Methods report

Natural Language Processing in Evaluation

Reflections, lessons learned, and further analysis

Department for Evaluation

Prepared for Norad

Authors:
Enrique Young, Jake Barrett, Ingela Ternström

ISBN: ISBN 978-82-8369-212-9
Published: 04.11.2024

norad.no/evaluation

Contents

1. Introduction

The context and purpose of this report

In 2023, Norad's Department of Evaluation commissioned an evaluation of cross-cutting issues (cross-cutting issues) in Norwegian development cooperation. The purpose of the evaluation was to provide evidence about how cross-cutting issues were implemented in the Norwegian aid administration and whether their consideration ultimately contributes to better results. The cross-cutting issues were those identified by the Norwegian government for consideration in all aspects of development cooperation:

- Human rights
- Women's rights and gender equality
- Climate change and environment
- Anti-corruption

The evaluation adopted a variety of methods to generate evidence relating to four evaluation questions. One requirement of the evaluation, established by its Terms of Reference, was the development and implementation of a machine learning approach to answer the evaluation's second question:

> "How is the Norwegian development administration implementing the four cross-cutting issues into the management of its programmes and projects? And to what extent is this implementation successful"?

As a basis for answering this question, the evaluation team was provided with a copy of Norad's digital archives[1], which were contained in a data dump of approximately 400,000 files dating back to 2003[2]. The evaluation team's task was to develop a machine-learning approach to extract relevant project and programme documentation from this data dump, examine it for cross-cutting issue implementation, and use the results to answer the evaluation's second question.

Overall, the results of the process indicated that only a small proportion of programme and project documents reflected a 'do no harm' approach to cross-cutting issues required by the Grant Management Assistant, which guides the administration of Norway's development cooperation. However, the results also indicated that a significantly larger proportion of programme and project documents included measures to proactively address cross-cutting issues[3]. These results were supported by the evaluation's broader findings, which were informed by evidence obtained through other methods, including interviews, focus group discussions, a survey, and document review.

This learning-focused report reflects on the process that the team followed to develop a machine-learning approach to answering the second evaluation question. It describes the phases that were

---

[1] This component of the evaluation related to Norad-administered grants only, unlike the rest of the evaluation which also considered grants managed by the Ministry of Foreign Affairs and the Ministry of Climate and Environment.
[2] Though the scope of the evaluation was 2018-2022, the team requested files that were archived from 2003 to ensure that design documentation relating to projects implemented within the evaluation's time frame were included in the sample.
[3] Note that these findings relate to how cross-cutting issues were addressed in documentation, and do not necessarily indicate how cross-cutting issues were addressed on the ground.

followed, the challenges that were encountered and how these were addressed. It also identifies lessons that were learned along the way and offers practical advice and suggestions to evaluators and commissioners on the potential use of similar methods in future evaluations. The report also contains some additional analysis of the results obtained for this evaluation question that were not included within the main evaluation because of time and space limitations.

What is Natural Language Processing?

Natural Language Processing (NLP) is a "set of methods for making human language accessible to computers"[4]. It has a variety of applications, but put simply, it is a set of techniques used to programme computers to understand and process human language to automatically undertake tasks that would be highly laborious for humans at scale. For example, it can be used to summarise, classify, or translate large quantities of text in a relatively short time. Artificial Intelligence platforms that generate responses to user prompts (i.e. questions or instructions), such as ChatGPT, Claude, and Microsoft Copilot are also based on NLP.

For this evaluation, NLP methods were used to develop a custom computer programme (henceforth referred to as an NLP 'model') to automatically process documents from Norad's digital archives, and classify them according to the manner in which they integrate cross-cutting issue considerations.

[4] Eisenstein, Jacob (2018) *Natural Language Processing,* University of California San Diego Publication, https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf, accessed 01/08/24

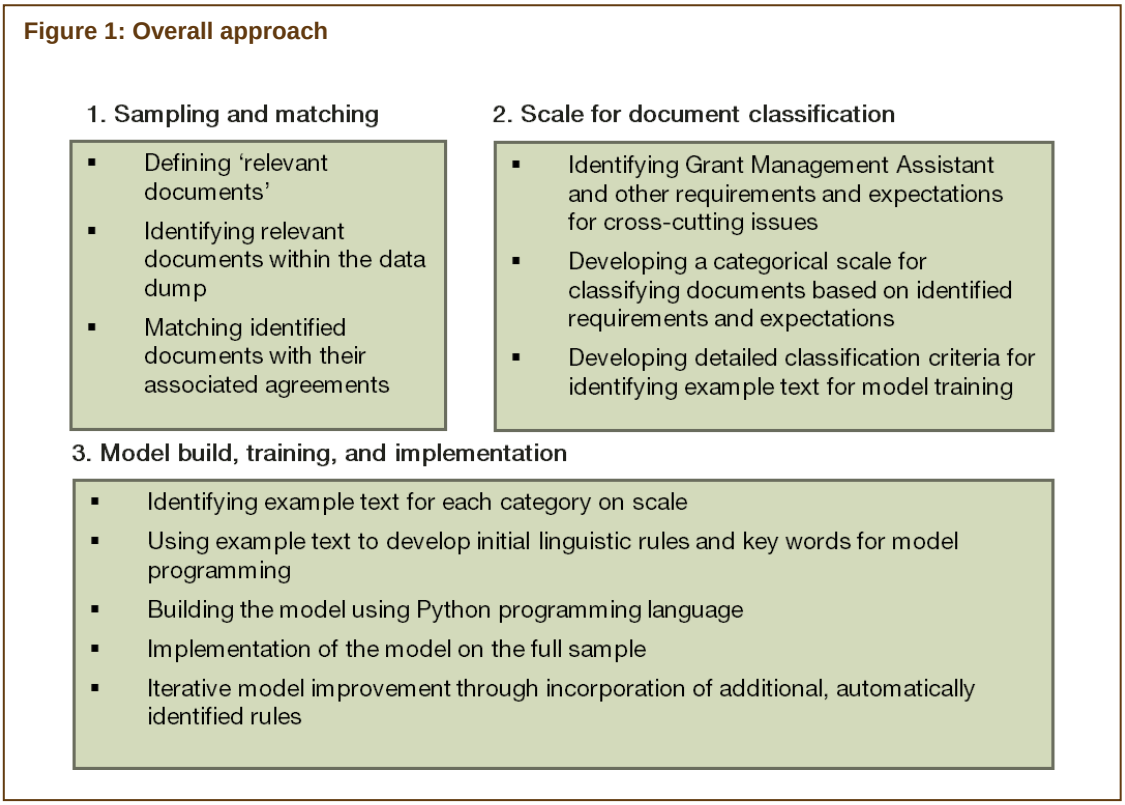
An overview of the approach

The approach adopted by the team involved three consecutive stages, each with a series of steps. These stages and their steps are represented in Figure 1 below, and form a basis for the structure of this report. While the approach adopted by the team is not representative of all NLP processes, it does provide a good reflection of many of the steps that evaluators and commissioners will likely encounter when applying these methods.

The first part of this report is focused on discussing each of these stages. The discussion for each stage includes:

- A concise description of the overall purpose of the stage and of each step, focusing on; a) why it was necessary; and b) how it was done.
- The main challenges encountered, and how these were overcome.
- Key lessons learned, and practical advice for similar exercises in the future.

Figure 1: Overall approach



2. Sampling and matching

The steps involved and their purpose

The purpose of this stage was to make sure that the NLP model would look for evidence of cross-cutting issue implementation in the right documents. This meant identifying a *relevant* sample of documents. The purpose was matching these documents to their associated project or programme where possible.

The first step was to **define the relevant documents** for sampling. This was necessary because the data dump that the team received hundreds of thousands of individual files representing a wide variety of documents, including email exchanges, internal and external research documents, budgets and other financial documents, administrative instructions, and unsuccessful funding applications, among others. Clearly, not all of these files were relevant for searching for cross-cutting issue implementation. Therefore, from among this archive of documents, the team needed to define the relevant documents, meaning those where it made sense to search for evidence of cross-cutting issue implementation.

Relevant documents were defined as the key documents, including mandatory documents, associated with the management of different stages of the project cycle. The relevant document types were identified from the Norwegian development administration's Grant Management Assistant (GMA), and through consultations with Norad and MFA staff. Relevant document types were defined for both the design and the follow up-phases of the project cycle.

Once the relevant document types were defined, the team needed to **identify the relevant documents in the data dump**. This was made challenging because of the very large number of files in the data dump, as well as its lack of a useful structure and file-naming convention.

In addition to identifying relevant documents, the team also needed to identify whether or not these could be linked to an agreement number. This was important for two reasons. Firstly, it was necessary for **matching documents to their associated project.** This was to enable analysis of relationships between project characteristics (i.e. sector, implementing partner group) and the quality of cross-cutting issue

Measures to ensure data security

The NLP process involved access to a large number of files contained in Norad's digital archives. which contained information relating to Norway's development cooperation. Although classified material was removed before the team was granted, the contents of the data dump were likely to have still included potentially sensitive (commercially, politically etc.) material, as well as personal data

For this reason, it was crucial that steps were followed to ensure that the data was kept safe throughout all stages of the process. To achieve this, protocols were developed to guide transportation, storage, and access. The data was saved locally on a secure device, and access to specific files was only granted to team members who had signed confidentiality and access agreements.

To completely ensure the safety of the data, it was processed entirely offline. This meant that commercially available NLP-based models, which require cloud access, could not be used for the task. Instead, the team developed a customised model using the Python programming language, which could process the data offline. Though custom offline models are more laborious to develop, they are more secure and transparent than commercial, online alternatives.

integration. Secondly, the presence of an agreement number indicated that the documentation in question was related to a project that had successfully received funding, which helped to ensure that unsuccessful applications and their associated documentation were not included in the sample.

The main challenges encountered

Going into this stage, the team held three important assumptions about the nature of the data-dump:

- **Assumption 1:** while the ToRs for the evaluation indicated a need for devising a method for identifying relevant document types, the team assumed that this process would be more straightforward than it was. Specifically, it was assumed that the location of files within the data dump would make it easy to identify which project or programme they were associated with, and that individual files would be named or tagged in a way which would clearly identify the type of document.

- **Assumption 2:** due to information received during the tendering process, the team assumed that the size of the data dump would be a maximum of 50 gigabytes, corresponding to approximately 10 gigabytes per archival year. This was expected to correspond to a maximum of approximately 200,000 documents, assuming 40,000 per archival year.

Both of these assumptions turned out to be wrong. Firstly, the structure of the data dump and the naming conventions used for its files were not helpful for determining either a document's type, or the project or programme to which it was related. Files were structured by year, and 'case number', which had no relation to agreement number. File names were numbers comprising several digits, and which provided no information relating to a what type of content the file contained (neither type of document, agreement, project or case number). A separate dataset containing meta-data on the contents of the data dump also provided minimal useable insights as to its content.

This meant that the process of identifying relevant documents and matching them to associated projects or programmes was challenging and time consuming. To identify relevant documents, the team had to rely on key document templates provided by Norad and the MFA. Using these, it was possible to identify documents in the data dump that followed a similar layout. However, because templates are likely to change over time and are unlikely to be always strictly followed, this method would be unlikely to identify all relevant documents if used alone. To address this limitation, the team supplemented the template search with an additional search for key words relating to the targeted document types. Key word searches were confined to the first few hundred characters of a document to target document titles, and included, for example "annual report", "progress report", "decision document", etc.

The structure of the dataset did not enable identification of which documents related to which overarching agreements (programmes/projects)[5]. This meant that the team needed to devise an alternative approach to matching documents to their projects/programmes. For the matching process, all documents in the data dump were searched for alpha-numeric codes that matched the

[5] Note that the team was provided with metadata relating to the contents of the archive, but this also did not provide a disaggregated indication of the type of document, or the overarching programme/project that it was related to.

format of Norad's agreement numbers. Where documents did contain agreement numbers, these were used when analysing the results to match the documents to their associated agreements and agreement-characteristics (sector, implementing partner type etc.) using Norad's aid statistics.

In order to ensure that documentation relating to projects that were designed before but were implemented during the evaluation's scope (2018-2023), the team requested access to documents that were archived before 2018. This increased the size of the data dump received by the team[6]. The total size of the entire archive was 753 gigabytes, and it contained approximately 400,000 files, some of which were made up of multiple documents. The large size of the data dump meant that computing times for the initial processing of the data to identify and match documents were very long.

The combined result of the unstructured nature of the data dump and its very large size was an unavoidable delay to the sampling and matching stage of the NLP process.

Key lessons and practical guidance

Lesson: the quality of an archive's structure and file-naming conventions has a considerable impact on the efficiency of an NLP processes. Efficiency and reliability are maximised when the structure and file-naming conventions for an archive make it easy to clearly identify:

- An individual document's 'type'. For example, document 'types may include project applications, business cases, annual reports, evaluations, budgets etc.
- An individual document's relationship to one or more broader categories. In this context, the most important broader category is the project or programme that a document is associated with. In other contexts, other categories may be relevant, for example theme, year, etc.

Guidance for future processes

When creating an archive or library of documents that may or will be used for NLP processing, aim to ensure that its structure and file naming conventions enable easy identification of a document's type and its relationship to relevant, broader categories. The filing structure of an archive or document library can assist greatly in this. For example, folder hierarchies can indicate the year, department and parent project/programme of the document held within. Equally, document names themselves can follow consistent conventions to enable easy identification of document type. In addition to naming conventions, a system for 'tagging' important features of documents, such as what type of document they are and which project/programme they relate to, can also facilitate easy identification and extraction.

When considering the application of NLP techniques to an existing archive or document library, it is helpful to first carefully consider its structure and naming conventions, and the implications that these have for the method. Specifically, it is important to question whether the way that files are organised facilitates easy identification of what they are and what they relate to. If this is the case,

[6] By way of illustration, the archive year 2020 alone contained 132 gigabytes of data (i.e. more than ten times the estimated 10 gigabytes per year.

then NLP techniques can be applied with less need for pre-processing, implying positive gains for both efficiency and reliability. If it is not the case, however, then it is likely that a considerable amount of pre-processing will be required to identify and categorise documents before the method can be applied. This may be costly and will likely make the overall process slower. It may also affect the reliability of the process, because although techniques can be applied to identify and categorise documents, these may not fully offset the challenges that arise from an archive whose structure and naming conventions do not facilitate rapid identification. In such a context, it is worth carefully considering whether the use of NLP techniques represents the best use of resources. If a decision is taken to use NLP techniques on a poorly organised archive or document library, then it is important to take the additional time and resources that will be necessary for pre-processing the data into consideration in the planning of the NLP process as well as the overall process it is intended to contribute to.

Lastly, it is important that evaluation teams are provided with early access to the dataset as early as possible to enable an assessment of size, structure etc. which can inform planning. This can help to minimise the risk of delays later on in the process. The size of an archive is particularly important to assess, especially if there is a choice relating to how many documents to include. This is because there is an inherent trade-off between depth and breadth. A larger archive will permit a complete sample, but will likely take longer to pre-process, particularly if its structure and naming conventions are not NLP-friendly.

Stage 1 Summary: Sampling and matching

Purpose: ensure that the NLP model is looking in the right place by selecting a sample of relevant documents, and identifying how these documents relate to projects/programmes

Challenges:

- Archive with a structure and naming conventions that did not facilitate rapid, automated identification of document type and links to associated projects
- Extremely large archive causing slow processing times

Lessons and guidance:

- Efficiency and reliability is highest when archive/document library structure facilitates identification of document type and context
- Archive/document library structure and naming conventions should follow clear and useful organising principles
- Decisions about whether to use NLP techniques should be informed by an assessment of quality of the targeted archive or document library's structure
- When NLP techniques are used to process archives with structures that do not facilitate rapid identification of document type and other relevant features, sufficient resources should be allocated for pre-processing

3. Classification scale and criteria

The steps involved and their purpose

The aim of this stage was to clearly define what the model would look for in the sampled documents and how. In this context, this meant defining a systematic approach to categorising text relating to cross-cutting issues in the sampled documentation.

The first step in his stage was to **identify the Norwegian government's expectations and requirements for cross-cutting issue implementation** in order to ensure that these were reflected in the approach to measurement. The main document that was used to identify requirements for cross-cutting issue implementation was the Grant Management Assistant which provides instructions to staff at Norad, the Ministry of Foreign Affairs, and the Ministry of Climate for managing grants. While this contains some guidance relating to the implementation of a 'do no harm' approach to cross-cutting issues at both the design and the follow-up stages of the project cycle, the detail is limited, and there is no guidance relating to proactive implementation. As such, less specific expectations relating to the proactive measures to positively address cross-cutting issues were identified through discussions with Norad and Ministry of Foreign Affairs staff, and though a wider review of relevant documentation.

The next step of the stage was to use the identified expectations and commitments to **develop a categorical classification scale for measuring the type of cross-cutting issue implementation.** This scale would be applied by the NLP model to classify the sampled documents based on an automatic assessment of their content. The scale that was developed included three categories, adaptable to each cross-cutting issue and applicable to both design and follow up phase documents.

To train the NLP model to apply the scale to automatically classify the sampled documentation, examples of text representing each category for each cross-cutting issue had to be identified first. This required the development of a more **detailed set of assessment criteria for identifying examples for each category.** These criteria clearly established what features a document would need to display in order to be assigned to each category. For example, the criteria established what it would mean for a design document to contain a "substantial analysis of risks"[7]. As described in

Summary of the classification scale

The scale developed for classifying documentation contained three broad categories, which could be adapted to each cross-cutting issue and to design or follow-up documentation:

- **Insufficiently implemented**: the sampled document
  - **Design documents:** lacks a substantial analysis of risks to the cross-cutting issue
  - **Follow-up documents:** lacks substantial reporting on risks to the cross-cutting issue
- **Do No Harm:** the sampled document
  - **Design documents:** contains a substantial analysis of risks to the cross-cutting issue
  - **Follow-up documents:** contains substantial reporting on risks to the cross-cutting issue
- **Proactive:** the sampled document
  - **Design documents:** contains cross-cutting issue-specific objectives in the intervention's results chain
  - **Follow-up phase:** includes reporting on cross-cutting issue-specific objectives in the intervention's results chain

[7] For an example of what this means, see chapter 4 below.

chapter 4, these assessment criteria were used by the team to manually identify examples of text from the document library representing instances of each category.

The main challenges encountered

The most important challenge that the team encountered at this stage was the limited availability of explicit information and guidance relating to requirements for how cross-cutting issues should be integrated within different types of project documentation. As noted, the Grant Management Assistant did contain some guidance on risk analysis for cross-cutting issues at various stages of the project cycle, and this was used to inform the 'do no harm' category of the classification scale and the associated assessment criteria for identifying example text. However, there was no equivalent guidance for 'proactive' implementation. As such, the development of assessment criteria for the 'proactive' category involved inputs from thematic specialists on the evaluation team. This challenge related to a broader conclusion of the evaluation, which found that the limited guidance available for cross-cutting issue integration was a key barrier to effective implementation.

Key lessons and practical guidance

Lesson: when designing an NLP process, it is necessary to develop clear answers to the following two questions:

- What are we looking for in the sampled documentation?
- How do we know it when we see it?

In some cases, the answers to these questions will be readily available before the evaluation begins. This should not be taken for granted, however, because in other cases, developing an answer to these questions will comprise a key component of the evaluation process. Such was the case for this evaluation, which included a component that required the team to clarify Norway's expectations and commitments for addressing cross-cutting issues. This represents an instance of good practice, because the structure of the proposed evaluation included a component requiring the team to address these important questions.

Lesson: there is distinction between identifying what to look for and developing an approach to measuring it in a valid and reliable manner. Developing a valid and reliable approach to measurement can be particularly challenging when the issues involved are complex and multidimensional. Categorical measures underpinned by clear criteria for classification, such as the approach utilised by this

Ensuring integration of the NLP method

NLP is a relatively new method in the world of evaluation which demands specific technical expertise for application. Because of this, it can be challenging to ensure that non-machine learning experts in the evaluation team are sufficiently involved in the design and application of the method. This raises the real risk that the method becomes 'siloed' from the rest of the evaluation, and that there will be a resulting mismatch between the evidence that it produces and the evaluation's needs.

To address this risk, it is essential that the non-machine learning experts on the evaluation team are meaningfully engaged in the method at every possible opportunity. This helps to ensure that there is alignment between the NLP process and the evaluations needs. Luckily, there are several stages in the design and implementation of the NLP method which represent excellent opportunities for input by thematic experts. This includes identifying the criteria for sampling relevant documents, defining 'what is being looked for' and how, and identifying example text for model training.

evaluation, are a useful way of empirically capturing examples of what is being looked for. Inputs from thematic experts can help to ensure the validity and reliability of measures. In this context, validity means that the measure should adequately represent the concept being measured, and reliability means that the measure should be equally applicable across contexts.

Lesson: it is important to ensure that thematic experts are involved in answering these two questions, as well as in the development of the associated method. This is to help ensure that the NLP element produces evidence that is relevant for and aligned with the requirements, theory, and concepts guiding the broader evaluation process.

Guidance for future processes

Determining what to look for, and the criteria for identifying it, are among the most important questions to answer for this type of NLP application. Clarity on what to look for improves efficiency by making the search more focused. Having a detailed set of criteria for accurately identifying the type of evidence that is being looked for helps to ensure reliability, and greatly assists in subsequent stages of the process. When planning this type of NLP process, it is important that these discussions are incorporated into the early stages of the workplan. Crucially, answering the two questions above is an excellent opportunity for collaboration between the thematic and the technical members of the evaluation team. Input from thematic experts helps ensure that the process is responding to the evidential needs of the evaluation, and that the criteria guiding the search for evidence are aligned with the broader conceptual definitions included in the evaluation. For technical members of the evaluation team who will be involved in building and applying the NLP model, interaction with thematic experts is crucial for developing an understanding of the context and aims of the evaluation and can also improve their ability to sense-check emerging results during subsequent stages. As such, when implementing this type of NLP process, it helps to ensure as much interaction between technical and thematic members of the evaluation team at this stage.

Stage 2 Summary: Classification scale and criteria

Purpose: ensure that there is clarity around what the NLP model will look for, and the criteria that will be used to identify it

Challenges:

- Clearly identifying what evidence is being looked for in the sampled documentation, and ensuring that this aligns with the wider conceptual needs of the evaluation.

Lessons and guidance

- Clarity on what to look for and how improves the efficiency, reliability, and relevance of the process
- This stage is an excellent opportunity for collaboration between the thematic and technical members of the evaluation team. This helps to keep the focus of an NLP component aligned with the broader requirements of the evaluation.

4. Model Construction, Training and Processing

The steps involved and their purpose

The aim of this stage was to programme and train a customised NLP model that could automatically process the sampled documentation and classify it according to the measurement categories developed in the previous stage. This stage was the most technically demanding of the process, requiring intensive inputs from the team's machine-learning expert.

The first step in this stage involved **compiling a training dataset**. This refers to a collection of sections of text extracted from the sample which fulfil the criteria for each category in the measurement categories outlined above and for each cross-cutting issue. For example, in this case, the training dataset needed to contain several examples of text representing a 'do no harm' approach to gender equality, as well as several examples of text representing a 'do no harm' approach to human rights, and so on for the other two cross-cutting issues. It also needed to contain several examples of text representing 'proactive' approaches to each of the cross-cutting issues. In general, the more examples that a training dataset contains, the better. In this case, the training dataset complied by the team was made up of 174 individual examples of text extracted from the documents in the sample.

Training Dataset Examples

Below are examples of text included in the training dataset, representing different categories in the classification scale. Text identifying the associated implementing agency has been removed.

Do No Harm – Women's Rights and Gender Equality

Specific risk related to project delivery: "....support for women's empowerment and the promotion of gender equality leads to increases in gender-based violence (physical and/or psychological) as women are abused by men that do not agree with their participation in forums and/or increased expression of their rights.

Associated mitigation measure: "gender equality policy in place that guides our decision making and informs our work with partners; [implementing agency] works with gender-focused organisations in most of our country programmes and draws upon their expertise to programme directly towards addressing gender-related discrimination and violence, as well as to provide capacity building and support to other partners... ... [implementing agency] will continue to focus efforts on tackling gender-based discrimination and violence in society through our programming and advocacy, and will identify additional means of increasing this focus across all programmes we support.

Proactive – Climate Change and Environment

Proactive measure in results chain: "the project will target local and ethnic communities to improve land management skills, establishing silvopastoral systems and diversify family orchards to improve productivity and food security without adding pressure to the forest. Periodical CO2 measuring will be taken to monitor the CO2 storage using innovative remote sensing techniques and promote knowledge transfer".

Insufficiently implemented – Human Rights

Absence of substantial risk analysis: the program will not have negative consequences on the 4 cross cutting issues. Phase IV of the Joint Program will be based on the principle of respect for human rights for all. The proposal promotes the need to invest and be accountable for protecting and promoting the rights of boys and girls, and young men and women.

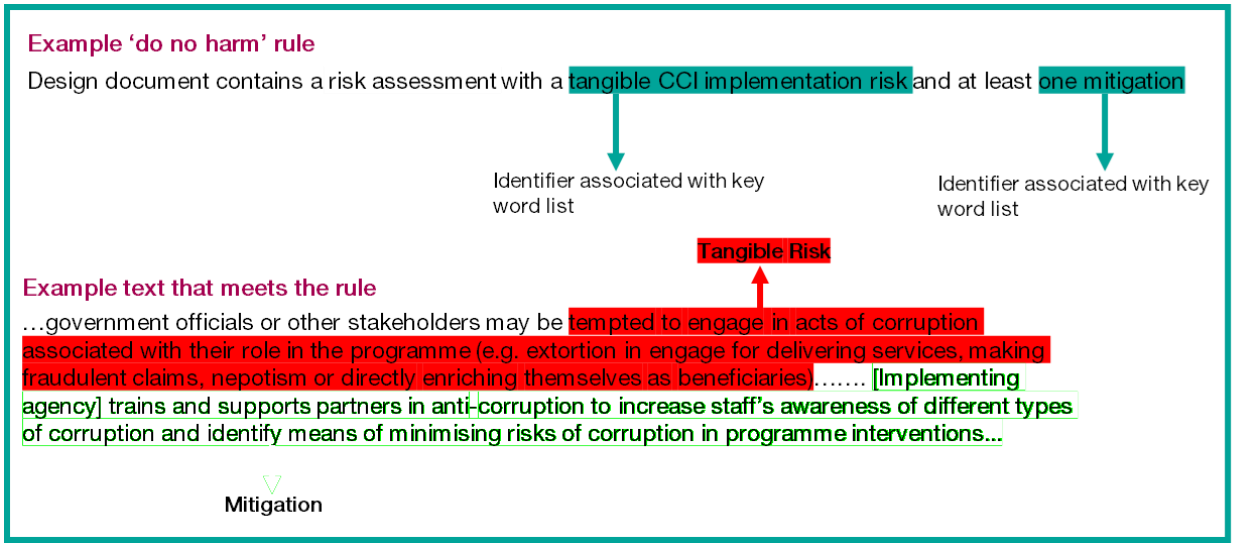
Once the training dataset was compiled, the process of building the NLP model began. This was a complex process, that is best described as a series of consecutive steps.

Developing initial model rules

The foundation of the NLP model was a set of linguistic rules that it would use to categorise sampled documents against the measurement approach described above. These rules were derived from the example text that was compiled for the training dataset during a workshop between team's machine learning expert and thematic experts. A total of 15 initial rules were identified, corresponding to the different categories of the measurement approach. Of these, 9 rules were for identifying text representing 'proactive' implementation, and 6 rules were for identifying text representing 'do no harm' implementation. The rules represent the diversity of ways in which a document can demonstrate either 'do no harm' or 'proactive' implementation. The NLP model was programmed to automatically process the contents of each sampled document and determine whether or not any of the rules were satisfied. If the document satisfied at least one rule, then it would be classified as belonging to the category associated with that rule (i.e. as 'do no harm' or 'proactive' for the cross-cutting issue in question)[8].

Individual rules were made up of a specific linguistic requirement and one or more identifiers, associated with separate lists of key words and phrases. This is illustrated by figure 2 below.

Figure 2: Example rule for 'do no harm' identification



[8] Note that documents could be classified as belonging to more than one category. For example, a document might display proactive measures *and* do no harm measures. Alternatively, a document could display proactive measures but no do no harm measures.

Several different classes of key words were included in the lists to facilitate accurate identification:

- Cross-cutting issue-specific: usually nouns or short phrases, reflecting key concepts and phenomena related to each cross-cutting issue. For example, cross-cutting issue-specific words and phrases for climate and environment included 'climate change', 'biodiversity', 'climate change adaptation', 'climate change mitigation' etc.

- Action: verbs signifying that an action has or will be taken. The sentiment of these words was flagged as either positive or negative depending on the noun that they appear close to. For example, the verb 'increase' would be flagged as positive if it appeared next to the noun 'participation', but negative if it appeared next to the noun 'discrimination'.

- Risk-specific key words: verbs, nouns, or phrases that signify that risks are being discussed. Examples include; "increase risk", "due to", "susceptible", "higher chance" etc.

- Result-specific key words: nouns that signify that a project or programme's anticipated results are being discussed. Examples include "activity", "output", "component", "workstream", "target" etc.

Programming and training the model

This step involved developing a custom NLP model to implement the rules on the sampled documentation, using the Python programming language. Python is one of several programming languages that can be used to design such models. The form of the model was a script containing code, which can be run on any device with a Python interpreter installed.

Before the model was applied to the whole sample, it was first run and re-run on the training dataset through an iterative process involving minor adjustments to the code and initial rules to ensure that the example text was being categorised accurately by the model. This process is known as **model training**. Once the model was accurately categorising the example text, it was ready to run on the full sample of documents.

A key feature of the model was that it was programmed to **iteratively improve itself** as it was run on the sampled documents. It did so by routinely suggesting additional keywords for each CCI to augment initial expert-derived lists, based on common terms in document areas that flagged model rules. Before these **automatic rules** were

Online vs. Offline Custom NLP models

There are several commercially available online models that could be applied to tasks such as this. Examples include Chat GPT and Microsoft Co-Pilot. The benefits of using such online models are that they are relatively inexpensive, easy to use, fast, and often very sophisticated in terms of what they can search for and the answers that they can generate. However, using these models means that the documents are processed online. Although the documents can generally be processed in a private and secure online space, there is no guarantee that the data will be immune from inadvertent or malicious data breaches. Furthermore, the methods employed by these online models are typically not public. Their results are therefore less explainable, which reduces the overall transparency of the evaluation process.

The alternative is using custom-built, offline models, such as the one that was developed for this evaluation. These are time consuming and resource intensive to develop, and require highly specialised technical expertise. However, such models are becoming increasingly powerful, and can closely emulate many of the functions of online commercial models can. Because they can be developed and implemented entirely offline, the data that they process is less vulnerable to breaches. Furthermore, because they are custom built, each step of the process as well as the results can be more consistently explained, which enhances transparency.

incorporated into the model, their relevance to the CCI area was verified by the team.

The output of the model was a large dataset where each row represented a single document. The columns of the dataset contain information about the characteristics of each document, as determined by the model. These characteristics included how the document was categorised by the model using the measurement approach. Using the agreement numbers contained in documents and identified by the model, documents were matched to associated agreements described in Norad's aid statistics. Agreement characteristics described in Norad's aid statistics, such as implementing partner type, sector, region, target area etc. were added to the dataset. This meant that for each document in the dataset, there was information about the characteristics of its associated agreement, which could be used when analysing the results.

The main challenges encountered

The team went into this stage with the expectation that programme managers and thematic experts within Norad would be able to identify projects demonstrating high quality cross-cutting issue implementation, from which it would be possible to extract example text for the training dataset. Early in the process, however, it became clear that this was not a viable approach, as personnel consulted during interviews were unable to assist by suggesting projects or programmes that might demonstrate high quality cross-cutting issue implementation. The team therefore had to devise an alternative, manual approach to identifying example text for the training dataset from within the large data dump that was provided.

The alternative approach involved computerised scanning of all documents within the data dump for a set of key words associated with each of the cross-cutting issues. The aim of this was narrow the scope of a manual search, by identifying those documents that were most likely to contain substantial material relating to cross-cutting issues. Once the scan was completed, documents with the highest number of keyword hits were manually reviewed for text examples for the training dataset. Although this process did enable the team to identify examples, it was very time consuming. The large size of the data dump, which contained numerous documents of considerable length, meant that the initial key word scan was slow, and once candidate documents were identified these had to be reviewed manually by team members to identify example text using a set of pre-agreed assessment criteria as described in chapter 3 above.

A further challenge experienced during this stage related to the integration of the NLP process into the wider evaluation. This refers to the need to ensure that the method was aligned with the conceptual framework and evidential needs of the evaluation. This was challenging, because the highly technical and specialised nature of this stage meant limited opportunities for participation by other team members. To address this challenge, the team invested time in ensuring that those elements of the process which did provide opportunities for involvement by non-technical team members were taken full advantage of. One such opportunity was the compilation and verification of the training dataset. Non-machine learning specialists with relevant thematic expertise were involved in developing the assessment criteria used for identifying relevant examples, manually reviewing documentation for examples, and then verifying the examples that were identified. The process for initial rule development also involved collaboration between different team members. This involved a workshop between the team's machine-learning specialist and other team members to go through each piece of extracted example text and agree exactly why and how it fulfilled the assessment

criteria. The justifications arrived at during this workshop were then used by the machine-learning specialist to develop the initial set of model rules.

The final challenge that the team faced in this stage was the considerable computational requirements of running the model on the sampled documentation. The sample comprised over 61,000 files, and the model itself was complex. Initial run-time estimates suggested that the model would take at least two weeks to process all the documents in the sample. The need to ensure data security meant that cloud processing options, which would have been considerably faster, were not available. To address this issue, the team split the sample up into three sets of documents and ran the model separately but simultaneously on each using three separate machines. This reduced the total run time to just under one week.

Key lessons and practical guidance

Lesson: model development is a highly technical process that requires a very specialised skill set. It can be challenging for team members without these skills to understand the method and how it works. This raises the risk that NLP becomes a siloed method within the evaluation. It is important to be aware of this risk and to manage it, because if it materialises then it can result in a misalignment between the method and the conceptual framework for and evidence needs of the wider evaluation.

Lesson: a high-quality training dataset made up of several examples of relevant text extracts is essential but can be difficult to compile. The training data should have enough material to represent the linguistic diversity of the type of material that the model will need to identify and classify. At the same time, it is important that the training data is compiled and organised using a coherent analytical framework. Various examples of text belonging to one category need to be conceptually and/or linguistically linked in a meaningful way. This balance between diversity of examples and analytical coherence can be difficult to achieve.

Lesson: processing times can be long, especially for offline models. The amount of time needed to run a model is influenced by several factors, including the complexity of model, the number and length of the documents that it needs to process, and the processing power of the available computers. If feedback loops are planned between an NLP process and other data collection/analysis methods, there needs to be sufficient time to allow for the different steps in the NLP process as well as backup plans in case of significant delays.

Guidance for future processes

To reduce the risk that an NLP process becomes siloed from the rest of the evaluation, it is crucial to identify and take advantage of every opportunity to involve other team members during the key stages of the method. This will help to ensure that the concepts employed by the model and the evidence that it is designed to search for align with the broader evaluation. There are several stages in the process where non-machine-learning specialists can become meaningfully involved, despite its complex and technical nature. These include devising a scale, categories and criteria, coding framework, or other scheme for categorising data, the development and verification of a training dataset, and the development of model rules.

It is important to devise a strategy for curating a training dataset early on in the evaluative process, preferably during the inception phase. It helps greatly if there are already examples available of the

type of evidence that the model should be looking for within the target archive or document library. These may have been identified by evaluation stakeholders, other evaluations, internal review documents, or other means. If such examples are not available, then the team will have to devise a means of identifying examples from within the archive or document library. While this may be more time consuming, having the evaluation team (rather than the commissioner) identify examples can help to uphold the independence of the process. The identification of examples should be facilitated by clear criteria that outline the preferred characteristics of example text and should be aligned to the evaluation's broader conceptual framework. Curating a dataset in this manner is likely to be time consuming, and it is important that this is built into the evaluation's workplan.

Be aware that the more documents that are in a sample, and the more complex the model, the longer it will take to run. Running time will also depend on the available hardware. A single, business-grade laptop may take several days to process what a more powerful machine could achieve in a matter of hours. Depending on the extent to which data security concerns are a priority, the use of cloud-based processing power may also be an option. This might be the case, for example, if the underlying data consisted of publicly available documents, such as published evaluations. If local processing is the only option, for security or other reasons, then it is important to account for potentially lengthy processing times in the evaluation's workplan. If commissioners expect lengthy processing times due to the volume of data involved, then one option is to clearly specify this in ToRs, and explicitly request that service providers have the requisite hardware available. In general, the technical expert on the team will be able to provide estimates of processing time, if given accurate information regarding the parameters of the likely sample the number of documents, archive structure and size, and file naming conventions.

Stage 3 Summary: Model Build, Training and Processing

Purpose: develop a custom NLP model to process and classify the sampled documentation according to the classification scale and criteria developed in stage 2.

Challenges:

- Limited assistance from evaluation stakeholders for identifying example text for training dataset
- Ensuring meaningful involvement by other evaluation team members without machine-learning expertise
- Long processing times because of large sample and model complexity

Lessons and guidance

- To help ensure alignment between the NLP method and the conceptual and evidential needs of the evaluation, ensure other evaluation team members are involved in:
    - Guiding and overseeing the process of identifying example text for the training dataset
    - Developing the initial linguistic rules for the NLP model
- At an early stage in the evaluation, devise a clear and realistic strategy for identifying example text for the training dataset, and ensure that there is adequate time in the workplan to implement this
- Once the sample and classification scale has been determined, it is usually possible to develop an estimate of the processing time needed for running the model on the documents. Be aware that processing times can be long (i.e. several days), especially if the model has to be run offline and if there is limited hardware available. Ensure that the evaluation's workplan provides ample time for this stage of the process.

5. A closer look at the model's design and results

The primary analysis of the model's output involved calculating the proportion of documents falling in each category of the measurement approach. Analysis also included an assessment of whether or not there were relationships between how documents treated cross-cutting issues and the meta-characteristics of their associated agreements. The main evaluation report had limited space, and the team had limited time to analyse the data produced by the model. Its focus was therefore on the headline results that were most relevant to the evaluation.

This section presents a closer look at the model's output which could not be incorporated into the main report due to these time and space limitations, but which are nevertheless interesting. It also illustrates the type of analysis that can be made of the data that results from the NLP analysis.

A closer look at the model rules and results

The rules for 'do no harm' implementation

For a document to be classified as 'do no harm' it had to identify (for design-phase documents) or report on (for follow-up phase documents) at least one **specific tangible risk to a cross-cutting issue** as well as **measures to mitigate that risk.** The NLP model used several separate rules to identify 'do no harm' implementation in documents. For a document to be categorised as 'do no harm' it had to satisfy at least one of the following rules:

Design phase rules

- **Rule 10:** The design document includes a risk assessment which describes a tangible risk to a cross-cutting issue from implementation, which is linked to a least one mitigation measure.
- **Rule 11:** The design document includes a risk assessment which describes categorised risks to cross-cutting issues from implementation (i.e. by likelihood and impact), linked to mitigation measures.
- **Rule 12:** [Decision-document specific]: The decision-document describes tangible risks to cross-cutting issues from implementation in the 'impact assessment' area of the document.
- **Rule 14:** [Decision-document specific]: The risk assessment section of the decision-document describes tangible risks to a cross-cutting issue from implementation, linked to at least one mitigation measure.

Follow up phase rules

- **Rule 13:** [Progress-report template specific]: the section requiring a description of the project's effects on cross-cutting issues or identified risk factors discusses a tangible risk to a cci, and measures that are being/have been taken to address it.
- **Rule 15:** [any follow-up document]: in any reporting document, there is a section that discusses tangible risks to cross-cutting issues as well as mitigations.

One important thing to note regarding these rules is that Rule 14 should be relatively easy for decision documents to fulfil. This is because of the specificity of the template in question, but also because of the clear requirements in the Grant Management Assistant for Decision Documents to include an analysis of risks to cross cutting issues. As illustrated in figure 3 below, all Decision Document templates contain a section explicitly requiring grant managers to specify the various risks to cross-cutting issues identified by funding applicants.

Figure 3: Section of decision documents requiring overview of risks to cross cutting issues

In addition to this very specific template requirement, the Grant Management Assistant contains relatively detailed guidance for grant managers to fill in this section.

While a majority (77%) of the Decision Documents[9] in the sample contained at least some discussion of risk to at least one cross-cutting issue, only a small minority (~5% when averaged across the four cross-cutting issues) of these documents discussed risk in a manner that fully met the criteria for do no harm. This means that Decision Documents are not consistently providing the required information about tangible risks to cross-cutting issues relating to project implementation alongside related mitigation measures.

Rule number 10 looked for examples of text indicating 'do-no-harm' in other types of design documents. This included project proposals and risk assessments when these were appended to agreement documents. In comparison to decision documents, a smaller proportion (33%) of these documents contained a discussion of risk to at least one cross-cutting issue. This suggests that Decision Document templates are useful in ensuring that grant managers include discussion of risks, even if they cannot guarantee that this is done in a way that fulfils the 'do-no-harm' requirements.

Documents that did fulfil the rules for do no harm included text that identified a clear risk associated with the implementation of the project, as well as a clearly linked mitigation measure. For example, one document noted that a specific risk posed by the project's implementation was that it could "potentially cause serious damage to key wildlife corridors and biodiversity hotspots if not well planned". A specific mitigation measure was then proposed to address this, to ensure that the damage limited to the ecosystem in question was within the thresholds specified by the Aichi

[9] N.B. The percentages presented here differ from those in the main evaluation report, because here they refer to decision documents only, whereas in the main evaluation report they refer to all types of design documents included in the sample.

Biodiversity Targets set out in the Convention on Biological Diversity. Another document pointed to the potential that the project could lead to "environmentally unsound handling of agrochemicals and farming techniques not in line with climate change mitigation and adaptation". The specific mitigation measure in this case involved training on the safe handling and disposal of agrochemicals. Another design document highlights the risks of "sexual exploitation and abuse of program participants and community members by [agreement partner] and partner's programme staff". The linked mitigation in this case involves the development of a "sexual exploitation and abuse policy", and reporting mechanisms including a "free call line, complaint boxes, and SMS". A further example highlights the risk of "more cases of defamation abuse or gender-based violence as men oppose women's increased participation in the public sphere", and has a mitigation of creating "solidarity networks through the use of social media and online communities.. ..to confront harmful gender norms and practices".

How did discussions of risk fail to meet the 'do no harm' criteria?

In both decision documents and other design documents, most discussions of risk were found to fall short of the criteria identified for the 'do no harm' approach. There are several examples available which illustrate how discussions of risk in decision documents and other design-phase documents can fail to meet the criteria for do no harm:

- **Indicating that there is no risk at all to the - cutting issue:** in some cases the documents suggested that the focus of the project meant that there were no risks to cross-cutting issues. For example, one reviewed decision document indicated that it was not applicable to conduct a risk assessment for human rights, because "the project supports gender mainstreaming and empowerment, making it fully aligned with the human rights theme". Another decision document simply indicated that "The project will not have any negative impact on the environment".

- **Indicating that no risk assessment had been carried out:** in some cases, the decision documents indicated that no risk analysis relating to potential negative impacts on cross-cutting issues had been carried out. For example, one document stated that "the possibility of negative programme-related impacts on climate change and the environment are not considered". Another stated that the possibility of programme-related "discrimination against particular marginalised groups (persons with disabilities, ethnic minorities), is not considered explicitly". The presence of agreement numbers in these documents indicates that they related to programmes that did receive funding, despite the absence of a thorough risk assessment relating to these cross-cutting issues.

- **Reporting only on risks from cross-cutting issues to project, rather than risks to cross-cutting issues from the project:** in other cases, decision documents misinterpreted the 'do no harm' requirement by reporting on risks to programme implementation that were driven by cross-cutting issues. For example, one decision document reported that "negative changes in the climate and environmental degradation may bring about significant natural hazards... ...This may greatly affect attendance of children in schools, hence, poor learning outcomes". Another stated that a "national context risk is the risk of fraud and corruption. This is a well-known risk for [agreement partner] and everyone working with development issues". In both examples, the potential for the programme to have a negative impact on these cross-cutting issues (climate change and environment, and anti-corruption) was not discussed.

- **Broad and intangible discussions of risk and mitigations:** in some cases, design documents discussed cross-cutting issues broadly, but failed to identify specific, tangible risks associated with the implementation of the project. For example, one document vaguely mentioned "attitudes and behaviour in the disposal of household waste" as a risk, but didn't not specify whether or how this was related to the implementation of the project. In other cases, there were discussions relating to implementation measures that were relevant to cross-cutting issues and which implied a risk, but without explicitly identifying the risk. For example, one document stated that the programme would take "measures to promote women's participation, and ensure that both women and men have equal opportunities in participation in community engagement activities, training, workshops, and advocacy campaigns". It did not, however, link this broad mitigation measure to any specific risk associated with the exclusion of women from programme activities. Similarly, another project document indicated in its risk assessment section that the agreement partner "uses a Rights Based Approach through securing the rights of women as rights holders, while at the same time supporting Governments and other institutions that provide services to be accountable duty bearers". Again, the adoption of a rights-based approach, which could be conceived as a broad mitigation measure or safeguard, was not linked to any tangible risks posed by the projects to either human rights broadly, or women's rights more specifically.

The rules for proactive implementation

For a document to be classified as 'proactive' it had to identify (for design documents) or report on (for follow-up phase documents) at least one **cross-cutting issue specific objective within its results chain.** The NLP model used several rules to identify 'proactive' implementation in design-phase documentation. For a document to be categorised as 'proactive' by the model, it had to satisfy at least one of the following rules:

Design phase rules:

- **Rule 1:** The design document articulates an anticipated positive result relating to a cross-cutting issue in its overall project description/impact statement
- **Rule 2:** The design document articulates a specific *outcome* relating to a cross-cutting issue in a results matrix
- **Rule 3:** The design document contains prose which includes declarative language around promoting positive outcomes relating to a cross-cutting issue
- **Rule 4:** The design document articulates a specific *output* relating to a cross-cutting issue in a results matrix
- **Rule 5:** The design document contains tangible, measurable (including numeric identifiers) outputs relating to a cross-cutting issue in a results matrix

Follow up phase rules

- **Rule 6:** cross-cutting issue outcomes and proactive past tense identifiers (i.e. covered, developed, delivered etc.) are mentioned in close proximity

- **Rule 7:** [progress report specific] tangible cross-cutting issue-specific actions (planned or completed] are described in the section of the template requiring a "brief description of the project's effects on gender equality, human rights, and the environment and climate change so far

- **Rule 8:** prose suggesting a beneficial outcome (past tense or declarative future tense) in relation to a specific cross-cutting issue

- **Rule 9:** past tense analysis of a cross-cutting issue featuring quantifiable outputs/results.

How did documents meet the criteria for proactive implementation?

The model's results indicate that proactive measures to address cross-cutting issues are considerably more common in project documentation than risk analyses that fulfil the 'do-no-harm' criteria. To some extent, this may be attributable to the difficulty in implementing a 'do-no-harm' approach in the correct manner. As the examples highlighted, the 'do-no-harm' requirement is often misinterpreted by grant managers/project implementers, and thorough assessments of risks to cross-cutting issues are often not conducted due to time and capacity limitations. By comparison, proactive measures to address cross-cutting issues are more straightforward to interpret. Indeed, given the broad scope of the cross-cutting issues, it is likely that many projects will contain measures that will satisfy the proactive rules.

There are various examples of proactive integration at various levels of the results chain[10]. Examples of proactive text in the sampled documents included:

- **Quantifiable targets/results:** for example, one document highlighted an ambition to scale up forest conservation to increase forest cover by 30%. Another quantifiable target/result for a different project was the "establishment of a tree nursery with a capacity of more than 4000 tree seedlings to be distributed".

- **Output statements:** for example, one document included a planned result to "support women homebased workers and excluded groups' access to social security benefits, legal support in labour courts" etc. Another identified an output relating to 'improved gender-friendly ecotourism infrastructure.. ..including with private sector engagement".

- **Broader outcome statements:** for example, one document indicated that the project intended to contribute to ensuring that China's "commodity markets are transformed so that they do not contribute to deforestation and conversion in South America and Southeast Asia".

[10] Note that in post-processing, agreements which had a cross-cutting issue as a primary objective were filtered out, so that these were not assessed for proactive implementation. For example, climate change projects were not assessed for proactive measures to promote climate and environment. They were, however, assessed for their treatment of risks to climate and environment.

Thematic analysis

This section provides an analysis of the text surrounding instances when the above rules are flagged. The aim is to provide some insights into the themes being discussed in relation to cross-cutting issues in project documentation.

Common non-thematic keywords

In this context, non-thematic keywords refer to keywords that appeared in passages discussing cross-cutting issues, but which were generic as opposed to thematically linked to a specific cross-cutting issue. These could be verbs, such as 'support', 'improve', or they could be generic nouns or adjectives that have plausible associations with multiple thematic issues, such as 'training', or 'local'.

Across all four cross-cutting issues, **awareness** and **capacity** were among the most common keywords that were not related to the theme of a specific cross-cutting issue, appearing in passages that satisfied model rules**.** This strongly suggests that where project documentation does discuss cross-cutting issues, it is often in relation to raising awareness of cross-cutting issue-related issues, or strengthening the capacity of organisations and other groups to address these.

Policy was another non-thematic keyword that appeared very frequently in rule-flagging passages for climate change and environment , anti-corruption, and women's rights and gender equality. From a proactive perspective, this could imply that efforts to proactively address these issues frequently focus on influencing, improving, and/or developing policies. From a do-no-harm perspective, the high prevalence of this key word is likely to indicate frequent references to (or ambitions to develop) policies to ensure that risks to cross-cutting issues are mitigated.

For passages that flagged rules for human rights, women's rights and gender equality, and climate change and environment, **protection** was a non-thematic keyword that appeared frequently. This is an intuitive finding, suggesting that efforts to address these cross-cutting issues often explicitly focus on the protection of rights (whether women's rights, or human rights more broadly), or environmental protection (for the climate change and environment cutting area). **Community** was a non-thematic keyword that appeared frequently for rules flagged for human rights and women's rights and gender equality. This may suggest that efforts to address this issue (whether proactively or by reducing risks) are frequently focused at the community level.

Keywords signifying actions that appeared particularly frequently across all cross-cutting issues included:

- Strengthen
- Raising
- Support
- Improve

Notably, these action keywords are logically linked to many of the frequently appearing nouns, including awareness, capacity, protection, and policy.

Common thematic keywords

In this context, thematic keywords are those that appear in documents discussing cross-cutting issues, and which are clearly thematically linked to that issue. Analysis of the frequency of these

keywords provides insights into the focus of efforts to address each cross-cutting issue. The five most common thematic keywords for each cross cutting issue were:

- **Women's rights and gender equality:** 'women', 'gender equality', 'enrolment/enrolment, 'girls', and 'sexual'. The presence of 'girls' and 'enrolment' on this list implies a strong focus on education (i.e. girls education/enrolment in schools) in efforts to address women's rights and gender equality. The presence of 'sexual' on this list may imply a strong focus either on sexual and reproductive health, or on the prevention/treatment of sexual violence.

- **Climate change and environment:** 'deforestation', 'environment', 'emission', 'climate', 'conservation'. The presence of 'deforestation' here indicates a strong focus on the forestry sector. It is possible that the presence of 'emission' and 'conservation' on this list is linked to the focus on deforestation (i.e. reducing emissions through forest conservation by addressing deforestation').

- **Human rights:** 'human rights', 'children', 'with disabilities', 'activists', 'access...service'. This list indicates a focus on different themes, including children's rights, and the rights of people with disabilities. Notably, combined with the fact that 'protection' is included as one of the most frequent non-thematic keywords in passages flagged for human-rights rules, this indicates a focus on the protection of the rights of children and people with disabilities. The inclusion of the word 'activists' indicates that the focus is often on supporting or protecting human rights activists. The presence of 'access.. ...service' suggests a focus on projects that aim to improve (or remove barriers to) access to services relating to human rights.

- **Anti-corruption:** 'fraud', 'anti corruption', 'financial management', 'accountab...(accountable, accountably, accountability etc.)', 'CSOs'. Most of the key words on this list are unsurprising, and don't imply a specific focus within the broader theme of 'anti-corruption'. The inclusion of 'CSOs' on this list is interesting, however, and may suggest a focus on support to CSOs engaged in efforts to tackle corruption, or support to CSOs to tackle corruption.

An analysis of the proportion of 'do no harm' and 'proactive' documents[11] featuring these key words over time was conducted. This can also be compared with the wider sample of documents, including those which failed to meet the criteria for 'do no harm' and 'proactive'. It is important to compare to the benchmark across all documents to ensure that any trends are a feature of document quality, rather than exogenous factors. For example, we might expect instances of the keyword "COVID-19" to increase dramatically over the period, which is due to external factors rather than implementation quality. If keyword usage increases faster (or declines slower) in high-quality implementation documents, we can be confidence that these keywords are associated with the higher quality, which may be because they capture emergent trends in the field that indicate quality implementation. Conversely, if the increase is slower (or the decline faster) in 'do no harm' and 'proactive' documents than it is in all documents, this implies that documents discussing the keyword are of a proportionately lower implementation quality: i.e. the keyword is mentioned, but not in a context which triggers the do no harm or proactive rules.

[11] This means documents that triggered at least one rule for 'do no harm' or 'proactive'

The most notable trends that emerged from this analysis were:

- **Climate Change and Environment:** between 2018 and 2023 there was a decline in the number of 'do no harm' and 'proactive' documents mentioning the term 'deforestation'. This decline was more muted among all documents. This implies that 'deforestation' is becoming less relevant to implementation quality. This may be because differential terms with greater nuance and situational relevance have replaced the umbrella term of 'deforestation' in high-quality documents, for example references to specific practices or forestry protection programmes.

- **Women's Rights and Gender Equality:** overall there are no major changes between 2018 and 2023 in the frequency of terms related to this cross-cutting issue. There was a slight increase in the overall frequency of each of the five most frequent key words for this cross-cutting issue. This trend was most pronounced for the term 'sexual'.

- **Anti-corruption:** the proportion of 'do no harm' and 'proactive' documents mentioning "CSOs" increased slightly between 2018 and 2023. Among all documents, however (i.e. including those that were not categorised as 'do no harm' or 'proactive'), the frequency of the term "CSO" fell over the same period. This implies that discussions of CSOs are correlated with higher implementation quality, and therefore a higher likelihood of meeting either 'do no harm' or proactive criteria.

- **Human Rights:** the terms "human rights" and "children" declined in frequency in both sets of documents between 2018 and 2023. Meanwhile, the frequency of the phrase "with disabilities" rose slightly during the same period in both document sets. The phrase "access... ..service" appears in over 10% of all documents, but only in 0.5% of documents classified as 'do no harm' or 'proactive'. This indicates that most documents that discuss access to services are not doing so in contexts that trigger the 'do no harm' or the 'proactive' rules relating to human rights.

Thematic Analysis of 'Premium' Documents using Latent Dirichlet Allocation

There were 3,603 documents in the sample that were either do no harm or proactive for three or more cross-cutting issues, out of a possible total of eight (do no harm and proactive for all four cross-cutting issues). These are referred to as 'premium' documents. This subset of documents was analysed using Latent Dirichlet Allocation (LDA), a machine-learning technique that is used to identify the underlying topics in a collection of documents, by automatically grouping together words that frequently appear together in similar contexts across different documents. Each topic is represented by a set of words, and each document is a mix of these different topics.

Using LDA, we can arrive at the following three outputs:

- **Topic words:** each topic is associated with a set of words that are most representative of the topic (i.e. the topic 'sports' might include topic words such as 'game', 'team', and 'score').
- **Document topics:** each document is broken down into a mixture of topics, to determine what proportion of each document is related to each topic (i.e. a document might be 70% about politics and 30% about economics)

- **Topic Distribution:** this considers how many documents are primarily about each topic, creating a picture of which topics are more prevalent in the data set.

Determining how many topics to identify through LDA is not always straightforward, and there is not always a one-size-fits-all answer. Selecting too few topics can result in topics that are overly broad, while choosing too many can produce results that are too granular for meaningful analysis. The trick is therefore to strike the right balance between 'high-level' and 'granular'. One common approach to doing this is calculating a coherence score, which measures how consistently the words within each topic co-occur in the documents. By comparing coherence scores for different numbers of topics, it is possible to choose the number of topics that makes the most sense for the data in question. This approach was applied to the 3,603 premium documents in the sample, and the best fitting number of topics was 8.

These eight topics represent the topics that were most frequently discussed in this subset of premium documents. This provides us with insights as to the type of themes that the best performing topics were discussing. In descending order of prevalence (i.e. the number of documents discussing the topic), these eight topics were:

- **Women, peace and security:** key words associated with this topic were "women", "peace", "security", "civil", "programme", and "rights". The finding that this topic features heavily among the 'premium documents' in the dataset strongly aligns with the evaluations' broader finding that the "governance, civil society, and conflict prevention" target area was the best performing in terms of 'do no harm' implementation for all cross-cutting issues apart from climate change and environment. The evaluation also found that this target area performed well for proactive implementation of cross cutting issues.

- **Agriculture:** key words associated with this topic were "agriculture", "farmer", "food", "rural", and "market. That agriculture is among the most prominent topics discussed in the set of 'premium' documents may indicate that cross-cutting issues are well implemented in this area.

- **Women and education:** the top key words associated with this topic were "girl", "edu(cate/ation)", "school", "service", "learn", "child", and "HIV". The prominence of this topic among premium documents also aligns strongly with the evaluation's finding that education was the best performing target area for proactive implementation for all cross-cutting issues. In particular, more than half of the agreements in the sample that were within the education target area included proactive measures to address women's rights and gender equality, and human rights. Education was also among the top performing target areas for 'do no harm' implementation.

- **Climate finance:** the top key words represented in this topic was "bank(ing)", "invest(ment)", "energies", "climate", "green", "resili(ant/ence)".

- **Forestry:** the top key words represented in this topic were "forest", "land", "indigenous", "environment", "REDD", "conserv(e/ation)", "natur(e/al)". Again, this was not a target sector analysed in the evaluation, but the topic's presence among the premium documents indicates that the area performs well in terms of cross-cutting issue implementation.

- **Budgets and grant management:** the top key words represented in this topic were "requirement, "account", "cost", "grant", and "office". It is not surprising that this topic features prominently in the premium documents. Information relating to budgets and grant management is a standard component of design and reporting documentation.

- **Monitoring and Evaluation:** the top key words represented in this topic were "evalu(ate/ation)", "programme(e/ing)", "review", "analysing", "Norad", and "learn". Again, it is not surprising that the topic of monitoring and evaluation appears frequently in premium documents. This is likely to reflect the inclusion of results matrices and monitoring and evaluation plans in design phase documentation, and structured results reporting in follow up-phase documentation.

- **Energy infrastructure:** the top key words represented in this topic were "cost", "construct(ing/ion)", "water", "energies", "power", "bank", "environment", "plant". Note that the evaluation found that the 'energy and environment' target area performed reasonably well in terms of cross-cutting issue integration, in particular with respect to 'do no harm' in design phase documentation for the climate and environment issue, where it out-performed other target areas.

6. Conclusion

The use of NLP to assess cross-cutting issue implementation in project documentation illustrates one potential application of the method to evaluation processes. In this case, NLP was used to conduct a systematic review of a very large sample of documentation, in search of something highly specific. Given the circumstances, this task would have been too laborious for humans to undertake efficiently in the same manner. A human-driven approach would, at the very least, have required a far smaller sample of documents alongside concerted efforts to ensure inter-coder reliability[12]. NLP is therefore a method that is worth considering for other evaluative exercises which would benefit from similar systematic reviews of large bodies of documentation. This is, of course, not the only potential application of NLP in evaluations. There is a large and growing body of NLP tools available to evaluators, which can be applied to numerous different tasks, including summarising, translating, and even generating text. Given the expanding capabilities of NLP-supported methods, it is likely that their usage in evaluations will increase over the coming years.

As discussed in this report, however, the application of NLP (and in particular, the development of custom NLP models), can be very challenging. It is crucial, therefore, that evaluators carefully consider the benefits of the method, and weigh these against potential challenges and risks, the evidence needs of the evaluation, and the resources and time that are available. In addition, there are a set of pre-conditions which, if in place, make NLP more suitable for integration within evaluation processes. These include:

- **Clarity of purpose:** this means that there should be a clear idea of what needs to be looked for in the available documentation/data, and an understanding of what it looks like (i.e. criteria or guidelines for identifying it). Ideally, there should also be some examples of what is being looked for, although if these are not available they can be identified at an early stage in the process. There should also be clarity and agreement about what the expected outputs of the NLP process are, and how these will fit into the wider evidence needs of the evaluation.

- **A well-structured archive/document library:** sometimes, an unstructured or minimally structured archive/document library is all that is available to work with. This does not mean that NLP techniques cannot be applied, but it does mean that the set-up time for process will be longer. NLP will be more efficient when there is a well-structured archive/document library to work with. This means that important characteristics of the documents within the archive/document library should be easy to identify from their positioning within its structure and/or their naming. For this evaluation, for example, the NLP process would have been considerably more efficient if the structure of the data dump had made it easy to associate documents with their parent agreements (i.e. if documents were filed by their associated agreement number), and if naming conventions allowed for a straightforward identification of document type (i.e. decision document, progress report, final report etc. were clearly labelled as such in their file names).

[12] Meaning that all humans involved in the task would be classifying documents in the same way

In addition to the lessons discussed above that related to specific steps in the NLP process implemented for this evaluation, there are a number of broader lessons that evaluation practitioners would benefit from considering when deciding whether to use of NLP in future projects.

Assessing the need for data security

The level of data security required will influence what NLP tools can be used for the evaluation. If a high level of security is necessary, for example if the documents for processing are confidential and/or governed by strict access agreements, then the available suite of tools for NLP is more restrictive. In such cases, it is unlikely that the evaluation team will be able to make use of powerful online models, which are highly efficient and effective at supporting sophisticated queries about the content of document libraries/archives. This is because although such platforms have security arrangements in place, they nevertheless tend to involve third-party processing or storage of the underlying data, and are therefore vulnerable to the same type of threats that face other online platforms. When data security is a major issue therefore, custom-built models which can be implemented locally are a good alternative. These can also perform sophisticated queries on documents, but are more time consuming and resource intensive to develop. If the underlying data is not confidential or sensitive (i.e. published evaluation/research reports), then it is worth considering the use of large commercial NLP platforms.

Integration with the wider evaluation process

NLP is a relatively new method in the field of evaluation, and one that requires a specific skill set to understand and implement. Because of this, it is likely that many evaluators who are accustomed to more traditional methods may find it challenging to engage meaningfully with the method. This raises the risk that an NLP process, when included as part of an evaluation design, becomes a siloed and isolated method. If this happens, there is a greater likelihood of divergence between the output of the NLP process and the evidence needs and conceptual framing of the evaluation. For this reason, if the decision is made to use NLP, it is crucial that measures are taken to ensure that non-machine learning experts are involved at all possible steps in the design and implementation of the method.

There are numerous opportunities for interaction between thematic and machine-learning experts during the design of the NLP process, which can help to improve both its relevance to the wider evaluation as well as the validity of its results. Each step of the process discussed in this report benefitted from inputs by thematic and sector experts. During sampling, inputs focused primarily on efforts to identify the most relevant documents to include for analysis. This was also informed by relevant documentation from Norad and MFA, as well as by interviews with staff responsible for overseeing grant management.

Another key area for input by thematic and sectoral experts involves determining 'what to look for' and 'how' using NLP. This requires clear definition of concepts, aligned with the approach adopted by the broader evaluation, as well as the development of a valid, reliable, and realistic approach for measuring these in the relevant documentation. Input from thematic experts helps to ensure that the definition of concepts adopted by the NLP process is consistent with that adopted by the wider evaluation, and that the approach to measurement adequately captures the various dimensions of these concepts. Interaction between machine learning specialists and thematic experts can also help to ensure that the proposed measurement approach is realistic. This is important because although

many of the concepts that evaluations seek to assess are highly complex and multidimensional, approaches to measurement need to be implementable in a machine-learning environment. This necessarily involves some degree of simplification. Because of this, it is crucial that machine learning processes are not used in isolation. Instead, they should be deployed alongside other more traditional methods, which are better-suited to capturing the full complexity of the issues being addressed.

Expectation management

There is great optimism among evaluation practitioners about the potential applications of machine learning and AI-supported methods. While these methods can indeed be applied to support the efficiency, scope, and quality of evaluations, it is important to bear in mind that they are not a silver bullet, and that they do not alter the fundamental and challenging nature of evaluation. It is therefore important to manage expectations about what these methods can and cannot deliver. When applied appropriately, they can provide useful insights into the questions that evaluators seek to answer. They cannot answer everything, however, and should be used in combination with more conventional evaluation and research methods. When designing an evaluation which includes NLP or other AI-supported processes, it is important to determine exactly why they are used, clearly define the expected outputs, and how these fit into the wider process.

Explainability

Because NLP and other AI-supported methods are relatively new and difficult to understand, it is important to be aware that there may also be scepticism among stakeholders about their place in evaluation processes. Some stakeholders may question the validity or accuracy of the results that these methods can produce. There may also be concerns about the transparency or impartiality of their results. These concerns are entirely legitimate; NLP and other AI-supported methods should be held to the same levels of scrutiny as other evaluation methods. Evaluation teams that use these methods should therefore prepare to invest time and resources in clearly communicating to stakeholders how and why they are being used, and should aim to ensure that the process is as well documented and as transparent as possible. Importantly, the degree of transparency that is possible is also influenced by the specific tools that are used. The inner workings of large, commercially available NLP platforms are often guarded as intellectual property. This limits the ability of users to explain exactly how data is being stored and analysed, thereby reducing the explainability of results. Explainability is one of the advantages of custom-built, locally implemented models. Although they are difficult to build and slower than their large commercial counterparts, they enable transparency of process by allowing their developers to communicate how they arrived at their results.

Department For Evaluation

evaluation@norad.no
norad.no/evaluation