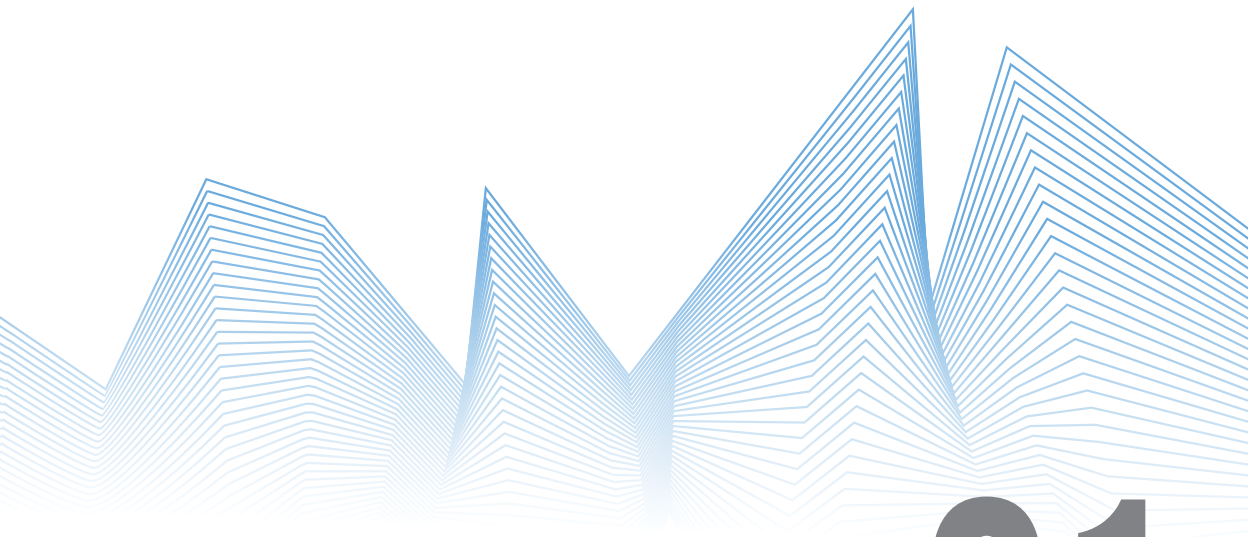
EBA



01
2021

DATA SCIENCE METHODS IN DEVELOPMENT EVALUATION: EXPLORING THE POTENTIAL

Gustav Engström, Jonas Norén

Data Science Methods in Development Evaluation: Exploring the Potential

Gustav Engström and Jonas Norén

Report 2021:01

to

The Expert Group for Aid Studies (EBA)

Gustav Engström holds a PhD in economics at Stockholm's University. For more than 10 years he has applied statistical and mathematical methods with a focus on environmental economic research. He has co-authored 15+ articles in top level academic journals including Science and Nature.

Jonas Norén holds a master degree (M.Sc.) in political science and economics. For the course of the last 15 years he has been engaged as an analyst, evaluator and data scientist within the fields of international development cooperation and private sector development.

Please refer to the present report as: Engström, G., Norén, J. (2021), *Data Science Methods in Development Evaluation: Exploring the Potential*, EBA Report 2021:01, The Expert Group for Aid Studies (EBA), Sweden.

This report can be downloaded free of charge at www.eba.se

ISBN 978-91-88143-63-1
Printed by Elanders Sverige AB
Stockholm 2021

Cover design by Julia Demchenko

Acknowledgements

We would first and foremost like to express our gratitude towards EBA that granted us the opportunity to conduct this study. The staff at the EBA secretariat – Lisa Hjelm, Markus Burman and Jan Pettersson – have been engaged and supportive throughout the process. We would also like to thank the appointed reference group that has given us valuable and constructive feedback throughout this process, which has made our work and report better in so many ways. The reference group was composed of: Torgny Holmgren (EBA), Kerstin Borglin (SPIDER), Charlotta Bredberg (Sida), Katarina Perrolf (Sida), Magnus Sahlgren (RISE) and Gustav Peterson (Vetenskapsrådet).

Table of Contents

Foreword by the EBA **1**

Sammanfattning **3**

Summary **8**

Acronyms and definitions **12**

Methods 12

Open-source packages 13

Other 15

1 Introduction **18**

2 Background and literature **23**

2.1 Requirements and general skill set needed 25

3 Methodology **27**

3.1 Natural language processing 27

3.2 Detailed walkthrough of study processes 34

3.3 Evaluating performance 37

4 Results **42**

4.1 Data collection and parsing of documents 42

4.2 Geography and time 48

4.3 Funding and donors 56

4.4 OECD/DAC evaluation criteria 63

4.5 Thematic area 74

5 Discussion 80

5.1 General strengths and weaknesses of a machine-based approach 80

5.2 Observed limitations in the study 84

5.3 Moving forward 86

6 Concluding remarks 90

References 93

Appendix 1 – Analytical framework 98

Previous EBA reports 106

Foreword by the EBA

As the flow of digital information has increased exponentially, the development of analytical methods to capture this information has followed. For some time now, we have heard about the potential of big data, artificial intelligence and machine learning also in the field of development cooperation, which is a field where analysts and evaluators often struggle with scarcity of basic data and lack of information, but there also exists numerous evaluations. It is clear that data science methods provide a huge potential in making use of this digital information, but are the methods user-friendly and accurate enough to be applied widely in development evaluation? In addition, use demands that evaluators and commissioners of evaluations are aware of available methods and the type of questions they can address.

This EBA report is an exploration into the potential of data science methods to identify, compile and analyse information from previously published evaluations of development cooperation projects. This opens for the possibility to quickly gain an overview of what previous evaluations of Swedish development cooperation say about for example the relevance or the sustainability of the evaluated projects. The advantages of machine-based methods, such as the transparency, replicability, and possibility to analyse large amounts of data in a short amount of time, are known. This report is testing what is feasible in practice.

The focus in the report is on analysis of text, using natural language processing (NLP) methods, and compare machine-based analysis with results from a manual assessment. As a large amount of the digital information is available in the form of unstructured text this is a method with potential for use in evaluation of development cooperation.

The authors find that descriptive statistics can be collected rapidly and effectively and that the levels of accuracy for this type of

statistics in general is in line with that of a manual assessment. Challenges occurred in more complex interpretation of results, such as if projects were deemed to be sustainable or not. The report also found that these, more complex types of interpretations varied in the manual, human interpretations pointing out how challenging it can be to interpret a varied and complex language even for experts in the field. The authors suggest that the results could be improved by further fine tuning the methods and adjusting how the questions were posed and by limiting the number of available response options. However, the authors also draw attention to the high energy consumption used for this type of analysis.

When should data science methods be considered then? The authors conclude that this depends on the available resources, requirements for transparency and replicability, and the need to be able to scale up the analysis. We hope that this report can contribute to a discussion between practitioners, evaluators, and data scientists around how these methods can be used and how challenges can be overcome in order to open up for better use of machine based methods in evaluation of development cooperation.

The study has been conducted with support from a reference group chaired by Torgny Holmgren, member of the Expert Group. The authors are solely responsible for the content of the report and its conclusions.

Gothenburg May 2021

Helena Lindholm

Sammanfattning

Mängden digital information som finns tillgänglig i dagens värld har expanderat kraftigt under de senaste åren. Många bedömare förväntar sig dessutom att utvecklingen kommer att fortsätta accelerera och följa en exponentiell tillväxttakt där volymen tredubblas under de kommande fem åren. Minskade kostnader för att processa information, ett ökat antal internetanslutna individer samt en samhällsomfattande digitalisering är några av drivkrafterna bakom denna trend. Den här utvecklingen skapar nya möjligheter för att generera kunskap och insikter för de myndigheter, organisationer och företag som besitter den praktiska kompetensen och den kapacitet som krävs för att kunna bearbeta stora mängder information.

För att kunna dra full nytta av den här utvecklingen så krävs både nya tekniker som bygger på beräkningsalgoritmer och maskinbaserad metodik, samt individer med rätt kompetens och förståelse för hur dessa tekniker och metoder kan och bör nyttjas. Inom yrkesområdet data science efterfrågas personer med bred kompetens för hantering, bearbetning, analys och visualisering av stora datamängder och ofta även en god förståelse, förmåga att kommunicera slutsatser och metodik, samt kapacitet att tillämpa de senaste metoderna som utvecklas kontinuerligt inom forskningsfältet. Eftersom mycket av den digitala informationen är tillgänglig i form av ostrukturerad text, behöver tillämpare också kunna använda sig av metoder inom språkteknologi och bearbetning av naturligt språk.

Den digitala utvecklingen har haft enorma effekter på stora delar av samhället, med applikationer som sträcker sig från klassificering och organisering av texter till maskinbaserad språköversättning och chattbotar. Utvecklingen har också påverkat hur forskning kan bedrivas samt vilka insikter som kan genereras. Detta gäller även i hög grad utvärderingsprofessionen där behovet av en vidgad analytisk verktygslåda för utvärderingar har lyfts fram som essentiell, framförallt om den ökande tillströmningen av information inom

fältet ska kunna hanteras och nyttjas på ett bra sätt. Detta antas även vara av vikt inom området för internationellt utvecklingssamarbete, där utvärderingar utgör ett viktigt instrument för att generera insikter som ligger till grund för ansvarsutkrävande och sund styrning av biståndsmedel. Inom detta område så har dock användandet av metoder inom data science hittills varit relativt begränsat.

Syftet med den här studien är att utforska potentialen med att använda metoder inom data science och språkteknologi i utvärdering av internationellt utvecklingssamarbete. Detta innefattar att både testa hur dessa metoder kan användas för att sammanställa och analysera data från tidigare utvärderingar inom internationellt utvecklingssamarbete samt att utvärdera styrkor och svagheter med dessa metoder jämfört med en manuell analys.

I studien används metoder och strategier för att hantera och bearbeta text från tidigare publicerade utvärderingar med målet att få fram resultat som replikerar utfallet från traditionella, manuella utvärderingsmetoder.

De metoder som använts bygger på språkteknologi eller Natural Language Processing (NLP) vilket är en metod för att bearbeta mänskligt språk. NLP bygger på tekniker för att exempelvis klassificera texter, sammanfatta texter eller identifiera nyckelord. Metoder inom NLP delas ofta upp i två olika typer av ansatser, en som bygger på förbestämda regler och en som bygger på maskininlärning. I realiteten kombineras ofta de två approacherna, vilket också är fallet i den här studien där vi för varje frågeställning har utarbetat en specifik strategi med en kombination av metoder för att kunna besvara en specifik fråga.

Grunden för utvecklingen av metoderna utgörs av tidigare manuella bedömningar som kommer från en studie som täcker innehåll och slutsatser från 128 decentraliserade utvärderingar som Sida beställt mellan 2012–2014. De tidigare bedömningarna har således fungerat som en slags träningsdata som väglett utvecklandet av maskinella metoder och algoritmer. De faktiska frågorna i denna övning

fokuserar på vissa aspekter av de decentraliserade utvärderingarna såsom geografi, finansiering, tematiskt område och projektens hållbarhet. Slutligen har de utvecklade metoderna, testats i en automatiserad process där alla Sidas decentraliserade utvärderingar mellan 2012–2020 (>300) inhämtas och analyseras.

Resultaten visar att det finns en stor potential i användandet av maskinbaserad språkteknologi för att på ett snabbt, effektivt och tillförlitligt sätt generera insikter och beskrivande statistik med en rimlig felmarginal. Värdefulla insikter kan genereras på ett sätt som gör att analysens omfattning och tidsåtgång inte längre är faktorer som behöver beaktas på samma sätt som vid manuella analysprocesser.

De initiala förväntningarna på denna studie var att dessa metoder skulle kunna appliceras för att få inblick i vad tidigare utvärderingar har kommit fram till vad gäller ett antal frågor som är relevanta för att styra framtida biståndsprojekt.

En majoritet av de metoderna som utvecklats och testats i denna studie har presterat bra och nästan alla har presterat bortom de ursprungliga förväntningarna. Detta innebär dock inte att ett maskinbaserat tillvägagångssätt är utan brister. Några bedömningar visade sig vara svårare och mer komplexa än vad vi ursprungligen trodde, främst på grund av ett komplext språkbruk och begränsningar i antalet tidigare bedömningar som fanns att tillgå som träningsdata. För att lyckas med de mer komplexa bedömningarna så skulle mer arbete och data behövas för att generera bättre resultat.

Vi har också noterat att det finns utmaningar och brister med manuellt utförda bedömningar av de slag som ligger till grund för denna studie. Vi lät granska ett slumpmässigt urval av bedömningar en oberoende expert med god sakkunskap med syfte att testa om frågorna hade formulerats på ett sätt så att svaren lätt kunde urskiljas av en tredje part med god sakkunskap. Denna valideringsövning visade att det fanns relativt stor skillnad i bedömningarna mellan den ursprungliga manuella bedömningen och den oberoende experten.

Detta var i synnerhet fallet för frågor som var mer komplexa eller vaga i sin formulering och därför försvårade en enhetlig bedömning mellan granskarna. Detta hände dock även för enklare frågor vilket kan tolkas som att uppgiftens repetitiva utformning kan ha varit ansträngande för en mänsklig bedömare. Från detta drar vi slutsatsen att särskild omsorg krävs när man formulerar frågor i dessa typer av metastudier för att undvika oenighet i bedömningar på grund av t.ex. vagt formulerade frågor.

Fördelen med automatiserade metoder jämfört med manuella bedömningar, är att de förutom att kunna generera snabb och intressant beskrivande statistik, har en potential att producera mer tillförlitlig statistik. Med detta menar vi att de automatiserade metoderna ofta är mer robusta i sina bedömningar i jämförelse med mänskliga bedömningar, vilka kan variera mer beroende på en rad faktorer som exempelvis tidpunkt på dagen, humör eller nattsömn. Eftersom både människor och dataalgoritmer sannolikt kommer att generera fel vid ett eller annat skede, är det viktigt att ha en förståelse för graden och typen av fel som kan uppstå. En annan viktig fördel med de maskinbaserade metoderna är möjligheten att snabbt kunna korrigera fel som upptäcks sent i en process. Om ett fel skulle upptäckas efter att en manuell arbetsprocess slutförts så skulle det troligen kräva mycket resurser för att åtgärda felet. Med ett maskinbaserat tillvägagångssätt kan den underliggande källkoden istället justeras, och analysprocessen kan sedan upprepas och generera nya resultat med en enkel knapptryckning.

Sammanfattningsvis anser vi att denna studie har påvisat att det finns potential för maskinbaserade metoder gällande bearbetning av mänskligt språk när det kommer till att sammanställa beskrivande statistik av utvärderingar inom internationellt utvecklingssamarbete. Värdet av detta tillvägagångssätt beror dock till stor del på de krav som ställs, eller med andra ord vilken grad av felmarginal som vi är villiga att acceptera. Från vårt perspektiv finns inga enkla svar på dessa frågor. Svaren varierar sannolikt utifrån vilken typ samt vikten av de beslut som planeras tas baserat på dessa bedömningar. Vissa

beslut och områden kan och kommer troligen kräva en liten felmarginal, medan andra kanske kan nöja sig med mindre robust statistik.

Summary

The amount of digital information in today's world has increased to unimaginable proportions, and the volume is expected to triple just over the next five years. Many expect that this development will continue, pick up pace, and proceed along an exponential growth path into the foreseeable future. Cost reductions for computational processing, increasing internet connectivity and society-wide digitalisation are examples of drivers behind this seemingly ever-increasing generation of digital content.

This development thus creates an opportunity for governments, organisations and companies to process this information in order to derive knowledge and new insights from it. Taking full advantage of this information, however, requires both the adoption of new techniques, relying on computational or machine-based approaches with quick processing, and individuals with the skill set that allows them to efficiently understand how to harness these techniques and methods. Over the last couple of decades, this has led to the creation of a new profession often referred to as data science, which captures the broad skill set involved in handling, processing, analysing and visualising large quantities of data. This often also includes understanding, communicating and applying state-of-the-art methods from computer science, machine learning and artificial intelligence. Furthermore, since much of the available data comes in the form of unstructured text, data scientists also need to make use of the most recent advances in research on language understanding and processing, known as computational linguistics or natural language processing. These methods are particularly important since they provide the foundations for gaining novel insights from large quantities of unstructured texts in ways that were not previously possible.

This development has had huge impacts on large parts of society, with applications ranging from the classification and organisation of texts to machine-based language translation and chatbots. It has also

affected how research can be conducted and which insights can be drawn from it. The practice of evaluation is no exception, not least due to its heavy reliance on data to conduct analysis and retrieve insights. Recently, calls have been raised relating to the need for a broadened analytical toolbox for evaluations if increases in the volume, velocity and variety of data are to be handled and taken full advantage of. This is believed to be the case particularly within the field of international development cooperation, where evaluations are an important instrument for obtaining insights as well as fostering accountability and sound governance.

The purpose of this study is to explore the potential of data science and natural language processing methods, and to assess how these methods may be applied to derive meta or secondary data from readily available evaluations within the field of international development cooperation. This entails developing custom strategies for data handling and processing in order to generate results that can replicate traditional evaluation methods that rely on manual labour.

In order to make progress, we use a manually annotated or so-called labelled dataset to guide the machine-based methods that are developed and tested. The labelled dataset comes from a meta study covering the content and conclusions of 128 decentralised evaluations commissioned by Sida between 2012 and 2014. These evaluations form a central instrument in the follow-up of Swedish projects and programmes within the realm of international development cooperation. The questions in the labelled dataset are directed at certain aspects of the decentralised evaluations, with a bearing on aspects such as geography, funding, thematic area and project sustainability. A final component of this study relates to testing a scaled-up exercise, in which the developed methods that are deemed successful are tested in an automated analytical process. All decentralised evaluations between 2012 and 2020 (>300) are fetched from Sida's web-based archive and automatically analysed in real-time over the course of a few minutes.

The initial expectations of this study were that, if successful, these methods could be readily applied to derive insights into what past evaluations have concluded regarding a number of questions of relevance for steering future aid projects and programmes. The advantage, in comparison to a manual assessment, would be that in addition to generating interesting descriptive statistics and a quick analytical turnaround, these methods would also give us better information about the reliability of the statistics produced. By this, we mean that since these types of methods are computational in nature, they are also typically more stable in their predictions and errors over time than human beings, who may vary more unpredictably in their assessments depending on a range of factors such as time of day, overall mood or hours of sleep. And since both humans and computer algorithms are likely to produce errors at one stage or another, having an understanding of the size and type of error is thus important. Another advantage with the machine-based approach is the possibility to correct for discovered processing errors. If an error were to be detected in a product from a manual labour process, this would require extensive resources to redo the work, while with a machine-based approach the underlying code could be adjusted for the detected error and the analytical process could then be quickly repeated with updated results in a matter of minutes, in many cases.

From our results, we have concluded that there is considerable potential when it comes to designing computational approaches that can derive valuable insights and descriptive statistics from past evaluations in a quick, efficient and reliable manner with a reasonable rate of error. Valuable insights can be generated in ways where scale, scope and timeliness are no longer factors of concern, allowing for the automation of many repetitive and strenuous tasks. The majority of the designed strategies/methods in this study have performed relatively well, and almost all of them have performed beyond initial expectations. This does not, however, imply that a machine-based approach and the tested methods are without flaws. A few questions proved to be harder and more complex than initially thought, mainly

due to a highly complex language context and limitations in the quality and availability of training data. In these cases, more work and higher quality data would be needed to produce desirable results.

This study has also revealed discrepancies among human-based assessments. These discrepancies were revealed using a random sample from the labelled dataset, to test whether a third-party expert evaluator would assign the same label as in the original dataset. The purpose of this was to test whether the questions were phrased in such a way that the answer could be easily distinguishable by a third party with good domain knowledge. This validation exercise revealed that the third-party and original assessment did not generally agree on the appropriate labels to a much greater extent than the machine-based approach did with the original assessment. In particular, this was the case for tasks which were complex in nature and thus required careful judgement from the evaluator, but it also occurred for simpler questions, indicating that the task of conducting a large number of manual assessments may indeed be strenuous for humans. From these results, we conclude that particular care is needed when phrasing questions in these types of meta studies in order to avoid annotator disagreement due to indistinct or complex labelling that blurs interpretability.

By and large, we believe this study has revealed that there is indeed a potential for data science and machine-based approaches to compile descriptive statistics for meta evaluations. However, the value of this approach depends largely on the standards to which we want it to adhere, or in other words what degree of error we are willing to accept. From our perspective, there are no straightforward answers to these questions. Answers are likely to vary with the type of decisions that are expected to be taken based on the estimations. Some decisions are likely to require a high degree of accuracy, while others can perhaps settle for less.

Acronyms and definitions

Methods

BERT – Bidirectional Encoder Representations from Transformers.

NER – Named-entity recognition is a sub-task of natural language processing (NLP), which aims to identify and classify named entities mentioned in unstructured text into predefined **categories.**

Sentiment classifier – Applies sentiment analysis system for text analysis which combines natural language processing (NLP) and machine learning techniques to classify text as positive, negative or neutral, based on a computed sentiment score.

Tf-Idf – Frequency–inverse document frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

Transformers – A type of deep neural network designed mainly to handle sequential data such as natural language. The transformers technology has paved the way for efficient pre-trained models of language, such as BERT and GPT.

Web scraper/scraping – A process used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler.

Word embeddings – Word embeddings are efficient, dense vector representations of words in which similar words have a similar encoding. They are capable of capturing the context of a word in a document, semantic and syntactic similarity, relationship with other words, etc.

Word2vec – A family of model architectures and optimisations that can be used to learn word embeddings from large datasets.

Zero-shot learning – The approach when the neural network is forced to make classifications for classes it was never trained for. In other words, the ability to detect classes that the model has never seen during training. This resembles our ability as humans to generalise and identify new things without explicit supervision.

Open-source packages

Python dependencies and packages that have been used in the study include:

country converter – country converter is a Python package to convert and match country names between different classifications and between different naming versions. (https://pypi.org/project/country-converter/)

FuzzyWuzzy – FuzzyWuzzy string matching package. This uses Levenshtein distance to calculate the differences between sequences. (https://github.com/seatgeek/fuzzywuzzy)

Gensim – Gensim is an open-source library for unsupervised topic modelling and natural language processing, using modern statistical machine learning. (https://radimrehurek.com/gensim/)

Hugging Face's transformers – State-of-the-art natural language processing library which provides thousands of pre-trained models to perform tasks on texts such as classification, information extraction, question answering, summarisation, translation, text generation, etc. in 100+ languages. (https://huggingface.co/)

JSON – JavaScript Object Notation is an open standard language-independent data format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and array data types (or any other serialisable value). (https://en.wikipedia.org/wiki/JSON)

Jupyter – Project Jupyter is a non-profit organisation created to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". (https://jupyter.org/)

NLTK – NLTK is a leading platform for building Python programs to work with human language data. (https://www.nltk.org/)

NumPy – NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. (https://numpy.org/)

PyMuPDF – A package for reading PDF files. The package opens PDF documents page by page, saves all its content in a block and identifies the text size, font, colour and flags. (https://pypi.org/project/PyMuPDF/)

Scikit-learn – Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms and more. (https://scikit-learn.org/stable/)

Scrapy – Scrapy is a fast high-level web crawling and web scraping framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing. (https://scrapy.org/)

spaCy – spaCy is a library for advanced natural language processing in Python and Cython. (https://spacy.io/)

Streamlit – Streamlit is an open-source app framework for machine learning and data science teams. (https://www.streamlit.io/)

SQLite – SQLite is a relational database management system contained in a C library. In contrast to many other database management systems, SQLite is not a client–server database engine. (https://www.sqlite.org/index.html)

Pandas – Pandas is a fast, powerful, flexible and easy-to-use open-source data analysis and manipulation tool built on top of the Python programming language. (https://pandas.pydata.org/)

PDFMiner – PDFMiner is a text extraction tool for PDF documents. (https://github.com/euske/pdfminer)

Plotly – Plotly creates and stewards the leading data visualisation and UI tools for machine learning, data science, engineering and sciences. (https://plotly.com/)

pycountry – pycountry provides the ISO databases for the standards relating to languages, countries, deleted countries, subdivisions of countries, currencies and scripts. (https://pypi.org/project/pycountry/)

PyMuPDF – MuPDF can access files in PDF, XPS, OpenXPS, epub, comic and fiction book formats, and it is known for its top performance and high rendering quality. (https://pypi.org/project/PyMuPDF/)

Python – Python is an interpreted, high-level and general-purpose programming language. (https://www.python.org/)

Other

AI – Artificial intelligence encompasses a wide range of subfields, and even though there is no universal definition, a common feature is the design of intelligent systems, which stretches in a continuum from simple tasks such as regulating indoor temperature to designing general intelligence in line with human-level intelligence (Russel and Norvig, 2016).

Data science – A discipline at the intersection of mathematics, statistics and computer science, where data is the underlying driver for analysis and for retrieving insights.

EBA – The Expert Group for Aid Studies.

HLP – Human-level performance is a benchmark technique where the machine learning model's accuracy is compared to human performance on a given task.

LME dataset – Labeled Meta-Evaluation Dataset from the study EBA2017:12, which has enabled this study with an analytical framework used as a training and test dataset.

Machine-based approach – Refers to an automated approach that is made possible by the capacity of computers, computer programming and storage capacities.

ML – Machine learning refers to processes where computer algorithms have been crafted to learn and make predictions based on previous observations.

NLP – Natural language processing.

Natural language annotation – Refers to the process of establishing metadata or descriptive labels for underlying observations (or any kind of data) with the purpose of augmenting an algorithm's capability to give accurate predictions (Pustejovsky and Stubbs, 2012).

ODA – Official Development Assistance.

Open-source packages – Publicly available libraries containing plug and play code. There are currently over 137 000 Python libraries with a vast variety of use cases and functionality that can be accessed and used free of charge. For additional information, see https://www.pypi.org

OECD/DAC – Organisation for Economic Co-operation Development and the Development Assistance Committee.

OECD/DAC evaluation criteria – OECD/DAC evaluation criteria are a set of standards for structuring and designing an evaluation within the field of international development cooperation. For details, see https://www.oecd.org/dac/evaluation/daccriteriaforevaluatingdevelopmentassistance.htm

Project repository – The complete database of this study. This includes the full source code and all written excerpts in this report, as well as the developed web-based dashboard with descriptive statistics from the study results (see https://github.com/dav-consulting/eba-study).

QSS – Designed question-specific strategy composed of various data science techniques and natural language processing methods to establish automated and robust machine-based approaches for responding to the selected questions from the LME dataset.

Sida – The Swedish International Development Cooperation Agency.

1 Introduction

Recent technological developments have made vast quantities of digital texts and recordings available in a seemingly ever-increasing share of human communication and online interaction. For example, recent estimates suggest that the total amount of global data will grow from 60 to 175 zettabytes between 2020 and 2025 (Reinsel et al., 2020). Furthermore, it is also believed that 60% of the world's population has access to and is using the internet as of 2020 (ITU, 2020), and that the global mobile phone usage – based on subscription numbers – has surpassed the population of the world (ITU, 2018). These estimates hint that huge changes are coming, and the projections for these trends are that they will continue and – in many cases – pick up in pace. The digitalisation and amount of information generated in society today is also expected to grow exponentially into the foreseeable future (Kurzweil, 2004).

In order to take advantage of and efficiently process this vast amount of information, various computational or machine-based approaches are needed. The development of such approaches has been underway ever since the introduction of computers, and is increasingly being applied in all areas of daily life where it is believed to be capable of freeing up human labour from tedious and repetitive tasks. This has also led to the emergence of a new profession often referred to as data science, which captures the broad skill set involved in gathering, processing, extracting value, visualising and communicating information from data (Varian, 2009). This often involves building complex quantitative algorithms and models to organise and synthesise large amounts of information, a process which typically relies on state-of-the-art research in computer science, such as the fields of artificial intelligence (AI) or machine learning (ML). In particular, since much of this data comes in the form of unstructured text, data scientists also make frequent use of methods developed in the field of computational linguistics or natural language processing (NLP) to derive metadata for their analyses.

This development has also had a major impact from a social science perspective. In particular, the information encoded in text has turned out to be a rich complement to the more structured kinds of data traditionally used in research, and recent years have seen an explosion of research using text as data in areas such as political science and economics (Gentzkow et al., 2019). Likewise, the described development is believed to have a direct effect on the practice of evaluations, not the least due to its heavy reliance on data to conduct analysis and retrieve insights. Here, calls have been raised relating to the need for a broadened analytical toolbox for evaluations if increases in volume, velocity and variety of data are to be dealt with and taken full advantage of (see Petersson et al., 2017). This is believed to be the case within the field of international development cooperation in particular, where evaluations are an important instrument for retrieving insights and for fostering accountability and sound governance.

The purpose of the present study is therefore to explore the potential of data science and NLP methods in order to assess how these methods can be applied and used to derive metadata for a systematic review[1] for evaluations within the field of international development cooperation. This entails an assessment of the performance of these machine-based approaches in relation to traditional evaluation methods that rely on manual labour. More specifically, we aim to i) apply a selection of these methods and test how they can be used to derive secondary (meta) data for a systematic review of what past evaluations have concluded about aid projects and programmes; and ii) as part of the process, evaluate what the strengths and weaknesses of these methods are compared to standard approaches relying on manual labour.

[1] This study's use of the term *systematic review* refers to the application of computational techniques for the replication and automation of collecting, analysing and synthesising larger volumes of secondary data. This study's use of this term should thus not be confused with systematic reviews that aim to scrutinise aspects such as the quality of the study design, data sources used, etc.

To address these questions, the study relies on a wide range of methods with a heavy emphasis on computational linguistics, or so-called NLP. This field has experienced rapid development during recent years, and has now proven to have the ability to reliably perform a variety of labour-intensive and analytical tasks, ranging from document summarisation to text classification at scale. It may thus be expected that these methods could also bring new insights into the field of international development cooperation by assessing what past evaluations have concluded regarding a number of indicators of relevance for steering future aid projects and programmes. The advantage, in comparison to manual assessments, would be that in addition to generating interesting descriptive statistics and a quick analytical turnaround, these methods would also give us better information on the reliability of the statistics produced. By this, we mean that since the methods are computational in nature, they may also be typically more stable in their predictions and errors over time than human beings, who may vary more unpredictably in their assessments depending on e.g., time of day, overall mood or hours of sleep (Ng, 2020). More specifically, once a computational methodology has been implemented, a manual assessment of the results can give us an indication of the size of the margin of error, as well as insights into how it might be decreased. Since both humans and computer algorithms are likely to produce errors at some stage, having an understanding of the size and type of error is thus important. Another major advantage, once a reliable machine-based approach has been developed, is that the scale (i.e. the number of evaluations included in the meta-analysis) is of little importance for the overall effort required to conduct the analysis.

In order to make progress on using these methods, we use a dataset provided by EBA featuring a manually annotated or so-called labelled dataset. The labelled dataset (henceforth referred to as the

LME dataset) comes from a meta study[2] covering the content and conclusions from 128 decentralised evaluations commissioned by Sida between 2012 and 2014. These evaluations form a central instrument in the follow-up of Swedish projects and programmes within the realm of international development cooperation. The LME dataset thus constitutes the part of the abovementioned EBA meta study that subjects each of the 128 evaluations to a battery of questions[3] covering areas such as geography, funding, thematic area, project sustainability, etc. More specifically, the questions are directed at certain aspects of the decentralised evaluations, which have been deemed to be relevant in terms of what to incorporate in this meta evaluation. The aim is to generate an overarching assessment of what the evaluations include, and to a certain extent what they conclude, in an attempt to bring value and inform interested stakeholders and policymakers of progress made or lack thereof.

Based on the questions in the LME dataset, we have singled out a subset of questions where our initial hypothesis was that a strategy based on data science and NLP methods might have success in terms of its ability to generate reliable answers to the questions in the LME dataset. The analytical framework encompassing the selected questions (Appendix 1) thus provides a coherent structure for assessing the potential in a variety of computational approaches that we call question-specific strategies, where the output from each such strategy could be compared against the manual assessments made in the LME dataset.

[2] Livslängd och livskraft: Vad säger utvärderingar om svenska biståndsinsatsers hållbarhet? (Burman, 2017).

[3] This type of labelled dataset is often used in supervised learning (an area of machine learning), where algorithms are fed with predefined answers to questions from which it can learn to make informed guesses to previously unseen questions.

The study and the methods used in it are, however, by no means intended to be exhaustive, and the study should certainly not be viewed as a review of the field of NLP, but rather as case-based study in which various machine-based methods with the ability to scale have been tested and applied to questions of relevance to international development cooperation. It should also be mentioned that it is the evaluator's language per se that is analysed in this applied science study. In particular, this should not be confused with an assessment or appraisal of the performance of the actual projects/programmes that have been subject to the evaluations conducted.

The study is structured as follows. Section 3 will elaborate on the study's overall work approach and detailed steps, as well as explaining and defining the specific methods and benchmarks used throughout the study. This includes the analytical methods as well as the conventions on how to measure progress and the quality of the work conducted. Section 4 will present the designed question-specific strategies for each selected question, as well as showing the results/predictions from an exercise in which the designed strategies are applied to all available Sida evaluations between 2012 and 2020. This section is furthermore divided into subsections based on the selected question's thematic focus: Data collection and parsing of documents; Geography and time; Funding and donors; Thematic area; and OECD/DAC evaluation criteria. Section 5 focuses on observed strengths and weaknesses in the designed and applied strategies and concludes with a broader discussion on future research questions and responses to the study's research questions. The final section contains the concluding remarks drawn from this study.

2 Background and literature

NLP methods, developed within the field of computational linguistics, have grown increasingly popular during recent years due to their applicability to a variety of labour-intensive and analytical tasks ranging from document summarisation to sentiment classification. These methods are also being applied to other areas of research. For example, within the field of finance, text from financial news, social media and company filings have been used to predict asset price movements and study the causal impact of new information (Tetlock, 2007).

In macroeconomics, unstructured text has been used to forecast variation in inflation and unemployment, and to estimate the effects of policy uncertainty (Scott and Varian, 2015). In media economics, similar texts from news and social media have been used to study the drivers and effects of political slant (Gentzkow and Shapiro, 2010). NLP methods have also been used in text analysis, using speech as a metric of differences in partisan language between groups. For example, Lauderdale and Herzog (2016) used these methods to quantify political polarisation by extracting features from speeches given in the US Senate from 1995 to 2014, finding that party differences in speech have increased faster than party differences in roll-call voting. Likewise, partisanship has recently been measured using the predictive accuracy of several machine-learning algorithms, resulting in similar conclusions with respect to increasing polarisation (see Peterson and Spirling, 2018 and Gentzkow, Shapiro and Taddy, 2019).

A recent state-of-the-art review of the existing and future applications for economic and political research can be found in Gentzkow et al. (2019). These methods have also been used to conduct systematic reviews. The idea here is that technologies and methods for NLP have the potential to speed up the production of systematic reviews by reducing the amount of manual labour needed and hence partially automating the process. Marshall and Wallace

(2019) provide an overview of current machine learning methods that have been proposed to expedite evidence synthesis, including their strengths and weaknesses, and how a systematic review team might go about using them in practice. They conclude that research into machine learning for systematic reviews has begun to mature, but many barriers to its practical use remain and systematic reviews require very high accuracy in their methods, which may be difficult for automation to attain. Further, in areas with a high degree of subjectivity, it is also pointed out that readers are more likely to be reassured by the subjective but considered opinion of an expert human versus a machine.

The field of international development cooperation also has its fair share of cases where new ways to conduct research, meta evaluations and systematic reviews are being tested and explored. OECD/DAC presented a working paper in 2019 that outlines how both the OECD and the World Bank are applying machine learning to a range of areas, such as topic modelling for the classification of reports, tracking migration flows, and applying poverty prediction models. The working paper focused on using unsupervised machine learning to predict how international donors target the sustainable development goals (SDGs) with their projects (Pincet et al., 2019).

Another central actor in this field is the UN Global Pulse, a UN initiative that works with and supports projects with a focus on big data and artificial intelligence for development, humanitarian action and peace. Projects such as making Ugandan community radio machines readable using speech recognition – a collaboration project between UN Global Pulse, Makerere University and Stellenbosch University – are good examples of novel methods for processing and utilising unstructured data. There are also recent studies that have looked into the possibility of bringing data science methods into the realm of evaluations of international development cooperation in order to improve quality and reduce the time and cost of evaluations (see York and Bamberger, 2020 or Petersson et al., 2017).

Other endeavours that have received attention use high tech installations for data collection and analysis, for example using remote sensing and satellite imagery to improve responses to humanitarian situations (Logar et al., 2020), or mobile network data as a way to inform policymaking, in this case using data on how people live and move when planning the construction of health facilities (Knippenberg et al., 2019). A clear momentum has been seen in recent years, and the desire to use such approaches seems to go hand-in-hand with improved performance in many of these technologies.

2.1 Requirements and general skill set needed

The requirements, and the general skill set, needed for a study of this sort can be boiled down to capacity in three separate fields: computer science and/or programming; mathematics and/or statistics; and contextual knowledge (Grus, 2017). The computer science and programming aspects relate to knowledge about computers and computational software systems. In this study, a wide range of open-source packages have been used. The value of these packages cannot be understated in the development of strategies similar to the ones applied in this study. Many of the packages used have taken many years for large teams to develop. The open-source packages and their utility range from basic data environmental support to data management and computational support, as well as visualisation of output. Database knowledge is also something that is required in order to store and run these kinds of systems. Requirements relating to mathematics and statistics are mostly tied to linear algebra, probability theory and inferential statistics. A final but important knowledge-based requirement is that of contextual knowledge, which boils down to intuition of the context at hand. Such an understanding of the fabric behind the numbers and patterns produced by the designed methods is very useful. In this study, this

translates into context knowledge and an understanding of international development cooperation in general and the practice of evaluations in particular.

On top of the abovementioned knowledge-based requirements, it is also important to emphasise that a machine-based approach – as with most analytical approaches – requires time and dedication to be optimised in order to achieve its full potential. Rule-based approaches and pre-trained models take less time to set up and can be deployed without much preparation or pre-processing. In many cases, a basic analytical structure can be set up and produce robust analytical predictions in a matter of hours. However, in more complex cases where a specific model needs to be trained to obtain satisfactory results, more resources are needed and it is not unusual for entire teams to spend years optimising specific models. A central aspect that requires more resources when training models entails access to labelled training data (a source of truth that is discussed further in the Methodology section). Training data can be obtained either through secondary sources or through the process of natural language annotation. This process refers to the generation of metadata or descriptive labels for underlying observations (or texts in the case of this study), and the purpose is to augment an algorithm's capability to give accurate predictions (Pustejovsky and Stubbs, 2012). A final requirement that should not be forgotten is the importance of a constructive dialogue between the designer and the intended users of a machine-based approach. Understanding the end-users' needs and intended use cases is crucial for establishing a practical and feasible machine-based approach.

3 Methodology

This chapter elaborates on the overall work approach and the analytical steps taken in this study. The ambition is to enable a high degree of transparency that can allow for a good intuition for the analytical steps taken. A project repository containing all the code and documentation is also made available, allowing for full reproducibility of the study results. This chapter also introduces the computational techniques that have been applied, as well as the conventions for how to evaluate the performance of the analytical results.

3.1 Natural language processing

Natural language processing (or NLP) is a collective phrase or catchall term for general approaches to "processing" natural or human language. In practice, NLP involves the use of a wide variety of computational algorithms and techniques, which allow us to identify linguistic rules, uncover the structure of a text and extract meaning, for instance. Common tasks to which such algorithms are applied include areas such as text classification, text similarity, text summarisation and keyword extraction.

As humans, we process language pretty well, but we are not perfect. Misunderstandings are relatively common among humans, and we often interpret the same texts or language differently. In other words, language processing is not deterministic, and something that might be interpreted in one way by one person may have a different meaning to another. A common example where this occurs frequently is when irony is used in texts or speech.

This inherent non-deterministic nature of language processing makes it an interesting and difficult problem to develop machine-based algorithms for. In this sense, understanding language is like creating a new form of intelligence in an artificial manner that can

understand how humans process language, which is also why NLP is a subfield of artificial intelligence. Importantly, if humans do not agree fully on NLP tasks, such as text classification or language translation, it is not generally possible to model an algorithm to perform these tasks without some degree of error. In machine learning, this peer-to-peer human understanding on a given subject or question is commonly referred to as inter-annotator agreement. A typical example in NLP where the level of inter-annotator agreement tends to be large is the problem of text classification, i.e. recognising which category a specific text belongs to (for example, whether a novel should be categorised as a thriller or a drama). In general, the level of inter-annotator agreement tends to form an upper boundary or benchmark for what to expect in terms of performance from a machine-based approach to a specific task (see e.g. Artstein, 2017 or Bobicev and Sokolova, 2017).

The methods applied in the field of NLP are often separated into two different sets of approaches. One relies on a hand-crafted set of rules, and the other on statistical or machine-learning techniques. In practice, however, NLP typically comprises a combination of these two approaches, where parts of both approaches are used with the intention of finding a potent mix that can optimise the level of accuracy. It is this mixed-methods approach that we have taken for most part in this study, where statistical models have been used for some parts and rule-based techniques for others. Below, we provide a brief overview of these approaches and some of their most central methods and features.

3.1.1 Rule-based methods

Rule-based systems are the earliest approach to NLP and consist of hand-crafted linguistic rules for text analysis. Each rule is formed by an antecedent and a prediction. So, when the system finds a matching pattern, it applies the predicted criteria. Since the rules are determined by humans, this type of system is easy to understand and

can sometimes provide accurate results with little effort. However, manually crafting and enhancing rules can be a difficult and cumbersome task, and often requires a linguist or a knowledgeable engineer with deep knowledge of the intrinsic details of the domain being analysed. Also, adding too many rules can lead to complex systems with contradictory rules. The rule-based approach has been chosen in this study for cases/questions where the task has been relatively straightforward, such as identifying unique text passages, assessing word frequency or extracting keywords from documents. An example of such an application in this study is an algorithm known as Tf-Idf (term frequency – inverse document frequency), which is typically used to extract keywords from a text. In a nutshell, the algorithm counts the frequency of occurrence of each word in a document, and then weighs these frequencies based on how common they are in other documents within the same corpus.[4]

Although the analysis that can be done with these types of methods is limited, a major advantage is that they do not require any labelled training data or cumbersome model estimation, which is typically the case when statistical or machine learning methods are applied. As a result, rule-based methods are a good option if you do not have much data and are just starting out on an analysis and need to conduct an exploratory analysis of the language used.

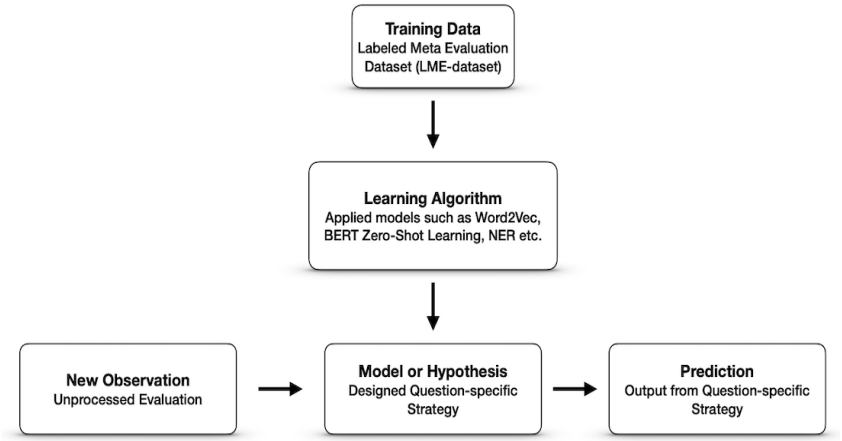
3.1.2 Statistical/machine learning-based methods

By statistical approaches and machine learning methods we mean algorithms which have been crafted to learn and make predictions based on previous observations. This process is typically referred to as the training of a model or establishing a hypothesis. In other words, this is the process where the model learns to make associations between a particular input and its corresponding output.

[4] This method was for instance applied in the first of our approaches to question 17 of the analytical framework.

In Figure 1 below, we have depicted a generic model that has been applied in this study. In most cases, the applied models or hypotheses have been derived from pre-trained models, but we have also trained models of our own for a few of the questions in this study. In short, and as depicted by the figure below, the vertical process constitutes the training process where labelled training data is fed into a learning algorithm. In a labelled dataset, the labels are typically defined manually, and simply constitute a question/answer sheet from which the algorithm can learn the correct answers to a predefined set of questions. The algorithm then calibrates a model by seeking out a set of parameters such that the model produces the best guesses of the correct answers to the questions in the labelled dataset. Translated to this study, the LME dataset has been used to train various algorithms. Once the model has been trained, the horizontal process or prediction of new or unprocessed data can be executed. In the case of this study, new evaluations or identified text paragraphs of relevance have been fed into the trained algorithms, which have made predictions based on this training.

Figure 1: Generic representation of statistical/machine learning model



The displayed learning process is the biggest advantage of these methods, since their ability to learn on their own implies that there is no need to define manual rules. What is, however, needed is accurate training data, which constitutes the foundation and the relationships upon which we want the model to learn. Machine learning models typically perform better than rule-based systems over time, and the more training data they are fed, the more accurate they often become. However, the algorithms are typically data hungry in the sense that they need enough training data that is relevant to the specific problem to be solved in order to produce an accurate model.

Training data can often be difficult to acquire, and usually involves many hours of manual work labelling the data where people with expert domain knowledge are usually needed to ensure good quality. This is the case when training models to predict aspects such as a word's part-of-speech tag or the linguistic relationships between words within a sentence. Another example where this is crucial is within the task commonly referred to as named-entity recognition (NER). This includes the process of finding specific types of entities within a text, where an entity can be a word or a series of words with a bearing on, for example, a personal name, an organisation, a location or a product, as well as date-related expressions, money and more.[5] NER was introduced in the mid-1990s in an attempt to find solutions for extracting data on entities within the field of information extraction (Nadeau and Sekin, 2007).

Text embeddings

A central concept in NLP research where machine learning approaches have yielded much success is that of text embeddings. Text embeddings is a process where words or phrases from a vocabulary are mapped to vectors of real numbers. These numerical

[5] This method was for instance applied in questions 3–5 in the analytical framework.

vectors thus become valued representations of text strings, where the numbers in the vectors are chosen so that vectors lying close to each other in a vector space represent text strings that appear in similar contexts in documents.

Text embeddings are considered a good starting point for many complex NLP tasks. They allow deep learning to be effective on smaller datasets, and are one of the most popular ways of doing transfer learning in NLP, where knowledge from training on one problem is transferred to another. A clear advantage with these algorithms is that they do not require manually labelled training data, but instead rely on large volumes of text from common data sources, such as Wikipedia, in their training process. The labelling (of the training data) in this case is extracted from pre-existing relationships in the way sentences and words have been spelled out in relation to each other.

One of the most popular text embedding algorithms in recent years has been Word2vec, which was created and published in 2013 by a team of researchers led by a group of scientists at Google (Mikolov et al., 2013). An interesting revelation when this research emerged was that the word vectors produced by the algorithm could actually be used to mathematically solve word riddles. One of the most noteworthy was the riddle: "King - man + women = ?". Replacing these words with their mathematical representations (i.e. text embeddings) produced by the Word2vec algorithm resulted in an output vector which was most close in the vector space to the numerical vector for the word queen, which many viewed as the most logical answer to the above riddle.

In this study, pre-trained embedding models have been frequently applied, and context-specific Word2vec embeddings were also evaluated for a few questions. In these cases, the Word2vec embeddings were trained using a training corpus consisting of 311 evaluations downloaded from Sida's website. The idea here was that the algorithm would learn the common language often used in evaluations. However, for most cases these text embedding

techniques have recently been surpassed by a new class of model algorithms often referred to as transformers, which have also been used and applied in this study.

Contextual text embeddings

A major challenge with Word2Vec is that it provides a single representation for a word that is the same regardless of its context. This means that words like "bank" that are used in several different senses, for example river bank and investment bank, will end up with a representation that is an average of the senses and thus does not represent either of the two particularly well.

For this reason, subsequent research focused on the idea of training separate language models to produce better contextual word representations. This has led to the development of the abovementioned transformer networks. The transformer is essentially a deep learning model proposed in a paper by researchers at Google and the University of Toronto in 2017, and used primarily in the field of NLP (see Vaswani et al., 2017). Since their introduction, transformers have become the model of choice for tackling many problems in NLP, in particular due to their capacity to differentiate between words based on their context.

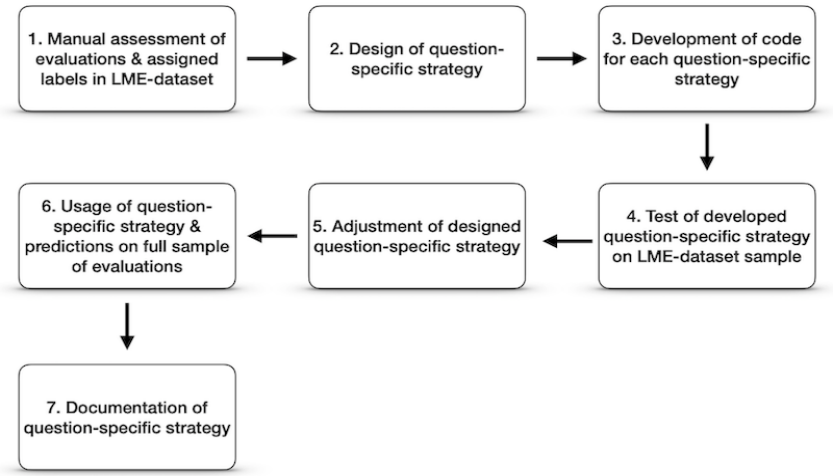
This enhancement has been something of a revolution, and led to the development of a wide variety of pre-trained systems all building off the abilities made possible by transformer networks. Examples include the BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) models, both of which have been trained on very large language datasets, and can be further fine-tuned to specific language tasks. These models are typically available to the general public as open-source code which can be readily downloaded through various channels.

In this study, we have relied on these transformer networks to solve specific challenges. The models have been accessed via an open-source package called Hugging Face's transformers. This package provides thousands of pre-trained models to perform tasks on texts, such as classification, information extraction, question answering, summarisation, translation, text generation, etc. in 100+ languages, with the aim of contributing to the public good and making cutting-edge NLP easier to use for anyone with an internet connection.

3.2 Detailed walkthrough of study processes

For each selected question phrased in the LME dataset (see Appendix 1 for the questions), different approaches have been designed, developed and practically tested in what we will refer to as the question-specific strategies (henceforth the QSSs) of this study. Each designed strategy is based on what was deemed the most suitable path to take in order to ensure a solid approach that would yield good results. In addition, and depending on the estimated accuracy of the designed strategies, two possible applications have been executed in this study. First, strategies estimated to have high performance, and thus to correlate well with the results from LME dataset, have been used in a scaled-up exercise where the designed strategy has been applied to the full set of available evaluations between 2012 and 2020. Second, strategies that perform poorly and thus have low correlation with the LME dataset have not been deemed to be suitable for generating predictions on unprocessed evaluations. Instead, the challenges and flaws of the poor strategies are discussed together with possibilities to improve the strategies (on a theoretical level). Figure 2 below outlines the generic steps taken in the development of all the QSSs in this study.

Figure 2: Generic process for development of question-specific strategies



Specific details of the steps taken in the development of all the QSSs in this study are outlined below:

First, a manual assessment of a random sample of evaluations processed in the LME dataset was carried out to get an idea of which methods may be useful to apply. In general, each question required a unique focus in order to grasp how to design and later deploy each of the developed QSSs. An important part of this step was to review and use the LME dataset as a point of departure for how to design each strategy.

Second, for each question a choice was made on the most appropriate design to use to address it. This included the following sub-steps, searching for methods that had been successfully applied to similar questions in the past as well as looking for available open-source packages that could be used.

Third, the design and coding of each strategy was implemented in order to allow for automated processing of multiple evaluations.

The fourth step involved testing the developed code by applying it to a sample of evaluations from the LME dataset. This step also

included performance testing, where the results of the developed strategy/algorithm were compared with the labels provided in the LME dataset – the applied tests at this stage are carefully described in the subsection *Evaluating performance* below. A manual assessment of these results was then done in order to follow-up on any significant discrepancies between the output produced by the algorithm and the labels provided in the LME dataset.

Fifth, based on details and insights from the manual assessment, the strategy was adjusted in order to improve its performance. The majority of designed strategies also required a number of iterations of steps 4 and 5 in order to arrive at a final strategy where we did not see any immediate options for quick improvements.

The sixth step involved a decision between two possible paths forward based on the strategy's overall performance. Strategies with good performance, in the sense that the developed strategy produced results which aligned well with the LME dataset labels, were used to extrapolate the analysis and include all evaluations from 2012 onwards (>300).

For well performing strategies, the following steps were then executed:

a. Scale-up of process and deployment of the QSS on the full sample of evaluations.
b. Descriptive statistical analysis and visualisation of results. The recorded estimations were compiled in a dashboard for easy access for the whole assessed period.
c. Comparative analysis between the LME dataset, the results/output from the designed strategies and estimations from a third party/independent validation assessment based on a random sample of 30 evaluations (see below for further details).

Strategies with lower accuracy – and thus not deemed to have enough potential to generate reasonable estimations – followed the following steps:

a. Thorough elaboration of the challenges and problems with securing high enough accuracy.
b. Additional research and theoretical discussion on possibilities to improve the strategy.
c. Comparative analysis between the LME dataset, the results/output from the designed strategies and estimations from a third party/independent validation assessment based on a random sample of 30 evaluations (see below for further details).

Finally, the last step involved documenting the full process for each QSS. All strategies have been thoroughly documented with the applied method(s), results and caveats, and are readily available in the project repository.

3.3 Evaluating performance

As described above, each QSS involved a test of various methods and techniques. In order to evaluate their performance, a benchmark or baseline for comparison needed to be established. As mentioned, one such baseline for comparison has been the labelled data from the LME dataset. The designed strategies' predictions have thus been compared to the manually labelled data in the mentioned dataset for each question. The idea is that when designing algorithms one can use these manually crafted labels as a source of truth when evaluating the performance of the algorithm. However, other potential benchmarks are also possible. Below, we describe our performance benchmarks in greater detail.

Performance metric

Several metrics exist for evaluating the performance of machine learning models. In this report, we will adopt one of the most common metrics called "accuracy". The accuracy of a model or

algorithm relates to the relative proportion of labels in the dataset that were correctly predicted. For example, imagine we have a dataset consisting of 100 sports referees from tennis and hockey, of which 30 have been labelled as hockey and the remaining as tennis. If our model, which is designed to predict the correct labels, manages to label 20 of the 30 hockey referees as hockey and 40 of the 70 tennis referees as tennis, then the accuracy of its predictions would be calculated as 60 percent - ((20+40)/100)).[6]

Evaluating accuracy scores

When comparing accuracy scores of different classification problems, one must also account for the difficulty of the classification problem. Clearly, achieving high accuracy scores is easier when predicting the correct outcome in a classification problem involving only two labels, compared to a classification problem where ten possible labels exist. For this reason, we may also want to evaluate our model based on the difficulty of the classification problem. This can be done in various ways. The most straightforward approach may be to calculate the expected outcome of a random choice where equal probabilities are given to each label. For example, if only two labels exist there is a 50-50 chance that we would guess the correct outcome. For a case with three labels, the corresponding probability is 33 percent, and so on. With this information we can at least judge whether an algorithm or model is performing worse than if labels were simply chosen randomly.

[6] For the sake of transparency and in order to make our results easily digestible for an audience that is unfamiliar with machine learning methodology, we have relied on accuracy as our sole metric of performance in this study. It is, however, important to note that accuracy is not the only possible evaluation metric. Other notable examples include precision, recall and f-score (see e.g. Grus (2019) for further details of these metrics).

Frequency-adjusted random accuracy scores

The abovementioned type of comparison can work well in many cases, but there are some cases where it may be misleading. For example, if we were training a machine learning model to perform a classification exercise, where all the input data has been replaced with nonsensical data, the training exercise may start to make guesses based on the frequency of label occurrence instead of producing a model which tries to make sense of the input data. In other words, the model learns the empirical frequency of a label in the training dataset and makes guesses based on what it expects the probability of occurrence to be. To account for this, one could instead evaluate a model's performance based on a probability-adjusted random guess, which takes into account the empirical distribution of the labels. For example, in the sports referee example above, if we were to choose a referee at random, the probability of picking a hockey referee would be 30%. Hence, if we picked out a random sample of ten referees and counted the number of tennis referees versus hockey referees, we would on average find three hockey referees and seven tennis referees if we repeated this experiment enough times. The machine learning model might thus learn that the distribution of labels in the training data is skewed in this way, and make use of this information to inform its predictions. If this is the case, this may imply that we wrongfully conclude that the model has learned to predict labels well based on the underlying training data when in fact it has only learned how many of each label exist in the training dataset. Hence, if machine learning models were only evaluated based on how much better they perform than a purely random guess, this would favour models that were trained on datasets with highly skewed labels. For this reason, a probability-adjusted random choice based on empirical frequencies of the labels in the training dataset is often a more appropriate benchmark comparison when evaluating machine learning models.

Third party validation

Another benchmark used in this study aims to compare our results with those of an independent assessment by a third-party evaluator.[7] The purpose of this step was threefold. First, it allowed us to evaluate the performance of an extended set of questions with regard to other OECD/DAC evaluation criteria apart from sustainability.[8] Second, it allowed us to compare the designed strategies' results to a second *source of truth*. Third, as mentioned above, human-based assessments are typically not perfect and the third-party assessment could thus give us an idea of what the upper bounds for accuracy scores might be in accordance with an inter-annotator agreement of sorts. That is, if for example the third party agreed with the LME dataset labels around 90% of the time for a specific question, we should not expect any better performance from our machine-based approach.

Presentation of findings

In this study we have calculated all of the abovementioned accuracy scores for all selected questions in order to shed light on the designed strategies' estimated performance. The scores are reported in tables in the upcoming results sections, together with our initial assessment of the difficulty and confidence in developing a good strategy. The table columns are defined as follows:

[7] The third-party evaluator Cecilia Ljungman was chosen in dialogue with EBA, and has 25 years of experience working in the field of international development cooperation in general and with evaluations in particular.

[8] A key question in the LME dataset focused on the OECD/DAC criteria sustainability. The concept of this criteria and the attempts to design a machine-based approach that can automate parts, or the whole process, of assessing the evaluator's judgments and conclusions relating to this criterion is therefore of central importance in this study.

- **QSS accuracy:** Comparison of predictions by designed question-specific strategies (QSSs) against the manually assessed LME dataset labels.
- **Label counts:** The number of plausible labels for answering a specific question, i.e. the number of potential answers to a question.[9]
- **Random adj. accuracy:** Theoretical accuracy scores of a random guess with probabilities for each outcome adjusted to match the empirical frequencies of the LME dataset labels.
- **Third party accuracy:** Accuracy scores for predictions from designed QSSs against the third-party validation assessment.
- **LME vs. third party:** Comparison of third-party validation assessment with that of LME dataset. This comparison is based on 15 evaluations and computed in the same way as a standard accuracy score.
- **Anticipated difficulty:** Our initial pre-study assessment of the difficulty of finding a good strategy.[10]
- **Assessed difficulty:** Our ex-post assessment of how difficult it was to find a good strategy.

[9] For example, a question that can only be answered with a *yes* or a *no* has two labels, while a *yes*, *no* and *maybe* question implies three labels.

[10] The scale for difficulty levels includes Very low, Low, Moderate, High and Very high. Our anticipated confidence level of success was also assessed ex ante as Highly confident, Confident, Fairly confident or Unconfident. These judgments are provided in the appendix.

4 Results

This section presents the results of the question-specific strategies (QSSs) for each selected question targeted in this study (see Appendix 1 for details). The section is structured according to the thematic focus of the selected questions and describes the steps taken in each strategy, its individual performance and the descriptive statistics from the scaled-up exercise for the QSS when deemed accurate enough. In those cases where the QSS was concluded to perform less well, the observed flaws, challenges and potential solutions are discussed.

4.1 Data collection and parsing of documents

This section outlines all the processes involved in the collection, pre-processing and organisation of the data in this study. All these steps are vital parts in all NLP studies. Finally, some descriptive statistics are showcased to illustrate how unstructured data can be synthesised and used to provide insights with regard to the current context.

4.1.1 Findings

Accuracy of designed strategies

In total, 318 of Sida's decentralised evaluations that covered the period 2012–2020 were downloaded in this study. After removing some deviating[11] cases, 311 documents remained and were included

[11] A few deviations have been observed among the downloaded evaluations. A few turned out to belong to earlier years (2009 and 2010). We also found documents of the "Sida Review" type which had been wrongly labelled as a "Sida decentralized evaluation". There were also examples where the evaluation per se

in the scaled-up analysis (as mentioned in the Methodology chapter above). Of these evaluations, we were able to identify a table of contents in 309 of the evaluations (99%). Given that this was such a common feature, we crafted a specific method for parsing the table of contents, which gave us an overview of the most frequent paragraphs in the downloaded evaluations (see Table 1 for details), and then used the page numbers as an index for how they could be found systematically.

Many of these paragraphs were particularly important to be able to answer certain questions in the analytical framework that targeted specific evaluation paragraphs. This included questions with a bearing on the executive summary, recommendations and terms of reference, as well as sections addressing specific OECD/DAC criteria. The table below presents summary statistics with regard to the frequencies with which these sections were found among the 309 documents parsed.

Table 1: Identified paragraphs in assessed evaluations between 2012 and 2020

Section	Count	Percent
Table of contents	309	99%
Executive summary	302	97%
Recommendations	283	91%
OECD/DAC – sustainability	283	91%
Terms of reference	262	84%
OECD/DAC – relevance	209	67%
OECD/DAC – effectiveness	197	63%
OECD/DAC – efficiency	163	52%
OECD/DAC – impacts	123	40%

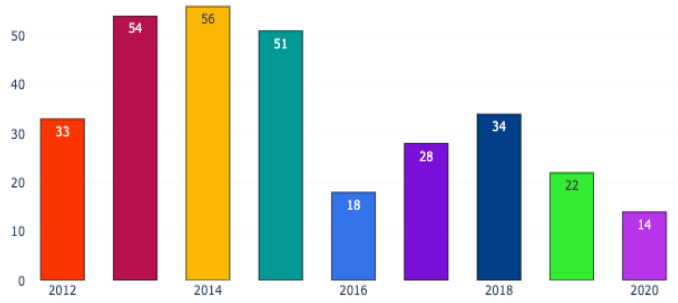
had the wrong content – the front page and the table of contents seemed to be accurate, but the actual document was a consultancy tender.

The parsing process also allowed us to identify document-specific characteristics such as title, authors, commissioning agency, publication date, series number, article number and publisher. This data was extracted with what we believe to be one hundred percent accuracy (no errors were found). However, these processes revealed some discrepancies between the data available on Sida's website and the data in the actual published evaluations. A plausible explanation for these discrepancies is believed to be linked to human error occurring at the time when the evaluations were uploaded to Sida's publication database and evaluation details were transcribed.

Descriptive statistics from scaled-up analysis

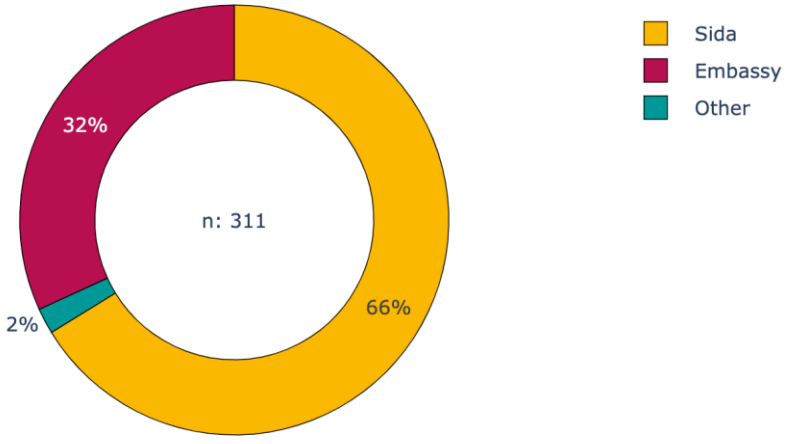
Even at this stage, the parsed data allowed us to derive some descriptive statistics for the full set of evaluations between 2012 and 2020. Figure 3 below shows the number of evaluations conducted for each individual year during the relevant period. There seems to have been a general decline in commissioned decentralised evaluations in recent years, compared to a few relatively busy years between 2013 and 2015.

Figure 3: Number of evaluations in Sida's database between 2012 and 2020



An assessment of the commissioning agency for the evaluations shows that two thirds (66%) were commissioned by Sida HQ, and almost 32% were commissioned by Swedish embassies, as shown in Figure 4.

Figure 4: Estimation of commissioning agency between 2012 and 2020



Another example of statistics that can be directly produced at this stage, and which are particularly important to consider (since it is the language of the evaluations that is being analysed), is the number of evaluators or authors involved. This can give an estimate of the diversity of language used in the processed evaluations. Using the collected data to assess this angle revealed roughly 550 unique names among the list of authors in the 309 processed evaluations. This suggests that there is a large number of individuals involved in drafting evaluations. A closer look showed that five percent of the authors are appear relatively often, i.e. they have been involved in six or more evaluations. The most frequent evaluator had taken part in no fewer than 53 evaluations, constituting 17% of the complete set of evaluations in our study.

4.1.2 Applied methods

A key component and necessary condition for this study to succeed was the design of the method that could extract relevant evaluations from the internet, and more importantly to parse content of key importance within these evaluations. The lion's share of the data underlying the analysis in this study was embedded in PDF documents available via Sida's publication database.[12] Given that the scope of our study stretched between 2012 and 2020, during which period several hundred evaluations had been produced and published, we decided to use an open-source package for web scraping named Scrapy. After having customised this package in line with the task at hand, we were able to systematically extract all available evaluations during the abovementioned period in a couple of minutes. This is a process which can easily be repeated and/or scheduled once new evaluations are published, should these processes need to be repeated and or the results updated.

[12] https://www.sida.se/English/publications/publicationsearch/

After collecting the evaluations and storing the documents, the text they contained needed to be extracted and converted into a format that allowed for the deployment of the designed strategies (as described in the Methodology chapter above). The open-source package PyMuPDF was used for this task, and all the collected evaluations were converted from PDF documents into JSON formatted documents. In this process, each row of the evaluations was labelled with metadata containing auxiliary information about the text, such as font size, text colour, position on page, etc.

This auxiliary information was of particular importance since it allowed for rule-based algorithms to identify specific sections within each of the evaluations, which was a crucial component when answering some of the selected questions in this study. For instance, the font size of a specific paragraph was important for parsing out specific sections within the evaluations. These algorithms made use of handcrafted rules that were based on document attributes such as the fact that document headers in the evaluations usually had a larger font size than the body text. Similarly, the most common font size in the evaluations allowed us to identify the body text of evaluations. Another aspect used in this approach was the position of text in the evaluations. In particular, page numbers and footnotes are typically located at the bottom of documents and numbered sequentially. Using knowledge of these types of characteristics of the texts allowed us to derive additional rules for how to identify relevant text passages. In short, the success of this parsing algorithm allowed us to parse and extract specific sections from the evaluations, which was a prerequisite for responding to several of the selected questions in the LME dataset.

4.2 Geography and time

This section presents the results for selected questions with a bearing on geographical and time-related issues. More specifically, we cover countries and geographical areas, time periods that are being evaluated and at what phase of the contribution the evaluation took place. The specific questions that are addressed in this section are as follows:

- Q3. Country (include all countries that have been studied in the evaluation)?
- Q4. Geographical region?
- Q5. Geographical focus area (country/local; region; global)?
- Q6. Time period that is being evaluated?
- Q14. At what phase of the contribution is the evaluation being conducted?

4.2.1 Findings

Accuracy of designed strategies

At the outset of the study, estimates of the difficulty and confidence levels for finding successful strategies for each question were made. Compared against these estimates, all of the QSSs in this section performed relatively well. The most noteworthy strategies, in this light, are the ones dealing with geographical focus areas (question five), estimations of the time period (question six) and the type of evaluation (question fourteen). All these questions were believed to be difficult to find proper solutions to, and the initial expectations have been surpassed with the results from the developed QSSs. Table 2 below depicts performance estimates for all questions in line with the benchmarks described in section 2.3. As shown in the table below, the accuracy level ranges from 63 percent to 86 percent.

When the results from the developed strategies are compared with the third-party validation data, the correlation is lower. This is also the case when comparing the LME labels against the third-party validation, which makes it difficult to set an upper boundary for each QSS.[13] In fact, the QSS tends to correlate higher with the LME than the third-party validation data does, which implies that the manual assessments are less consistent than those of the machine-based approach and the LME dataset.

A likely explanation for the somewhat scattered correlations between the different comparisons may be the rigour used to determine the accuracy. The result from the QSS needs to match exactly with that of the benchmark dataset in order to register as equivalent. This can lead to reduced accuracy measures when the result actually consists of multiple observations, which for example was the case with question three where the answer consisted of a list of countries. For comparison, if we were to settle with comparing the most frequently extracted country name in each evaluation, the accuracy level between the developed QSS for question 3 and the LME dataset would increase to 97 percent. This suggests that there is higher correlation if parts of the predicted data are used for comparison rather than using all observations or countries. In this case, it might be more valuable to use the most observed countries since most evaluations usually focus on one or more countries. However, we have settled on applying a strict metric where only exact matches will register as a success in this study.

[13] In a few cases, the accuracy levels between the validation dataset and the LME dataset are based on few observations (<5 evaluations). This is an effect of the LME dataset's varying coverage – not all questions have 128 observations – and since the validation dataset is based on a random sample of the full sample (128 evaluations), the number of comparative observations varies between questions.

Table 2: Results and performance estimates for QSSs with a bearing on geography and time

Question	QSS accuracy	Label counts	Random adj. accuracy	Third party accuracy	LME vs. third party	Anticipated difficulty	Assessed difficulty
Q3	73%	-	-	54%	86%	Moderate	Moderate
Q4	79%	6	25%	55%	67%	Moderate	Low
Q5	86%	3	51%	79%	80%	High	Low
Q6	63%	-	-	57%	100%	High	High
Q14	76%	4	52%	72%	60%	High	Moderate

Notes:

QSS accuracy: Comparison of designed question-specific strategy's predictions against LME dataset labels.

Label counts: Number of plausible labels for answering a specific question.

Random adj. accuracy: Theoretical accuracy scores of a random guess with probabilities for each outcome adjusted to match the empirical frequencies in the LME dataset

Third party accuracy: Accuracy scores for predictions from QSS against the third-party validation assessment.

LME vs. third party: Comparison of third-party validation assessments with LME dataset.

Anticipated difficulty: Our initial pre-study/ex-ante assessment of the difficulty of finding a good strategy.

Assessed difficulty: Our ex-post assessment of how difficult it was to find a good strategy.

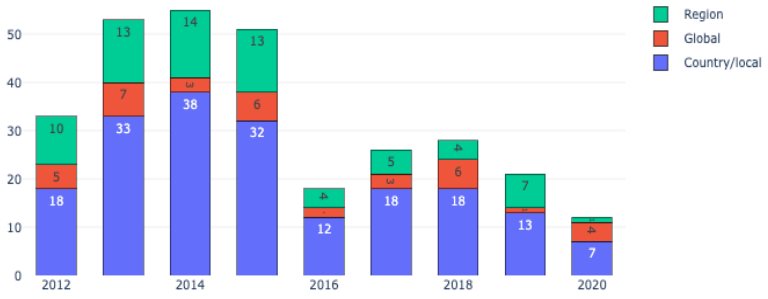
Descriptive statistics from scaled-up analysis

All of the QSSs in this section have been deemed accurate enough to be used in a scaled-up exercise where each designed strategy was used on the full dataset of >300 evaluations. The applied strategies are believed to give a fair overview for the assessed period (2012–2020), and have given some insights into the questions at hand. However, it is important to note that the numbers and figures below should be viewed in the light of the displayed accuracy levels for each specific strategy.

A total of 115 countries were recorded in the full dataset. The top ten ODA recipient countries – those mentioned most frequently – are as follows: Kenya (68 evaluations or 21.3%), Tanzania (60 or 18.8%), Uganda (54 or 16.9%), South Africa (37 or 11.6%), Turkey (34 or 10.7%), Serbia (31 or 9.7%), Bosnia and Herzegovina (31 or 9.7%), Rwanda (30 or 9.4%), Ethiopia (29 or 9.1%), Zambia (28 or 8.8%) and Georgia (28 or 8.8%). It is noteworthy that East African countries stand out and are highly overrepresented in the full dataset.

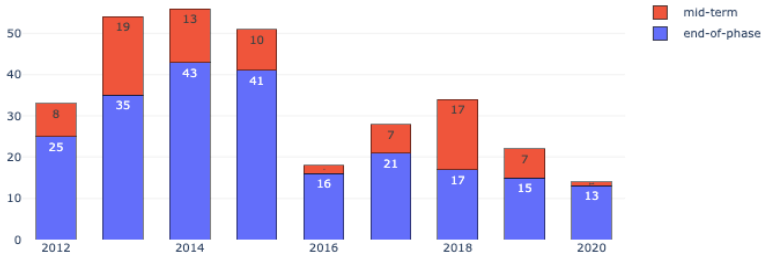
The results for geographical region (based on UN regions) gave an estimate that the top ten regions evaluated during the last ten years are as follows: Eastern Africa (137 or 43.3%), Western Asia (60 or 18.9%), Southern Europe (50 or 15.8%), Southern Asia (34 or 10.7%), South-Eastern Asia (32 or 10.1%), Western Africa (29 or 9.1%), Southern Africa (26 or 8.2%), Eastern Europe (21 or 6.6%), Northern Africa (20 or 6.3%) and South America (18 or 5.7%). For the final geographical question and the geographical focus area, the results for the full sample estimated that, in line with Figure 5 below, most evaluations focused on the country/local level (192 or 60.1%). The regional level (76 or 24%) was the second most common level, with the global (37 or 11.7%) level trailing. The chart below details the number of evaluations each year grouped by the estimated regional focus area (question five). The relative proportions between the three regional categories are relatively stable over the course of the assessed period.

Figure 5: Estimation of evaluations by geographical focus area between 2011 and 2020



The time-related questions also gave some interesting estimations. The average time period for the evaluated projects/programmes is 4.1 years, with a distribution ranging from less than a year up to 24 years. As shown in Figure 6, the most common evaluation type for the entire period is the end-of-phase evaluations, with 72.3% (230), and the remaining share of 27.6% (88) relates to mid-term evaluations. The chart below shows the distribution of estimations according to the type of evaluation over the whole assessed period. Years that deviate with considerably fewer mid-term reviews are 2016 (12%) and 2020 (7%).

Figure 6: Estimation of evaluation group by type between 2011 and 2020



4.2.2 Applied method(s)

A mixed-methods approach was used to develop the QSSs in this section. At an initial stage, all the QSSs relied on the text parsing mechanism described above. Identified text passages with content that had a bearing on geographical and time-related aspects were extracted and singled out for additional analysis.

All of the strategies for handling the geographical queries share a basic analytical structure, where relevant parsed text excerpts from the evaluations were processed using a three-pronged approach. First, the pre-trained spaCy model for named-entity recognition (NER) was utilised to extract geographical entities throughout each processed evaluation.

Second, the identified entities were normalised and cross-checked against manually established validation lists. A central example is the normalisation process for country names – variation in country names and misspellings were common in the underlying dataset, and needed to be sorted before any further analysis could be conducted. The ISO standard 3166 for country names was used for this, and a nomenclature for country names was established using the open-

source package pycountry. During this process, identified countries were also matched against a second validation list of countries that received ODA support from OECD countries. All positive matches with these queries were counted and recorded as positive observations for question three.

Third, for the two questions relating to geographical region (question four) and geographical focus area (question five), yet another matching exercise was conducted where recorded countries were matched against UN regions using the open-source package country-converter. The geographical focus area (question five) required an additional step that accounted for the number of recorded countries and their geographical spread to estimate the geographical focus area. The estimation of accuracy for all three of these strategies was determined using mapping exercises, where the strategy's output (for each processed evaluation) was compared against the corresponding data in the LME dataset.

Regarding the questions with a bearing on time-related issues, two separate rule-based QSSs were developed. Both strategies were designed based on a thorough review of samples of evaluations and how the time period and the type of evaluation (i.e. whether it was a mid-term or end-of phase evaluation) are commonly expressed. Both these QSSs were limited to process text excerpts of the evaluation's terms of reference, executive summary and introduction.

The QSS for the evaluation period (question six) used a predetermined text matcher that recorded all observations that followed patterns where two years (or more) were observed in close proximity to each other. The most common combination found in the document was deemed to be the most likely estimation, and was assumed to be the correct answer. The last strategy, with the purpose of assessing the type of evaluation (question fourteen), also used text matching as a method. This strategy added the document title to sub-sections that were analysed. Both strategies furthermore included validation steps that compared the estimations against available benchmarks, such as a comparison with the publication date of the

evaluations (i.e. the publication date is likely to fall within the estimated project phase in cases of mid-term evaluations).

4.2.3 Caveats

The most obvious limitation of the QSSs in this section is the inflexibility inherent in the rule-based approach. The underlying rules are deterministic, and are mostly designed based on expectations of how the underlying documents are structured. There will most likely always be deviating observations, and it is almost impossible to derive a rule-based approach that yields perfect results when dealing with these relatively complex documents, which feature a high degree of variation in terms of content, structure and individual writing style.

False predictions or inaccurate results are thus unavoidable to a certain degree. One example of this shortcoming is limitations in grasping the importance of a key country when many different countries are frequently mentioned. Another example, with a bearing on time period, is references to older programme periods for the same object. These are examples of context complexity and settings that are difficult for the QSS to handle. Yet another challenge is cases where there is no data recorded to be assessed in the QSS (e.g. no county names were found). There are also limitations when it comes to capturing more abstract notions such as the theoretic scope rather than a practical one. For instance, if operations in an evaluated project or programme are centred on a few countries but the project's objective(s) suggests a wider geographical scope beyond the core/mentioned countries, the designed strategies have been observed to make the wrong predictions on occasion.

However, the developed strategies also have advantages. Besides the typical factors relating to advantages in speed and consistency, the geographical strategies have harnessed the advantages to collect larger amounts of data by default. The designed strategies record additional metadata besides the necessary country names, including

information on for example geographical region and focus area. During the text parsing exercise, additional entities were singled out, such as OECD/DAC donor countries and donor organisations, which can give additional insights into how the evaluated projects/programmes are funded. This data could also be used to produce deductive estimations with regard to whom Sida and Sweden collaborate with. This could in turn fuel an analysis of key objectives in the Paris Declaration on Aid Effectiveness relating to donor alignment and harmonisation from a Swedish perspective, for instance.

4.3 Funding and donors

This section focuses on questions that in one way or another are tied to the funding of the evaluated projects and programmes. More specifically, it focuses on content with a bearing on funding and donor-related issues that are discussed in the processed evaluations. The aim has been to design QSSs that can yield reliable estimations for the questions below:

Q9. Is Sida the sole financier?
Q21. Does the evaluation assess the importance of Sida's funding relating to the contribution's sustainability/lack of sustainability?
Q22. Does the evaluation analyse whether the contribution is dependent on funds from international donors?

4.3.1 Findings

Accuracy of designed strategies

The expectations of designing solid models for these questions were low at the outset of this study. However, the designed QSSs have performed relatively well for all questions in this section. Compared

to the LME dataset, the accuracy levels range from 68 percent for the assessment of Sida's importance (question 21) to 78 percent for dependency on international donors (question 22). When the results are compared to the third-party validation data, the accuracy is suggested to be lower for all the listed questions. It is also somewhat surprising that the accuracy levels between the LME and the third-party validation datasets are even lower.[14]

[14] In a few cases, the accuracy levels between the validation dataset and the LME dataset are based on few observations (<5 evaluations). This is an effect of the LME dataset's varying coverage – not all questions have 128 observations – and since the validation dataset is based on a random sample of the full sample (128 evaluations), the number of comparative observations varies between questions.

Table 3: Results and performance estimates for QSSs with a bearing on funding and donors

Question	QSS accuracy	Label counts	Random adj. accuracy	Third party accuracy	LME vs. third party	Anticipated difficulty	Assessed difficulty
Q9	72%	2	64%	55%	40%	High	High
Q21	68%	4	55%	55%	50%	Very high	Very high
Q22	78%	2	59%	59%	80%	High	High

Notes:
QSS accuracy: Comparison of designed question-specific strategy's predictions against LME dataset labels.
Label counts: Number of plausible labels for answering a specific question.
Random adj. accuracy: Theoretical accuracy scores of a random guess with probabilities for each outcome adjusted to match the empirical frequencies in the LME dataset
Third party accuracy: Accuracy scores for predictions from QSS against the third-party validation assessment.
LME vs. third party: Comparison of third-party validation assessments with LME dataset.
Anticipated difficulty: Our initial pre-study/ex-ante assessment of the difficulty of finding a good strategy.
Assessed difficulty: Our ex-post assessment of how difficult it was to find a good strategy.

Descriptive statistics from scaled-up analysis

All of the developed strategies were deemed to be reliable enough [15] to be used and deployed on the full sample of collected evaluations (>300 evaluations). A large majority of the evaluated projects/programmes/organisations were found to have references to several donors' organisations and/or OECD donor countries. In more than 8 out of 10 (80%) of the processed evaluations, the developed QSS for assessing whether Sida was the sole donor (question nine) found one or more donors (besides Sida/Sweden) that were referenced in relation to the funding of the project/programme. In the remaining twenty percent of the cases, Sida/Sweden was the only entity mentioned in the same context. The vast majority of the evaluated projects and programmes are thus believed to receive funding from multiple donors.

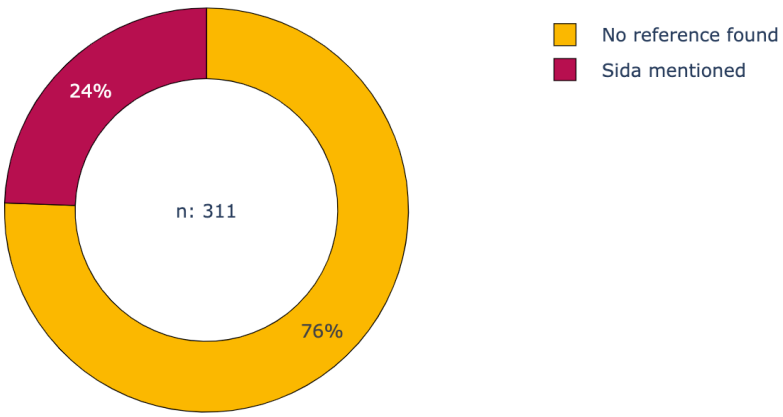
The same strategy also collected data on the specific donor organisation and donor country mentioned in the evaluations. Sweden was referenced for funding in relevant text paragraphs in 98 percent of the processed evaluations, which is not surprising given that Sida is undertaking the evaluations and is thus likely to be a donor. The large percentage for this observation – Sida being mentioned in the relevant context – adds support to the strategy's ability to find and extract relevant text passages and is a validation of the overall performance level. Other commonly mentioned OECD donor countries[16] are the USA (18 percent), Norway (8%), the UK (6%), France (5.5%), Belgium (5.5%), Netherlands (5%), Denmark (3.5%), Canada (3%) and Germany (3%).

[15] The estimation of the reliability of the QSS prediction capability was determined by the displayed accuracy levels. In general, and in most cases, 70% was used as a lower limit for when to include the designed QSS in the scaled-up analysis of >300 evaluations.

[16] The displayed estimates also include observations of national donor organisations for each country.

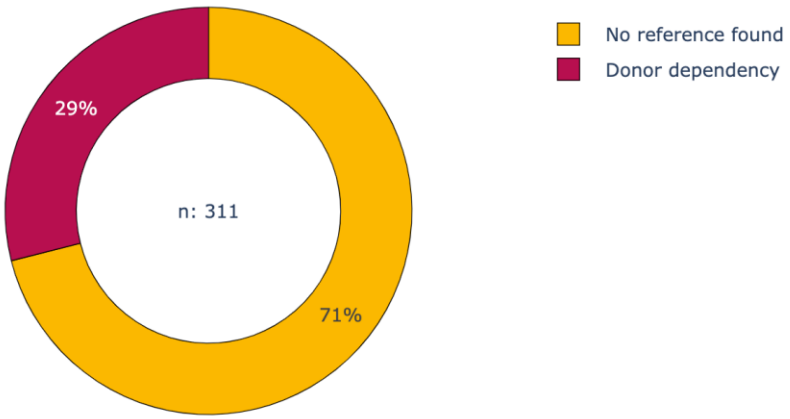
The second developed QSS in this section aimed to estimate whether or not Sida's funding was discussed in relation to its importance for the evaluated entity's sustainability (question 21). This strategy found relevant content in almost one in four (25%) of the processed evaluations, as shown in Figure 7 below. This would suggest that, in one way or another, roughly a quarter of the evaluations discuss the importance of Sida's funds in proximity to discussions relating to the sustainability of the entities being evaluated. Figure 7 depicts this result for the entire assessed period.

Figure 7: Estimation of Sida's importance for sustainability 2011–2020



The final strategy was designed to give an estimate of the extent to which the concept of donor dependency is covered in the evaluations. As shown in Figure 8, roughly 30 percent of the processed evaluations had one or more text passages with content that semantically matched with what the current QSS was designed to record, i.e. donor dependency. The chart below shows the relative shares for all processed evaluations.

Figure 8: Estimation of discussion of donor dependency between 2011 and 2020



4.3.2 Applied method(s)

Each QSS has required a unique mixed-methods approach. Initially, a model with pre-trained word embeddings using word2vec was used in the strategies for all three questions. The purpose of this initial step was to identify semantically similar words to the key words in the spelled-out questions (i.e. which semantically similar words for *donor*, *funding*, *sustainability*, *importance*, etc. were used in the processed evaluations). This pre-processing was executed, for each question, on the entire corpus (i.e. all words in the >300 evaluations were scanned). The resulting output was a set of words that share contextual and semantic similarity to key words in each of the presented questions.

The next step included designing a flexible analytical structure for applying the various sets of semantically similar words that could be fitted to the requirements for each individual question. This flexibility allowed each of the QSSs to undertake targeted searches for text paragraphs with content of particular relevance for each question. The final step, which was the same for all the strategies,

included designing a rule-based approach that located text paragraphs with a bearing on *funding*. All positive observations were extracted for additional analysis.

At this stage of the analysis, the three strategies diverged and applied unique rules depending on the specific requirements for each question. However, all strategies followed the same logic with identical analytical steps – processed sentence by sentence, and recorded all sentences where words in the selected text paragraphs matched with the semantic word sets tied to each question. For example, if a text paragraph with a reference to funding was found, the paragraph was then selected for additional analysis. If one or more sentences in the selected text paragraph included words with semantic similarity to *Sida/Sweden* as well as *important* and *sustainability*, this would register as a positive observation for question 21.

4.3.3 Caveats

Despite the fairly good results, the developed strategies have several limitations. First, the strategies in this section were designed – in line with the selected questions from the LME dataset – to pick up on content and assess whether or not a topic of interest is discussed. The current design of the strategies does not account for how or in what way it is discussed. This has implications for the sort of conclusions that can be drawn from this assessment (i.e. the conclusion can, for instance, be that the concept of *donor dependency* was discussed but not whether the project or programme per se was dependent on funds from external parties).

The developed validation lists that the rule-based approaches have drawn upon, relating to donors and semantic words of relevance for the questions at hand, are by no means believed to be exhaustive. It is, hence, likely that the performance of all the strategies in this section can be improved by adding more data on relevant entities as well as semantic words of relevance. The strategies are furthermore

ill equipped to handle text paragraphs where past and present implementation periods and/or funding of other projects are mixed. Examples have been found where references in the processed evaluations were made to past implementation periods, where for instance there were more donors involved, which could then be mistaken for the evaluated period and thus yield erroneous conclusions.

A clear advantage that extends to all the designed strategies in this section is their wider scope and the possibility to process the full texts of the evaluations. Manual follow-up and scrutiny of diverging results showed that some results or observations in the LME dataset seem in some cases to have been extracted from a limited part of the processed evaluations, for instance the sustainability chapter. Hence, and due to the wider scope applied in the QSS, which stretched beyond obvious parts of the evaluations in the search for specific content, accurate observations with relevant content were able to be identified and recorded.

4.4 OECD/DAC evaluation criteria

This section presents the results for the developed QSSs in terms of assessing what evaluations have concluded regarding the OECD/DAC evaluation criteria sustainability as well as the potential in terms of scaling up to an assessment of what the evaluations have concluded about other OECD/DAC evaluation criteria. The specific questions that are addressed in this section are:

- Q17. Is the contribution (and/or its results) deemed to be sustainable?
- Q23. Does the evaluation mention the contribution's sustainability in the evaluation's summary?
- Q24. Does the evaluation mention the contribution's sustainability in the evaluation's recommendations?
- Q25. Does the evaluation give recommendations for how the contribution can improve its sustainability?

4.4.1 Findings

Accuracy of designed strategies

The results show that the initially developed QSSs, which used a pre-trained sentiment classifier to address question 17, performed quite poorly when evaluated against the LME dataset. The predictions of our sentiment classifier aligned with the manual assessments in the LME dataset in only 41 out of a total of 126 assessed evaluations. This corresponds to an accuracy score of approximately 33%.

As detailed in the Applied methods section below, the poor results triggered an attempt to establish a second strategy for question 17. This strategy was designed to test whether the predictions could be improved if we instead trained a model of our own for the specific task at hand. This approach is what the NLP community would generally resort to when the accuracy of models is of central importance. High accuracy rates are, however, typically produced when large volumes of labelled training data are available, and the 126 evaluations that were available in this case should be regarded as a very small dataset, which at the outset thus dimmed our hopes of successfully training an accurate model. Interestingly, the trained model did however perform significantly better than our previous strategy. Based on a fivefold cross-validation (see Applied methods below), our model aligned with the labels derived from the LME dataset, on average, in 11 out of 27 evaluations. This corresponds to approximately 40 percent of the predictions correctly aligning with the LME dataset.[17] Although this result is not likely to be sufficiently high to be of any practical use, it still suggests that this strategy has potential if more time and resources are spent on optimising the model parameters, as well as extending the size of the training

[17] The corresponding standard deviation for the cross-validation was approximately 6%.

dataset, for instance by annotating additional labels that could be used to improve the model's accuracy.

It is also noteworthy that the correlation between the LME dataset and the validation dataset is slightly higher (47%) in this case, which underlines the difficulty involved in finding a successful approach for this question. Further, when evaluating the models trained on the 14 evaluations assessed by the independent third-party expert, the score looks better and reaches an accuracy score of 49%, which thus appears even better in contrast to the comparison between the LME dataset and the third-party assessments. Importantly, it should however be noted that the size of the evaluation dataset is very small, implying that one should be careful when drawing far-reaching conclusions from these results.

An important benchmark when evaluating these results is what outcome one should expect if the accuracy was nothing more than a random guess. In this light, the results of the first approach are slightly better than reported above, and the results of the second approach are significantly better. A purely random guess would on average align with the LME labels about 25% (¼) of the time, indicating that both methods still manage to produce some information of value. As previously mentioned, a comparison with a random guess may however not be the most adequate baseline. Instead, a frequency-adjusted random guess which takes into account the empirical distribution of the labels may be a better comparison (see the Applied methods section below). For the full sample of 126 evaluations, the LME dataset reports 24 as sustainable, 46 as partially sustainable, 46 as unsustainable and 14 as non-applicable. Based on these numbers, the random guess could thus be adjusted and – instead of having an equal probability for each outcome – base these probabilities on each label's frequency of occurrence in the underlying dataset. Such frequency adjusted probabilities would result in an algorithm that guesses unsustainable or partially sustainable ~36.5% of the time (i.e. they both occur in 46 out of the 126 evaluations) and sustainable ~16% of the time

(occurring in 20 out of 126 evaluations) and non-applicable ~11% of the time (occurring in 14 out of 126 evaluations). On average, such an algorithm would align with the LME dataset approximately 30% of the time (38 out of 126 evaluations if repeated enough times). This is a relevant comparison for the second question-specific strategy, where a unique model was trained. The reason for this is simply that if our training data contained no information of value for predicting the correct label, the training algorithm may simply adjust its parameters so that predictions are made based entirely on each label's frequency of occurrence in the training dataset instead of making predictions based on the prediction data.

For questions 23–25, regarding whether the OECD/DAC evaluation criteria was mentioned in the text, the results were much better. This is partly due to a lower level of difficulty. For question 23, we replicated the results from the LME dataset in 83% of the cases, while for questions 24–25 the results matched in approximately 76% of the cases. Table 4a summaries these results (see section 2.3 for detailed explanations of columns).

Table 4a: Results and performance estimates for QSSs with a bearing on OECD/DAC evaluation criteria

Question	QSS accuracy	Label counts	Random adj. accuracy	Third party accuracy	LME vs. third party	Anticipated difficulty	Assessed difficulty
Q17	40%	4	30%	49%	47%	High	Very high
Q23	84%	2	69%	76%	60%	Moderate	Low
Q24–25	67%	2	71%	66%	60%	High	Moderate

Notes:

QSS accuracy: Comparison of designed question-specific strategy's predictions against LME dataset labels.

Label counts: Number of plausible labels for answering a specific question.

Random adj. accuracy: Theoretical accuracy scores of a random guess with probabilities for each outcome adjusted to match the empirical frequencies in the LME dataset.

Third party accuracy: Accuracy scores for predictions from QSS against the third-party validation assessment.

LME vs. third party: Comparison of third-party validation assessments with LME dataset.

Anticipated difficulty: Our initial pre-study/ex-ante assessment of the difficulty of finding a good strategy.

Assessed difficulty: Our ex-post assessment of how difficult it was to find a good strategy.

Questions 23–25 further used an approach that allowed us to also evaluate the results from these questions against other OECD/DAC evaluation criteria (i.e. apart from sustainability). For these criteria areas, the accuracy was assessed using the assessment from our third-party validation dataset. The results turned out to be more or less equally favourable as the results for sustainability. The accuracy scores are summarised in Table 4b below.

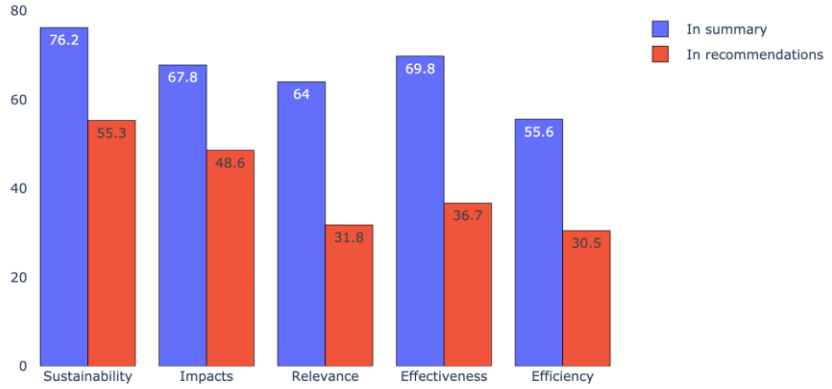
Table 4b: Results and performance estimates for QSSs with a bearing on OECD/DAC evaluation criteria

Question	Relevance	Efficiency	Effectiveness	Impacts
Q23	86%	76%	83%	45%
Q24–25	79%	69%	86%	59%

Finally, and given the higher level of accuracy for the strategies developed for questions 23–25, it may also be of interest to include some statistics when deploying the model on all the collected evaluations (>300 evaluations). The results are shown below in Figure 9, including all OECD/DAC evaluation criteria in the processed evaluations' chapters/sections, covering the summary section as well as the recommendation section.

This data gives an indication of what the evaluations focus on in terms of their analytical scope, and can be cross-referenced with the Terms of Reference to assess the degree of compliance and/or the changes the evaluations undergo compared to the original plan from the Terms of Reference.

Figure 9: Estimation of occurrence (%) of OECD/DAC evaluation criteria between 2012 and 2020



4.4.2 Applied methods

Of these questions, number 17 proved to be the most challenging and several different statistical-based strategies were developed and tested. First, in order to address the question in a way that could be scaled to other OECD/DAC evaluation criteria, our first strategy relied on building a so-called classifier that would classify each evaluation criteria under one of the four labels applied in the LME dataset for this question. The idea was that this classifier would build upon previously trained language classifiers from the open-source community that could be modified for this purpose. In particular, we hypothesised that a so-called sentiment classifier (positive, negative or neutral labelling) could be applied to the sentences or phrases from the evaluation that mentioned sustainability or other OECD/DAC evaluation criteria, and then weigh them together to produce an overall document sentiment score.

Support for our hypothesis came from other attempts to classify tweets from Twitter in a similar way, which have achieved accuracies as high as 60% (Elbagir and Yang, 2019). To accomplish this, we made use of a pre-trained sentiment classifier provided by Hugging

Face's transformers. The approach was thus simply to extract sentences containing variations of the word *sustainability* and then, after some pre-processing (removing e.g. common words and casing), to apply the sentiment classifier to these sentences.

The classifier gave us a label (positive or negative) and a score between one and zero indicating the strength of the sentiment. An overall score was then obtained by averaging the individual sentence scores. Finally, we experimented with various threshold levels for when a score should indicate that a document belonged to the "partially" or "not applicable" category as labelled in the LME dataset. As reported in the results section below, the performance of this approach was unsatisfactory.

For this reason, we also looked into a second strategy for addressing the question. This method relied on state-of-the-art methods for text classification using another modelling pipeline that was also provided by Hugging Face's transformers. This strategy made explicit use of the manually labelled data from the LME dataset. More specifically, we used the labels provided to train a machine-learning model with the specific objective of predicting the label given to a specific evaluation.

The training process involved designing and setting up a neural network model featuring pre-trained parameters from a large-scale state-of-the-art language model, and then fine-tuning that model for the task of classifying our evaluations under one of the four categories in the LME dataset.[18] The pre-trained parameters we used for this task were based on the well-known language representation model BERT (Devlin et al., 2018). This model was adopted due to both its high-performance level and the fact that it has become a benchmark model that many subsequent models have been

[18] The LME dataset included four labels for the assessment of question 17: "yes", "no", "partially" and "not applicable".

evaluated against.[19] The implementation of this approach was done using the spaCy package. In order to accurately evaluate the model, we split the full dataset into a training dataset consisting of approximately 80 percent randomly chosen evaluations that were represented in the LME dataset (100 evaluations) and a test dataset consisting of the remaining 20 percent (27 evaluations).[20] The test dataset was then used to evaluate the predictions of the model that we trained using the training dataset. This procedure was repeated five times with different test datasets in a procedure known as cross-validation. The accuracy score reported below is an average of these five models' performance on their respective test datasets.

For questions 23–25, it was deemed that strategies should be based on rule-based methods to search for predetermined patterns, and that this would give satisfactory estimations. All these strategies furthermore relied heavily on the successful parsing mechanism and the above-described capability to parse certain sections in the evaluations (i.e. the executive summary for question 23 and the recommendation section for questions 24–25). All positive observations were recorded and counted. The third-party validation data also provides some additional insights into the performance and accuracy of the developed strategies. When it comes to question 23 – whether sustainability is mentioned in the summary – the developed strategy correlates well with both the LME dataset (84%) and the third-party validation dataset (76%). Questions 24 and 25 – whether the evaluation recommendation handles sustainability – obtained decent scores, with roughly two thirds of the predictions matching for both datasets.

---

[19] Since the publication of BERT, several models have been published that have a slightly higher performance in benchmark datasets than the BERT model. Examples include models such as RoBERTa, XL-NET and T5. However, since these models are not as common benchmarks as the BERT model, we did not use them in this study.

[20] The random choice was done in a way that ensured an approximately equal balance of labels in both the training data and the test data.

4.4.3 Caveats

The challenges in this section, and in particular for question 17, were mainly due to the complexity of the language used in many of the processed evaluations. Based on our own manual assessment of excerpts of text passages on sustainability, we have observed that a relatively large proportion tends to refrain from using clear statements such as *"the project is not sustainable"*, instead relying on more vague formulations. The text passages below are anecdotal examples from our manual assessment[21] where there have been discrepancies between the guesses produced by our developed strategies and the LME labels.

> "NBE has a strong economic situation, which means that the cost of continuing activities introduced in the NBE/SEA cooperation will not be a major threat to sustainability."

> "It is impossible to provide any general conclusion about sustainability of knowledge gained from the programme. […] which all point in a positive direction regarding sustainability of capacity development. Its sustainability will partly depend on what parts of the Turkish judicial reform programme go forward and to what extent such knowledge is applied. […]. The general conclusion is hardly surprising: that the likelihood

[21] Our manual assessment showed that in many cases our assessments are in consensus with the labels in the LME dataset. However, disagreement is not uncommon in cases where the LME dataset has deemed that there is no support for sustainability or in cases where information was concluded to be non-existent. A likely explanation for the latter is that discussions on sustainability were not always available in a specific sustainability section, but rather were incorporated in a conclusions or summary chapter – something that is easily overlooked when the whole document is not scrutinised for each evaluation.

> for sustainability depends on varying conditions within and outside the programme."
>
> "Sida funding covered activities between 2012 and 2013, and clearly, it would be too early to judge sustainability at this stage. […] It is difficult to discuss the potential for sustainability given the absence of follow-up and monitoring on ERRC's activities. […] Sustainability is also likely to be enhanced by a coherent human rights-based approach that prioritizes processes as much as results, including more focus on building capacity of partners."
>
> "We also assess the capacity built through the municipal-level working groups to be sustainable; this is also the case for some of the working groups, which we expect to operate beyond the project intervention. […] Nevertheless, the sustainability of results of the project depends largely on the ongoing commitment to JJ by government counterparts, which will, to some extent, be a result of continuing advocacy work by the international community."

These excerpts demonstrate the complexity when it comes to passing a judgment on whether or not they advocate for sustainability. This partly explains why both of our strategies – the sentiment approach to classification and the model training approach – underperform for question 17. This complexity furthermore transcends, in our view, the limit between human and computer. In many cases, because there are no simple answers to these questions, and it is this background that foreshadowed the discrepancies between the output from the developed strategies for question 17 and the LME data.

The overall conclusion from the above analysis is thus that the second strategy for question 17, involving training a model, seems to be the most effective approach. However, despite gradual improvements, the model accuracy was deemed too low for inclusion in the scaled-up exercise to cover the full dataset. This accuracy level could however potentially be improved by further fine-tuning the BERT model to the specific vocabulary used within the field of international development cooperation. Another limitation in this case is that the strategy could only scale to assessing more evaluations with respect to the OECD/DAC evaluation criteria sustainability. If another OECD/DAC were to be included, a completely new model would have to be trained for this task alone. The amount of training data would likely need to be more than the available 126 observations in order to produce good results.

4.5 Thematic area

This section presents the results of applying methods in an attempt to develop a QSS capable of generating reliable estimations on the thematic area of an evaluation. The LME dataset contains 16 thematic areas for labelling the evaluations where each evaluation was given a single label. Translated from Swedish to English, the labels were as follows: "Democracy", "Human rights", "Gender equality", "National, regional or local government", "Market entrepreneurship trade innovation", "Agriculture forestry fishing land", "Education", "Research higher education", "Humanitarian aid", "Climate", "Environment and water", "Sexual and reproductive health and rights", "Conflict peace security", "Sustainable community building infrastructure" and finally "several categories in one". Thematic labelling was executed on all 128 evaluations included in the LME dataset.

4.5.1 Findings

Accuracy of designed strategies

The first strategy, relying on word2vec word embeddings (detailed in the Applied methods section below), performed the poorest. Compared to the LME dataset, the strategy predicted the same category in only 29 cases out of 126 (~23%). The second strategy, relying on pre-trained sentence transformers, performed significantly better. Compared to the LME dataset, this model predicted the same category in 44 out of 126 cases (~35%). The third strategy, relying on zero-shot learning, had the highest accuracy scores and performed the best. By using the pre-trained zero-shot learning algorithm, we managed to make predictions corresponding to the LME dataset in 55 cases out of a total of 126 (~44%).

Given that there are 15 distinct labels to choose from in total, a random guess would be expected to make on average eight correct predictions that match the LME dataset after 126 trials. In this light, our first strategy performs more than three times as well as the random guess' the second strategy performs more than five times as well, and the third strategy is more than six times as good. Similarly, a random guess with probabilities adjusted to account for the frequency of occurrences in the dataset would on average guess in line with LME dataset in 15 out of a total of 126 cases (~11%). Although these strategies perform quite well in comparison to random guesses, the results are – in our opinion – still not sufficiently good to replace manual labour, and hence lack the necessary performance to be included in the scaled-up analysis for all collected evaluations.

4.5.2 Applied methods

The task of labelling text and placing it into different bins is typically referred to as topic or text classification. Several methods exist for

approaching this problem, ranging from purely rule-based methods to training models adapted to the specific task at hand. The success of these methods also varies depending on context, and more specifically the extent to which the classes/labels are regarded as exclusive in the sense of being very distinct in terms of the texts in which they are nested.

We tested three different approaches for classifying the evaluations. These approaches all involve the use of word embeddings. The first approach relies on the word embeddings called "en_core_web_lg", which is part of spaCy's NLP package. These vectors are based on Levy and Goldberg (2014), which involves a specific implementation of the word2vec model. The approach also makes use of Tf-Idf vectors, i.e. a method for extracting keywords from a document.

The analytical procedure for this strategy was as follows:

- For each evaluation, we extracted the evaluation keywords using the Tf-Idf algorithm.
- For each category, we manually create a list of ten synonyms for the topic.
- We computed a similarity score between each category synonym and each keyword for each evaluation.
- We matched each evaluation to the category which had the highest average similarity score.
- If two or more topics had very similar average scores, we assumed that the evaluation covered more than one topic.

The second strategy relied on the use of pre-trained sentence transformers from a spaCy extension package. This package wraps sentence transformers (also known as sentence-BERT) directly into spaCy (see Reimers and Gurevych, 2019). The intention of this algorithm is that when the similarity of the pair of sentence embeddings is computed, it should accurately represent the semantic similarity of the two sentences. This differs from standard measures

of sentence similarity, where similarities are computed by simply averaging the similarity among the different words in a sentence. Using the spaCy sentence transformers, we calculated the sentence similarity between the title of the evaluations and the translated categories above.

The procedure for the second strategy was as follows:

- First, calculate the sentence embeddings for evaluation titles using spaCy sentence transformers.
- Second, compute the similarity between these embeddings and the topic descriptions.
- Match each evaluation to the topic which had the highest average similarity score.
- If two topics had very similar average scores, we assumed that the evaluation covered more than one topic.

The third strategy designed to estimate predictions for this question relied on an implementation of zero-shot learning using Hugging Face's transformers. This is an unsupervised machine learning approach which can be used to solve text classification problems when there is no training data available to train a model. Instead, this approach relies on the use of large-scale pre-trained transformer models, similar to what we applied in our previous approach for developing a strategy to estimate the sustainability of the evaluated projects/programmes (question seventeen). Further, as in our second approach, we assumed that the title of the evaluation would provide enough information about which category the evaluation belongs to.

Thanks to Hugging Face's transformers, the procedure for implementing this method is fairly straightforward. The procedure involves feeding the algorithm with a list of evaluation titles, as well as a list of potential categories to which each title may belong. The algorithm returns a score for each potential category, where numbers close to one indicate a high degree of similarity between category and

title. The category that receives the highest score is then chosen as the best guess, unless the score is below a certain threshold which indicates that it may belong to more than one topic.

4.5.3 Caveats

A major caveat for classifying the thematic focus was the fact that there were so many topics and that they tend to be semantically very similar. For example, topics such as democracy and human rights often appear in similar contexts and hence tend to lie close to each other in the vector space. This creates a challenge for these types of algorithms, which tend to work better when categories are semantically more distinct, e.g. sports and politics. A simple way to improve the performance of our third strategy would thus simply be to reduce the number of categories and ensure that their intrinsic meanings are distinct.

Another issue for this section aligns with comments or caveats from earlier sections in this study, namely disagreement – in some cases – relating to the labels applied in the LME dataset. One obvious reason for this discrepancy in the case of assigning a thematic area is likely due to the relatively large variety of labels which may be applied to each of the processed evaluations.

Two examples where we disagreed with the LME dataset and agreed with the labels assigned by the zero-shot learning algorithm were the evaluations with series numbers 2012:2 and 2014:54. In 2012:2, the title of the evaluation was "Review of the Sida-funded Project Education for Sustainable Development in Action (ESDA)". In this case, the zero-shot learning labelled this as the thematic area "Education" while in the LME dataset it was labelled as "Climate change". From the title and reading of the executive summary, we find no evidence in support of labelling the thematic area as "Climate change", but instead find quite compelling evidence for the label "Education". For 2014:54, the title was "MidTerm Review of The LVEMP II Civil Society Watch project of the East African

Sustainability Watch Network". Here, the LME dataset is "Agriculture forestry fishing land", while zero-shot learning gave it the label "Environment and water". On manually reviewing this evaluation, we would also have labelled it as the latter.

On the other hand, our manual assessment is clearly dependent on context and/or the original intentions of the creators of the typology. Our subjective judgements may therefore be incorrect. These examples were thus not chosen to point out that the LME dataset is flawed; rather, the point we are trying to make is that assigning labels is a tricky context-dependent task which is easy to get wrong and thus perhaps not suitable for a fully machine-based assessment.

5 Discussion

A central component of this study has been to observe and take note of strengths and weaknesses in data science and NLP methods when it comes to the challenges involved in the extraction of metadata from evaluations. This chapter discusses these challenges in more detail, with a particular focus on comparing the performance of human and machine-based approaches to data processing. In order to better understand the results of this study, and to put them in perspective, we also focus on possible initial expectations one might have with regard to a machine-based approach. At the end of the section, more detailed limitations of this study are spelled out.

5.1 General strengths and weaknesses of a machine-based approach

There is no shortage of vivid examples where machine-based approaches are outperforming humans [22] in almost any field of operations, and this has led to expectations for the future capabilities of computers and the utilities that they can bring (see e.g. Brynjolfsson et al., 2018). It is hard to argue against the fact that machine-based approaches are on the rise. However, and regardless of recent advances, the future is always difficult to predict. Early pioneers such as John McCarthy promised back in the 1960s that general intelligent machines were within grasp (Marr, 2017). Statements like this tend to leave out intrinsic details (or are lifted out of context). The fact of the matter is that there are many factors that need to come together for a computer to beat a human in even

[22] Two well-known examples are IBM's question-answering software Watson, which outperformed human-level performance when answering questions posed in natural language in 2011 (IBM 2020), and the more recent example of Google's Deepmind AlphaGo, which was able to take another step and beat humans in a game considered to be intrinsically difficult for computers to master (Deepmind, 2019).

the simplest of tasks, let alone when it comes to human language as in the case of this study. There are simply no shortcuts for truly *understanding* human language.

In this light, what is the current state of affairs when it comes to the undisputed advantages of applying a machine-based approach to language understanding? There are a few aspects that cannot be overlooked. The most obvious are those of speed, consistency and endurance. Speed simply relates to the time it takes to complete an assigned task. A computer operates at blazing speeds for the most part, and will thus always outperform humans on repetitive simple tasks. Consistency, on the other hand, entails the prowess not to deviate from an assigned task, which is often a difficult issue for humans. The last general advantage – endurance – relates to the fact that a computer never tires. In the context of this study, these traits are key factors that underlie the potential embedded in a machine-based approach.

Among these traits, consistency is perhaps the most important one. Apart from our third-party comparison, which shows variation in responses among experts, this phenomenon has also been recognised by other studies revealing inconsistencies both over time and between individuals as an inherently human feature.[23] The speed of both collecting and analysing data are orders of magnitude faster (once up and running), implying that the number of evaluations processed in this study has thus not really been an issue. Similarly, advantages with a bearing on endurance have not played a major role in this study, since the volume of evaluations is relatively low and the pace of incoming evaluations is also modest. However, for processes

[23] The differences or discrepancies obtained by repeating a survey or replicating some questions in a survey has received much attention (Biemer, 2004). Several studies have highlighted shifts in preferences, depending on how the same problem is framed (Tversky and Kahneman, 1981). Inconsistencies of choice have also been observed and modelled in the brain (Kurtz et al., 2019).

with larger volumes of data streams and where timeliness of delivery is important, these traits are of the essence.

What about the general weaknesses of the machine-based approach involving computational techniques such as NLP? For starters, a typical machine-based approach can be seen as a competent expert in a very narrow field. As soon as the context is altered, these kinds of systems are potentially in trouble (Alcorn et al., 2019). Another central weakness is that of a lack of contextual understanding. Humans typically – and by default – tend to use preconceived notions with regard to their surroundings, and to the problems that need to be solved. This has proven to be a highly successful strategy for solving problems of many different sorts. And this general flexibility or creativity has been proven to be very difficult to mimic with a machine-based approach. What might seem like the most trivial task which humans take for granted, such as body motor skills, is often an extremely complex task for a computer (see e.g. Minsky, 1986 or Moravec, 1988).

For many of the strategies developed in this study, this has proven to be at the very centre of what has been challenging. The variation in the language used in evaluations is considerable, and as with most NLP tasks it cannot easily be fitted into a rule-based approach. Further, the amount of training data has not been nearly enough to be able to effectively train statistical models for every task. However, even if more data had been available, the task would still have been challenging due to the discrepancies in how humans interpret language. In particular, this is something that was revealed in our third-party validation assessment, which showed that the answers to several questions in the analytical framework were not always aligned between our independent third-party expert and the LME dataset. These factors are clearly important to consider when training machines to complete these tasks, since any errors or biases in the data will of course also be reflected in the model utilising this data in its training process (Mitchell, 2019).

Another aspect which is worth considering when evaluating machine-based approaches is the energy that goes into training these models. For example, in a widely cited study by Strubell et al. (2019), it was estimated that training a single deep learning model like the BERT or GPT-2 model can generate CO_2 emissions corresponding to about a single passenger flight between New York and San Francisco. Given that these types of models are actively being developed and continuously growing in size, it has thus become increasingly important that they are also made publicly available via open source in order to avoid unnecessary carbon dioxide emissions from model training.[24]

A final important aspect is the question of when the methods developed and tested in this study, and NLP in general, are preferable as an approach for drawing inferences regarding a large corpus of evaluations as opposed to drawing inferences based on a smaller sample assessed by a contextual expert. This is a difficult question. First, it is connected to the statistical discourse on the appropriate sample size needed to avoid detecting false effects when there are none (often referred to as type I errors), but also to reduce the probability of not detecting an effect when one exists (type II errors). For example, if we wish to assess how often evaluation studies conclude on average that the evaluated project(s) is deemed sustainable, then the question is how many studies we need to assess in order to feel confident that the sample is representative of the full population of studies. The answer to this question depends on several factors, such as the extent to which we are willing to accept the occurrence of type I and II errors, as well as how far the actual population mean deviates from our ex-ante hypothesis of the correct mean. Second, the question also relates to the finding in this and several other studies that inter-annotator agreements can be low, even among experts. In particular, we have found that when

[24] For example, the current GPT-2 is about one hundred times smaller compared to its successor the GPT-3 model, and therefore requires a substantial amount of additional training cycles and energy.

questions are not well defined and when answers to questions (labels) tend to overlap, this creates problems both for assessments made by contextual experts and for machine-based algorithms, which is also in line with other studies showing that the precision of the results relies heavily on the way the question has been phrased and how distinct the answers tend to be (Artstein, 2017). Finally, the verdict on when to use a machine-based approach versus contextual expertise comes down to a question of organisational resources and requirements, in particular issues such as the transparency of methods and the replicability and scalability of the assessment. If an organisation aims to repeat an assessment at regular intervals, machine-based approaches are highly advantageous, while if an assessment is a one-shot exercise, it depends more on requirements such as transparency and weighing in the issues raised above. In economic jargon, one could say that machine-based approaches are typically associated with high fixed costs but low running costs, while human-based approaches involve low fixed costs but high running costs.

5.2 Observed limitations in the study

Before addressing the limitations, it is first important to emphasise that it is the language in the processed evaluations that has been analysed, which is not to be confused with the data reflecting the performance of actual evaluated projects/programmes per se. In other words, this study reviews the evaluator's language and their (sometimes subjective) assessments of the project's/programme's performance. This means that personal writing styles may (or may not) affect the outcome of the results in this study. And as we have learned (from the results section), there is a relatively large number of individuals involved in the processed evaluations. In aggregation, this means that the context that has been analysed is relatively stochastic and the outputs from the designed strategies should be viewed as more or less qualified estimations, and thus not to be

confused with absolute certainties. Given this context, we believe the following limitations are particularly worth discussing.

First, after the initiation of the study, we detected limitations in the quality of the data which were not known to us at the outset. This realisation came as a result of reviewing parts of the LME dataset, which made us realise that this dataset is not guaranteed to be fully accurate in terms of reflecting a unanimous agreement on the correct answer to all the questions posed. This in turn created a challenge for both training and testing the strategies developed in this study. This is particularly true for a few of the more complex questions dealing with what are often difficult judgments in the evaluations (e.g. judgements on whether or not the project/programme is deemed sustainable, and which thematic field a project/programme should be sorted under). It should, however, be emphasised that this is not an uncommon issue and tends to arise quite frequently when similar exercises are attempted at scale (see e.g. Kurtz et al., 2019). In particular, human-conducted annotations for statistical models and machine learning have proven not to be without flaws. Machine learning practitioners have observed contradictory labelling decisions in human performance, sometimes by the same person over the course of a single day. This has raised questions about the robustness of accuracy tests that rely heavily on human-level performance (HLP) (Ng, 2020). It is, in other words, hard to ensure flawless training and test data. The uncertainty this brings has clearly affected this study, and there is thus an unknown and inherent margin of error that affects both training and testing.

Second, this study has not conducted any normative-based assessment of the questions or their relevance to the field of international development cooperation. The questions have rather been a benchmark for assessing and testing the developed strategies, and have hence served as a guide for assessing the potential of data science and NLP methods for producing meta evaluations. Many of the selected questions from the LME dataset settle on checking for content rather than estimating insights. For instance, several

questions focus on aspects such as whether a concept is discussed in the evaluations, rather than what is concluded about the concept.

Third, the formulation of the questions and the predetermined response categories have in many cases been challenging. For example, the sixteen different thematic response labels included in question 11, which in several cases are semantically similar, make it a difficult NLP task to separate one from the other with the word embedding used in this study. There is also a certain degree of inconsistency in the available labels in the LME dataset, which have brought limitations in terms of how some of the questions could be answered.

A final and central limitation in this study is tied to the scope. All questions, without exception, would benefit from additional adjustments and the accuracy would surely increase if more time could be spent on fine-tuning the developed strategies. In hindsight, it might have been more beneficial for showcasing and exploring the potential of data science and NLP if the number of questions had been reduced, thus allowing more time to be spent on a smaller subset of the selected questions. This could, for example, have allowed for proper context-bound natural language annotations to be incorporated into the study, thus increasing the volume of training data for the more complex questions.

5.3 Moving forward

A nationwide assessment of the current state of affairs for using AI techniques in Sweden stipulated that there is both the potential and a pressing need to automate existing operations in order to improve organisational efficiency and data reliability in many organisations (Vinnova, 2018).[25] In fact, the ability to harness these techniques is stressed as being important for any organisation's future

[25] Among 170 targeted Swedish authorities, only 6% stated that they had ongoing projects at the time involving AI technology (Vinnova 2018).

competitiveness. In line with these suggestions, this study has tried to explore techniques that can be used by organisations to take a step in this direction.

Apart from the results reported and discussed above, several additional research questions have come to light which we think would be of general interest for future studies. In several cases, these questions would be of particular interest for Sida as areas to follow-up on in terms of processing and assessing its decentralised evaluations.

- Explore ways to combine the advantages of a machine-based approach with the strengths of a manual labour approach. Both approaches have clear advantages, often for different types of tasks, and we think it would be valuable to research how a mixed approach could be set up and to assess how this can bring explicit value to an organisation. This study and the developed strategies, such as the generation of metadata from unstructured narrative texts, could help facilitate manual assessments in terms of speed, accuracy and consistency. For instance, the designed algorithm that extracts text passages with a certain content can dramatically reduce the workload of a manual process, and thus establish a good balance where the advantages of a machine-based approach are combined with the strengths of a manual validation assessment. Approaches like this are sometimes referred to as intelligence augmentation, and the most common and widely used example is online search engines, which bring vast improvements to manual work processes (see e.g. Jordan, 2019). In the case of this study, and the stated example of extracting relevant content, a dashboard could ideally be developed where data obtained using a machine-based approach is made available for additional and final manual scrutiny and/or usage.
- Take advantage of generated metadata and explore possibilities for merging it with other existing datasets. For example, make cross-references between the evaluation frequency of countries

and the number of ODA volumes – is there a correlation between funding and follow-up? This could be viewed from a commonly accepted perspective where 1% of the contribution funds should be used for evaluation. How close to or far off from this are current operations? Are there regional variations? It should be noted that in this study we initially pursued the possibilities to find ways to connect the processed evaluations with contribution statistics. This would be relevant for question 12 in the LME dataset, for instance. Based on discussions with Sida staff from the evaluation unit, the majority of the decentralised evaluations were however not tied to Sida's contribution statistics, and it was concluded that there was no easy way to connect the two datasets. This could be worth pursuing, particularly since it would likely bring significant value for policymakers. It is also possible that this could be used to train more accurate statistics-based models, since the labels – at least for some variables – are deemed to be unquestionable in the contribution statistics.

- The overrepresentation of Eastern African countries among the processed evaluations can serve as an interesting follow-up to assess the reasons for this pattern. One possible question may be: How does this result compare the amount of ODA to individual countries?
- Several of the strategies developed in this study recorded additional metadata that was not used in the study. This was the case for the developed strategies that extracted data on geographical entities, where we also recorded data on OECD/DAC donor countries and donor organisations. This could serve as a point of departure for looking into the possibility of researching how projects/programmes are funded. This data could also be used to give estimates of who Sida and Sweden is collaborating with, and thus feed an analysis or estimations on key objectives in the Paris Declaration on Aid Effectiveness relating to, for instance, donor alignment and harmonisation.

- Contemplate adjusting the formulation of research questions and pre-coded response alternatives to better suit a machine-based approach. As mentioned in the results section, there are low hanging fruits for improving the accuracy of a machine-based approach, for instance by limiting the number of thematic labels and making sure that selected categories are semantically diverse. There is also a possibility for this kind of approach to be set up to estimate the level of adherence to several thematic areas (i.e. many contributions are probably at the intersection of two or more thematic categories, and can thus be considered to be labelled as more than one).
- Last, but not least, consider allocating resources for natural language annotation and the development of a high-quality training dataset(s) for more complex tasks within the field of international development cooperation. This would improve the accuracy of the trained models, even for more complex tasks such as labelling evaluations based on thematic focus and/or OECD/DAC evaluation criteria.

6 Concluding remarks

In this study, we have applied data science and NLP methods to extract meta information on specific topics of interest with regard to what past evaluations have concluded about aid projects and programmes. We have found that these techniques can indeed provide useful insights which, on several occasions, appear to be on a par with the errors one might expect from a manual (human-based) assessment of evaluations.

As well as accuracy, the advantages of machine-based approaches also include speed, consistency and endurance in processing the relevant data. Descriptive statistics can be compiled in a quick, efficient and reliable manner with a reasonable rate of error for most of the tested areas and questions. In short, insights can be generated in ways where scale, scope and timeliness are no longer factors of concern. This further implies that there is scope for automation when it comes to deriving interesting insights from evaluation studies.

This does not, however, imply that a machine-based approach and the developed strategies are without flaws. Several limitations have been observed when applying these methods. Obvious challenges with a machine-based approach include its inability to process outliers appropriately. These methods also do not fare well in terms of imperfections in training data and/or larger deviations in the structure and/or content of the evaluations. We have also concluded that the applied methods require careful tuning to effectively answer the questions in this study, which sometimes imply larger costs to get up and running. In particular, we found some of the selected questions were difficult to tackle with a machine-based approach, mainly due to the complexity of the language in the evaluations. Generally, these methods work best when the language is distinct and clear, but struggle when processed texts contain many subtle statements and/or ironies (which, however, also holds true for humans).

The study has also revealed discrepancies among human-based assessments of this kind. This conclusion was drawn from a manual assessment by a third-party, independent evaluator, based on a random sample of evaluations which was also part of this study. The results from this validation exercise revealed, in line with our conclusions regarding machine-based performance, that the third-party answers to the questions also deviated in many cases from the originally conducted assessment (LME dataset). In particular, this was the case for the specific questions which were more complex in nature and thus required careful judgement on behalf of the annotator, but it also occurred with simpler questions, indicating that the task of conducting a large number of manual assessments may be strenuous for humans. These insights thus call for particular care when it comes to phrasing questions for these types of meta studies in order to avoid large variations in annotator agreement due to indistinct or complex labelling which poses a challenge to interpretability, for example. Ideally, a pilot or pre-study test could be conducted that brings in several annotators to ensure questions are phrased in such a way that the answers remain consistent across the annotators. Another potential advantage of a machine-based approach, which can help mitigate situations like this, is the ability to build in checks and balances in order to better understand which types of errors may arise. This also includes the possibility to easily correct for errors discovered at later stages in a process. That is, if an error were to be detected in a product from a manual labour process, this would require extensive resources to redo the work, while with a machine-based approach the underlying code could be adjusted for the detected error, and the analytical process could then be quickly repeated with updated results in a matter of minutes, in many cases.

To summarise, this study has found that machine-based methods have performed quite well for the majority of the questions addressed herein. Importantly, this conclusion is not solely based on the extent to which the machine-based assessment aligned with the original manual assessment, but also on the fact that the third-party assessment did not align much better with the original assessment

than a machine-based approach did for many of the questions. We thus believe there is a potential upside to adopting machine-based approaches for compiling descriptive statistics for meta evaluations, for example, but the value depends largely on the standards to which we want the results to adhere, or in other words what degree of error we are willing to accept in our assessments. From our perspective, there are no straightforward answers to these questions. Answers are likely to vary with the type of decisions that are expected to be taken based on the estimations. Some decisions are likely to require a high degree of accuracy, while others can perhaps settle for less.

References

Alcorn, M.A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.S. and Nguyen, A., 2019. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4845–4854).

Appa Rao et al. 2018. A partial ratio and ratio based fuzzy-wuzzy for characteristics mining of mathematical formulas from documents. IJSC vol 8.

Artstein, R., 2017. Inter-annotator agreement. In Handbook of linguistic annotation (pp. 297–313). Springer, Dordrecht.

Beckett, C., 2019. New Powers, New Responsibilities: A Global Survey of Journalism and Artificial Intelligence, London, The London School of Economics and Political Science. Polis.

Bobicev, V. and Sokolova, M., 2017, September. Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective. In RANLP (Vol. 97).

Brynjolfsson, E., Mitchell, T. and Rock, D., 2018, May. What can machines learn, and what does it mean for occupations and the economy?. In AEA Papers and Proceedings (Vol. 108, pp. 43–47).

Burman, M. 2017. Livslängd och livskraft: Vad säger utvärderingar om svenska biståndsinsatsers hållbarhet. Expertgruppen för biståndsanalys (EBA), 2017.

Choi, S., Fisman, R., Gale, D. and Kariv, S., 2007. Consistency and heterogeneity of individual behavior under uncertainty. American economic review, 97(5), pp.1921–1938.

Cisco. 2019. White paper, Cisco visual networking index; forecast and trends, 2017–2022.

DeepMind. 2019. AlphaGo, a case study. 18 June 2019. https://deepmind.com/research/case-studies/alphago-the-story-so-far

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

EBA (2017), Livslängd och livskraft: Vad säger utvärderingar om svenska biståndsinsatsers hållbarhet?, EBA Rapport 2017:12, Expertgruppen för biståndsanalys.

Elbagir, S. and Yang, J., 2019. Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. In Proceedings of the International MultiConference of Engineers and Computer Scientists (pp. 122–16).

Gentzkow, M., Kelly, B. and Taddy, M., 2019. Text as data. Journal of Economic Literature, 57(3), pp.535–74.

Gentzkow, M. and Shapiro, J.M., 2010. What drives media slant? Evidence from US daily newspapers. Econometrica, 78(1), pp.35–71.

Gentzkow, M., Shapiro, J.M. and Taddy, M., 2019. Measuring group differences in high-dimensional choices: method and application to congressional speech. Econometrica, 87(4), pp.1307–1340.

Goyal, A. Gupta, V. Kumar, M. 2018. Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review, Volume 29, August 2018.

Grus, J., 2019. Data science from scratch: first principles with python. O'Reilly Media.

Gupta, M. 2018, A Review of Named Entity Recognition (NER) Using Automatic Summarization of Resumes, Towards Data Science 9 July 2018.

Honnibal, M. & Montani, I., 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

IBM. 2020. A Computer Called Watson. https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/

International Telecommunication Union, 2020. World Telecommunication/ICT Indicators Database, Individuals using the Internet (% of population), Mobile cellular subscriptions.

Jordan, M.I., 2019. Artificial intelligence – the revolution hasn't happened yet. Harvard Data Science Review, 1(1).

Knippenberg, E. Sibande, R. Chirwa, E. Smith, T. Subramanian, S, K. Sinha, A. K, Raza S. 2019. Using Mobile Phone Data to Make Policy Decisions - A study in how new data sources optimized health facility placement in Malawi. Cooper/smith and Digital Impact Alliance (dial) 2019.

Kurtz-David, V., Persitz, D., Webb, R. and Levy, D.J., 2019. The neural computation of inconsistent choice behavior. Nature communications, 10(1), pp.1–14.

Kurzweil, R., 2004, The Law of Accelerating Returns. in: Alan Turing: Life and Legacy of a Great Thinker. Springer Berlin Heidelberg. pp.381–416.

Lauderdale, B.E. and Herzog, A., 2016. Measuring political positions from legislative speech. Political Analysis, 24(3), pp.374–394.

Logar, Tomaz. Bullock, Joseph. Nemni, Edoardo. Bromely, Lars. Quinn, John A. Luengo-Oroz, Miguel. 2020. PulseSatellite: A tool using human-AI feedback loops for satellite image analysis in humanitarian contexts.

Manning, C, D. Raghavan, P. Schütze, H. 2008. Introduction to Information Retrieval, Cambridge University Press. 2.

Marr B. 2017. 12 AI Quotes Everyone Should Read. Sep 22, 2017 at www.forbes.com

Marshall, I.J. and Wallace, B.C., 2019, Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 8, 163.

Varian H. 2009, How the Web challenges managers. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers

Mitchell, M. 2019. Melanie Mitchell on Artificial Intelligence and the Challenge of Common Sense. October, 14, 2019. https://www.preposterousuniverse.com/podcast/2019/10/14/68-melanie-mitchell-on-artificial-intelligence-and-the-challenge-of-common-sense/

Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Minsky, M., 1987, April. The society of mind. In The Personalist Forum (Vol. 3, No. 1, pp. 19–32). University of Illinois Press.

Moravec, H., 1988. Mind children: The future of robot and human intelligence. Harvard University Press.

Ng, A. 2020, DeepLearning.AI, the Batch - essential news for deep learners, https://blog.deeplearning.ai/blog/the-batch-government-ai-falls-short-face-recognition-for-bears-research-papers-in-one-sentence-counting-crowds

OECD/DAC. 2019. Better Criteria for Better Evaluation Revised Evaluation Criteria Definitions and Principles for Use OECD/DAC Network on Development Evaluation.

Petersson, G.J. and Breul, J.D. eds., 2017. Cyber society, big data, and evaluation: Comparative policy evaluation. Routledge.

Peterson, A., A. Spirling 2018: Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems, Political Analysis, 26, 120–128.

Pincet, Arnaud. Okabe, Shu. Pawelczyk, Martin. 2019. Linking Aid to the Sustainable Development Goals - A Machine Learning Approach. OECD Development Co-Operation Working Papers 52.

Pustejovsky, J. and Stubbs, A., 2012. Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc.".

Reinsel, D. Gantz, J. Rydning, J. 2020, The Digitization of the World, From Edge to Core. International Data Corporation (IDC) and Seagate.

Russel, S. Norvig, P., 2016. Artificial Intelligence – A Modern Approach 3rd Ed. Berkeley.

Scott, S. L., and Varian, H. R., 2015, Bayesian Variable Selection for Nowcasting Economic Time Series. In Economic Analysis of the Digital Economy, edited by Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, 119–35. Chicago: University of Chicago Press.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.

Sun, C., Qiu, X., Xu, Y. and Huang, X., 2019, How to fine-tune bert for text classification?. In China National Conference on Chinese Computational Linguistics (pp. 194–206). Springer, Cham.

Tversky, A. and Kahneman, D., 1981. The framing of decisions and the psychology of choice. science, 211(4481), pp.453–458.

Tetlock, P.C., 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of finance, 62(3), pp.1139–1168.

UN Global Pulse, Making Ugandan Community Radio MachineREadable Using Speech Recognition Technology, Tool Series, no.1, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. arXiv 2017. arXiv preprint arXiv:1706.03762.

Vinnova, 2018. Artificiell intelligens i svenskt närinsglic och samhälle - Analys av utveckling och potential, Serie: Vinnoa Rapport VR 2018:08.

York, P. Bamberger, M. 2020. Measuring Results and Impact in the Age of Big Data: the Nexus of Evaluation, Analytics, and Digital Technology. The Rockefeller Foundation, March 2020.

Appendix 1 – Analytical framework

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
1	Title of evaluation	Rule-based approach	Low	Low	Highly confident	≈ 100%	-	-	-	-
2	Evaluation number	Rule-based approach	Low	Low	Highly confident	≈ 100%	-	-	-	-
3	Country (include all countries that have been mentioned in the evaluation)	Mixed-methods approach, using text embeddings, unique rules and validation lists	Moderate	Moderate	Confident	73%	-	-	54%	86%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
4	Region (geographical)	Mixed-methods approach, using text embeddings, unique rules and validation lists	Moderate	Low	Confident	79%	6	25%	55%	67%
5	Evaluation's geographical focus (country/local; region; global)	Mixed-methods approach, using text embeddings, unique rules and validation lists	High	Low	Unconfident	86%	3	51%	79%	80%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
6	Time period that is being evaluated	Rule-based approach	High	High	Fairly confident	63%	-	-	57%	100%
9	Is Sida a sole financier?	Mixed-methods approach, using text embeddings, unique rules and validation lists	High	High	Fairly confident	72%	2	64%	55%	40%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
11	Thematic area	Test of various models, including pretrained pre-trained classification model such as zero-shot learn.	Moderate	Very high	Fairly confident	44%	16	-	-	-
14	At what phase of the contribution is the evaluation being conducted?	Rule-based approach	High	Moderate	Fairly confident	76%	4	52%	72%	60%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
17	Is the contribution (and/or its results) deemed to be sustainable?	Test of various models, including rule-based sentiment model, use of pretrained transformer models such as BERT and use of pre-trained classification model such as zero-shot learn.	High	Very high	Confident	40%	4	30%	49%	47%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
21	Does the evaluation assess the importance of Sida's funding relating to the contribution's sustainability/ lack of sustainability?	Mixed-methods approach, using text parsing, text embeddings, similarity assessments and unique rules	Very high	Very high	Unconfident	68%	4	55%	55%	50%
22	Does the evaluation analyse whether the contribution is dependent on funds from international donors?	Mixed-methods approach, using text parsing, text embeddings, similarity assessments and unique rules	High	High	Fairly confident	78%	2	59%	59%	80%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
23	Does the evaluation mention the contribution's sustainability in the evaluation's summary?	Text parsing and unique rule-based approach	Moderate	Low	Confident	84%	2	69%	76%	60%
24	Does the evaluation mention the contribution's sustainability in the evaluation's recommendations?	Text parsing and unique rule-based approach	Moderate	Moderate	Confident	67%	2	71%	66%	69%

No	Question	Approach	Anticipated difficulty*	Assessed difficulty*	Confidence of success**	QSS accuracy	Label counts	Random adj. accuracy	Strategy vs. third party	LME vs. third party
25	Does the evaluation give recommendations for how the contribution can improve its sustainability?	Text parsing and unique rule-based approach	High	Moderate	Fairly confident	67%	2	71%	66%	69%

* The scale for difficulty levels is Very low; Low; Moderate; High; Very high.

** The scale for confidence for success is Highly confident; Confident; Fairly confident; Unconfident

Previous EBA reports

2020:07 *Effects of Swedish and International Democracy Aid*, Miguel Niño-Zarazúa, Rachel M. Gisselquist, Ana Horigoshi, Melissa Samarin and Kunal Sen

2020:06 *Sextortion: Corruption and Gender-Based Violence,* Åsa Eldén, Dolores Calvo, Elin Bjarnegård, Silje Lundgren and Sofia Jonsson

2020:05 *In Proper Organization we Trust – Trust in Interorganizational Aid relations,* Susanna Alexius and Janet Vähämäki

2020:04 *Institution Building in Practice: An Evaluation of Swedish Central Authorities' Reform Cooperation in the Western Balkans*, Richard Allen, Giorgio Ferrari, Krenar Loshi, Númi Östlund and Dejana Razić Ilić

2020:03 *Biståndets förvaltningskostnader För stora? Eller kanske för små*?, Daniel Tarschys

2020:02 *Evaluation of the Swedish Climate Change Initiative, 2009–2012,* Jane Burt, John Colvin, Mehjabeen Abidi Habib, Miriam Kugele, Mutizwa Mukute, Jessica Wilson

2020:01 *Mobilising Private Development Finance: Implications for Overall Aid Allocations*, Polly Meeks, Matthew Gouett and Samantha Attridge

2019:09 *Democracy in African Governance: Seeing and Doing it Differently,* Göran Hydén with assistance from Maria Buch Kristensen

2019:08 *Fishing Aid – Mapping and Synthesising Evidence in Support of SDG 14 Fisheries Targets,* Gonçalo Carneiro, Raphaëlle Bisiaux, Mary Frances Davidson, Tumi Tómasson with Jonas Bjärnstedt

2019:07 *Applying a Masculinities Lens to the Gendered Impacts of Social Safety Nets*, Meagan Dooley, Abby Fried, Ruti Levtov, Kate Doyle, Jeni Klugman and Gary Barker

2019:06 *Joint Nordic Organisational Assessment of the Nordic Development Fund (NDF)*, Stephen Spratt, Eilís Lawlor, Kris Prasada Rao and Mira Berger

2019:05 *Impact of Civil Society Anti-Discrimination Initiatives: A Rapid Review*, Rachel Marcus, Dhruva Mathur and Andrew Shepherd

2019:August *Migration and Development: the Role for Development Aid*, Robert E.B. Lucas (joint with the Migration Studies Delegation, Delmi, published as Delmi Research overview 2019:5)

2019:04 *Building on a Foundation Stone: the Long-Term Impacts of a Local Infrastructure and Governance Program in Cambodia*, Ariel BenYishay, Brad Parks, Rachel Trichler, Christian Baehr, Daniel Aboagye and Punwath Prum

2019:03 *Supporting State Building for Democratisation? A Study of 20 years of Swedish Democracy Aid to Cambodia,* Henny Andersen, Karl-Anders Larsson och Joakim Öjendal

2019:02 *Fit for Fragility? An Exploration of Risk Stakeholders and Systems Inside Sida*, Nilima Gulrajani and Linnea Mills

2019:01 *Skandaler, opinioner och anseende: Biståndet i ett medialiserat samhälle*, Maria Grafström och Karolina Windell

2018:10 *Nation Building in a Fractured Country: An Evaluation of Swedish Cooperation in Economic Development with Bosnia and Herzegovina 1995–2018*, Claes Lindahl, Julie Lindahl, Mikael Söderbäck and Tamara Ivankovic

2018:09 *Underfunded Appeals: Understanding the Consequences, Improving the System,* Sophia Swithern

2018:08 *Seeking Balanced Ownership in Changing Development Cooperation Relationships*, Nils Keizer, Stephan Klingebiel, Charlotte Örnemark, Fabian Scholtes

2018:07 *Putting Priority into Practice: Sida's Implementation of its Plan for Gender Integration*, Elin Bjarnegård, Fredrik Uggla

2018:06 *Swedish Aid in the Era of Shrinking Space – the Case of Turkey*, Åsa Eldén, Paul T. Levin

2018:05 *Who Makes the Decision on Swedish Aid Funding? An Overview*, Expertgruppen för Biståndsanalys

2018:04 *Budget Support, Poverty and Corruption: A Review of the Evidence*, Geske Dijkstra

2018:03 *How predictable is Swedish aid? A study of exchange rate volatility,* Númi Östlund

2018:02 *Building Bridges Between International Humanitarian and Development Responses to Forced Migration*, Alexander Kocks, Ruben Wedel, Hanne Roggemann, Helge Roxin (joint with the German Institute for Development Evaluation, DEval)

2018:01 *DFIs and Development Impact: an evaluation of Swedfund,* Stephen Spratt, Peter O'Flynn, Justin Flynn

2017:12 *Livslängd och livskraft: Vad säger utvärderingar om svenska biståndsinsatsers hållbarhet?* Expertgruppen för biståndsanalys

2017:11 *Sweden's Financing of UN Funds and Programmes: Analyzing the Past, Looking to the Future*, Stephen Browne, Nina Connelly, Thomas G. Weiss

2017:10 *Seven Steps to Evidence-Based Anticorruption: A Roadmap*, Alina Mungiu-Pippidi

2017:09 *Geospatial analysis of aid: A new approach to aid evaluation*, Ann-Sofie Isaksson

2017:08 *Research capacity in the new global development agenda*, Måns Fellesson

2017:07 *Research Aid Revisited – a historically grounded analysis of future prospects and policy options*, David Nilsson, Sverker Sörlin

2017:06 *Confronting the Contradiction – An exploration into the dual purpose of accountability and learning in aid evaluation*, Hilde Reinertsen, Kristian Bjørkdahl, Desmond McNeill

2017:05 *Local peacebuilding – challenges and opportunities*, Joakim Öjendal, Hanna Leonardsson, Martin Lundqvist

2017:04 *Enprocentmålet – en kritisk essä*, Lars Anell

2017:03 *Animal health in development – it's role for poverty reduction and human welfare,* Jonathan Rushton, Arvid Uggla, Ulf Magnusson

2017:02 *Do Anti-Discrimination Measures Reduce Poverty Among Marginalised Social Groups?* Rachel Marcus, Anna Mdee, Ella Page

2017:01 *Making Waves: Implications of the irregular migration and refugee situation on Official Development Assistance spending and practices in Europe*, Anna Knoll, Andrew Sherriff

2016:11 *Revitalising the policy for global development,* Per Molander

2016:10 *Swedish Development Cooperation with Tanzania – Has It Helped the Poor?* Mark McGillivray, David Carpenter, Oliver Morrissey, Julie Thaarup

2016:09 *Exploring Donorship – Internal Factors in Swedish Aid to Uganda,* Stein-Erik Kruse

2016:08, *Sustaining a development policy: results and responsibility for the Swedish policy for global development* Måns Fellesson, Lisa Román

2016:07 *Towards an Alternative Development Management Paradigm?* Cathy Shutt

2016:06 *Vem beslutar om svenska biståndsmedel? En översikt,* Expertgruppen för biståndsanalys

2016:05 *Pathways to change: Evaluating development interventions with Qualitative Comparative Analysis (QCA),* Barbara Befani

2016:04 *Swedish responsibility and the United Nations Sustainable Development Goals,* Magdalena Bexell, Kristina Jönsson

2016:03 *Capturing complexity and context: evaluating aid to education,* Joel Samoff, Jane Leer, Michelle Reddy

2016:02 *Education in developing countries what policies and programmes affect learning and time in school?* Amy Damon, Paul Glewwe, Suzanne Wisniewski, Bixuan Sun

2016:01 *Support to regional cooperation and integration in Africa – what works and why?* Fredrik Söderbaum, Therese Brolin

2015:09 *In search of double dividends from climate change interventions evidence from forest conservation and household energy transitions,* G. Köhlin, S.K. Pattanayak, E. Sills, E. Mattsson, M. Ostwald, A. Salas, D. Ternald

2015:08 *Business and human rights in development cooperation – has Sweden incorporated the UN guiding principles?* Rasmus Klocker Larsen, Sandra Atler

2015:07 *Making development work: the quality of government approach,* Bo Rothstein and Marcus Tannenberg

2015:06 *Now open for business: joint development initiatives between the private and public sectors in development cooperation,* Sara Johansson de Silva, Ari Kokko and Hanna Norberg

2015:05 *Has Sweden injected realism into public financial management reforms in partner countries?* Matt Andrews

2015:04 *Youth, entrepreneurship and development*, Kjetil Bjorvatn

2015:03 *Concentration difficulties? An analysis of Swedish aid proliferation*, Rune Jansen Hagen

2015:02 *Utvärdering av svenskt bistånd – en kartläggning*, Expertgruppen för biståndsanalys

2015:01 *Rethinking Civil Society and Support for Democracy*, Richard Youngs

2014:05 *Svenskt statligt internationellt bistånd i Sverige: en översikt*, Expertgruppen för biståndsanalys

2014:04 *The African Development Bank: ready to face the challenges of a changing Africa?* Christopher Humphrey

2014:03 *International party assistance – what do we know about the effects?* Lars Svåsand

2014:02 *Sweden´s development assistance for health – policy options to support the global health 2035 goals,* Gavin Yamey, Helen Saxenian, Robert Hecht, Jesper Sundewall and Dean Jamison

2014:01 *Randomized controlled trials: strengths, weaknesses and policy relevance,* Anders Olofsgård

Den digitala utvecklingen har bidragit till nya möjligheter att generera information och insikter genom maskinbaserad analys. Inom utvärderingsprofessionen har man lyft behovet av en vidgad analytisk verktygslåda för att den ökande tillströmningen av data ska kunna hanteras och nyttjas på ett bra sätt. Syftet med den här studien är att utforska potentialen med att använda metoder inom data science och språkteknologi i utvärdering av internationellt utvecklingssamarbete.

The digital development has provided new possibilities to generate information and insights through machine-based analysis. Evaluators have raised the need for a broadened analytical toolbox if increases in the volume, velocity and variety of data are to be handled and taken full advantage of. The purpose of this study is to explore the potential of data science and natural language processing methods for use in evaluations within international development cooperation.

EBA
www.eba.se

Expertgruppen för biståndsanalys (EBA) är en statlig kommitté som oberoende analyserar och utvärderar svenskt internationellt bistånd.

The Expert Group for Aid Studies (EBA) is a government committee with a mandate to independently analyse and evaluate Swedish international development aid.