Semantic Searching for Efficient Assessment of Sustainable Development in National Plans

Jonathan Galsurkar[1], Science for Social Good Fellow, Columbia University Data Science Institute

Under guidance of Aditya Vempaty[1], Kush R. Varshney[1], Lingfei Wu[1], Mikhail Sushkov[2], and Moninder Singh[1]

In Partnership with the United Nations Development Programme: Devika Iyer[3] and Serge Kapto[3]

[1] IBM Research [2] IBM Watson [3] UNDP

***Abstract*— Much of the application of technology to societal problems is currently gated by government processes that are still largely mechanical and labor-intensive. The United Nations Development Programme helps countries implement the UN's sustainable development goals, which tackle major societal issues such as poverty, hunger, and education. The first step of this implementation is the Rapid Integrated Assessment (RIA) methodology, which consists of reviewing national development plans and attempting to match specific pieces of text to one or more of the 169 targets of the 17 sustainable development goals (SDG). The RIA methodology requires policy experts to review hundreds if not thousands of pages of documents, taking weeks to accomplish. We apply Natural Language Processing and Machine Learning, specifically sentence embedding and semantic search techniques to automate the RIA, properly assigning sentences of national development text to SDG targets, ultimately helping governments be more effective and develop more sustainably.**

I. Introduction

On August 2, 2015, governments united behind an ambitious agenda that features 17 sustainable development goals (SDGs) and 169 targets that are a part of them, aiming to end poverty, combat inequality, and promote prosperity. Agreed by consensus, the draft outcome document "Transforming Our World: the 2030 Agenda for Sustainable Development" [1], was formally adopted by world leaders at the United Nations Summit. Those SDGs fall closely in line with the objectives of country leaders, trying to improve the state and wellbeing of their country in various sectors. The first step to any actionable improvement is creating National Development Plans, which outline a systematic path of growth and prioritize the actions and legislation that must be fabricated to propel the country in a course of sustainable development. If executed properly, the citizens of those countries will be closer to living in an atmosphere of peace, harmony, and political stability. Creating national plans to ensure endeavors such as these is no simple task.

After writing their National Plans, country leaders turn to the United Nations Development Programme (UNDP) for further analysis and evaluation on how well their priorities and ideas work toward the SDG targets. The first step of this evaluation is the Rapid Integrated Assessment (RIA) methodology, which consists of reviewing the national development plans and attempting to match specific pieces of text to one or more of the 169 targets of the 17 sustainable development goals. Much of the application of technology to societal problems is currently gated by government processes that are still largely mechanical and labor-intensive. The RIA methodology is no different, requiring policy experts to review, by hand, hundreds if not thousands of pages of documents. This crucial task not only takes weeks to accomplish, but currently requires the knowledge only policy experts possess.

In order to address the aforementioned challenges, we apply natural language processing and machine learning methods, specifically, we propose to make full use of recently presented word and document embedding techniques to effectively develop a semantic searching system for automating the RIA, properly assigning sentences of national development text to SDG targets. The proposed system has three phases: training our model, finding the sentences/paragraphs of new national plans that match the UNDP targets, and returning the top matches for each target. In this paper, we shall

1) Study various word/document embeddings techniques and investigate the effectiveness of these techniques.
2) We propose to augment the current symantic search system with additional knowledge from previously completed RIAs for various countries.
3) Conduct extensive experiments to demonstrate the efficiency and effectiveness of the proposed method and system.

II. Data

We have access to the national development plans of countries whose RIAs have been previously conducted by the UNDP. These documents were in pdf format and therefore, text extraction was necessary to utilize the information within them. Similarly, we have access to those previously completed RIAs which contained sentences and the targets that policy experts had matched them to. Let's

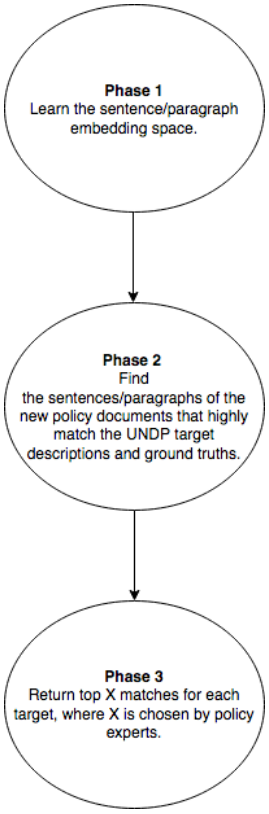
call these sentences our "ground truth." These RIAs came in several formats from xlsx to docx files. We pre-processed those files to retrieve the ground truth sentences along with the target they matched.

Target	Target Match from National Plan Text
3.1 By 2030, reduce the global maternal mortality ratio to less than 70 per 100,000 lives birth.	• Equip, upgrade and expand a network of health facilities providing quality emergency obstetric care (EmOC) to secure a fair distribution of and access to services. • Adequate training of district teams and training of doctors and nurses for comprehensive EmOC at all health facilities providing basic and comprehensive EmOC.
5.2 Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation	• Enforce legislation and increase accountability of perpetrators of domestic violence against women. Strengthen inter-agency cooperation on domestic violence. Continue to raise awareness among medical practitioners on domestic violence and provide measures for early identification and intervention. • Reforms Institutions to ensure that victims of domestic violence are given immediate shelter in a Government institution and provided with a job and a house within a reasonable time frame to lead a normal life anew.

Table 1. Example portion of a RIA

III. Methodology

To find specific sentences/paragraphs of national plans that match the targets of the SDGs, we need a model that can discern the semantics and context of a given sentence and be able to match it to a target of similar meaning and intention. Regardless of model choice, we can break our general technique up into three phases:



Classic representations of text such as a bag of words model [2] or tf-idf model [2] mainly take word frequencies into account, disregarding word order which therefore disregards the context behind that text. These types of models would perform poorly in the setting of national development plans due to the many varying ways there are to write and convey legislation or plans of actions that ultimately have the same goal or meaning. To that end, we focus on sentence embedding techniques, meaningful vector representations of a sentence/paragraph that capture and preserve its semantic and syntactic relationships. We investigate the utilization of Word2Vec [3], [4] in combination with numerical statistics popular in text mining/information retrieval such as tf-idf or nbow as well as Doc2Vec [5], [6].

All policy related documents are used as input to learning the models, including that of the country currently being tested.

Note that during phase 2, we exclude the ground truths of the country we are currently testing.

A. *Doc2Vec Model*

Doc2vec is an unsupervised model to generate vectors for sentence/paragraphs/documents. For phase 1, we train the Doc2vec model with our policy related documents. We can now embed the ground truth and target descriptions which will comprise our vector space.

B. *Word2Vec Based Model with Scaling*

Word2vec is an unsupervised model that is used to produce word embeddings. This model is a shallow, two-layer neural network that is trained to reconstruct linguistic contexts of words. For Phase 1, we have two options:

- We can train our own Word2vec model by having Word2vec take as input all the policy related documents and produce a vector space, currently set to around two thousand dimensions, with each unique word in the corpus of policy documents being assigned a corresponding vector in the space.

- We can use Google's pre-trained Word2vec model, which includes word vectors for a vocabulary of 3 million words and phrases that were trained on roughly 100 billion words from a Google News dataset. The vector length is 300 dimensions.

Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

Next, to learn our sentence/paragraph embedding space, we infer the embedding for a sentence/paragraph as follows:

$$\sum_{w_j \in s_i} word2vec(w_j) * scale\_factor(w_j)$$

where the scaling factor can be the word's tf-idf or normalized term frequency (nbow). The next step is to embed the target descriptions and ground truths which will comprise our vector space.

C. *Utilizing Existing RIA Data*

In most semantic search problems, each query corresponds to a relatively small search text. With our application of semantic searches to national plans, we utilize the ground truth information from prior RIAs, appending the ground truth sentences to their corresponding target descriptions, allowing a query to relate to multiple aspects of targets, greatly enhancing our semantic searches and improving the quality of our matches. In essence, although our models are unsupervised, because we have access to the "class" (target) of our ground truth sentences and many examples of that target, we are able to capture additional perspectives of the target that the semantics of the target description alone would not provide. We will later compare the results of utilizing the ground truth sentences versus only relying on target descriptions as a general semantic search would do. With every RIA conducted, we will have new sentences that match the corresponding target. With evaluation from policy experts, we can incorporate the results that policy experts agree are good matches and utilize these results in the next RIA conducted.

Phases 2 and 3 will stay the same regardless of technique. Once we have our model and vector space of ground truths/target descriptions, we embed each sentence of the new documents. For each embedded sentence, find the k nearest neighbors, where the distance measure is the cosine similarity [7]. Assign the sentence to each of the targets of the k nearest neighbors. For Phase 3, sort the results for each target by cosine similarity, and return the top X results.

IV. Experimental Evaluations

After reading the SDG target matches obtained with the Doc2Vec model, many matches were poor and did not reflect the targets matched to them. The rest of our analysis pertains to variations of our Word2vec based model with scaling.

A. *Numerical Analysis/Comparison with Policy Experts*

The main procedure of evaluation of our results is to see how many of the same matches we find with policy experts for RIAs that have been completed previously. Of course, as we increase the number of sentences outputted, we will eventually get a 100% match with policy experts.

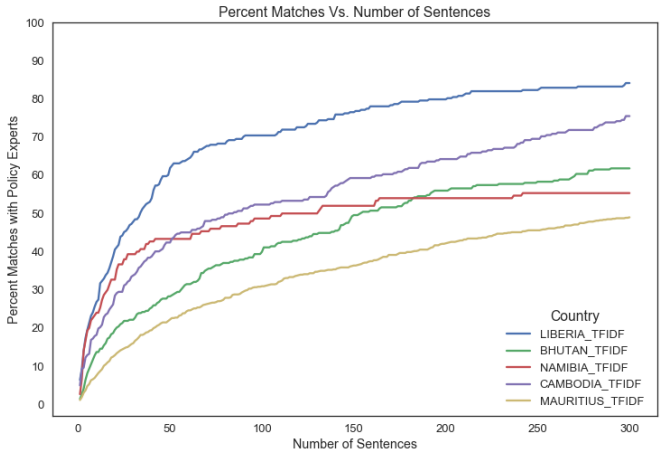


Fig 1. Percent Matches with policy experts (using tf-idf scaling) as we increase the number of sentences outputted to 300.

More realistically however, policy experts will only wish to see the top 30 or so results per target. An interesting finding, as policy experts confirmed, was that the rank ordering of percent matches after 30 sentences directly reflected the relative difficulty of conducting the RIA for those policy experts.

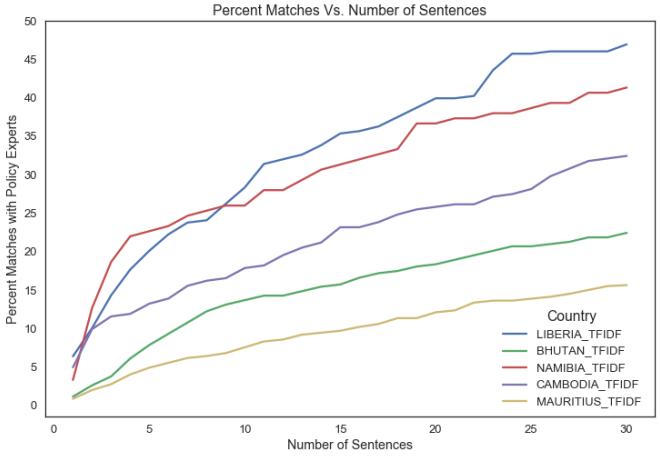


Fig 2. Percent Matches with policy experts (using tf-idf scaling) as we increase the number of sentences outputted to 30.

It is important to note that the results above are based upon the percent match across all targets. There are major differences in the difficulty of finding matches for the various targets. The targets in which a high percent match was found early on reflect the difficulty of finding matches for that particular target within the given national plans. It is interesting that the "level of difficulty" for the program and policy experts to find matches for the various targets is similar.

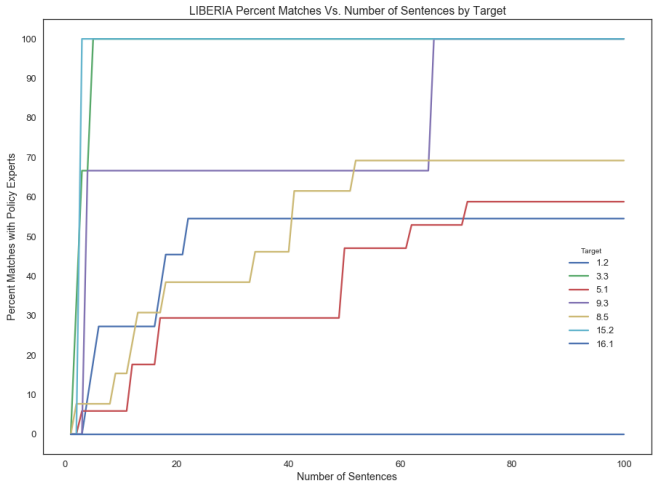


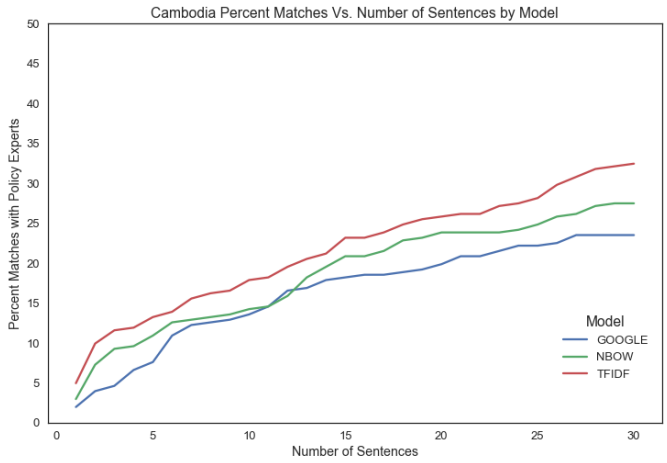
Fig 3. Variance of Percent Matches with policy experts (using tf-idf scaling) by target for Liberia.

We have also found that Google's pre-trained Word2vec model performs worse than training our own Word2vec model with tf-idf scaling in all test cases when using the target descriptions. This is not surprising for two reasons:

1. Google's pre-trained model has 300 dimensions while ours has around 2000.

2. Google's pre-trained model is learned from the words of google news data set, which adds noise due to the variety of text in those documents. Our Word2vec model was learned strictly using policy documents, capturing the context of the text with much less noise.

The difference in results between NBOW scaling and Google's Word2vec model is not significant. Liberia and Cambodia performed slightly better with the NBOW scaling, while the other three countries performed slightly better with Google's model. We plan to continue evaluating the results of the scaling options as well as try new scaling techniques.



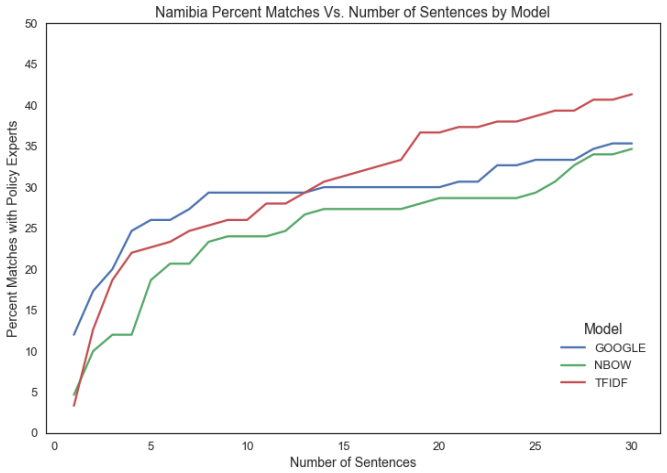


Fig 4a, 4b. Impact of model type/scaling.

It is worth mentioning once again that the uniqueness of this semantic search problem is derived from the fact that we have ground truth sentences that are known to match certain targets. If we only relied on the target descriptions, we would generally have worse results.

	Percent Matches after 30 sentences	
Country	*Using Target Descriptions and Ground Truth*	*Using only Target Descriptions*
Bhutan	22.45 %	19.24 %
Cambodia	34.11 %	38.04 %
Liberia	47.26 %	35.67 %
Mauritius	16.39 %	15.28%
Namibia	41.33 %	33.56%

Table 2. Percent Matches with policy experts (using tf-idf scaling) using the ground truth matches vs. Only using the target descriptions.

After every RIA conducted, policy experts will evaluate our matches for each target. Those sentences that policy experts deem are good matches will be appended to the target descriptions in phase 1, further improving our model. It is worth to note that our word2vec model will improve as well due to the addition of policy documents to train the model with. With this in mind, with every RIA conducted, we expect the quality of our matches to increase, something that would not be as apparent when only using the target descriptions

B. Case Study: Liberia

Liberia's RIA has been previously conducted by policy experts at the UNDP, meaning that we know what sentences/paragraphs of Liberia's national plans have been matched. Reviewing some of the sentences found with policy experts from the UNDP, it is clear that we find many high-quality matches for each of the targets when such matches exist. We have also found matches that were not present in the RIA previously conducted. While policy experts are still evaluating those matches, initial evaluation suggests that some of those sentences are relevant to the target.

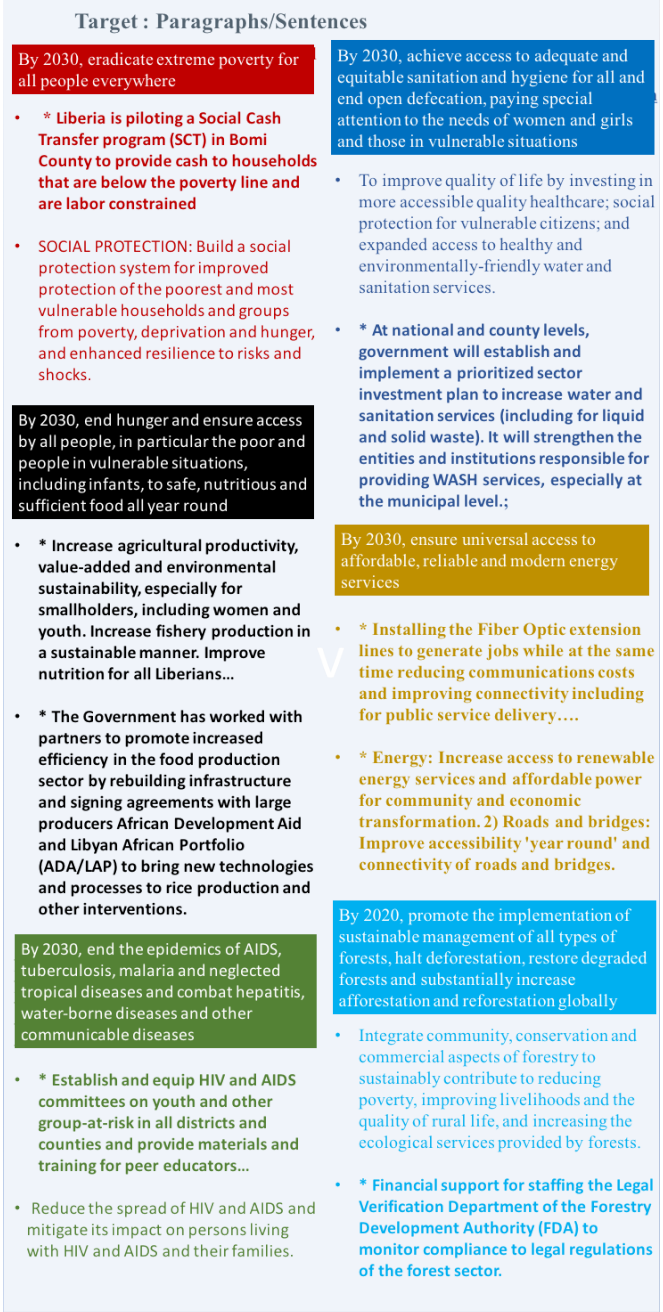


Fig 5. Results for Liberia for selected targets. Bolded Results with a * in front are matches we found that are not present in Liberia's RIA.

V. Impact and Further Research

The main impact of our methodology is the time it takes to conduct a RIA. As an example, the UNDP has not yet attempted to conduct a RIA for Papua New Guinea, estimating that it would take a few weeks to get done. There are 17 policy documents, totaling about 1500 pages that need to be read. We provided a RIA with high-quality target matches within a few hours.

Moving forward, we will:

- Explore different scaling factors

- Work with the UNDP to evaluate our results for new RIAs in larger scale studies.

- Incorporate the supervised nature of our data into our model. i.e. augment the embedding space itself by the ground truths for each target.

VI. Conclusion

With our current Word2vec based embedding scheme and semantic search technique, we automate the RIA, allowing UNDP policy experts to drastically decrease the amount of time necessary to conduct one. For each target match outputted, the page number and national plan the text originated from are provided as well, allowing policy experts to verify the matches as well as be directed to the pages with relevant text for a particular target. This in turn helps countries more quickly ensure coherence of various legislative frameworks, national plans, and ultimately, make those governments more effective. We will continue experimenting and improving our methodology to further increase the quality of matches outputted.

VII. References

[1] UN General Assembly, Transforming our world : the 2030 Agenda for Sustainable Development, 21 October 2015, A/RES/70/1, available at: http://www.refworld.org/docid/57b6e3e44.html

[2] Salton, Gerard and Buckley, Chris Term Weighting Approaches in Automatic Text Retrieval. , Cornell University , Ithaca, NY, USA (1987).

[3] Mikolov, Tomas, Chen, Kai, Corrado, Greg and Dean, Jeffrey. "Efficient Estimation of Word Representations in Vector Space." *CoRR* abs/1301.3781 (2013).

[4] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Gregory S. and Dean, Jeffrey. "Distributed Representations of Words and Phrases and their Compositionality.." Paper presented at the meeting of the NIPS, 2013.

[5] Le, Quoc V and Mikolov, Tomas. "Distributed Representations of Sentences and Documents.." Paper presented at the meeting of the ICML, 2014.

[6] Dai, Andrew M., Olah, Christopher and Le, Quoc V.. "Document Embedding with Paragraph Vectors.." *CoRR* abs/1507.07998 (2015).

[7] Subhashini R., Jawahar V., Kumar S., Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval, First International Conference on Integrated Intelligent Computing, IEEE, 2010