Advanced Content Analysis

Can Artificial Intelligence Accelerate Theory-Driven Complex Program Evaluation?

Samuel Franzen
Cuong Quang
Lukas Schweizer
Alexander Budzier
Jenny Gold
Mercedes Vellez
Santiago Ramirez
Estelle Raimondo

IEG
Independent Evaluation Group
WORLD BANK GROUP
World Bank • IFC • MIGA

IEG Methods and Evaluation Capacity Development Working Paper Series

1818 H Street NW
Washington, DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

ATTRIBUTION
Please cite the report as: Franzen, Samuel, Cuong Quang, Lukas Schweizer, Alexander Budzier, Jenny Gold, Mercedes Vellez, Santiago Ramirez, and Estelle 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.

MANAGING EDITORS
Jos Vaessen
Ariya Hagh

EDITING AND PRODUCTION
Amanda O'Brien

GRAPHIC DESIGN
Luísa Ulhoa

Advanced Content Analysis

Can Artificial Intelligence Accelerate Theory-Driven Complex Program Evaluation?

Samuel Franzen, Cuong Quang, Lukas Schweizer, Alexander Budzier, Jenny Gold, Mercedes Vellez, Santiago Ramirez, Estelle Raimondo

Independent Evaluation Group

January 2022

CONTENTS

AUTHORS

Samuel Franzen[1]

Cuong Quang[2]

Lukas Schweizer[3]

Alexander Budzier[1]

Jenny Gold[4]

Mercedes Vellez[4]

Santiago Ramirez[4]

Estelle Raimondo[4]

Corresponding author

Samuel Franzen: sam.franzen@oxfordglobalprojects.com

Author Affiliations

[1] Oxford Global Projects

[2] Octant AI

[3] Deepreason.ai

[4] World Bank Independent Evaluation Group

ABBREVIATIONS

AI	artificial intelligence
IEG	Independent Evaluation Group
NLP	natural language processing
SML	supervised machine learning
ToC	theory of change
t-SNE	t-distributed stochastic neighbor embedding
UML	unsupervised machine learning

All dollar amounts are US dollars unless otherwise indicated.

ACKNOWLEDGMENTS

The paper was prepared by the Independent Evaluation Group for international development specialists, with both evaluation professionals and nutrition sector practitioners in mind. This paper is offered as part of the methodological paper series sponsored by the Independent Evaluation Group's Methods Advisory Function. The authors would like to thank the editors and staff of this series for their comments and contributions: Jos Vaessen, Ariya Hagh, Sylvia Otieno, and Maurya West Meiers. A special thanks goes to Amanda O'Brien and Luísa Ulhoa for their support in editing, production, and graphic design.

All authors reviewed and contributed to the paper. The findings, interpretations, and conclusions expressed in this paper are those of the authors and should not be attributed in any manner to the World Bank Group, members of its Board of Executive Directors, or the countries they represent.

OVERVIEW

This paper presents the methodology and results used to pilot and test the applicability, usefulness, and added value of using artificial intelligence for advanced theory-based content analysis. Traditionally, qualitative synthesis would be used to perform a theory-driven structured analysis of project reports. This pilot sought to assess the efficiency gains generated by artificial intelligence–assisted content analysis in labeling and classifying text according to an outcome-based conceptual framework. The approach used a set of interventions associated with the World Bank's stunted growth and chronic malnutrition evaluation portfolio, consisting of 392 unique project reports from 64 countries.

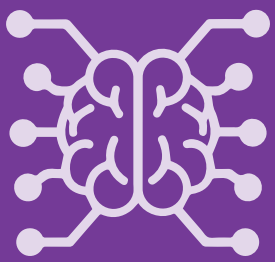
First, supervised machine learning was used to deductively label content under three main categories: nutrition challenges addressed, interventions, and outcome indicator achievement. Although performance at predicting exact sublabels (n = 74) was modest, the high level of accuracy achieved in predicting top-level categories suggested that the possibility of developing a text classifier model with acceptable coding accuracy is promising.

Second, unsupervised machine learning was used to identify emergent insights from text labeled "factors affecting intervention success." Overall, the topic model showed excellent performance in identifying inductive topics that not only were novel and domain relevant but proved to be key predictors of project performance and good practices. Semantic similarities between machine learning–labeled text were then visualized using t-distributed stochastic neighbor embedding. This proved effective at identifying important patterns in the data that would not be obvious to a human analyst, facilitating the establishment of unique country program characteristics.

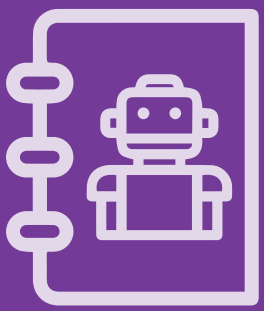
Finally, knowledge graph approaches were used to structure machine learning outputs according to the conceptual framework and explore relationships among components of the theory of change. Rule-based reasoning successfully performed simple statistical analyses on the success rates of interventions, but further research is required before knowledge graphs can enable a theory-based evaluation of program performance.

1

THE CHALLENGE



Challenges in development evaluation and the promise of artificial intelligence



Could artificial intelligence be used to accelerate theory-driven complex portfolio evaluation?

Challenges in Development Evaluation and the Promise of Artificial Intelligence

The increasing use of systems approaches and adaptive management within development programming has led to growing demand for complexity-responsive evaluation methods (Bamberger, Raimondo, and Vaessen 2016). These methods must be able to meet the challenge of providing deeper insights into development project performance while working within the constraints of decreasing budgets and ever-increasing expectations for real-time evidence. In this context, the increasing volume and accessibility of evaluation evidence presents both a solution and a problem to resolving this challenge.

The innumerable data and countless evaluation reports currently available present an incredible opportunity for learning through evidence synthesis. But although current evaluation methods can be highly effective at generating both context-specific and transferable learning, the manual nature of traditional analyses preclude the comprehensive and systematic examination of big evaluation data. Evaluation practitioners simply cannot rigorously synthesize such large evidence bases. This results in missed opportunities in terms of the breadth and scale of learning that big evaluation data offers.

Artificial intelligence (AI) methodologies have revolutionized our ability to make sense of big data. If applied to the field of development evaluation, these innovations may offer much faster, more comprehensive, and systematic synthesis of evaluation data and reports. This could not only generate powerful learning opportunities for development practice but also be timely and cost-efficient. Automated analyses could be regularly updated and scaled as new data and evidence are generated, ensuring that the knowledge remains current and justifying the initial investment.

However, AI for evaluation purposes remains an emergent technology. Most work at this frontier has focused on applying AI to data collection, cleaning, and modeling for primary research (McKenzie 2018; Korenblum 2017; Bravo et al. 2021). There has been some pioneering work exploring the use of AI for evaluation synthesis, and the World Bank is beginning to use AI to identify portfolios of work and classify content such as risks (Bravo et al. 2021).[1] Nevertheless, the use of AI for content synthesis and theory-driven analyses is limited. More experimentation and learning are required to establish the potential benefits of AI for evaluation purposes.

As the organization responsible for promoting innovation in evaluation analyses at the World Bank, the Independent Evaluation Group's (IEG) Methods Advisory Function commissioned a study to test the feasibility and usefulness of using AI to support and accelerate its thematic and country-focused evaluations. These evaluations assess the performance of World Bank projects and activities that were provided over 5- to 10-year periods.

This white paper presents the results from the feasibility study and offers conclusions on the use of AI for evaluation syntheses and deepening evidence and learning on development project results. This white paper provides sufficient detail for a general overview of the technical methods used but is not intended to meet the higher experimental reproducibility standards typical of academic and related work. Further details are available in a separate technical report.

Could Artificial Intelligence Be Used to Accelerate Theory-Driven Complex Portfolio Evaluation?

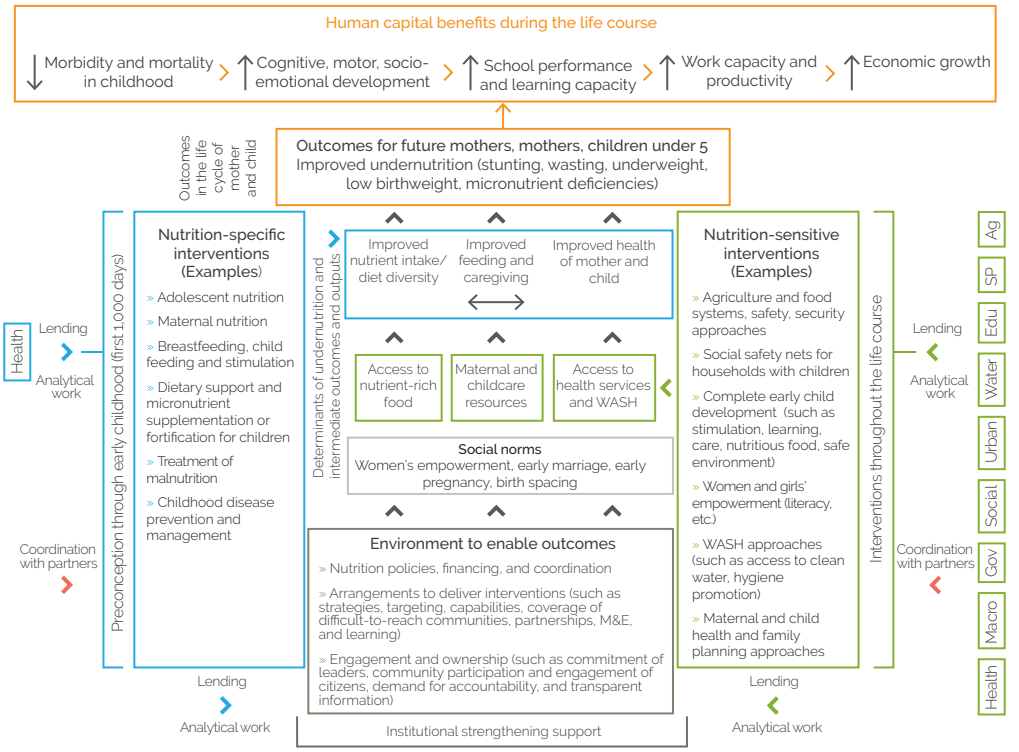
The overall objectives of the feasibility study were to pilot and test the applicability, usefulness, and added value of using AI, including machine learning and knowledge graph methodologies, for advanced theory-based content analysis in the framework of IEG's thematic evaluations. The pilot focused on a set of interventions from 64 countries that were within the scope of the IEG thematic evaluation *The World Bank's Support to Reducing Child Undernutrition* (World Bank 2021). One pilot country from within this portfolio was selected as a test case for deeper analyses to determine if the AI methodologies could also support country-focused evaluations.

IEG evaluations typically use a predefined conceptual framework to enable a structured portfolio review. The conceptual framework maps out the latest approaches to improving outcomes in the thematic area in the form of a theory of change (ToC). The ToC presents a hypothesized results chain among intervention inputs, outputs, outcomes, and impacts, and factors considered important for intervention success.

The World Bank portfolio of activities and results are then compared with this ToC to determine whether World Bank support is consistent with best practice approaches, and whether these approaches have proven effective in the way intended. This kind of theory-driven approach is often used to determine program contribution when experimental methods are not possible.

For the AI methodologies to support and accelerate IEG evaluations, the pilot needed to determine if advanced content analysis using machine learning could identify intervention inputs, outputs, outcomes, impacts, and contributory factors, based on a conceptual framework developed by the IEG team. This conceptual framework is presented in figure 1.1. The pilot also sought to explore whether the AI methodologies, specifically knowledge graphs, could identify relationships among elements of the ToC as evidence for program contribution.

Figure 1.1. Conceptual Framework of Child Undernutrition



Source: Independent Evaluation Group, adapted from Maternal and Child Nutrition Study Group 2013 and UNICEF 1990.

Note: Ag = Agriculture; Edu = Education; Gov = Governance; Social = Social Development; Macro = Macroeconomic; M&E = monitoring and evaluation; SP = Social Protection WASH = water, sanitation, and hygiene.

Endnotes

[1] The Independent Evaluation Group's evaluations are available at https://ieg.worldbankgroup.org/evaluations.

2

HOW WAS ARTIFICIAL INTELLIGENCE USED TO AUTOMATE CONTENT ANALYSIS AND QUALITATIVE SYNTHESIS?

Traditional methods used for evaluation synthesis

Artificial intelligence approaches to automating content analysis and qualitative evaluation synthesis

Traditional Methods Used for Evaluation Synthesis

IEG thematic evaluations draw on project design and end-of-project evaluation documents. These are long narrative documents written by project teams and evaluation experts that draw on quantitative and qualitative data from multiple sources of experimental and nonexperimental methods. The reports are therefore qualitative and evaluative in nature and lack a standardized structure or labeling of specific results, except where logframes of key performance indicators are used.

Traditionally, systematic qualitative synthesis would be used to analyze this kind of evidence. This would normally involve reading all relevant documents and using thematic content analysis to code sections of the text against a conceptual framework. Coding may be done according to a predefined conceptual framework that seeks to confirm or reject hypotheses or evaluation questions (deductive). Alternatively, data may be coded from the ground up for exploration and understanding of what themes emerge from the content (inductive).

In practice, deductive and inductive coding can be used simultaneously or in sequence. However, inductive approaches tend to be less common because they require greater domain-specific expertise and deeper analysis, which is often only feasible on relatively few documents given the time required. Once all content is coded, the evidence within each code can be organized according to the conceptual framework. The framework is then reviewed to draw interpretations and conclusions.

This approach can be very time-consuming, even when using computer-assisted qualitative analysis software, and requires domain-specific expertise and methodological skills. It can therefore be costly and prohibitive for systematically analyzing large numbers of documents.

Artificial Intelligence Approaches to Automating Content Analysis and Qualitative Evaluation Synthesis

The pilot used natural language processing (NLP) methodologies to automate the process of thematic content analysis and explore the potential for NLP to complement traditional evaluation syntheses through greater speed, scope, and efficiency. NLP is a subset of AI technologies that enables computers to process and structure human language as natural language data, which can then be analyzed and interpreted in a meaningful way.

The AI methodologies used for NLP in the study included vector space modeling, supervised machine learning (SML), and unsupervised machine learning (UML), combined with knowledge graph approaches to perform an advanced theory-based content analysis of the project documents. A summary of how these methods compare with traditional evaluation synthesis methodologies is presented in figure 2.1.

Vector space modeling is a technique used in understanding the structure of language in documents so that text can be transformed into numerical representations (for example, vectors). Representing the text as a numerical vector allows the computer to perform calculations for machine learning.

SML is a **deductive** approach that uses training material to teach AI models to code content in a similar way. Once these patterns are learned from sufficient training samples for each code, the learned patterns are then applied to make coding predictions on new data outside of the training set. In the case of this study, subject matter experts prepared labeled examples of project documents and the SML made coding predictions on blocks of text from these documents. The performance of the AI model is then assessed by benchmarking the AI coding predictions against new material that is unseen during model training. When the AI performance reaches an acceptable threshold, the model can be deployed to undertake automated analysis on the full body of data.

UML contrasts with SML in that it is an **inductive** and data-driven approach. This study used topic modeling, a UML technique that allows related concepts to emerge from the data based on self-learning and semantic clustering. This approach makes few initial hypotheses about the concepts or meanings in the data and does not use training material to emulate the coding of human experts. Instead, topics are discovered based on statistical similarity measures in the word and sentence fragments. The usefulness of the UML approach is assessed through expert judgment on whether the emergent topics and codes are coherent and provide new insights that add to the interpretation of the evaluation evidence.

Knowledge graphs organize and integrate data on an entity of interest and structure it according to an ontological model known as a knowledge graph schema. Mapping data onto a model in this way allows connections and relationships among the data to be identified, analyzed, and better understood. Inferences can be drawn that rely on complex domain knowledge (represented by the knowledge graph schema), which is not possible when using machine learning classification methods alone. Knowledge graphs therefore work in complement with machine learning approaches to add context, depth, and reasoning ability that help explain the data-driven outputs of machine learning.

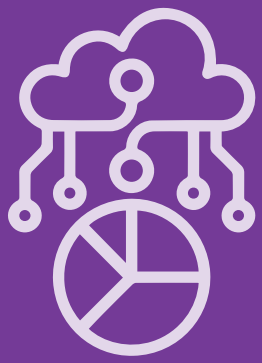
Figure 2.1. Comparison of Traditional and AI Approaches to Qualitative Evaluation Synthesis

Content analysis-method	Analytical framework	Data processing	Deductive analysis	Inductive analysis	Interpretation and conclusions
	Coding framework of themes and codes representing evaluation questions that test the results framework and hypotheses Framework structure is linear and hierarchical.	Experts systematically read all documentary evidence, ideally in electronic format by importing to analysis software.	While reading, experts subjectively label sections of text with predefined codes or write evidence summaries within framework. Experts identify patterns in coding and their meaning.	While reading, experts notice new emergent themes not considered in coding framework. New themes are assigned a code that is then systematically applied through rereading all documents.	Text coded under each section of the framework is reread in the context of the evaluation question or hypothesis tested. Experts triangulate this evidence, look for relationships, and draw evaluative judgments.
	Topics and labels representing individual components of the results framework, for example, inputs, outputs, and outcomes Label organization is multidirectional and continuous.	Vector space modeling is used to understand the structure of the language in documents so that text can be transformed into a numerical representation (for example, vectors). This allows the computer to perform calculations for the purposes of machine learning.	Supervised machine learning model trained to emulate human coding. Once trained, the AI model objectively labels documents. But training material may involve coders' manual and subjective judgment. Patterns are identified by visualization and semantic similarity statistics.	Unsupervised machine learning. AI uses topic modeling to identify similarities in text and group them as related topics. Experts read sample text and assign a new topic name. Visualization identifies semantic relationships among topics.	Knowledge graphs. Labeled content is mapped into a schema representing the results framework. Experts use pattern mining to query the data for characteristics or relationships. Statistical evidence is used as a basis for conclusions.

Source: Independent Evaluation Group.

Note: AI = artificial intelligence.

3
RESULTS



Can supervised machine learning automate theory-driven content analysis?

Can unsupervised machine learning offer new and important emergent insights into project data?

Can knowledge graphs organize data into a theory of change and help determine program contribution?

Can Supervised Machine Learning Automate Theory-Driven Content Analysis?

The first learning goal was to assess whether SML can use text classification models to accurately replicate manual content analysis and to consider whether these models can perform analyses quickly and efficiently. Project reports from a set of interventions within the World Bank's stunted growth and chronic malnutrition portfolio were selected as the body of evidence to pilot this approach. This included 392 unique project reports from 64 countries, with a total commitment of $28.8 billion.

The first step involved the production of training material in the form of correspondence tables. World Bank IEG experts reviewed the content from the various types of project reports and their relevant sections to select text of interest. This text was then pasted into the correspondence table using a simple data extraction template and labeled with hierarchical classification codes (levels 1 through 4) that fell under five label categories. To make the labeling process theory driven, the classification categories and codes related to the various components of the portfolio ToC. There were 74 labels under the three main label categories: "nutrition challenges addressed," "interventions," and "outcome indicator achievement." Text placed under the "factors affecting success or failure" label category was left unclassified except for labeling the content as either a "success factor" or "failure factor." Examples of the categories and labels used are presented in table 3.1.

The content from the correspondence table guidance material was then used to train the (multiclass) text classification model. Data from all countries and label types were used to train the model. The exceptions were data from the pilot country, which was retained for use as an unseen test case to assess the performance of the model, and data under the "factors affecting success or failure" label, which was left uncoded for use in the UML model.

Initial exploratory analysis suggested that training the text classifier model using these data would be challenging. This was because the data were highly imbalanced by class. This occurs when some labels are used frequently (more than 200 inputs per label) and others are rarely used (fewer than 10 inputs per label). As a result, there are many types of labels to predict but few examples of less frequent labels to learn from.

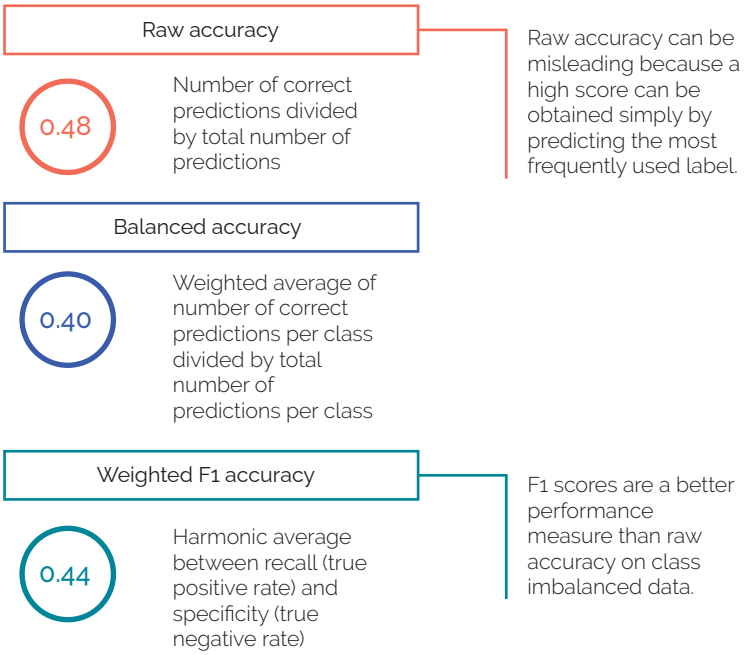
Table 3.1. Example of Correspondence Table Guidance Material Categories and Labels

Example	Project Identifier Information	Nutrition Challenges Addressed	Interventions	Outcome Indicator Achievement	Factors Affecting Success or Failure
1	Project ID	Underlying determinants of undernutrition	Institutional strengthening	Underlying determinants of undernutrition	Factors affecting success or failure
2	Country	Access to health services	Support to improve nutrition service delivery	Access to health services	Failure factor
3	Document type	Utilization of child health care		Utilization of child health care	
4	Section of document			Outcome achieved	

Source: Independent Evaluation Group.

Despite the challenges inherent in highly class-imbalanced data, the model showed promising results in its ability to predict expert labels from training data. This was achieved by using multiple classification algorithms and different preprocessing and feature engineering methods to optimize the AI model. Once the model was optimized, it was tested on the unseen project data from the pilot country. Performance was modest in terms of the model's ability to predict the exact classification label, with a weighted F1 accuracy score of 0.44 (figure 3.1).

Figure 3.1. Supervised Machine Learning Accuracy Scores for Predicting Exact Labels in the Pilot Country Test Case



Source: Independent Evaluation Group.

However, this performance must be understood within the context of selecting from 74 different class-imbalanced and hierarchical sublabels at the lowest classification level, among three label categories at the top level (nutrition challenges addressed, interventions, and outcome indicator achievement). This is a particularly challenging task.

When assessed by whether the model predicted the correct top-level category (even if it got the sublabel wrong) the algorithm performed much better, with an average accuracy of 192 out of 202 predictions. The model

was able to predict the correct top-level category with an average accuracy of 90–95 percent (F1 and raw or balanced accuracy score). The average size of the training set was 274 projects, and the average size of the held-out test set was 118 projects.

The high accuracy in predicting top-level categories suggests that the results achieved in this preliminary work could be improved. This would enable the text classifier model to provide direct statistical evidence to test the ToC by tallying results for labels that correspond to different components of the ToC.

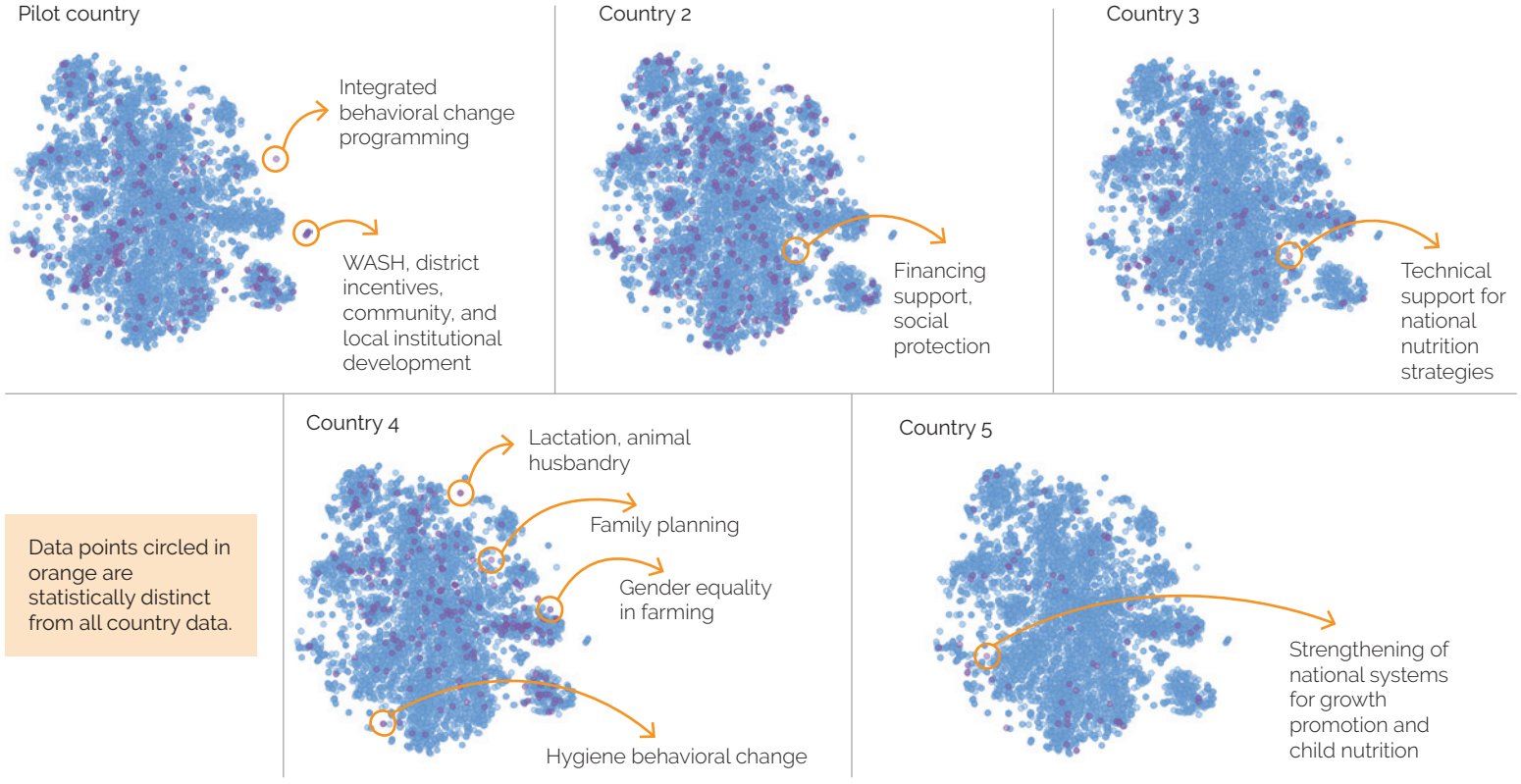
However, even at this early stage of development, the SML offers important insights that cannot be obtained through traditional methods. For example, to determine whether the pilot country would be an appropriate test case for the text classifier model, we converted the text data into numerical form and then used t-distributed stochastic neighbor embedding (t-SNE) visualization to determine if the pilot country and 11 other countries were statistically distinct from all other countries in terms of their nutrition challenges addressed, interventions, and outcome indicators achieved.

We used NLP and vectorization methods to translate each high-dimensional data point into a simple two-dimensional representation that preserves valuable information about the original meaning of the text; data points that are similar to each other in original meaning are closer together than data points that have different meanings. We represent these two-dimensional data points in a t-SNE visualization where the data points that are close to each other refer to similar issues or topics.

Using this t-SNE visualization, we were quickly able to provide a powerful visual assessment that showed only 17 data points from five countries were distinct, out of a data set of 5,468 data points. The t-SNE visualizations from these five countries are shown in figure 3.2. The diagrams show how we can project high-dimensional text data onto a two-dimensional plane where points with similar meanings are close to each other.

The results suggest relatively little variability among World Bank country programs except in specific areas. For example, the pilot country showed a statistically distinct focus on water sanitation and hygiene. The pilot country's water, sanitation, and hygiene cluster is distinct because it is the most geometrically distant of all clusters and only contains pilot country data, unlike other (less separated) clusters, which contain data from multiple countries.

Figure 3.2. t-SNE Visualization of Statistically Distinct Challenges, Interventions, and Outcome Indicator Achievement for the Five Selected Country Programs



Source: Independent Evaluation Group.

Note: t-SNE = t-distributed stochastic neighbor embedding; WASH = water, sanitation, and hygiene.

It is important to note that our pilot text classifier model cannot yet be considered automated. This is because it was trained and tested on text that was manually extracted into correspondence tables by World Bank IEG experts. To avoid this manual and time-consuming step, a model that can identify, extract, and format content of interest ready for labeling directly from raw project documents would need to be built.

The effectiveness of NLP in doing this will be affected by the quality of the documents being read. Inappropriate pagination, inconsistent chapter headings, unusual text formatting, and so on would degrade the performance of the text extraction. Given these intricacies, developing an algorithm that identifies content of interest for labeling was out of the scope of this pilot.

However, IEG has now developed an automated document section extraction routine that can identify and extract specific sections from raw documents with a high degree of accuracy. Combining this approach with the text classifier model could help overcome the manual step of extracting text of interest before labeling, thus bringing the approach closer to wider-scale automated label prediction. This should be explored further in subsequent research.

Can Unsupervised Machine Learning Offer New and Important Emergent Insights into Project Data?

The second learning goal was to determine whether UML using topic modeling could generate new and important emergent insights from a large data set. The data set used for the topic model included all text classified under the "factors affecting success or failure" label category. These data were chosen for the pilot because they represented an opportunity to determine whether topic modeling could be used to generate new knowledge from this important data set without the use of more resource-intensive manual methods.

Inductive approaches like topic models or traditional ground-up analyses make no prior hypotheses or expectations of what themes or topics will be found within the data. Instead, topics are identified or emerge from the data while the content is being read. As such, inductive approaches are exploratory and hypothesis generating (What themes will be found in the data and how might they be important?), and deductive approaches are confirmatory (What evidence is there that intervention X affects project outcomes through mechanism Y?).

Our first analysis involved performing topic modeling and t-SNE visualization on all the data in the "factors affecting success or failure" category (for topic modeling, see Whye Teh et al. 2006; Sai-hung and Flyvbjerg 2020). By pooling both the success factors and failure factors we could explore whether there were any common themes or relationships between success and failure factors, such as the failure factors being the inverse or absence of success factors.

The topic model algorithm clusters the success and failure factor data together and extracts statistically significant guide words for each cluster. Texts with the highest scores against the guide words in each cluster were then selected. These prime examples of guide word text were then reviewed by a domain-specific expert to determine if they represented a coherent topic. The expert then used their domain-specific knowledge and the model-provided guide words to interpret the content and develop topic descriptors. These topic descriptors therefore represent factors that are hypothesized to be important to both the success and failure of projects. The 10 topics identified through the topic modeling and their hypothesized topic descriptors are presented in table 3.2.

Once the topics were validated, they were mapped as data points in a continuous space using a t-SNE model to visualize similarity among topics. Simply put, the closer topics are to each other, the more meaning they have in common. But by preserving a holistic view, as opposed to a one-to-one comparison, we can see that the different clusters vary smoothly over the space in a continuous way. This allows graduations in meanings to be interpreted. The results of the t-SNE visualization are shown in figure 3.3.

As can be seen in figure 3.3, topics running in a north-south direction (y-axis) tend to transition from underlying sociopolitical themes to more technical themes. Topics running in an east-west direction (x-axis) tend to transition from project-specific themes to more contextual country- or system-specific themes. Proceeding clockwise from topic 1, the topics also approximately correspond to the key elements of a project implementation cycle, especially when topics are grouped into broader quadrants.

The performance of the topic model in generating such coherent and domain-relevant inductive topics is very promising, especially given that the inductive topics were novel and provided greater depth to the analysis than was originally envisaged within the predetermined labels of the deductive SML approach.

Table 3.2. Artificial Intelligence–Identified Inductive Topics and Their Hypothesized Topic Descriptors

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Systems, Procurement, Monitoring, Budgets	Outcome Ratings, Revising, Objectives	Operations, Communication, Finance	Coordination, Timeliness, Gender	Donors, Finance, Community, Service Providers
Program Design and Setup	**Adaptive Management**	**Performance Improvement Strategies**	**Implementation**	**Risks**
» Multisectoral project design » Objectives appropriate to scope » Promoting equitable access » Budgeting cycle and disbursements » Procurement	» Dynamically adjusting project objectives and M&E framework to project context	» Flexibility to adapt » Multisectoral coordination » Continuity of program and staff » Training and communication » Use of third parties to fill capacity gaps » Performance-based financing and incentives	» Technical implementation and adaptation » Government commitment and operational capacity » Needs-based targeting of beneficiary groups	» Community-level interventions and beneficiary inclusion » Donor-institution alignment in objectives and financing » Political instability » Natural disasters

(continued)

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Capacity, Quality, Risk	M&E	Local, Community, and Social Context and Services	Implementing Agency Performance, Coordination, Staff	World Bank and Government Performance, Political Commitment and Stability, Indicators
Risk Mitigation	**Leadership and Management**	**Operations**	**Sustainability Factors**	**Evaluation and Performance Review**
» Aligning program to context » Implementation planning » Capacity assessment and availability » Risk assessment	» Working with local leaders and institutions—benefits and risks » Project management: scope, risk, quality, monitoring, stakeholder management, and lessons learned	» Planning and coordination » Monitoring, information systems, and risk management » Capacity, team commitment, and flexibility » Inclusivity, empowerment, collaboration, and trust » Community leadership » Multisectoral, systems, and community-based approach to primary health care	» Government-led implementation, coordination, and capacity » Stakeholder and participant coordination » Financial sustainability	» Specific and meaningful KPIs for managing project performance » Civil unrest affecting project delivery » Accountability

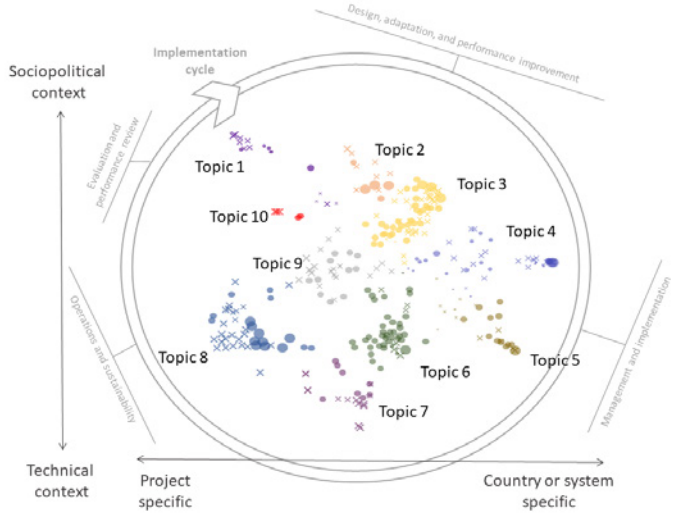
Source: Independent Evaluation Group.

Note: Artificial intelligence–identified inductive topics are shown in gray rows and their hypothesized topic descriptors are shown in blue rows. KPI = key performance indicator; M&E = monitoring and evaluation

The spatial visualization also shows significant value added in terms of its insights, not just over deductive approaches but also over traditional inductive approaches, which would have struggled to identify these patterns.

As a secondary analysis, we performed topic modeling on the data labeled as "success factors" and "failure factors" separately (Chapman and Quang, 2021). This was to determine whether the success factor data and failure factor data share common topic descriptors and guide words within the 10 topics or whether themes within the 10 topics are specific to either success or failure. This analysis also allows us to check whether the 10 topics identified in the pooled analysis comprehensively explained all the data, or whether additional, unidentified topic clusters may emerge. This is best practice in traditional thematic analysis, where the goal is to reach thematic saturation.

Figure 3.3. t-SNE Visualization of Artificial Intelligence–Identified Inductive Topics and Hypotheses for Spatial Patterns



Source: Independent Evaluation Group.

Note: t-SNE = t-distributed stochastic neighbor embedding.

When topics were analyzed as separate cohorts, no new topic clusters emerged. This confirmed that the analysis had reached thematic saturation. As hypothesized, the success factor data and failure factor data were not distinct within topics. Instead, the two sets of data shared strong similarities and overlaps, confirming that the failure factors were essentially the inverse or absence of success factors. This also explains why the 10 topics in

the pooled analysis explained all the data. For instance, the largest cluster in the failure factor data included topics related to misalignment of project designs to country contexts, overly optimistic objectives, insufficient local coordination, civil instability, and natural disasters. This failure factor topic is covered by topic 5 in the pooled analysis and is the opposite of a success factor topic.

This observation lends itself to the hypothesis that the 10 topics represent prerequisites, conditions, and actions that are necessary for project success; when they are in place, projects usually succeed; when they are not in place, projects often fail.

It may seem remarkable that such an important hypothesis can be generated through a UML model with no guidance except for labeling texts as success factors or failure factors. However, this hypothesis appeared highly plausible given the obvious alignment among the 10 topics and widely recognized good practices for international development programs (Ika, Diallo and Thuillier, 2021; Ika and Donnelly, 2017). As suggested by our topic model, the international development community also regards these good practices as prerequisites for program success.

Subsequent statistical analyses performed by IEG found that the UML-identified topics turned out to be key predictors of project performance, as measured by nutrition indicator achievement. This confirmed the usefulness of topic modeling for identification of success factors, providing supporting empirical evidence for good practices in international development. Since these factors were identified at the project rather than the intervention level, this exercise can likely be replicated in other thematic evaluations.

In our final analysis, we explored whether project-specific characteristics determined the importance of any of the 10 prerequisites for project success and ultimately the likelihood that the project would achieve its objectives. It was hypothesized that the multidimensionality of a project may be an important characteristic in determining challenges faced and project success. Therefore, we calculated a multidimensionality score for each project, which measured the variety of interventions used to tackle stunted growth and chronic malnutrition; the higher the score, the more interventions applied.

Using t-SNE clustering that analyzed success and failure factors separately, we found that clusters relating to topics 3, 7, and 9 were common across the full range of multidimensionality scores. However, clusters relating to topics 1, 4, 6, and 8 were particularly prominent in projects with low multidimen-

sionality scores. This suggests that performance improvement strategies (topic 3), monitoring and evaluation (topic 7), and sustainability factors (topic 9) are universally important. But projects that use a limited number of interventions should pay particular attention to project design and setup (topic 1), implementation (topic 4), risk mitigation (topic 6), and operations (topic 8).

The potential importance of using different project management strategies depending on the multidimensionality of the project is supported by further analysis that showed that high- and low-dimensional projects had different success and failure factors in common, particularly those relating to institutional strengthening.

This demonstrates that although the topics for success factors and failure factors all fall within the 10 prerequisites for project success, the project-specific factors affecting success do vary by multidimensionality. However, although we observed a positive relationship between multidimensionality score and outcome indicator achievement, this was not statistically significant. Therefore, we cannot conclude that projects with a greater variety of interventions are more likely to be successful.

Overall, the UML approach showed excellent performance in identifying inductive topics that were not only coherent and domain relevant but also novel and insightful. This added extra depth to the SML analysis, including identifying patterns that a human-generated analysis may have missed. The key advantage of the topic modeling or UML approach is its scalability. Manual inductive analyses can be very time-consuming, which means that they are often limited in the number of documents they can include.

Although the UML approach required significant time investment to extract text of interest for the topic model, the analytical work and interpretation of the outputs was completed very quickly. Therefore, although topic modeling or UML may not result in significant time savings for smaller studies, it quickly becomes advantageous for larger studies. If the process becomes more automated, these time savings could become even more significant.

However, the reality is that UML approaches will always be dependent on and strengthened by deploying them in tandem with manual methods. Traditional approaches are needed to provide the domain expertise to identify the need for a topic model through in-depth smaller studies and to interpret and report the output. The AI approach then offers the analytical power to scale these studies beyond the possibilities of normal qualitative analyses.

Can Knowledge Graphs Organize Data into a Theory of Change and Help Determine Program Contribution?

The final learning goal was to understand the potential of knowledge graphs to structure machine learning outputs according to a ToC and enable a theory-based evaluation of program performance.

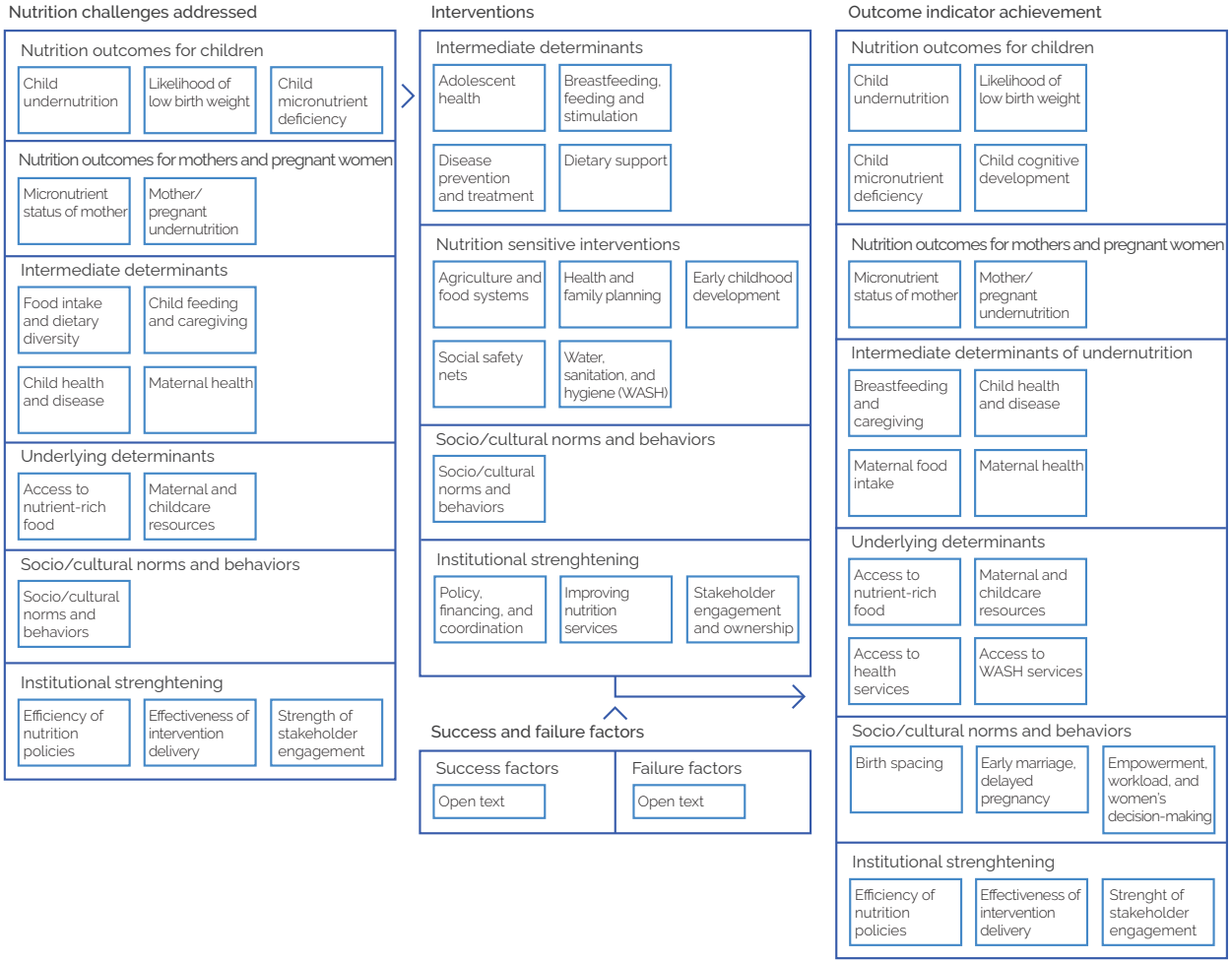
Within the field of evaluation, knowledge graphs could act as a smart ToC to streamline the portfolio review process by facilitating retrieval and interrogation of multiple sources of evidence on a specific domain of interest. For instance, content on intervention performance that has been coded by machine learning can be mapped onto a knowledge graph schema based on the program ToC. Evidence on each component of the ToC can then be reviewed, compared, and analyzed for causal relationships, thus helping to determine program contribution and answer evaluation questions.

To undertake these analyses, we needed to develop a knowledge graph schema that mapped the semantic relationships between key categories of the data to corresponding concepts in the stunted growth and chronic malnutrition ToC. Arranging the data into this logical flow was a simple exercise because the machine learning labels already represented components of the ToC: nutrition challenges addressed, interventions, outcome indicator achievement, and factors affecting success or failure. A high-level diagrammatic representation of the knowledge graph schema, or ontological model, is presented in figure 3.4.

Once the conceptual pathways between label categories were established, data from the machine learning outputs were used to generate a knowledge graph. This smart ToC can be statistically interrogated to directly explore relationships in the data that might suggest contributing factors. Analyses can also be run to provide evidence for programmatic decision-making or specific evaluation questions.

In this pilot, we focused on identifying success rates of different interventions at country and regional levels. The performance of interventions was tested by identifying the outcomes that an intervention was hypothesized to result in (as per the conceptual framework), and then testing whether the outcome's indicator was labeled as achieved or not achieved.

Figure 3.4. High-Level Diagrammatic Representation of the Knowledge Graph Schema

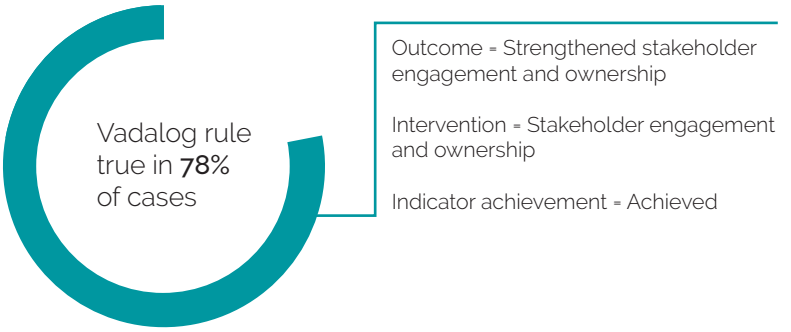


Source: Independent Evaluation Group.

Note: WASH = water, sanitation, and hygiene.

This kind of pattern mining uses Vadalog rule-based reasoning to test the rule "when outcome = X and intervention = Y, indicator achievement = achieved." Each rule has a confidence, which represents the ratio of how often the rule has been found to be true. For example, figure 3.5 shows that the outcome "strengthened stakeholder engagement and ownership" has been achieved in 78 percent of the cases when "stakeholder engagement and ownership" interventions are used.

Figure 3.5. Rule-Based Pattern Mining



Source: Independent Evaluation Group.

The results from a similar set of knowledge graph queries on all closed projects in the pilot country data set are presented in figure 3.6.

This initial investigation was limited to specific pairings between interventions and outcomes. Therefore, although we can determine if intervention^{X} from project^{Y} resulted in outcome^{Z}, this offers limited insight into the totality of outcomes that a project or intervention may have achieved. This is because, in reality, the data are more complex. One intervention may have multiple outcome indicators with different achievement status, and outcome indicators are not mutually exclusive because the same indicator label can be targeted by different interventions.

This multilabel complication can be overcome through the development of more granular labels and knowledge schema based on a more detailed ToC. This would enable the mapping of more specific relationships among nutrition challenges addressed, interventions, and outcome indicator achievement. This is similar to the context, mechanism, and outcome configurations used in realist analysis to determine program contribution. Other labels could be added to provide evidence on intervention delivery such as implementation fidelity.

Figure 3.6. Interventions and Outcome Achievement Status for Closed Projects in the Pilot Country Portfolio

Source: Independent Evaluation Group.

Note: Data from active projects are not included. One project may employ multiple interventions, and one intervention may have several outcomes. WASH = water, sanitation, and hygiene.

In this pilot, indicator achievement status had to be identified from a separate internal database containing baseline, target, and final values because project reports did not explicitly state outcomes for each indicator. This meant that a more granular labeling approach would be too time-consuming.

However, future work could explore the possibility of a classifier model being trained to read logframes or other similar outcome frameworks. This could enable automated capture of quantitative results and outcome achievement status, making it feasible to populate a more granular knowledge schema. The inclusion of quantitative results could also help establish the size of outcome achievements and may provide benchmarks for comparing projects.

If successful, such logframe analyses could open possibilities for tracking project implementation status and highlighting risks. Similar AI-enabled project management systems are routinely used in the infrastructure sector to predict projects at risk of failure but have not been adopted in the development sector despite increasing project size, budgets, and complexity.

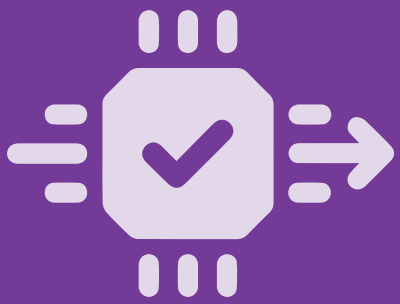
Currently, our knowledge graph is useful for structuring and exploring simple patterns in machine learning outputs that could be used to provide evidence for evaluations on intervention effectiveness and program contribution. However, one of the key outstanding challenges is dealing with incomplete evidence linking elements of the conceptual framework. For instance, an intervention may not have achieved its original objective but may have resulted in alternative outcomes, or project adaptation may make outcome pathways harder to trace. This incomplete evidence trail can make it difficult to build a unified knowledge graph because different parts of the network can be left stranded or orphaned from logical precursors or successors.

It is possible that additional data could be added to the knowledge graph, or graph reasoning methods could be used to populate missing nodes or identify alternative pathways—much like how an expert human evaluator would approach this problem. But as in traditional approaches, it would take time to build a more granular schema that incorporates all causal paths and the output of the knowledge graph will always depend on the breadth, depth, and quality of available data.

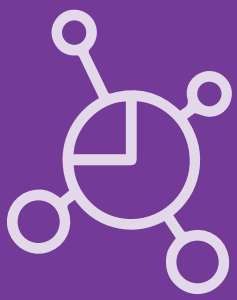
4

CAN ARTIFICIAL INTELLIGENCE ACCELERATE THEORY-DRIVEN COMPLEX PROGRAM EVALUATION?

Supervised machine learning with text classification models



Unsupervised machine learning to generate novel emergent findings



Knowledge graphs for structuring machine learning outputs

This pilot study set out to test the applicability, usefulness, and added value of using AI for advanced theory-based content analysis within the framework of IEG's thematic evaluations.

Our first learning goal was to assess whether SML can use text classification models to accurately replicate manual content analysis and to consider whether these models can perform analyses quickly and efficiently. Performance at predicting exact sublabels was modest and would require more time to optimize the model before being applied independently to new data sets. However, the high accuracy in predicting top-level categories suggests that the results achieved in this preliminary work could be improved. With some further work, the possibility of a text classifier model with coding accuracy that is acceptable when compared with expert coding is therefore promising.

The current approach requires significant effort on the part of subject matter experts to extract sections of text of interest ready for labeling. However, IEG's new automated document section extraction routine could be used to overcome this manual step. Therefore, further research aimed at combining the document section extraction routine with the classifier model could bring the approach closer to automated label prediction and the goal of faster and more efficient analyses.

Beyond speed and efficiency, the exploratory data analysis and visualization capabilities used for SML offer valuable additional insights to complement traditional methods. The ability to identify and visualize unique characteristics within country programs could prove useful for informing program management decisions and evaluation designs. Contextual relevance of interventions and evaluation findings is a perennial concern, so being able to establish the distinctiveness of a country program could help determine the transferability and suitability of lessons learned.

The second learning goal was to determine if UML approaches could generate novel and important emergent insights from a large and rich data set. Overall, the topic model showed excellent performance in identifying inductive topics that were not only coherent and domain relevant but also novel and insightful.

At this stage of development, it is particularly promising that the AI was able to generate and provide evidence supporting hypotheses on good practices for international development. That these topics were statistically shown to be key predictors of project performance by IEG proves not only that the

UML findings were robust and meaningful but also that this approach could provide important evidence for programmatic decision-making.

Although the UML approach will always require domain expertise from a human expert to validate and interpret the topic modeling results, it offers significant potential to complement traditional methods by providing an opportunity to scale inductive analyses to much larger data sets. The visualization methods could also help identify patterns that would not be obvious to a human analyst.

The final learning goal was to understand the potential of knowledge graphs to structure machine learning outputs according to a ToC and enable theory-based evaluation of program performance. Our preliminary investigations show that rule-based reasoning can be used to identify simple relationships among components of a knowledge schema representing a ToC. However, difficulties remain with multilabel modeling and the complexities of setting up a more granular and unified knowledge graph schema to comprehensively represent possible causal pathways.

In its current form, rule-based reasoning can provide only supporting quantitative information on project outcomes achievement. Therefore, further research is required before knowledge graphs can enable a theory-based evaluation of program performance.

An overarching limitation of all the AI approaches used in this pilot study is the significant front-end investment required to establish the analytical framework and prepare data for machine learning. Although this may be reduced by developing and integrating the approach with IEG's document section extraction routine, the use of AI for theory-driven content analysis will currently only be worthwhile for larger-scale studies, particularly those that would benefit from being regularly updated. However, where organizations use standardized reporting templates, the initial investment could be leveraged by applying the SML and UML models to multiple portfolios with minimal tailoring.

Although cognizant of these limitations, we must remember that the findings presented in this white paper are part of only a small body of work exploring the use of AI for evaluation synthesis purposes. For such early-stage work, the preliminary results are promising, especially those for SML and UML technologies. Given the acceleration in the deployment and sophistication of AI technologies in other sectors, we may be optimistic about what the future holds for the use of AI technologies in the development evaluation sector.

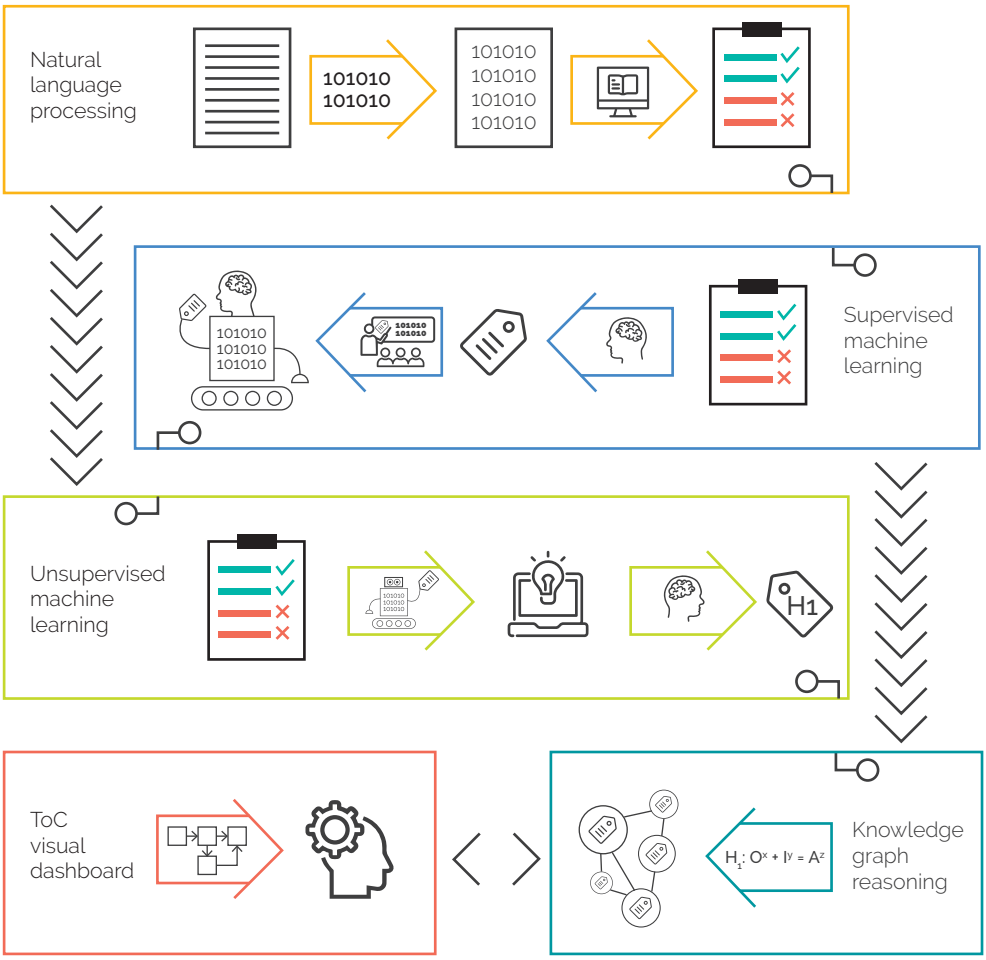
In this context, and to inform future research priorities, it is worthwhile to consider what a fully functional AI-enabled theory-driven content analysis might look like. Figure 4.1 provides a representation of the sequence and articulation of the multiple AI models and capabilities that need to be refined to maximize the potential applications of AI for theory-based portfolio analysis. First, the model would automate identification and extraction of text of interest from raw documents using NLP approaches, accurately predict deductive labels from extracted text using SML technologies (text classification model), identify emergent labels and patterns in the data using UML technologies (topic modeling), and then structure and interrogate the evidence using theory-based analysis in the form of a knowledge graph and Vadalog rule-based reasoning. The inclusion of a simple, engaging dashboard that could be used to easily drill down into the ToC could further improve the interpretability of the findings.

This paper represents one of the few and earliest investigations into this field and clearly lays out pathways for further research and refinement. If significant further research were to build on this work, development timelines could arguably be accelerated. Developments in the field of evaluation can be inspired by other sectors where AI technologies are already used for complex project management and risk quantification purposes (Sai-hung and Flyvbjerg 2020, Aldana et. al. 2021).

In conclusion, the results of this feasibility study show that NLP approaches can be usefully applied to theory-based content analysis and add significant value to complement existing synthesis methods. SML and UML are promising, particularly topic modeling and t-SNE visualization, and knowledge graphs can identify simple relationships in the data according to a conceptual framework. However, additional work is needed to automate document extraction, optimize text classification models, and develop more sophisticated knowledge graphs capable of analyzing more complex theories of change.

At their current stage of development, these AI technologies can be a useful tool for increasing the scope, speed, and depth of larger portfolio review processes. But further research addressing the limitations identified in this study could enable wider adoption of AI within the field of development evaluation and leverage big evaluation data.

Figure 4.1. Maximizing the Applications of Artificial Intelligence for Theory-Based Portfolio Analysis



Source: Independent Evaluation Group.

Note: The figure shows the sequence and articulation of the multiple artificial intelligence models and the capabilities that need to be refined to maximize the potential applications of artificial intelligence for theory-based portfolio analysis.

REFERENCES

Aldana, A., K. Hay, C. LeGrand, P. Schmidt, and M. Bereni. April 2021. "O18:Exploring the Use of Artificial Intelligence (AI) Solutions to Improve the Accuracy of Project Delivery Forecasts". *National Asset Centre of Excellence, Brisbane, Australia*. https://www.nacoe.com.au/wp-content/uploads/2021/04/015402-NACOE-O18_AI-application-for-project-forecast.pdf.

Bamberger, M., E. Raimondo, and J. Vaessen. 2016. *Dealing with Complexity in Development Evaluation: A Practical Approach.* Thousand Oaks, CA: Sage Publications.

Black, R. E., C. G. Victora, S. P. Walker, Z. A. Bhutta, P. Christian, M. de Onis, M. Ezzati, S. Grantham McGegor, J. Katz, R. Martorell, R. Uauy, and the Maternal and Child Nutrition Study Group. 2013. "Maternal and Child Undernutrition and Overweight in Low-Income and Middle-Income Countries." *The Lancet* 382 (9890): 427-51. http://dx.doi. org/10.1016/S0140-6736(13)60937-X

Bravo, L., A. Hagh, Y. Xiang, and J. Vaessen. 2021. "Machine Learning in Evaluation Synthesis—Lessons from Private-Sector Evaluation in the World Bank Group." In *Evaluation of International Development Interventions: An Overview of Approaches and Methods*, edited by J. Vaessen, S. Lemire, and B. Befani. Independent Evaluation Group. Washington, DC: World Bank.

Chapman, Paul, and Cuong Quang. 13 February 2021. "Major Project Risk Management: Reconciling Complexity During Delivery with the Inside View in Planning". *Engineering Project Organization Journal* 10, no. 1. https://doi.org/10.25219/epoj.2021.00104.

Korenblum, J. 2017. "Three Ways Artificial Intelligence Can Reshape Monitoring and Evaluation." ICTworks. https://www.ictworks.org/artificial-intelligence-can-reshape-monitoring-evaluation/.

Ika, Lavagnon A., and Donnelly, Jennifer. 'Success Conditions for International Development Capacity Building Projects'. January 2017. *International Journal of Project Management* 35 (1): 44–63. https://doi.org/10.1016/j.ijproman.2016.10.005.

Ika, Lavagnon A., Diallo, Amadou, and Thuillier, Denis. January 2021. 'Critical Success Factors for World Bank Projects: An Empirical Investigation'. *International Journal of Project Management* 30(1): 105–16. https://doi.org/10.1016/j.ijproman.2011.03.005.

McKenzie, D. 2018. "How Can Machine Learning and Artificial Intelligence Be Used in Development Interventions and Impact Evaluations?" *Development Impact* (World Bank Blogs), March 5, 2018. https://blogs.worldbank.org/impactevalua-

tions/how-can-machine-learning-and-artificial-intelligence-be-used-development-interventions-and-impact.

Sai-hung, Ir Lam, and Bent Flyvbjerg. 2020. "AI in Action: How the Hong Kong Development Bureau Built the Project Surveillance System—An Early Warning System for Public Works Projects." White Paper, Hong Kong Development Bureau and Oxford Global Projects, Hong Kong and Oxford.

UNICEF. 1990. "Strategy for Improved Nutrition of Children and Women in Developing Countries." New York: UNICEF.

Whye Teh, Yee, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101 (476): 1566–1581.

World Bank. 2021. *The World Bank's Support to Reducing Child Undernutrition*. Independent Evaluation Group. Washington, DC: World Bank.

The World Bank
1818 H Street NW
Washington, DC 20433