**Rapid Assessment of UN National Planning Documents using Machine-based Text Analytics**

Executive Summary

**Group: Brahms**

# Table of Content

[**Table of Content**](#_l5yq1c5nd3kd) **2**

[**Abstract**](#_qw392b8h3qq8) **3**

[**Introduction**](#_2bb5d7p69mnb) **4**

[**Data Processing**](#_jpki374z8h09) **6**

[**Methodology**](#_863nidk35ywn) **7**

[Feature Lexicon](#_byf7pceat84k) 7

[Semantic Similarity Model](#_jsabel4ae1z0) 8

[TF-iDF Ranking](#_gw1tnk1goenm) 10

[**Result and Evaluation**](#_3a88n5uaviff) **12**

[**User Interface**](#_29km779lih4c) **13**

[**Limitation and Further Stage**](#_aqe4leftzieh) **14**

[**Appendix**](#_pe4g5h24pu0n) **15**

[Special Statement for Contribution](#_sshmgiqxp63b) 15

# **Abstract**

Each country submits their national planning documents to the united nations every year and one of the UN’s work is to figure out whether the national plan addressed the goal, to be more specific, which sentence in national plans mentioned sustainable development goals. Since there are large number of documents and each file contains thousands of pages, manually doing this work is quite time-consuming and inefficient.

The goal of the project is using machine-based text analysis to match sentences from national planning documents with Sustainable Development Goals (SDG) in a more efficient way.

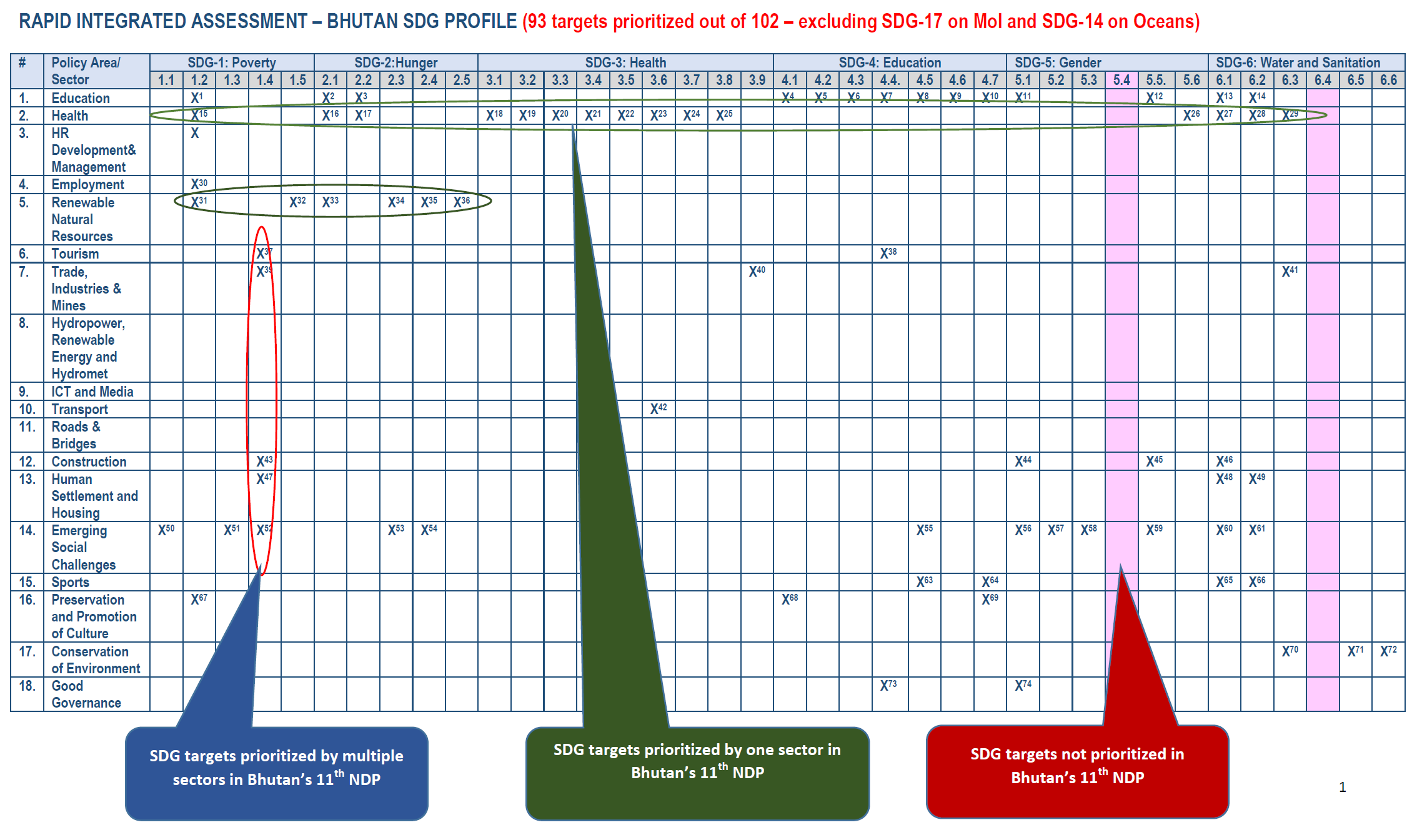
A two-part model was built in our project to achieve our objectives and the final results reached a high accuracy of 88.9%. In addition, to help the user effectively communicate with the machine and manage the output, an user interface was designed to combine the model’s choices with human judgement thus to get the most satisfied outcome.

# **Introduction**

The Sustainable Development Goals, or SDGs, is a part of a wider 2030 Agenda for Sustainable Development. It is a set of 17 "Global Goals" with 169 targets to coordinate efforts to end poverty and hunger, combat inequality and disease, protect the planet, and ensure prosperity for all.

The national planning documents usually contains sectors such as education, health, employment, tourism, transportation etc. However, each country may use various expressions in their documents, which increases the difficulty of this project.

This project aims to match paragraphs or sentences from national planning documents with the SDG targets and come out a final template to visualize the results.



This manually evaluated national plan of Bhutan was made in a template shown above. The objective of this project is to conduct a similar analysis but on a computer-assisted base. This machine-based analysis will serve as a rapid analysis of a national plan which then can be refined by humans.

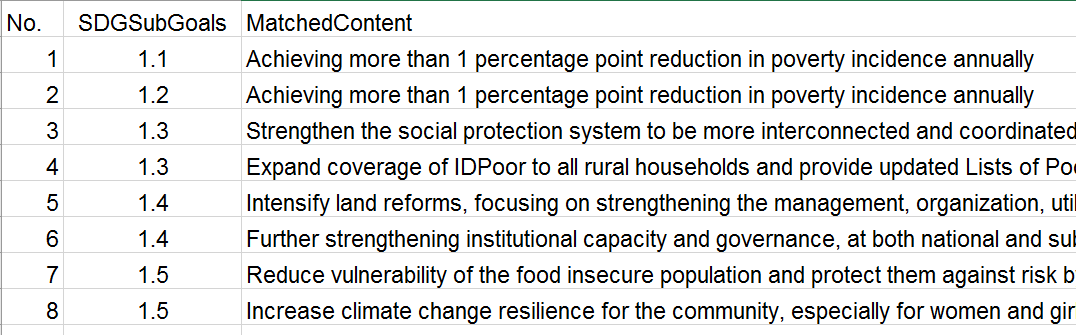
# **Data Processing**

The dataset includes SDG (sustainable development goals) and national plan for each country.

SDG has total 169 subgoals that each of them will be a document needs to match.

For national plans,since the format is pdf, first step of processing is to transform the pdf to txt format and keep the structure at the same time. Then splitting the each national plan into separate sentences and manually labelling sector for them. The sentences are tokenized, stemmed and stopword removal are applied to the words.

The national plan was re-structured into separate sentences shown as below:



The team use one of the National Plans’ reviewed result as golden standard to further evaluate each matched sentence.

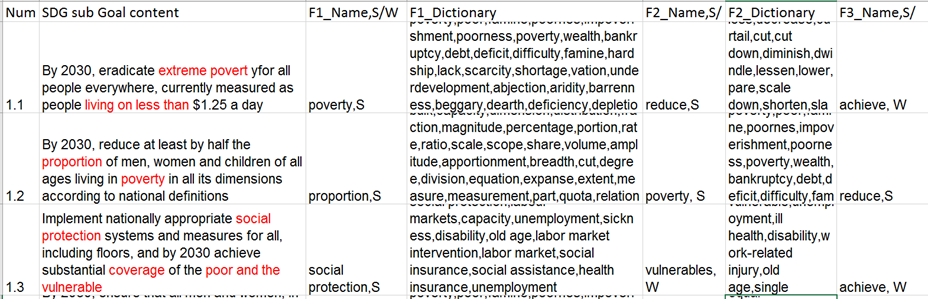
# **Methodology**

After considering applying algorithms in classification, clustering and realizing that no single algorithm can be implement and reach out high accuracy, the team decided to combine both dictionary-based approach with machine-learning approach.

## **Feature Lexicon**

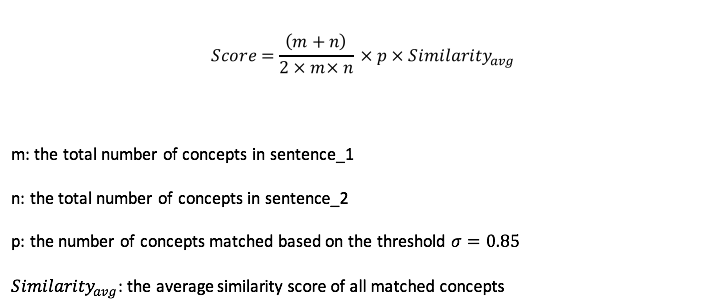
A dictionary of SDG goals was created with at most four keywords are listed for each subgoal as features. Each feature will have a lexicon that any words imply this feature should be contained. Lexicon has two part, one is feature word’s synonyms, the other is the words implied the feature, for example “hurricane” implied feature “extreme climate”.

Features are defined as strong features or weak features. Any sentence extracted should contain one of the words in each strong features’ lexicon simultaneously, that is, if one goal has two strong features, the sentence extracted should contain one of the words from feature number one and another words from feature number two. And the team use weak feature to further capture the sentence matching the goal.



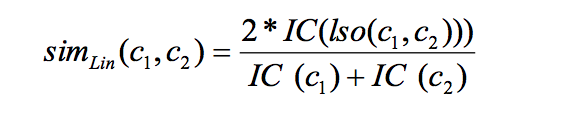
## **Semantic Similarity Model**

The semantic similarity model was to calculate the similarity score between two sentences. Each sentence includes numbers of concepts and the similarity of each concept and the number of similar concepts were used to calculate the final similarity score.

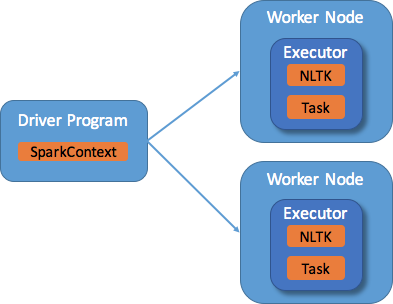


To calculate the similarity of each concept between two sentences, first of all, the concepts of sentence would be extracted by using the part of speech tagging approach. All types of nouns and adjective were extracted. Then bi-grams was conducted to further generate the concept using phrases. The goal is to catch some noun + noun or adjective + noun phrases that have the same meaning of a single words. For example, in our training sample, the word “poverty” and “poor individuals” may share some similarity. However, if they were compared separately, the result score will be zero.

After extracting the concepts, for each word in the concept, a similarity score was calculated between each two words based on the wordnet similarity methods. Lin Similarity Method was used here and it would return a score denoting how similar two word senses are, based on the Information Content (IC) of the Least Common Subsumer and that of the two input Synsets in WordNet. Below is the formula:



As the cost of computing each word’s similarity is extremely high and the purpose of this target is to rapid assessment, thus a big data technology was introduced here to speed up the process. Spark, as the most popular big data technics, works with the file system to distribute data across the cluster and process it in parallel. It would only take 0.1 second to process thousands of sentences.



## **TF-IDF Ranking**

TF–IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. In the case of the term frequency tf(t,d), the simplest choice is to use the raw count of a term in a document, i.e. the number of times that term t occurs in document d. If we denote the raw count by ft,d, then the simplest tf scheme is tf(t,d) = ft,d.The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

.\begin{displaymath} \mbox{tf-idf}_{t,d} = \mbox{tf}_{t,d} \times \mbox{idf}_t. \end{displaymath}

\begin{displaymath} \mbox{Score}(q,d)=\sum_{t\in q} \mbox{tf-idf}_{t,d}. \end{displaymath}

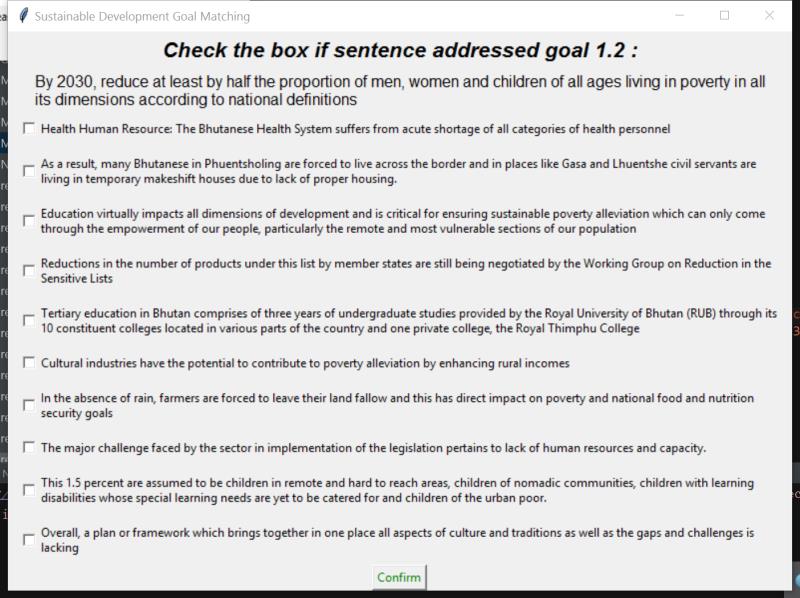
After running semantic similarity model, sentences selected were all about certain SDG subgoals but still too many, the team use TF-iDF to rank the similarity between selected sentences and certain SDG subgoal content. In the meanwhile, the team imported golden standard to this model, calculated the similarity between golden standard and SDG subgoals and took this similarity as a dynamic threshold. Any sentence with score that below the threshold will not be considered in the final results, while others would be ranked and only top 10 would be selected.

# **Result and Evaluation**

This text mining application is tested by using Namibia national plan as input. The goal 3.1 to 3.9 are tested. 8 goals out of 9 are correctly matched. The accuracy is 8/9, which is pretty satisfying in this goal. So the final results reached a high accuracy of 88.9%.

**User Interface**

A user interface is designed for users to manually selected satisfying sentences from the sentences recommended by the computer. After the user chooses the 10 most related sentences, the sentences will come back to the dataset and a template will be created. The template looks similar to what United Nations staff manually did before. The sectors which selected sentences belong to will be marked.



# **Limitation and Further Stage**

The weak features do not function well in this algorithm. The weak features can be further narrow the number of sentences. In the national plan, the countries sometimes mention about current situation, not the plan for the next step, the team needs to recognize which sentences refer to the plan. Moreover, the sectors for sentenced should be labelled automatically. For the next stage, our team aims to put weak features into our analysis, figure out a way to classify current situation and future plan, and design a method to label sectors.