

Machine Learning in Evaluative Synthesis

Lessons from Private
Sector Evaluation in the
World Bank Group

Leonardo Bravo
Ariya Hagh
Roshin Joseph
Hiroaki Kambe
Yuan Xiang
Jos Vaessen



IEG
INDEPENDENT
EVALUATION GROUP

WORLD BANK GROUP
World Bank • IFC • MIGA

Machine Learning in Evaluative Synthesis

Lessons from Private Sector
Evaluation in the World Bank Group

Leonardo Bravo, Ariya Hagh, Roshin Joseph,
Hiroaki Kambe, Yuan Xiang, Jos Vaessen

Independent Evaluation Group

June 2023

CONTENTS

Authors	iv
Abstract	vi
Abbreviations	viii
Introduction	x
1. Machine Learning Applications in Evaluation	2
What Is Machine Learning?	4
Previous Applications	5
Potential in Evaluation	8
2. Classification and Synthesis of Evaluative Evidence in The Independent Evaluation Group's Finance and Private Sector Evaluation Unit	12
Objectives	14
Problem Statement	14
Methodology	14
Model Refinement	21
Summary of Results	22
Limitations	28
Conclusion	30
Bibliography	34

AUTHORS

Leonardo Bravo¹

Ariya Hagh²

Roshin Joseph³

Hiroaki Kambe⁴

Yuan Xiang⁵

Jos Vaessen⁶

Corresponding Author

Leonardo Bravo, lbravo@ifc.org

Author Affiliations

¹ World Bank Independent Evaluation Group

² World Bank Group

³ International Finance Corporation

⁴ Japan International Cooperation Agency

⁵ World Bank Independent Evaluation Group

⁶ World Bank Independent Evaluation Group

ABSTRACT

The analysis of the implementation challenges private sector projects face has traditionally involved manual identification and categorization of project documents by evaluation officers. An approach of this type offers nuance, but that nuance comes at a significant cost in terms of time and effort expended. The labor required to manually classify project performance parameters and assess the factors that explain why a particular project did (or did not) successfully achieve its intended development outcomes is both intensive and extensive and calls for a more efficient approach. Such an approach should take advantage of evaluators' established experience in diagnosing critical challenges and impediments to project performance as well as recent advances in machine learning. These advances allow practitioners to overcome the challenges manual classification presents by extracting and classifying vast quantities of text in ways that would otherwise be prohibitively laborious. As a demonstration of this concept, we discuss the use of automated content analysis to identify and classify factors and issues commonly faced in the implementation of private sector projects, sorting them according to a curated taxonomy. We describe our approach, which started with the development of a taxonomy of project factors and issues identified by subject area experts. This subsequently provided the basis for employing a combination of machine learning algorithms to iteratively fine-tune the taxonomy. The factors and issues were then classified into 5 overarching categories and 51 subcategories. We show that once machine learning models are sufficiently well trained, they are able to correctly identify the majority of factors and issues under consideration in the taxonomy, including not only their probability of occurrence in a particular paragraph, but also whether those factors and issues affected a particular project positively or negatively. The experiment suggests new avenues for machine-assisted classification of large corpora of documents for use in portfolio analysis and evaluative synthesis.

ABBREVIATIONS

IEG Independent Evaluation Group
IFC International Finance Corporation
LDA latent Dirichlet allocation

All dollar amounts are US dollars unless otherwise indicated.

INTRODUCTION

Faced with an ever-growing pool of evidence-rich text reports, evaluators are increasingly interested in extracting and synthesizing insights from these reports in a more efficient and reliable manner. A shift from manual identification and extraction of information to a more automated process is warranted in many cases, specifically in an institutional environment with a steady accumulation of reports that follow fairly standardized formats and types of content.

Three issues necessitate such a shift. First, manual categorization can be time consuming, which can limit evaluators to classifying either a smaller number of evaluation documents or a smaller number of factors and issues within the documents than they otherwise would. Second, differences among evaluators' backgrounds and individual classification decisions can introduce inconsistencies in how insights of the same type are classified. These inconsistencies can result in potential over- or underestimation of the prevalence of certain factors and issues, introducing unintended differences in classification that bias the resulting output. Third, manual classification does not readily lend itself to updating existing data sets with new documents and inputs that might become available after the initial classification has been completed. Machine learning for text classification provides an intuitive solution to these problems.

The automation of information extraction and classification opens up exciting avenues for streamlining evaluative synthesis, enabling evaluators to render in seconds what would otherwise require hours or even days of labor-intensive manual identification and coding. Machine learning methods can accelerate content extraction, provided that practitioners train the extraction tool properly. In the context of the text analytics explored in this paper, machine learning involves a combination of unsupervised and supervised text-mining techniques that transform raw text data into a matrix of terms, which is then classified according to a taxonomy of issues pertinent to the analysis at hand. Integration of existing theoretical priors and evaluator experiences can ensure an appropriate balance between the granularity and generalizability of the insights extracted from project documents. Such an approach offers evaluators a powerful analytical tool for better understanding the various determinants of project success, potential challenges to project implementation, and practical lessons for future projects, among other matters.

Automated methods provide three main advantages over conventional approaches. First, they permit faster and more systematic analysis of a set of documents than manual coding alone can achieve. Machine learning does not invalidate systematic manual review; rather, automatic classification and extraction of knowledge can provide a first step to inform further analysis. Second, automated methods can place a larger quantity of relevant data at the disposal of evaluation officers, who can then draw insights from a broader set of inputs than would have been available had manual approaches alone been used. Third, once properly trained, classification algorithms can form the underlying infrastructure for real-time or just-in-time analysis to inform decision-making, whereas using a purely manual process would not produce the required analysis for weeks or even months. Such algorithms can allow faster and custom manipulation of elements included in analyses based on user needs. In fact, providing real-time insights (for example, to the chair of an investment review meeting) could be the next use for this approach. The integration of machine learning into evaluative synthesis would represent a relatively low-cost intervention that would provide economies of scale for both current and future evaluations. As an investment, the approach would offer a tool that can be reused and modified for future analyses.¹ Machine learning can catalyze positive feedback loops, translating insights from identified project challenges into lessons that feed into project design and improve the quality of project implementation and project performance in the long term.

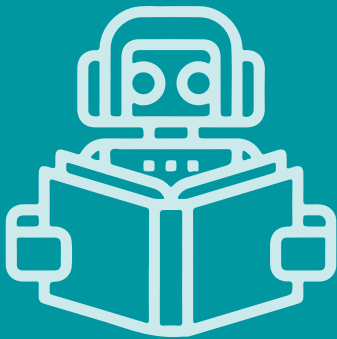
This paper builds on these points as follows. Chapter 1 provides an overview of machine learning and discusses relevant applications in the field of evaluation, briefly outlining previous work and potential future applications. In chapter 2, we use the case of the Finance and Private Sector Evaluation Unit of the Independent Evaluation Group as an example to illustrate the benefits of machine learning for text classification in evaluation. A summary of the results of this experiment and a brief discussion of potential next steps conclude the paper.

Endnotes

¹ The diagnosis of delivery challenges, their rank-ordering by salience, and viable strategies for iterative amelioration of future projects are some examples of ways in which multiuse machine learning applications can be employed.

1

MACHINE LEARNING APPLICATIONS IN EVALUATION



What Is Machine Learning?



Previous Applications



Potential in Evaluation

What Is Machine Learning?

Machine learning is based on pattern recognition and the theory that computers can autonomously learn to perform certain well-defined tasks (Samuel 1959). The procedure employed usually relies on algorithms, a set of unambiguous mathematical rules used to perform classification and data processing and draw basic inferences. At its core, machine learning is a Bayesian endeavor in which prior beliefs are updated based on new data introduced into analysis. Though the philosophy underlying this approach dates back to the eighteenth century, recent improvements in the efficiency and accessibility of computational methods have allowed scholars and practitioners to apply the tools of machine learning methods to a wide array of complex problems.

Training on a subset of the data, a machine learning algorithm extracts generalizable lessons from new data, becoming more precise as more information is inputted. Human classification often faces an upper limit on both efficiency and scalability. There are also limits on how perceptive human coders can be in regard to patterns hidden in very large data sets; given the complexity of the underlying phenomena under observation, more nuanced insights based on fewer observations might be lost in the sea of available data. The same shortcomings that limit the performance of manual methods can, however, serve as a source of strength in automated content analysis. Automated classification tends to become more accurate as the quantity of information increases and does not neglect more nuanced patterns, provided that the training data used are sufficiently well ordered.

Machine learning applications can involve both supervised and unsupervised methods, as well as a mixture of the two. Supervised-learning algorithms rely on human-coded training sets to train a classification tool to generate predictions from a broader sample of data. Such algorithms are given a set of latent parameters to search for a priori, classifying raw data into categories according to those parameters. Among other uses, they can be trained to categorize text, detect spam, diagnose health issues, and discover fraudulent spending activity. The accuracy of supervised methods relates to how well the parameters for information classification are vetted and the quality of the manual classification of information that is used as a training set for the algorithms employed. In short, supervised classification methods require essential inputs from human sources to function properly. However, they tend to make up for the initial time investment needed to provide these inputs once they have been properly calibrated, parsing and categorizing textual data that are relevant to a particular topic of interest faster and more accurately than manual approaches.

Unsupervised approaches, conversely, do not rely on human input. Instead, they independently search input data for potential correlates and clusters based on different underlying features. Both approaches offer unique advantages specific to different applications. Unsupervised methods can best be thought of as tools that support a Popperian “logic of discovery,” serving as an exploratory probe for detecting clusters and patterns in texts (Aggarwal and Zhai 2012).¹ However, though unsupervised classification tools can successfully detect patterns in complex and multidimensional data corpora, they can also be susceptible to misclassification errors and overfitting.

In rare cases, unsupervised approaches may unintentionally extrapolate substantively meaningless but statistically “significant” quirks in the data they are analyzing. Not every hidden association within data is useful in regard to a particular research topic. Human intervention is therefore needed to ensure that unsupervised training algorithms generate results that are substantively meaningful and not driven by stochastic noise in the underlying data. Such intervention becomes more pertinent as the complexity of the data increases. In practice, unsupervised learning can often be used with great success to detect hitherto unclassified clusters in data, highlight potential outliers in data sets, or reduce dimensionality within a complex framework.² But practitioners should not rely on unsupervised learning to produce consistent and meaningful outputs without some degree of vetting by those with substantive knowledge of the underlying phenomena of interest.

Previous Applications

Practical applications of machine learning and text analytics in the realm of evaluation have primarily focused on three areas: automatic coding of key implementation challenges, risk identification, and impact evaluation. Though different machine learning methods can offer a variety of efficiencies related to the practice of evaluation, arguably the most pertinent method has involved supervised or semi supervised classification of large quantities of text. Previous applications have taken advantage of tools for such supervised classification in several different contexts. A variety of studies that have applied machine learning to data in health care, pharmaceutical research, transportation, energy, and labor, among other areas, have noted the benefits of such an approach.

Cimiano et al. (2005) use machine learning to categorize a large corpus of heterogeneous data, extracting common text features and examining interrelationships among the various terms identified. Tanguy et al. (2016) use support vector machine learning to classify and evaluate safety event records

and archival documents, which enables them to categorize incident reports in the aviation sector. The resulting output improves the accuracy and reliability of analysis conducted by aviation experts, providing insights relevant to facets of aviation incidents. Schmidt, Schnitzer, and Rensing (2015) similarly take advantage of an automated classification algorithm for text-heavy source data, in this case a catalog of job offers based on hours of work, modes of employment, and functional work areas. The resulting output consists of a domain-specific search engine that enables subject-specific knowledge to be exploited more efficiently using a set of supervised subject filters.

Plmanabhan (2015) applies a battery of supervised multilabel classifiers and natural-language-processing techniques to analyze policy documents and survey data on psychological counseling for military servicemembers. He then uses the resulting output as a framework for explaining how the policies of the United States' Military Health System influence servicemembers' access to psychological services. Burscher, Vliegenthart, and De Vreese (2015) use a supervised-learning algorithm to categorize policy issues, political articles, and parliamentary discourse by salience and topic. The authors then use the results to investigate the generalizability of policy issue classifiers, testing the relevance of different machine-coded topics relative to those yielded by hand-coded training sets.

In regard to risk assessments, machine learning can help policy makers identify category-specific risk factors and quantify their impact, drawing on insights from challenges and obstacles encountered in earlier projects. In this context, Rona-Tas et al. (2019) use supervised learning in the field of food safety to assess the two main issues related to food hazards, helping practitioners better understand underlying ambiguities and emergent risks related to monitoring and inspection practices. Quantification of risk factors provides specific benefits in this context, as the output of the model employed (assessing the need for potential safety warnings and recalls) demands accurate and timely assessments of food risk parameters. Similarly, Abdellatif et al. (2015) and Ali (2007) use neural networks to assess flood risks and river water quality, generating output that helps manage urban water systems and minimize loss of life and property after water-based disasters. Galindo and Tamayo (2000) apply supervised-learning algorithms such as classification and regression tree models and neural networks to evaluate risk among financial intermediaries, generating an important diagnostic tool for assessing institutional risks and volatility.

Okori and Obua (2011) apply machine learning techniques to predict famines in Uganda, using data from the country's northern region to train their tool on inputs from other regions. They employ a combination of support vector machine, *k*-nearest neighbors, naïve Bayes, and decision tree analysis to highlight meaningful

relationships related to food security and famines, yielding output beneficial for evaluating causal variables related to theorized causes of food scarcity. Ofli et al. (2016) combine crowdsourcing and real-time supervised machine learning to evaluate large quantities of aerial and satellite imagery for time-sensitive disaster response. Jean et al. (2016) similarly apply machine learning to survey data and satellite imagery from Malawi, Nigeria, Rwanda, Tanzania, and Uganda, training a convolutional neural network to identify variations in local economic outcomes. The resulting output offers a scalable tool for predicting poverty according to a combination of data sources. Likewise, McBride and Nichols (2015) implement stochastic ensemble methods such as quantile regression forests to improve the accuracy of beneficiary targeting in poverty reduction, generating economies in areas in which conventional means testing can be prohibitively costly.

Impact evaluation has also benefited from advances in applied machine learning techniques. Counterfactual designs determine the effect of a policy intervention by comparing a treatment group with a control group over time, using experimental or quasi-experimental techniques to control for observable and non-observable causal factors. However, this type of comparison is not always feasible or desirable. In practice, achieving a proper balance among treatment and control groups is no easy feat, particularly when the active samples (such as specific social groups or geographical areas) tend to be structurally diverse. Matching techniques, including unsupervised learning, can be used in this area (see, for example, Gertler et al. 2016). In one example, Ruz, Varas, and Villena (2013) use *k*-means clustering algorithms to identify the common characteristics of households lacking internet access as a means of evaluating whether an unconditioned broadband and subsidiary campaign had a significant effect on broadband penetration in Chile.

Zheng, Zheng, and Ye (2016) also use machine learning methods to assess the development impact of environmental tax reform in China. Niu, Wang, and Duan (2009) rely on support vector machine analysis to evaluate the impact of power plant construction projects in China, and Burlig et al. (2017) examine, via machine learning, the impact of energy efficiency upgrades in primary and secondary schools. Machine learning can also yield useful meta-analytical insights. Mueller, Gaus, and Konradt (2016) note that progress in evaluation research depends on establishing a productive cycle of scholarly knowledge generation, dissemination, and implementation. Examining the uneven proliferation of scholastic work on evaluation, they employ a cross-national design for predicting evaluation research output, assessing the relative impact of country-specific research output in evaluation research.

In recent years, applications of machine learning and (more complex) deep learning models in the practice of evaluation have become more widespread. For example,

the Independent Evaluation Group (IEG), one of the early adopters of data science applications in evaluation, has applied these tools in the analysis of textual data in portfolio identification exercises and content analysis (for example, Franzen et al. 2022), as well as of imagery data in poverty mapping and geospatial impact evaluation (for example, Ziulu et al. 2022).

Potential in Evaluation

The use of machine learning approaches in evaluation is still in its early stages but shows significant potential, not only as part of advanced text analytics but also in the use of other data such as imagery data. Regarding advanced text analytics, machine learning techniques can be used to process and analyze text documents by automatically coding and categorizing key issues in the documents. For example, machine learning can be used to extract common challenges across various sectors studied and map the evolution of obstacles over time. Machine learning applications can provide at least two significant advantages over manual approaches in the context of evaluation. First, they can systematically explore large or growing data sources (such as, archives or document repositories), analyzing quantities of information that would be prohibitively time consuming for human coders. They can do this systematically, without a bias toward or against certain issues over others. The impact of various traits these applications discover in the data will therefore be directly related to the presence or absence of those traits in the data. This attribute of machine learning applications is quite valuable in evaluation, as assessments should reflect as closely as possible the underlying features of the evidence examined, without subjective biases or unintended variations of the type different human coders might introduce.

Second, automated machine learning applications can continue to improve their assessments as new evaluative data are introduced. As a result, their output represents a “living” classifier: new categories and implementation challenges will be added, updated, and removed as the body of data assessed changes over time. In the case of the work presented in this paper, for example (see chapter 2), the use of machine learning applications allows real-time learning and adaptation by the model in response to evaluator output and the integration of project lessons in practice. Over time, as new data are integrated into supervised analysis, a positive feedback loop can develop between evaluation and practice, allowing future projects to integrate generalizable and context-specific lessons into their design and implementation. This ability to learn and adapt can provide notable efficiency gains relative to manual coding.

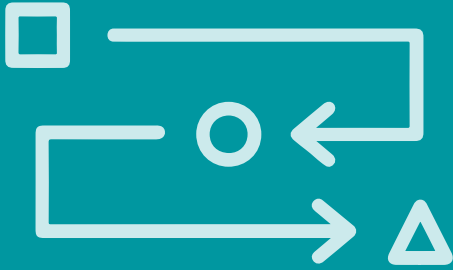
The application presented in this paper focuses on the extraction and classification of implementation challenges from private sector evaluation reports using machine learning techniques. In many ways it is similar to the Delivery Challenges in Operations for Development Effectiveness platform developed for public sector operations by the Global Delivery Initiative. The Delivery Challenges data set uses Implementation Completion and Results Reports from completed projects to generate a taxonomy of common issues that have an impact on project performance. Practitioners can then use insights from the data set to improve implementation and supervision outcomes.³ The experiment outlined in this paper offers a similar output for private sector operations, generating a set of implementation challenges representing specific obstacles encountered in the project cycle.

Endnotes

- ¹ For example, one particular type of unsupervised method (topic modeling) can be used to extract central themes and topics from documents, something that can be useful for parsing as well as classification (Blei 2012).
- ² For example, unsupervised methods can be used to identify a latent construct represented in clusters of text that contain common words related to a particular construct, such as women's empowerment, poverty, or democracy.
- ³ For more on the taxonomy, see Ortega Nieto, Hagh, and Agarwal (2022).

2

CLASSIFICATION AND SYNTHESIS OF EVALUATIVE EVIDENCE IN THE INDEPENDENT EVALUATION GROUP'S FINANCE AND PRIVATE SECTOR EVALUATION UNIT



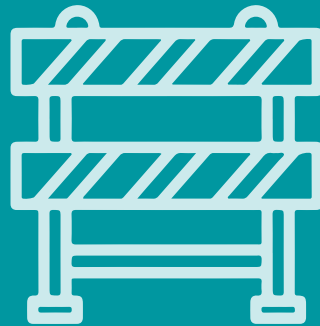
Methodology



Summary of Results



Model Refinement



Limitations

Objectives

IEG, an independent department within the World Bank Group, is charged with evaluating the activities of the World Bank (the International Bank for Reconstruction and Development and the International Development Association), the International Finance Corporation (IFC), and the Multilateral Investment Guarantee Agency. Specifically in regard to IFC's work, IEG conducts desk-based exercises to validate IFC's Investment Project Reports (Expanded Project Supervision Reports) and its Advisory Project Reports (Project Completion Reports). Three objectives drive the analysis outlined in this paper: (i) to support accountability by assessing the relevance, efficiency, and effectiveness of IFC's projects; (ii) to support organizational learning by identifying lessons from experience to improve IFC's operational performance; and (iii) to reinforce corporate objectives and values among IFC staff members.

Problem Statement

Automating the analysis of private sector project evaluations serves two major goals. First, we aim to build an automatic classifier so that the vast quantity of existing information in evaluation documents can be efficiently categorized according to distinct clusters of issues and challenges encountered in project implementation. Given the issues raised in chapter 1 related to the inefficiency of manual categorization, such an endeavor represents an intuitive next step in the parsing of evaluative evidence. Second, properly trained machine learning applications can help overcome issues related to intercoder reliability and evaluator subjectivity in classification. Based on the challenges of efficiency and accuracy discussed in the preceding chapter, automated classification and synthesis of project insights presents a viable solution for optimizing both the reliability and the objectivity of project analysis. The following sections summarize our methodological strategy, outline our implementation, and summarize our results.

Methodology

Using a combination of human expert knowledge and unsupervised- and supervised-learning algorithms (including naïve Bayes, random forest, support vector machine, and multilayer neural network methods), we generated a taxonomy of factors and issues that private sector projects typically encounter in regard to implementation. Approximately 1,600 documents evaluating private sector projects, produced between 2008 and 2022, provided our source input data for generating this taxonomy.

First, experts (IEG sector leaders with subject area expertise in the evaluation of projects in the financial, infrastructure, manufacturing agriculture and services, and funds sectors) discussed and shared the main factors and issues they faced in the development sectors in which they worked. We took the list of issues produced by the IEG sector leaders to conceptually account for the bulk of implementation issues private sector projects face throughout their life cycle. The sector leaders then manually classified these issues into five broad categories (country, market, sponsor, project specific, and other). Table 2.1 summarizes the taxonomy.

Table 2.1. Taxonomy of Project Insight (Categories)

Categories
Country and macro factors
Market, sector, and industry factors
Sponsor or client (management, sponsorship, and leadership)
Project-inherent challenges
Other


Source: Independent Evaluation Group.

Drawing on evaluation documents (specifically, IEG Evaluative Notes), we then extracted terms and concepts that were relevant to the issues identified, creating a matrix of keywords that was used to refine the experts’ draft taxonomy. In parallel, we applied automated text categorization to a list of more than 10,000 paragraphs to uncover potential subcategories from the corpus of supplied text. We used two unsupervised methods to complement the manual identification of conceptual categories. First, we used latent Dirichlet allocation (LDA) to find mixtures of terms for salient topics in the text. An evaluation officer compared the topics and key terms generated by LDA to the existing categories in our taxonomy, and it was found that four LDA-generated categories matched concepts identified by subject area experts. Keywords from those topics were added to the list of terms that would be used to identify those categories.¹



Second, we used Google’s Word2vec model, which presents each term as a unique vector.² The model can easily identify similar word combinations in common contexts by measuring their spatial proximity to generate clusters of concepts that are relevant to the analysis being undertaken. Figure 2.1 shows the Word2vec cluster for the concept of “expertise.” Using the interactive dashboard in the *TensorBoard* application, we then inputted keywords from our LDA and visualized the resulting word-proximity vectors in three-dimensional manifolds. Next, we compared the con-

This approach was used to classify paragraphs in the corpus of 1,600+ documents, with the system generating some 85,000 classified paragraphs overall. To allow categorization of paragraphs to more than one theme, the classification assigned a primary, secondary, and tertiary subcategory alongside a probability of assignment to each.⁴ As an additional measure to aid categorization, we also used a sentiment analysis to assign a score to each paragraph, ranging between -1 (totally negative; paragraph includes information on a factor or issue that is a barrier or impediment to project implementation) and +1 (totally positive; paragraph includes information on a factor or issue that contributes to success in project implementation). This analysis was carried out using polarity scores from Python’s Natural Language Processing Package.



Table 2.2. Taxonomy of Project Insight (Categories and Subcategories)

Categories	Subcategories	Definition
Country and macro factors 	Civil unrest and armed conflict	Factors related to civil unrest, armed conflict, and war
	Economic factors	Factors related to the macroeconomic environment, inflation, monetary policy, or austerity measures
	Epidemics and COVID-19	Factors related to epidemics (human, animal, and plant) and COVID-19
	Expropriation, nationalization, and transferability	Factors related to expropriation, nationalization, transfer, and convertibility
	Foreign exchange and local currency factors	Factors related to currency fluctuation, exchange rate and local currency issuance instruments
	Legal or regulatory factors	Factors related to regulatory policies, government, legislation, and bureaucratic mechanisms
	Natural disasters	Factors related to natural disasters such as hurricanes and earthquakes
	Political factors	Factors related to the political environment, including legislative and electoral dynamics

(continued)

Categories	Subcategories	Definition
Market, sector, and industry factors 	Business factors	Factors related to business model, cyclical business, or the operating environment
	Competition	Factors related to market competition: barriers to entry, monopolies, market dominance, and penetration
	Customers	Factors related to identifying correct target markets and clientele
	Market share	Factors related to market share
	Pricing	Factors related to price elasticity, supply, and marginal gains
Sponsor or client (management, sponsorship, and leadership) 	Capacity, capitalization, leverage	Factors related to sponsor capacity, capitalization, and leverage
	Commitment and motivation	Factors related to the strength and valence of strategic alignment, including compatibility, motivation, and ownership
	Conflicts of interest, corporate governance	Factors related to minority interest, conflicts of interest, and corporate governance
	Integrity, transparency, fairness, reputation	Factors related to integrity and transparency, such as disclosures of sensitive ethical issues, irregularities, and negative public perceptions
	Organizational structure	Factors related to organizational culture, institutional procedures, policies, and accountability
	Technical expertise, track record, and capacity	Factors related to the quality and expertise of the management team, their technical skills and track record, and contractor competency, familiarity, and acumen
	Succession	Factors related to succession, especially in family-owned businesses

(continued)

Categories	Subcategories	Definition
Project-inherent challenges 	Asset quality	Factors related to asset quality
	Cost overruns and delays	Factors related to overruns or delays
	Earnings and profitability	Factors related to earnings and profitability
	Environment and sustainability	Factors related to environmental standards, social health and safety parameters, or other safety standards
	Expansion	Factors related to acquisition, modernization, and expansion
	Funding	Factors related to funding
	Greenfield	Factors related to greenfield projects
	Gender	Factors related to gender
	Liquidity	Factors related to liquidity
	Technology	Factors related to changes in technology that affected project performance
	Training, know-how, and implementation	Factors related to training and know-how
Other 	Additionality principle and catalytic role ^a	Factors related to additionality and added value
	Coordination and collaboration with World Bank Group, other DFIs, donors, and other external stakeholders	Factors related to combined partnership and collaboration among the various stakeholders: the World Bank Group, donors, DFIs, and other external stakeholders
	Coordination and collaboration within IFC: AS-IS	Factors related to use of investment and advisory services to enhance IFC roles and contributions
	Project scoping and screening; country and stakeholder assessment; client needs assessment	Factors related to ex ante market analysis, due diligence, and consumer preferences

(continued)

Categories	Subcategories	Definition
	Client selection, commitment, and capacity	Factors related to client or implementing-partner selection (appropriateness) and client commitment and involvement
	Project design	Factors related to project design
	Financial model, project cost, and sensitivity assumptions	Factors related to financial modeling assumptions, including issues regarding overambitious objectives, deviations from forecasting estimates, and scaling
	Market assessment	Factors related to market assessment, market analysis, and consumer preferences
	Resources and timeline	Factors related to staffing, budget, and timeline
	Supervision and reporting	Factors related to (i) supervision and reporting; and (ii) taking measures to enhance these, as well as proactive client and stakeholder follow-up
	Sensitivity analysis	Factors related to sensitivity analysis, worst-case scenarios, stress tests, and risks to achieving development outcomes
	Documentation	Factors related to the quality of monitoring, documentation, and reporting
	Loan issues	Factors related to loan agreements, operating policies, breaches, and technical defaults
	Relationship management	Factors related to the quality and scope of relationship management, including fruitful and proactive engagements with on-site staff
	Debt issues	Factors related to debt issues, such as syndication, repayment, security, and refinancing
	Equity issues	Factors related to equity, valuation, and shareholder rights
	Financial risk mitigation	Factors related to risk-mitigation mechanisms such as guarantees, securities, prepayment penalties, and restructuring mechanisms
	Prepayments	Factors related to prepayments

(continued)

Categories	Subcategories	Definition
	Monitoring and evaluation	Factors related to compliance, monitoring including measurement, reporting, auditing, monitoring and evaluation plan and framework, appropriate indicators and targets, and clarity of data collection and evaluation approach
	Other issues	Factors related to other issues

Source: Independent Evaluation Group.

Note: a. The latest guidance on additionality can be found at <https://km.ifc.org/sites/pnp/MainDocumentMigration/DI716AdditionalityFramework.pdf>.

AS = advisory services; DFI = development finance institution; IFC = International Finance Corporation; IS = investment services.

Model Refinement

It should be noted that the initial classification exercise yielded low-accuracy results. This may be related to two possible causes. First, the unrefined taxonomy originally included 81 subcategories, before the manual validation described earlier. This meant that many subcategories were too sparsely populated to enable accurate identification of themes. Second, some of the keywords selected for use in classification occurred too commonly in evaluation documents to provide meaningful information for the models. By their nature, some of the themes included in the taxonomy overlapped conceptually. For example, the subcategories “client selection, commitment, and capacity” and “monitoring and evaluation” could be considered integral parts of the category “project-inherent challenges” as well as of the category “other,” where they appear in our taxonomy. This required manual review to separate the themes (where possible) and refine the keywords.

Given the large number of subcategories generated in the taxonomy, several steps were taken to iteratively refine it to improve classification precision and relevance. This yielded the smaller taxonomy of 51 categories shown in table 2.2. First, the subject area experts addressed deficiencies by either formulating new subcategories or deleting irrelevant or less-frequently occurring ones, expanding or consolidating categories when needed, and updating definitions. This helped us to avoid including

catchall categories that would make the resulting classifications of issues discussed in project documents less meaningful.⁵ Likewise, the removal of subcategories with very few observations helped make the taxonomy more manageable.⁶ At the same time, the training set was refined to eliminate catchall words and phrases to improve classification precision. For example, manual classification led to more than 15 percent of the initial paragraphs being assigned to the subcategory “IFC work quality.” We therefore assessed this subcategory as a catchall and divided it into several different subcategories, such as “market assessment,” “sensitivity analysis,” and “financial model, project cost, and sensitivity assumptions.”

Streamlining and refinement of model subcategories also involved additional diagnostics like cosine similarity. Cosine similarity analysis is a heuristic method of the distinctiveness of the vocabulary associated with a particular concept and can be used to identify categories that are problematically correlated with each other. Cosine similarity was used to find areas where underlying keywords or phrases used in conceptually distinct topics created issues in regard to classification accuracy: although the topics themselves might be conceptually distinct, the use of similar terms to identify relevant passages would result in overlaps among groups that reduce classification accuracy. In the case of high similarity scores, we checked keywords and categories to ensure that the groups identified in the taxonomy were (to the extent possible) mutually exclusively defined. After a few iterations, we were able to eliminate several categories with problematic overlaps, further improving the subcategories in the taxonomy.

The model refinement process offered three main benefits. First, it ensured that most categories were reasonably well balanced with respect to the number of paragraphs classified into them. Second, it improved the quality and informativeness of text tags and examples used in classification. Third, it generated sufficient observations per subcategory to allow for the exploratory and descriptive statistical analysis of lesson categories. After this recalibration, the subcategory with the maximum number of paragraphs represented about 6 percent of the total population of paragraphs, and the average subcategory included about 2 percent. Classification accuracy improved to an average of about 70 percent across the refined subcategories.

Summary of Results

The results of the automated classification and synthesis procedure were compared against hand-coded samples generated by subject experts. Table 2.3 provides an illustration of the results of this analysis.

Table 2.3. Comparison of Hand-Coded (Human) and Machine-Coded Classification

Factor	Subcategory	Text
Human Coding		
Factor 1	Legal or regulatory factors	Lack of a properly regulated public transportation system led to uncertainty and high risk regarding the setting of fares and payment of subsidies.
Factor 2	Political factors	Effective nationalization of [Company X] within the country operation. Cancellation of license (Country CDE Operation).
Machine Coding		
Factor 1	Legal or regulatory factors	Lack of a properly regulated public transport system (at the national or municipal level) leads to uncertainty and therefore high risk regarding the setting of fares and payment of subsidies. The project was expected to have a demonstration effect for other governments and municipalities and encourage similar public private partnerships.
Factor 2	Legal or regulatory factors	An attempt could be to have the legal agreement (between the government agency and the company) subject to an outside jurisdiction. It needs to ensure that there is a functioning regulatory authority that determines the amount and timing of fare increases and subsidy payments. This should be (and act) as legally independent of local and/or national governments.

(continued)

Factor	Subcategory	Text
Factor 3	Legal or regulatory factors	Subsidy turned out to be critical for the project. Take the form of international law governing the documents, or the presence of a strong independent regulatory authority in an environment where the judiciary is also strong and independent. If no effort to protect the project is undertaken, then it is subject to the changing whims of local regulators.
Factor 4	Legal or regulatory factors	[Company X] could not meet its performance targets owing to "operational and regulatory difficulties with the regulator" as the government refused to pay the subsidies agreed upon or increase the agreed-upon tariffs.
Factor 5	Political factors	Nationalization of [Company X] and cancellation of the license smacks of political interference and sets a lasting, negative effect which would deter future private investment in the public transport sector in both countries.
Factor 6	Political factors	The project was structured through the parent operation and provided some insulation against project-level risks. Nevertheless, from a development perspective this oversight exposed the project to high and unmitigated political risk.

(continued)

Factor	Subcategory	Text
Factor 7	Political factors	The political movement had a significant political and financial impact on the country, with (among other things) several national government changes. It is very difficult to structure a project so that it achieves its development objectives while going through a once-in-a-generation political and social revolution.
Factor 8	Political factors	The project relied on two important factors: (i) subsidies from FGH and (ii) implementation of agreed tariff increases. The subsidy only amounted to a small portion of receipts from traffic violations and thus this was not seen as an issue. Without control mechanisms, the project was entirely reliant on political will which is uncertain at best and was completely lacking after the political movement.
Factor 9	Expansion	[Company X] was to invest approximately US\$[X] million to modernize their facilities and expand their fleet. The loan was disbursed in two tranches.

(continued)

Factor	Subcategory	Text
Factor 10	Expansion	[Company X] planned to invest US\$[X] million, most of it in the form of a capital increase. Additional investment as well as capital provided by the existing shareholders to modernize its facilities and expand its fleet.
Factor 11	Expansion	[Company X], as the part of an expansion plan, signed an agreement to invest US\$[X] million through a capital increase. The capital increase would be used toward financing a capital expenditure program over the coming years with modern maintenance facilities, as well as a major fleet renewal and expansion.
Factor 12	Additionality principle and catalytic role	The project went ahead without adequately mitigating development risks (as distinct from the credit risks) as both deserve equal attention given the corporate mandate and purpose.
Factor 13	Additionality principle and catalytic role	It was expected that the project would have a strong developmental impact with increased transport access to the urban poor and the disabled, leading to improvements in service levels overall. In addition, the project was expected to encourage other governments and municipalities to create public-private frameworks.

Source: Independent Evaluation Group.

Note: Firm names and specific dollar amounts are withheld for reasons of confidentiality.

As expected, the model showed a high degree of accuracy in classifying content into well-defined subcategories such as “legal or regulatory factors,” “political risk,” and “market share,” whereas classifications into less well-defined categories such as “commitment and motivation” yielded a higher number of false positives. Overall, classification according to supervised machine learning techniques offered clear advantages over manual classification of factors and issues in project implementation. Manual classification relies on individual practitioners, each drawing on a set of unique theoretical priors, influenced by knowledge and experience that could potentially affect the way they search evaluation documents for factors and issues in implementation.

Furthermore, human coders focus on high-level or highly salient issues with greater frequency, potentially ignoring substantively meaningful but more subtle features that evaluation documents may also discuss. Drawing on a vetted training subset, supervised learning generated considerably higher classification efficiency than human coding with a comparable degree of accuracy. Properly calibrated machine analysis produced faster and more efficient synthesis of evaluative evidence.

After this initial test was undertaken, IEG undertook a wider analysis, with both human coders and algorithms classifying content in more than 170 Evaluative Notes published between 2020 and 2022 across four industries in which IFC funds projects (Financial Institutions Group; Manufacturing, Agribusiness, and Services; Infrastructure and Natural Resources; and Disruptive Technologies and Funds). Human coders were asked to include (i) the top three factors (taxonomy subcategories) that explained the success or failure of a project in terms of achieving its desired development outcome, organized from most important to least important; (ii) the direction in which each factor (subcategory) affected project success (+1 if the factor supported project success, –1 if the factor presented a risk affecting a project); and (iii) a copy of the paragraph from the Evaluative Note that supported why a factor (subcategory) was chosen.

Once the initial coders had classified the content in their project documents, a specialist or sector leader validated the classifications, as a form of peer review intended to make classification consistent across the four IFC industry groups. There was also an additional review across industries to make sure that classifications were consistent over the total portfolio of Evaluative Notes analyzed.

After human coders had classified the content in the Evaluative Notes and their classifications had been reviewed as discussed in the preceding paragraph, the same machine learning protocol was applied to the content. The average accuracy of machine-generated classification was about 70 percent across the subcategories evaluated, with classification in some subcategories such as “economic factors,” achieving greater than 90 percent accuracy.⁷

To ensure the relevancy and adaptability of our machine learning model against the evolving risk landscape, model performance is assessed periodically to reflect new evidence and adapt the subcategories in our taxonomy. Rigorous quality and change control procedures are in place to ensure the robustness, stability, and reliability of the model output.

Limitations

No methodology is without flaws, and machine learning is no exception to that rule. This section outlines some of the limitations to the approach explored in this paper. As discussed earlier, the inclusion of many overly granular subcategories resulted in low accuracy rates, especially in areas where there were very few observations to help classify a particular concept. We addressed issues of excessive granularity through a refinement of problematic subcategories. In addition, the use of diagnostics like cosine similarity ensured that the remaining categories were conceptually exclusive. However, this also implied that some of the nuances requested by subject experts and practitioners had to be omitted from the taxonomy. In those cases, the subcategories were often too subtle or complex to allow for accurate classification.

The output of a supervised model is only as good as the reliability of training data inputted. There are numerous pathways to suboptimal machine classification, but sufficient diligence and meticulous calibration of input parameters can guard against more pernicious errors and biases. If overarching categories in the taxonomy were not well defined or not mutually exclusive, the machine learning algorithm had difficulty in categorizing content into them accurately. Two examples illustrate this point. First, the model initially omitted the classification of factors and issues related to advisory services projects. When it became clear that the initial taxonomy was insufficiently equipped to classify such factors and issues, we modified the subcategories to address the omission. Once pertinent examples of such factors and issues had been provided to train the model, machine learning was then successfully used to identify other instances of similar issues. Second, the model initially used overly broad keywords, such as, “commitment” as a keyword in the subcategory “commitment and motivation.” This resulted in an overestimation of challenges related to that subcategory, as commitment can mean “the state or quality of being dedicated to a cause, activity, and so on,” but can also mean “obligation to provide a pledged amount of capital.” Its prevalence in evaluation reports therefore made it an inefficient classifier for machine learning applications. In both cases, we identified and corrected for this type of error through cross-validation of the output data and providing the machine learning algorithm with examples instead of keywords.

Endnotes

- ¹ Though relatively efficient, the latent Dirichlet allocation approach often generated groupings without a clearly interpretable significance. While these clusters could have represented potential categories, they were more likely a by-product of random associations without significant substantive meaning. We therefore omitted them from the analysis.
- ² Google developed Word2vec to reconstruct the linguistic context of sentence fragments. It maps inputted text data into a vector space.
- ³ We used the four algorithms to classify paragraphs that human experts had previously classified, and the algorithm with results closest to those of the manual classification was naïve Bayes.
- ⁴ For example, in cases in which a paragraph spoke exclusively about “economic factors,” then the probability for that subcategory would be 100 percent, and the probability for the next two categories would be 0 percent. In one example in which the majority of the paragraph was about economic factors, the probabilities assigned were 70 percent for “economic factors,” 20 percent “foreign exchange and local currency factors” and 10 percent “legal or regulatory factors.”
- ⁵ After refinement, average per-subcategory inclusion rates approached 2.0 percent, and the most broadly defined subcategory had an inclusion rate of 6.0 percent. To correct for the inclusion of frequent but substantively uninformative categories, we normalized the frequency with which categories were predicted by dividing the number of predictions for a particular category by the overall distribution of the predicted categories in the universe of coded keyword tags. We then chose the categories that had greater than 1.2 times the average along the distribution. This yielded a workable hierarchy of the most salient factors included in each document.
- ⁶ We eliminated any subcategories that included fewer than 50 paragraphs or merged them with conceptually proximate categories to increase identification accuracy. For example, we merged the subcategories “conflicts of interest” and “corporate governance,” as we found that they were both capturing similar concepts and each accounted for less than 1 percent of total paragraphs classified.
- ⁷ It should be noted that certain subcategories continued to perform suboptimally, even after modeling refinements were applied. In several cases, machine learning generated a substantial volume of false positives that required additional manual validation. Part of this relates to the trade-off between completeness and classification accuracy: although the lower performing subcategories may be of conceptual interest, there are certain limits to the quality of categorization output that are highly dependent on the keywords and phrases that can be used to correctly identify a concept. In some cases, these nuances are too subtle to be picked up by machine coding.

CONCLUSION

This paper has discussed the advantages and challenges of using machine learning in evaluative synthesis; more specifically, it has looked at the identification and classification of project-level implementation factors and issues. Our analysis showed that with the right combination of manual and automated approaches, machine-learning-based information classification can lead to significant efficiency gains without the loss of accuracy in information extraction and classification. Indeed, the incorporation of quality control practices can even result in gains in accuracy in certain cases. We discussed the concrete experience of IEG’s Financial and Private Sector Micro Unit as a basis for a systematic discussion of this process. We first discussed the principles for generating a taxonomy for classification. We then applied a combination of unsupervised- and supervised-learning techniques to generate word clusters, keywords, and examples from evaluation documents as features for classification. These were integrated into a taxonomy and used to classify the features into multiple categories of factors and issues.

Following several rounds of cross-validation and calibration, we were able to achieve accuracy rates for classification comparable to those achieved by human coders in this field (about 70 percent accuracy) but at substantially higher levels of efficiency, because the model we designed can perform the classification task at a much faster rate than human coders. As expected, our model classified features into well-defined subcategories such as “legal or regulatory factors,” “political factors,” and “market pricing” with much higher accuracy (that is, fewer incorrect classifications) than it did into broader subcategories such as “commitment and motivation.” In instances in which we specified subcategories imprecisely, the model faced greater difficulties in converging on the correct subcategories into which to classify the features. Furthermore, the use of overly broad keywords also initially resulted in misclassification errors. Subsequent refinements to the model and inputs from subject experts helped improve the training data, enabling the model to efficiently generate more relevant tags for features it classified.

Currently, the output of our extraction and classification process is captured in a data visualization tool (based on Microsoft’s Tableau platform), which generates descriptive statistics on implementation factors and issues disaggregated by geographic area and private sector industry. In addition, the output is used for writing synthetic evaluative analyses. The inclusion of readily accessible and searchable parameters for factors and issues allows project practitioners in the

Bank Group to observe commonalities and patterns across large numbers of successful and unsuccessful projects and disaggregate the output according to sectoral or regional factors where useful. Such a combination of features thus allows the model to be used to leverage decades of institutional experience in project implementation and apply it to both evaluative synthetic analysis and project design more efficiently and systematically than has been possible before.

As with any other form of analysis, the accuracy of our model's results is contingent on the quantity and quality of inputted data, as well as the presence of adequate supervision and cross-validation. Given these conditions, automated parsing and tagging of project information shows promise as an intuitive improvement over a manual approach. The output from our taxonomy allows evaluators to access the entire universe of project insights from all available project evaluations and learn about salient factors influencing project performance. With future revisions and refinements to the taxonomy (particularly with the inclusion of more examples in the training set), the classification accuracy rates achieved by the model will continue to improve. Taken together, the gains in efficiency and benefits in regard to data accessibility that result from the use of machine learning techniques will allow evaluators and practitioners to better incorporate lessons from the past into future practice.

BIBLIOGRAPHY

- Abdellatif, M., W. Atherton, R. Alkhaddar, and Y. Osman. 2015. "Flood Risk Assessment for Urban Water System in a Changing Climate Using Artificial Neural Network." *Natural Hazards* 79 (2): 1059–77. <https://doi.org/10.1007/s11069-015-1892-6>.
- Aggarwal, C. C., and C. Zhai, eds. 2012. *Mining Text Data*. New York: Springer Science and Business Media. <https://doi.org/10.1007/978-1-4614-3223-4>.
- Ali, M. Z. 2007. "The Application of the Artificial Neural Network Model for River Water Quality Classification with Emphasis on the Impact of Land Use Activities: A Case Study from Several Catchments in Malaysia." PhD thesis, University of Nottingham, Nottingham, UK. <https://eprints.nottingham.ac.uk/11867/>.
- Bail, C. A. 2015. "Commentary: Lost in a Random Forest; Using Big Data to Study Rare Events." *Big Data & Society* 2 (2): 2053951715604333. <https://doi.org/10.1177/2053951715604333>.
- Blei, D. M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. <https://doi.org/10.1145/2133806.2133826>.
- Burlig, F., C. Knittel, D. Rapson, M. Reguant, and C. Wolfram. 2017. "Machine Learning from Schools about Energy Efficiency." NBER Working Paper 23908, National Bureau of Economic Research, Cambridge, MA. <https://www.nber.org/papers/w23908>.
- Burscher, B., R. Vliegthart, and C. H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *Annals of the American Academy of Political and Social Science* 659 (1): 122–31. <https://doi.org/10.1177/0002716215569441>.
- Camillo, F., and I. D'Attoma. 2010. "A New Data Mining Approach to Estimate Causal Effects of Policy Interventions." *Expert Systems with Applications* 37 (1): 171–81. <https://doi.org/10.1016/j.eswa.2009.05.072>.
- Cimiano, P., A. Pivk, L. Schmidt-Thieme, and S. Staab. 2005. "Learning Taxonomic Relations from Heterogeneous Sources of Evidence." In *Ontology Learning from Text: Methods, Evaluation and Applications* (Frontiers in Artificial Intelligence and Applications, vol. 123), edited by P. Buitelaar, P. Cimiano, and B. Magnini, 59–73. Amsterdam: IOS Press.
- Dayan, P., M. Sahani, and G. Deback. 1999. "Unsupervised Learning." In *The MIT Encyclopedia of the Cognitive Sciences*, edited by R. A. Wilson and F. C. Keil. Cambridge, MA: MIT Press.

- Franzen, S., C. Quang, L. Schweizer, A. Budzier, J. Gold, M. Vellez, S. Ramirez, and E. Raimondo. 2022. *Advanced Content Analysis: Can Artificial Intelligence Accelerate Theory-Driven Complex Program Evaluation?* IEG Methods and Evaluation Capacity Development Working Paper Series. Independent Evaluation Group. Washington, DC: World Bank. <https://ieg.worldbankgroup.org/methods-resource/advanced-content-analysis-can-artificial-intelligence-accelerate-theory-driven-complex>.
- Galindo, J., and P. Tamayo. 2000. "Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications." *Computational Economics* 15 (1): 107–43. <https://doi.org/10.1023/A:1008699112516>.
- Gertler, P. J., S. Martinez, P. Premand, L. B. Rawlings, and C. M. J. Vermeersch. 2016. *Impact Evaluation in Practice*. 2nd ed. Washington, DC: Inter-American Development Bank and World Bank. <http://hdl.handle.net/10986/25030>.
- Ghahramani, Z. 2004. "Unsupervised Learning." In *Advanced Lectures on Machine Learning*, edited by O. Bousquet, U. Luxburg, and G. Rätsch, 72–112. Berlin: Springer. <https://link.springer.com/book/10.1007/b100712>.
- Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. <https://www.jstor.org/stable/24572662>.
- Hillard, D., S. Purpura, and J. Wilkerson. 2008. "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research." *Journal of Information Technology and Politics* 4 (4): 31–46. <https://doi.org/10.1080/19331680801975367>.
- Ittoo, A., L. M. Nguyen, and A. van den Bosch. 2016. "Review: Text Analytics in Industry; Challenges, Desiderata and Trends." *Computers in Industry* 78: 96–107. <https://doi.org/10.1016/j.compind.2015.12.001>.
- Jean, N., M. Burke, M. Xie, M. Davis, D. B. Lobell, and S. Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94. <https://doi.org/10.1126/science.aaf7894>.
- McBride, L., and Nichols, A. (2018). "Retooling Poverty Targeting Using Out-of-Sample Validation And Machine Learning." *The World Bank Economic Review*, 32(3), 531–50.
- Mueller, C. E., H. Gaus, and I. Konradt. 2016. "Predicting Research Productivity in International Evaluation Journals across Countries." 12 (27): 79–92. <https://doi.org/10.56645/jmde.v12i27.459>.
- Ofli, F., P. Meier, M. Inran, C. Castillo, D. Tuia, N. Rey, J. Briant, et al. 2016. "Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response." *Big Data* 4 (1): 47–59. <https://doi.org/10.1089/big.2014.0064>.

- Okori, W., and J. Obua. 2011. "Machine Learning Classification Technique for Famine Prediction." In *Proceedings of the World Congress on Engineering 2011* (London, July 6–8, 2011), vol. 2, edited by S. I. Ao, L. Gelman, D. W. L. Hukins, A. Hunter, and A. M. Korsunsky, 991–96. Hong Kong SAR, China: International Association of Engineers. <https://www.iaeng.org/publication/WCE2011/>.
- Ortega Nieto, D., A. Hagh, and V. Agarwal. 2022. "Delivery Challenges and Development Effectiveness: Assessing the Determinants of World Bank Project Success." Policy Research Working Paper 10144, World Bank, Washington, DC. <http://hdl.handle.net/10986/37902>.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 (February): 2825–30. <https://doi.org/10.5555/1953048.2078195>.
- Plmanabhan, J. P. 2015. "Applying Machine Learning Techniques to the Analysis of Policy Data of the Military Health Enterprise." Masters' thesis, Massachusetts Institute of Technology, Cambridge, MA. <https://dspace.mit.edu/handle/1721.1/106270>.
- Popper, K. 2002. *The Logic of Scientific Discovery*. London: Routledge. <https://doi.org/10.4324/9780203994627>.
- Rennie, J. D. M., L. Shih, J. Teevan, and D. R. Karger. 2003. "Tackling the Poor Assumptions of Naive Bayes Text Classifiers." In *ICML '03: Proceedings of the Twentieth International Conference on Machine Learning* (Washington, DC, August 2003), edited by T. Fawcett and N. Mishra, 616–23. Washington, DC: AAAI Press. <https://doi.org/10.5555/3041838.3041>.
- Rona-Tas, A., A. Cornuéjols, S. Blanchemanche, A. Duroy, and C. Martin. 2019. "Enlisting Supervised Machine Learning in Mapping Scientific Uncertainty Expressed in Food Risk Analysis." *Sociological Methods & Research* 48 (3): 608–41. <https://doi.org/10.1177/004912411772970>.
- Ruz, G. A., S. Varas, and M. Villena. 2013. "Policy Making for Broadband Adoption and Usage in Chile through Machine Learning." *Expert Systems with Applications* 40 (17): 6728–34. <https://doi.org/10.1016/j.eswa.2013.06.039>.
- Samuel, A. L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development* 3 (3): 210–29. <https://doi.org/10.1016/10.1147/rd.33.0210>.
- Schmidt, S., S. Schnitzer, and C. Rensing. 2016. "Text Classification Based Filters for a Domain-Specific Search Engine." *Computers in Industry* 78: 70–79. <https://doi.org/10.1016/j.compind.2015.10.004>.
- Sebastiani, F. 2002. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys* 34 (1): 1–47. <https://doi.org/10.1145/505282.505283>.

- Tanguy, L., N. Tulechki, A. Urieli, E. Hermann, and C. Raynal. 2016. "Natural Language Processing for Aviation Safety Reports: From Classification to Interactive Analysis." *Computers in Industry* 78: 80–95. <https://doi.org/10.1016/j.comp-ind.2015.09.005>.
- Tong, S., and D. Koller. 2002. "Support Vector Machine Active Learning with Applications to Text Classification." *Journal of Machine Learning Research* 2 (March): 45–66. <https://doi.org/10.1162/153244302760185243>.
- Zheng, Y., H. Zheng, and X. Ye. 2016. "Using Machine Learning in Environmental Tax Reform Assessment for Sustainable Development: A Case Study of Hubei Province, China." *Sustainability* 8 (11): 1124. <https://doi.org/10.3390/su8111124>.
- Ziulu, V., J. Meckler, G. Hernández Licona, and J. Vaessen. 2022. *Poverty Mapping: Innovative Approaches to Creating Poverty Maps with New Data Sources*. IEG Methods and Evaluation Capacity Development Working Paper Series. Independent Evaluation Group. Washington, DC: World Bank. <https://ieg.worldbankgroup.org/evaluations/poverty-mapping-innovative-approaches-creating-poverty-maps-new-data-sources>.



IEG
INDEPENDENT
EVALUATION GROUP

WORLD BANK GROUP
World Bank • IFC • MIGA

The World Bank
1818 H Street NW
Washington, DC 20433