The Journal of Open Source Software

Machine Reading of Hypotheses for Organizational Research Reviews and Pre-trained Models via R Shiny App for Non-Programmers

Victor Zitian Chen[1], **Evan Canfield**[2], **Felipe Montano Campos**[3], **and Wlodek Zadrozny**[4]

1 Belk College of Business, University of North Carolina at Charlotte **2** Allstate Insurance **3** The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute **4** College of Computing and Informatics, University of North Carolina at Charlotte

DOI: 10.21105/joss.0XXXX

Software

- Review
- Repository
- Archive

Editor: Editor Name

Submitted: 01 January XXXX
Published: 01 January XXXX

ABSTRACT

The volume of scientific publications in organizational research becomes exceedingly overwhelming for human researchers who seek to timely extract and review knowledge. This paper introduces natural language processing (NLP) models to accelerate the discovery, extraction, and organization of theoretical developments (i.e., hypotheses) from social science publications. We illustrate and evaluate NLP models in the context of a systematic review of stakeholder value constructs and hypotheses. Specifically, we develop NLP models to automatically 1) detect sentences in scholarly documents as hypotheses or not (Hypothesis Detection), 2) deconstruct the hypotheses into nodes (constructs) and links (causal/associative relationships) (Relationship Deconstruction), and 3) classify the features of links in terms causality (versus association) and direction (positive, negative, versus nonlinear) (Feature Classification). Our models have reported high performance metrics for all three tasks. While our models are built in Python, we have made the pre-trained models fully accessible for non-programmers. We have provided instructions on installing and using our pre-trained models via an R Shiny app graphic user interface (GUI). Finally, we suggest the next paths to extend our methodology for computer-assisted knowledge synthesis.

STATEMENT OF NEED

TKTKTK

INTRODUCTION

Knowledge accessibility is a significant constraint in synthesizing the scientific literature in organizational research (Chen &; Hitt, 2021; Larsen, Hekler, Paul, &; Gibson, 2020; Li, Larsen, &; Abbasi, 2020). A scientific study typically starts with a systematic review of the existing literature, extracting and connecting the published causes-and-effects relationships among constructs of interest. The information extraction work is recognized widely as one of the most challenging and time-consuming activities for research reviews (Felizardo &; Carver, 2020). The volume of scientific publications is exceedingly overwhelming for human researchers to synthesize the existing knowledge timely (Antons, Breidbach, Joshi, &; Salge, 2021). For instance, a keyword search of "organizational performance" in Web of Science generated about 9,000 papers between 1980-2020, half of which were published in the last five years alone.

Researchers often have to spend limited resources and professional time on tedious manual work of knowledge detection and extraction, yet these efforts may not be sufficiently thorough and timely. It is thus no surprising that recently Antons et al. (2021) call for accessible new methods of computational literature reviews (CLRs) for organizational researchers. They suggest that new methods and tools are needed to engage machine learning algorithms to automatically extract and analyze the content of the text corpus, rather than topics, effect sizes, meta-information, or bibliometric analysis (Antons et al., 2021).

While significant advances have been made in recent years in the field of natural language processing (NLP) to train computers to read and comprehend textual data (e.g., OpenAI's GPT-3) [for a review, see, e.g., Zhang, Yang, Li, and Wang (2019)], there have been limited developments of NLP models to solve the knowledge inaccessibility problem in reviewing the theoretical content of social science papers. Several efforts were made outside social sciences to extract findings, hypotheses, and descriptive information from scientific publications to assist systematic reviews (Felizardo &; Carver, 2020). However, these models are typically built on pre-trained language representations by domain experts and have limited generalizability outside the specific domains where they are developed. So far, almost all the machine reading models for systematic reviews have been developed in biomedicine (Jonnalagadda, Goyal, &; Huffman, 2015; Valenzuela-Escárcega et al., 2018). Despite a growing interest in such tools by social and organizational researchers (Chen &; Hitt, 2021; Larsen et al., 2020; Li et al., 2020), the development of machine reading models for literature reviews in social sciences, especially in organizational research, has been profoundly limited. The current approaches of computational literature reviews focus primarily on topic modeling and sentiment analysis (for a review, see Antons et al., 2021).

The purpose of this research is thus to introduce to organizational researchers interpretable machine reading approaches to reading and organizing theoretical insights from organizational research papers. We develop NLP models to accelerate the detection, classification, and deconstruction of hypotheses from organizational research publications. This paper, to our knowledge, represents the first efforts to develop machine-aided techniques for theoretical knowledge extraction from scientific publications in organizational research. We focus on techniques of detecting hypothesis statements, classifying the causal and associative relationships in these statements, and deconstructing these relationships into entities and links. It is essential to distinguish associations and causal relationships, the latter of which is a stronger statement about the cause-and-effect logic (Pearl, 2009). It is crucial to detect and extract causal knowledge in organizational research, so that researchers and practitioners can draw evidence-based causation to design managerial and policy interventions.

Specifically, we developed machine reading models to complete three sequentially related tasks. The first task was ***hypothesis detection***. We tried to identify whether a statement in a scholarly paper is a hypothesis or not, that is, whether this relationship was deliberately developed as a hypothesis for empirical testing. For this task, we used a model from the **fastText** library. **fastText** is an open-source library that does both word representations and text classification. This type of model has similar performance (e.g., accuracy, precision, etc.) as deep learning models but faster (Zolotov &; Kung, 2017).

Our second task was ***relationship deconstruction***. Specifically, we deconstructed a hypothesis into cause entities, outcome entities. We used a two-layer stacked bi-directional *Long-Short Term Memory (LSTM)* architecture for the model, along with pre-trained GloVe word vectors (Pennington Socher, &; Manning, 2014) for the text embeddings, which yielded good overall performance.

Our third task was ***feature classification*** (***causality and direction***). We classified a hypothesis as to whether it is stating a causal relationship or simply an association and classified the direction of the relationship in the hypothesis (positive, negative, nonlinear). We compared

The Journal of Open Source Software

multiple models and pre-processing methods and found that the logistic regression model outperforms other methods. Furthermore, similar to prior works (e.g., Catalyst Team, 2016), we found that models using bag-of-words (BOW) features outperformed those using other features.

MACHINE READING FOR LITERATURE REVIEWS

Machine reading for literature reviews is to engage NLP models to automate knowledge discovery and extraction from the scientific literature. As an emerging subfield of NLP, machine reading for literature reviews has been developed almost entirely in biomedical research, notably *Textpresso* (Müller, Kenny, &; Sternberg, 2004), *GATE* (Cunningham, Tablan, Roberts, &; Bontcheva, 2013), *Spá* (Kuiper, Marshall, Wallace, &; Swertz, 2014), and *Reach* (Valenzuela-Escárcega et al., 2018). These programs are built on pre-trained language representations of biomedicine, such as a taxonomy of biomedical entities (e.g., proteins) and events (e.g., biochemical interactions) of interest. Most machine reading models work on relatively simple jobs of extracting key findings from paper abstracts [for reviews, see, e.g., Marshall and Wallace (2019) and Jonnalagadda et al. (2015)]. As an exception, Reach (Reading and Assembling Contextual and Holistic mechanisms from text), recently developed by Valenzuela-Escárcega et al. (2018), adapts pre-trained NLP models to read full texts of biomedical databases, extracting biomedical entities (e.g., proteins) and the mechanisms linking these entities (e.g., "influences"). However, as *Reach* is built on biomedical taxonomies and corpus, it has limited application for social science papers.

Our approach follows the general principles of Reach and combines domain-specific rules and machine learning techniques to read the full texts of social science papers. Specifically, our approach takes four steps: Data preparation, hypothesis detection, relationship deconstruction, and relationship classification. Figure 1 illustrates the steps, which are discussed in detail in the following sections.

Figure 1: Overview of Methodological Approach

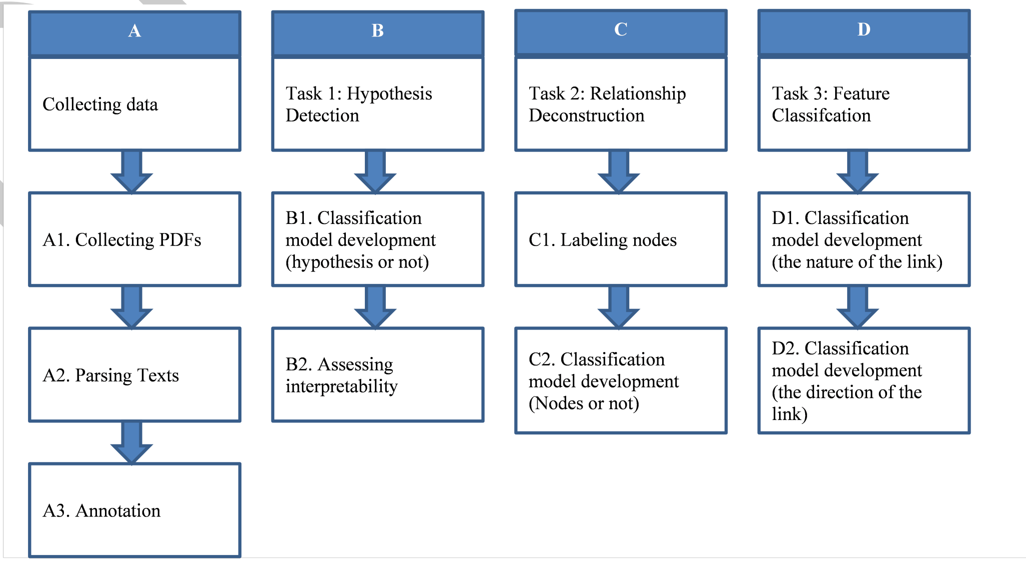


Figure 1: Overview of Approach

DATA PREPARATION

Our approach started with data collection for a sample textual data from publications (Section A in Figure 1). After extracting the hypothesis sentences and manually classifying them as explained above, we then randomly selected a relatively identical sample size of non-hypothesis statements from the same publications. As mentioned above, for hypothesis statements, we labeled each of the extracted sentences with four features: the cause, the outcome, the direction of the relationship, and whether this relationship is causal or not (causality). This labeling practice generally mimics the process of information reduction by human researchers. By reducing a large volume of publications into an annotated corpus, researchers can analyze and organize the four features to briefly understand the main findings of the literature.

Collecting a Corpus

To ground the NLP models into the domain of organizational research (Section A1 in Figure 1), we started by collecting a sample of papers related to organizational research in social sciences. We restrict our search of papers based on the explicit inclusion of organizational performance as part of the research question. In line with the new paradigm of multi-stakeholder and multi-dimensional conceptualization of corporate purpose (Harrison, Phillips, &; Freeman, 2020), we defined organizational performance as an organization's effectiveness in meeting the expectations of two or more stakeholder groups (investors, employees, customers, and communities).

Based on the ISI Web of Science database of publications, all empirical publications (excluding meta-analysis) were first downloaded and read, as long as at least one keyword was directly suggesting a stakeholder group. The keywords indicating stakeholders were: *stakeholder**, *investor**, *shareholder**, *owner**, and *financ** for investors; *customer**, *consumer**, and *user** for consumers; *employee**, *worker**, *workforce**, *labor**, *labour**, and *human resource** for employees; and *communit**, *societ**, *environment**, *climate**, *natural resource**, *responsib**, and *social performance** for the community. A snowball approach was adopted, in which each newly found performance construct will be added as a new keyword for the next search until no new construct was found. With the pool of papers collected above, we further shortlisted papers that included theoretical developments related to performance measures concerning at least two stakeholder groups. This sample represents high-quality scientific journal articles and offers a viable corpus of testable knowledge (i.e., hypotheses) concerning organizational performance.

The primary studies included two stakeholder groups for measuring organizational performance: the correlations between a factor, and at least two stakeholder values. In total, we have identified and downloaded 138 peer-reviewed articles published between 1990 and 2018. We further removed 13 papers of which the PDFs were of poor quality for optical character recognition (OCR). The remaining 125 papers represent cross-disciplinary literature in social sciences in 1990-2018 to explain different organizational performance dimensions. The complete reference of these papers is listed in **Supplementary Materials S1**.

Developing a Sample for Hypothesis Detection

We prepared a corpus for NLP model development (Section A2 in Figure 1). First, we converted each PDF (e.g., "paper.pdf") to raw text ("paper.txt"). We removed any tables, figures, and commonly used stop-words from articles (using the built-in dictionary by Python NLTK package). We then continued developed an algorithm to identify which statements are likely hypotheses. Specifically, the algorithm works like the following. It detects any statements in a format similar to the following:

> *"Hypothesis 1: …"*
>
> *"H1: …"*

We trained the algorithm to search for sentences that included targeted expressions such as "Hypothesis" (or "Proposition") or "H" (or "P") followed by a number. This gave us 2,230 sentences that potentially contained hypotheses. We ended up with many false-positive extractions (i.e., sentences that contained the targeted expressions related to hypotheses but were not the original hypothesis statements, simply explanations or mentions of them). For instance, researchers often refer to a hypothesis when discussing the evidence. We screened all the 2,230 sentences manually and kept actual hypothesis statements. This yielded 643 hypothesis statements across our 125 papers.

Below is an example of extracted hypothesis sentences:

> *"H1. Commitment configuration is positively associated with firm performance."*

Finally, we constructed a relatively balanced corpus of 1,300 sentences by randomly drawing from the same publications 657 non-hypothesis sentences that also included the word "Hypothesis" (or "Proposition") or "H" (or "P") followed by a number. Essentially, we aimed to train classification models to distinguish the original hypotheses from the in-text mentions of them (e.g., discussion of empirical findings for a hypothesis).

Annotating Features of Hypothesis Statements

The next task was to develop models to extract information from each hypothesis statement. The objective was to reduce each hypothesis to its four key features: node 1 (a construct), node 2 (another construct), the direction of the link (positive, negative, or nonlinear), and the nature of this link (causal or associative statements). Below are two sets of examples that were classified as causal statements and association statements, respectively.

Examples of causal statements:

> *"H1: The environmental legislation exerts a positive influence on the manager's perception about the environment as a competitive opportunity."*
>
> *"H1: Stakeholder management will have a positive effect on CEO compensation levels."*

Examples of association statements:

> *"H1: Stakeholder relations are negatively associated with the persistence of inferior financial performance."*
>
> *"H1: The grafting of new management team members after venture start-up is positively related to venture performance."*

We manually classified each hypothesis sentence into nodes, the direction of the link, and the nature of the link. We use these features as inputs to perform classification tasks later. Six well-trained graduate students in data science from an elite university completed the feature coding work. Each statement was coded by two different students independently.

The Journal of Open Source Software

The inter-coder agreement was 95%, with the remaining disagreements fully resolved after a direct conversation. A co-author who specializes in organizational research played quality control to make sure the final coding was 100% correct. As an example, the last hypothesis statement cited earlier was annotated into the following features: Node 1 ("the grafting of new management team members after venture start-up"), Node 2 ("venture performance"), the direction of the link (positive), and the nature of the link (association). In total, we have manually completed these annotations for the 643 hypotheses that we extracted.

A Summary of the Annotated Corpus

The 643 hypothesis statements reported a mean of seven hypotheses per article and a standard deviation of five hypotheses. Typically, a set of hypothesis statements is one or three sentences long. As a reference, generally, an English sentence has on average 15 to 20 words (Plain English Campaign, 2004). Thus, we censored extractions by dropping sentences with more than 60 words, assuming they are not hypotheses in any organizational research papers. As Figure 2 illustrates, after this censoring, each hypothesis statement's number of words was approximately following a normal distribution, with a mean of 18.5 words and a standard deviation of 9.8 words. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Figure 2: Number of Words per Sentence in Our Corpus

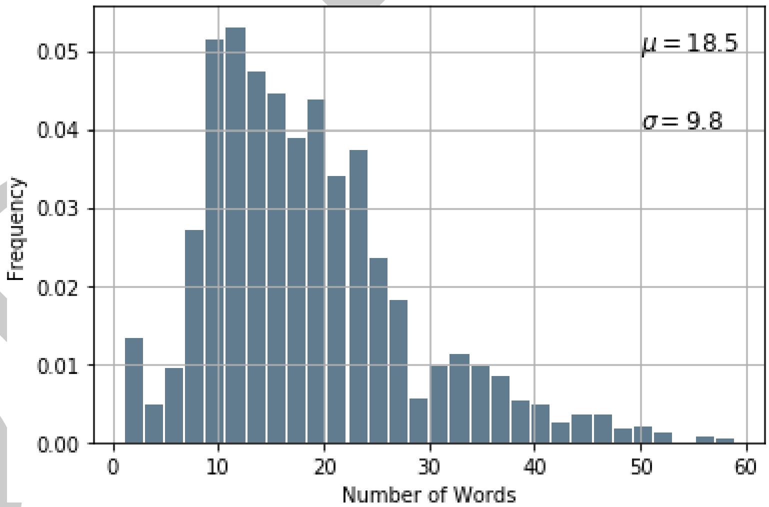


Figure 2: Sentence Word Count

TASK 1: HYPOTHESIS DETECTION

After constructing the corpus, we develop text classification models to detect whether a sentence is a *hypothesis sentence* or not (Section B in Figure 1). As mentioned earlier, our corpus contained our final sample contained 1,300 sentences (including 643 hypothesis statements and randomly extracted 657 non-hypothesis sentences from the same sample of publications). This corpus was then divided for 10-fold cross-validation. Specifically, the whole sample was randomly split into ten subsamples. In each testing, nine subsamples (90% of the entire sample) were used as the training set to train the text classification models for identifying hypothesis statements. The remaining subsample (10% of the whole sample) was

The Journal of Open Source Software

used to measure the training model's out-of-sample performance. We repeated this process ten times, in each of which we used a different 10% subsample as the test. We then reported the average out-of-sample performance as the overall performance of the training model. By averaging performance in ten sets of testing in different sets of subsamples, we would avoid overfitting bias. We also replicated the division to 75% training set and 25% test set and received highly consistent results.

Text classification models do not need to understand the meanings or grammatical structures within texts. Instead, we let statistical models predict the classification (1 for hypothesis and 0 for non-hypothesis). We fed text classification models with features of a sentence to find statistical relationships between features (inputs) of a raw sentence and the classification of this sentence (output). For example, if the word "associated"were more related to hypothesis sentences than non-hypotheses, the model would be more likely to classify sentences with the word "associated" as a *hypothesis* without knowing its meaning. Our sample met the requirement for successful text classification models requirement, as it covered a wide range of possible hypothesis-related words.

We used the text classification models from the **fastText** library – a supervised machine learning model –to classify sentences as a hypothesis or not (Section B1 in Figure 1). Facebook's AI Research group created this algorithm to learn word embeddings and perform text classification. This model has been shown to have similar performance (e.g., accuracy, precision, and f1-scores) as more complex deep learning models but at a significantly faster speed (Zolotov &; Kung, 2017). Thus, it meets the purpose of our project, that is, saving time for research reviews.

Specifically, the algorithm of **fastText** model works as the following:

1. A sentence apart into separate tokens. Each token is a commonly used clause term or a word;
2. Every token in the training sample is assigned an n-dimensional numerical vector (word embedding);
3. Every sentence is assigned an n-dimensional numerical vector that averages the values of every dimension of the word's vectors in the sentence (sentence embedding);
4. The sentence embeddings are used as features (inputs) into a supervised classification model to predict the classification (hypothesis or non-hypothesis).

After comparing the preliminary performance of different linear and nonlinear supervised models, we used a neural network with one hidden layer and iterated through word and sentence embeddings. The embedding for a given sentence and its associated label vector were very close to each other in a vector space. Finally, sentence embeddings were used as features for the final prediction.

Table 1: Evaluation of Hypothesis Detection Models

Model	N-grams	Learning Rate	f1-score
SoftMax	Neg Sampling		
—	—		
Parametrization 1	1	0.1	87.10%
—	—	—	—
Parametrization 2	2	0.1	84.60%
Parametrization 3	5	0.1	85.10%
Parametrization 4	1	0.3	95.70%

Note: We used 120-dimensional vectors. The f1-Score was calculated using two loss functions: Soft Max and Negative Sampling. We used word N-grams (N=1, 2, and 5).

We trained the **fastText** model and tuned model parameters (parametrization). Table 1 presents the four best-performing parametrizations alignments. We find the order of words played no effect on the results of identifying hypothesis sentences. Furthermore, models using bi-grams, compared to those using uni-grams, reported a lower accuracy under all specifications. Also, the negative-sampling loss provided a better accuracy under most specifications. The best specification was Parametrization 4 in Table 1, which used uni-grams, a learning rate of 0.3, a 120-dimension vector to represent words, and the negative-sampling loss function. As presented in Table 1, we achieved an F-1 score of 96.7% for this specification on the test data, where the F-1 score is a comprehensive measure of model accuracy combining Precision and Recall. *Precision* is the ratio between the true positives (correctly predicted hypotheses) and all the positives (correctly and incorrectly predicted hypotheses), and *Recall* is the measure of our model correctly identifying true positives (percentage of correctly predicted hypotheses among all actual hypotheses). The f-1 score is the harmonic mean of *Precision* and *Recall*, providing a balanced metric for optimization between the two values.

Assessing the Interpretability of the Model

One limitation of machine learning models is that they are often difficult to interpret. As a result, it cannot be trusted that these models have picked up the data's meaningful features. For instance, if hypothesis sentences are on average shorter (or longer) than non-hypothesis sentences in our sample, then the model might have classified short (or long) sentences as hypotheses and others as non-hypotheses. In this case, the model would report a high accuracy but is not based on meaningful features that define a hypothesis and thus may not perform effectively in new samples.

Ribeiro, Singh, and Guestrin (2016) introduced an approach to interpreting complex machine learning models, named Local Interpretable Model-Agnostic Explanations (LIME). Following LIME, we need to explain how the **fastText** model predicts by training a simpler stand-in model, then use this simpler stand-in model to explain the original **fastText** model's prediction (Section B2 in Figure 1). Even though the simpler model cannot capture all of the **fastText** model's complexity, it helps to understand the logic the complex model might have used. Instead of training the stand-in model on the entire sample, we used a subsample of the data for the stand-in model to classify one sentence correctly. As long as the stand-in model used the same logic as the **fastText** model, we would understand and explain the predictions made by **fastText**.

To construct the stand-in model's training set, we created many variations out of each sentence, each time removing specific words. In hypothesis detection, we classified a hypothesis sentence multiple times by removing a different word each time from the sentence. In this way, we estimated each word's relative importance in the final prediction. By making several predictions for many variations of the same sentence using **fastText** (i.e., missing different words), we were essentially capturing how the model weighted different words as a way of "understanding" that sentence. Finally, we used the sentence variations and classification predictions as the training set to train the stand-in model using the Simple Linear Classification Model.

We want to note that this approach's shortcoming is the implicit focus on only the importance of single words, not phrases or n-grams. However, as we will show, this limitation does not prevent us from making reasonable interpretations of fastText. Specifically, our stand-in model's outputs were the weights assigned to each word in the hypothesis sentence, where the weights represent how much that word affected the final prediction.

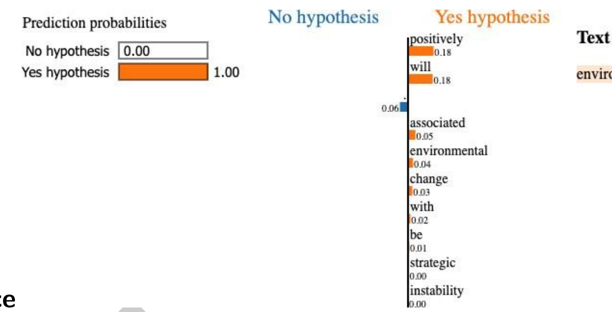


Figure 3: An Example of Hypothesis Sentence

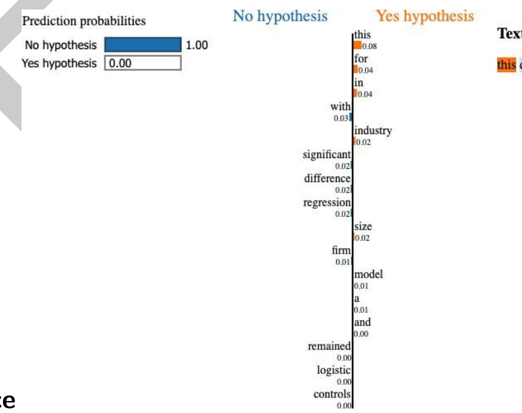


Figure 4: An Example for a Non-Hypothesis Sentence

Figure 3 shows that the words "positively" and "associated" were among the most important words as they contributed the most to the classification of a sentence as a hypothesis. Figure 4 shows that the words contributing the most to classifying a sentence as a *non-hypothesis* were "significant" and "regression." They are usually not part of the original hypothesis sentence but were used to discuss the empirical test for or against the hypothesis. However, no word in this sentence was strongly associated with a hypothesis sentence. Therefore, from these two figures, it seems clear that the **fastText** model was valuing the correct words to make predictions regarding hypothesis detection.

TASK 2: RELATIONSHIP Deconstruction

We then developed our NLP model to extract the key features in a relationship from each hypothesis, including two nodes (constructs) and the link between them from a sentence (Section C in Figure 1). For example, if we have an association statement, *"Node1 is related to Node2,"* we want to extract both **Node1** and **Node2**. But if we have a causal statement, *"Node 1 causes Node 2,"* then we need to not only extract **Node1** and **Node 2**, but also identify **Node 1** as the cause and **Node 2** as the outcome.

First, we labeled the nodes in our sample data. For each hypothesis sentence, we labeled non-nodes as "0," the "cause" node as "1," and the "outcome" node as "2." In the case of atypical hypotheses such as more than two nodes (e.g., multiple causes or outcomes) and more than one link (e.g., moderators), we aggregated multiple nodes of the same level together to form a Node 1 -link- Node 2 structure. Specifically, in the case of more than two nodes, such as "A would reduce B and C," we treated "A" as Node 1 and "B and C" together as Node 2. In the case of multiple links, such as "A is moderating the relationship between B and C," we treated "A" as Node 1 and "the relationship between B and C" together as Node 2.

Six well-trained graduate students in data science completed the feature coding work. Each statement was coded by two different students independently. The inter-coder agreement was 90%, with the remaining disagreements fully resolved after a direct conversation. A co-author who specializes in organizational research played quality control to ensure the final coding was 100% correct.

We padded each of the sentences, so they were formatted to have the same dimension of 50 (i.e., the vector dimension). We then fitted the data to a model with the following architecture listed:

1. Text vectorization layer, which standardizes each text and utilizes only uni-grams;
2. Embedding layer, which applies the pre-trained words vectors based on the GloVe dataset;
3. One dimensional spatial dropout layer with a dropout rate of 0.5;
4. Two-layer stacked bi-directional LSTM, with 32 units on the first layer, and 128 units on the second, both with a recurrent dropout rate of 0.1;
5. Time-distributed dense output layer;

In addition, we use the RMSprop back-propagation optimizer, with loss calculated by categorical cross-entropy. Complete visualization of the model can be seen in Appendix 1, generated via Net2Viz (Alex Bäuerle &; Timo Ropinski, 2019).

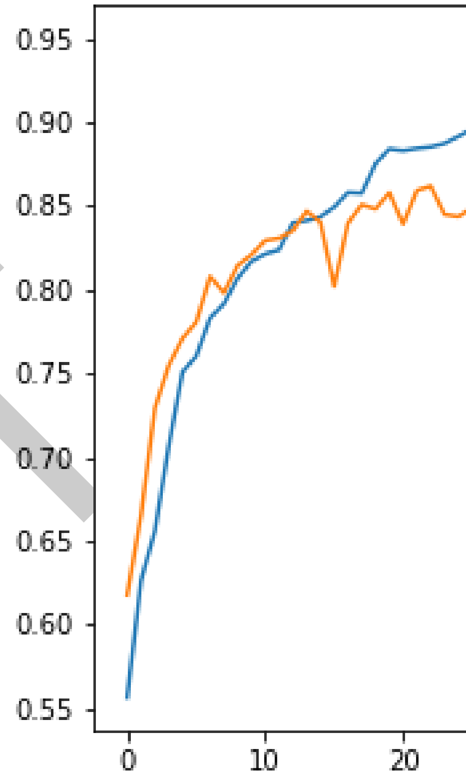


Figure 5: Performance of Relationship Deconstruction

Table 2: Evaluation of Relationship Deconstruction Models

Precision	Recall	f1-Score	
Overall (All Nodes)	92.4%	91.9%	92.2%
—	—	—	—
Non-Label (0)	98.6%	98.7%	98.6%
Cause (1)	88.8%	89.9%	89.4%
Outcome (2)	89.8%	87.2%	88.5%

We ran the model with a batch size of 32 and 50 epochs to minimize training overfitting. Figure 5 shows the training and test accuracy over the number of epochs. We received an accuracy 97.2% from the test data, measuring the total percentage of true positives (correctly predicted nodes and links) and true negatives (correctly predicted non-nodes and non-links). However, the dataset is highly imbalanced, with approximately 90% of all tokens representing non-node or non-link entities, 5% representing cause entities, and 5% representing outcome entities. Thus, accuracy may be an inappropriate indicator of model performance, and we need to rely on additional performance metrics, including precision, recall, and f1-score, to evaluate the model. Table 6 shows the additional metrics on different predictions, all of which are satisfactory – significantly over 90% in all measures.

TASK 3: FEATURE CLASSIFICATION

Classifying the Nature of the Link (Causality or Association)

We moved on to develop a model to classify if a sentence made a causal statement or not (Section D1 in Figure 1). We created two different representations from each hypothesis. The first representation was word embedding based on **BOW** features. Specifically, we identified the frequency of uni-gram, bi-gram, and tri-grams against the complete corpus (1,300 sentences). The second representation was a sentence embedding using **Doc2Vec (D2V)**. With these two different representations, multiple classification models were evaluated. For both **BOW** and **D2V** features, we used and evaluated the following classification models: logistic regression, random forest, and support vector machine (SVM). We also used synthetic oversampling methods **SMOTE** and **ADASYN**, which did not exhibit any significant model improvements. Thus, synthetic data was not included in the model.

Table 3: Evaluation of Models using BOW Features to Classify the Nature of the Link

Model	Feature Normalization	Accuracy	Precision	Recall	f1-Score
Logistic Regression*	Stemming	93.7%	93.5%	91.4%	92.4%
—	—	—	—	—	—
Random Forest	Lemmatization	90.6%	94.0%	84.4%	87.6%
Support Vector Machines	Stemming	93.1%	92.5%	90.9%	91.6%

* Model with the greatest f1-score as the overall performance measure.

Table 4: Evaluation of Models using D2V Features to Classify the Nature of the Link

Model	Feature Normalization	Accuracy	Precision	Recall	f1-Score
Logistic Regression	Stemming	73.6%	68.7%	62.2%	63.0%
—	—	—	—	—	—
Random Forest*	Lemmatization	77.4%	78.5%	64.9%	66.3%
Support Vector Machines	Lemmatization	70.4%	85.1%	51.0%	43.3%

* Model with the greatest f1-score as the overall performance measure.

Prediction performance metrics of different classification models with **BOW** and **D2V** are reported in Tables 3 and 4, respectively. Models using **BOW** representation generally performed better than **D2V** representation. Among all evaluations, logistic regression using **BOW** features produced the greatest f1-score. We further tuned the hyperparameters on this model, using stratified 10-fold cross-validations and three repeats. This hyperparameter tuning yielded a further improved f1-score as high as 92.4% (see Table 3).

Classifying the Direction of the Link (Positive, Negative, or Nonlinear)

Table 5: Evaluation of Models using BOW Features to Classify the Direction of the Link

Model	Feature Normalization	Accuracy	Precision	Recall	f1-Score
Logistic Regression*	Stemming	91.3%	87.6%	84.6%	85.9%
—	—	—	—	—	—
Random Forest	Stemming	85.7%	89.3%	67.5%	72.0%
Support Vector Machines	Stemming	85.7%	80.9%	70.7%	74.4%

* Model with the greatest f1-score as the overall performance measure.

Table 6: Evaluation of Models using D2V Features to Classify the Direction of the Link

Model	Feature Normalization	Accuracy	Precision	Recall	f1-Score
Logistic Regression*	Lemmatization	70.2%	47.4%	38.1%	38.7%
—	—	—	—	—	—
Random Forest	Stemming	73.9%	58.2%	35.7%	33.6%
Support Vector Machines	Lemmatization	73.9%	24.6%	33.3%	28.3%

* Model with the greatest f1-score as the overall performance measure.

We then trained a model to classify the direction of the link in a hypothesis (positive, negative, or nonlinear) (Section D2 in Figure 1). This process used the same feature representations (BOW and D2V), models, and oversampling methods as the feature classification model. Prediction performance metrics for different classification models with **BOW** and **D2V** are reported in Tables 5 and 6, respectively.

Models using **BOW** representation generally performed better than **D2V** representations. Logistic regression using **BOW** features produced the greatest f1-score. We further tuned the hyperparameters on this model, using stratified 10-fold cross-validations and three repeats. This hyperparameter tuning yielded a further improved f1-score as high as 85.9% (see Table 3).

A USER'S GUIDE

In this project, we constructed an interdisciplinary corpus of hypothesis statements from a set of high-quality peer-reviewed papers in social sciences. Then we used this data to train models that perform three different tasks that mimic how human researchers typically extract theoretical insights from the literature for research reviews. We recognize that most organizational researchers would be direct users of the existing pre-trained models for machine-reading, rather than those who have the programming background to re-train the models for a different task. For this large audience, we have made several efforts to make our models fully accessible, that is, a simple drag-and-drop with minimum coding. First, we have developed a free R Shiny app as the graphic user interface (GUI). On this GUI, users can upload an unlimited volume of papers as PDFs to initiate the pre-trained models to automatically parse texts into a corpus and then play all three tasks. Second, we have connected the R Shiny app through r-reticulate package to convert the Python programs into R programs. The R Shiny app then runs on both R and Python programs on the back end.

Now we illustrate how users without a programming background can install and use our pre-trained models via an R Shiny app in detail.We have developed the R package HypothesisReader and stored it on Github for remote installations. The package implements the methodology outlined in this paper and automatically launches the pre-trained models for users' own PDF data. The following software should be pre-installed in a user's computer.

1. Java 8 or OpenJDK 1.8
2. R and R package "devtools"

Installation Steps

1. Open R and install R package from GitHub repository by typing the following:

   devtools::install_github("canfielder/HypothesisReader")

When prompted "Enter one or more numbers, or an empty line to skip updates:" simply hit the Enter key;

1. Execute the function below to launch the R Shiny app GUI:

   HypothesisReader::LaunchApp()

2. Upload PDFs on the GUI to initiate the text processing and install Python package;
3. At the prompt in the console, select **y** to install Miniconda;
4. Restart R session (Session > Restart);
5. The pre-trained models are now ready for use.

Troubleshooting

If any of the required Python packages do not automatically install (which would yield an error), installation can be forced with the following function in R:

   HypothesisReader::InstallHypothesisReader()

Usage

Finally, we provide a step-wise illustration of using our pre-trained models via an R Shiny app GUI. As shown in Appendix 2, using the tool takes three simple steps:

1. Launch the GUI through R
2. Upload the PDF data
3. Download the deconstructed data in CSV

DISCUSSION

We suggest several directions of future research are valuable for improving our models. First, our models currently force each hypothesis into a three-part structure – two nodes and one link. The majority (82%) of the hypotheses in our sample follow this structure to contain two separate constructs. However, there are exceptions, such as moderators and multiple causes or outcomes. Currently, our models would aggregate nodes or links at the same level to force a hypothesis into three parts. Such cases include a) more than two nodes or b) more than two links (moderators and, in rare cases, mediators). As an example for more than two nodes, take the following sample hypothesis which has multiple outcomes:

> "increased use of high-performance work systems results in increased labor productivity, increased workforce innovation, and decreased voluntary employee turnover."

Currently, our pre-trained models would deconstruct it into Node 1 (*"increased use of high-performance work systems"*), Node 2 (*"increased labor productivity, increased workforce innovation, and decreased voluntary employee turnover"*), and a link (nature=positive; causality=1). However, the ideal outputs should be three relationships with a shared Node 1 (*"use of high-performance work systems"*):

a) Node 2 as *"labor productivity"* with a link (nature=positive; causality=1);
b) Node 2 as *"workforce innovation"* with a link (nature=positive; causality=1);
c) Node 2 as *"voluntary employee turnover"* with a link (nature=negative; causality=1).

As an example for more than two links, our sample contains some hypotheses on moderating effects like *"the positive relationship between corporate philanthropy and a firm's financial performance increases with its advertising intensity."* Currently, our pre-trained models would deconstruct this relationship into:

- Node 1: *"the positive relationship between corporate philanthropy and a firm's financial performance"*,
- Node 2: *"advertising intensity"*, and
- Link: nature=positive; causality=0.

The ideal outputs should generate an additional relationship with:

- Node 1: *"corporate philanthropy"* ,
- Node 2: *"a firm's financial performance"*, and
- Link: nature=positive; causality=0.

As another example for more than two links, our sample contains hypotheses that combine two causal relationships through a mediating process, such as *"marketing competence mediates the relationship between CSR toward society and firm performance."* Currently, our pre-trained models would deconstruct this relationship into:

- Node 1: *"marketing competence"*,
- Node 2: *"the relationship between CSR toward society and firm performance"*, and
- Link: nature=nonlinear; causality=1.

However, the ideal outputs should divide this relationship into two causal relationships. The first relationship should have:

- Node 1: *"CSR toward society"*,
- Node 2: *"marketing competence"*, and
- Link: nature=positive; causality=1.

The second relationship should have:

- Node 1: *"marketing competence"*,
- Node 2: *"firm performance"*, and
- Link: nature=positive; causality=1.

Currently, our training is limited by the small sample of such exceptional cases. We propose to increase the size of our sample by annotating a more extensive corpus that contains significantly more atypical hypotheses, including more than two nodes, moderators, and mediators. A larger sample would also significantly improve the training and the out-of-the-sample performance.

Second, we suggest future studies should also develop clustering models to sort and aggregate extracted nodes into a standardized taxonomic hierarchy. For instance, after deconstruction, our sample contains expressions of nodes like "CSR towards society," "social performance," and "social responsibility." Currently, the outputs would export the original forms of each, and thus would treat them as different constructs. We propose to develop a standardized taxonomy of commonly used terms in organizational research to sort and aggregate semantically similar constructs into the same new construct. For instance, the three mentioned examples could be grouped into a new construct called "firm performance towards the society." As the literature continues to evolve and grow, a challenge is that many constructs may be introduced to the field without precisely fitting into an existing taxonomy. We suggest a highly valuable approach would be to use unsupervised learning to cluster contructs automatically without a pre-defined taxonomy. We suggest that researchers draw a larger corpus of research documents such as company reports, Wikipedia, and textbooks to triangulate each construct's semantically adjacent words (e.g., N-grams) and use adjacent words to cluster constructs together.

Finally, we suggest that researchers with advanced NLP training can further refine our methodology and re-train our models for different tasks. Currently, our approach applies only to hypotheses, that is, testable theoretical statements. As literature reviews are often accompanied by empirical syntheses such as meta-analysis and meta-regressions, researchers often would like to detect and extract the empirical findings. Researchers could go beyond hypotheses and focus on detecting and comparing empirical evidence by focusing on a different set of trigger words. Rather than using only "Hypothesis" (or "Propositions) or"H" (or "P") followed by a number, we could combine them and with other trigger words indicating empirical evidence, such as "support," "supportive," "evidence," "significant," and so on. In this way, we could train models to detect empirical findings and classify each hypothesis as "supported" or "unsupported." This, however, would be more challenging to develop, as not all empirical evidence is mentioned in the text. Many empirical details, such as coefficients and p-values, are only reported in Tables without specific mentions in the paper. However, for meta-analytic reviews, it would also require that the machine reading models extract the same information in papers where a focal variable was tested only as a control variable and thus unmentioned specifically as hypotheses anywhere in the paper.

REFERENCE

Antons, D., Breidbach, C. F., Joshi, A. M., &; Salge, T. O. (2021). Computational literature reviews: Method, algorithms, and roadmap. *Organizational Research Methods*, In-Press.

Bäuerle, A., &; Ropiski, T. (2019). Transforming deep convolutional networks into publication-ready visualization. *arXiv preprint,* arXiv:1902.04394.

Catalyst Team. (2016). *Corpus to graph ML*. Accessible at https://github.com/CatalystCode/corpus-to-graph-ml.

Chen, V. Z., &; Hitt, M. A. (2021). Knowledge synthesis for scientific management: practical integration for complexity versus scientific fragmentation for simplicity. *Journal of Management Inquiry, 30*(2), 177-192.

Cunningham, H., Tablan, V., Roberts, A., &; Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open-source text analytics. *PLoS Computational Biology, 9*(2), e1002854.

Felizardo, K. R., &; Carver, J. C. (2020). Automating systematic literature review. *Contemporary Empirical Methods in Software Engineering*, 327-355.

Harrison, J. S., Phillips, R. A., &; Freeman, R. E. (2020). On the 2019 business roundtable "statement on the purpose of a corporation." *Journal of Management, 46(7)*, 1223-1237.

Jonnalagadda, S. R., Goyal, P., &; Huffman, M. D. (2015). Automating data extraction in systematic reviews: a systematic review. *Systematic reviews, 4*(1), 78.

Kuiper, J., Marshall, I. J., Wallace, B. C., &; Swertz, M. A. (2014). *Spá: A web-based viewer for text mining in evidence-based medicine.* Paper presented at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Larsen, K. R., Hekler, E. B., Paul, M. J., &; Gibson, B. S. (2020). Improving usability of social and behavioral sciences' evidence: a call to action for a National Infrastructure Project for mining our knowledge. *Communications of the Association for Information Systems, 46*(1), 1.

Li, J., Larsen, K., &; Abbasi, A. (2020). TheoryOn: A design framework and system for unlocking behavioral knowledge through ontology learning. *MIS Quarterly*, 1-48.

Marshall, I. J., &; Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews, 8*(1), 163.

Müller, H.-M., Kenny, E. E., &; Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biology, 2*(11), e309.

Pearl, J. (2009). *Causality*. Cambridge, UK: Cambridge University Press.

Pennington, J., Socher, R., &; Manning, C. D. (2014). GloVe: *Global vectors for word representation*. Accessible at https://nlp.stanford.edu/projects/glove/. Stanford, CA: Stanford University.

Plain English Campaign. (2004). *How to write in plain English*. Kent, UK: The University of Kent.

Tulio Ribeiro, M., Singh, S., &; Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*, arXiv-1602.

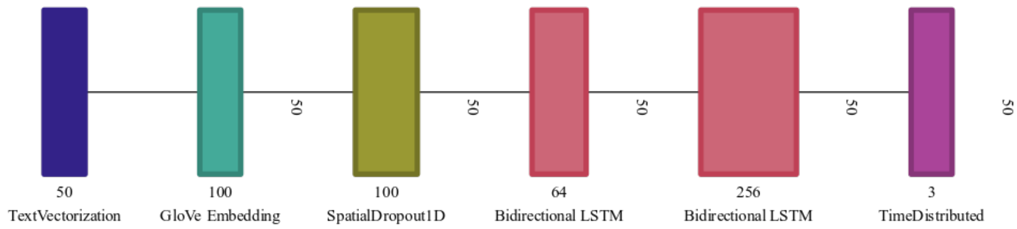
Valenzuela-Escárcega, M. A., Babur, Ö., Hahn-Powell, G., Bell, D., Hicks, T., Noriega-Atala, E., . . . Morrison, C. T. (2018). Large-scale automated machine reading discovers new cancer-driving mechanisms. *Database, 2018*, 1-14.

Zhang, X., Yang, A., Li, S., &; Wang, Y. (2019). Machine reading comprehension: a literature review. *arXiv preprint,* arXiv:1907.01686.

Zolotov, V., &; Kung, D. (2017). Analysis and optimization of **fastText** linear text classifier. *arXiv preprint,* arXiv:1702.05531.

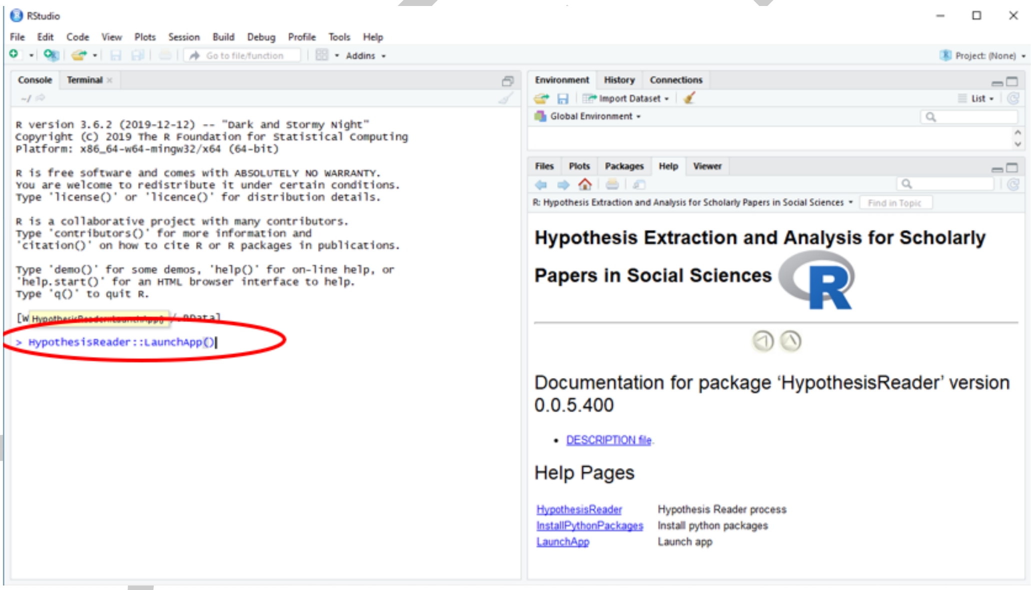
The Journal of Open Source Software

Appendix 1: Relationship Deconstruction Model Structure

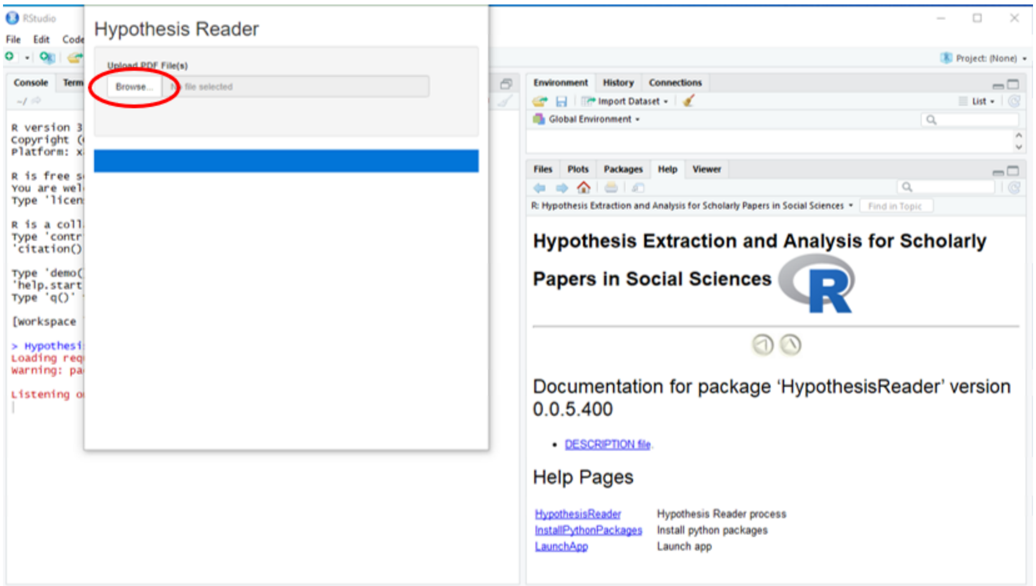


Appendix 2: Usage of Pre-trained Models via R Shiny App

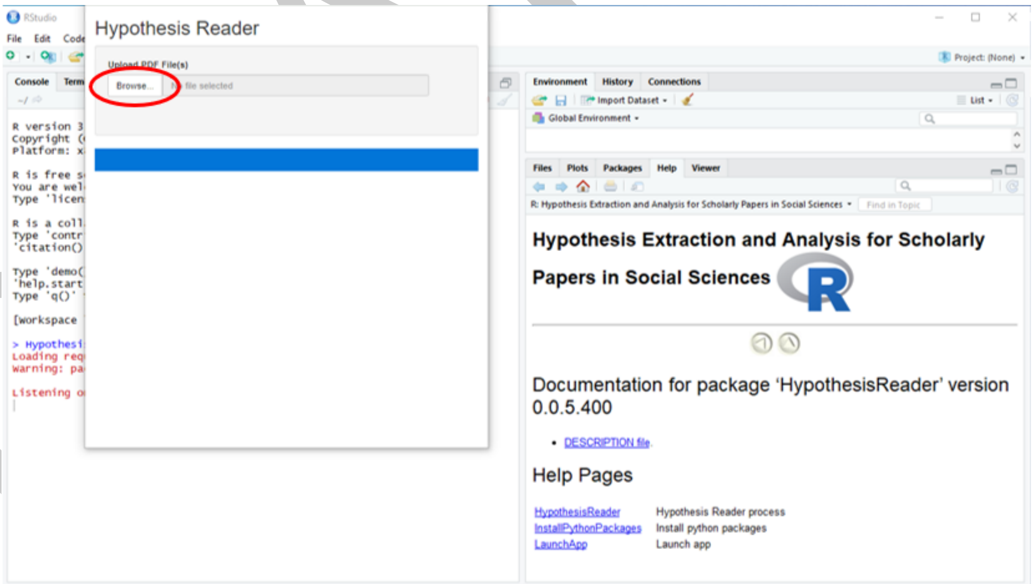
Step 1: Launch Pre-trained Models via R Shiny GUI



Step 2: Upload all PDFs by clicking the Browse button



Step 3: Download the Deconstructed Data of Hypotheses



Supplementary Materials

S1: Studies included in the corpus

1. Abbott, W. F., & Monsen, R. J. 1979. On the measurement of corporate social responsibility: Self-reported disclosures as a method of measuring corporate social involvement. *Academy of Management Journal*, *22*: 501–515. http://doi.org/10.5465/255740

2. Abdullah, N. A. H. N., & Yaakub, S. 2014. Reverse logistics: Pressure for adoption and the impact on firm's performance. *International Journal of Business and Society*, *15*: 151–170.

3. Akhtar, S., Ding, D. Z., & Ge, G. L. 2008. Strategic HRM practices and their impact on company performance in Chinese enterprises. *Human Resource Management*, *47*: 15–32. http://doi.org/10.1002/hrm.20195

4. Alexander, G. J., & Buchholz, R. A. 1978. Corporate social responsibility and stock market performance. *Academy of Management Journal*, *21*: 479–486.

5. Angle, H. L., & Perry, J. L. 1981. An empirical assessment of organizational commitment and organizational effectiveness. *Administrative Science Quarterly*, *26*: 1–14.

6. Aragón-Correa, J. A., Hurtado-Torres, N., Sharma, S., & García-Morales, V. J. 2008. Environmental strategy and performance in small firms: A resource-based perspective. *Journal of Environmental Management*, *86*: 88–103. http://doi.org/10.1016/j.jenvman.2006.11.022

7. Armstrong, C., Flood, P. C., Guthrie, J. P., Liu, W., MacCurtain, S., & Mkamwa, T. 2010. The impact of diversity and equality management on firm performance: Beyond high performance work systems. *Human Resource Management*, *49*: 977–998. http://doi.org/10.1002/hrm.20391

8. Arthur, J. B. 1994. Effects of human resource systems on manufacturing performance and turnover. *Academy of Management Journal*, *37*: 670–687. http://doi.org/10.2307/256705

9. Audea, T., Teo, S. T. T., & Crawford, J. 2005. HRM professionals and their perceptions of HRM and firm performance in the Philippines. *The International Journal of Human Resource Management*, *16*: 532–552. http://doi.org/10.1080/09585190500051589

10. Bae, J., & Lawler, J. J. 2000. Organizational and HRM strategies in Korea: Impact on firm performance in an emerging economy. *Academy of Management Journal*, *43*: 502–517. http://doi.org/10.2307/1556407

11. Bai, X., & Chang, J. 2015. Corporate social responsibility and firm performance: The mediating role of marketing competence and the moderating role of market environment. *Asia Pacific Journal of Management*, *32*: 505–530. http://doi.org/10.1007/s10490-015-9409-0

12. Baker, W. E., & Sinkula, J. M. 2005. Environmental marketing strategy and firm performance: Effects on new product performance and market share. *Journal of the Academy of Marketing Science*, *33*, 461–475. http://doi.org/10.1177/0092070305276119

13. Batt, R., & Colvin, A. J. 2011. An employment systems approach to turnover: Human resources practices, quits, dismissals, and performance. *Academy of Management Journal*, *54*, 695–717. http://doi.org/10.5465/amj.2011.64869448

14. Beltrán-Martín, I., Roca-Puig, V., Escrig-Tena, A., & Bou-Llusar, J. C. 2008. Human resource flexibility as a mediating variable between high performance work systems and performance. *Journal of Management*, *34*: 1009–1044. http://doi.org/10.1177/0149206308318616

15. Ben Brik, A., Rettab, B., & Mellahi, K. 2010. Market orientation, corporate social responsibility, and business performance. *Journal of Business Ethics*, *99*: 307–324. http://doi.org/10.1007/s10551-010-0658-z

16. Bernhardt, K. L., Donthu, N., & Kennett, P. A. 2000. A longitudinal analysis of satisfaction and profitability. *Journal of Business Research*, *47*: 161–171. http://doi.org/10.1016/S0148-2963(98)00042-3

17. Bhattacharya, M., Gibson, D. E., & Doty, D. H. 2005. The effects of flexibility in employee skills, employee behaviors, and human resource practices on firm performance. *Journal of Management*, *31*: 622–640. http://doi.org/10.1177/0149206304272347
18. Bingley, P., & Westergaard-Nielsen, N. 2004. Personnel policy and profit. *Journal of Business Research*, *57*: 557–563. http://doi.org/10.1016/S0148-2963(02)00321-1
19. Bird, A., & Beechler, S. 1995. Links between business strategy and human resource management strategy in U.S.-Based Japanese subsidiaries: An empirical investigation. *Journal of International Business Studies*, *26*: 23–46. http://doi.org/10.1057/palgrave.jibs.8490164
20. Brammer, S. J., & Pavelin, S. 2006. Corporate reputation and social performance: The importance of fit. *Journal of Management Studies*, *43*: 435–455. http://doi.org/10.1111/j.1467-6486.2006.00597.x
21. Brammer, S., & Millington, A. 2005. Corporate reputation and philanthropy: An empirical analysis. *Journal of Business Ethics*, *61*: 29–44. http://doi.org/10.1007/s10551-005-7443-4
22. Brammer, S., & Pavelin, S. 2004. Voluntary social disclosures by large UK companies. *Business Ethics: a European Review*, *13*: 86–99. http://doi.org/10.1111/j.1467-8608.2004.00356.x
23. Brammer, S., Millington, A., & Pavelin, S. 2009. Corporate reputation and women on the board. *British Journal of Management*, *20*: 17–29. http://doi.org/10.1111/j.1467-8551.2008.00600.x
24. Brown, B., & Perry, S. 1994. Removing the financial performance halo from Fortune's 'most admired' companies. *Academy of Management Journal*, *37*: 1347–1359. http://doi.org/10.5465/256676
25. Brown, M. P., Sturman, M. C., & Simmering, M. J. 2003. Compensation policy and organizational performance: The efficiency, operational, and financial implications of pay levels and pay structure. *Academy of Management Journal*, *46*: 752–762. http://doi.org/10.5465/30040666
26. Cabello-Medina, C., López-Cabrales, Á., & Valle-Cabrera, R. 2011. Leveraging the innovative performance of human capital through HRM and social capital in Spanish firms. *International Journal of Human Resource Management*, *22*: 807–828. http://doi.org/10.1080/09585192.2011.555125
27. Carmeli, A., & Tishler, A. 2005. Perceived organizational reputation and organizational performance: An empirical investigation of industrial enterprises. *Corporate Reputation Review*, *8*: 13–30. http://doi.org/10.1057/palgrave.crr.1540236
28. Chandler, G. N., & Lyon, D. W. 2009. Involvement in knowledge–acquisition activities by venture team members and venture performance. *Entrepreneurship Theory and Practice*, *33*: 571–592. http://doi.org/10.1111/j.1540-6520.2009.00317.x
29. Chen, K. H., & Metcalf, R. W. 1980. The relationship between pollution control record and financial indicators revisited. *The Accounting Review*, *55*: 168–177.
30. Chen, Y. J., Wu, Y. J., & Wu, T. 2015. Moderating effect of environmental supply chain collaboration. *International Journal of Physical Distribution & Logistics Management*, *45*: 959–978. http://doi.org/10.1108/IJPDLM-08-2014-0183
31. Cheng, C. C. J., Yang, C.-L., & Sheu, C. 2014. The link between eco-innovation and business performance: A Taiwanese industry context. *Journal of Cleaner Production*, *64*: 81–90. http://doi.org/10.1016/j.jclepro.2013.09.050

32. Choi, J., & Wang, H. 2009. Stakeholder relations and the persistence of corporate financial performance. *Strategic Management Journal*, *30*: 895–907. http://doi.org/10.1002/smj.759

33. Choi, J.-S., Kwak, Y.-M., & Choe, C. 2010. Corporate social responsibility and corporate financial performance: Evidence from Korea. *Australian Journal of Management*, *35*: 291–311. http://doi.org/10.1177/0312896210384681

34. Chow, I. H. S., & Liu, S. S. 2009. The effect of aligning organizational culture and business strategy with HR systems on firm performance in Chinese enterprises. *The International Journal of Human Resource Management*, *20*: 2292–2310. http://doi.org/10.1080/09585190903239666

35. Chow, I. H., Huang, J.-C., & Liu, S. 2008. Strategic HRM in China: Configurations and competitive advantage. *Human Resource Management*, *47*: 687–706. http://doi.org/10.1002/hrm.20240

36. Chuang, C.-H., & Liao, H. 2010. Strategic human resource management in service context: Taking care of business by taking care of employees and customers. *Personnel Psychology*, *63*: 153–196. http://doi.org/10.1111/j.1744-6570.2009.01165.x

37. Clarkson, P. M., Li, Y., Richardson, G. D., & Vasvari, F. P. 2008. Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Accounting, Organizations and Society*, *33*: 303–327. http://doi.org/10.1016/j.aos.2007.05.003

38. Cole, M. A., Elliott, R. J. R., & Shimamoto, K. 2006. Globalization, firm-level characteristics and environmental management: A study of Japan. *Ecological Economics*, *59*: 312–323. http://doi.org/10.1016/j.ecolecon.2005.10.019

39. Collins, C. J., & Smith, K. G. 2006. Knowledge exchange and combination: The role of human resource practices in the performance of high-technology firms. *Academy of Management Journal*, *49*: 544–560. http://doi.org/10.5465/amj.2006.21794671

40. Combs, J. G., & David J Ketchen, J. 1999. Explaining interfirm cooperation and performance: Toward a reconciliation of predictions from the resource-based view and organizational economics. *Strategic Management Journal*, *20*: 867–888. http://doi.org/10.1002/(SICI)1097-0266(199909)20:9/%3C867::AID-SMJ55>3.0.CO;2-6

41. Coombs, J. E., & Gilley, K. M. 2005. Stakeholder management as a predictor of CEO compensation: main effects and interactions with financial performance. *Strategic Management Journal*, *26*: 827–840. http://doi.org/10.1002/smj.476

42. Cormier, D., & Gordon, I. M. 2001. An examination of social and environmental reporting strategies. *Accounting, Auditing & Accountability Journal*, *14*: 587–617. http://doi.org/10.1108/EUM0000000006264

43. De Carolis, D. M. 2003. Competencies and imitability in the pharmaceutical industry: An analysis of their relationship with firm performance. *Journal of Management*, *29*: 27–50. http://doi.org/10.1177/014920630302900103

44. De Castro, G. M., López, J. E. N., & Sáez, P. L. 2006. Business and social reputation: Exploring the concept and main dimensions of corporate reputation. *Journal of Business Ethics*, *63*: 361–370. http://doi.org/10.1007/sl0551-005-3244-z

45. Deephouse, D. L. 2000. Media reputation as a strategic resource: An integration of mass communication and resource-based theories. *Journal of Management*, *26*: 1091–1112. http://doi.org/10.1177/014920630002600602

46. Deephouse, D. L., & Carter, S. M. 2005. An examination of differences between organizational legitimacy and organizational reputation. *Journal of Management Studies*, *42*: 329–360. http://doi.org/10.1111/j.1467-6486.2005.00499.x

47. Delery, J. E., & Doty, D. H. 1996. Modes of theorizing in strategic human resource management: Tests of universalistic, contingency, and configurational performance predictions. *Academy of Management Journal*, *39*: 802–835. http://doi.org/10.2307/256713

48. Detert, J. R., Treviño, L. K., Burris, E. R., & Andiappan, M. 2007. Managerial modes of influence and counterproductivity in organizations: A longitudinal business-unit-level investigation. *Journal of Applied Psychology*, *92*: 993–1005. http://doi.org/10.1037/0021-9010.92.4.993

49. Douglas, T. J., & Judge, W. Q., Jr. 2001. Total quality management implementation and competitive advantage: the role of structural control and exploration. *Academy of Management Journal*, *44*: 158–169. http://doi.org/10.5465/3069343

50. Dowell, G., Hart, S., & Yeung, B. 2000. Do corporate global environmental standards create or destroy market value? *Management Science*, *46*: 1059–1074. http://doi.org/10.1287/mnsc.46.8.1059.12030

51. Eng Ann, G., Zailani, S., & Abd Wahid, N. 2006. A study on the impact of environmental management system (EMS) certification towards firms' performance in Malaysia. *Management of Environmental Quality*, *17*: 73–93. http://doi.org/10.1108/14777830610639459

52. Englmaier, F., Kolaska, T., & Leider, S. 2016. Reciprocity in organizations: Evidence from the UK. *Discussion paper.*

53. Ethiraj, S. K., Kale, P., Krishnan, M. S., & Singh, J. V. 2004. Where do capabilities come from and how do they matter? A study in the software services industry. *Strategic Management Journal*, *26*: 25–45. http://doi.org/10.1002/smj.433

54. Feng, T., Di Cai, Wang, D., & Zhang, X. 2016. Environmental management systems and financial performance: the joint effect of switching cost and competitive intensity. *Journal of Cleaner Production*, *113*: 781–791. http://doi.org/10.1016/j.jclepro.2015.11.038

55. Flanagan, D. J., & O'Shaughnessy, K. C. 2005. The Effect of layoffs on firm reputation. *Journal of Management*, *31*: 445–463. http://doi.org/10.1177/0149206304272186

56. Fombrun, C., & Shanley, M. 1990. What's in a name? Reputation building and corporate strategy. *Academy of Management Journal*, *33*: 233–258. http://doi.org/10.5465/256324

57. Gelade, G. A., & Ivery, M. 2003. The impact of human resource management and work climate on organizational performance. *Personnel Psychology*, *56*: 383–404. http://doi.org/10.1111/j.1744-6570.2003.tb00155.x

58. Gilley, K. M., Worrell, D. L., Davidson, W. N., III, & ElJelly, A. 2000. Corporate environmental initiatives and anticipated firm performance: the differential effects of process-driven versus product-driven greening initiatives. *Journal of Management*, *26*: 1199–1216. http://doi.org/10.1177/014920630002600607

59. Glebbeek, A. C., & Bax, E. H. 2004. Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal*, *47*: 277–286. http://doi.org/10.2307/20159578

60. Gould-Williams, J. 2003. The importance of HR practices and workplace trust in achieving superior performance: A study of public-sector organizations. *The International Journal of Human Resource Management*, *14*: 28–54. http://doi.org/10.1080/09585190210158501

61. Guest, D. E., Michie, J., Conway, N., & Sheehan, M. 2003. Human resource management and corporate performance in the UK. *British Journal of Industrial Relations*, *41*: 291–314. http://doi.org/10.1111/1467-8543.00273

62. Hassel, L., Nilsson, H., & Nyquist, S. 2005. The value relevance of environmental performance. *European Accounting Review*, *14*: 41–61. http://doi.org/10.1080/0963818042000279722

63. Huselid, M. A. 1995. The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal*, *38*: 635–672. http://doi.org/10.2307/256741

64. Janssen, O., & Van Yperen, N. W. 2004. Employees' goal orientations, the quality of leader-member exchange, and the outcomes of job performance and job satisfaction. *Academy of Management Journal*, *47*: 368–384. http://doi.org/10.5465/20159587

65. Judge, W. Q., & Douglas, T. J. 1998. Performance implications of incorporating natural environmental issues into the strategic planning process: An empirical assessment. *Journal of Management Studies*, *35*: 241–262. http://doi.org/10.1111/1467-6486.00092

66. Jung, H.-J., & Kim, D.-O. 2016. Good neighbors but bad employers: Two faces of corporate social responsibility programs. *Journal of Business Ethics*, *138*: 295–310. http://doi.org/10.1007/s10551-015-2587-3

67. Kacmar, K. M., Andrews, M. C., Van Rooy, D. L., Steilberg, R. C., & Cerrone, S. 2006. Sure everyone can be replaced… but at what cost? Turnover as a predictor of unit-level performance. *Academy of Management Journal*, *49*: 133–144. http://doi.org/10.5465/amj.2006.20785670

68. Katou, A. A., & Budhwar, P. S. 2006. Human resource management systems and organizational performance: a test of a mediating model in the Greek manufacturing context. *The International Journal of Human Resource Management*, *17*: 1223–1253. http://doi.org/10.1080/09585190600756525

69. Kaynak, H. 2003. The relationship between total quality management practices and their effects on firm performance. *Journal of Operations Management*, *21*: 405–435. http://doi.org/10.1016/S0272-6963(03)00004-4

70. Kim, J. H., Youn, S., & Roh, J. J. 2011. Green Supply Chain Management orientation and firm performance: evidence from South Korea. *International Journal of Services and Operations Management*, *8*: 283–23. http://doi.org/10.1504/IJSOM.2011.038973

71. King, A., & Lenox, M. 2002. Exploring the locus of profitable pollution reduction. *Management Science*, *48*: 289–299. http://doi.org/10.1287/mnsc.48.2.289.258

72. Lai, C. S., Chen, C. S., & Yang, C. F. 2012. The involvement of supply chain partners in new product development: The role of a third party. *International Journal of Electronic Business Management*, *10*: 261–273.

73. Lam, L. W., & White, L. P. 1998. Human resource orientation and corporate performance. *Human Resource Development Quarterly*, *9*: 351–364. http://doi.org/10.1002/hrdq.3920090406

74. Laosirihongthong, T., Adebanjo, D., & Tan, K. C. 2013. Green supply chain management practices and performance. *Industrial Management & Data Systems*, *113*: 1088–1109. http://doi.org/10.1108/IMDS-04-2013-0164

75. Lee, J., & Miller, D. 1996. Strategy, environment and performance in two technological contexts: contingency theory in Korea. *Organization Studies*, *17*: 729–750. http://doi.org/10.1177/017084069601700502

76. Lee, S. M., Tae Kim, S., & Choi, D. 2012. Green supply chain management and organizational performance. *Industrial Management & Data Systems*, *112*: 1148–1180. http://doi.org/10.1108/02635571211264609

77. Liden, R. C., Wayne, S. J., Liao, C., & Meuser, J. D. 2014. Servant leadership and serving culture: Influence on individual and unit performance. *Academy of Management Journal*, *57*: 1434–1452. http://doi.org/10.5465/amj.2013.0034

78. Lin, R.-J., Tan, K.-H., & Geng, Y. 2013. Market demand, green product innovation, and firm performance: evidence from Vietnam motorcycle industry. *Journal of Cleaner Production*, *40*: 101–107. http://doi.org/10.1016/j.jclepro.2012.01.001

79. Liouville, J., & Bayad, M. 1998. Human Resource Management and Performances. Proposition and Test of a Causal Model. *Human Systems Management*, *12*: 337–351. http://doi.org/10.1177/239700229801200304

80. Llach, J., Perramon, J., del Mar Alonso-Almeida, M., & Bagur-Femenías, L. 2013. Joint impact of quality and environmental practices on firm performance in small service businesses: an empirical study of restaurants. *Journal of Cleaner Production*, *44*: 96–104. http://doi.org/10.1016/j.jclepro.2012.10.046

81. Love, E. G., & Kraatz, M. 2009. Character, conformity, or the bottom line? How and why downsizing affected corporate reputation. *Academy of Management Journal*, *52*: 314–335. http://doi.org/10.5465/amj.2009.37308247

82. López-Gamero, M. D., Molina-Azorín, J. F., & Claver-Cortes, E. 2011. The relationship between managers' environmental perceptions, environmental management and firm performance in Spanish hotels: a whole framework. *International Journal of Tourism Research*, *13*: 141–163. http://doi.org/10.1002/jtr.805

83. Magness, V. 2006. Strategic posture, financial performance and environmental disclosure. *Accounting, Auditing & Accountability Journal*, *19*: 540–563. http://doi.org/10.1108/09513570610679128

84. Makni, R., Francoeur, C., & Bellavance, F. 2009. Causality between corporate social performance and financial performance: Evidence from Canadian firms. *Journal of Business Ethics*, *89*: 409–422. http://doi.org/10.1007/s10551-008-0007-7

85. Marquis, C., & Qian, C. 2014. Corporate social responsibility reporting in China: Symbol or substance? *Organization Science*, *25*: 127–148. http://doi.org/10.1287/orsc.2013.0837

86. Menguc, B., & Ozanne, L. K. 2005. Challenges of the "green imperative": a natural resource-based approach to the environmental orientation–business performance relationship. *Journal of Business Research*, *58*: 430–438. http://doi.org/10.1016/j.jbusres.2003.09.002

87. Menguc, B., Auh, S., & Ozanne, L. 2010. The interactive effect of internal and external factors on a proactive environmental strategy and its influence on a firm's performance. *Journal of Business Ethics*, *94*: 279–298. http://doi.org/10.1007/s10551-009-0264-0

88. Miller, D., & Lee, J. 2001. The people make the process: commitment to employees, decision making, and performance. *Journal of Management*, *27*: 163–189. http://doi.org/10.1177/014920630102700203

89. Miller, T., & Del Carmen Triana, M. 2009. Demographic diversity in the boardroom: Mediators of the board diversity–firm performance relationship. *Journal of Management Studies*, *46*: 755–786. http://doi.org/10.1111/j.1467-6486.2009.00839.x

90. Mishra, S., & Suar, D. 2010. Does corporate social responsibility influence firm performance of Indian companies? *Journal of Business Ethics*, *95*: 571–601. http://doi.org/10.1007/s10551-010-0441-1

91. Ngo, H.-Y., Turban, D., Lau, C.-M., & Lui, S.-Y. 1998. Human resource practices and firm performance of multinational corporations: influences of country origin. *The International Journal of Human Resource Management*, *9*: 632–652. http://doi.org/10.1080/095851998340937

92. Perry-Smith, J. E., & Blum, T. C. 2000. Work-family human resource bundles and perceived organizational performance. *Academy of Management Journal*, *43*: 1107–1117. http://doi.org/10.2307/1556339

93. Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. 2010. A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, *53*: 1131–1152. http://doi.org/10.5465/amj.2010.54533222

94. Ployhart, R. E., Weekley, J. A., & Ramsey, J. 2009. The consequences of human resource stocks and flows: A longitudinal examination of unit service orientation and unit effectiveness. *Academy of Management Journal*, *52*: 996–1015. http://doi.org/10.5465/amj.2009.44635041

95. Rettab, B., Brik, A. B., & Mellahi, K. 2008. A study of management perceptions of the impact of corporate social responsibility on organisational performance in emerging economies: The case of Dubai. *Journal of Business Ethics*, *89*: 371–390. http://doi.org/10.1007/s10551-008-0005-9

96. Russo, M. V., & Fouts, P. A. 1997. A resource-based perspective on corporate environmental performance and profitability. *Academy of Management Journal*, *40*: 534–559. http://doi.org/10.5465/257052

97. Schadewitz, H., & Niskala, M. 2010. Communication via responsibility reporting and its effect on firm value in Finland. *Corporate Social Responsibility and Environmental Management*, *17*: 96–106. http://doi.org/10.1002/csr.234

98. Shaw, J. D., Duffy, M. K., Johnson, J. L., & Lockhart, D. E. 2005a. Turnover, social capital losses, and performance. *Academy of Management Journal*, *48*: 594–606. http://doi.org/10.5465/amj.2005.17843940

99. Shaw, J. D., Gupta, N., & Delery, J. E. 2005b. Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. *Academy of Management Journal*, *48*: 50–68. http://doi.org/10.5465/amj.2005.15993112

100. Sheehan, M. 2014. Human resource management and performance: Evidence from small and medium-sized firms. *International Small Business Journal: Researching Entrepreneurship*, *32*: 545–570. http://doi.org/10.1177/0266242612465454

101. Shen, W., & Cannella, A. A., Jr. 2002. Revisiting the performance consequences of CEO succession: The impacts of successor type, postsuccession senior executive turnover, and departing CEO tenure. *Academy of Management Journal*, *45*: 717–733. http://doi.org/10.5465/3069306

102. Shortell, S. M., Zimmerman, J. E., Rousseau, D. M., Gillies, R. R., Wagner, D. P., Draper, E. A., et al. 1994. The performance of intensive care units: Does good management make a difference? *Medical Care*, *32*: 508–525.

103. Shrader, R., & Siegel, D. S. 2007. Assessing the relationship between human capital and firm performance: Evidence from technology-based new ventures. *Entrepreneurship Theory and Practice*, *31*: 893–908. http://doi.org/10.1111/j.1540-6520.2007.00206.x

104. Siebert, W. S., & Zubanov, N. 2009. Searching for the optimal level of employee turnover: A study of a large U.K. retail organization. *Academy of Management Journal*, *52*: 294–313. http://doi.org/10.2307/40390289?refreqid=search-gateway:9b4a973beabea66247ecfc0fa891127a

105. Skaggs, B. C., & Youndt, M. 2004. Strategic positioning, human capital, and performance in service organizations: a customer interaction approach. *Strategic Management Journal*, *25*: 85–99. http://doi.org/10.1002/smj.365

106. Subramony, M., & Holtom, B. C. 2011. Customer satisfaction as a mediator of the turnover- performance relationship. *Journal of Organizational Psychology*, *11*: 49–62.

107. Swink, M., Narasimhan, R., & Wang, C. 2007. Managing beyond the factory walls: Effects of four types of strategic integration on manufacturing plant performance. *Journal of Operations Management*, *25*: 148–164. http://doi.org/10.1016/j.jom.2006.02.006

108. Tagesson, T., Klugman, M., & Ekström, M. L. 2013. What explains the extent and content of social disclosures in Swedish municipalities' annual reports. *Journal of Management & Governance*, *17*: 217–235. http://doi.org/10.1007/s10997-011-9174-5

109. Takeuchi, R., Lepak, D. P., Wang, H., & Takeuchi, K. 2007. An empirical examination of the mechanisms mediating between high-performance work systems and the performance of Japanese organizations. *Journal of Applied Psychology*, *92*: 1069–1083. http://doi.org/10.1037/0021-9010.92.4.1069

110. Ton, Z., & Huckman, R. S. 2008. Managing the impact of employee turnover on performance: The role of process conformance. *Organization Science*, *19*: 56–68. http://doi.org/10.1287/orsc.1070.0294

111. Tzafrir, S. S. 2006. A universalistic perspective for explaining the relationship between HRM practices and firm performance at different points in time. *Journal of Managerial Psychology*, *21*: 109–130. http://doi.org/10.1108/02683940610650730

112. Van Jaarsveld, D. D., & Yanadori, Y. 2011. Compensation management in outsourced service organizations and its implications for quit rates, absenteeism and workforce performance: Evidence from Canadian call centres. *British Journal of Industrial Relations*, *49*: s1–s26. http://doi.org/10.1111/j.1467-8543.2010.00816.x

113. Vanhala, S., & Tuomi, K. 2006. HRM, company performance and employee well-being. *Management Revue*, *17*: 241–255. http://doi.org/10.2307/41783520

114. Wang, H., & Qian, C. 2011. Corporate philanthropy and corporate financial performance: The roles of stakeholder response and political access. *Academy of Management Journal*, *54*: 1159–1181. http://doi.org/10.5465/amj.2009.0548

115. Wang, M., Qiu, C., & Kong, D. 2011. Corporate social responsibility, investor behaviors, and stock market returns: Evidence from a natural experiment in China. *Journal of Business Ethics*, *101*: 127–141. http://doi.org/10.1007/s10551-010-0713-9

116. Way, S. A. 2002. High performance work systems and intermediate indicators of firm performance within the US small business sector. *Journal of Management*, *28*: 765–785. http://doi.org/10.1177/014920630202800604

117. Wiersema, M. F., & Bantel, K. A. 1993. Top management team turnover as an adaptation mechanism: The role of the environment. *Strategic Management Journal*, *14*: 485–504. http://doi.org/10.2307/2486714

118. Wright, P. M., Gardner, T. M., Moynihan, L. M., & Allen, M. R. 2005. The relationship between HR practices and firm performance: Examining causal order. *Personnel Psychology*, *58*: 409–446. http://doi.org/10.1111/j.1744-6570.2005.00487.x

119. Wright, P. M., Mccormick, B., Sherman, W. S., & Mcmahan, G. C. 1999. The role of human resource practices in petro-chemical refinery performance. *The International Journal of Human Resource Management*, *10*: 551–571. http://doi.org/10.1080/095851999340260

120. Xun, J. 2013. Corporate social responsibility in China: A preferential stakeholder model and effects. *Business Strategy and the Environment*, *22*: 471–483. http://doi.org/10.1002/bse.1757

121. Yu, S.-H. 2007. An empirical investigation on the economic consequences of customer satisfaction. *Total Quality Management*, *18*: 555–569. http://doi.org/10.1080/14783360701240493

122. Zahra, S. A., & Nielsen, A. P. 2002. Sources of capabilities, integration and technology commercialization. *Strategic Management Journal*, *23*: 377–398. http://doi.org/10.1002/smj.229

123. Zatzick, C. D., & Iverson, R. D. 2006. High-involvement management and workforce reduction: competitive advantage or disadvantage? *Academy of Management Journal*, *49*: 999–1015. http://doi.org/10.5465/amj.2006.22798180

124. Zeng, S. X., Meng, X. H., Zeng, R. C., Tam, C. M., Tam, V. W. Y., & Jin, T. 2011. How environmental management driving forces affect environmental and economic performance of SMEs: a study in the Northern China district. *Journal of Cleaner Production*, *19*: 1426–1437. http://doi.org/10.1016/j.jclepro.2011.05.002

125. Zhu, Y., Sun, L.-Y., & Leung, A. S. M. 2014. Corporate social responsibility, firm reputation, and firm performance: The role of ethical leadership. *Asia Pacific Journal of Management*, *31*: 925–947. http://doi.org/10.1007/s10490-013-9369-1