Eliecer Rodrigo Díaz Díaz

AI summarizer assistant

Metropolia University of Applied Sciences

Bachelor of Engineering

Degree Programme in Information Technology

Bachelor’s Thesis

26 November 2021

Abstract

Author:	Eliecer Rodrigo Díaz Díaz
Title:	AI summarizer assistant
Number of Pages:	36 pages
Date:	26 November 2021

Degree:	Bachelor of Engineering
Degree Programme:	Information Technology
Professional Major:	Cloud computing and IoT
Supervisors:	Janne Salonen, Senior Lecturer and Head of School Taavi Erkkola, Senior adviser at UNAIDS, heading the workstream responsible for delivering the annual Global AIDS Report

This thesis was done in cooperation with the Joint United Nations Program on HIV/AIDS or UNAIDS. The mission of UNAIDS is to inspire the world in achieving universal access to HIV support. UNAIDS collects data from 193 United Nations' member states, and tracks progress towards global AIDS targets writing annual reports. UNAIDS manifested the need for assistance on this task seeking expertise in implementing efficient AI and NLP solutions to improve working with vast amounts of unstructured data, which is typically highly laborious. A solution in the form of PoC was built using AI tools such as Natural Language Processing NLP and Optical Character Recognition OCR. These tools work by summarizing text and extracting numerical information from tables embedded in reports, respectively. These functions were integrated with a user interface UI to facilitate the operation by UNAIDS experts. In addition, visualization functions were also developed to monitor monetary investments per country and the number of people living with HIV. Similarly, sentiment analysis and population pyramids charts were also included as visual capabilities in the solution. Finally, the critical functionality developed was the "integration" tab which puts the results from all the other functionalities in a logical output resembling a dashboard; experts can evaluate the HIV situation for a specific year and country. As a final deliverable result, a PoC was presented to UNAIDS experts. UNAIDS indicated that using this AI summarizer assistant solution and further AI implementations might save an estimated 30–40% of time spent on data analysis.

Keywords: Artificial intelligence, AI, natural language processing, NLP, transformers, BERT, HIV, UNAIDS

Table of Contents

List of Abbreviations

AI	Artificial intelligence
ARV	Antiretrovirals
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Networks
DL	Deep Learning
RNN	Recurrent Neural Networks
GAM	Global AIDS Monitoring (GAM)
GPT	Generative Pre-training Model
GRU	Gated Recurrent Unit
LSTM	Long Short Term memory
JSON	JavaScript Object Notation
NASCOP	National AIDS/STI Control Programme
NER	Named-entity recognition techniques
NLO	Natural Language Processing
PoC	Proof of Concept
OCR	Optical Character Recognition
SDK	Software Development Kit
S3	Storage version 3 from AWS
STI	Sexual transmitted infection
UNAIDS	Joint United Nations Programme on HIV/AIDS.

Acknowledgments

Primarily, I want to thank for this collaboration with Taavi Erkkola, senior representative and advisor from UNAIDS, for allowing me to do this thesis in such an inspiring working team as UNAIDS in Geneva. I also would like to manifest my gratitude to Kai Knuutila and Katariina Mahkonen from Silo AI for endless support and exciting discussions during this thesis development. To Janne Salonen, my supervisor, who always trusted me on this endeavour. Finally, to my wife Johanna, for absolutely endless comprehension and support along this study journey.

1 Introduction

Originally, Artificial intelligence (AI) was conceived as a field of computer science concerned with constructing intelligent agents as computer algorithms and understanding the generation of intelligence as a technological phenomenon [1]. Nowadays, the search for this level of intelligence involves several fields such as mathematics, linguistics, philosophy, psychology, and biological sciences. Within the linguistic scientific field, the ultimate objective is to develop software that can understand and generate human languages [2].

On the technical side, AI methods can be grouped into two main groups: 1. the classic or rule-based, where an algorithm is constructed based on boolean logic. 2. the statistical and machine learning approach: where examples of data with results are presented to the algorithm, which produces an internal representation (learning) of the relationship input-result and then is ready to predict [3] (Figure 1).

1.1 Thesis structure

Chapter One, “Introduction”, presents the concepts of Artificial Intelligence and Natural Language Processing NLP. There is a short literature review on the most relevant techniques for NLP. Additionally, the study case is presented.

Chapter two presents and describes the technology used and the solution developed “AI summarizer assistant”.

Chapter three presents the operation of the “AI summarizer assistant” with examples of procedures and screenshots.

Chapter four discusses the solution performance, the impression from the client UNAIDS and further points to improve.

1.2 Review: From machine learning to natural language processing

Machine learning models are usually described as black boxes with an understanding of the structural component parts such as matrices, decision trees, optimization algorithms, but less on the emergent properties from these algorithms exhibited when they predict [1]. A classic example of this is Deep Learning DL algorithms, whose body plan is a dense neural network (Figure 2A).

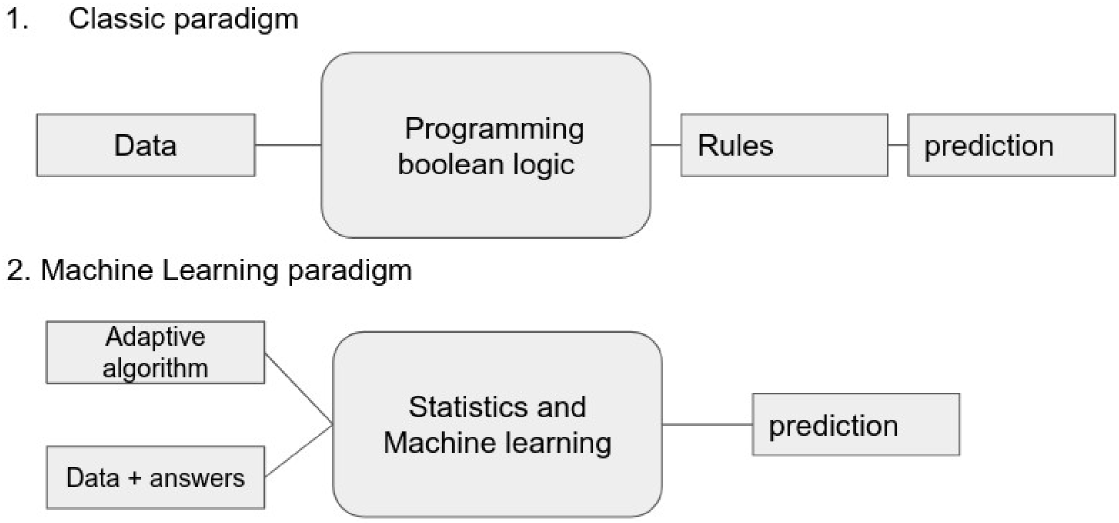

1. Classic paradigm
Data
Programming boolean logic
Rules
prediction
2. Machine Learning paradigm
Adaptive algorithm
Data + answers
Statistics and Machine learning
prediction

Figure 1. Diagram of paradigms to develop adaptable models and AI. The first diagram shows the classic paradigm pipeline, data is fed, and the user needs to program the logic to achieve expected predictions, e.g. programming chess game. The second diagram on the bottom shows that data and output feed an adaptable algorithm. Then the algorithm changes its internal parameters and generates predictions on new data.

These neural networks comprise neurons grouped in layers of neurons. Each neuron contains an activation function that acts as a threshold propagating or stopping the information to the subsequent layers of neurons (Forward propagation). The network ends with a cost function, which evaluates a given prediction using an optimizer; if the forecast deviates from the target, information is sent back through the network in reverse order, from the last to the first layer (Backward propagation). On each iteration, including a forward and backward propagation, the neuron parameters (weights values) get corrected, improving the accuracy of the prediction. Finally, the correction ends when the prediction matches the labels presented; at this point, it is said that the neural network has been trained [3, 4]. Deep Learning uses programming API libraries such as TensorFlow, PyTorch, Transformers, etc., which makes the training of models for software such as R and Python [2]. The primary uses for Deep Learning are for computer vision problems such as recognizing images, faces and detecting objects; these problems are tackled with a specific neural network architecture known as Convolutional Neural Network (CNN) (Figure 2B) [5, 6]. Additionally, when DL deals with languages, such as spelling correction, grammar checking, text translation, text understanding, and text generation, they use a more sophisticated architecture known as Transformers [7] Figure 3.

Language processing is not an easy task to be handled by computers, which is caused by the intrinsic features of any language, such as the presence of multiple grammar exceptions and fast evolution and creation of new words. The first non-deep learning attempts to interpret language data were based on text fragmentation; this is to break paragraphs into sentences and sentences into words (parsing), for example, using the ELIZA NLP tool [8]. The objective was to analyze the dependency probabilistically among words and generate new sentences. Later, the “Frames” framework emerged, which is a further development from ELIZA that incorporates a contextual data structure of texts to the previous method, refining the sentence generation [8]. A parallel attempt to sentence generation was a technique called “Fine State Machines”, which utilized gates containing sentences connected deterministically.

For each input gate, there is a specific output gate for text generation. Notice that output gates can act later as input gates creating a closed loop of input-output without generating a new text [2].

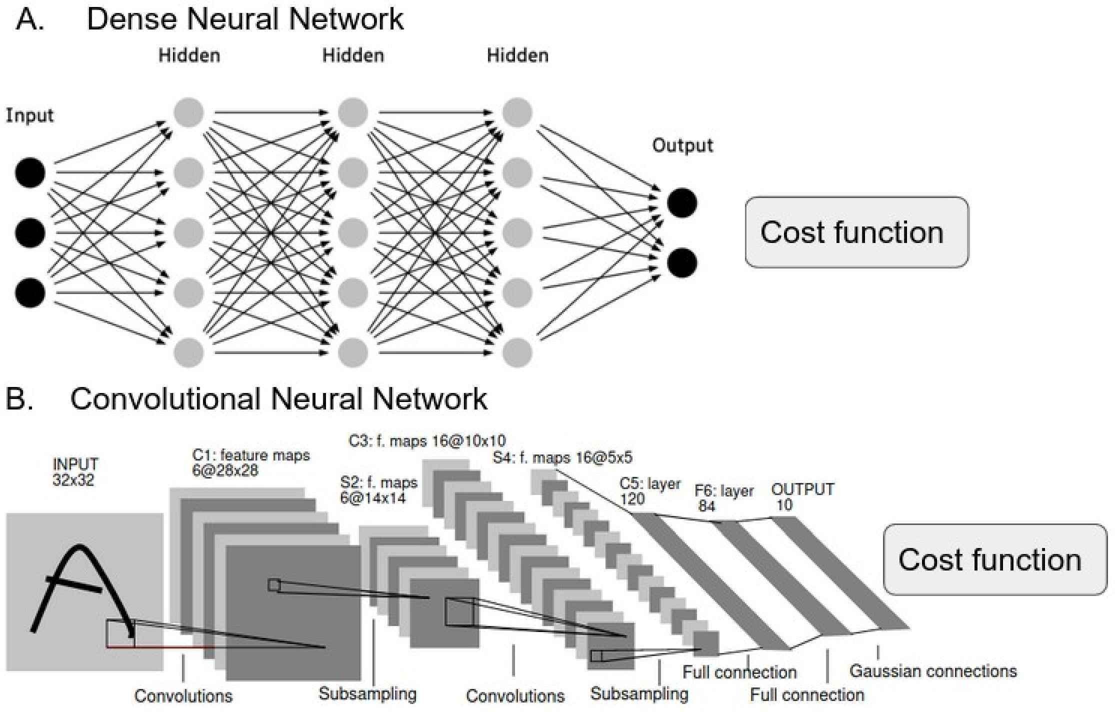


Figure 2. A. The diagram shows an example of a Dense Neural network, where circles are neurons, and arrows indicate their connection. **B**. This diagram represents the Convolutional Neural Network. The photo was taken from LeCun et al. [5].

Often text analysis in the form of meaning extraction or text generation exhibit some inherent complexities related to the position of words within a sentence; this implies that a change in the order of the word may change the meaning of a sentence. Similarly, sentences are in a specific semantic order within a paragraph and apply paragraphs within the text. Additionally, there exist the problem of using synonyms and the problem of using different expressions that have the same meaning within sentences. The traditional deep learning DL methods, such as dense and convolutional neural network architectures, have been tested to handle these problems [3]. These techniques exhibit some success for text classification tasks, but they fail on text summarization and generations. The problem seems to be that these techniques use text data as a bag of words and not as a sequence of words [2, 3].

For this reason, a new set of neural networks were designed to handle sequences, time-series data such as recurrent neural networks RNN including ‘Long-short term memory’ LSTM and gated recurrent unit GRU [2]. RNN incorporates feedbacks between gates to allow and stop the flow of information. LSTM add memory gates to maintain the local and global semantic of a text. GRU is a variation of LSTM whose “units” and “gates” control how much information is kept from previous inputs within the neural network. Nevertheless, they all failed to maintain the hierarchical spatial dependency of words in the text so that the semantics do not get lost [2, 3, 9, 10]. All these neural networks are based on the idea of “supervised learning”, where algorithms are presented with examples of the correct output they have to achieve. The counterpart of supervised learning is unsupervised learning, where algorithms can find patterns within a non-labelled data set. For NLP, a particular set of algorithms have been developed named “autoencoder”. These comprise hidden layers that encode input vectors and add size constraints to them, for example, compressing the output per layer like PCA operation [2].

These techniques led to the development of the “Transformers”, a more complex and sophisticated type of neural network architecture, which also takes the new elements developed for LSTM, GRU and autoencoders. They are made of two stacks, the encoder and the decoder stack [7, 9, 10]. The encoder stack receives the text information where each word is converted as a token forming a numeric vector. Then, the encoder assigns importance to each word received according to the sentence’s order and context. The “attention layer” unit does this. Later, data is normalized and passed through a dense neural network to end up in the decoder stack [9, 10, 11].

Transformers architecture handles language type data such as text, speech among others. They were initially designed to do text translations, but their success has promoted the development of advanced language models, such as BERT and GPT models. These two are successful examples based on transformer architecture and whose applications are related to text generation and summarization. In this project, I utilize the summarizing capabilities of a type of transformer to build a summarizer functionality [9].

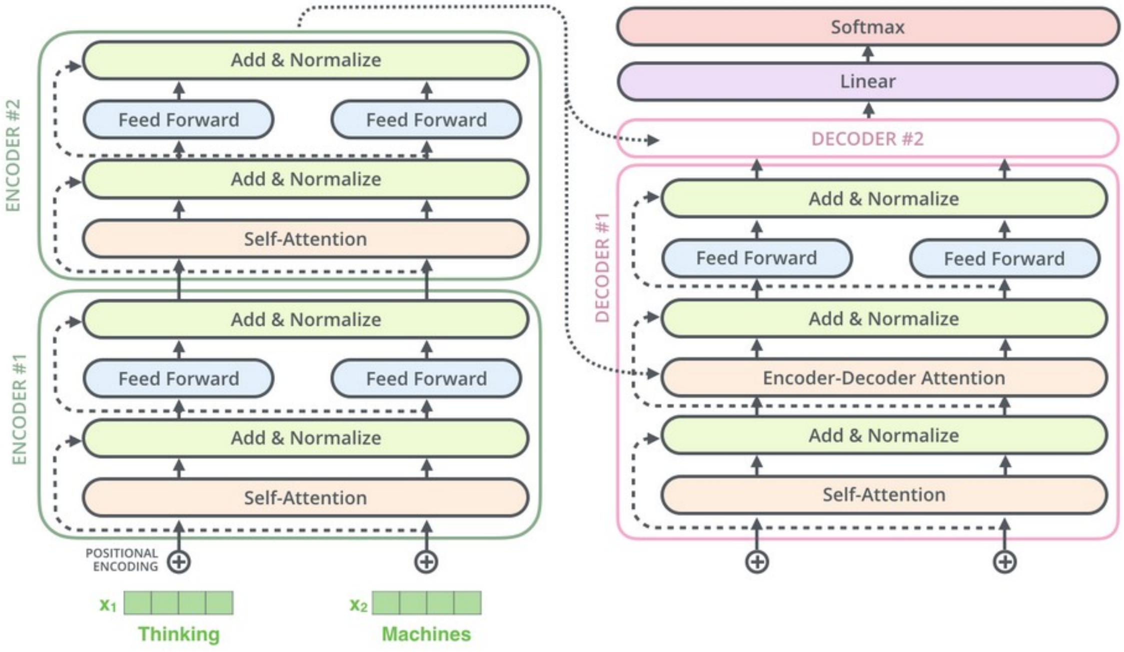


Figure 3. Schema of a transformer. Encoder and decoder same components: self-attention, Normalization, Feed Forward layers. The encoder stack is located on the left, while the decoder stack is located on the right side—Diagram is taken from Alammar [9].

1.3 Optical Character Recognition (OCR)

In addition to text summarization, there is a need to extract digital information from documents and convert them into a more manipulative format. This need has promoted the development of Optical Character Recognition (OCR). Often OCR systems utilize a camera. Usually, these systems take a photo and parse information from the image to text. Modern versions of OCR software use NLP language tools such as LSTM architectures, e.g. “tesseract 4” trained to recognize lines of text instead of single characters [12]. For example, some OCR systems can recognize document sections and digitally reconstruct the text in the same format. This is the case of Amazon Textract OCR, which uses the tesseract engine as a back-end to read and process any document, accurately extracting text, handwriting, tables, and other data without manual effort. Developers can automate document processing and take action on the information extracted using SDK Boto3 [13] and embedded in any software.

1.4 Study case UNAIDS

This thesis aims to collaborate with The Joint United Nations Programme on HIV/AIDS or UNAIDS developing AI technologies to improve data summarization and analysis specifically for their reporting tool Global AIDS Monitoring GAM [14]. UNAIDS must annually deliver this strategic report to diagnose the status and progress made of the AIDS epidemic at the local, national, regional and global levels. This reporting process comprises extensive data collection on different topics related to AIDS such as epidemiology, finance, discrimination and legislation. For example, to build reports, it is necessary to search the internet for relevant official and non-official information such as scientific papers and NGO reports; once the information is gathered, UNAIDS analysts read and assess whether or not the contained information is relevant. This has been described as a demanding and time-consuming task, and there would be a great benefit in building an intelligent search function looking for relevant paragraphs over documents.

Additionally, and with the benefits of AI tools, documents considered helpful must be gathered to be later summarized. For example, intelligent recognition of essential papers and paragraphs can be implemented by a keyword search engine implemented using libraries such as SpaCy in Python [15], or tools like ElasticSearch, coupled with a language model, e.g. BERT [10, 15, 17]. In addition, numerical information needs to be extracted from tables and figures. Here, implementing an OCR system that allows the extraction of tables from PDF files or any other document would reduce the time in passing information from one document to another.

1.5. Objectives

This thesis describes the development and operation of an AI software solution that is delivered as a Proof of Concept PoC to UNAIDS. This solution aims to facilitate data scanning, extraction, text summarization and reporting for UNAIDS officers. Together with UNAIDS, experts agreed on the AI-specific functionalities that will help them in their reporting tasks. The identified AI-functionalities become the following objectives:

- Keyword search development
- Text summarization development engine based in a AI language model
- Table extraction OCR development tool transforms it into an Excel file.
- Quantitative trends display
- Development of an user interface where all these functionalities are integrated.

2 Methods

A solution was developed in a direct dialogue with UNAIDS experts. This dialogue was organized by a series of meetings where experts presented their current methods to gather information from several reports received from UNAIDS representative countries. Together we discussed AI possibilities (meeting notes in Appendix 1). Finally, an iterative solution was presented and named “AI Summarizer Assistant”.

2.1 List of technologies to be used in the thesis

1. The work will be developed in Python software version 3.8.
2. An user interface UI as a web page was developed using Flask 2.0 development framework [19] comprising the following functionalities: text search, text summarization, table extraction, quantitative analyses and integration.
3. An advanced pretrained Natural Language Processing library was used to develop a summarization text based on BERT: “distilbert-base-uncased-fine-tuned-sst-2-english” [20].
4. AMAZON Cloud Services, Textract OCR service, was used to parse tables and charts from documents [13].
5. MongoDB to save an important document and summaries [18].

The source code for this solution can be found in my repository in Github [28].

2.2 Solution design and description

In Figure 4, the solution developed is presented, which starts with 1. the document search and summarizer pipeline start with the user who searches and collects relevant documents from the internet specifying keywords, e.g., HIV, discrimination, and Kenya. The papers found will be stored on a local computer, a server folder or S3 buckets from the AWS cloud provider. Then, the users can request the extraction of relevant paragraphs from documents found. Those paragraphs are later stored in a NoSql database, i.e. MongoDB. Later, the users can request a summary of all sections calculating at the same time the sentiment of each one of them. 2. The second functionality corresponds to the ability to parse tables to spreadsheet formats. This is achieved using a connection between the AI summarizer solution and the AWS cloud provider via SDK, e.g., BOTO3; the resulting excel files will be stored either in a local or a dedicated S3 storage bucket. This function utilizes a keyword search engine to find relevant documents and tables within records found. 3. The third section represents the outputs offered by the AI summarizer: i. automatic report summarization; ii. quantitative trends regarding prices, investments, or epidemiological data. Finally, "trends validation" is also represented as the result in the AI summarizer assistant.

2.3 User interface description

All the functionalities presented in Figure 4 were integrated into a simple Python user interface which also acted as API (Figure 5). This API was developed using the Python library Flask, and the following functionalities were embedded inside acting as individual tabs:

2.3.1 Keyword search

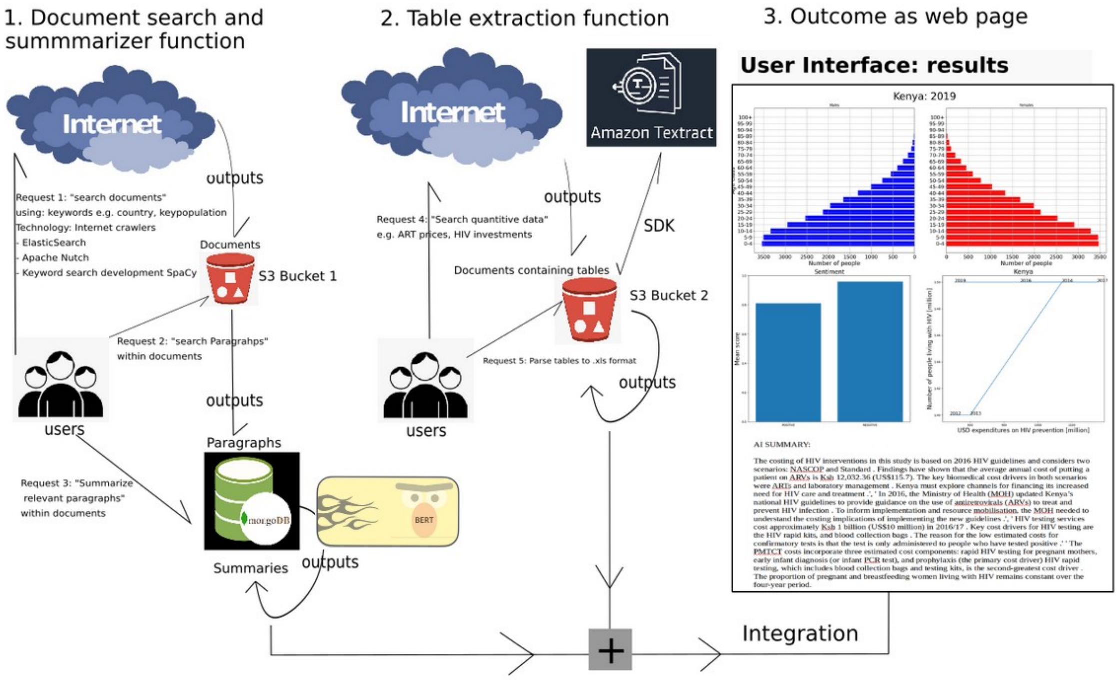
This is the core function of the solution. It was built in Python software using the library SpaCy. This function allows the user to specify keywords for searching paragraphs where the combination of words are located in a document; the user interface retrieves a list of the paragraphs stored in a NoSql database MongoDB (Figure 6).

2.3.2 Text summarization

The list of paragraphs found and stored in MongoDB with the Keyword search function was summarized by the trained transformer (NLP algorithm) "*distilbert-base-uncased-fine-tuned-sst-2-english*" [20, 21]. The model was embedded in the API tab under the tab "summarization". The operation for this function starts when the user chooses one paragraph, and the summarizer retrieves a synthesis of the text. The summaries are stored in a NoSql database MongoDB (Figure 5. point 1 and Figure 6).

2.3.3 Table extraction

To identify correct tables containing the appropriate information, the keyword search function was modified to identify relevant tables within PDF documents (Figure 5, point 2). This functionality was embedded in a "table extraction" route within the flask API developed. The functionality is connected to the AWS Textract service, which takes a table from a PDF document and transforms it into an Excel file stored in a local folder. The tables in Excel format are stored in S3 Bucket within AWS web services, with a copy in the user's local directory (Figure 4. section 2).


1. Document search and summmarizer function
2. Table extraction function
3. Outcome as web page
Internet
outputs
Request 1: "search documents"
using: keywords e.g. country, keypopulation
Technology: Internet crawlers
- ElasticSearch
- Apache Nutch
- Keyword search development SpaCy
Documents
S3 Bucket 1
Request 2: "search Paragrahps"
within documents
users
outputs
Paragraphs
Request 3: "Summarize
relevant paragraphs"
within documents
morgoDB
BERT
Summaries
outputs
Internet
Amazon Textract
outputs
Request 4: "Search quantitive data"
e.g. ART prices, HIV investments
SDK
Documents containing tables
S3 Bucket 2
Request 5: Parse tables to .xls format
users
outputs
User Interface: results
Kenya: 2019
Number of people
Sentiment
Kenya
Mean score
Number of people living with HIV [million]
2012
2013
USD expenditures on HIV prevention [million]
Integration
+

AI SUMMARY:

The costing of HIV interventions in this study is based on 2016 HIV guidelines and considers two scenarios: NASCOP and Standard . Findings have shown that the average annual cost of putting a patient on ARVs is Ksh 12,032.36 (US$115.7). The key biomedical cost drivers in both scenarios were ARTs and laboratory management . Kenya must explore channels for financing its increased need for HIV care and treatment .', ' In 2016, the Ministry of Health (MOH) updated Kenya's national HIV guidelines to provide guidance on the use of antiretrovirals (ARVs) to treat and prevent HIV infection . To inform implementation and resource mobilisation, the MOH needed to understand the costing implications of implementing the new guidelines .', ' HIV testing services cost approximately Ksh 1 billion (US$10 million) in 2016/17 . Key cost drivers for HIV testing are the HIV rapid kits, and blood collection bags . The reason for the low estimated costs for confirmatory tests is that the test is only administered to people who have tested positive .' ' The PMTCT costs incorporate three estimated cost components: rapid HIV testing for pregnant mothers, early infant diagnosis (or infant PCR test), and prophylaxis (the primary cost driver) HIV rapid testing, which includes blood collection bags and testing kits, is the second-greatest cost driver . The proportion of pregnant and breastfeeding women living with HIV remains constant over the four-year period.

Figure 4. Illustration of the AI summarizer assistant solution. For simplicity, the diagram is divided into three sections: 1. Document search (manually) and AI summarizer functions (using BERT NLP model), 2. Table extraction function pipeline, and 3. the outputs from the solutions.

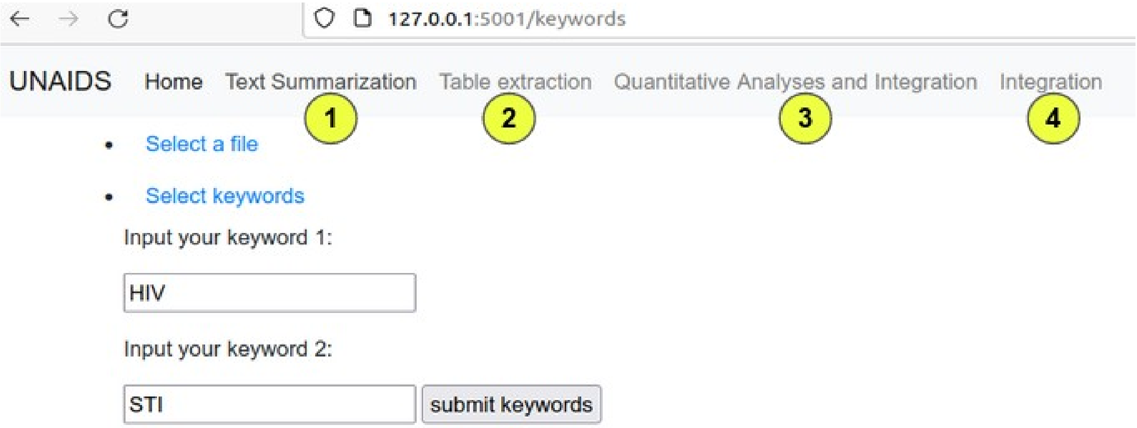
2.3.4 Quantitative analysis

Two analytic functions were embedded in the “quantitative analyses” tab. The first function was developed to extract quantitative data in the form of tabular datasets; for example, expenditures for HIV prevention taken from GAM expenditure data [22]; and the number of new infections over time, whose source was Epidemiology and treatment GAM [23]. The objective was to create connections between these two datasets and visualize the relationship between investment and the number of people living with HIV per country. Such analysis and visualizations were defined together with UNAIDS experts to evaluate the effect of investments of HIV prevalence in the population. Additionally, to complement the previous analysis, population pyramids were presented; users can choose the country and have an idea about the age distribution per gender per country during a particular year. This last function gathers population data from The Department of Economic and Social Affairs and the United Nations [24] and plots population pyramids for a country and a year specified by the user (Figure 5, point 3).

2.3.5 Integration

This This page aims to provide a quantitative view of the population structure and the relationship between mitigation and the number of people living with HIV within a selected country. A sentiment representation on the chosen keywords is presented, and a text summary is provided (Figure 5, point 4). The paragraph saved in MongoDB is extracted, and the sentiment is calculated. The sentiment was calculated using another functionality from the same API provided by Huggingface [20] using the same algorithm "*distilbert-base-uncased-fine-tuned-sst-2-english*" [20, 21].

This integration, the last function, is the most important. It starts clicking the tab "integration" since it takes the results from the previously developed functionalities and puts them together in an organized manner.



Input your keyword 1: HIV

Input your keyword 2: STI

Figure 5. The User Interface was developed using the Flask framework in Python. The UI is simple and contains four tabs: 1. Text summarization, 2. Table Extraction, 3. Quantitative Analyses, and 4. Integration.

3 Operation and results

This section presents the actual operation of the AI Summarizer assistant solution applied to documents concerning HIV general status for Kenya and South Africa.

3.1 Text summarization operation with an example

The summarization was tested using the report by the Kenyan ministry of health for the year 2016 [25], while the table extraction was tested using a report from the Treasury Ministry from South Africa [24]. The document was uploaded to the AI summarizer solution and paragraphs were extracted using the following keywords “HIV” and “ART”. The backend engine searches for the relevant paragraph where these keywords occur and then an AI transformer algorithm “BERT: ***distilbert-base-uncased-fine tuned-sst-2-english***” synthesizes the paragraph. A detailed summary achieved by the transformer is here:

One original paragraph found:

"*The costing of HIV interventions in this study is based on 2016* ***HIV*** *guidelines and considers two scenarios: NASCOP and Standard. The NASCOP scenario was based on national HIV programme targets towards achievement of 90-90-90, and the Standard scenario was based on the assumption that the guidelines would be fully implemented as spelt out in the 2016* ***ART*** *guidelines. In the NASCOP scenario, the population base is assumed to be constant for the four-year costing projections, while the Standard scenario adjusts its population in need to consider incidence, mortality, and population growth rates. However, both scenarios considered two key assumptions: the gains obtained from the reduction in HIV mortality and the reduction in* ***HIV*** *incidence rates. Findings have shown that the average annual cost of putting a patient on ARVs is Ksh 12,032.36 (US$115.7). The Standard scenario costs more than the NASCOP scenario, although differences vary across programme areas. The key biomedical cost drivers in both scenarios were* ***ART****s and laboratory management. Although non-biomedical interventions seemed to be a cost driver, these cost estimates were derived as a proportion of the total cost based on the report of a cost study of* ***HIV*** *treatment conducted in 2013 (U.S. Centers for Diseases Control and Kenya Ministry of Health, (2013), which had an in-depth micro-costing of both medical and non-medical* ***HIV*** *programme interventions. In conclusion, there is an expected escalation in future costs with the implementation of the new 2016* ***ART*** *guidelines. This is due to inclusion of more people living with* ***HIV*** *in care and treatment programmes, and further enhanced by a reduction in* ***HIV****-related mortality. The inclusion of PrEP and PEP as prevention measures might have a significant impact on reducing the* ***HIV*** *incidence rate. This will, however, come with a considerable increase in the resources needed to fund interventions. Therefore, the key policy interventions* *introduced in the new guidelines center on prevention and acceleration of treatment aimed at reducing **HIV** incidence rates and related mortalities. This calls for resources to cover an increased number of people targeted for care and treatment programmes. A lowered mortality rate achieved through an **ART** programme implies increased **HIV** costs, as the number of people living with **HIV** who require **HIV** care will also increase. Kenya must explore channels for financing its increased need for HIV care and treatment. These could include coverage of **HIV** services in health insurance benefits packages (social or private health insurance), increased domestic resource mobilisation (including allocations from national and county governments), and engaging the private sector to play a bigger role in **HIV** financing. " [25]*

And the corresponding summary achieved by the NLP algorithm in JSON format directly pulled from MongoDB database is:

"*[{'summary_text': ' The costing of **HIV** interventions in this study is based on 2016 HIV guidelines and considers two scenarios: NASCOP and Standard. Findings have shown that the average annual cost of putting a patient on **ARV**s is Ksh 12,032.36 (US$115.7) The key biomedical cost drivers in both scenarios were ARTs and laboratory management. Kenya must explore channels for financing its increased need for **HIV** care and treatment.', '_id': ObjectId('61193161c1642b63a200d057')}]*"

3.2 Table extraction operation using OCR

To demonstrate this function, UNAIDS experts suggest extracting tabular data from South African National Treasury [24]. The user operation starts clicking on the "Table extraction" tab function in the UI. The user utilizes the keyword search function to find relevant tables in the documents. When a table containing relevant information is found, the user proceeds to take a screenshot of the table and then submit it using the UI. The backend engine connects to AWS Textract service (OCR) via a SDK (BOTO3). Textract converts the image into a .csv spreadsheet which is delivered to a S3 bucket or to a local folder. An example of the result of this process is illustrated in the figure 7.

3.3 Quantitative analysis operation and result

This functionality attempts to show the population structure for Kenya and South Africa (Figure 8a). It is presented in figure 8 as an illustration of the quantitative analyses carried out. Additionally, it shows the changes in the investment in HIV mitigation and their connection to the number of people living with HIV. (Figure 8b). It was clear that the number of people living with HIV was five times larger in South Africa than Kenya (1.4 million). This number stabilized in 2014, therefore the monetary investments too. In contrast, South Africa has about 7 million people living with HIV, with a faster propagation rate than Kenya and a required ascending mitigation investment.

3.5. Integration

This functionality puts together results from the functionalities developed: i. pyramid charts, ii. sentiment analysis bar charts, iii. investment vs the number of people living with HIV trends and a synthesis extracted from the uploaded document made by DistilBert algorithm. In this sense, the user can know the HIV status in the chosen country (Figure 9).

4. Discussion

The PoC solution developed as PoC and presented to UNAIDS as "AI summarizer assistant" successfully implemented an NLP engine. The summaries were generated using an Artificial Intelligent language model, "distilbert-base-uncased-fine-tuned-sst-2-english", which is one of the many BERT versions available at Hugginface [20]. This is a particular type of neural network model that maintains the semantic coherence in different elements in a text. Additionally, the BERT models allowed extracting the sentiment of paragraphs, opening the opportunities to create a quantitative and graphical comparison of issues across countries or regions.

This engine contains two significant functionalities: i. searching paragraphs function based on keywords within a document. ii. Summary function: the solution was able to produce summaries of the paragraphs found, and even more, make a summary from all the resumes by clicking the “integration” tab. These first two capabilities will reduce the reading workload for UNAIDS experts and speed up identifying documents and paragraphs of interest. According to UNAIDS experts, these functions can be applied to search for discriminatory statements against minorities within different countries’ legislation. In connection to this functionality, the AI summarizer assistant can calculate the sentiment index per paragraph, and get an estimation, without reading the document, how negative or discriminatory the paragraph found is. Therefore, text summarization is possible and to transform the semantic data (summaries) into numerical scores, allowing to perform statistical comparisons among countries and regions.

In addition to the NLP capabilities, a table extraction functionality based on OCR technologies from AWS Textract service was successfully implemented. This function will facilitate data gathering from PDFs and allow data manipulation and combining from different extracted tables to perform new analyses. In addition, UNAIDS experts said that this tool might be helpful for financial data triangulation since it allows to extract and contrast the use of financial prevention investments and their utilization from different sources.

The implementation of quantitative analyses visualization functionality addressed the relationship between national costs for HIV prevention against epidemiological. Specifically, it allows tracking the effects of annual investment in prevention on the number of people living with HIV per country. In this way, governments can estimate and regulate their budgets for next year. UNAIDS experts remarked that the number of people living with HIV is very dependent on their age group in the population of certain countries. To further examine the population structure, population pyramids were added in the visualization. Furthermore, a future improved visualization that can be implemented is the population pyramids for people living with HIV.

The latest functionality acted as a dashboard, putting together all information from previous functions on one web page. In this way, the UNAIDS experts can select a country and then get an overview about the population structure using population pyramids, the investment made and the number of people living with HIV, a sentiment score graph indicating how positive or negative the situation is in that country and a text summary.

4.1 Future directions and improvements

4.1.1 Scale-up

Currently, the solution presented can process one single document at a time (present status). Nevertheless, the idea is to scale it up to process various documents simultaneously. The “AI summarizer assistant” future should first be scaled to read whole storage, which will contain directories named after each country. Within these directories, there will be reports provided by each country. The AI summarizer should loop through each directory and document, extract relevant paragraphs, and save them into the database, tagging the country and keyword used. This means the production of text summaries and charts per country, which can later be utilized for analyses.

4.1.2 Validation capabilities

Validation capabilities refer to the ability to contrast one source of information against another. For example, UNAIDS experts mentioned that they have to examine official and non-official reports regarding HIV. Sometimes the information agrees or disagrees. It is said that the data is validated when unofficial sources agree with official ones. The issue arises when there is an apparent disagreement. In some cases, one can argue that the official source is reliable, while the opposite may be true in controversial government systems. Using historical data of agreement and disagreement, AI or NLP techniques can be trained to distinguish between genuine disagreements and false disagreements, considering factors such as type of government, years of president in power, history of human rights violations, religion, GDP, among other.

4.1.3 Internet searches

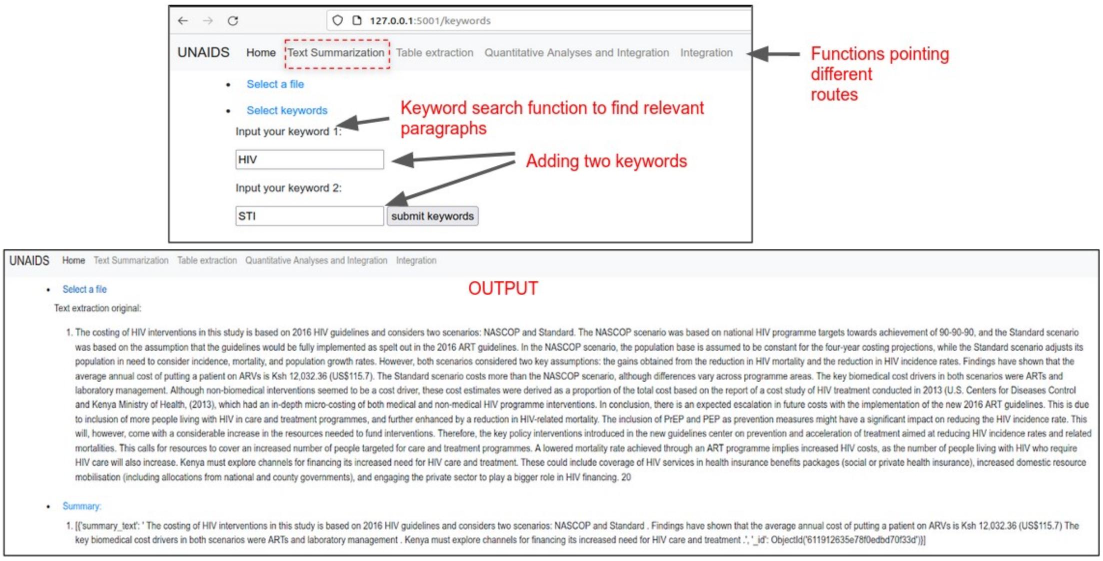
The current solution presented was not implemented to search documents on the internet. Instead, the demo PoC was tested on a reliable source of PDF documents provided by UNAIDS experts. It was identified that a technology such as “elastic search” [17] could be suitable to find documents on the internet. The advantages are the possibility of retrieving large amounts of documents, and additionally, this will also avoid country and regional biases imposed by popular search engines.

Finally, UNAIDS experts expressed their future interest in utilizing NLP capabilities to identify disparities in the testing and treatment cascade based on socio-economic status, race and education. This goal is outside of the scope of this thesis. Still, it is possible to be implemented using named-entity recognition techniques NER [27], which can be trained to find specific terms and values, e.g. unemployment and associated values, such as a percentage. With this type of method, one could maintain accountability in the number of people living with HIV concerning key populations. Adding this future envisioned functionality to the current AI summarizer assistant. After scanning and extracting values from relevant documents, the UNAIDS officer can automatically see the difference between an expected value for the population group (e.g. 20% more cases of people with HIV in the highly educated sector) and determine if the value is realistic or not.

4.2 Conclusion

This project presents reliable proof of the application of AI-based on NLP such as text search, extraction and summarization. In addition, table extraction from PDF was successfully based on OCR techniques from AWS. The user interface developed allowed an easy interaction between the user and the AI functions developed. The functionalities were developed to select a country of interest and year. This may enable fast comparison between years and countries, making it easier to get an overview of the HIV situation in a country.

Finally, a point to keep in mind when discussing capabilities and possibilities with AI for NLP is that these methods evolve rapidly. What is possible now was possibly infeasible or performing poorly a few years back, and what seems complicated now might be very doable after a few years. In this thesis, we bring up the currently available methodologies which could help execute tasks of UNAIDS GAM processes.



Functions pointing different routes

Keyword search function to find relevant paragraphs

Adding two keywords

Input your keyword 1: HIV

Input your keyword 2: STI

OUTPUT

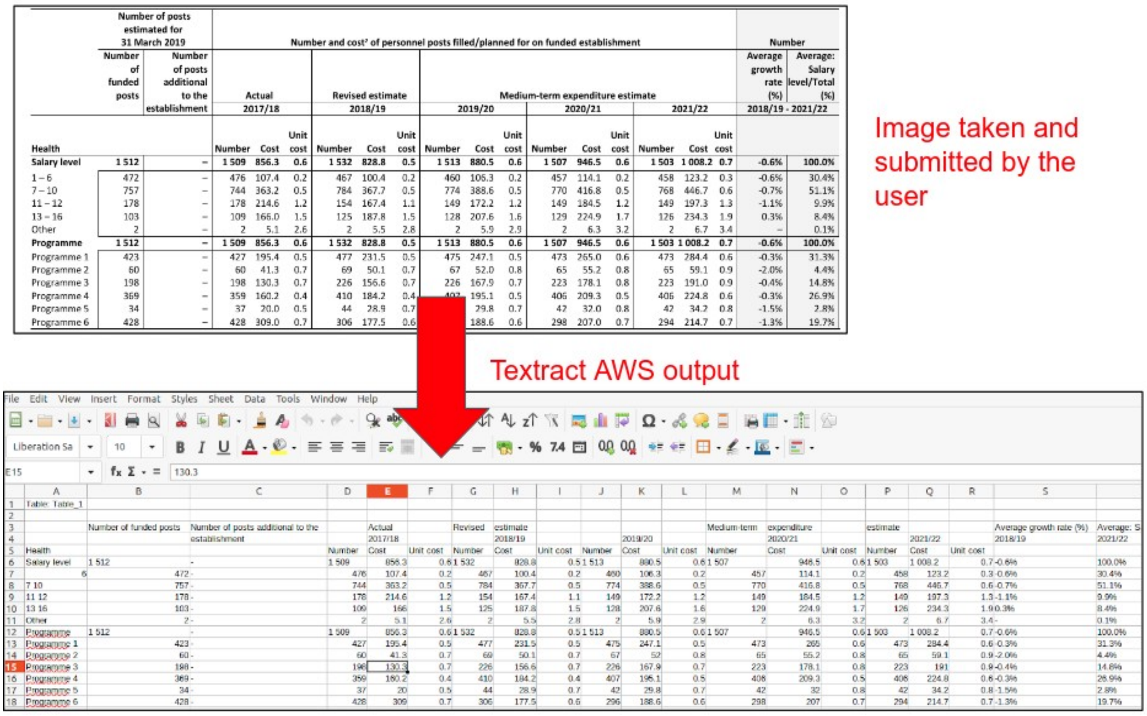
Text extraction original:

1. The costing of HIV interventions in this study is based on 2016 HIV guidelines and considers two scenarios: NASCOP and Standard. The NASCOP scenario was based on national HIV programme targets towards achievement of 90-90-90, and the Standard scenario was based on the assumption that the guidelines would be fully implemented as spelt out in the 2016 ART guidelines. In the NASCOP scenario, the population base is assumed to be constant for the four-year costing projections, while the Standard scenario adjusts its population in need to consider incidence, mortality, and population growth rates. However, both scenarios considered two key assumptions: the gains obtained from the reduction in HIV mortality and the reduction in HIV incidence rates. Findings have shown that the average annual cost of putting a patient on ARVs is Ksh 12,032.36 (US$115.7). The Standard scenario costs more than the NASCOP scenario, although differences vary across programme areas. The key biomedical cost drivers in both scenarios were ARTs and laboratory management. Although non-biomedical interventions seemed to be a cost driver, these cost estimates were derived as a proportion of the total cost based on the report of a cost study of HIV treatment conducted in 2013 (U.S. Centers for Diseases Control and Kenya Ministry of Health, (2013), which had an in-depth micro-costing of both medical and non-medical HIV programme interventions. In conclusion, there is an expected escalation in future costs with the implementation of the new 2016 ART guidelines. This is due to inclusion of more people living with HIV in care and treatment programmes, and further enhanced by a reduction in HIV-related mortality. The inclusion of PrEP and PEP as prevention measures might have a significant impact on reducing the HIV incidence rate. This will, however, come with a considerable increase in the resources needed to fund interventions. Therefore, the key policy interventions introduced in the new guidelines center on prevention and acceleration of treatment aimed at reducing HIV incidence rates and related mortalities. This calls for resources to cover an increased number of people targeted for care and treatment programmes. A lowered mortality rate achieved through an ART programme implies increased HIV costs, as the number of people living with HIV who require HIV care will also increase. Kenya must explore channels for financing its increased need for HIV care and treatment. These could include coverage of HIV services in health insurance benefits packages (social or private health insurance), increased domestic resource mobilisation (including allocations from national and county governments), and engaging the private sector to play a bigger role in HIV financing. 20

- Summary:

1. [{'summary_text': ' The costing of HIV interventions in this study is based on 2016 HIV guidelines and considers two scenarios: NASCOP and Standard . Findings have shown that the average annual cost of putting a patient on ARVs is Ksh 12,032.36 (US$115.7) The key biomedical cost drivers in both scenarios were ARTs and laboratory management . Kenya must explore channels for financing its increased need for HIV care and treatment .', '_id': ObjectId('611912635e78f0edbd70f33d')}]

Figure 6. Above the User Interface showing the search functionality through the input of two keywords. The second box shows the result from the search. The first and more extensive paragraph represents the text where both keywords are present, while the second paragraph represents the summary of the first text.



	Number of posts estimated for 31 March 2019		Number and cost[1] of personnel posts filled/planned for on funded establishment															Number	
	Number of funded posts	Number of posts additional to the establishment	Actual			Revised estimate			Medium-term expenditure estimate									Average growth rate (%)	Average: Salary level/Total (%)
			2017/18			2018/19			2019/20			2020/21			2021/22			2018/19 - 2021/22	
Health			Number	Cost	Unit cost	Number	Cost	Unit cost	Number	Cost	Unit cost	Number	Cost	Unit cost	Number	Cost	Unit cost		
Salary level	**1 512**	**–**	**1 509**	**856.3**	**0.6**	**1 532**	**828.8**	**0.5**	**1 513**	**880.5**	**0.6**	**1 507**	**946.5**	**0.6**	**1 503**	**1 008.2**	**0.7**	**-0.6%**	**100.0%**
1 – 6	472	–	476	107.4	0.2	467	100.4	0.2	460	106.3	0.2	457	114.1	0.2	458	123.2	0.3	-0.6%	30.4%
7 – 10	757	–	744	363.2	0.5	784	367.7	0.5	774	388.6	0.5	770	416.8	0.5	768	446.7	0.6	-0.7%	51.1%
11 – 12	178	–	178	214.6	1.2	154	167.4	1.1	149	172.2	1.2	149	184.5	1.2	149	197.3	1.3	-1.1%	9.9%
13 – 16	103	–	109	166.0	1.5	125	187.8	1.5	128	207.6	1.6	129	224.9	1.7	126	234.3	1.9	0.3%	8.4%
Other	2	–	2	5.1	2.6	2	5.5	2.8	2	5.9	2.9	2	6.3	3.2	2	6.7	3.4	–	0.1%
Programme	**1 512**	**–**	**1 509**	**856.3**	**0.6**	**1 532**	**828.8**	**0.5**	**1 513**	**880.5**	**0.6**	**1 507**	**946.5**	**0.6**	**1 503**	**1 008.2**	**0.7**	**-0.6%**	**100.0%**
Programme 1	423	–	427	195.4	0.5	477	231.5	0.5	475	247.1	0.5	473	265.0	0.6	473	284.4	0.6	-0.3%	31.3%
Programme 2	60	–	60	41.3	0.7	69	50.1	0.7	67	52.0	0.8	65	55.2	0.8	65	59.1	0.9	-2.0%	4.4%
Programme 3	198	–	198	130.3	0.7	226	156.6	0.7	226	167.9	0.7	223	178.1	0.8	223	191.0	0.9	-0.4%	14.8%
Programme 4	369	–	359	160.2	0.4	410	184.2	0.4	407	195.1	0.5	406	209.3	0.5	406	224.8	0.6	-0.3%	26.9%
Programme 5	34	–	37	20.0	0.5	44	28.9	0.7	[illegible]	29.8	0.7	42	32.0	0.8	42	34.2	0.8	-1.5%	2.8%
Programme 6	428	–	428	309.0	0.7	306	177.5	0.6	[illegible]	188.6	0.6	298	207.0	0.7	294	214.7	0.7	-1.3%	19.7%

Image taken and submitted by the user

Textract AWS output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Table: Table_1																			
2																				
3		Number of funded posts	Number of posts additional to the		Actual		Revised	estimate					Medium-term	expenditure		estimate			Average growth rate (%)	Average: S
4			establishment		2017/18			2018/19			2019/20			2020/21			2021/22		2018/19	2021/22
5	Health			Number	Cost	Unit cost	Number	Cost	Unit cost	Number	Cost	Unit cost	Number	Cost	Unit cost	Number	Cost	Unit cost		
6	Salary level	1 512	-	1 509	856.3	0.6	1 532	828.8	0.5	1 513	880.5	0.6	1 507	946.5	0.6	1 503	1 008.2	0.7	-0.6%	100.0%
7	6	472	-	476	107.4	0.2	467	100.4	0.2	460	106.3	0.2	457	114.1	0.2	458	123.2	0.3	-0.6%	30.4%
8	7 10	757	-	744	363.2	0.5	784	367.7	0.5	774	388.6	0.5	770	416.8	0.5	768	446.7	0.6	-0.7%	51.1%
9	11 12	178	-	178	214.6	1.2	154	167.4	1.1	149	172.2	1.2	149	184.5	1.2	149	197.3	1.3	-1.1%	9.9%
10	13 16	103	-	109	166	1.5	125	187.8	1.5	128	207.6	1.6	129	224.9	1.7	126	234.3	1.9	0.3%	8.4%
11	Other	2	-	2	5.1	2.6	2	5.5	2.8	2	5.9	2.9	2	6.3	3.2	2	6.7	3.4	-	0.1%
12	Programme	1 512	-	1 509	856.3	0.6	1 532	828.8	0.5	1 513	880.5	0.6	1 507	946.5	0.6	1 503	1 008.2	0.7	-0.6%	100.0%
13	Programme 1	423	-	427	195.4	0.5	477	231.5	0.5	475	247.1	0.5	473	265	0.6	473	284.4	0.6	-0.3%	31.3%
14	Programme 2	60	-	60	41.3	0.7	69	50.1	0.7	67	52	0.8	65	55.2	0.8	65	59.1	0.9	-2.0%	4.4%
15	Programme 3	198	-	198	130.3	0.7	226	156.6	0.7	226	167.9	0.7	223	178.1	0.8	223	191	0.9	-0.4%	14.8%
16	Programme 4	369	-	359	160.2	0.4	410	184.2	0.4	407	195.1	0.5	408	209.3	0.5	406	224.8	0.6	-0.3%	26.9%
17	Programme 5	34	-	37	20	0.5	44	28.9	0.7	42	29.8	0.7	42	32	0.8	42	34.2	0.8	-1.5%	2.8%
18	Programme 6	428	-	428	309	0.7	306	177.5	0.6	296	188.6	0.6	298	207	0.7	294	214.7	0.7	-1.3%	19.7%

Figure 7. Representation of Table extraction functionality. On top is the targeted table whose numerical information wishes to be parsed into a spreadsheet after calling the Textract function from the UI. One can see a good correspondence between the original and the resultant table.

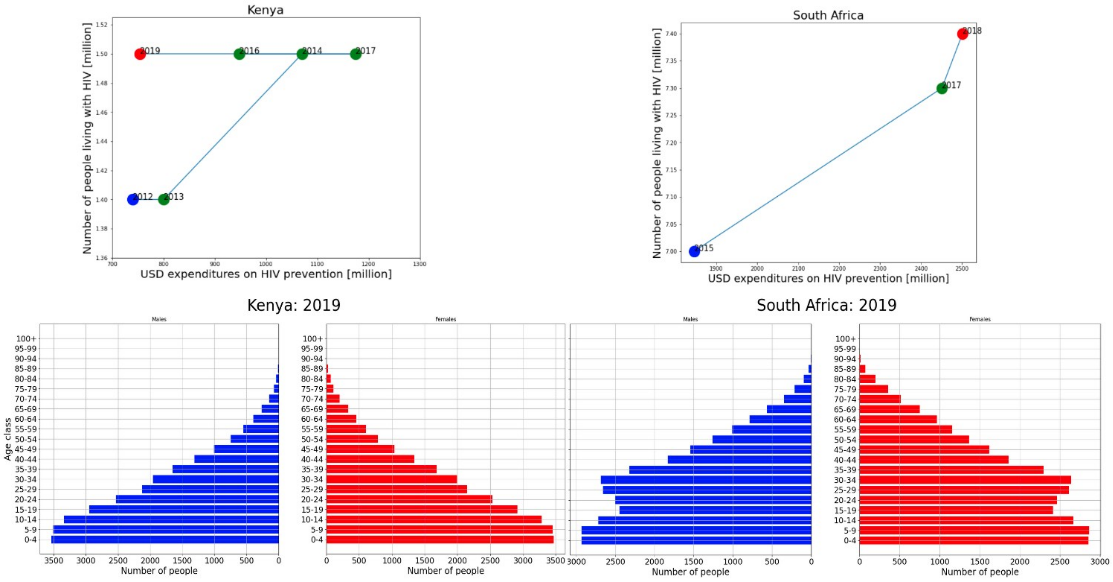
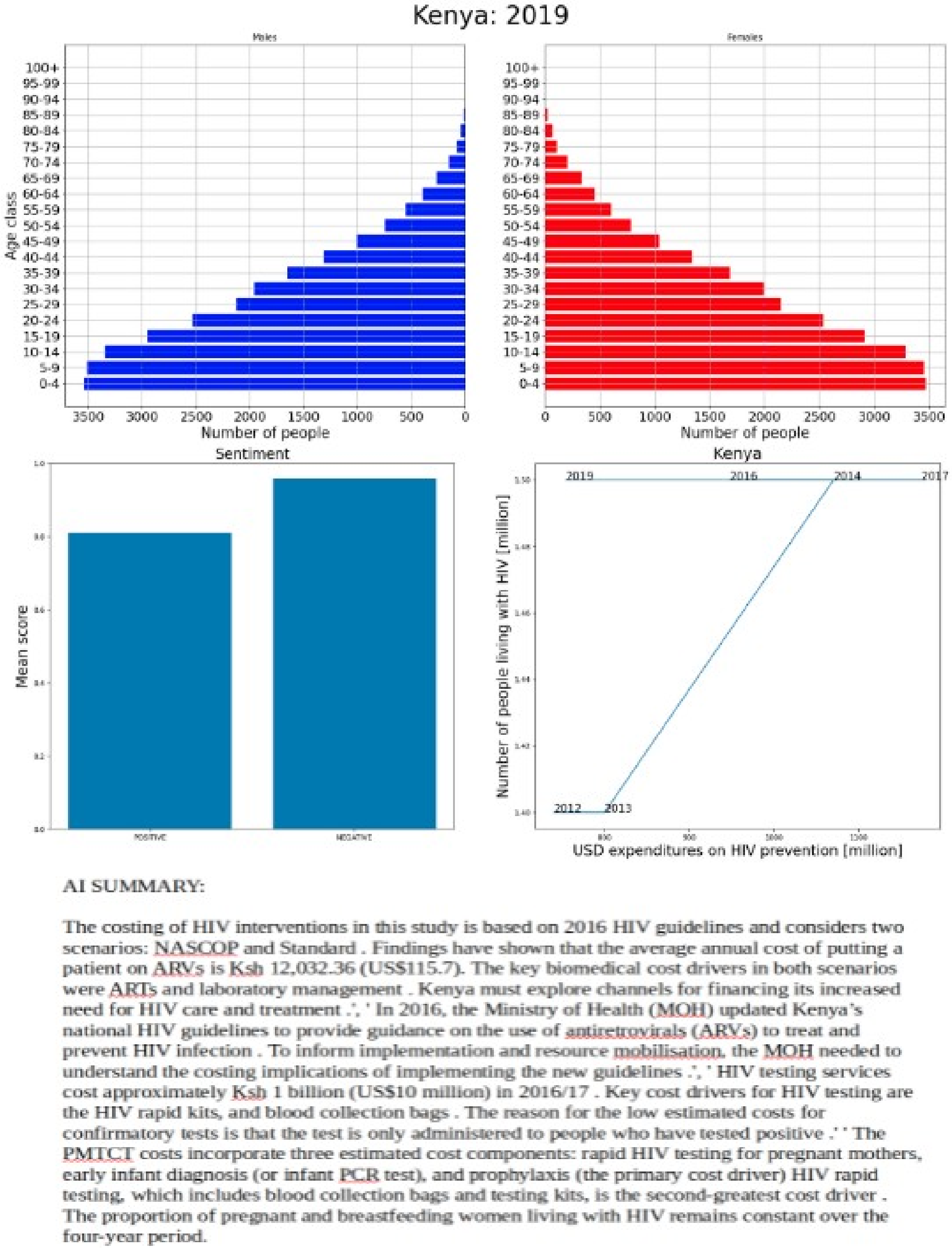


Figure 8. Result of quantitative analyses tab functionality. The figure shows the results for Kenya and South Africa regarding the investments and number of people living with HIV during the period 2011-2019. Notice, for South Africa, the investment for HIV mitigation is about twice the investment for Kenya. At the same time, the number of people living with HIV in South Africa is about four times larger than in Kenya. In addition, the UI shows a comparison between pyramid charts per country and gender, representing the distribution of the population classes per age. It is possible to observe that the population of South Africa has more young adults than (age:25-45) than in Kenya.



AI SUMMARY:

The costing of HIV interventions in this study is based on 2016 HIV guidelines and considers two scenarios: NASCOP and Standard . Findings have shown that the average annual cost of putting a patient on ARVs is Ksh 12,032.36 (US$115.7). The key biomedical cost drivers in both scenarios were ARTs and laboratory management . Kenya must explore channels for financing its increased need for HIV care and treatment .', ' In 2016, the Ministry of Health (MOH) updated Kenya's national HIV guidelines to provide guidance on the use of antiretrovirals (ARVs) to treat and prevent HIV infection . To inform implementation and resource mobilisation, the MOH needed to understand the costing implications of implementing the new guidelines .', ' HIV testing services cost approximately Ksh 1 billion (US$10 million) in 2016/17 . Key cost drivers for HIV testing are the HIV rapid kits, and blood collection bags . The reason for the low estimated costs for confirmatory tests is that the test is only administered to people who have tested positive .' ' The PMTCT costs incorporate three estimated cost components: rapid HIV testing for pregnant mothers, early infant diagnosis (or infant PCR test), and prophylaxis (the primary cost driver) HIV rapid testing, which includes blood collection bags and testing kits, is the second-greatest cost driver . The proportion of pregnant and breastfeeding women living with HIV remains constant over the four-year period.

Figure 9. The final and more critical function of the AI summarizer assistant is "integration". The diagram shows on top the population pyramids for Kenya in 2019 (males and females). The sentiment analyses carried out for Kenya are presented on the bottom left, while at the bottom right is the chart showing the relationship between expenditures and people living with HIV. Finally, a summary of all paragraphs found is presented at the bottom.

References

1. Feldman J. 2001. ‘Artificial Intelligence in Cognitive Science’ in the International Encyclopedia of the Social & Behavioral Sciences, 2: 792–796
2. Håkansson A and RL Hartung. 2020. Artificial Intelligence. Concepts, areas, techniques and applications. Stundentlitteratur AB. Lund, Sweden. 372 pp.
3. Chollet F. 2017. Deep learning with Python. Manning Publications Co. USA. 384 pp.
4. Aloysius N and M Geetha 2017. A review on deep convolutional neural networks. International Conference on Communication and Signal Processing (ICCSP). doi: 10.1109/ICCSP.2017.8286426.
5. LeCun Y, Bottou L, Bengio Y and P Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86: 2278-2324
6. Yamashita R, Nishio M, Do RKG and K Togashi. 2018. Convolutional neural networks: an overview and application in radiology. Insights into Imaging 9:611–629
7. Chernyavskiy A, Ilvovsky Dm and P Nakov. 2021. Transformers: “The End of History” for NLP? arXiv:2105.00813
8. Bassett, Caroline (2019). The computational therapeutic: exploring Weizenbaum's ELIZA as a history of the present. AI & Society 34: 803–812
9. Alammar J. 2019. The Illustrated GPT-2 (Visualizing Transformer Language Models). https://jalammar.github.io/illustrated-gpt2/ Accessed 11 September 2021
10. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and I Polosukhin. 2017. Attention is all you need. ArXiv:1706.03762
11. Radford A, Wu J, Child R, Luan D, Amodei D, and I Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog 1, no. 8

https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf Accessed 11 September 2021

12. Tesseract OCR. https://github.com/tesseract-ocr/tesseract Accessed 11 September 2021
13. Amazon Textract. https://docs.aws.amazon.com/textract/latest/dg/what-is.html Accessed 11 September 2021
14. Global AIDS UPDATE 2021. Confronting inequalities UNAIDS. https://www.unaids.org/en Accessed 11 September 2021
15. SpaCy Industrial-strength Natural Language Processing. https://spacy.io/ Accessed 11 September 2021
16. Lopez LA, Duerr R and SJ Khalsa. 2015. Optimizing apache nutch for domain specific crawling at large scale. In 2015 IEEE International Conference on Big Data (Big Data). 1967-1971 pp
17. Gormley C and Z Tong. 2015. Elasticsearch: the definitive guide: a distributed real-time search and analytics engine. O'Reilly Media, Inc. 196 pp
18. MongoDB. The application data platform. https://www.mongodb.com/ Accessed 11 September 2021
19. FLASK, web development one drop at the time. https://flask.palletsprojects.com/en/2.0.x/ Accessed 11 September 2021
20. The AI community building the future. https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english Accessed 11 September 2021
21. Devlin J, Chang MW, Lee L and K Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv:1810.04805
22. Expenditures GAM. 2021. https://hivfinancial.unaids.org/hivfinancialdashboards.html Accessed 11 September 2021
23. Epidemiology and treatment GAM. 2021. https://aidsinfo.unaids.org
24. Department of Economic and Social Affairs United Nations. 2020. https://population.un.org/wpp/Download/Standard/CSV/ Accessed 11 September 2021

25. Muriithi M, Muchiri S, Maina T, Wanjiru M and C Barker. 2017. Costing the Implementation of the 2016 HIV Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection in Kenya. Ministry of Health. National AIDS and STI Control Programme. 38 pp.
26. Mogojane D. 2019. Estimates of National Expenditure. National Treasury. Republic of South Africa. 60 pp.
27. Song HJ, Jo BC, Park CY, Kim JD and YS Kim. 2018. Comparison of name entity recognition methodologies in biomedical documents. BioMed Eng OnLine 17: 158. https://doi.org/10.1186/s12938-018-0573-6
28. Díaz Díaz ER. AI summarizer assistant source code. Github https://github.com/eliecerecology/Summary_assistant2

Appendices

Appendix. Minutes taken during the meetings with the client

1. Meeting 1: Kick-off meeting 8th April 2021

Participants:
Eliecer Diaz (Project Owner and developer), Katariina Mahkonen (Silo AI), Taavi Erkkola (UNAIDS)

Goal: Introduce the parties, SILO AI and UNAIDS, to initiate the dialogue in which UNAIDS experts expose their different methodologies and constraints to generate reports. UNAIDS is organized in various teams: Financial, Key Populations, and Legislation, which have to report the status of AIDS annually in each country. Participants manifested a burden on collecting, validating and analyzing the data, and they complain that the workload is increment year by year. Therefore, It is pressing to explore new computer methods, i.e. Artificial Intelligence technologies that assist them on data gathering from public sources and platforms, reducing the manual reporting burden.

Notes:

Specifically, the dialogue asks about the possibility that AI can handle the following needs:

- to Synthesize articles (Peer-reviewed or reports)
- to fill up forms and spreadsheets with specific indicators and values
- To compare and contrast different sources of information
- To extract financial information from articles and fill tables

As a reply, SILO AI team led by Eliecer Diaz suggested the following:

- The use of NLP on social media
- Revise Financial Time Series using machine learning
- Check these suggestions grouping the countries in regions, e.g. South America, Africa, European countries, etc.

Conclusion

The meeting concludes with the schedule of three more sessions where the UNAIDS team will present their methods on how they produce reports. The Financial team will deliver the first meeting, then the Key Populations team, and the Legislation team.

2. Introduction to study cases to apply AI. 16April 2021

Meeting 2: Financial Case first meeting.

Participants: Eliecer Diaz (SILO AI), Katariina Mahkonen (SILO AI), Kai Knuutila (SILO AI), Taavi Erkkola (UNAIDS), Deepak Mattur (UNAIDS).

Goal: The meeting had the objective to inform about the common issues gathering financial data from reports in different countries.

Notes:

There are about 95 countries involved in the UNAIDS program. From this program, the financial department is engaged in monitoring prices of antiretroviral treatment ART medicines and the investment efforts made by each government to make accessible AIDS treatment. They analyze these pieces of information, for example, correlating the investment amount with epidemiological data over time, hereafter to evaluate the progress made by each country in reducing AIDS.

Within the issues mentioned by the financial department, the extraction of information from tables within documents is perhaps the most time-consuming. Therefore to parse table contents manually to a spreadsheet is a daunting task. Additionally, NGOs and government reports may show similar or divergent values regarding, for example, ART and national investments within countries. This makes it difficult for UNAIDS to triangulate the acceptable use of the money within a country.

Conclusion

To develop a functionality which allow customer to search relevant tables and parse the information to spreadsheets.

3. Meeting 3: Key Populations. 19 April 2021

Participants: Participants: Eliecer Diaz (SILO AI), Katariina Mahkonen (SILO AI), Kai Knuutila (SILO AI), Taavi Erkkola (UNAIDS), Sonia Arias Garcia (UNAIDS).

Goal: The meeting had the objective to inform about the common issues gathering and summarizing data from reports in different countries regarding advances in access to HIV testing, ART access in minority populations, i.e. gay, under age, racial minorities, low education groups.

Notes:

The client raises the question about automatically detecting possible disparities in HIV testing and treatment cascade among groups of different sexual orientation, socio-economic status, race or education level per country, and access to HIV treatment. For example, this might be initiated by developing a search engine that uses keywords such as STD, malaria, race and AIDS, and extracts paragraphs from specific documents where there is a match. Then these paragraphs should be synthesized and delivered to the officer. In this way, a UNAIDS officer can repeat the same procedure for papers from different countries. With this summarized information, a specialist from UNAIDS can report possible discrepancies in HIV access and treatment among countries.