hiring coders, which increases training requirements and can also increase coding errors, reduce reliability, and be cost prohibitive (Christensen, Nørskov, Frederiksen, & Scholderer, 2017; Crowston, Allen, & Heckman, 2012). Thus, despite the abundance of unstructured text data available to scholars, we face constraints in our ability to code complex latent constructs from those data on a large scale.

In this article, we explore the use of supervised machine learning (ML) as a method to overcome these constraints and facilitate construct measurement in the organizational sciences. When applied as a measurement technique, supervised ML methods can use pre-coded data as an input and learn how to reproduce the coding scheme based on features of the text that need not be defined a priori (e.g., word pairings, semantic similarity, syntax) (Gentzkow, Kelly, & Taddy, 2019). Once the algorithm achieves an acceptable level of accuracy, it can be applied to text data of essentially infinite size to produce accurate and reliable measures, overcoming the primary constraints scholars face in studying constructs that require hand coding. Scholars in the computer sciences have used ML methods to classify various kinds of texts, including group communication among software developers (Crowston et al., 2012; Crowston, Liu, & Allen, 2010), ideas proposed in online communities (Christensen et al., 2017), and speculative language in academic papers (Medlock & Briscoe, 2007). A few recent studies in strategic management have also begun to use ML-based methods for other purposes, such as topic modeling, facial recognition, and pattern recognition (e.g., Choi, Menon, & Tabakovic, 2021; Choudhury, Allen, & Endres, 2021; Choudhury, Wang, Carlson, & Khanna, 2019; Miric, Jia, & Huang, 2022). Still, such applications have been relatively limited in the organizational sciences. As recently expressed by Choudhury et al. (2021: 31: 31), "ML methods represent an exciting but underutilized toolkit for strategy and management researchers." Nevertheless, "greater adoption of these methods could be facilitated by illustrating relevant applications of ML to research in our fields."

We answer this call and contribute to the limited work that has applied ML-based methods in strategy research by outlining and demonstrating an approach to apply supervised ML to scale prior human-coded data. Our approach consists of four overarching stages: (a) identification of one or more relevant constructs and obtaining a training dataset of pre-coded text data; (b) development and selection of ML-based models using the training dataset; (c) application of the final selected model(s) to create a large-scale dataset; and (d) use of the resulting dataset to generate new scholarly insights. We outline and demonstrate this approach using board leadership as an illustrative context, specifically by using supervised ML to develop a large-scale dataset of CEO duality and board chair orientations (Krause, 2017; Oliver, Krause, Busenbark, & Kalm, 2018). We also make our ML scripts and dataset available online for future use.

Our study contributes to strategy and management research by enhancing scholars' awareness, understanding, and ability to apply ML as a construct measurement technique within our fields. Through the process and demonstration that we present, we aim to increase the accessibility of ML methods for organizational scientists, specifically as a means of scaling prior human-coded data. Making hand-coded data scalable using supervised ML can significantly advance strategy and related fields by increasing our ability to conduct replicable research and develop a cumulative body of knowledge (Ethiraj, Gambardella, & Helfat, 2016). As we outline in more detail later, the method has the potential to be applied to countless theoretical and empirical settings, allowing scholars to address new research questions and develop large-scale, novel datasets that otherwise would be unobtainable. This contribution is further enhanced by the nature and public availability of our resulting ML script and dataset. As we describe throughout our demonstration, we employ a wide range of ML techniques when developing our specific models, including various feature generation techniques, estimation methods, and n-gram lengths, as well as multiple metrics to assess different aspects of model performance. As such, the resulting script is not just applicable to our context, but may also be adapted to assess other constructs that can be gleaned from unstructured text data. Overall, our proposed process, demonstration, and publicly available script provide a strong foundation for future work seeking to develop large-scale datasets of difficult-to-measure constructs.

2 | USING SUPERVISED ML TO SCALE HUMAN-CODED DATA

Table 1 provides an overview of our approach to scale human-coded data using supervised ML. As outlined in the table, our approach consists of four overarching phases and involves a number of key considerations to support appropriate application of supervised ML methods for this purpose. In the first stage, a scholar identifies one or more relevant constructs and obtains a training dataset consisting of pre-labeled text data associated with those constructs. In the second stage, the training dataset is used to develop ML models that predict the constructs, after which final models are selected based on an assessment of their performance. In the third stage, the selected models are applied to additional text data to scale the prior human-coded dataset. Finally, the scaled dataset may be merged with other data and used in statistical analyses to conduct robust and replicable research. In the following sections, we detail each phase and highlight key considerations via a demonstration in the board leadership context.

3 | PHASE I: CONSTRUCT IDENTIFICATION AND TRAINING DATASET

3.1 | Identifying a relevant construct

The first step in our process is to identify a relevant construct for ML-based data scaling. When doing so, it is important to consider how the construct has been measured in the past and whether ML-based methods can enhance measurement. For instance, some constructs can be accurately and reliably measured using a dictionary-based approach, given the ease of identifying indicators that map onto those constructs a priori. Examples from past research include using lists of temporally oriented words to assess individuals' temporal focus (Nadkarni & Chen, 2014) and using word sets reflecting different types of strategies to assess strategic topics (Eklund & Mannor, 2021). In such cases, ML methods can add unnecessary complexity and reduce transparency relative to simpler methods. However, many constructs in our field are not easily assessed using simple approaches. These include constructs that have previously required human coding or where existing dictionaries (or other proxies) do not adequately capture the construct. In these cases, as long as the construct can be reliably assessed by human raters from unstructured text data, our approach has great potential to enhance measurement and future research related to the construct.

As an illustrative case, we focus on the board leadership context, and specifically on the control and collaboration orientations of the board chair position (Krause, 2017). As part of the Dodd-Frank Act of 2010, publicly traded U.S. firms are required to include a passage in their proxy statements (i.e., DEF14a filings) justifying their board leadership structure (Securities and Exchange Commission, 2010). Corporate governance scholars have recently relied on these passages to hand code the control and collaboration orientations of non-CEO board chairs (Krause, 2017; Oliver et al., 2018). Given research on board leadership has generally been limited to examining CEO duality, with some equivocal findings associated with its effects (Krause, Semadeni, & Cannella, 2014), the addition of more nuanced measures is an important extension. Yet, the labor intensity of hand coding these measures has limited both the number and sample sizes of studies exploring these or other more nuanced aspects of board leadership. Indeed, only two studies of which we are aware have explored board chair orientations to date,

TABLE 1 Using supervised ML to scale human-coded data

Phase/step	Example considerations
Phase I: Construct identification and training data	
1 Identify a relevant construct for ML-based data scaling	• Can ML-based models improve measurement relative to simpler methods?
2 Obtain the training dataset for the ML task	• Can the construct be accurately and reliably coded by human raters from unstructured data? • Is the training sample sufficiently large to designate a holdout set and apply cross-validation?
Phase II: Model development and selection	
3 Implement the ML task	
a. Text preprocessing	• How can the text data be simplified for training? • What range of n-grams will be used?
a. Feature extraction	• What type(s) of features will be generated? (e.g., frequency, semantic similarity, topic modeling)
a. Algorithm selection and regularization	• Is the construct binary, categorical, or continuous? • What algorithm(s) will be used? • How will the models be regularized? • What hyperparameters will be specified for estimation?
a. Model training and cross-validation	• What train-test split will be used? • What method will be used for cross-validation?
4 Assess performance metrics and select final ML model(s)	• Do models appear to be overfit to the training data? • What combination of features, algorithms, and hyperparameters most accurately predict the construct, with an acceptable level of loss? • What model-specific performance metrics can be used to assess performance?
Phase III: Data scaling	
5 Collect additional data	• How will additional text data be gathered and parsed? (e.g., manual, semi-automated, automated)
6 Apply the final selected ML model(s) to the new data	• Which model(s) will be used to predict the focal construct(s)?
7 Perform additional post hoc validation	• How much agreement is there between the predicted values and manual, post hoc checks?
Phase IV: Research application	
8 Use the scaled dataset to examine research questions	• Is there sufficient variance in the ML-based measure(s) to mitigate classification error? • Do the ML-based measure(s) replicate past findings? • What additional research questions can be answered with the scaled dataset?
9 Fully report on ML methods applied, validation steps, and any relevant limitations	• How were the training data compiled? How reliable were the human codings? • What specific steps were followed for preprocessing, feature extraction, model training, and validation? • What metrics were used to assess model performance?

TABLE 1 (Continued)

Phase/step	Example considerations
	• How were the final models selected? • What steps were taken for post hoc validation? • What are the limitations of the methods used?

Abbreviation: ML, machine learning.

each limited to samples of a few hundred firms (Krause, 2017; Oliver et al., 2018). Board leadership thus represents a useful case for our demonstration, as a setting where ML methods have strong potential to facilitate measurement.

3.2 | Building the training sample

After identifying a relevant construct, the next step is to build or otherwise obtain a training dataset, which consists of example input–output pairs that will be used to train and test the algorithm. In our case, because board chair orientations are only relevant in cases where the CEO and board chair roles are separated, it was necessary to obtain data that would allow us to train models to first identify the basic structure of board leadership (i.e., whether the CEO and board chair positions were separated or combined) and then to identify the orientations of the separate board chairs.[1] We did so by obtaining and combining the manually coded board leadership datasets previously used by Krause and colleagues (Krause, 2017; Oliver et al., 2018) and then supplementing those data with additional hand codings that we completed to independently verify the previously coded values and increase the size of the training sample.[2] Increasing the training sample was important so that we could designate a holdout set that was representative of the broader training sample and apply cross-validation during the training process. As we will describe and demonstrate in more detail below, each of these techniques is used to optimize the training process and ensure that the model does not become overfit to the training data.

Another important aspect of building the training sample is to ensure that the human codings are accurate and reliable prior to introducing ML methods. This is because the quality of the output provided by any given ML model is intrinsically linked to the input data (Domingos, 2012). Poor quality inputs will limit the quality of any ML-based models that can be developed using those data. In our case, because the final training sets were compiled from various separate datasets, we are unable to directly calculate interrater reliabilities for these data. However, in each of the initial studies, interrater reliabilities were between .75 and .90 and all discrepancies had been resolved prior to us receiving those data. Also, observations we added to those initial data were checked by multiple coauthors of this article, and only included

[1]While CEO duality is available through some databases (e.g., MSCI), the annual designations the databases assign to a given firm in a given year are not always aligned with the qualitative information in the proxy statement, particularly in years when either the CEO or board chair changes (Gove & Junkunc, 2013). Our dataset specifies the value of the CEO duality variable that is referred to in the proxy statement passage used to code the text-based variables. As such, we ensure that all our variables are coded based on the same information.

[2]Table S1 in the Online Appendix provides detailed descriptions of the board chair orientations, which were used by independent raters to code the variables both here and in Krause and colleagues' initial studies.

in the training sample once consensus was achieved regarding their coding. Altogether, the training data consist of 2,024 manually coded firm-year observations for CEO duality and 2,008 manually coded firm-year observations for control and collaboration orientations between 2010 and 2019. The full training datasets include the final coding of each variable as well as the unstructured text from the board leadership section of company proxy statements that were used to code the variable. All variables are dichotomous, taking the value of 1 for the presence of the construct in the analyzed text and 0 otherwise.

4 | PHASE II: MODEL DEVELOPMENT

4.1 | Implementing the ML task

After building the training sample, ML methodologies can be applied to execute the learning task. Typically, supervised ML tasks reflect either classification, where the machine learns to classify text into predetermined categories based on the training data, or regression, where the machine learns to predict values of a given construct by comparing features of the underlying text to values of the training data. Given our focal variables are dichotomous, our ML exercise reflects a binary classification task. We employed a multi-step process in Python to develop our classification algorithms, which broadly followed the steps used in other recent work employing supervised ML (see Choudhury et al., 2021). First, we preprocessed the text to reduce its dimensionality (Gentzkow et al., 2019). Second, we extracted features from the text to use in the ML task. Third, we defined the classification algorithms to use in the training process and defined regularization parameters. Finally, we used the human-coded data to train models to classify each of our focal variables based on the associated unstructured text (Christensen et al., 2017; Medlock & Briscoe, 2007). We describe each of these steps in greater detail below.

4.1.1 | Text preprocessing

Preprocessing consists of organizing and simplifying the training data by reducing its dimensionality and removing extraneous information. In our setting, because references to "chairman of the board" varied and comprised different token lengths (e.g., chairperson, chair of the board, board chair, the board chairman), we began by replacing all variants with "COTB." This simplified references to the board chair to a single token that could be used in subsequent steps. As part of the training, we also included code for (a) removing stop words (e.g., "the," "is," "at," "which," etc.), punctuation, extra whitespaces, and accents; (b) stemming, or reducing words to their stems; and (c) tokenization, or splitting the text within each document by words. We also generated various n-grams, which reflect the range of sequenced terms to use when representing features of the text in each document. N-gram lengths tested include bigrams (1–2 words), trigrams (1–3 words), 5-g (1–5 words), 10-g (1–10 words), and 20-g (1–20 words).[3]

[3]Tests using unigrams (single words) were insufficient to produce the desired level of accuracy. This provides further evidence that a simple dictionary-based approach is likely unsuitable for our variables of interest.

TABLE 2 Hyperparameters specified for each feature extraction technique

Model	Specifications
CV	Max features: 1 million Max iterations: 1,000
TF-IDF	Max features: 1 million Max iterations: 1,000
W2V	Vector size: [100, 200, 300] Model type(s): CBOW and skip-gram Minimum word frequency: 1
D2V	Vector size: [100, 200, 300] Model type(s): DM + DBOWs Minimum word frequency: 1 Max iterations: 100
LDA	LDA topics: [100, 200, 300] Learning method: Batch

Note: Values in brackets reflect ranges used to tune vector sizes and number of topics.
Abbreviations: CBOW, continuous bag of words; CV, count vector; D2V, Doc2Vec; DBOW, distributed bag of words; DM, distributed memory; LDA, latent Dirichlet allocation; TF-IDF, term frequency - inverse document frequency; W2V, Word2Vec.

4.1.2 | Feature extraction

Computer scientists have developed various methods for feature extraction, including simpler methods such as those relying on word frequencies and/or weights (see Gentzkow et al., 2019; Kaur, Kumar, & Kumaraguru, 2020), as well as more complex methods such as those based on topic modeling (e.g., Blei, Ng, & Jordan, 2003) or semantic similarity (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). As with most supervised ML, it was impossible to know a priori which specific method would yield the best results for our focal constructs (Choudhury et al., 2021). Thus, we employed a variety of these techniques, which we later compare before selecting final models. We used (a) count vectorizer (CV); (b) term frequency - inverse document frequency (TF-IDF) vectorizer; (c) Word2Vec (W2V) embedding; (d) Doc2Vec (D2V) embedding; and (e) latent Dirichlet allocation (LDA).[4] The hyperparameters used for each feature extraction method are outlined in Table 2.

4.1.3 | Classification models and regularization

We used two different types of classifiers: (a) a logistic regression (log) classifier and (b) a random forest (rf) classifier. When applied in ML, logistic regression is a simple but powerful classification method designed to identify the most relevant features for a given construct, reducing the weight assigned to less meaningful features and classifying an observation by measuring the probability that it falls into a given category. Cutoff points may then be used to classify a given

[4]The general logic behind each feature generation technique is described in Table S2 in the Online Appendix.

TABLE 3 Hyperparameters specified for each classification method[a]

Model	Specifications
Logistic regression	Solver(s): [Lib-Linear, L-BFGS] Penalty: [L1 (for Lib-Linear), L2 (for L-BFGS)] Class weight: [balanced, None] C: [.001, .01, .10, 1.0]
Random forest	Bootstrap: [True, False] Max depth: [10, 50, None] Estimators: [100; 500; 1,000]

[a]Values are provided only for hyperparameters we directly specified in the training script. Values in brackets reflect ranges or options selected for hyperparameter tuning. All other hyperparameters used default values.

observation dichotomously.[5] Moreover, in regression-based ML models such as logistic regression, regularization terms are added to the objective function and act as "penalties" to prevent overfitting of the model to the training data (Pekhimenko, 2006). When implementing the logistic classifier, we applied both L1 and L2 regularization, depending on the solver used (see Table 3). L1 regularization works by adding the absolute value of the weight parameters to the error function (i.e., attempting to estimate the median of the data), whereas L2 adds the squared value of weights (i.e., attempting to estimate the mean) (Pekhimenko, 2006). These and all other hyperparameter values used while training the logistic models are provided in Table 3. During training, we used cross-validated grid search to automatically tune these values to identify the best combination of hyperparameters.

rf models are another common ML technique, which operate by taking multiple random independent samples from the training data and then developing a decision tree for each sample based on how the data are classified within the training set (Breiman, 2001a). The result is a collection (or "forest") of decision trees, which the algorithm can use to classify new data by having each tree independently code the data and producing predicted probabilities to classify the data based on the category most chosen by the collection of trees. Like regularization in regression-based ML models, this process mitigates overfitting of the model, while also improving predictive accuracy. Hyperparameters used for the rf models are provided in Table 3 and were also tuned using cross-validated grid search.

4.1.4 | Model training and cross-validation

To train our models, we first randomly split the training data to designate a holdout set, which the machine never encountered during the training process. By comparing the predictive accuracy of the model in the training set (train accuracy) to that of the holdout set (test accuracy), a scholar can determine how accurate the model is on out-of-sample data and whether the model is overfit to the training data. Similar to past work, we employed an 80/20 train-test split (Harrison, Thurgood, Boivie, & Pfarrer, 2019). We then implemented the script to train separate models for each distinct combination of features (CV, TF-IDF, W2V, D2V, and LDA), n-gram

[5]For variables that are roughly balanced between categories, the convention is to use a cutoff value of .5 (Prati, Batista, & Silva, 2015). Given substantial balance among the categories for duality, control, and collaboration orientations, this is what we use for our focal variables.

lengths (bigrams, trigrams, 5-g, 10-g, and 20-g), and classification methods (log and rf). Given we also tested a range of vector sizes (for W2V and D2V) and LDA topic lengths, this yielded 110 different models to choose from for each variable, or 330 total models. To further optimize the training process for each model, we also employed k-fold cross-validation, a resampling procedure used to evaluate model performance in limited samples (James, Hastie, Tibshirani, & Witten, 2013; Kuhn & Johnson, 2013). The method is analogous to designating a holdout set from the full training sample, but is employed while working with the training set (i.e., in our case, 80% of the full training sample). It works by splitting the data into k groups of approximately equal size and then cycling through each group, holding it out as a test sample while using the remaining $k - 1$ groups to train the models (James et al., 2013). This process provides the machine with information regarding model accuracy while it is training the model. Most computer scientists use a k of 5 or 10 (Kuhn & Johnson, 2013). We use $k = 5$ to ensure that the cross-validation test set is large enough to be representative of the broader training set. Following best practice, we also use a shuffle-split strategy to randomly shuffle the data prior to splitting it into the five groups used for cross-validation.

4.2 | Assessing model performance and selecting final models

After training the models, the next step is to assess model performance and determine which models to apply to scale the human-coded data. Computer scientists have developed various metrics to assess performance for classification models, including accuracy statistics, error and loss statistics, and indicators of true and false positive rates (Seliya, Khoshgoftaar, & Van Hulse, n.d.; Huang & Ling, 2005). Given each performance metric provides different information, it is generally best to use multiple metrics to get an overall sense of model performance. Here, we consider model accuracies, train-test loss, and area under the curve (AUC) values.

Accuracies reflect the percent of observations in the training or test data that the model is able to accurately classify. In our case, the machine uses accuracies for hyperparameter tuning during training and validation and then outputs the accuracies for both the training and test sets once the model has been optimized. The difference between these two values is the train-test loss, which reflects the amount of accuracy that is lost when extrapolating the model from the training set to the test set. In general, some degree of loss is to be expected between the training and test accuracies; however, high train-test loss values indicate that a model may be overfit to the training data. Additionally, while higher accuracies typically reflect superior models, models that perfectly or nearly perfectly predict the training data are likely to be overfit.

As a final indicator of performance, we also plotted the receiver operating characteristic (ROC) curve and calculated AUC values for each model (Huang & Ling, 2005). ROC curves illustrate the diagnostic ability of a classification model by plotting the model's true positive rate against its false positive rate across all possible thresholds (i.e., cutoff levels). The AUC is then calculated as an indication of the model's ability to distinguish between the categories of the variable, with higher values reflecting superior models. An AUC score between .7 and .8 is generally considered good diagnostic ability, a score between .8 and .9 is excellent, and a score above .9 is outstanding (Gallop, Crits-Christoph, Muenz, & Tu, 2003).

Performance metrics are provided for all models in the Online Appendix (see Tables S3 through S6). To decide on a final set of models for our focal variables, we first examined train accuracies and train-test loss values to identify models that were likely overfit and excluded

TABLE 4 Final model specifications and performance by board leadership variable

	Board leadership variable		
Model specifications	**Duality**	**Control**	**Collaboration**
Feature(s)	CV	CV	CV
n-grams (g)	5-g	2-g	10-g
Classifier	rf	log	log
Estimation hyperparameters	Bootstrap: False Max depth: 50 Estimators: 500	Solver: L-BFGS (L2) Class weight: Balanced C: .10	Solver: L-BFGS (L2) Class weight: None C: .01
Train accuracy	.986	.994	.990
Test accuracy	.856	.861	.858
Train/test loss	.130	.133	.132
AUC	.915	.904	.931

these models from consideration. We considered a model overfit if it had a train accuracy of .99 or above and/or a train-test loss value above .15. Nearly all models based on the rf classifier in the more advanced models (i.e., those using W2V, D2V, and LDA) appear to be overfit, so we only considered models using the logistic classifier for those sets of models. Many of the simpler, count-based models (CV and TF-IDF) also appear to be overfit, although these seem to be more sensitive to the variable under consideration (i.e., logistic models are generally overfit for duality, rf models are overfit for control, and models for collaboration are mixed). After excluding these models, we compared accuracies and AUC scores across the remaining models. For the remaining models, we found that models based on CV and TF-IDF generally outperformed models based on the more complex features. Train accuracies were consistently above .9 for the count-based models compared to between .6 and .9 for W2V, D2V, and LDA, and test accuracies were consistently above .8 compared to between .6 and .8 for W2V, D2V, and LDA. Similarly, AUC scores for the count-based models were consistently above .9, whereas those of models using the more complex features often fell below .8 or .7. This led us to select our final models from the count-based models. Among these, where accuracy and diagnostic ability were roughly equivalent, we prioritized simpler models (e.g., smaller n-grams, logistic over rf, and CV over TF-IDF) for parsimony and because simpler models tend to be more interpretable than more sophisticated models (Rudin, 2019). As scholars apply our approach in the future, we suggest that they follow a similar pattern by selecting the simplest set of models that optimize model performance. Doing so will not only increase the transparency of the models used in subsequent testing, but should also increase their accessibility for future application by other scholars.

Table 4 provides model specifications (i.e., classifier type, estimation hyperparameters, n-gram length, and feature type) and performance metrics (i.e., accuracies, loss values, and AUCs) for the best overall models for each variable, which we use to build our final dataset.[6] Because the text data and ML script used to code these variables accompany this article, future research may implement any of the 330 different models to generate measures of either our focal variables or

[6]As supplemental performance indicators, ROC curves and confusion matrices for the final models are provided in Table S7 of the Online Appendix.

other variables of interest from prior hand-coded data. Importantly, as scholars apply our script, they should be aware that models will perform differently for different constructs. In our case, duality and the control and collaboration orientations are relatively simple constructs, though not so simple that we could identify a specific word list for them a priori. This likely explains why the count-based models outperform those using more sophisticated features and why the rf classifier generally overfits the models to the training data. However, more complex constructs (e.g., psychological traits) may require the use of more sophisticated methods, so scholars should test and compare multiple options before deciding on final models.

5 | PHASE III: DATA SCALING

5.1 | Additional data collection

After training ML algorithms to accurately predict the desired variables, scholars may then apply those models to new data to scale the human-coded dataset. For natural language processing exercises such as ours, this will involve collecting and parsing a much larger set of text data than was used to develop the hand-coded dataset. We gathered additional text data from company proxy statements that we collected from SEC EDGAR using central index keys (CIKs) for a vast number of publicly traded companies. We compiled our initial list of CIKs by identifying all firms listed among the S&P 1500 at any time between 2010 and 2018. We added to this initial list any other firm included in the MSCI database over that time frame for which a CIK was available. Altogether, our final list included 5,503 unique CIKs of firms listed among the S&P 1500 and/or within the MSCI database between 2010 and 2018.

To facilitate gathering company filings, we utilized a Python "fetcher" script to look up each company by its CIK and download all available DEF14a filings for the company beginning in 2010. We were able to download filings for 4,450 firms, or approximately 88% of the initial list. Next, we developed an "extractor" script to identify the paragraphs in the filings that contained the board leadership text. The extractor script works by detecting key phrases (e.g., "board leadership structure") that would be at the beginning of the relevant section. The script also relied on html tags that indicated headers when present. Because some documents are less structured than others, we iterated this process, inspecting documents for which no leadership section was extracted and determining what keywords or html tags were present that could indicate the start of the appropriate section. Using this script, we extracted the most relevant section of DEF14a filings for 3,303 unique CIKs, or about 74% of covered firms.[7]

5.2 | Application of ML models to create measures

With additional text data gathered and relevant sections extracted, scholars can use their final selected ML models to predict their construct(s) of interest. We did this by applying the final models represented in Table 4 to the additional data we gathered. The final generated dataset consists of 22,388 firm-year observations and includes the predicted probabilities and binary

[7]When conducting subsequent checks on a random sample from the final dataset, we note that approximately 7% of extracted texts did not provide sufficient information for a human coder to clearly identify board chair orientation. Our algorithms coded these values as zero, which is as close to human coding as we could expect.

coding of all study variables, text from the board leadership section of DEF14a filings used to create those variables, and unique identifiers for companies and text passages to facilitate future use and matching to other datasets. A full summary of the database is provided in the Online Appendix (see Table S9). Within the dataset, variables related to CEO duality are coded for all 22,388 firm-year observations. Because control and collaboration orientations are only possible for firms with a separate CEO and board chair, these variables are only coded for the 13,900 firm-years where the binary coding of CEO duality takes a value of 0.

5.3 | Additional validation using the dataset

After creating the final dataset, scholars should perform additional validation checks to ensure that the models are functioning as expected when applying them to new data. We did this by hand coding a random sample of 200 observations that had not been included during the initial training and testing process and then comparing those codings to predicted values from our final models. We found 87% agreement for CEO duality, 82% agreement for control, and 87% agreement for collaboration.[8] These values are comparable to the test accuracies found during the training exercise, indicating that the models are performing consistently when applied to the additional data. Moreover, given interrater agreement for these variables ranged between .75 and .90 in the original studies by Krause and colleagues (Krause, 2017; Oliver et al., 2018), the post hoc values demonstrate that our algorithms are at least as reliable, and in some cases more reliable, than human coding. This may be generally true when scaling human-coded data, as long as rigorous ML methods are applied, since computers do not suffer from error due to fatigue as do human coders (Crowston et al., 2012; Medlock & Briscoe, 2007). This provides another indication of the utility of our method for scaling prior human-coded data.

Still, the lack of perfect consistency indicates that at least some observations will be misclassified in the final dataset. Classification error is likely to exist at least to some extent in both hand-coded and computer-generated datasets. Recent research suggests that this can cause bias when machine-generated variables are used in subsequent econometric analysis, either as an independent or dependent variable (Ge, Huang, & Png, 2016; Hausman, Abrevaya, & Scott-Morton, 1998). At the same time, this bias decreases with the variability of the machine-generated variable (Ge et al., 2016). Fortunately, our variables each exhibit a high degree of variability, as evidenced by coefficient of variation values (duality = 1.11; control = 1.14; collaboration = 1.26). This suggests that classification error should not be biasing our predictive tests below. However, scholars should be aware of this potential limitation when applying the method to variables that exhibit less variance, and should consider applying modified estimators or using other techniques that can mitigate bias from classification error in these instances (e.g., Hausman et al., 1998).[9]

[8]Full confusion matrices and Cohen's Kappa values for each variable are provided in Table S8 of the Online Appendix.

[9]A related issue that has received some attention in past research is that generated regressors can sometimes exhibit low consistency and efficiency, and may limit the validity of inferences that can be made using those regressors (Pagan, 1984). However, this issue is less relevant to ML-based measures and more specific to predicted values or residuals from traditional statistical analyses (e.g., OLS regression). This is because of certain unique features of ML-based methods compared to traditional statistics: ML-based methods focus on predictive accuracy rather than causal inference, can accommodate larger and more complicated sets of features, rely on far fewer assumptions than traditional approaches, and are better equipped to handle nonlinear relationships (Breiman, 2001b; Bzdok, Altman, & Krzywinski, 2018). As a result, ML-based methods are more comparable to the human codings they are meant to mimic than generated regressors using traditional statistical analyses.

6 | PHASE IV: RESEARCH APPLICATION

Once the data have been scaled using the selected ML models, scholars can merge the resulting variables with other data to conduct replicable research and contribute to knowledge in their discipline (Ethiraj et al., 2016). As an illustration and to provide a predictive test of our method, we merged our dataset with other existing datasets to examine the extent to which our ML-based variables moderate the well-established negative relationship between firm performance and CEO dismissal (Finkelstein, Hambrick, & Cannella, 2009). Corporate governance research has consistently shown that poor firm performance is the primary predictor of CEO dismissal, with all other factors either moderating or at least operating within the context of that core relationship (Wiersema & Zhang, 2011; Wowak, Hambrick, & Henderson, 2011). Considerable research in corporate governance has focused on identifying the factors that either enhance a board's ability to dismiss poorly performing CEOs or, conversely, help CEOs avoid dismissal for poor performance (e.g., Flickinger, Wrage, Tuschke, & Bresser, 2016; Shen & Cannella, 2002). Among the most frequently studied factors are CEO duality and other factors reflecting controlling or collaborative governance approaches. Thus, CEO dismissal represents a useful context to illustrate our method's predictive validity.

As a basis for the predictive test, we combined our ML-based dataset with CEO dismissal data from Gentry, Harrison, Quigley, and Boivie (2021) as well as firm, executive, and other data from Compustat, Execucomp, Institutional Brokers Estimates System, and Institutional Shareholder Services. After excluding observations with missing data and accounting for time lags, the sample for the predictive tests includes 6,963 firm-year observations (and 1,249 unique firms) from 2010 to 2018. Of these, the CEO and board chair positions were separated in 3,733 observations (and 874 unique firms). Using these data, we then employed random-effects logistic regression models to examine the relationships among firm performance (using prior year return on assets, or ROA), our ML-based measures, and CEO dismissal, while also controlling for other factors that have been previously shown to be relevant in this context.[10] Table 5 presents the results of the random effects logistic regression models.[11] Below, we describe our findings within the context of existing corporate governance research.

6.1 | CEO duality

Agency theorists have regularly argued that CEO duality should protect CEOs from dismissal for poor performance. Yet, one of the more persistent findings in governance research is that

[10]We follow prior work and use random effects models both because we are conceptually interested in between-firm variance in our focal independent variables and because fixed effects models would exclude any firm that did not experience CEO dismissal during the sample period (approximately 80% of our sample), leading to biased estimates (Flickinger et al., 2016; Gentry et al., 2021; Wiersema & Zhang, 2011). A limitation of this approach is that it does not allow us to account for stable, unobserved properties of firms. To address this, in supplemental tests we also estimated (a) multilevel mixed effects logistic regression models, which account for general firm-level effects (Snijders & Bosker, 2011) as well as (b) generalized estimating equations (GEEs), which take into account within-firm correlation in error terms (Liang & Zeger, 1986; Snijders & Bosker, 2011). For the GEE models, we specified a binomial distribution family with a logit link function because of the binary outcome variable and exchangeable correlation structure. The results from both sets of models were consistent with those reported below.

[11]Variable descriptions, descriptive statistics, and correlations for all variables included in our analysis are provided in Tables S10 and S11 of the Online Appendix.

TABLE 5 Random effects logistic regression models predicting CEO dismissal[a,b]

	Full sample						Separate CEO and board chair					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	b	*p*	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*
Constant	−2.784	(.016)	−3.276	(.004)	−3.317	(.003)	−3.080	(.007)	−3.043	(.009)	−2.971	(.010)
Total stock return	−0.930	(.193)	−0.692	(.320)	−0.700	(.315)	−1.381	(.002)	−1.381	(.002)	−1.371	(.003)
Average analyst recommendation	−0.192	(.253)	−0.124	(.449)	−0.119	(.468)	−0.008	(.968)	0.008	(.970)	−0.012	(.951)
Leverage	0.000	(.730)	−0.001	(.635)	−0.001	(.646)	−0.002	(.762)	−0.002	(.751)	−0.003	(.736)
Firm size	−0.013	(.853)	0.050	(.469)	0.048	(.484)	0.096	(.252)	0.090	(.279)	0.098	(.244)
Ln (R&D expenses)	0.097	(.026)	0.094	(.028)	0.094	(.027)	0.112	(.057)	0.107	(.064)	0.112	(.052)
CEO age	0.011	(.373)	0.013	(.265)	0.014	(.245)	−0.004	(.835)	−0.001	(.974)	−0.001	(.947)
CEO tenure	−0.031	(.014)	−0.024	(.058)	−0.024	(.057)	−0.030	(.107)	−0.030	(.096)	−0.030	(.103)
CEO is female	0.194	(.585)	0.168	(.641)	0.148	(.686)	−0.148	(.776)	−0.186	(.722)	−0.217	(.681)
Board size	0.025	(.626)	0.019	(.706)	0.021	(.677)	0.057	(.323)	0.057	(.326)	0.045	(.450)
Inside directors	−0.163	(.039)	−0.170	(.030)	−0.172	(.028)	−0.185	(.059)	−0.207	(.037)	−0.201	(.042)
Female directors	−0.070	(.430)	−0.065	(.467)	−0.064	(.471)	−0.156	(.210)	−0.149	(.227)	−0.133	(.282)
ROA			−2.405	(.000)	−2.047	(.002)	−1.792	(.005)	−1.633	(.010)	−1.814	(.028)
Duality			−0.335	(.038)	−0.301	(.061)						
Duality × ROA					−1.210	(.192)						
Control									−0.405	(.050)	−0.318	(.130)
Collaboration									0.200	(.279)	0.165	(.373)
Control × ROA											−2.750	(.061)
Collaboration × ROA											1.396	(.185)
Industry fixed-effect	Yes		Yes		Yes		Yes		Yes		Yes	
Year fixed-effects	Yes		Yes		Yes		Yes		Yes		Yes	
N	6,967		6,967		6,967		3,733		3,733		3,733	

TABLE 5 (Continued)

	Full sample						Separate CEO and board chair					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	b	*p*	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*	*b*	*p*
Number of firms	1,247		1,247		1,247		874		874		874	
χ^2	75.2	(.104)	116.1	(.000)	124.0	(.000)	132.4	(.000)	149.3	(.000)	153.0	(.000)

[a]Two-tailed *p*-values are provided in parentheses; all models include robust *SE*s, clustered by firm.
[b]Except for industry and firm fixed effects, all independent variables are measured 1 year prior to the year for which we assess CEO dismissal.

CEO duality has a negative main effect on dismissal, but no interaction effect with firm performance (Cannella & Lubatkin, 1993; Ocasio, 1994). This finding has confounded scholars for years, but is remarkably robust. One possible reason for this finding may be that the protective effect of CEO duality is so strong that it interferes with any possible dismissal antecedents, including performance. Whatever the reason, if our measure of CEO duality using ML is valid, we expect to find a negative relationship between duality and CEO dismissal, and we should expect no interactive effect between duality and performance on dismissal.

As expected, we find a consistent negative relationship between firm performance and CEO dismissal. We also find that our ML measure of CEO duality is negatively related to dismissal (Model 2, $\beta = -.335$, $p = .038$) and that there is not a meaningful interaction between CEO duality and firm performance on CEO dismissal. But because our models are nonlinear, interaction terms are not sufficient as a test of the moderating effect of CEO duality (Hoetker, 2007). Thus, to further examine this relationship, we calculate and compare the marginal effects of ROA on CEO dismissal for firms with and without dual CEO-board chairs.[12] Our results show similar negative effects for ROA on CEO dismissal when the CEO and board chair positions are combined (-0.078, $p = .001$) and when they are separated (-0.067, $p = .003$), providing additional evidence that duality does not moderate the relationship. These findings are consistent with expectations, demonstrating predictive validity of the ML-based measure of duality.

6.2 | Control and collaboration orientations

Whereas past work has often failed to find an interaction between duality and performance on dismissal, prior work does suggest that a more control-minded board is more likely to dismiss a CEO for poor firm performance, and a more collaborative board is less likely to dismiss a CEO for poor firm performance (Guo & Masulis, 2015; Shen & Cannella, 2002). Although the control and collaboration orientations of the board chair position, in particular, have not yet been linked to CEO dismissal, if our measures of board chair control and collaboration orientations are accurate, we would expect similar relationships using these measures. That is, a control-oriented board chair should be more likely to dismiss a CEO for poor performance, and a collaboration-oriented board chair should be less likely to dismiss a CEO for poor performance (Sundaramurthy & Lewis, 2003).

In our data, while we do not find immediate evidence of an interactive effect for ROA and a collaboration board chair orientation (Model 6, $\beta = 1.396$, $p = .185$), we do find that a control board chair orientation strengthens the negative relationship between firm performance and CEO dismissal (Model 6, $\beta = -2.750$, $p = .061$). As before, we performed marginal effects analysis for both interactions. We find that the marginal effect of ROA on dismissal for firms adopting a control orientation (-0.102, $p = .011$) is about 2.5 times that of firms not adopting this orientation (-0.042, $p = .135$). Conversely, the effect is the opposite for a collaboration board chair orientation. We find no strong evidence for a marginal effect of ROA on dismissal for firms with a collaboration board chair orientation (-0.046, $p = .194$) but we do for firms that do not have a collaboration orientation (-0.081, $p = .003$). Overall, these results are consistent with the well-established theoretical argument that a control approach to governance should strengthen the negative performance-dismissal relationship, whereas a collaboration approach should weaken this relationship. These findings provide further evidence of the predictive

[12]Detailed results of this analysis are provided in Table S12 of the Online Appendix.

validity of our method and measures of board chair orientation, demonstrating that supervised ML techniques can be used to produce variables needed to test key relationships in much broader samples than those typically used in studies that require human coding.

7 | FUTURE RESEARCH

In this study, we present and demonstrate a process for using supervised ML to build large datasets of complex constructs in the organizational sciences, by using prior human-coded text as training data. In our demonstration, we employ a wide range of ML methods, including various feature extraction techniques, n-gram lengths, and estimation methods, to arrive at an optimal set of algorithms to assess three board leadership variables (i.e., CEO duality, control orientation, and collaboration orientation), from company proxy statements. We show the validity of our method both in terms of its ability to accurately assess these variables, as well as its ability to predict well-established theoretical relationships in the governance literature. The Python script we used to train these algorithms and score the unstructured text data, along with the resulting board leadership dataset of 22,388 firm-year observations of public U.S. companies from 2010 to 2018, are available at 10.5281/zenodo.7304697. As such, we contribute to strategic management research by providing both a method and dataset that may be used by scholars in the future to pursue research questions that otherwise may be unachievable.

7.1 | Future research using the method

Moving forward, our method may be applied to any theoretical or empirical setting where scholars seek to build large-scale datasets of complex or context-dependent constructs from publicly available text data. One broad area where we see vast potential is in research on behavioral and cognitive aspects of strategic leadership and corporate governance. Despite a recent surge of research in these areas (Bromiley & Rau, 2016; Westphal & Zajac, 2013), scholars have generally had to rely on predefined dictionaries (e.g., Boudt & Thewissen, 2019; Gamache et al., 2015; Graf-Vlachy, Bundy, & Hambrick, 2020) or human coding (Nadkarni et al., 2016; Wowak et al., 2016) to measure their constructs of interest. In the former case, applying our method may allow for finer-grained or more accurate measures of complex, context-specific constructs. For instance, while the dominant approach for measuring sentiment in a firm setting has been to use dictionaries of positive and negative emotion words, many words are unlikely to translate across settings (Choudhury et al., 2019; Loughran & McDonald, 2011). Applying supervised ML methods may facilitate the creation of better validated, context-specific algorithms across a wide range of data sources, from 10-K filings and shareholder letters, to interviews, to social media or blog posts. In the latter case, applying our method to prior human-coded data may improve the replicability of past research by allowing scholars to assess certain variables in much larger datasets than have been used in the past. For example, in order to study the relationship between CEO charisma and firm strategy, Wowak et al. (2016) created dossiers of media and analyst coverage for 113 CEOs of S&P 1500 firms and manually coded these dossiers for CEO charisma. Applying our method, scholars could use these data as a training sample to develop algorithms to efficiently code CEO charisma and build a much larger dataset on CEO charisma across a greater number of firms.

Overall, applying our method may allow scholars to improve the accuracy of prior methods, expand the scale of prior human-coded data, and also develop new measures or datasets that have yet to be conceived. This method will also allow scholars to avoid the major shortcomings of human-coded data, including scalability as well as accuracy and reliability issues. As future research seeks to apply our method in these ways, a few additional points of guidance are warranted. First, scholars should be aware that ML-based tools tend to be highly source- and sample-specific. Thus, any application of those tools should be consistent with the nature of the original training data. For instance, given our board leadership tool was developed to assess duality and board chair orientations from public, U.S. company proxy statements, it would be inappropriate to apply it to evaluate text sources other than proxy statements, or filings submitted to regulators in other jurisdictions, without first modifying it for those purposes. Any time a scholar seeks to apply an ML-based tool to data that differ significantly from the original input (e.g., due to differences in geographic or cultural context, systematic changes in reporting conventions over time, a difference in how the text was produced, etc.), additional training data would need to be collected and the tool retrained to be able to account for those differences.

Second, it is important to note that while we found that simpler, count-based models functioned better for our board leadership variables (similar to Crowston et al., 2010), some constructs may require more complex modeling. In developing our code, we have built in flexibility to accommodate more complex word embeddings and different model types. The code can also be adapted to perform different ML tasks beyond the binary classification we develop in this article. This is because the Python modules we use include different classifiers for various outcome distributions. Still, additional development may be necessary depending on the constructs or data sources of interest. In particular, while we believe the same scale-up approach is possible with other types of underlying data, including audio, video, and informal text (e.g., social media posts), scholars may need to implement additional steps not currently included in our code to ensure suitability for those data types. Likewise, while our underlying data exhibited a high degree of reliability between raters, indicating a certain amount of objectivity when assessing our constructs of interest, additional steps may be needed if data used in future applications have a higher rate of disagreement between raters, which would indicate greater subjectivity or the potential presence of multiple evaluation schema.

7.2 | Future research using the dataset

Our open-source board leadership database also provides a number of opportunities for future governance research. Given only two prior studies have explored board chair orientation, many questions remain unanswered regarding its antecedents and consequences. For example, while Krause (2017) found that board chair control and collaboration orientations each affect firm performance, research has yet to explore more proximal outcomes. Elsewhere in the literature, studies have shown that governance-related factors can have important implications for firm strategies such as acquisitions, innovation, and corporate risk-taking (e.g., Goranova, Priem, Ndofor, & Trahms, 2017; Hoskisson, Hitt, Johnson, & Grossman, 2002; Lim & McCann, 2013). With firms increasingly separating the CEO and board chair positions (Spencer Stewart, 2021), the impact of the board chair's orientation on strategies like these will only become more important to study. Moreover, given the potentially sweeping implications of the control and collaboration orientations for firm strategy, governance, and performance, future research may

benefit from using our dataset to identify additional factors that drive a board chair to take on one or both of these orientations, such as CEO or board characteristics, firm characteristics, or external pressure or industry isomorphism.

We also expect that our coding of CEO duality will be of value to corporate governance scholars. While information on CEO duality is available in databases like MSCI, these datasets are proprietary and may be prohibitively expensive for many scholars to obtain. Because we have coded our data independently from publicly available SEC filings, we are able to provide CEO duality data for general use. Moreover, changes in the practice of CEO duality over time may create new opportunities for future work. While a vast literature exists on this topic (Dalton, Hitt, Certo, & Dalton, 2007; Krause et al., 2014), if duality is undergoing fundamental institutional change, some of these past relationships may differ today or in the future. Future research may benefit by using our dataset to explore the evolving meaning of CEO duality.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the National Center for the Middle Market (NCMM) at the Fisher College of Business, The Ohio State University, supporting their efforts to make governance data more readily available on a wide set of firms. The authors also appreciate contributions and coding expertise provided by Alexander Gedranovich, Hissaan Ali Shah, Muhammad Usama, and Umang Mehta at various stages of the project. Finally, the authors would like to thank Rajshree Agarwal and three anonymous reviewers for their valuable and developmental feedback on this article.

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 code and data are openly available at 10.5281/zenodo.7304697.

ORCID

Joseph S. Harrison https://orcid.org/0000-0002-4835-0566
Matthew A. Josefy https://orcid.org/0000-0003-1174-1878
Matias Kalm https://orcid.org/0000-0001-8333-6340
Ryan Krause https://orcid.org/0000-0001-8651-3074

REFERENCES

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

Boudt, K., & Thewissen, J. (2019). Jockeying for position in CEO letters: Impression management and sentiment analytics. *Financial Management*, *48*(1), 77–115. https://doi.org/10.1111/fima.12219

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Bromiley, P., & Rau, D. (2016). Social, behavioral, and cognitive influences on upper echelons during strategy process: A literature review. *Journal of Management*, *42*(1), 174–202. https://doi.org/10.1177/0149206315617240

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Cannella, A. A., & Lubatkin, M. (1993). Succession as a sociopolitical process: Internal impediments to outsider selection. *Academy of Management Journal*, *36*(4), 763–793. https://doi.org/10.2307/256758

Choi, J., Menon, A., & Tabakovic, H. (2021). Using machine learning to revisit the diversification–performance relationship. *Strategic Management Journal*, *42*(9), 1632–1661. https://doi.org/10.1002/smj.3317

Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57. https://doi.org/10.1002/smj.3215

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. https://doi.org/10.1002/smj.3067

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creativity and Innovation Management*, *26*(1), 17–30. https://doi.org/10.1111/caim.12202

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216. https://doi.org/10.1098/rsos.140216

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523–543. https://doi.org/10.1080/13645579.2011.625764

Crowston, K., Liu, X., & Allen, E. E. (2010). Machine learning and rule-based automated coding of qualitative data. In Proceedings of the American Society for Information Science and Technology.

Dalton, D. R., Hitt, M. A., Certo, S. T., & Dalton, C. M. (2007). The fundamental agency problem and its mitigation: Independence, equity, and the market for corporate control. *Academy of Management Annals*, *1*(1), 1–64. https://doi.org/10.5465/078559806

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

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. https://doi.org/10.1177/1094428106289252

Eklund, J. C., & Mannor, M. J. (2021). Keep your eye on the ball or on the field? Exploring the performance implications of executive strategic attention. *Academy of Management Journal*, *64*(6), 1685–1713. https://doi.org/10.5465/amj.2019.0156

Ethiraj, S. K., Gambardella, A., & Helfat, C. E. (2016). Replication in strategic management. *Strategic Management Journal*, *37*(11), 2191–2192. https://doi.org/10.1002/smj.2581

Finkelstein, S., Hambrick, D. C., & Cannella, A. A. (2009). *Strategic leadership: Theory and research on executives, top management teams, and boards*. New York, NY: Oxford University Press.

Flickinger, M., Wrage, M., Tuschke, A., & Bresser, R. (2016). How CEOs protect themselves against dismissal: A social status perspective. *Strategic Management Journal*, *37*(6), 1107–1117. https://doi.org/10.1002/smj.2382

Gallop, R. J., Crits-Christoph, P., Muenz, L. R., & Tu, X. M. (2003). Determination and interpretation of the optimal operating point for roc curves derived through generalized linear models. *Understanding Statistics*, *2*(4), 219–242. https://doi.org/10.1207/S15328031US0204_01

Gamache, D. L., McNamara, G., Mannor, M. J., & Johnson, R. E. (2015). Motivated to acquire? The impact of CEO regulatory focus on firm acquisitions. *Academy of Management Journal*, *58*(4), 1261–1282. https://doi.org/10.5465/amj.2013.0377

Ge, C., Huang, K. W., & Png, I. P. (2016). Engineer/scientist careers: Patents, online profiles, and misclassification bias. *Strategic Management Journal*, *37*(1), 232–253. https://doi.org/10.1002/smj.2460

Gentry, R. J., Harrison, J. S., Quigley, T. J., & Boivie, S. (2021). A database of CEO turnover and dismissal in S&P 1500 firms, 2000–2018. *Strategic Management Journal*, *42*(5), 968–991. https://doi.org/10.1002/smj.3278

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Goranova, M. L., Priem, R. L., Ndofor, H. A., & Trahms, C. A. (2017). Is there a "dark side" to monitoring? Board and shareholder monitoring effects on M&A performance extremeness. *Strategic Management Journal*, *38*(11), 2285–2297. https://doi.org/10.1002/smj.2648

Gove, S., & Junkunc, M. (2013). Dummy constructs? Binomial categorical variables as representations of constructs: CEO duality through time. *Organizational Research Methods*, *16*(1), 100–126. https://doi.org/10.1177/1094428112472000

Graf-Vlachy, L., Bundy, J., & Hambrick, D. C. (2020). Effects of an advancing tenure on CEO cognitive complexity. *Organization Science*, *31*(4), 936–959. https://doi.org/10.1287/orsc.2019.1336

Guo, L., & Masulis, R. W. (2015). Board structure and monitoring: New evidence from CEO turnovers. *The Review of Financial Studies*, *28*(10), 2770–2811. https://doi.org/10.1093/rfs/hhv038

Harrison, J. S., Thurgood, G. R., Boivie, S., & Pfarrer, M. D. (2019). Measuring CEO personality: Developing, validating, and testing a linguistic tool. *Strategic Management Journal*, *40*(8), 1316–1330. https://doi.org/10.1002/smj.3023

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, *87*(2), 239–269. https://doi.org/10.1016/S0304-4076(98)00015-3

Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal*, *28*(4), 331–343. https://doi.org/10.1002/smj.582

Hoskisson, R. E., Hitt, M. A., Johnson, R. A., & Grossman, W. (2002). Conflicting voices: The effects of institutional ownership heterogeneity and internal governance on corporate innovation strategies. *Academy of Management Journal*, *45*(4), 697–716. https://doi.org/10.2307/3069305

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299–310. https://doi.org/10.1109/TKDE.2005.50

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

Kaur, S., Kumar, P., & Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. *Soft Computing*, *24*(12), 9049–9069. https://doi.org/10.1007/s00500-019-04436-y

Krause, R. (2017). Being the CEO's boss: An examination of board chair orientations. *Strategic Management Journal*, *38*(3), 697–713. https://doi.org/10.1002/smj.2500

Krause, R., Semadeni, M., & Cannella, A. A. (2014). CEO duality: A review and research agenda. *Journal of Management*, *40*(1), 256–286. https://doi.org/10.1177/0149206313503013

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, NY: Springer.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lim, E. N. K., & McCann, B. T. (2013). The influence of relative values of outside director stock options on firm strategic risk from a multiagent perspective. *Strategic Management Journal*, *34*(13), 1568–1590. https://doi.org/10.1002/smj.2088

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In Proceedings of the Association of Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of the Conference on Neural Information Processing Systems.

Miric, M., Jia, N., & Huang, K. G. (2022). Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*. https://doi.org/10.1002/smj.3441

Nadkarni, S., & Chen, J. (2014). Bridging yesterday, today, and tomorrow: CEO temporal focus, environmental dynamism, and rate of new product introduction. *Academy of Management Journal*, *57*(6), 1810–1833. https://doi.org/10.5465/amj.2011.0401

Nadkarni, S., Chen, T., & Chen, J. (2016). The clock is ticking! Executive temporal depth, industry velocity, and competitive aggressiveness. *Strategic Management Journal*, *37*(6), 1132–1153. https://doi.org/10.1002/smj.2376

Ocasio, W. (1994). Political-dynamics and the circulation of power—CEO succession in United-States industrial corporations, 1960-1990. *Administrative Science Quarterly*, *39*(2), 285–312. https://doi.org/10.2307/2393237

Oliver, A. G., Krause, R., Busenbark, J. R., & Kalm, M. (2018). BS in the boardroom: Benevolent sexism and board chair orientations. *Strategic Management Journal*, *39*(1), 113–130. https://doi.org/10.1002/smj.2698

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, *25*, 221–247. https://doi.org/10.2307/2648877

Pekhimenko, G. (2006). Penalizied logistic regression for classification. Working Paper. Department of Computer Science, University of Toronto. Retrieved from https://www.cs.cmu.edu/gpekhime/Projects/CSC2515/project.pdf

Prati, R. C., Batista, G. E., & Silva, D. F. (2015). Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems*, *45*(1), 247–270. https://doi.org/10.1007/s10115-014-0794-3

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.48550/arXiv.1811.10154

Securities and Exchange Commission. (2010). Proxy disclosure enhancements. In 17 § 229.407h.

Seliya, N., Khoshgoftaar, T. M., & Van Hulse, J. A study on the relationships of classifier performance metrics. In Proceedings of the 21st IEEE International Conferrence on Tools with Artificial Intelligence.

Shen, W., & Cannella, A. A. (2002). Power dynamics within top management and their impacts on CEO dismissal followed by inside succession. *Academy of Management Journal*, *45*(6), 1195–1206. https://doi.org/10.2307/3069434

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: Sage.

Spencer Stewart. (2021). Spencer Stuart board index. Retrieved from https://www.spencerstuart.com/-/media/2021/october/ssbi2021/us-spencer-stuart-board-index-2021.pdf

Sundaramurthy, C., & Lewis, M. (2003). Control and collaboration: Paradoxes of governance. *Academy of Management Review*, *28*(3), 397–415. https://doi.org/10.2307/30040729

Westphal, J. D., & Zajac, E. J. (2013). A behavioral theory of corporate governance: Explicating the mechanisms of socially situated and socially constituted agency. *Academy of Management Annals*, *7*(1), 607–661. https://doi.org/10.1080/19416520.2013.783669

Wiersema, M. F., & Zhang, Y. A. (2011). CEO dismissal: The role of investment analysts. *Strategic Management Journal*, *32*(11), 1161–1182. https://doi.org/10.1002/smj.932

Wowak, A. J., Hambrick, D. C., & Henderson, A. D. (2011). Do CEOs encounter within-tenure settling up? A multiperiod perspective on executive pay and dismissal. *Academy of Management Journal*, *54*(4), 719–739. https://doi.org/10.5465/amj.2011.64869961

Wowak, A. J., Mannor, M. J., Arrfelt, M., & McNamara, G. (2016). Earthquake or glacier? How CEO charisma manifests in firm strategy over time. *Strategic Management Journal*, *37*(3), 586–603. https://doi.org/10.1002/smj.2346

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: Harrison, J. S., Josefy, M. A., Kalm, M., & Krause, R. (2023). Using supervised machine learning to scale human-coded data: A method and dataset in the board leadership context. *Strategic Management Journal*, *44*(7), 1780–1802. https://doi.org/10.1002/smj.3480