Assessing National Development Plans for Alignment with Sustainable Development Goals via Semantic Search

Jonathan Galsurkar,[1] Moninder Singh,[1] Lingfei Wu,[1] Aditya Vempaty,[1] Mikhail Sushkov,[2] Devika Iyer,[3] Serge Kapto,[3] Kush R. Varshney[1]

[1]IBM Research and [2]IBM Watson, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA

[3]United Nations Development Programme, 304 East 45th Street, New York, NY 10017, USA

Abstract

The United Nations Development Programme (UNDP) helps countries implement the United Nations (UN) Sustainable Development Goals (SDGs), an agenda for tackling major societal issues such as poverty, hunger, and environmental degradation by the year 2030. A key service provided by UNDP to countries that seek it is a review of national development plans and sector strategies by policy experts to assess alignment of national targets with one or more of the 169 targets of the 17 SDGs. Known as the Rapid Integrated Assessment (RIA), this process involves manual review of hundreds, if not thousands, of pages of documents and takes weeks to complete. In this work, we develop a natural language processing-based methodology to accelerate the workflow of policy experts. Specifically we use paragraph embedding techniques to find paragraphs in the documents that match the semantic concepts of each of the SDG targets. One novel technical contribution of our work is in our use of historical RIAs from other countries as a form of neighborhood-based supervision for matches in the country under study. We have successfully piloted the algorithm to perform the RIA for Papua New Guinea's national plan, with the UNDP estimating it will help reduce their completion time from an estimated 3-4 weeks to 3 days.

Introduction

In 2015, the governments of the world adopted an ambitious agenda containing 17 Sustainable Development Goals (SDGs) aiming to end poverty, combat inequality, and promote prosperity. The document agreed upon by consensus, *Transforming our World: the 2030 Agenda for Sustainable Development* (UN General Assembly 2015), contains 169 very specific measurable targets within the goals, e.g. "By 2030, end all forms of malnutrition, including achieving, by 2025, the internationally agreed targets on stunting and wasting in children under 5 years of age, and address the nutritional needs of adolescent girls, pregnant and lactating women and older persons" and "Adopt and strengthen sound policies and enforceable legislation for the promotion of gender equality and the empowerment of all women and girls at all levels".

Leaders of many of the member states of the United Nations (UN) are now aiming to align their own countries' national development plans and sectoral plans with the SDGs and their targets. (Plans typically outline a systematic path of growth and prioritize the actions and legislation that must be undertaken.) To help facilitate this objective, the UN Development Programme (UNDP) offers its policy experts as a resource to review drafts of national plans at the request of individual governments.

The first step of this evaluation by policy experts is the Rapid Integrated Assessment (RIA) methodology, which consists of manually reviewing the national development plans and assessing alignment of national targets with one or more of the 169 targets of the 17 SDGs. Although containing the word 'rapid,' the RIA methodology requires policy experts to review hundreds of pages of documents, taking weeks to accomplish, and requiring the knowledge that only policy experts possess. The specific task that the experts undertake is to read every paragraph and mark it as relevant to one or more of the 169 SDG targets (or to none).

In this work, we develop natural language processing and machine learning methods to help reduce the manual burden. Specifically, we use recently proposed 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: model training, finding sentences/paragraphs of new national plans that match the UNDP targets, and returning the top matches found for each target.

One unique feature of our problem is that we have access to previously-conducted RIAs from other countries. Because of this data, we have snippets of text known to match the different SDG targets and thus we do not have a completely unsupervised problem at hand. We take advantage of this information in the second phase of the proposed system by not only matching to SDG target descriptions, but also to extracts from previously conducted RIAs. This novel technical approach improves accuracy appreciatively and presents a general tactic for other similar problems that may be encountered having unique semantics but little data.

The system has been quantitatively evaluated on the national plans of 5 countries with previously conducted manual RIAs: Liberia, Bhutan, Namibia, Cambodia, and Mauritius. The results are promising to form the basis for a deployed system. Moreover, it has been piloted for a country

Table 1: Extract of a RIA

SDG 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
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.

whose RIA the UNDP had not conducted before: Papua New Guinea. Feedback on these results from UNDP policy experts has been very positive.

Data

Two different kinds of data was used to develop and test the system. The first kind of data consisted of the national development plans of 6 countries (Bhutan, Liberia, Cambodia, Mauritius, Namibia, and Papua New Guinea (PNG)). These plans ranged from 2 documents for Bhutan (totaling over 800 pages) to over a dozen documents for Cambodia (totaling around 1400 pages). These documents were in pdf format and, therefore, text extraction was necessary to utilize the information within them.

The second data source consisted of previously completed RIAs (for all the above countries, except Papua New Guinea for which a RIA has not yet been done by the UNDP), each of which contain plan-document sentences and the SDG targets that policy experts had matched them to. We refer to these sentences as our "ground truth". These RIAs came in several formats from xlsx to docx files and we processed them to retrieve the ground truth sentences along with the target they matched. Table 1 shows an extract of one such RIA.

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.

Classic information retrieval methods, such as normalized bag of words (nbow) or term frequency-inverse document frequency (tf-idf) (Salton and Buckley 1987), mainly take word frequencies into account while ignoring word order, thereby disregarding 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 action that ultimately have the same goal or meaning. To that end, sentence embedding techniques, meaning vector representations of a sentence/paragraph that capture and preserve its semantic and syntactic relationships, are better suited for this purpose. Word2Vec (Mikolov et al. 2013a; 2013b) and Doc2Vec (Le and Mikolov 2014; Dai, Olah, and Le 2015) are two such models that can be used for this purpose

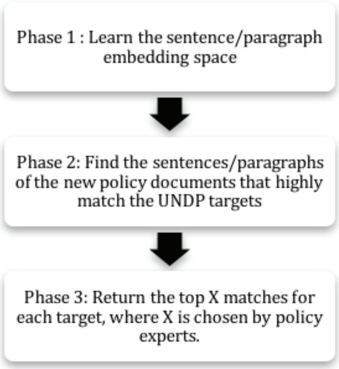
Phase 1 : Learn the sentence/paragraph embedding space
Phase 2: Find the sentences/paragraphs of the new policy documents that highly match the UNDP targets
Phase 3: Return the top X matches for each target, where X is chosen by policy experts.

Figure 1: Three main phases of our semantic searching system.

Word2Vec is an unsupervised model (two-layer neural network) that is used to produce word embeddings from a corpus of text by mapping words to vectors (typically several hundred dimensions) such that the word vectors of syntactically/semantically related words are located close to each other in the vector space. Once the model is trained, to learn the sentence/paragraph embedding space, the embedding for a sentence/paragraph S_j (comprising words $w_{ij} \in S_j$) can be inferred as

$$\sum_{w_{ij} \in S_j} Word2Vec(w_{ij}) * scale\_factor(w_{ij})$$

for some appropriate scaling factor (e.g. word's tf-idf or normalized term frequency (nbow) in the corpus).

Doc2Vec is an adaptation of Word2Vec. It is an unsupervised model that is used to directly generate vectors (embeddings) of sentences, paragraphs, or entire documents.

Both methods can be used to map sentences/paragraphs from policy documents as well as target descriptions to the corresponding sentence/paragraph embedding space, and policy document sentences "close" to various target descriptions in that vector space can be mapped to those descriptions.

As such, we propose the following method (outlined in Figure 1) for mapping policy document sentences/paragraphs to SDG target descriptions.

First, in phase 1, a Word2Vec model is trained using as input all available national plan documents as well as SDG target descriptions to produce a vector space, set to around two thousand dimensions. Note that since a RIA maps plan document sentences to target descriptions, all prior RIA data is automatically included in the input. While the word embeddings can then be used to learn the sentence/paragraph embedding space by simply averaging the word vectors, or by using a scaling factor, such as each word's tf-idf or nbow in the corpus (as described above), it does not utilize the information available in the mapped data (sentences to targets) in the prior RIAs. Since Word2Vec is an unsupervised model, it does not take the relationship between the mapped sentences and targets into account. If the text corpus was of sufficient size, it could presumably learn these relationships. Given that the corpus consists of only a handful of RIAs, it is imperative that this be coded explicitly. To that end, we create a corpus of documents where each document is the concatenation of a SDG target description with every sentence/paragraphs mapped to that target in any prior RIA. Given this corpus of 169 documents (corresponding to the 169 SDG targets), we then compute a scaling factor for each word present in a target description/RIA as the tf-idf value of that word in this corpus. Furthermore, for each word that exists in a national policy document but not in a RIA, we look for words that are close to it in the Word2Vec vector space, and take the tf-idf score of the most similar word that is available. If no such word is available, the scaling factor is set to 0.

The next step is to embed the target descriptions which will comprise our vector space. Once again 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 searching 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, which are examples of the corresponding target, we are able to capture additional perspectives of the target that the semantics of the target description alone would not provide. We later compare the results of utilizing the ground truth sentences versus only relying on the target description as a general semantic search would do.

Second, in phase 2, once we have our vector space of ground truth sentences combined to their corresponding target descriptions, we embed each sentence of the new documents. For each embedded sentence, we find the k nearest neighbors, where the distance measure is the cosine similarity (Subhashini, Jawahar, and Kumar 2010). Next, the sentence is assigned to each of the targets of the k nearest neighbors.

Third, in phase 3, we sort the results for each target by cosine similarity and return the top X results.

Experimental Evaluation

In order to evaluate the performance of the system, we followed a leave-one-out strategy. For each of the five countries for which we have a completed RIA, we trained the model using the RIAs of the other four countries along with the national plan documents of all five. The learned model was then applied to the national plan documents of the country in question to identify sentences in the document that were deemed to be relevant to each SDG target.

The main procedure of evaluation of our results is to see how many of the sentences mapped by the policy experts in the RIA for each SDG target are also picked by the system for that target. Of course, as we increase the number of sentences outputted by the system per target, we will eventually retrieve 100% of the sentences matched by policy experts to that target in the RIA.

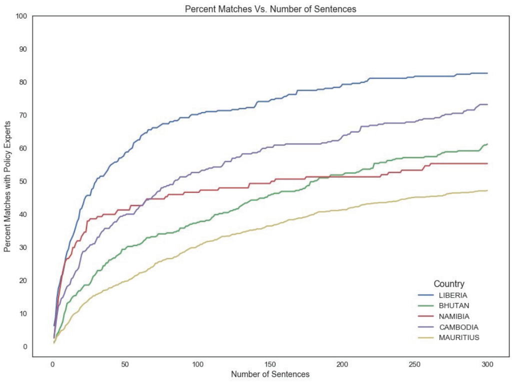

Figure 2: Average percent matches with policy experts (using tf-idf scaling) across all targets as the number of sentences outputted per target increases to 300.

The evaluation metric used was the average percentage (across all SDG targets) of RIA sentences that were also recovered by the system. Figure 2 shows the average percentage of matches with the RIA as a function of the number of sentences generated per target by the system. As can be seen from the figure, the performance increases fairly rapidly with a relatively small number of sentences generated per target (less than 50), and then increases much more gradually as the number of sentences generated per target is increased to 300.

After consulting with the UNDP policy experts, the number of sentences generated per target was set to 30. This number was deemed to be reasonable for policy experts to evaluate while still significantly faster and easier than finding matches themselves. Figure 3 shows the corresponding results. The best performance was attained for Liberia, followed by Bhutan, Namibia, and Cambodia. The worst performance was obtained for Mauritius. An interesting finding, as policy experts confirmed, was that the rank ordering of the countries by average percent matches after 30 sentences directly reflected the relative difficulty of conducting the RIA for those countries by the policy experts. Thus, the policy experts too found Mauritius to be the hardest country for carrying out the RIA analysis, while Liberia was considered to be the most straightforward of the lot.

It is important to note that the results above are based upon the average percent matches across all targets after the corresponding number of sentences outputted per target. There are, however, major differences in the difficulty of finding matches for various targets. The targets in which a high percent match was found early on (i.e. with fewer sentences) generally reflects the level of difficulty for policy

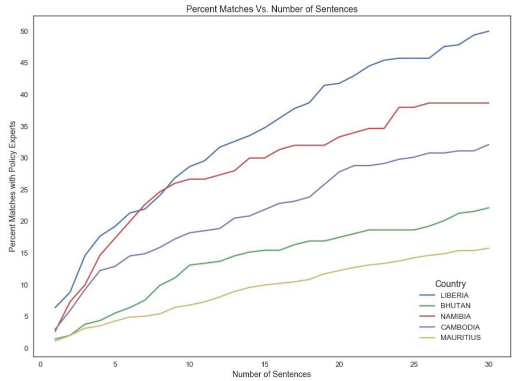

Figure 3: Average percent matches with policy experts (using tf-idf scaling) across all targets as the number of sentences outputted per target increases to 30.

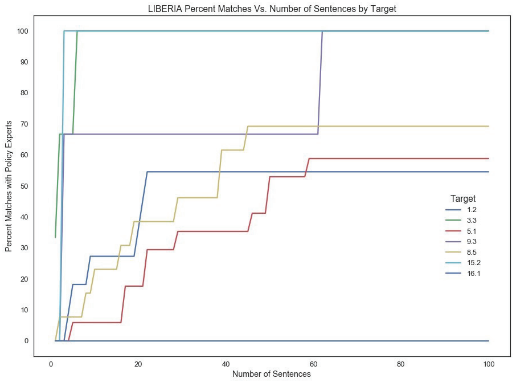

Figure 4: Liberia's variations of percent matches with policy experts (using tf-idf scaling) for individual targets.

experts for finding matches for that particular target within the given national plans. It is interesting that the "level of difficulty" for the system and policy experts to find matches for the various targets is similar. Figure 4, for example, shows the percentage of matches obtained for various targets for Liberia, as a function of the number of sentences outputted. Targets 15.2, 3.3 and 9.3 were considered to be easier to match by the policy experts, while 5.1 and 16.1 were considered to be relatively difficult.

We also compared the performance of our system to Google's pre-trained Word2Vec model, as well as our word2Vec model with nbow scaling. While we also did some preliminary comparisons with the Doc2Vec model, manual inspection of SDG target matches obtained with the Doc2Vec model showed that matches were relatively poor and did not reflect the targets matched to them; hence, we did not pursue it further. Figures 5 and 6 show the results for Cambodia and Namibia. While in the case on Cambodia,

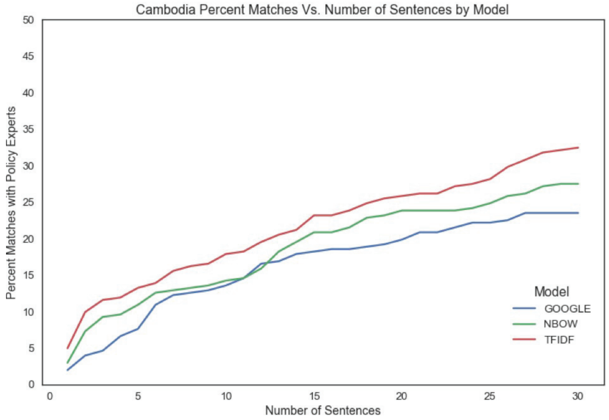

Figure 5: With some countries such as Cambodia, training our own Word2evc model with nbow scaling outperforms Google's pre-trained Word2Vec model.

Word2Vec with nbow scaling outperformed Google's pre-trained Word2Vec model, the opposite held true in the case of Namibia. In both cases (as well as the remaining 3 countries), our system performed better than both methods.

This outperformance with respect to the Google model is not surprising for two reasons:

1. Google's pre-trained model has 300 dimensions while ours has around 2000. With training our own model, we have the flexibility to choose the dimensionality and because the training set is relatively small, we can still efficiently use high dimensional vectors. To that extent, because our training set is relatively small, training our own model with vectors of 300 dimensions yields poor results.
2. Google's pre-trained model is trained on the words of a Google news data set, which adds noise due to the variety of text in those documents. Our Word2Vec model was trained strictly using policy documents, capturing the context of the text with much less noise.

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 (as mapped by policy specialists in the RIAs). Mapping the sentences by relying only the the target descriptions almost always resulted in worse performance. Using the historical RIA data to augment the target descriptions improved the performance substantially for 3 of the countries, improved slightly for one, and resulted in worse performance for one (Table 2).

Once our system conducts a RIA, policy experts have the ability to evaluate the sentences found for each target. Those sentences that policy experts deem to be good matches will be incorporated into subsequent RIAs, further improving the system performance. Note that our Word2Vec model will likely improve as well due to the addition of policy documents to train the model with. With every RIA conducted, we expect the quality of our matches to increase, something

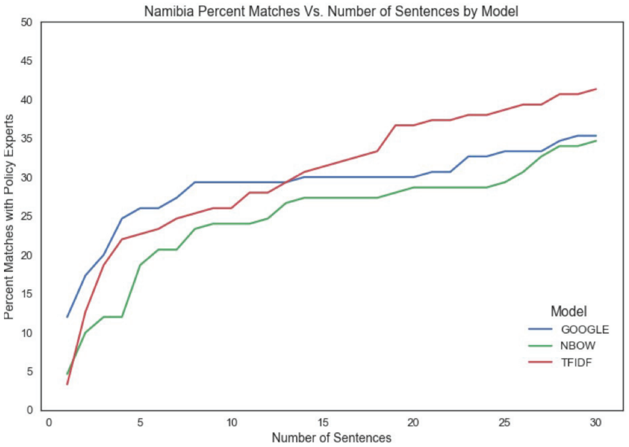

Figure 6: With other countries such as Namibia, Google's pre-trained Word2Vec model outperforms training our own Word2Vec model with nbow scaling Training our own Word2Vec model with tf-idf scaling does best in all cases when using prior RIA data.

that would not be as apparent when only using the target descriptions.

Finally, while the system performs well in terms of identifying sentences for various targets that match those found by policy experts in the RIA, it also often finds matches that were not present in the RIA previously conducted. Table 3 shows some such sentences that were picked by the system for Liberia. While policy experts are still evaluating these matches, initial evaluation suggests that at least some of these sentences are relevant to the target, showing that the system can pick up sentences that policy experts failed to match.

Table 2: Average percent matches across all targets with policy experts (using tf-dif scaling) using the ground truth sentences with target description vs only using the target description text.

	Average Percent Matches after 30 sentences outputted for each target	
Country	**Appending ground truth to target description text (%)**	**Only using target description text (%)**
Bhutan	23.03	19.24
Cambodia	33.11	38.04
Liberia	50.91	35.67
Mauritius	15.91	15.28
Namibia	39.33	33.56

Table 3: Sample results for Liberia for selected targets. Results shown were not present in Liberia's RIA conducted by policy experts.

SDG Target	**Match found that is not present in Liberia's RIA**
By 2030, eradicate extreme poverty for all people everywhere.	Liberia is piloting a social cash transfer program (SCT) in Bomi County to provide cash to households that are below the poverty line and are labor constrained.
By 2030, end hunger and ensure access by all people, in particular the poor and people in vulnerable situations, including infants, to safe, nutritious, and sufficient food all year round.	Increase agriculture productivity, value-added and environmentally sustainability, especially for smallholders, including women and youth. Increase fishery production in a sustainable manner. Improve nutrition for all Liberians.
By 2030, end the epidemic of AIDS, tuberculosis, ma-, laria, and neglected tropical diseases and combat hepatitis, water borne diseases and other communicable diseases.	Establish and equip HIV and AIDS committees on youth and other groups-at-risk in all districts and counties and provide materials for training for peer educators.
By 2030, achieve access to adequate and equitable sanitation hygiene for all and end open defecation, paying special attention to the needs of women and girls and those in vulnerable situations.	At national and country levels, government will establish and implement a prioritized sector investment plan to increase water and sanitation services (including for liquid and solid waste). It will strengthen the entitites and institutions responsible for providing WASH services, especially at the municipal level.

Pilot Study

As discussed previously, the main purpose of our system is to reduce the time taken by the UNDP to conduct a RIA for a new country. As such, we piloted the algorithm to conduct a RIA for Papua New Guinea for which the UNDP has yet to conduct a RIA. There are 17 policy documents, with almost 1500 pages, that need to be evaluated for alignment against the 169 SDG targets.

The system was used to identify up-to 30 relevant sentences for each one of the 169 targets. These sentences (along with the corresponding document/page number) were then provided to the UNDP. The time taken to generate the data was under an hour. UNDP policy experts then evaluated the results to see how well each one of the flagged sentence aligned with the corresponding SDG target. The evaluation was conducted by a UNDP policy expert and an intern, taking them between 1-2 days each (9 and 8 SDGs, respectively). The overall assessment of the results was very positive, with the experts stating that the system found many high-quality, relevant matches for the targets. With the help of this system, the UNDP estimates that the time taken to conduct a RIA assessment will drop down to around 3 days from the typical 3-4 weeks it takes them manually.

Going forward, in the coming months, the UNDP plans on demonstrating the system to a wider audience within the organization, followed by a deployment of the system as an integrated part of the RIA process.

Conclusions and Future Work

The UNDP helps nations assess the alignment of their national plans with the targets of the United Nation's Sustainable Development Goals by carrying out a Rapid Integrated Assessment of the plan documents. Currently done manually by UNDP policy experts, this process takes several weeks as it involves going over hundreds to thousands of pages of documents. In this paper, we described a natural language processing-based system to accelerate the workflow of policy experts. The system uses paragraph embedding techniques to find paragraphs in the documents that match the semantic concepts of each of the SDG targets, and incorporates prior RIA data (from other countries for which a RIA has been done) as a form of neighborhood-based supervision for matches in the country under study. Automating this process allows UNDP policy experts to drastically decrease the amount of time necessary to conduct a RIA. For each target match outputted, the page number and national plan document 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.

We have successfully piloted the algorithm to perform the RIA for Papua New Guinea's national plan, with the UNDP estimating it will help reduce their completion time from an estimated 3-4 weeks to 3 days. This reduction in time should, in turn, helps countries more quickly identify policy gaps and make changes to ensure coherence of their national development planning frameworks with the SDGs.

Based on UNDP feedback, we are working on several items to further improve the quality of the results generated. These include

- Improving filtering of generated results to remove duplicates, redundant matches, as well as better discrimination between similar sentences that are candidates for a given SDG target.
- Working with policy experts to remove ground truth sentences that may be poor matches to their corresponding target.
- Improving scaling factors to improve the quality of matches.
- Working with the UNDP to evaluate our results for new RIAs in larger scale studies. With every RIA conducted by our system, we expect the results to improve.

The proposed method can also be applied for semantic searching in other domains, specifically where the domain is specific with unique semantics, the corpus is limited, and some form of ground truth data is available.

Acknowledgements

This work was conducted under the auspices of the IBM Science for Social Good initiative.

References

Dai, A. M.; Olah, C.; and Le, Q. V. 2015. Document embedding with paragraph vectors. In *Proc. NIPS Deep Learning Workshop*.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *Proc. International Conference on Machine Learning*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *Proc. International Conference on Learning Representations*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Adv. Neural Information Processing Systems*.

Salton, G., and Buckley, C. 1987. Term weighting approaches in automatic text retrieval. Technical Report 87-881, Department of Computer Science, Cornell University, Ithaca, NY.

Subhashini, R.; Jawahar, V.; and Kumar, S. 2010. Evaluating the performance of similarity measures used in document clustering and information retrieval. In *Proc. IEEE International Conference on Integrated Intelligent Computing*.

UN General Assembly. 2015. Transforming our world: The 2030 agenda for sustainable development. https://sustainabledevelopment.un.org/content/documents/ 21252030 Agenda for Sustainable Development web.pdf.