Artificial Intelligence and Machine Learning Methods for Programme Evaluations in Global Affairs Canada

Literature Review

Paul Jasper, Søren Vester Haldrup, Slava Jankin Mikhaylov

March 2019

Oxford Policy Management

the Data atelier

About Oxford Policy Management

Oxford Policy Management is committed to helping low- and middle-income countries achieve growth and reduce poverty and disadvantage through public policy reform.

We seek to bring about lasting positive change using analytical and practical policy expertise. Through our global network of offices, we work in partnership with national decision makers to research, design, implement, and evaluate impactful public policy.

We work in all areas of social and economic policy and governance, including health, finance, education, climate change, and public sector management. We draw on our local and international sector experts to provide the very best evidence-based support.

Executive Summary

The International Assistance Evaluation Division (PRA) of Global Affairs Canada (GAC) has commissioned Oxford Policy Management (OPM) to carry out a literature review that explores the use of text analytics, Artificial Intelligence (AI), and Machine Learning techniques, including Natural Language Processing (NLP), to improve the process with which GAC implements programme evaluations.

To carry out this task, OPM reviewed the current process whereby GAC implements programme evaluations and identified three main steps where modern data science methods could most usefully be applied: i) the selection of projects to be included in programme evaluations ii) the selection of documents to be included in the review of the identified projects, and iii) the coding and analysis of document reviews. In addition, text analytics could help GAC teams at the implementation stage of projects.

A range of different NLP approaches could be applied in these three stages of GAC programme evaluations:

Project Selection: Using Topic Modelling

Topic modelling is a Machine Learning based NLP method that can be used to automatically identify unobserved themes or topics in large sets of text data and to label documents accordingly. GAC selects projects for programme evaluations in part based on which thematic focus areas they fall under. Evaluators currently read project documentation and manually extract information required to identify the thematic focus of a project. Hence, topic modelling can be used to partly automate and improve this process.

Document Selection: Using Information Retrieval Methods

Information Retrieval methods can be used to automatically find material (e.g. documents, names, or words) in large sets of text data. This covers a wide range of approaches. Search query systems that allow for logical operators in text databases are one example. Another example is “Word2Vec”, a method that uses Machine Learning to estimate the meaning of words based on how they co-occur with other words. Finally, Named Entity Recognition (NER) is a Machine Learning based approach that can be used to quickly label large volumes of text by whether known named entities (e.g. certain organisations or individuals) appear in them or not. In GAC’s programme evaluation process, evaluators usually select a pre-defined set of documents for review. In this context, GAC can use Information Retrieval techniques to improve the speed and ease with which documents are selected in programme evaluations, and to help avoid that documents with crucial information are overlooked.

Document Review: Sentiment Analysis

Sentiment analysis is a Machine Learning method that can be used to attach an emotional label to text. It is based on pre-defined dictionaries of words that are associated with positive, neutral, or negative emotions and can be applied to large amounts of text in situations where the equivalent task would take a long time for humans to complete. GAC evaluation officers can use this technique to identify sentiments expressed towards a particular project, hence helping GAC to identify implementation challenges, likely impact, and relevance (including contextual fit) of an initiative.

Document Review: Combining Methods

Generally, many applications of NLP involve making use of several methods in sequence or at the same time. Document reviews in the context of GAC programme evaluations would likely benefit from such a combination of methods. For example, topic modelling can help identify topics or themes covered in certain documents, and correlations between sentiments and topics appearing in documents can be used to identify topics likely to be associated with positive or negative outcomes. Furthermore, NER can help evaluators quickly find paragraphs in many different documents that refer to similar organisations or entities. Indeed, the true strength of different NLP methods materialises once these are applied together to a machine-readable database that comprises text found in GAC project documentation.

Requirements and prerequisites

Applying NLP and other Machine Learning techniques entails a set of technical requirements related to data pre-processing, data management, and data analysis. Most of these requirements can be met using free and open-source software. In addition to technical requirements, there are a number of “analogue” prerequisites for successful application of Machine Learning methods, including creating a diverse and interdisciplinary team to implement Machine Learning pilots, accounting for and adjusting to existing GAC programme evaluation processes, and taking the wider policy and regulatory environment into account.

Piloting and implementation

This review is the starting point of a longer process of trialling and, if successful, using Machine Learning and NLP methods in GAC programme evaluations. The initial process of piloting and implementation could proceed in four steps: i) an in-depth feasibility assessment, ii) appointment of an interdisciplinary team to lead the piloting process; iii) an agile “design thinking” inspired implementation process involving testing, learning, and iteration, and iv) the development of a plan to ensure sustainability and uptake within GAC.

Table of Contents

List of Tables, Figures, and Boxes

List of Abbreviations

AI	Artificial Intelligence
CTM	Correlated Topic Model
DAC	Development Assistance Committee
GAC	Global Affairs Canada
GDPR	General Data Protection Regulation
GE	Gender Equality
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MNCH	Maternal, New-born, and Child Health
MSR	Management Summary Report
MVP	Minimum Viable Product
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
OECD	Organisation for Economic Co-operation and Development
OPM	Oxford Policy Management
PIP	Project Information Profile
PMF	Performance Measurement Framework
PRA	Assistance Evaluation Division
PSLA	Probabilistic Latent Semantic Analysis
UNDP	United Nations Development Programme

1 Introduction

The International Assistance Evaluation Division (PRA) of Global Affairs Canada (GAC) has commissioned Oxford Policy Management (OPM) to carry out a literature review that explores the use of text analytics, Artificial Intelligence (AI), and Machine Learning techniques to improve the process with which GAC selects projects and implements programme evaluations.

The review was implemented in two steps: first, the team from OPM engaged with GAC and PRA to better understand how programme evaluations are currently being done. No interviews were carried out in this process. Instead, the team shared a list of questions with GAC and PRA to which written answers were provided. In addition, the team reviewed documentation and examples of evaluation steps shared by GAC.

In a second step, the team reviewed and summarised a range of literature (journal articles, grey literature, blog posts, and web sites) on Machine Learning, text analytics, and their application in practice. The team particularly focussed on reviewing examples related to applications in public policy. The objective was to provide answers to three broad questions:

> 1) What AI, Machine Learning, and text analysis methods could help to improve the way in which programme evaluations are implemented in GAC?
>
> 2) What are technical, logistical, and other requirements for successfully trialling and implementing these methods in the context of GAC programme evaluations?
>
> 3) What would a pilot and implementation plan for such a project look like?

The following sections answer these questions. Section 2 summarises how programme evaluations in GAC are implemented. Section 3 provides an overview of Machine Learning and Natural Language Processing (NLP) methods. Section 4 describes some of these methods in detail and discusses how they may be used in programme evaluations in GAC. Section 5 discusses requirements for successful implementation of these methods and describes what a piloting and implementation plan could look like.

A note on terminology: AI vs. Machine Learning vs. Natural Language Processing

Even though the title of this document refers to AI, we will not be using the term much in this review. The reason is that we consider Machine Learning to comprise a set of methods that can be used to build AI, in the sense of intelligent systems, rather than the two being alternative or interchangeable approaches. As such, Machine Learning may be seen as methods that can help to analyse data in smart ways, which in turn can help to build intelligent systems (AI). This review focusses on these Machine Learning methods, rather than on AI systems.[1]

[1] USAID (2018), Reflecting the Past, Shaping the Future: Making AI Work for International Development, Center for Digital Development USAID. See also Google Machine Learning Services, *What is machine learning*.

In contrast, we use the term **Natural Language Processing (NLP)** repeatedly (for details see section 3). In short, NLP can be used as an overarching term for any method employed to process human language (e.g. text) using computers. Some of these methods draw on Machine Learning. Because much of the programme evaluation work at GAC involves processing text, NLP is of particular importance. We may use “text analytics” as a shorthand for NLP and any quantitative analysis implemented on text data.

Finally, we use the terms **algorithms, models, and estimation procedures** interchangeably in this review. These terms all refer to statistical or mathematical procedures that are applied to data in order to perform some type of analysis, with the objective of estimating target values. The nature of these target values can vary greatly: they can be topics in a document, predicted contents of documents, average number of words, or coefficients that describe the relationship between different variables.

2 Programme Evaluations in GAC: an Overview

The overall objective of this review is to provide GAC and PRA with a description of the potential benefits, caveats, and practical implications of employing modern statistical and text analysis methods – such as Natural Language Processing and Machine Learning – in the context of PRA programme evaluations. PRA programme evaluations are based on reviews of large sets of project information and documentation, i.e. project reports and related documentation that describe the implementation and results of projects that were implemented under programmes funded by GAC. Figure 1 below depicts a simplified version of our understanding of the information flow around project implementation and evaluation of GAC funded programmes.

We understand GAC funded programmes to be strategic funding facilities that focus on specific themes and can span a substantial number of projects, across a large set of countries, several years of implementation, and involve a significant amount of funds. For example, the Canadian Maternal, New-born, and Child Health (MNCH) programming (MNCH 1.0 and MNCH 2.0) covers almost a decade of funding (from 2010 to 2018) across over 900 projects, with funds of over $3 billion CAD. This means that, within one GAC programme, a large variety of different projects that focus on certain sub-themes will be funded.

We structure the life of these programmes and projects into two large phases: the **implementation phase** and the **programme evaluation phase**. These two phases modulate how information is produced and used in the GAC programme implementation and evaluation process. In very general terms, such information can be stored using two different types of data: structured and unstructured data. (See Box 1 below.)

Box 1: Structured and Unstructured Data

The difference between unstructured and structured data is fundamentally a difference in how data is stored and hence how analytical software can 'read' and analyse data:

- **Structured data** generally refers to data that is stored in a table format, such as Excel. The rows and columns in the table refer to observations and variables. Such datasets can easily be processed by analytical software. For example, it is very easy to calculate the mean of a variable (column) in Excel, or filter such a calculation by certain observations (i.e. only taking into account certain rows).
- **Unstructured data** – in contrast – is data that does not come in this type of neat format. Examples of such data are electronically stored images, texts, or videos. These data are unstructured because – without any processing – variables and observations are not clearly defined. The processing of unstructured data can change this. For example, a document that contains several sections can be seen as an agglomeration of text (no structure). With some text analytics it is possible to count the times with which certain words appear in each section of the document. This can be transferred into a structured database, where each row is a section of the document and each variable corresponds to a word. Each cell in this table would then represent the number of times that a word appears in each section of the document. This structured data can then easily be analysed, e.g. by calculating the average number of times a certain word appears across the sections of the document.

Many of the methods that this review will touch upon can be applied to unstructured data and can help to provide structure to such datasets.

The way in which project information is handled in the two phases in the life of GAC projects and programmes can roughly be summarised as follows:

- **In the implementation phase**, information and data are collected about projects. It is important to note that in Figure 1 below, this implementation phase can apply to several projects at the same time. In addition, such projects can run in parallel but with shifted timelines (i.e. some projects might be at the start, while others might already have ended).
- **In the evaluation phase**, data from projects funded within the framework of a programme are analysed for evaluation purposes. This means that project data from a large set of projects funded by programmes will have to be compiled and analysed.

Based on information received from GAC and the PRA team, the following paragraphs describe each of these phases in more detail.

2.1 Implementation Phase: Producing and Collecting Data

For programmes, the implementation phase of an initiative or thematic priority comprises a large set of projects, their implementation, and their individual project-level evaluations. The paragraphs below describe this for a single project, though in practice these stages can occur for several projects at the same time and in parallel.

There are three separate stages in this phase:

First, the start of a project, in which a project officer in GAC processes a project proposal. Two key information processing actions take place in this phase:

- A project officer in GAC, i.e. a GAC team member responsible for managing and reviewing project proposals, transfers information from the proposal into a structured database that is held by the group responsible for department statistics in GAC ("CFO-Stats"). This database includes pre-defined requirements of information about projects on OECD DAC sector codes, budgets, countries covered, and other specific indicators that describe a project.

- In addition, the proposal itself is stored as a text document in a central location. This proposal contains additional information about the project that has not been coded by the project officer into a database, such as information on GAC internal thematic focus areas that the project falls under. These thematic areas can relate to topics such as strengthening health systems, prevention and treatment of diseases, or enhanced nutritional practices of mothers, new-borns, and children. In itself, as a text document, this proposal represents a set of unstructured text data about the project.

Second, the implementation stage of a project, in which regular progress and review reports are produced by the implementing agency. These documents serve the purpose of informing GAC about the progress and achievements of the project. These reports cover a wide variety of document types such as end of year project reports, evaluations, performance measurement frameworks (PMFs) and project information profiles (PIPs). These reports are not processed further into databases, but constitute a set of centrally stored unstructured text data about the project.

Third, the end stage of a project, in which GAC produces and/or reviews project-end reports, final evaluation reports, or final reports about the project. These documents provide a final summary of project implementation and the results achieved. Again, these reports are not further processed into a database, but constitute a set of unstructured text data that relates to the project.

2.2 Programme Evaluation Phase: Data Analysis

The programme evaluation phase starts towards the end of GAC funded programmes. Programme evaluations can be both summative and formative, which means that their objective is either to assess whether programmes achieved their objectives (e.g. for accountability purposes) or to derive lessons that can help to improve the implementation of similar programmes in the future. In both cases, the goal is to answer specific evaluation questions that allow for statements about how the programming worked, whether outcomes were achieved, or to derive lessons for the improvement of future programming.

Such programme evaluations are currently implemented in four key stages:

First, project selection, in which evaluators need to decide which projects to include in the evaluation. As discussed above, GAC programmes fund a very large number of projects, and not all of these can be included in an evaluation. For instance, by one estimate provided by GAC, reviewing all project documentation related to GAC's most recent MNCH programming (see beginning of section 2 above) would take over 1,000 years of one reviewer's time. Pre-selecting projects is therefore an important step in any GAC programme evaluation.

Currently, projects are purposefully selected based on a variety of different criteria, such as e.g. budget size and thematic focus of a project. Data used to assess projects against these criteria is stored in two places:

- **In the structured database held by the CFO Stats team into which project officers have previously coded information from projects**. Criteria that can be assessed with data from this database relate to indicators stored here by project officers (see Step 1 under "Implementation phase" above), such as country of implementation, OECD DAC sector codes, and project budgets. For example, the PRA might want to only include projects from certain focus countries and with budgets that exceed a certain amount in its evaluation.

- **In project documentation, with a focus on project proposals.** Criteria that can be assessed with information from these proposals relate mainly to thematic focus areas (or thematic pillars) of projects. These areas have not been pre-coded by project officers. There is therefore no structured database available that covers these thematic areas. This means that evaluators need to read project documentation, in particular proposals, and manually extract information required to identify the thematic focus of a certain project. Evaluators subsequently decide which projects to include in the programme evaluation, in part based on these thematic areas. The thematic areas can vary from one evaluation to another.

In the case of the MNCH programme evaluation, evaluators used information from both of these sources to purposefully select a set of 73 out of over 900 projects to be included in the evaluation.

Second, document selection. After having selected projects, evaluators select a subset of documents from each project that will be included in the evaluation. Given the large volume of text documents available across projects, the GAC evaluation team does not review all documents prior to document selection. Instead, they select documents based on pre-defined expectations of what kind of documentation are of relevance to the evaluation. The pre-defined list of project documentation assumed relevant includes:

> "*Evaluation reports, final project reports, year-end reports, PMF (performance measurement frameworks), PIP (project information profiles), PAD (project approval documents), MSRs (management summary reports), GE assessments (gender equality assessments), contracting agreements, or other documents describing project design, implementation, or results*." (From documentation shared by GAC.)

Third, document review, in which evaluators review the documents identified to be included from selected projects. Reviewing these documents entails reading them and manually extracting information that is of relevance to the evaluation. 'Relevance to the evaluation' means that information is useful to answer the evaluation questions set out at the beginning of the review. The document review is implemented in two steps. Both steps involve review and analysis of unstructured text data and extraction of information from them. This is done by the evaluation team responsible for the programme evaluation and requires significant time investment from team members.

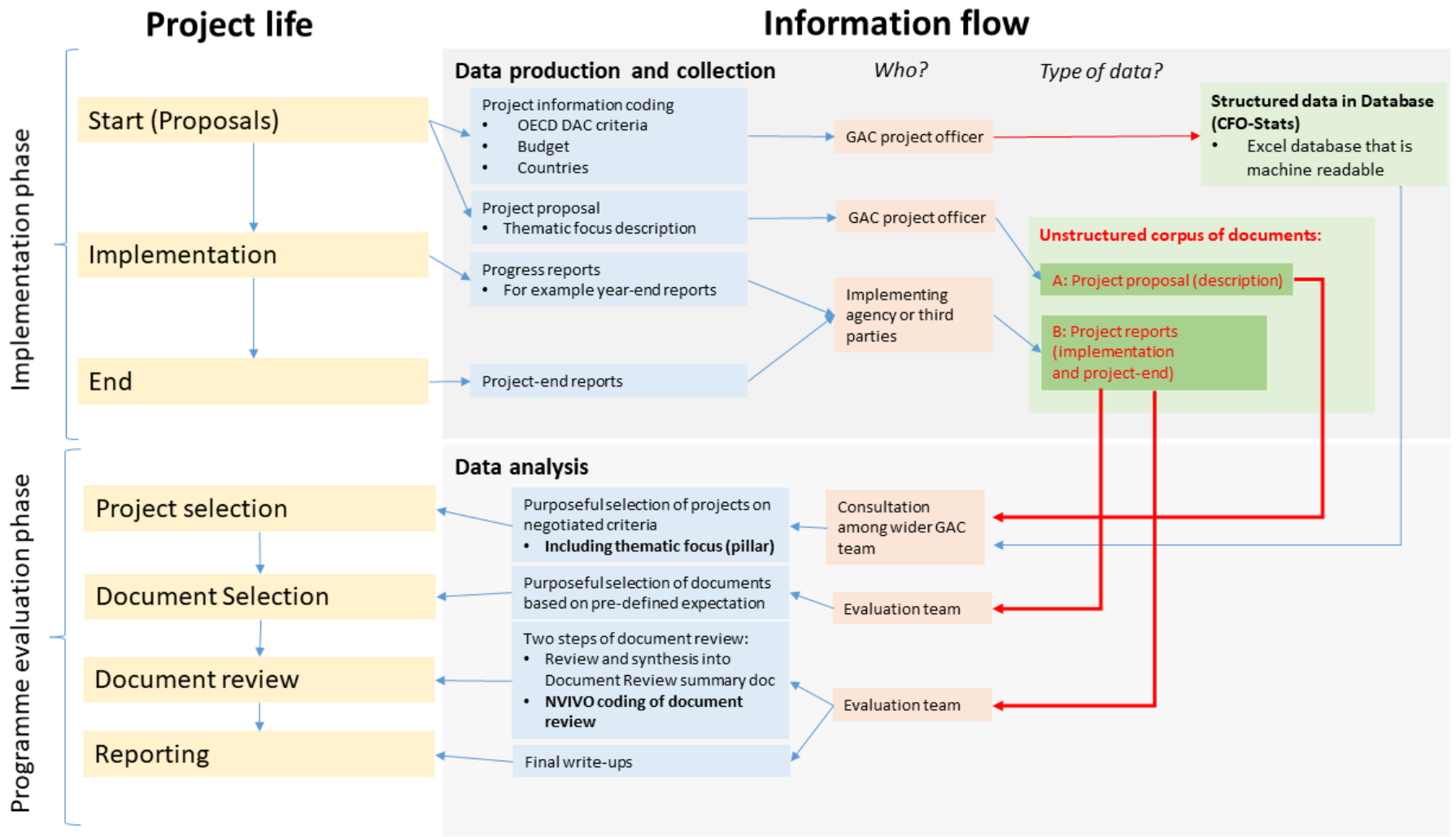
- In a first step a **template review summary** Word document is filled in by the evaluator while reviewing documentation. This template requires the evaluator to fill in information on a variety of specific indicators (e.g. project size, country

of implementation, etc.) and to describe how the documents reviewed provide answers to the evaluation questions as they relate to one specific project. Each project will have one document review summary that covers all reviewed documentation for that project.

- In a second step, the information summarised in this **Word document is coded into NVIVO**, a qualitative data analysis software. This coding consolidates information from different projects into one NVIVO database at the programme level. It appears that in some cases not only the summary review document is coded into NVIVO, but also information from other documents. This database provides structure to the qualitative information provided by data that is contained in (unstructured) text documents. It can be used, for example, to assess the frequency with which specific themes and sub-themes appear in project documentation and how they relate to each other across documents.

Fourth – the reporting stage. In this stage, the evaluation team summarises findings from the document reviews, either in a presentation or summary document.

Figure 1: The GAC Programme Information and Evaluation Cycle



2.3 GAC Programme Evaluations: Where Could Machine Learning and Natural Language Processing be Applied?

As described above, GAC programme evaluations are complex processes that involve several different steps of information and data processing by a variety of actors and at different levels (i.e. project and programme level). Figure 1 above provides a schematic overview of these processes.

While the following sections describe Machine Learning, Natural Language Processing, text analytics, and their application in more detail, the general idea behind applying these methods to the processes described above is that they could be used to improve the efficiency and effectiveness with which some manual steps are currently being implemented.

Efficiency could be improved because automated procedures would reduce the time and resource investment needed by the evaluation team to implement these steps. On the other hand, the effectiveness of these steps could be improved because automated, analytics-driven procedures could make the implementation of these steps more systematic and less prone to human errors, while allowing for a wider set of documents to be reviewed.

From our perspective, the largest improvements could potentially be gained where – in **the evaluation phase** – unstructured text data is being analysed by the GAC evaluation team. We highlighted these steps with thick, red arrows in Figure 1.

- **Project selection based on thematic focus areas (pillars)**: as described above, extracting this information from project documentation is currently done manually at the evaluation stage – i.e. the evaluation teams need to read documents and extract information to judge which thematic areas projects fall under. Extracting topics or themes from projects is something that NLP can help with.

- **Document selection**: this is currently based on pre-defined expectations of what documents will be of use to the GAC team. Information retrieval systems can help to sieve through large sets of documents to ensure that no additional sources of important information are missed.

- **Coding of document reviews**: the GAC evaluation team currently codes original documents and summary reviews into structured NVIVO databases that can then be analysed for the purposes of the evaluation. This coding involves, for example, counting whether certain themes or sub-themes appear in relation to specific projects. NLP can help with such structuring and analysis of information from texts.

Finally, at the **implementation stage**, text analytics could potentially also help GAC project officers with the coding of information from proposals into a structured database. This is highlighted with a thin red arrow in the figure above.

In the following paragraphs, we will present in more detail how Machine Learning and NLP techniques can be applied to the steps listed above. We will start with a general overview of these methods (section 3) and continue with a detailed description of potential applications to the GAC programme evaluation process (section 4). It is important to note that this paper proposes the use of these methods as complements

to the current evaluation process and as a way to free up time for evaluators to invest in parts of the evaluation process that require extensive human involvement. This paper does not propose the use of Machine Learning and NLP as replacements to what is currently being done by the evaluation team.

3 Machine Learning and Natural Language Processing: Overview and Applications in Development

In general terms, Machine Learning can be defined as a set of methods that enable computer software (algorithms) to detect patterns in data. In Machine Learning, this is done through the use of training data where an algorithm is trained to find patterns in data. Based on this, the algorithm produces a model that can make predictions for and detect patterns in new (unseen) data. This differs from traditional statistical analysis where an analytical model is pre-defined based on mathematical rules and subsequently applied to data. Machine Learning can be very effective in finding complex and nonlinear relationships in data, and for making sense of large amounts of unstructured data in various formats, including in particular text data.[2]

Machine Learning models can be classified into two groups according to how they "learn" from data. In **Supervised Machine Learning** you have input variables and an output variable and you train an algorithm to learn how to predict outputs based on inputs.[3] Supervised Machine Learning requires that the possible outputs are known for at least part of the data and that the data used to train an algorithm is already labelled with correct answers. For instance, an algorithm may learn how to identify dogs in photos, after being trained on a dataset of images that are properly labelled according to whether or not it includes a dog as well as with some identifying characteristics.[4] Supervised Machine Learning can be used to solve two types of problems: classification problems and regression problems. Classification aims to assign an instance to one of several distinct categories based on learning from past observations (e.g. whether an observation is a dog or a cat). Regression problems are defined as problems where Machine Learning is used to predict a continuous output (e.g. the weight of a person).[5]

In **Unsupervised Machine Learning** you only have input data and no corresponding output variables. The goal for unsupervised learning is therefore to model the underlying structure or distribution in the data in order to learn more about the data and to detect patterns. Unsupervised learning models therefore do not have a correct answer. Instead, the algorithm learns to identify complex processes and patterns without a human to provide guidance along the way.[6] **Many of the methods discussed further below in this review are examples and applications of unsupervised learning methods**.

See Box 2 below for an example of how supervised and unsupervised learning methods compare in the context of text analytics.

[2] USAID (2018).
[3] Brownlee, Jason (2016), *Supervised and Unsupervised Machine Learning Algorithms*, Machine Learning Mastery.
[4] Castle, Nikki (2017), *Supervised vs. Unsupervised Machine Learning*, ORACLE + DATASCIENCE.COM
[5] Note that this is different from the context in which the term 'regression' is usually applied (a statistical tool). USAID (2018) and Brownlee (2016).
[6] Caste (2017) and Brownlee (2016).

Box 2: Examples of Problems that Could be Addressed with Supervised and Unsupervised Learning

There is a fundamental difference in the types of problems or questions that are typically tackled using supervised or unsupervised learning methods. Supervised learning methods are good at solving prediction problems and estimating relationships between certain outputs and other data features. Unsupervised learning methods are good at detecting patterns and categorising observations in data into clusters that are similar to each other.

In the context of text analyses, examples of such applications could be:

- An analyst has access to a dataset that contains a long list of documents and texts that are labelled as describing projects that pertain to certain thematic categories, such as for example high-level OECD DAC sector codes: Education, Health, Water Supply & Sanitation, etc. The analyst could then use **supervised learning methods** with this pre-labelled dataset to develop an algorithm that uses the features of the dataset (the text contained in the documents) to predict which OECD DAC code a project falls under. The model built with this data (i.e. the 'seen' data, where we know the OECD DAC codes of projects), could then be used by the analyst to predict OECD DAC codes in a new dataset of documents, where these labels do not yet exist (i.e. classification).
- Alternatively, an analyst could have access to a dataset that contains the same documents as above, yet without the corresponding OECD DAC labels. The analyst could use **unsupervised learning methods** to identify whether there is any pattern in terms of the types of issues that these documents cover. For instance, the analyst could suspect that documents could be clustered or categorised with respect to the type of projects that they describe and hence experiment with unsupervised learning methods to identify these unobserved categories. The outputs of these unsupervised learning methods could be clusters that overlap with the OECD DAC labels discussed above. However, unsupervised learning methods may also find other patterns, such as relating to the geographical location of the projects or whether the projects are focused on collaboration with government or community organisations.

Machine Learning is being used to perform a range of tasks ranging from image and voice recognition, traffic predictions, and customized social media feeds, to chat bots, product recommendations and email spam filtering. **One inter-disciplinary field that draws heavily on Machine Learning is Natural Language Processing (NLP)**. NLP is about using computers to process a "natural" language spoken and written by humans.[7] The field draws on a mix of Computer Science, AI, and Computational Linguistics. NLP can be used to perform a variety of tasks such as automatic summarisation, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic modelling.[8] Essentially, NLP is about computer-based processing of human language interpreted in its widest sense.

NLP includes a wide range of text analytics methods. Some of these draw on Machine Learning. Some use supervised learning models, while others are unsupervised. NLP is already being used in a variety of contexts, ranging from common word processor operations and grammar correction, to voice recognition,

[7] USAID (2018).
[8] Kiser, Matt (2016), *Introduction to Natural Language Processing (NLP)*, Algorithmia.

converting speech to text and translating between languages. Google translate, Apple's *Siri* and predictive text (used on smartphones) are all examples of NLP in action. In the field of international development, NLP technology has been used to map public opinions expressed through radio to better respond to a refugee crisis,[9] and to understand immunisation awareness through analysis of social media and news content.[10] Again, it is important to emphasise that not all NLP methods rely on Machine Learning.

Nonnumeric formats such as unstructured text, images or audio usually require additional pre-processing to be converted into a format that can work with Machine Learning algorithms. In some cases, such as computer vision for image data or natural language processing for text data, these pre-processing steps can be complex and sophisticated — and can even themselves be augmented by Machine Learning.[11] For the context of GAC programme evaluations, we describe these in section 5.1 below.

[9] See UN Global Pulse, *Bringing in people's voices from radio content analysis to respond to a refugee crisis*, www.unglobalpulse.org/projects/bringing-peoples-voices-radio-content-analysis-respond-refugee-crisis [accessed on 21.03.2019].
[10] See UN Global Pulse, *Understanding Immunisation Awareness And Sentiment Through Analysis Of Social Media And News Content*, www.unglobalpulse.org/understanding-immunisation-awareness-through-social-media [accessed on 21.03.2019].
[11] USAID (2018).

4 Machine Learning and Natural Language Processing in GAC Programme Evaluations

Section 2 provided an overview of the process whereby GAC selects and evaluates projects under their thematic evaluations. This process involves a number of steps where unstructured text data is compiled and analysed. Machine Learning and other types of automation can potentially be used to improve this process. This section discusses how particular Machine Learning and NLP techniques may be used to increase the efficiency and effectiveness of three steps in the project selection process that involves the compilation and analysis of unstructured text data. Table 1 below provides an overview of these steps and the associated analytical techniques.

Table 1: NLP Applied to Steps in GAC's Program Evaluation Process

Step in program evaluation process	NLP techniques
Project selection: identification of thematic pillars	Topic modelling
Document selection: identification of documents for review	Search query based information retrieval systems Named entity recognition Word2vec
Document review: the coding of documents (currently undertaken in NVIVO)	Sentiment analysis Search query based information retrieval systems Topic modelling

4.1 Project Selection: Using Topic Modelling to Identify Thematic Pillars

Information received by the review team to understand the GAC programme evaluation process revealed that part of the project selection step relies on identifying thematic pillars that projects fall under. This information (e.g. Health System Strengthening) is currently not coded into the CFO-Stats database:

"These are internal thematic focus areas that projects fall under, however, this information is not captured by the project officer (and thus, CFO-Stats does not have the information). In order to ascertain which pillar(s) a project falls under, we made our best guess, and then this was verified during our consultations on the sample with each programming area […]*."*

Topic modelling can be applied to this procedure. Topic modelling is an unsupervised learning approach that can be used to identify topics that best describe a set of documents. It is a potentially useful tool for GAC because it can be used to automate and systematically capture themes that emerge from the variety of documents associated with a particular program.

When it comes to Natural Language Processing, there is a hierarchy of lenses that we can use to extract meaning from text. We can look at specific words, at sentences or paragraphs as well as whole documents. When looking at a whole document or collections of documents (the latter called a *corpus*), a useful way to understand text is to analyse its topics.[12] A topic contains a cluster of words that frequently occurs together. Topic modelling connects words with similar meanings and it can distinguish between uses of words with multiple meanings.[13] Topic modelling can be used to analyse large volumes of text where we want to discover patterns of word-use and to connect documents that share similar patterns. It can also help people automatically organise, search, index and browse large collections of documents once topics have been identified.

All topic models are based on the same assumption that each document consists of a mixture of (latent) topics and that each topic consists of a collection of words: "topic models are built around the idea that the semantics of our document are actually being governed by some hidden, or "latent," variables that we are not observing. As a result, the goal of topic modelling is to uncover these latent variables—topics—that shape the meaning of our document and corpus."[14] Expressed differently, this means that topic models assume that any document has been written with specific issues or topics in mind that shape the text that forms the document. These topics are not explicitly expressed in the document, but reveal themselves by reading the text. For humans, identifying these themes can be a straight forward process. In NLP, topic modelling aims to train an algorithm to do the same.

There are different methods for topic modelling. Alqhamdi and Alfalqui (2015) identify four:

- Latent Semantic Analysis (LSA),
- Probabilistic Latent Semantic Analysis (PLSA),
- Latent Dirichlet Allocation (LDA),
- Correlated Topic Model (CTM).

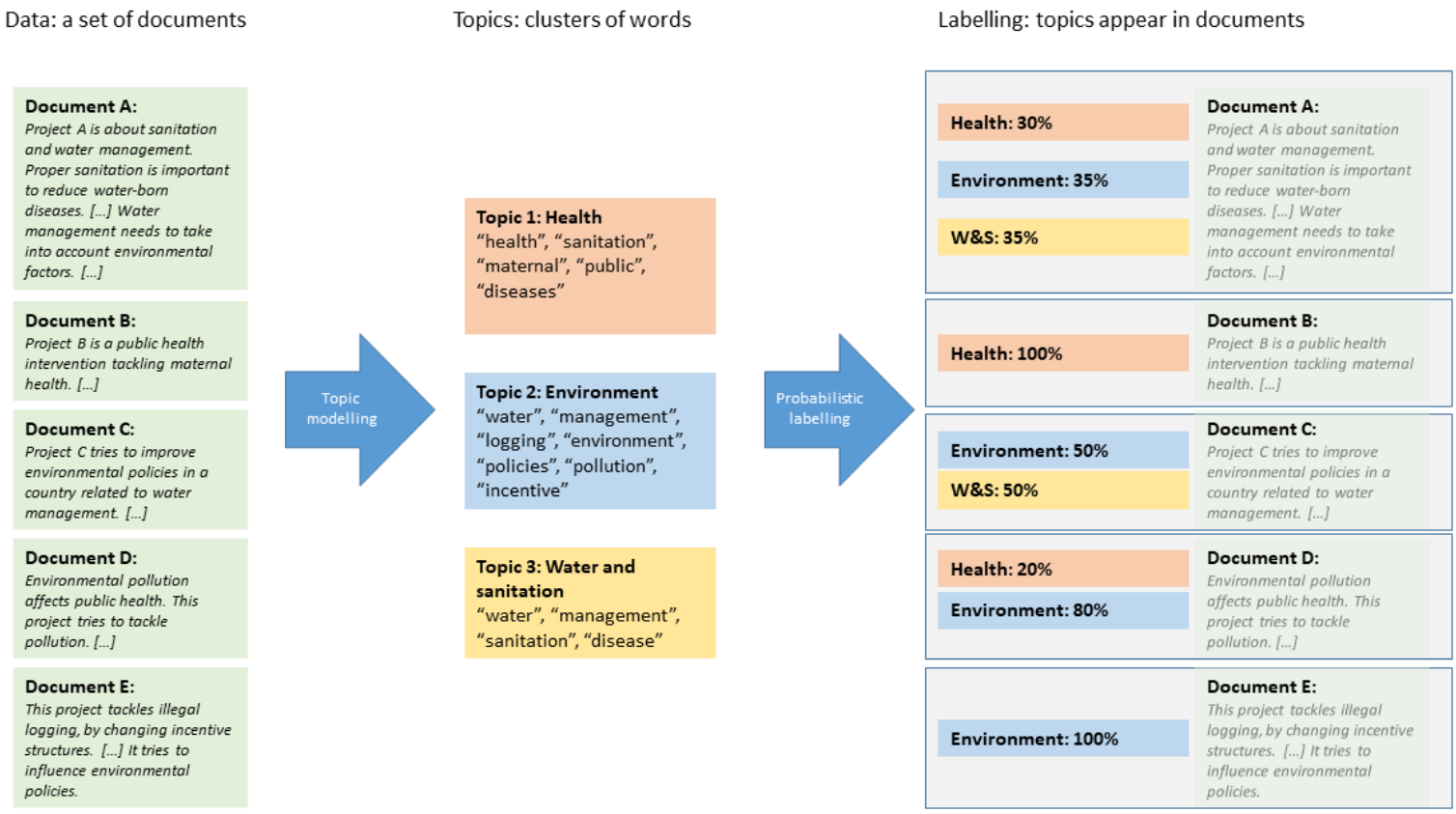
Each model, has a slightly own approach, strengths and weaknesses. For instance, LDA imagines a fixed set of topics across a corpus of documents, and each topic represents a set of words. The goal of LDA is to map all the documents to the topics in a way so that the words in each document are mostly captured by those imaginary topics. This is illustrated in Figure 2 where words are modelled by a set of topics (i.e. clusters of words associated with certain topics) and documents modelled by a set of topics (i.e. documents labelled by the topics that appear in them). Topics created in an LDA model are based on patterns of word co-occurrence in documents and they do not necessarily match up with theoretical concepts or themes (such as GAC's "thematic pillars"). Topics could, for instance, also reflect a certain writing or speaking style (e.g. words referring to emotions), events (such as natural disasters) or frames.[15] We will in this paper not dive into the technical differences between the various topic modelling

[12] Xu, Joyce (2018), *Topic Modelling with LSA, PLSA, LDA & Ida2Vec*, Medium.
[13] Alghamdi, Rubayyi and Khalid Alfalqui (2015), 'A survey of topic modelling in text mining', *International Journal of Advanced Computer Science and Applications*, 6(1), p 1.
[14] Xu (2018).
[15] Jacobi, Carina, Wouter van Atteveldt and Kasper Welbers (2016), 'Quantitative analysis of large amounts of journalistic texts using topic modelling', *Digital Journalism*, 4(1).

approaches. However, the literature referenced in this section provides additional details.

These types of topic models are being used in a range of contexts. For instance, in evaluation of scientific impact, trend analysis, and document search[16] or when (online) consumers of books and magazines are provided with reading suggestions.[17] One initiative also modelled topics of legal documents as a quick way to summarise their content and nature.[18] The box below provides an example of the types of inputs that can be used in topic modelling and what the output looks like.

Figure 2: Topic Modelling in Action



[16] Lau, Jey Han, David Newman and Timothy Baldwin (2014), *Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality*, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.
[17] For an easy-to-understand example of application watch this video: https://www.youtube.com/watch?v=3mHy4OSyRf0
[18] Oguejiofor Chibueze (2018), *NLP For Topic Modelling Summarization of Legal Documents*, Medium / Towards Data Science.

Box 3: Topic Modelling in Action

Dwivedi (2018)[19] provides an example of the application of LDA. The data used in this example is the 20 Newsgroups data set, a collection of approximately 20,000 newsgroup documents partitioned fairly evenly across 20 different topics such as space, politics (guns), politics (middle east), religion (Christian), and sports (baseball).

The raw data (newspaper articles) were first pre-processed. This involved tasks such as removal of stop words as well as lemmatisation and stemming of words (see 5.1.3 for explanation). The next step entailed creating an overview table with all the words in each article and a value indicating the number of times that word occurs in the entire corpus. This is sometimes referred to as creating a "bag of words". The same type of overview table were subsequently produced for each news article. This was all done with pre-written algorithms and therefore did not involve a lot of human labour.

The LDA model was subsequently applied to the ready data. This entailed specifying how many topics there are in the data set. You may, for instance, start with 8 unique topics and see what the output looks like. The output from the LDA model may look like this:

- Topic 1: (possibly space)
  Words: "space","nasa" , "drive" , "scsi" , "orbit" , "launch" ,"data" ,"control" , "earth" ,"moon"
- Topic 2: (possibly sports)
  Words: "game" , "team" , "play" , "player" , "hockey" , season" , "pitt" , "score" , "leagu" , "pittsburgh"
- Topic 3: (possibly politics)
  Words: "armenian" , "public" , "govern" , "turkish", "columbia" , "nation", "presid" , "turk" , "american", "group"
- Topic 4: (possibly gun violence)
  Words: "kill" , "bike", "live" , "leav" , "weapon" , "happen" , *"gun", "crime" , "car" , "hand"

The LDA model will not provide label names to the collections of words identified as topics. Assigning these labels will have to be done by a human (suggested labels have been provided in brackets in the example above).

Applying topic modelling in GAC's programme evaluations: GAC selects projects for evaluation in part based on which thematic focus areas they fall under. As described above, evaluators currently read various project documentation and manually extract information required to identify the thematic focus of a project. Topic modelling as described in this section could be used to automate this process, and give evaluators an idea of which topics best characterise project documents. Some of the topics emerging from topic modelling possibly coincide with the thematic pillars that evaluators would have identified manually, while others may be of less relevance. This analysis could help evaluators more quickly and systematically identify thematic pillars and, based on this, select projects for evaluation.

4.2 Document Selection: Using Information Retrieval Systems to Select Relevant Documents for Review

It can be a burdensome task to identify all the relevant documents for review in connection with a programme evaluation. As GAC evaluation officers gather project

[19] Priya Dwivedi (2018), *NLP: Extracting the main topics from your dataset using LDA in minutes*, Towards Data Science.

documents for review, they typically only ask for a subset of documents: "*Asking for all documents related to a project would not only take time and be a burden on the project officers, but would also involve searching through unimportant information.*" This is understandable, but entails a risk of sampling bias if important information is accidentally left out or overlooked.

Information Retrieval and Word2Vec

Machine Learning as well as simpler Information Retrieval systems may be used to improve the speed and ease with which documents are selected, and to help avoid that documents with crucial information are overlooked. One way to improve document selection is to use relatively simple search query Information Retrieval systems that do not draw on Machine Learning. Information Retrieval is about finding material (usually documents, but this can also refer to more specific pieces of information) from a wider pool of unstructured information.[20] This is a fairly broad definition, and it includes a range of activities – including the use of library catalogues, physically looking up something in a dictionary or "googling" information online. There are a variety of methods and systems for conducting Information Retrieval. Some methods are sophisticated and may include Machine Learning elements such as those discussed below. Simple Information Retrieval methods include physically reading through text to look for whether key pieces of information appear, or conducting a keyword search in Microsoft Word.

A fairly simple Information Retrieval system can be based on a model where a user poses a search query using operators such as AND, OR, and NOT. This query conveys to a computer what information is needed.[21] A simple keyword search function/system has the advantages that it is cheap and easy to develop and implement. This approach is also useful when users know exactly what they are looking for and are able to convey this to a computer in a simple concise query.

However, drawbacks include that regular keyword searches do not inherently take into account synonyms or more abstract terms related to the given query words. Users may therefore need to conduct a range of different searches using different words and combinations of words to ensure they capture all relevant aspects of a theme.[22] In the context of GAC's programme evaluations, a simple search query for documents that include "health" and "Ghana" but not "Nigeria" may take the following from: "health" AND "Ghana" NOT "Nigeria".

In contrast to a simple keyword search function, Machine Learning based methods are able to take account of a text as a whole, meaning of text, and not just the specific words. By considering the entire text of a document, more information can be taken into account, such as the context of keywords used.[23] For example, a GAC evaluation officer may only wish to identify documents that talk about "maternal health" in the

[20] Manning, Christopher D., Prabhakar Raghavan & Hinrich Schutze (2008), *Introduction to Information Retrieval*, Cambridge University Press.
[21] Manning et al (2008).
[22] For a discussion of the merits and drawbacks of keyword searches compared to Machine Learning models see Helmers, Lea, Franziska Horn, Franziska Biegler, Tim Oppermann, Klaus-Robert Muller (2019), *Automating the search for a patent's prior art with a full text similarity search*.
[23] Helmers et al (2019); and Skymind AI, *A beginners guide to Word2vec and neural word embeddings*.

context of Ghana (and not in general), or to only identify documents that mention the country “Turkey” and not the bird “turkey”.

Simple search query systems can also be combined with Machine Learning techniques. For instance, in case users wish to search for text in documents that are in an image or PDF format, Machine Learning could first be used to transfer data from such images into machine readable databases. (See more on this in section 5.1.1.)

Additionally, Machine Learning techniques such as the so-called “Word2vec” method can be used to help GAC evaluation officers continuously improve their search queries, by suggesting and expanding searches to identify the right documents and information pieces.[24] Word2vec is a group of methods that are used to produce word embeddings. It quantifies and categorises similarities between various words and sentences in a text. By doing so, it allows algorithms to estimate meanings of words and hence to identify words that are similar to each other. With enough data, usage and contexts, Word2vec can make very accurate estimates about a word’s meaning based on past appearances.[25] This means that searching for specific words or words with similar meanings can become more accurate than simple keyword queries. The box below provides an example for Word2vec in action.

[24] Helmers et al (2019); and Skymind AI, *A beginners guide to Word2vec and neural word embeddings*.
[25] Skymind AI, *A beginner’s guide to Word2vec and neural word embeddings*.. See also Wikipedia, Word2vec, https://en.wikipedia.org/wiki/Word2vec [accessed on 21.03.2019].

Box 4: Word2vec – An Example

Word2vec is a method to measure the similarity of meanings of words in a quantitative way. It does so by looking at the context (i.e. the surrounding words) in which words appear. The fundamental idea behind this is that words with similar meanings tend to appear in similar contexts. For example, the two words "house" and "building" have similar meanings (they can describe physical structures humans live in) and hence they tend to appear in similar contexts, e.g. together with other words that describe properties or dwellings. On the other hand, the word "bank" as a financial institution has the same form as "bank" in the "river bank" (so called homonymy) but the context in which these two words appear allows us to distinguish their meaning. Word2vec measures such word similarity in one figure. When searching through text, this allows to enter one word as search term and then also find other words or text components that have similar meanings.

Word	Distance
Norway	0.760124
Denmark	0.715460
Finland	0.620022
Switzerland	0.588132
Belgium	0.585835
Netherlands	0.574631
Iceland	0.562368
Estonia	0.547621
Slovenia	0.531408

Technically speaking, Word2vec produces word embeddings. Word embeddings are vector representations of a particular word – i.e. the position of one point (word) in space relative to another. The usefulness of Word2vec is that it can group the vectors of similar words together in a vector space. Words that have similar or otherwise related meanings (such as "good" and "great", or "man" and "boy") will occupy close spatial positions in the vector space. With enough data, usage, and contexts, this technique can make very accurate estimates about a word's meaning based on word co-occurrence in texts. For instance, Word2vec produces the list above (right-hand side) of words associated with "Sweden" in order of proximity.[26]

Named entity recognition

Named entity recognition (NER) is another potentially relevant, and fairly simple Machine Learning technique that may be used to improve document classification and identification. NER is used in many parts of NLP and it seeks to locate and classify named entities in text into pre-defined categories such as names of persons, organisations, and locations.[27] NER can be used to classify content for news providers by automatically scanning an article and reveal key people, organisations and locations discussed in it. Based on this information, the article can then be automatically categorised in some defined hierarchy and easily accessed. This adds a lot of semantic knowledge to a text. There are various standard libraries used to perform NER (e.g. Stanford NER, spaCy and NLTK) and these include a wide range of categories for classification. Categories include: i) people, ii) nationalities or religious or political groups, iii) companies, agencies, institutions, iv) countries, cities, states, v) dates, and vi) events such as hurricanes, wars or sports events.[28] In the context of document selection for GAC's programme evaluations, NER can be employed for different purposes. For instance, NER may be used to identify documents that include

[26] Karani, Dhruvil (2018), *Introduction to Word Embedding and Word2Vec*, Towards Data Science.
[27] Li, Susan (2018), *Named Entity Recognition with NLTK and SpaCy*, Towards Data Science / Medium.
[28] Banerjee, Suvro (2018), *Introduction to Named Entity Recognition*, Explore Artificial Intelligence / Medium.

certain geographical locations (e.g. Myanmar or Turkana County in Kenya) or actors (such as organisations like UNICEF). Furthermore, NER can be used by GAC officers to quickly identify the various people, actors or countries mentioned in a particular document.

Box 5: Named Entity Recognition – An Example

This box provides a simple example of how NER, in practice, helps to label text components with respect to the appearance of known or named entities.

The example is drawn from Li (2018). Li uses SpaCy's named entity recognition which has been trained on the OntoNotes 5 corpus. This corpus supports a wide range of entity types such as NORP (nationalities or religious or political groups), ORG (companies, agencies, institutions, etc.), MONEY (monetary values, including unit), DATE (absolute or relative dates or periods) and EVENT (named hurricanes, battles, wars, sports events, etc.).

SpaCy's NER is applied on the following sentence: "European authorities fined Google a record $5.1 billion on Wednesday for abusing its power in the mobile phone market and ordered the company to alter its practices", and the algorithm produces the following output:

[(#European', 'NORP'),
('Google', 'ORG'),
('$5.1 billion', 'MONEY'),
('Wednesday', 'DATE')]

The NER algorithm performs satisfactorily. European is identified as NORP (nationalities or religious or political groups), Google is identified as an organization, $5.1 billion is identified as a monetary value and Wednesday is identified a date. NER can be used to label large sets of text in this manner quickly. After labelling, simple search functions, such as e.g. for all organizations, can be deployed to look through this text for entities of interest.

Application in GAC's programme evaluations: In GAC's programme evaluation process, evaluation officers usually only select a pre-defined subset of project documents for review. GAC can use improved search query based Information Retrieval systems to improve the speed and ease with which documents are selected, and to help avoid that documents with crucial information are overlooked. Information Retrieval systems can be further strengthened with Machine Learning techniques such as the "Word2vec" method, which can be used to automatically suggest and expand search queries. Similarly, NER can be used to a help GAC evaluation officers better (and faster) categorise and understand the subject of any given text, and make a judgement as to whether it is relevant for document review. These methods are all potentially useful tools for improving and speeding up the document selection stage in GAC's programme evaluation process.

4.3 Document Review: Using Machine Learning to Improve Coding and Analysis of Documents

The document review stage involving NVIVO coding may also provide a number of opportunities for the use of Machine Learning and other types of automation. However, this will to a great extent depend on the coding criteria and evaluation questions applied in each type of evaluation. The techniques discussed above are all likely to be relevant in this phase, including search query Information Retrieval

systems, NER and topic modelling. Machine Learning powered Information Retrieval could, for instance, be used to quickly detect documents / projects that include certain themes or key words, such as projects that use innovation and new technology, documents that mention design issues, or staff turnover challenges.

Sentiment Analysis

In some instances, GAC evaluation officers may be interested in reviewing documentation that contains feedback on or expresses sentiments towards a particular programme. This can help GAC identify implementation challenges, likely impact and relevance (including contextual fit) of an initiative. Documents to be used for this type of analysis could be annual reports, external programme reviews or statements and emails with feedback from partners and stakeholders.

In these situations, Machine Learning powered sentiment analysis is a potentially useful method for efficiently capturing positive, negative or neutral opinions about a particular programme during the document review phase. [29] Sentiment analysis involves the use of algorithms to attach an emotional label to text.[30] This is particularly useful in cases with large amounts of textual data from a variety of sources – information that would take a long time for humans to read and code.

Sentiment analysis is based on dictionaries of words that are associated with positive or negative emotions[31] and it is being used in a variety of contexts. For instance, looking at restaurant reviews, Snyder and Barzilay (2007) have applied sentiment analysis to capture multiple related opinions about a restaurant in the same text, allowing for differences in opinion towards food compared to ambience and service.[32] Similarly, the Behavioural Insights Team (2018) have used words and phrases from user reviews of general medical practitioners (GPs) from the UK's National Health Service website to identify indicators of good or bad practice.[33]

Document Review at the Start of Projects

The discussion so far has focused on three key stages in GAC's programme evaluation process where there is potential for Machine Learning application. In addition to these main areas, Machine Learning may also be useful at the outset of the programme cycle: the start of projects. In this stage the project officer currently transfers information manually from the project proposal into a structured database (CFO-Stats). This includes assigning OECD DAC sector codes and countries covered, as well as entering budgets and other specific indicators relating to the project. A combination of the methods discussed above could be used to ease or automate some of these tasks. For instance, topic modelling and Information Retrieval

[29] Jain, Anuja P and Padma Dandannavar (2016), *Application of machine learning techniques to sentiment analysis*, 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).
[30] USAID (2018)
[31] Broniecki, Philipp, Anna Hanchar, and Slava J. Mikhaylov (2017), *Data Innovation for International Development: An overview of natural language processing for qualitative data analysis*, arXiv: computer science.
[32] Snyder, Benjamin and Regina Barzilay (2007), *Multiple Aspect Ranking using the Good grief Algorithm*, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, in HLT-NAACL.
[33] Behavioural Insights Team (2017), *Using Data Science in Policy. A report by the Behavioural Insights Team*, London: The Behavioural Insights Team.

systems may be used to identify OECD DAC sector codes and NER can be applied to help identify countries covered in the project.

Document Review for Comparative Analyses

Furthermore, there may be scope to use Machine Learning in a comparative analysis of projects. PRA is interested in conducting more rigorous comparative case studies in its evaluations - to examine the successes and failures of projects in order to distinguish contextual factors and mechanisms that can lead to successful outcomes. Machine Learning can help inform such comparison. For instance, Machine Learning could be used to look for underlying patterns in text describing projects deemed successful (or failures). This might entail looking for similarities and differences in terms of the themes that appear in successful/failed projects or the sentiments with which key stakeholders refer to the projects in stakeholder consultation documentation.

However, there are limitations for this type of analysis. Most fundamentally, the analysis will be limited by the available data. The data foundation for this analysis will be available project documentation. The availability, depth and quality of documents may vary from one project to another. This can make comparisons unreliable. Furthermore, while some important contextual factors and mechanisms may appear in these documents, it is likely that many others will not. For instance, changes in the political economy environments in which projects are implemented may not be captured in the available document corpus. To mitigate such weaknesses it may, in comparative analyses, be useful to expand the types and sources of data used.

This could entail drawing on “Big Data”, such as various publicly available data sources of relevance to a project, including news stories, social media chatter, parliamentary debates and macro-economic indicators from the country that a project is implemented in. Such alternative data sources could help reveal important insights about the wider political, social and economic context within which projects succeed or fail.[34]

4.4 Case Study: Combining Different Machine Learning and Natural Language Processing Methods in one Project

Sections 4.1 to 4.3 above describe different Natural Language Processing and Machine Learning methods separately that could be used by GAC in programme evaluations. It is important to emphasise, however, that in many applications such methods are used in conjunction and related outputs build on each other. The same would apply to a situation where these methods would be used to improve the processes depicted in Figure 3. To show how this could look like in practice, we present a case study in Box 6 below that provides an example of how Information Retrieval, topic modelling, and sentiment analysis have been applied in a project that aimed at analysing public opinions in Uganda.

[34] For a detailed discussion of “big data” see Hammer, Cornelia L., Diane C. Kostroch, Gabriel Quiros, and STA Internal Group (2017*), Big Data: Potential, Challenges and Statistical Implications*, IMF Staff Discussion Note, Washington D.C.; and UN Global Pulse (2016), *Integrating Big Data into the monitoring and evaluation of development programmes*, New York: UN Global Pulse.

Box 6: Case study 1: Using Information Retrieval and Machine Learning to Analyse Public Opinions in Uganda

In 2016, the first live televised Presidential debates were held in Uganda as a precursor to the general elections that took place later that year. In this context United Nations Global Pulse, a UN outfit tasked with harnessing big data for development, collaborated with the United Nations Development Programme (UNDP) to identify, in the aggregate, public perceptions of how the debates were organised and whether they were viewed as relevant to the electoral process.

To do this, the initiative looked at social media (Facebook) where debates were extensively discussed. This involved the collection, compilation and analysis of a large volumes of unstructured text. From January to February 2016, approximately 50,000 messages were shared in public pages by 25,000 unique individuals. These messages where extracted through keyword-based information retrieval that applied filters to anonymised Facebook messages. Filtering was done using a taxonomy of keywords such as 'presidential debate' or '#ugdebate16'. Essentially, messages were selected based on whether specific keywords were included in them or not.

The messages that made it through the filtering were then subjected to topic modelling. This analysis was used to categorise the messages into "general comments" and "comments about specific topics". For the latter category, the main topics identified were about "candidates", "moderators", "organisers" and "outreach". Based on this categorisation, Global Pulse proceeded to analyse trends in comments, including through the use of sentiment analysis. For instance, the analysis revealed that more than 90% of the discussions around the moderators were negative. Furthermore, the public expressed concerns about the choice of moderators, some moderators' conduct and the way they posed questions.

The graph below summarises how different NLP methods were used in this process to analyse the text data originally coming out of Facebook comments: the comments were first filtered using Information Retrieval systems. Then, topic modelling was used to identify the four key topics and to categorise comments with respect to the topics they dealt with. In a final step, sentiment analysis was used to assess sentiments towards these topics.

This case study[35] provides a good example of how relatively simple Information Retrieval methods and advanced Machine Learning techniques can be used to improve the efficiency and effectiveness of how large volumes unstructured natural language text is categorised and analysed. The example also illustrates how this work can help development actors better understand and navigate the context that they operate in. For additional examples of how Information Retrieval and Machine Learning techniques have been applied to categorise and analyse natural language text see Jacobi et al (2016) and Helmers et al (2019).

[35] For details about this project see UN Global Pulse, Informing governance with social media mining, debates.unglobalpulse.net//uganda/ [accessed on 21.03.2019].

5 Moving from Concept to Implementation

The following sections will provide some general observations about issues that need to be taken into account when organisations are planning to use any of the methods discussed in the previous sections. It should be noted that all discussions and descriptions of GAC processes presented in this document are based on second-hand information provided in writing to the team drafting this review.

In order to develop a more tangible and actionable action plan for testing and implementing the use of Machine Learning tools in GAC's programme evaluations, the team drafting this paper would need to spend time with the GAC evaluation team directly, asking questions and observing workflows in action. Without such direct observation and deeper consultation, it is likely that opportunities and difficulties that GAC might encounter when testing these methods are overlooked.

5.1 Technical Requirements and Considerations

Section 4 above presents a selection of analytical methods related to Machine Learning and NLP that could be included in the PRA programme evaluation process described in section 2 in order to improve the way in which this process is implemented. These methods are likely to be particularly helpful where large sets of unstructured text data need to be processed, either for management or analytical purposes. However, certain technical requirements have to be satisfied for implementation to happen and analyses to deliver useful results.

5.1.1 Machine-Readable Text: Encoding and Optical Character Recognition (OCR)

A key first step in any project that aims to analyse text using the NLP and Machine Learning methods described above is to make sure that the text databases on which the analyses are based are machine-readable. In practice, this means that text needs to be in a format so that standard statistical programming languages such as e.g. R or Python can read text into their internal storage system to then perform analyses on this data. The information received from PRA and GAC by the study team suggests that texts used in the process described in section 2 can come in different formats:

- Some documents will be stored in standard text-processing formats, such as e.g. Microsoft Word or Excel. Standard data analysis software can easily read such documents.
- Other documents will be stored as PDFs that are equally easily transferrable into text data given that they are stored in the right format. These are typically PDFs that have been produced (and subsequently converted) by text-processing software in the first place (e.g. Microsoft Word) and are stored with appropriate meta-data.
- Finally, some documents will be stored as **images or images within PDFs**. This can typically be the case where physical documents have been scanned or photographed and added to a repository.

While careful processing of text that comes in the first two formats will be required, it is technically easy to read such documents into datasets, and hence such text can be considered machine-readable.

Documents that are stored as images, on the other hand, cannot easily be read by machines. In order to analyse such text, a pre-processing step of Optical Character Recognition (OCR) will be required. This means that a computer will have to be given the task of 'reading' the images and copying the text identified into a format that can then be analysed.

Lucas et al. (2015) provide an overview of methods and software that can be used for OCR. [36] There are a significant set of open-source options available to researchers that wish to implement OCR. While this "software is rarely, if ever, 100% accurate, clearly-written texts can often achieve accuracy rates of above 90%, which is enough to understand the content of the text and to use automated content analysis."[37] Examples for software that could be used are Tesseract[38] and FreeOCR[39]. Once documents have been processed using such software, the text can be analysed by a computer. GAC documents could therefore fairly easily be processed into machine-readable formats.

5.1.2 Multilingual Modelling and Translation

Documentation reviewed in GAC programme evaluations can come in several different languages due to two main reasons. First, documents in Canada are provided in the two official languages of the country – English and French. Second, GAC funds projects across the globe, which means that implementing partners might sometimes provide certain documentation in a third language that is neither French nor English (e.g. Spanish). This is of relevance because many analytical methods applied to text data generally assume that documents are written in one language only.

While there are some explicitly multilingual methods to NLP, e.g. to implement multilingual topic modelling,[40] more commonly used approaches focus either on analysing documents within a specific language only (e.g. separating English and French texts and analysing them independently from each other) or on translating documents into one single language.

While translation of texts by professional (human) translators remains the Gold Standard, the volumes of texts to be translated in some instances prohibit doing this comprehensively for entire text corpora. Hence, researchers increasingly rely on

[36] Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, Dustin Tingley (2015), 'Computer-Assisted Text Analysis for Comparative Politics', *Political Analysis*, 23(25), pp. 254-277. See also: Anyline, What is Optical Character Recognition?, medium.com/@anyline_io/what-is-ocr-why-does-it-make-your-life-easier-209b9fcedec4 [accessed on 21.03.2019].
[37] Ibid. Online Appendix, p. 1.
[38] Tesseract-ocr, github.com/tesseract-ocr/ [accessed on 21.03.2019].
[39] Soda PDF anywhere, www.sodapdf.com [accessed on 21.03.2019].
[40] Boyd-Graber, Jordan and Philip Resnik (2010), 'Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation'. *Empirical Methods in Natural Language Processing*. See also: Hu, Yueying, Ke Zhai, Vladimir Eidelman, Jordan Boyd-Graber, 'Polylingual tree-based topic models for translation domain adaptation', *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers, pp. 1166-1176.

automated translation methods. Most prominently, Google and Microsoft have made their translation systems accessible to the public. These systems can easily be accessed and used to automatically translate large volumes of text from and into multiple languages. One example is *translateR,*[41] a package that can be used in the statistical programming software R and which uses Google's and Microsoft's translation applications. The benefit of being able to access this from within the statistical analysis software is that the translation step can be fully integrated into any analytical workflow.

It should be noted that automated translations are never as good as translations made by a competent human. However, it is important to emphasise that most NLP methods do not require fully perfect translations of documents. Rather it is only of relevance that important terms are translated appropriately, not all words, which the Google or Microsoft translation systems generally achieve well.[42] In fact, Vries et al. (2017) find that automated translations perform similar to official manual translations when comparing the two directly.[43]

5.1.3 Other Text Pre-Processing

Technically, both steps above relate to the phase of text pre-processing, where text data is prepared so that it can then be analysed statistically. In research projects and applications where text is used as data, there are a few further pre-processing steps that are generally implemented, once the text corpus is machine readable and available in one single language. In particular, these focus on preparing the text dataset in a way so that it only contains the most useful information. For the context of GAC programme evaluation – where texts are likely to be available in English or French – two specific steps are of particular relevance:

- **Stop word removal.** In this step, words that are very common in a language but that add very little information in analytical modelling are removed from datasets. In English, this could be words such as "the" or "that". Stop word removal can easily be implemented using common statistical software packages as lists of stop words (for many languages) are provided in such packages.
- **Stemming and lemmatization.** In this step, words are reduced to their 'stem' or base form (*lemma*), which means that conjugations, plurals, or other forms of transformations are stripped from words in order to return one common form of the word. For example, the words 'economics', 'economical', and 'economist' could all be related to the base form of 'economic'. Stemming and lemmatization helps to standardise words in texts so that they can then be analysed for their occurrence and co-occurrence using methods described in previous sections. However, it should also be noted that they need to be

[41] TranslateR, cran.r-project.org/web/packages/translateR/translateR.pdf [accessed on 21.03.2019].
[42] Lucas et al (2015)
[43] Vries, Erik de, Martijn Schoonvelde, Gijs Schumacher (2017), *Lost in Translation? Evaluating the usefulness of machine translation for bag-of-words text models*, The Euengage working paper series.

applied with care, given that certain meaning is stripped from words when implementing this processing step.

Depending on the language that a certain document has been written in, other pre-processing steps might be necessary.[44]

5.2 Analogue Requirements and Considerations Affecting Successful Implementation

In addition to technical requirements, the usefulness of digital technology hinges on a wide range of analogue factors.[45] In its report on the use of AI in international development, USAID notes that **"**each step of building a Machine Learning model requires making choices that can reflect personal biases and judgments, as well as expertise and insight. As Machine Learning models are developed into tools that inform decision-making, they become part of a larger system, interacting with people, organizations, social norms and policies. This social influence is reflected in the data the models consume, in choices that are made about how models are developed and refined, and in decisions about how the outputs of the model are used."[46] In addition, it is important to emphasise that the use of Machine Learning and NLP should be seen as a way to support and augment the work of humans, rather than to replace them. Efficiency and effectiveness gains from the use of Machine Learning may, for instance, free up GAC evaluators to then devote more time on the parts of programme evaluation that require intensive human involvement (such as creative tasks associated with interpreting findings and revisiting causal assumptions and theories of change in a project). All of these factors will influence the way in which digital technology will be used and how successful it can be.

We have identified three sets of "analogue" factors affecting Machine Learning initiatives. These apply not only to Machine Learning projects, but to public sector digitisation and reform initiatives more generally.

First, leadership and the skills-mix in the implementing team matters. Introducing Machine Learning in programme evaluations in GAC will require buy-in and skilful change management from senior leadership. Identifying and nurturing champions and addressing pockets of resistance will be a key task. It is also important that the group of people managing the implementation process has a mix of expertise, not only in Machine Learning and Data Science, but also subject matter expertise in relation to evaluations and the sectors that GAC works in. Similarly, given the usual product development process for AI solutions, it is advisable to have people with expertise in how to deliver projects in a more agile and experimental fashion based on design thinking principles.[47] Generally, ensuring diversity and interdisciplinarity in teams working on Machine Learning solutions is an important way to manage the risk that

[44] See Lucas et al (2015) for additional examples of pre-processing steps.
[45] World Bank (2016), *World Development Report 2016: Digital Dividends*, World Bank Group, Washington D.C.
[46] USAID (2018), p. 44.
[47] Plattner, Hasso, Christoph Meinel and Larry Leifer (Eds.) (2011), *Design Thinking. Understand – Improve – Apply*, London: Springer Heidelberg Dordrecht.

bias and ill-founded assumptions are built into analytical models.[48] Furthermore, it will be crucial for these people to understand the current work processes and incentives that programme evaluators work within. For instance, it will be important to understand and address any fears that staff may have as to whether their jobs and expertise become superfluous in the wake of computer driven automation.[49]

Second, existing organisational processes and systems affect success.[50] There must be a fit between the design and implementation of a Machine Learning initiative and the wider set of systems and processes in an organisation. At times, this may require the Machine Learning initiative to adjust to existing organisational structures and systems, while at other times there may be a need for changes to these structures and systems. Learning from past IT initiatives highlight this need. For example, (non-Machine Learning based) IT investments by police departments in the US have been shown to only be linked to improved productivity when they are complemented with organisational changes and adjustments in management practices.[51]

Existing GAC systems and processes that influence the success of a Machine Learning initiative may include formal structures such as decision making procedures (i.e. how and by whom projects are selected for evaluations), and arrangements for recruitment and professional development of staff (i.e. how new talent is attracted and how staff skills are developed in new technical areas). Informal procedures and operating practices in GAC will also affect how a Machine Learning initiative will be implemented in practice. For instance, established but informal practices around document selection (section 2) for evaluations in GAC may discourage evaluators from using new documents suggested by an improved Information Retrieval system.

Third, the wider institutional environment matters.[52] Existing institutions such as policies, laws and regulations influence what is desirable, possible and legal. For instance, data protection laws may affect how, when and for what purposes data (incl. data about beneficiaries and vulnerable populations) can be stored and used. The European Union's General Data Protection Regulation (GDPR) is a good example of how existing or new regulation can affect AI projects. Consumers interacting with AI-enabled services such as NLP powered personal assistants, robo-advisors providing automated financial advice, and movie recommendations on streaming services, will all be affected by this new law.[53] Similarly, GDPR will have ramifications for AI solutions in the public sector, though these effects will depend on the particular project and a country's legislation. In the context of using Machine Learning in GAC's programme evaluations, GAC will need to ensure that any new procedures for data collection, storage and processing are aligned with existing legislation and policy frameworks.

---

[48] Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz (2018), *AI Now Report 2018*, AI Now Institute, New York University, p. 36.
[49] Ostrof, Frank (2006), *Change Management in Government*, Harvard Business Review, May issue; Hallsworth, Michael, Mark Egan, Jill Rutter, Julian McCrae (2018), *Behavioural Government. Using behavioural science to improve how governments make decisions*, Behavioural Insights Team; Management Concepts (2016), *Successful Change Management Practices in the Public Sector. How governmental agencies implement organizational change management*.
[50] World Bank (2016), p. 180.
[51] Garicano, Luis and Paul Heaton (2010), 'Information Technology, Organization, and Productivity in the Public Sector: Evidence from Police Departments', *Journal of Labor Economics*, 28 (1).
[52] World Bank (2016), p. 180.
[53] Wallace, Nick and Daniel Castro (2018), *The Impact of the EU's New Data Protection Regulation on AI*, Center for Data Innovation, March 27.

5.3 Data Size, Infrastructure, Analysis, Security, and User Interface Requirements

Detailed technical requirements for the implementation of any automated processing of text data in the context of programme evaluations at GAC will need to be assessed and tested if and when such procedures are actually being piloted. In the context of this review such detailed assessment is not possible. However, in this section, we present some general observations on how to potentially tackle these issues in the future.

First, the size of text data available to GAC in programme evaluations is likely sufficient for a successful application of most methods reviewed in this report. Many Machine Learning applications, in particular predictive analytics, require large datasets (e.g. with millions of observations) to achieve high levels of accuracy. For most of the methods discussed above, however, such as e.g. topic modelling, document sizes as faced by GAC are likely sufficient to deliver useful results. For example, there were over 900 projects in GAC's MNCH programme, which means that the text corpus is likely to have comprised several thousand documents. Given that each document, in turn, contains large sets of text, a dataset containing a lot of information that could be used to implement topic modelling is easily achieved. In practice, required sample size would have to be assessed on a case by case basis, depending on the tasks at hand.

Second, all of the analytical methods and data management steps covered in this review can generally be implemented using free and open-source software. For statistical programming and text analysis, researchers commonly use either Python or R, both programmes that are open-source. This software covers even sophisticated Machine Learning and NLP approaches. For pre-processing tasks, as described above, similar open-source software is available. OCR can be implemented using Tessaract (see section 5.1.1.). Automated text translation can be done using Google and Microsoft platforms that are freely available. However, to implement the full analytical workflow in such set-ups, some specialist statistical programming or text analysis knowledge is required.

Third, it is possible to combine these statistical analyses with commercial solutions related to cloud computing, should computational capacity be an issue. Depending on the size of the datasets underlying an analysis, some Machine Learning and NLP approaches do require significant computing power that is not easily accessible from desktop or laptop computers. To solve this, R and Python analyses can be integrated in commonly available and scalable cloud computing solutions, such as Amazon AWS.[54] Note also that many governments in OECD countries have dedicated cloud solutions – which might also apply to the Government of Canada and GAC.

Fourth, secure data management that ensures the protection of sensitive data needs to be taken into account when planning the implementation of Machine Learning solutions. It is likely that GAC data needs to be stored either on local servers (i.e. servers of the Government of Canada) or needs to be fully encrypted when being transferred to and from different physical or cloud locations. There are several different solutions available to ensure such secure data handling. Again, Amazon AWS offers encrypted data handling. Whatever the chosen solution, secure

[54] Amazon Web Services, aws.amazon.com/ [accessed on 21.03.2019].

data management would be an important requirement for the successful implementation of this project.

Fifth, and finally, careful user interface development and testing will be important for the successful integration of any of the analytical methods discussed above into programme evaluations at GAC. In general, the teams currently in charge of these evaluations do not have the capacity to implement text analytics using statistical programming e.g. with R or Python. However, they would be the end users of the results of these analyses. Hence, a user interface that allows the evaluation teams to explore the benefits of Machine Learning and NLP in the context of the programme evaluations without needing to code any of the analytics themselves will have to be developed. From a technical point of view, there are several approaches that could be tested in order to achieve this. For example, Python[55] can be used to develop complex user interface set-ups that integrate sophisticated analytical approaches. User interface development will need to be thoroughly tested and piloted in order to ensure that the solution developed indeed serves the needs of the GAC evaluation teams and that it is easy for them to use in their day-to-day work.

5.4 How Could a Pilot and Implementation Plan Look Like?

Including and implementing innovative solutions into existing processes is not a one-off activity, but a complex, often non-linear, procedure that involves a series of steps and actions. The end result of such a process is difficult to predict and may often vary from what was envisioned at the beginning. This applies in particular to explorative projects piloting the use of new technology such as Machine Learning. Should GAC decide to move forward with piloting the use of Machine Learning in programme evaluations, a useful approach to the implementation could be based on the four steps outlined in Figure 3 and elaborated on below. We also provide an example for how a similar process was implemented in the past in a case study in Box 7 below.

Figure 3: Implementation Process

Step 1: In-depth feasibility analysis

Building on the initial analysis of GAC's procedures for programme evaluations (see section 2), an important first step will be to analyse in further detail the existing process whereby GAC generates, compiles, stores and utilises text data in programme evaluations. At this stage, GAC will also have to identify and agree on the steps where Machine Learning and automation could be used. This includes determining how Machine Learning techniques should augment, displace, or replace existing processes and how this is expected to feed into future decision-making procedures.[56] A key question at this stage is to ensure that Machine Learning is indeed the best solution to the problems at hand (e.g. in terms of cost, likely impact and ease of implementation).

[55] GUI Programming in Python, wiki.python.org/moin/GuiProgramming [accessed on 21.03.2019].
[56] USAID 2018.

If an external team is undertaking this analysis, there is a need for deep in-person consultation with GAC staff and observation of existing workflows and processes.

Step 2: Identify team and required capacity

Should GAC decide to proceed with implementation after the in-depth feasibility study, a next step will be to identify a team to lead the process. Generally, the teams that implement public digitisation projects vary in terms of their location, membership, and expertise. The right team structure will depend on the context. However, in all circumstances, GAC will need a broad set of skills and experience.[57] Expertise relating to the particular Machine Learning techniques in question is important, as is access to people with expertise in the variety of analogue components discussed above. The leadership of the project team will be crucial as it will have to mediate and build trust and understanding between people with very different perspectives and professional backgrounds. This mediation will happen within the project team, as well as more widely between different parts of GAC to ensure that new Machine Learning tools are supported and trusted across the organisation.

Should GAC decide to procure Machine Learning solutions from external providers, it is recommendable to push for transparency from technology partners, working to ensure that these providers explain their decisions to GAC staff and that GAC staff learns in this process. Close GAC oversight of the "tech solution providers", especially from GAC evaluation officers, will help ensure that the project remains focused on solving the problem at hand rather than getting lost in the technical solution: "Effective technology solutions require those familiar with the problem to be outspoken, well-informed, and focused on development challenges rather than exclusively on solutions".[58]

Step 3: implement, learn and iterate

The development of useful Machine Learning models in GAC is a complex implementation challenge because it involves the development and testing of new technology in a new setting. USAID provides an outline of a useful three-step implementation process for the development of Machine Learning models: First, you review data (identification of existing data, data cleaning, checking for bias, etc.). Second, you build the model (defining the modelling problem, selecting variables, identifying appropriate algorithms, building the actual model). Third, you integrate it into practice. This involves testing the model, collecting feedback, and going through several iterations. This applies not only to model development, but also to the development of the user interface.

The implementation process for Machine Learning solutions should be iterative and adaptive. Design thinking principles based on continuous testing, feedback, and iteration are helpful.[59] GAC evaluation officers, the users of the Machine Learning solutions, should be extensively involved in this process. To ensure this, the implementation process could draw on user centred design principles,[60] and could

---

[57] See section 5.2 as well as Ford Foundation (2018), *Making the case for a broader definition of "technologist",* Equals Change Blog, October 2018.
[58] USAID 2018.
[59] For an example see IDEOU, Design Thinking, www.ideou.com/pages/design-thinking [accessed on 21.03.2019].
[60] For an overview see David Benyon (2014), *Designing Interactive Systems: A comprehensive guide to HCI, UX and interaction design*, Harlow: Pearson Education Limited, 3rd ed.

include regular meetings between software developers, GAC evaluation officers, and other stakeholders across GAC. Furthermore, in order to save time and money it will be useful to develop Machine Learning tools based on the principle of the minimum viable product (MVP).[61] In an MVP approach, the new digital solution is developed with just enough features to satisfy early users, while the final, complete set of features is only designed and developed after considering feedback from the product's initial users. In this process it is important for GAC evaluation officers to ask about model errors and potential biases to ensure that these are identified and addressed.

Step 4: Ensure sustainability

In the process outlined above, GAC should develop a clear plan for ensuring sustainability of the project, including having access to the infrastructure, technology and skills needed for operation and maintenance of the models. This may initially entail a close collaboration with a technology partner (providing the technical solutions). However, GAC should ensure that they maintain access to the right set of skills and knowledge, and that they build crucial skills in-house over time. It will also be important to ensure that GAC evaluation officers have the skills and knowledge needed to appropriately interpret the outputs of the Machine Learning models developed.

[61] For details see Wikipedia, Minimum Viable Product, en.wikipedia.org/wiki/Minimum_viable_product [accessed on 21.03.2019]; and Eric Ries (2011), *The Lean Startup: How Constant Innovation Creates Radically Successful Businesses*, New York: Crown Publishing Group.

Box 7: Case Study 2: Developing, Testing, and Applying Machine Learning for Public Policy Purposes – the Example of Identifying at Risk Cases for Social Workers in the UK.

Social workers that work with children have a difficult job. They work on many cases simultaneously and are responsible for quickly assessing whether a child is at risk of harm and in need of protection. Often, cases are closed with a recommendation for 'no further action'. However, sometimes these closed cases later re-emerge and escalate in the system. The Behavioural Insights Team has collaborated with social workers in the UK to apply Machine Learning and NLP techniques to predict which closed cases are likely to re-emerge and escalate.[62]

The first step in this process entailed a series of semi-structured interviews with current social workers to understand their decision-making process and interpretation of existing case data. Based on this, the BIT identified the core predictive problem as follows: "given the text of the initial referral and assessment, and structured data relating to the case, could we predict whether the case would be re-referred and escalated if it were closed?"

Second, the BIT proceeded to identify and apply appropriate methods for analysing the available data. The most substantial data on a case are social workers' case notes. These notes are largely unstructured so traditional text analysis techniques such as counting frequency of words were of little use. Instead, the BIT applied topic modelling to extract themes from the case notes. These topics were, together with more traditional structured data on the case, fed into a Machine Learning algorithm[63] to identify cases which had a high risk of returning into the social care system after being closed (i.e. use of Machine Learning for classification). Structured data included child information (such as age, gender and school attendance) and referral information (such as the source of referral).

The third step in this process entailed testing, feedback and adaptation. The BIT sought feedback from and conducted another round of semi-structured interviews with social workers during the testing of the Machine Learning model. This allowed the BIT to ensure appropriate interpretation of the data and that findings were situated in real-world experiences. In was key that social workers understood the reasons behind the algorithm's suggestions on a case in order for them to be able to combine these insights with their own expertise.

Through the use of Machine Learning techniques the BIT was able to build a model that identified a small (6%) set of cases that were closed as being 'high risk'. These high-risk cases contained nearly half of the cases that would later return and escalate. The algorithm identified a very small (0.6%) set of false positives – meaning cases that the algorithm categorised as high-risk but which did not in fact return and escalate. Based on this initial pilot, the BIT has been working with social workers to build a digital tool that allows social workers to see the algorithm's estimated risk for a particular case. This tool can be used by social workers to "get a second opinion" and improve the evidence base to justify spending more time on potentially risky cases. Several stages of feedback and adaptation have been built into the tool development process in order to ensure that the tool offers practical insights for social workers and that it is easy for them to incorporate it into their day-to-day work.

[62] For a full overview see Behavioural Insights Team (2017), *Using Data Science in Policy. A report by the Behavioural Insights Team*, London: The Behavioural Insights Team. For an insightful view on the risks associated with the use of algorithms in child protection, see Virginia Eubanks (2017), *Automating Inequality. How high-tech tools profile, police, and punish the poor*, New York: St. Martin's Press, pp. 127-175.

[63] A gradient boosting decision tree algorithm. For details see Behavioural Insights Team (2017), p. 11.

6 Concluding Remarks

GAC's international assistance supports targeted investments, partnerships, innovation and advocacy efforts around the world to close gender gaps, flight poverty and improve everyone's chance for success. These various interventions produce a large amount of data that can inform GAC's evaluation work - both for identifying which projects to evaluate and to actually evaluate them. However, much of this data is unstructured and text based. Compiling, structuring and analysing this data is currently done manually by staff in PRA. This takes a long time and consumes many resources.

This report has explored the use of text analytics, Machine Learning, and other forms of Natural Language Processing to improve the process with which GAC implements programme evaluations. The review reveals that Machine Learning driven techniques such as topic modelling, sentiment analysis, Word2vec and NER are all likely to be useful. Non-Machine Learning based Information Retrieval systems are also potentially useful. When considering whether to pilot these techniques it is important to note that many of the tasks involved in implementing these techniques do not actually require Machine Learning. Most of the work involves non-Machine Learning based data extraction, cleansing, normalising and wrangling. In essence, most of the work relates to providing structure to unstructured text data, so that it can then be analysed. As in many data-related projects, most of the work is in preparing the data for analysis, rather than the analysis itself.

Moving forward it is important to consider when and how Machine Learning based models can be applied to other contexts and use other data sources. Generally, Machine Learning models reflect the data that they have been trained on. If there is bias in the training data then the model will reflect this bias. Similarly, the model may perform poorly if applied to data that are systematically different from the data it has been trained on. For instance, models trained on data from health focused projects may not be directly applicable to data from climate change focused projects.

The field of "transfer learning" can help GAC navigate these challenges. Transfer learning refers to methods that seeks to reuse models developed for one task as a starting point for a model to be applied to a different task. These methods does not remove the need for training new models, but they can speed up the time that it takes to develop and train a model by reusing existing parts of the model.[64] For example, some of the code from models trained on large data sets can often be used in other models. These new models are then trained on a new dataset. How much of an existing model can be reused will depend on the context, including differences in the tasks that the models will be applied to. Transfer learning is an emerging field and it is still unclear how more complex transfer learning can work. For the purposes of GAC programme evaluations, it is likely that some insights from transfer learning can be applied to contexts where Machine Learning models are applied to different sets of project data, e.g. relating to different GAC programmes.

Exploring and piloting the use of Machine Learning and other types of NLP is an exciting and cutting-edge field for GAC. But there is also a public good element to this work. The models, datasets, and learning that will emerge from this work will be useful not only for GAC but also for a wider audience in the development field and

[64] Curry, Brian (2018), *An Introduction to Transfer Learning in Machine Learning*, KC AI Lab.

beyond. For instance, a searchable database of projects, documents and associated topics could be useful for organisations working on similar projects. Furthermore, the implementation lessons from Machine Learning pilots in GAC are likely to be useful for other development actors exploring the use of Machine Learning, including USAID, DFID, and philanthropic organisations such as the Bill and Melinda Gates and Rockefeller Foundations. Pooling resources and sharing learning as these new technologies are applied in the development and humanitarian space, will be an important way to lower the costs of technology development and adoption, and it will help organisations achieve better results faster.

References

Alghamdi, Rubayyi and Khalid Alfalqui (2015), 'A survey of topic modelling in text mining', *International Journal of Advanced Computer Science and Applications*, 6(1).

Amazon Web Services, aws.amazon.com/ [accessed on 21.03.2019].

Anyline, What is Optical Character Recognition?, medium.com/@anyline_io/what-is-ocr-why-does-it-make-your-life-easier-209b9fcedec4 [accessed on 21.03.2019].

Banerjee, Suvro (2018), *Introduction to Named Entity Recognition*, Explore Artificial Intelligence / Medium, medium.com/explore-artificial-intelligence/introduction-to-named-entity-recognition-eda8c97c2db1 [accessed on 21.03.2019].

Behavioural Insights Team (2017), *Using Data Science in Policy. A report by the Behavioural Insights Team*, London: The Behavioural Insights Team.

Boyd-Graber, Jordan and Philip Resnik (2010), 'Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation'. *Empirical Methods in Natural Language Processing*.

Broniecki, Philipp, Anna Hanchar, and Slava J. Mikhaylov (2017), *Data Innovation for International Development: An overview of natural language processing for qualitative data analysis*, arXiv: computer science, arxiv.org/abs/1709.05563 [accessed on 21.03.2019].

Brownlee, Jason (2016), *Supervised and Unsupervised Machine Learning Algorithms*, Machine Learning Mastery, machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [accessed on 21.03.2019].

Castle, Nikki (2017), *Supervised vs. Unsupervised Machine Learning*, ORACLE + DATASCIENCE.COM, www.datascience.com/blog/supervised-and-unsupervised-machine-learning-algorithms [accessed on 21.03.2019].

Curry, Brian (2018), *An Introduction to Transfer Learning in Machine Learning*, KC AI Lab, medium.com/kansas-city-machine-learning-artificial-intelligen/an-introduction-to-transfer-learning-in-machine-learning-7efd104b6026 [accessed on 21.03.2019].

David Benyon (2014), *Designing Interactive Systems: A comprehensive guide to HCI, UX and interaction design*, Harlow: Pearson Education Limited, 3rd ed.

Eubanks, Virginia (2017), *Automating Inequality. How high-tech tools profile, police, and punish the poor*, New York: St. Martin's Press.

Ford Foundation (2018), *Making the case for a broader definition of "technologist",* Equals Change Blog, October 2018.

Garicano, Luis and Paul Heaton (2010), 'Information Technology, Organization, and Productivity in the Public Sector: Evidence from Police Departments', *Journal of Labor Economics*, 28 (1).

Google Machine Learning Services, *What is machine learning*, cloud.google.com/what-is-machine-learning/ [accessed on 21.03.2019].

GUI Programming in Python, wiki.python.org/moin/GuiProgramming [accessed on 21.03.2019].

Hallsworth, Michael, Mark Egan, Jill Rutter, Julian McCrae (2018), *Behavioural Government. Using behavioural science to improve how governments make decisions*, Behavioural Insights Team.

Hammer, Cornelia L., Diane C. Kostroch, Gabriel Quiros, and STA Internal Group (2017*), Big Data: Potential, Challenges and Statistical Implications*, IMF Staff Discussion Note, Washington D.C.

Helmers, Lea, Franziska Horn, Franziska Biegler, Tim Oppermann, Klaus-Robert Muller (2019), *Automating the search for a patent's prior art with a full text similarity search*, arxiv.org/abs/1901.03136 [accessed on 21.03.2019].

Hu, Yueying, Ke Zhai, Vladimir Eidelman, Jordan Boyd-Graber, 'Polylingual tree-based topic models for translation domain adaptation', Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers, pp. 1166-1176.

IDEOU, Design Thinking, www.ideou.com/pages/design-thinking [accessed on 21.03.2019].

Jacobi, Carina, Wouter van Atteveldt and Kasper Welbers (2016), 'Quantitative analysis of large amounts of journalistic texts using topic modelling', *Digital Journalism*, 4(1).

Jain, Anuja P and Padma Dandannavar (2016), *Application of machine learning techniques to sentiment analysis*, 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).

Karani, Dhruvil (2018), *Introduction to Word Embedding and Word2Vec*, Towards Data Science, towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa [accessed on 21.03.2019].

Kiser, Matt (2016), *Introduction to Natural Language Processing (NLP)*, Algorithmia, https://blog.algorithmia.com/introduction-natural-language-processing-nlp/ [accessed on 21.03.2019].

Lau, Jey Han, David Newman and Timothy Baldwin (2014), *Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality*, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.

LDA Topic Models, www.youtube.com/watch?v=3mHy4OSyRf0 [accessed on 21.03.2019].

Li, Susan (2018), *Named Entity Recognition with NLTK and SpaCy*, Towards Data Science / Medium, towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da [accessed on 21.03.2019].

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, Dustin Tingley (2015), 'Computer-Assisted Text Analysis for Comparative Politics', *Political Analysis*, 23(25).

Management Concepts (2016), *Successful Change Management Practices in the Public Sector. How governmental agencies implement organizational change management*.

Manning, Christopher D., Prabhakar Raghavan & Hinrich Schutze (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

Oguejiofor Chibueze (2018), *NLP For Topic Modelling Summarization of Legal Documents*, Medium / Towards Data Science, towardsdatascience.com/nlp-for-topic-modeling-summarization-of-legal-documents-8c89393b1534 [accessed on 21.03.2019].

Ostrof, Frank (2006), *Change Management in Government*, Harvard Business Review, May issue

Plattner, Hasso, Christoph Meinel and Larry Leifer (Eds.) (2011), *Design Thinking. Understand – Improve – Apply*, London: Springer Heidelberg Dordrecht.

Priya Dwivedi (2018), NLP: Extracting the main topics from your dataset using LDA in minutes, Towards Data Science, towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925 [accessed on 21.03.2019].

Ries, Eric (2011), *The Lean Startup: How Constant Innovation Creates Radically Successful Businesses*, New York: Crown Publishing Group.

Skymind AI, *A beginners guide to Word2vec and neural word embeddings*, skymind.ai/wiki/word2vec [accessed on 21.03.2019].

Snyder, Benjamin and Regina Barzilay (2007), *Multiple Aspect Ranking using the Good grief Algorithm*, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, in HLT-NAACL.

Soda PDF anywhere, www.sodapdf.com [accessed on 21.03.2019].

Tesseract-ocr, github.com/tesseract-ocr/ [accessed on 21.03.2019].

TranslateR, cran.r-project.org/web/packages/translateR/translateR.pdf [accessed on 21.03.2019].

UN Global Pulse (2016), *Integrating Big Data into the monitoring and evaluation of development programmes*, New York: UN Global Pulse.

UN Global Pulse, *Bringing in people's voices from radio content analysis to respond to a refugee crisis*, See www.unglobalpulse.org/projects/bringing-peoples-voices-radio-content-analysis-respond-refugee-crisis [accessed on 21.03.2019].

UN Global Pulse, Informing governance with social media mining, debates.unglobalpulse.net//uganda/ [accessed on 21.03.2019].

UN Global Pulse, *Understanding Immunisation Awareness And Sentiment Through Analysis Of Social Media And News Content*, www.unglobalpulse.org/understanding-immunisation-awareness-through-social-media [accessed on 21.03.2019].

USAID (2018), *Reflecting the Past, Shaping the Future: Making AI Work for International Development*, Center for Digital Development USAID.

Vries, Erik de, Martijn Schoonvelde, Gijs Schumacher (2017), *Lost in Translation? Evaluating the usefulness of machine translation for bag-of-words text models*, The Euengage working paper series.

Wallace, Nick and Daniel Castro (2018), *The Impact of the EU's New Data Protection Regulation on AI*, Center for Data Innovation, March 27.

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz (2018), *AI Now Report 2018*, AI Now Institute, New York University.

Wikipedia, Minimum Viable Product, en.wikipedia.org/wiki/Minimum_viable_product [accessed on 21.03.2019].

Wikipedia, Word2vec, https://en.wikipedia.org/wiki/Word2vec [accessed on 21.03.2019].

World Bank (2016), *World Development Report 2016: Digital Dividends*, World Bank Group, Washington D.C.

Xu, Joyce (2018), *Topic Modelling with LSA, PLSA, LDA & Ida2Vec*, Medium: medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 [accessed on 21.03.2019].