Department of Economic & Social Affairs

DESA Working Paper No. 159
ST/ESA/2019/DWP/159

May 2019

Art is long, life is short: An SDG Classification System for DESA Publications

*Author: Marcelo T. LaFleur**

ABSTRACT

Between the many resolutions, speeches, reports and other documents that are produced each year, the United Nations is awash in text. It is an ongoing challenge to create a coherent and useful picture of this corpus. In particular, there is an interest in measuring how the work of the United Nations system aligns with the Sustainable Development Goals (SDGs). There is a need for a scalable, objective, and consistent way to measure how similar any given publication is to each of the 17 SDGs. This paper explains a proof-of-concept process for building such a system using machine learning algorithms. By creating a model of the 17 SDGs it is possible to measure how similar the contents of individual publications are to each of the goals — their SDG Score. This paper also shows how this system can be used in practice by computing the SDG Scores for a limited selection of DESA publications and providing some analytics.

JEL Classification: O0 General Economic Development; O20 General Development Policy and Planning; C88 Other Computer Software

Sustainable Development Goals: 17

Keywords: SDG; publications; classification; topic models; machine learning, LDA

* The views expressed herein are my own and do not necessarily reflect the views of the United Nations.

CONTENTS

Typesetter: *Nancy Settecasi*

UNITED NATIONS
Department of Economic and Social Affairs
UN Secretariat, 405 East 42nd Street
New York, N.Y. 10017, USA
e-mail: undesa@un.org
https://www.un.org/development/desa/

I Introduction

The United Nations is a source of big data in the form of text. Between the many resolutions, speeches, meetings, conferences, studies, reports and internal regulations that exist and that are produced each year, the UN is awash in text. Even in a single department of the UN Secretariat, the amount of publications is significant. In the Department of Economic and Social Affairs (DESA), publications are central to its overall mission to support international cooperation in the pursuit of sustainable development for all. They inform development policies, global standards and norms on a wide range of development issues that affect peoples' lives and livelihoods: social policy, poverty eradication, employment, social inclusion, inequalities, demographics, indigenous rights, macroeconomic policy, development finance and cooperation, public sector innovation, forest policy, climate change and sustainable development.

However, very few people are in a position to see much more than a small sliver of specialized text. Even fewer can parse the various streams into a coherent and useful picture. What is needed is a quick and objective way to analyze large quantities of United Nations publications according to a desired criteria, namely the Sustainable Development Goals (SDGs).

This work provides a solution by introducing a proof-of-concept classification system that measures the alignment of publications with each of the SDGs. It uses a machine-learning approach to compute how much each of the 17 SDGs is represented in individual publications. This is the first time United Nations publications have been analyzed in this way.

Using machine learning algorithms to analyze digital texts has many advantages. Algorithms can be used at scale with objectivity and can help identify patterns across publications and over time. This approach can also serve as a tool to explore and discover new texts, and to inform the direction of future research. More importantly, this method hopefully inspires other efforts to use modern data analytics to better understand the body of work of the United Nations.

This paper is organized as follows. Following this introduction, the paper discusses how machine learning algorithms called topic models can be used to analyze text. The third section explains the process of building the SDG classification system and computing the SDG Scores for each publication. A fourth section presents the results and the insights from using this methodology on DESA publications. The last section concludes with suggested areas for future work.

II How Machine Learning can help us better understand UN publications

The problem of classifying texts is one of scale and objectivity. If you have a small number of books and wish to understand something of what they contain, there is no better way to do so than to sit down and read with interest. Human beings are capable of readily inferring the latent structure in the texts. It is easy to imagine someone reading a few books and identifying a handful of themes that best describe them. Readers of Charles Dickens may identify social class and poverty as central themes. For Mark Twain, the themes may be race, religion, and deception. For Franz Kafka, a reader may identify themes of identity, isolation and social class. Now imagine trying the same but with hundreds of books. How would a reader identify the three, fifteen, or fifty themes that best describe the collection?

There have been previous efforts to classify DESA and UN publications and facilitate document discovery and analytics. DESA's Working Papers have recently been manually classified according to individual SDGs. There have also been a number of recent in-depth analyses of UN texts. Le Blanc, Freire, and Vierros (2017)[1], for example, use a large collection of UN publications and academic sources to manually determine the connections among the ten targets of SDG 14. Vladimirova and Le Blanc (2015)[2] used 40 global reports to carefully examine the links between education and other SDGs in flagship publications of the United Nations system. Le Blanc (2015)[3] analyzed the targets in each of the 17 SDGs that refer to multiple goals and show the connections between some thematic areas. In each of these novel papers, the authors demonstrated the power of expert analysis and careful reading of individual texts to derive important insights.

However, there are limits to how well this methodology can scale and how it can be replicated with other texts. For any significant number of texts, the time and focus needed to understand them all becomes prohibitive. The problem gets worse as the number of documents continues to grow and as one discovers new connections between topics. For example, a publication that discusses inequality touches upon unemployment, gender, social protection, vulnerability, public policy, and many other relevant topics. Moreover, major publications like DESA's World Economic and Social Survey cover a broad range of topics related to development and simultaneously address multiple SDGs. As the Latin and Greek aphorism tells us, art is long, life is short.

Machine learning methods can make the problem tractable, combining the kind of close reading done by humans with a broader bird's-eye approach and revealing hidden patterns or trends in large collections of text. Scientific means and tools developed by academics are available that allow us to analyze large quantities of text, conducting hypothesis-testing, computational modeling, and quantitative analysis.

One technique in particular—topic modeling—makes it possible to classify texts according to some desired criterion. Topic models work in much the same way that humans identify topics in what they read. The algorithms extrapolate backward from a collection of documents to infer the discourses (themes or "topics") that could have generated them. These topics are then used to classify individual texts according to how well they are connected.

II.I A brief explanation of how topic models work

Humans are very good at understanding the content of what they read. It is no great difficulty for a person to read a book and, in a few sentences explain what themes or topics it discusses. Careful reading can identify multiple topics, and scholars can identify how some topics can be found in the works of multiple authors. Topic models work in the same way.

1 Le Blanc, David, Clovis Freire, and Marjo Vierros. 2017. "Mapping the Linkages between Oceans and Other Sustainable Development Goals: A Preliminary Exploration." *DESA Working Paper 149 (February).* https://www.un.org/development/desa/publications/working-paper/wp149.

2 Vladimirova, Katia, and David Le Blanc. 2015. "How Well Are the Links between Education and Other Sustainable development Goals Covered in UN Flagship Reports? A Contribution to the Study of the Science-Policy Interface on Education in the UN system." *DESA Working Paper 146 (October).* https://www.un.org/development/desa/publications/working-paper/education-and-sdgs-in-un-flagship-reports.

3 Le Blanc, David. 2015. "Towards Integration at Last? The Sustainable Development goals as a Network of Targets." *DESA Working Paper 141 (March).* https://www.un.org/development/desa/publications/working-paper/towards-integration-at-last.

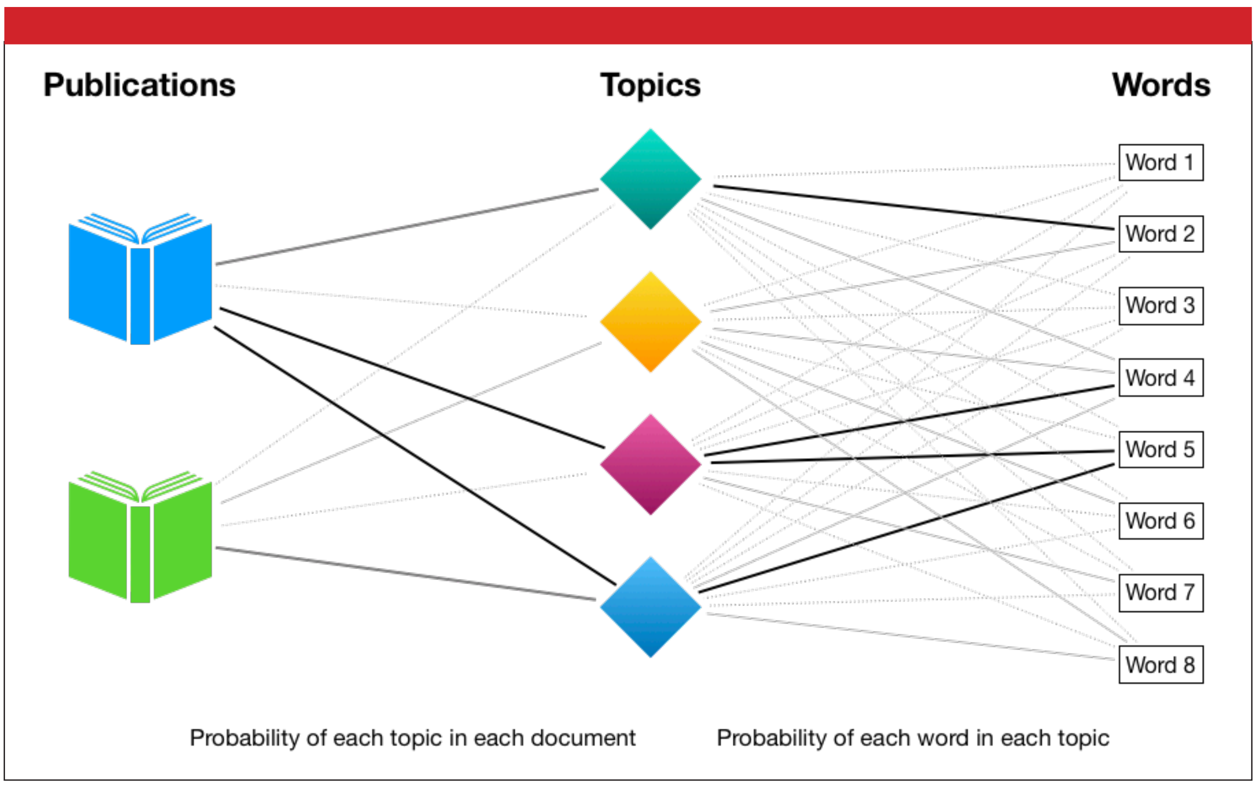
Topic modeling algorithms use statistical methods to partition a data set into subgroups. When applied to text, these algorithms can create semantically meaningful groupings from a collection of documents.[4] Put another way, topic model algorithms analyze the content and structure of a collection of texts, extrapolating backward to infer the discourses (themes or topics) that could have generated them.

The algorithm commonly used for topic modeling is called Latent Dirichlet Allocation, or LDA. What makes this algorithm useful for textual analysis, particularly the kind done in this paper, is that it results in a statistical model that can be applied to out-of-sample data. In other words, a model can be trained on pre-determined data and then used to classify a different data set. This means that using LDA to categorize a collection of texts according to the SDGs creates a model that can be used to then categorize other documents as needed.

LDA topic models start from the premise that texts are not only comprised of a set of words but are created from a set of topics. It is an author's creativity and inspiration that informs how each of the topics is used in the final text. LDA assumes that the collection of documents can be represented by a given number of topics, each of which is associated with a variety of words. Each individual document is, therefore, the result of the probabilistic sampling over the topics that describe the corpus and over the words that comprise each topic (see Figure 1).

The LDA algorithm, therefore, represents documents as combinations of all the topics in the corpus. This makes sense if one considers that texts are rarely about a single subject. A report about stagnant wages and

Figure 1
Graphical representation of a topic model (LDA)



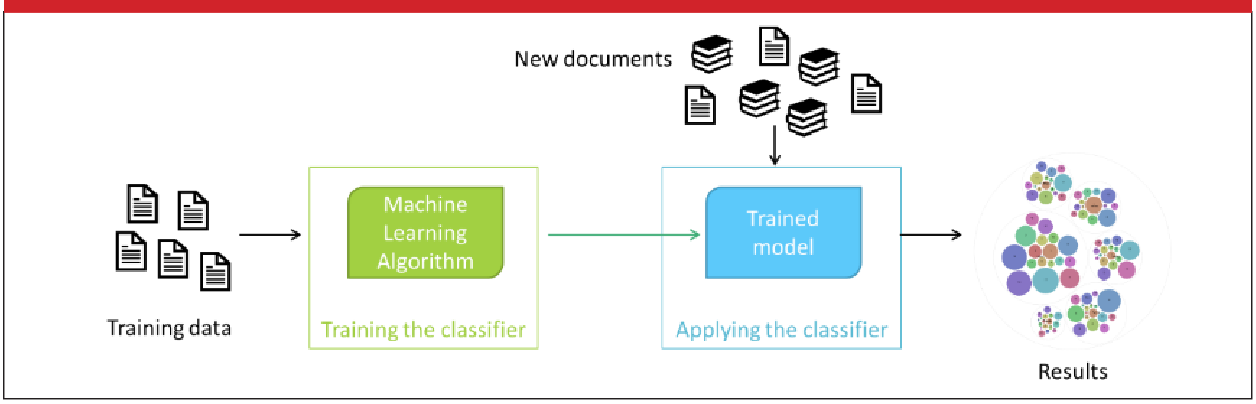
4 For an overall introduction to topic modeling, see Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

poverty could reasonably be assigned topics like "development", "jobs", "government", "the economy", "globalization" and others even though it likely mostly discusses jobs and incomes.

Using pre-determined topics to analyze the SDGs

Topic models can be applied to a collection of texts without any prior knowledge of the content of the text. In this unsupervised mode, the algorithm will dutifully identify the requested number of topics that statistically describes the text corpus (see appendix 1 for an example of a fully unsupervised 15-topic model of DESA's publications). While an unsupervised topic model gives us some interesting insights, it leaves something to be desired. The goal of this work is to understand the connections between each publication, and the connection between the entire corpus and specific topics: the SDGs. This requires a model of the SDGs that can be compared against each publication. The topic model must be pre-determined using a pre-selected collection of texts that represent the 17 SDGs. These pre-determined themes are then used to calculate the connection between each publication and each SDG (Figure 2).[5]

Figure 2
Classification workflow



The following section will describe in detail each of the steps in the classification workflow.

III Building an SDG classifier for DESA publications

This section explains how an SDG classifier can be constructed using a carefully selected training data set, and how the classifier can be applied to DESA publications. A less technical but more conceptual explanation is given in appendix 2.

A good classifier must be able to differentiate between each of the 17 SDGs. The key difficulty lies in not having labelled data in sufficient quantity to train a classifier.[6] Without labelled data, the strategy of building

5 It is an advantage here that the texts to be analyzed are expected to have certain themes. After all, they are all produced under the same Department, DESA, and support a common mission under the United Nations. It would make little sense to classify texts from a different discipline according to a pre-determined classifier that discriminates among the SDGs.

6 A supervised classifier using labelled data would be the preferred approach, but this requires a significant amount of labelled data. Tests confirmed that training a classifier using the available labelled data produced poor results when the classifier was applied to unlabelled documents.

the classifier relies on a key insight of how the algorithm builds the various topics. The LDA algorithm creates a specified number of word groupings—topics—that best represent the entire corpus (Figure 1). Each topic is comprised of a list of words with individual weights. Words with high weights are more likely to be selected in a random draw. The topics themselves are also given a weight. A topic with a large weight will contain the words that are common and frequent in most documents, while topics with small weights have more unusual and distinctive words. By having sufficiently unique texts for each of the 17 SDGs, it is possible to estimate a topic model with 17 narrow topics.

The process has two steps. The first step is to collect 17 unique and balanced texts that each represent one of the SDGs. The LDA algorithm is used to estimate an 18 topic model. The extra topic acts as a filter and should capture the common words among all the 17 representative texts. The end result should be 1 "general" topic and 17 "specific" and unique topics that are the representation of each of the SDGs. The general topic acts like a filter of common words and is excluded from the final classifier, which is then used to infer SDG scores. The process of inference answers the question: how much of each of the 17 topics is needed to maximize the similarity with the vocabulary of the document? A document about poverty is expected to be written using a vocabulary similar to the one in the poverty topic, but it may also include vocabulary from the others. The final weights should reflect this proportion.

The steps needed to build the classifier and infer the SDG Scores are described in detail below.

III.1 Building the training and target data sets

The classifier is trained on a corpus of 17 representative texts, one for each SDG. The representative texts are chosen to maximize their uniqueness, but have common structures and lengths to ensure balance. Two sources were selected for each representative text:[7]

1. The text for the UN webpage that describes each SDG. For example, the representative text for SDG 1 includes the text for the webpage found at www.un.org/sustainabledevelopment/poverty/.
2. The relevant section of the Secretary-General's annual report "Progress towards the Sustainable Development Goals" for the years 2016, 2017 and 2018 (for example, http://undocs.org/E/2018/64).

The target data is a selection of 267 individual DESA publications of various types, including flagships, working papers, policy notes, and other reports (see Table 1). The coverage was determined by the availability of digitized versions of each publication in time for the completion of this analysis. Future updates will include new publication types and additional years. Future updates to this work will aim to have the full collection of DESA publications classified in this form. With a full dataset it will be possible to answer interesting questions such as how the launch of the new 2030 Agenda in 2015 has affected the direction of publications.

7 The HLPF Review of SDG Implementation and the description of various targets in the E-Handbook on SDG Indicators was also considered to be included in the 17 representative texts. However, at the time of this writing, HLPF texts were not available for every SDG and the E-Handbook did not include every target. Tests show that using these sources resulted in an unbalanced corpus and a biased classifier.

Table 1
Number of each publication type classified

Publication	Count	First	Last
GSDR	2	2015	2016
WPSR	7	2001	2018
RWSS	7	2005	2018
Other report	7	2009	2014
NDS Policy Note	7	2007	2007
WYR	8	2003	2018
CDP Policy Note	11	2005	2017
WESS	23	1995	2018
DESA Policy Note	43	2007	2014
DESA Working Paper	152	2005	2019
TOTAL	267	1995	2019

Each publication is classified as a single item without any adjustment based on length. As a result, smaller publications like DESA Working Papers and Policy Notes are disproportionately represented in the corpus. Including more flagships and larger DESA reports in the analysis will mitigate this effect in future updates. It should also be noted that some publications are not officially mandated. An analysis of the results should be cognizant of this distinction. Notably, the entire collection of the WESS (starting in 1947) has recently been digitized while only the last 23 are reflected in this work.[8]

In order to apply the classifier to the content of each publication, the texts were stripped of all figures, footnotes, preambles, tables, bibliography, and uninformative text. The focus was kept, to the extent possible, on the main text of the publication to ensure a more direct focus on the relevant content. The algorithm also removes common, uninformative words like "*the*", "*and*", or "*a*". Additionally, the names of countries are filtered for this exercise.

The LDA algorithm uses a "bag of words" approach to analyze text. That is, the algorithm calculates the frequency of individual words across documents. Words are determined based on space-separation, and each word is examined independent of where it occurs in a document and independent of nearby words. For this reason it is not important to ensure a clear formatting of the text. However, removing uninformative words does help the algorithm identify the latent topics that comprise the corpus. The time spent identifying and removing stop words must be weighed against the increase in accuracy of the estimated model.

III.2 Training and validating the classifier

To train the classifier, 17 representative texts (training data) are used to estimate an 18-topic model using the open source tool Mallet.[9] As discussed above, 18 topics are needed instead of 17 to consider the commonality that exists across all 17 representative texts. This 18th topic thus represents a filter of terms that are common to all texts.

8 A future update to this work will expand the coverage of the WESS as far back as possible, and will also add the WESP to the corpus. The WESP and the WESS have been published as separate reports since 2005. Prior to that, a single publication combined the macroeconomic monitoring of the WESP and the thematic survey of development topics of the WESS into a single publication.

9 http://mallet.cs.umass.edu/

Evaluating the accuracy of the topic model is an important step before interpreting the results. In a supervised classification, the true labels of a subset of the data are known a priori, making it possible to check the results. In topic models generally, and in this exercise specifically, there is no prior "true" classification. Instead, the LDA algorithm statistically identifies logical subgroupings in the data without any prior knowledge of the best grouping.[10] For this reason, the approach taken here is semi-supervised in that the classifier is given a set of training data that is designed to maximize the ability to discriminate each of the 17 SDG topics.

To validate this approach, it is important to verify that the classifier can discriminate among the 17 desired topics, that it is comprised of words that correctly describe each topic, and that make sense when applied to new texts.

The first step in validating the classifier is to ensure that each topic represents one of the representative texts. The results are shown below (Table 2). The top row lists each of the 17 topics (the 18th topic is not shown). The first column shows each of the representative texts. As the table shows, each of the topics is strongly associated with only one of the SDG texts, as expected.

Table 2
Correspondence between estimated topics and SDG-representative texts

	Topics estimated using the LDA algorithm ("filter" topic is not shown)																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
sdg-1.txt	60%	0%	0%	0%	0%	1%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-2.txt	0%	66%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%
sdg-3.txt	0%	0%	74%	0%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-4.txt	0%	0%	0%	71%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-5.txt	0%	0%	0%	2%	78%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-6.txt	0%	0%	0%	0%	0%	71%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-7.txt	0%	0%	0%	0%	0%	0%	69%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-8.txt	1%	0%	0%	0%	0%	0%	0%	70%	0%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-9.txt	0%	0%	0%	0%	0%	0%	0%	0%	70%	0%	0%	0%	0%	0%	0%	0%	0%
sdg-10.txt	0%	0%	0%	0%	0%	0%	0%	0%	0%	67%	0%	0%	0%	0%	0%	0%	0%
sdg-11.txt	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	68%	0%	0%	0%	0%	0%	0%
sdg-12.txt	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	78%	0%	0%	0%	0%	0%
sdg-13.txt	0%	0%	0%	0%	0%	0%	3%	0%	0%	0%	1%	0%	80%	1%	0%	0%	0%
sdg-14.txt	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	72%	5%	0%	0%
sdg-15.txt	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	79%	0%	0%
sdg-16.txt	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	74%	0%
sdg-17.txt	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%	0%	0%	0%	0%	69%

A second step to validate the classifier is to examine the most frequent words in each group formed by the LDA algorithm. In the table below, the words that represent each topic are compared and judged based on how well they represent each of the 17 SDGs (Table 3). The table lists the top 20 terms that comprise each of the 18 topics. A visual inspection of each of the themes confirms that each topic can be reasonably associated with a single SDG and can be used to differentiate between them.

10 For a discussion on evaluating topic models, see Blei, David M. 2012. "Probabilistic Topic Models." Communications of the ACM 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Table 3
Evaluation of the SDG classifier using the 20 most important words of each topic

Label	Estimated topic (20 most important words)
Filter	sustainable development people global developing access goal world sustainable_development including regions economic developed africa asia national increase progress increased globally
GOAL-1	poverty social protection poor social_protection extreme disasters extreme_poverty poverty_line day end_poverty line losses living disaster cash protection_systems person poor_vulnerable disaster_risk
GOAL-2	food hunger agricultural agriculture children malnutrition production genetic prices export markets hungry food_production subsidies food_security nutrition undernourished breeds aid insecurity
GOAL-3	health deaths diseases people mortality births live_births maternal children hiv age live care rate years_age reproductive deaths_live worldwide risk women
GOAL-4	education primary children school primary_education quality secondary schools learning quality_education skills secondary_education reading primary_school proficiency mathematics teachers minimum saharan basic
GOAL-5	women girls gender women_girls equality gender_equality violence sexual age marriage female married work rights genital_mutilation mutilation female_genital genital partner globally
GOAL-6	water sanitation management water_sanitation people drinking_water drinking population improved hygiene facilities safely water_resources global_population wastewater freshwater water_scarcity scarcity resources water_stress
GOAL-7	energy electricity renewable renewable_energy clean affordable modern cooking fuels access energy_efficiency energy_consumption energy_intensity intensity consumption access_electricity reliable efficiency technologies affordable_reliable
GOAL-8	growth labour employment unemployment work decent financial productivity decent_work financial_services productive men adults working youth economic_growth child labour_productivity jobs developed
GOAL-9	manufacturing infrastructure developing added manufacturing_added industrialization innovation gdp developed employment industries industrial mobile research job research_development intensity resilient_infrastructure resilient emissions
GOAL-10	inequality developed income developing duty exports oda money duty_free developing_states tariff remittances migration treatment reducing inequalities products migrant policies island
GOAL-11	cities urban waste air pollution slums urban_population solid_waste solid land urbanization management disasters air_pollution rapid safe resilient housing inclusive risk
GOAL-12	consumption production material consumption_production sustainable sustainable_consumption water convention material_consumption food domestic_material domestic impacts patterns natural capita production_patterns environmental pollutants wastes
GOAL-13	climate change climate_change agreement paris paris_agreement action global parties emissions adaptation convention temperature framework nations framework_convention united_nations determined climate_action degrees
GOAL-14	marine oceans ocean coastal resources areas fisheries ecosystems pollution protected_areas marine_resources fish overfishing biodiversity protected management stocks eutrophication ocean_acidification acidification
GOAL-15	biodiversity land species forests areas loss forest degradation wildlife desertification protected ecosystems terrestrial conservation resources halt land_degradation management covered biodiversity_loss
GOAL-16	institutions rights justice violence inclusive victims access_justice children human_rights data human societies trafficking effective peaceful levels sexual forms birth_registration registration
GOAL-17	developed development data oda developing capacity registration partnerships capacity_building building regions trade received agenda statistical enhance debt areas complete death_registration

These results confirm that the classifier is behaving as expected and is correctly and uniquely identifying individual SDGs. But how well will it do on the target data? The classifier is further validated by doing a spot check of its performance when used with the target data. The classifier is used to compute SDG Scores for some publications and the results are subjectively compared with the actual contents of the publications. This validation was done for three publications: one flagship and two shorter reports.

The 2013 World Economic and Social Survey titled "Sustainable Development Challenges" covers many of the SDGs and makes for a good test of how well-balanced and accurate the results are. The findings are shown in Table 4. According to the classifier, that WESS was principally concerned with Hunger (SDG 2), climate action (SDG 13), sustainable cities and communities (SDG 11), responsible consumption and production (SDG 12), and affordable and clean energy (SDG 7). The titles of the five chapters of the WESS are listed on the right column, matched with the corresponding SDG. As the results show, the top five SDGs identified by the classifier are a good reflection of the actual contents of the WESS.

A similar comparison is conducted on smaller publications that have a narrower focus to confirm the ability of the classifier to identify the main themes. The results of classifying DESA Working Papers number 107 and 124 are shown in Table 5. The classifier gives much higher weights to the main results, reflecting the more specialized nature of these publications compared to the results for the WESS discussed above. A subjective comparison of the results against the descriptions of the texts confirms that the scores are a good reflection of the contents of each publication.

Table 4
Comparing the SDG Scores with the contents of "WESS 2013: Sustainable Development Challenges"

SDG Score	Goals	WESS 2013 chapter titles
8.3%	GOAL 2: Zero Hunger	4) Ensuring food and nutrition security
7.2%	GOAL 13: Climate Action	1) Global trends and challenges to sustainable development post-2015
7.1%	GOAL 11: Sustainable Cities and Communities	3) Towards sustainable cities
6.1%	GOAL 12: Responsible Consumption and Production	2) Strategies for development and transformation
5.0%	GOAL 7: Affordable and Clean Energy	5) The energy transformation challenge
3.0%	GOAL 9: Industry, Innovation and Infrastructure	
2.6%	GOAL 17: Partnerships to achieve the Goal	
2.0%	GOAL 15: Life on Land	
1.9%	GOAL 10: Reduced Inequality	
1.4%	GOAL 6: Clean Water and Sanitation	
1.3%	GOAL 8: Decent Work and Economic Growth	
1.1%	GOAL 1: No Poverty	
0.6%	GOAL 14: Life Below Water	
0.0%	GOAL 3: Good Health and Well-being	
0.0%	GOAL 4: Quality Education	
0.0%	GOAL 16: Peace and Justice Strong Institutions	
0.0%	GOAL 5: Gender Equality	

Table 5
Comparing the SDG Scores with the contents of two DESA Working Papers

DWP 107: Food Crises and Gender Inequality		DWP 124: Innovative Development Finance: The Latin American Experience	
SDG Score	Goals	SDG Score	Goals
21.47%	GOAL 2: Zero Hunger	10.15%	GOAL 17: Partnerships to achieve the Goal
11.10%	GOAL 5: Gender Equality	9.27%	GOAL 13: Climate Action
4.47%	GOAL 15: Life on Land	8.71%	GOAL 10: Reduced Inequality
3.73%	GOAL 13: Climate Action	2.68%	GOAL 6: Clean Water and Sanitation
3.42%	GOAL 8: Decent Work and Economic Growth	2.06%	GOAL 15: Life on Land
3.28%	GOAL 9: Industry, Innovation and Infrastructure	1.73%	GOAL 1: No Poverty
2.26%	GOAL 10: Reduced Inequality	1.55%	GOAL 7: Affordable and Clean Energy
1.28%	GOAL 16: Peace and Justice Strong Institutions	1.53%	GOAL 9: Industry, Innovation and Infrastructure
1.22%	GOAL 17: Partnerships to achieve the Goal	1.50%	GOAL 8: Decent Work and Economic Growth
1.08%	GOAL 12: Responsible Consumption and Production	1.14%	GOAL 3: Good Health and Well-being
0.10%	GOAL 1: No Poverty	1.06%	GOAL 4: Quality Education
0.01%	GOAL 3: Good Health and Well-being	0.78%	GOAL 16: Peace and Justice Strong Institutions
0.01%	GOAL 6: Clean Water and Sanitation	0.10%	GOAL 2: Zero Hunger
0.01%	GOAL 4: Quality Education	0.02%	GOAL 12: Responsible Consumption and Production
0.00%	GOAL 7: Affordable and Clean Energy	0.00%	GOAL 11: Sustainable Cities and Communities
0.00%	GOAL 14: Life Below Water	0.00%	GOAL 5: Gender Equality
0.00%	GOAL 11: Sustainable Cities and Communities	0.00%	GOAL 14: Life Below Water
Description of DWP 107: "This paper examines the nature of the current *food crises*, the projected effect of *climate change* on food availability in developing countries, the vulnerabilities created by regional concentrations of food production, imports and exports, and the significant *role of women* as food producers, consumers and home food managers. A substantial body of work demonstrates that bridging the productivity differentials between male and female farmers, by helping women farmers overcome the production constraints they face, could significantly increase *agricultural growth and output*. This also becomes an imperative, given that the proportion of women in the total agricultural work force has been growing across the developing world, thus tending toward a *feminization of agriculture*."		Description of DWP 124: "This paper assesses to what extent the developing Latin America and the Caribbean (LAC) region benefited from IDF since its emergence in the early 2000s. It focuses on *global health* and *climate funds* through which most widely recognized IDF flows have been channelled. However, the paper also discusses *remittances*, flows ensuing from South-South Cooperation (SSC), and *financial transactions taxes*, in view of their importance on the ground as new sources of external finance, no matter whether these are recognized as IDF or not."	

The combined results of the various validation tests confirm that the classifier is doing a good job in correctly identifying the themes of publications. The following section gives the results of applying the classifier to a larger dataset to illustrate the type of analysis that is possible.

IV Classification results for DESA Publications

As mentioned previously, one of the advantages of an algorithmic approach to classifying text is its scalability to many texts. This section provides the results of the classification of DESA publications included in this study. The results for the full corpus are summarized first, followed by more detailed results of the DESA Working Papers, of the WESS, and of the RWSS. To the extent possible, different methods of presenting the results are used to illustrate the insights that can be gleaned from the rich set of results.

IV.1 Results for all DESA publications covered in this study

Table 6 below shows the average SDG Scores for each publication type included in this analysis. Goal 13 (climate action) is best represented in many publications. Most of the publications also discuss Goal 17 (partnership to achieve the goal). Goal 8 (decent work and economic growth) is of focus in the World Youth Reports as well as the RWSS. Goal 10 (inequality) has the second highest overall average and is particularly well represented in DESA Working Papers and in the RWSS.

Table 6
Average SDG Scores for each publication type

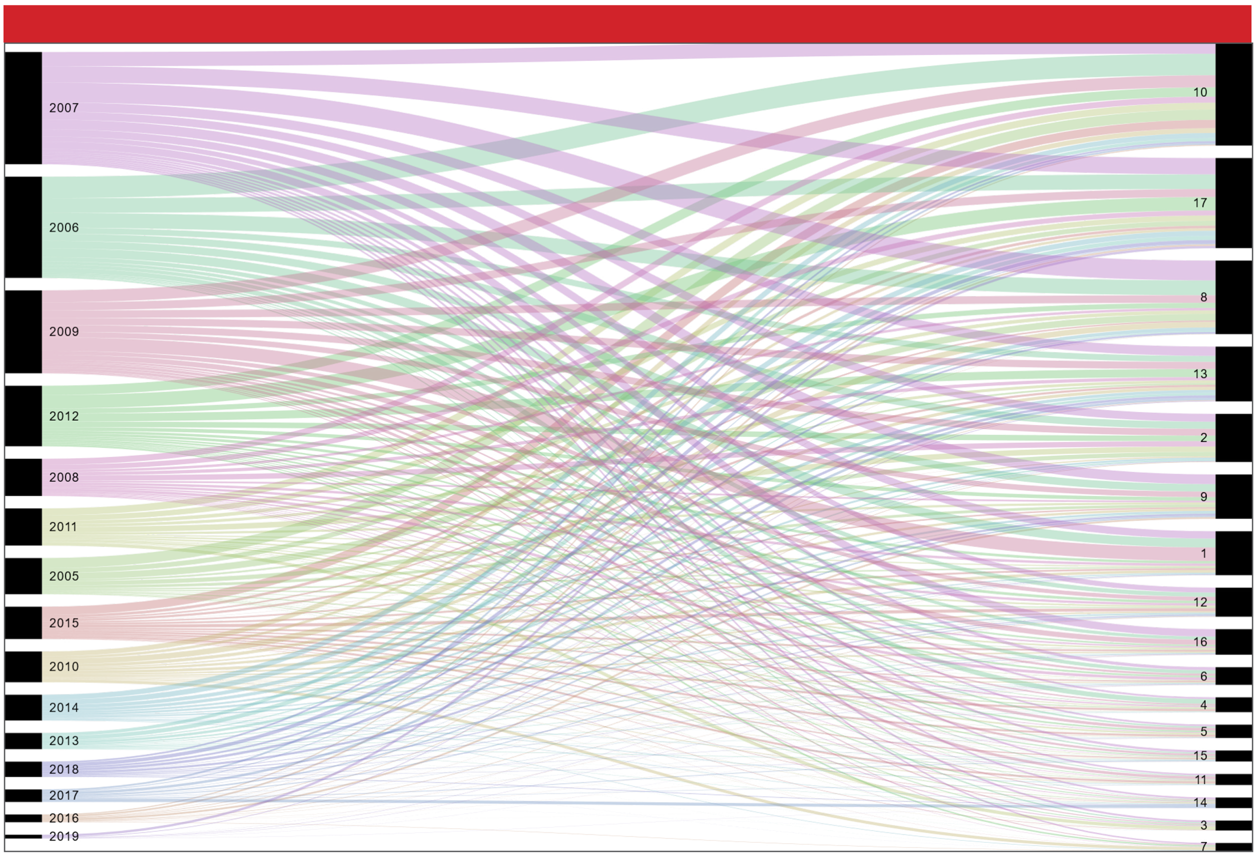
Publication	Goal 1	Goal 2	Goal 3	Goal 4	Goal 5	Goal 6	Goal 7	Goal 8	Goal 9	Goal 10	Goal 11	Goal 12	Goal 13	Goal 14	Goal 15	Goal 16	Goal 17
CDP Policy Note	1.95	2.57	2.02	0.56	0.70	1.14	1.69	3.34	2.85	5.63	0.42	1.92	9.86	0.63	0.70	2.72	11.34
DESA Working Paper	3.83	4.19	0.87	1.29	1.17	1.54	0.70	6.41	3.87	9.25	0.96	2.53	4.78	0.92	0.96	2.23	7.85
GSDR	4.40	0.56	0.48	1.84	1.92	2.92	1.28	0.67	5.43	1.81	3.10	3.13	5.90	5.55	2.40	2.17	8.43
NDS Policy note	2.90	5.17	0.27	0.98	0.56	0.78	0.55	5.46	5.17	7.48	0.81	3.68	3.41	0.78	0.62	3.00	10.27
DESA Policy note	2.19	5.41	1.78	0.02	0.13	1.42	4.76	2.88	2.83	3.11	1.67	1.29	9.97	1.73	2.80	0.73	7.04
Other Report	2.34	0.65	1.91	0.77	6.98	0.90	0.02	3.62	2.44	4.01	5.69	0.69	4.87	0.23	0.60	2.36	11.54
RWSS	10.53	2.84	1.94	3.59	3.24	0.28	0.00	11.33	1.35	8.13	0.98	0.45	1.57	0.04	0.15	2.57	2.02
WESS	2.50	4.66	1.28	0.96	0.51	0.95	1.96	4.30	4.35	6.89	1.27	2.51	6.83	0.79	1.19	1.88	7.49
WPSR	2.73	0.87	0.49	1.23	3.03	2.72	0.11	1.91	2.04	3.85	2.04	1.31	6.08	0.44	0.78	11.06	13.54
WYR	2.42	0.35	2.84	8.71	4.88	0.56	0.18	17.26	1.73	3.93	0.66	1.01	5.53	0.23	0.38	2.96	3.07
AVERAGE	3.42	4.03	1.18	1.29	1.28	1.39	1.44	5.73	3.53	7.34	1.23	2.16	5.95	0.99	1.22	2.26	7.85

The large number of publications limits how much detailed information can be shown as a table or graph. Instead, other forms of visualization can be useful to gain insights. Appendix 1 gives full results using network graphs that are able to give insights into the entire corpus as well as how each publication is related to each of the SDGs.

IV.2 Results for DESA Working Papers

The results for each of the 152 DESA Working Papers analyzed are shown below (Figure 3). The papers are grouped by year in an alluvial (or Sankey) diagram. As the figure shows, SDG 10 and SDG 17 were the most well represented in the collection (shown by the height of their black bars on the right column). For the WP series, 2006 and 2007 were particularly productive years. SDG 8 was much more represented between 2005 and 2007, and much less so in subsequent years.

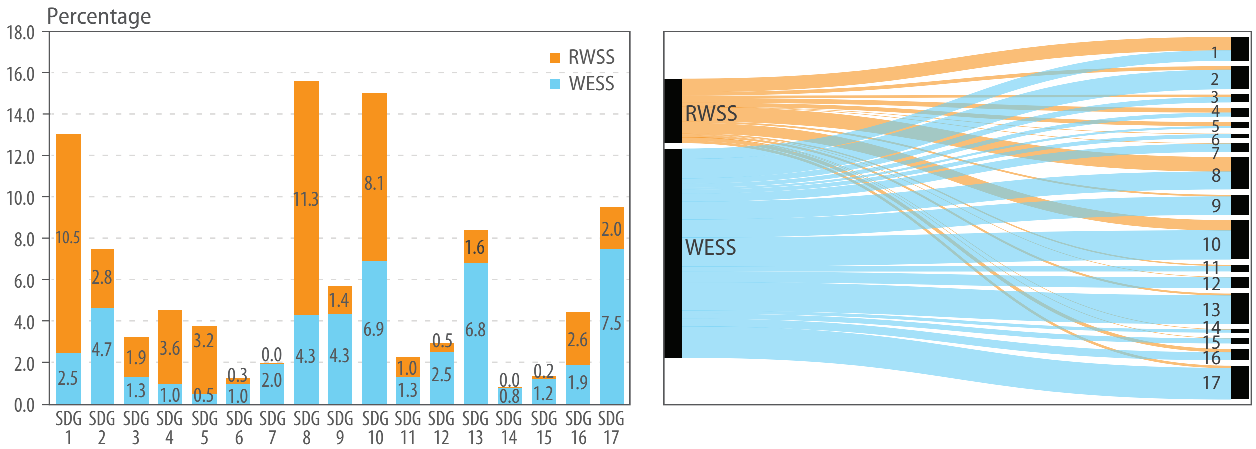
Figure 3
SDG Scores for each DESA Working Paper, by year (left) and by SDG (right)



IV.3 Results for the WESS and RWSS

The average SDG Scores for the RWSS and the WESS are shown below (Figure 4). The results show how these publications complemented each other in many ways. In the case of Goal 1, for example, the RWSS has much higher coverage than the WESS. The reverse was true for other Goals. WESS coverage of Goal 13, for example, was much higher. Goal 10 was well covered by both publications, while Goals 6 and 14 featured very little in either of these flagships. The alluvial diagram on the right gives additional insight and understanding of how each of the publications is linked to each of the 17 SDGs, and how they complement each other.

Figure 4
Average SDG Scores of WESS and RWSS

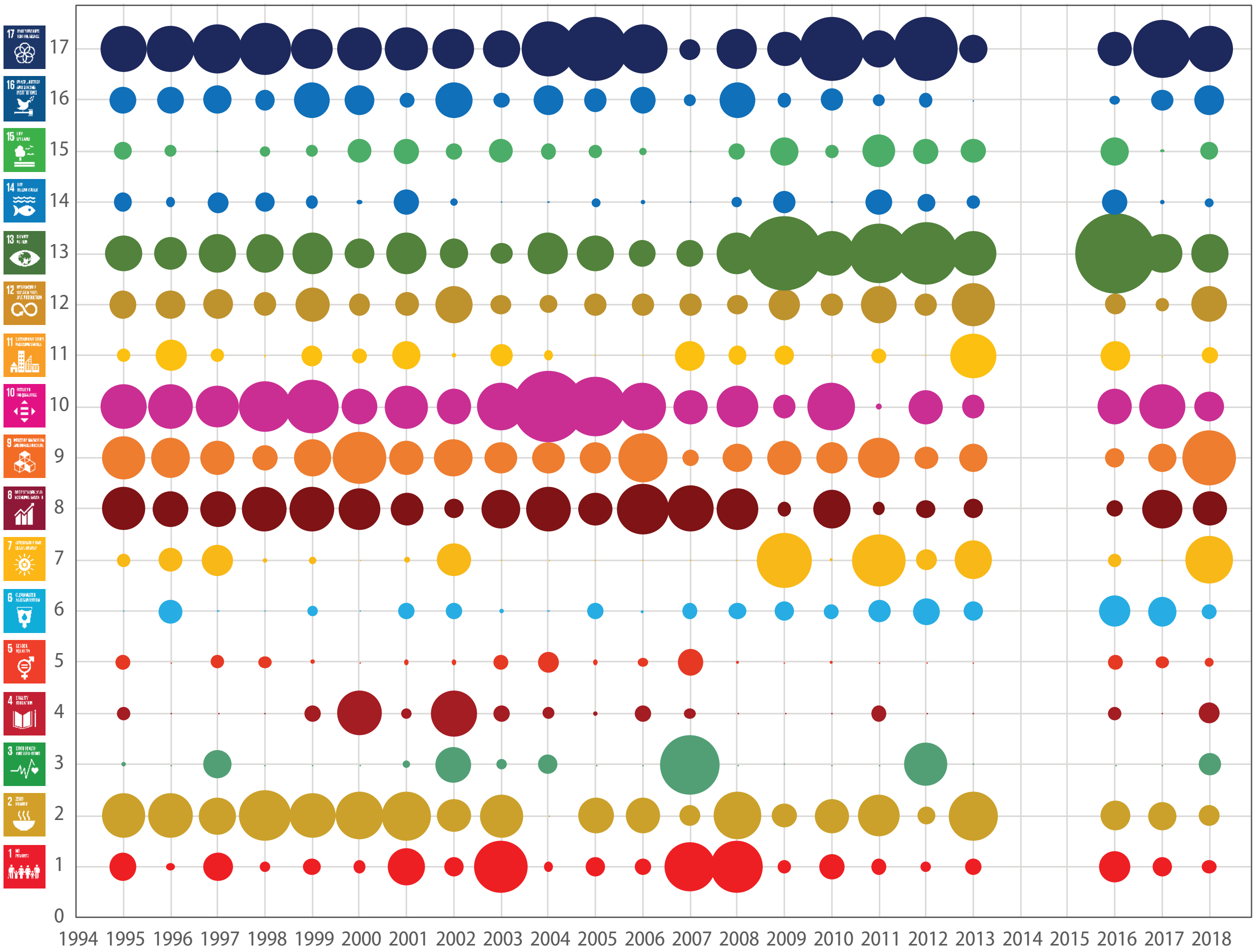


The diagram below shows how the SDG Scores for DESA publications have changed over time (Figure 5). Each publication is listed according to the year it was released on the horizontal axis. Each SDG is listed on the vertical axis in order. Each bubble shows how the text of any publication matches the text of any SDG. It is possible then to visualize how each SDG has been covered over time or how the focus of each publication shifts over time across the SDGs. It is also possible to visualize SDGs that are more consistently covered or not covered.

The figures show that the WESS has been more closely associated with Goals 2, 8, 9, 10, 13 and 17. Coverage of the other Goals has been more sporadic and, in the case of Goals 4, 5, 6, 14 and 15, almost non-existent. Overall, the DESA flagships analyzed here have consistently covered SDG 17. The WESS was particularly broad. A few SDGs were less represented by the publications analyzed here. DESA has specific Divisions that focus on these and this is likely a function of not having their outputs represented in the corpus analyzed in this study.[11]

11 It should also be noted that DESA is not expected to address all SDG-related themes equally. Other parts of the UN System have specific mandates linked to certain SDGs. There is no value judgment made in these results as to what should be the expected outcome. The results are reported here solely to illustrate the types of analysis that can be made available to administrators.

Figure 5
SDG Score of WESS, by year



V Conclusion and next steps

The use of topic modeling techniques to analyze vast quantities of text is commonly used in the humanities, but this is the first time these techniques have been used to classify DESA's publications. Developing this data-driven SDG Classification System adds to the previous efforts to better understand how the goals and targets of the 2030 Agenda are interconnected, and how the publications of the United Nations address the goals of the organization.

A measure of how well each of the 17 SDGs is represented in DESA's published work sheds light on which issues are well covered, and which issues are sparsely discussed. For decision makers, it can help to identify how to maximize the impact of publications and how to improve the alignment between DESA's work and SDG implementation—one of the ten points of the vision articulated by USG Liu in response to the call for DESA reform. Such an analysis should consider DESA's objectives which may not be complete coverage of all SDGs, but rather a prioritization based on its mission. A careful analysis of DESA publications to answer these questions is in progress and is left to a future paper.

This method hopefully inspires other efforts to use modern data analytics to better understand the body of work of the United Nations. New approaches and new ways to present the results can amplify their messages and help bring the work of the United Nations closer to the policymakers and to the interested public. Attractive and well-designed visualizations are powerful tools. They can provide an intuitive indication of the gaps or opportunities for research, motivate a more careful analysis of the results, and help users explore and discover DESA's work.

This work can be expanded in simple but powerful ways.

1. There are technical improvements to the classifier that can improve its accuracy and make the system fully supervised. This change will likely have the largest impact on the results as it would change the classification model.

2. The number and types of publications can be constantly expanded and classified, adding to the results. While the main aim of this paper was not to present definitive results of DESA's work, this is an important next step. The first priority should be to include the work of the Divisions that are not yet reflected in the current dataset. Research questions must be carefully defined in coordination with the needs of DESA leadership. There are also ongoing efforts to use this classification system on speeches by senior UN officials, and to explore its wider use.

3. Finally, the visual presentation of the results in charts, tables, and dynamic websites is an area where innovation can have a significant impact. A proof-of-concept website showing the results of the topic model has been tested and aims to make the results easily accessible and to give users a way to interact with the latest data and discover DESA publications.

Appendix I

Using network diagrams for insights into the results of topic models

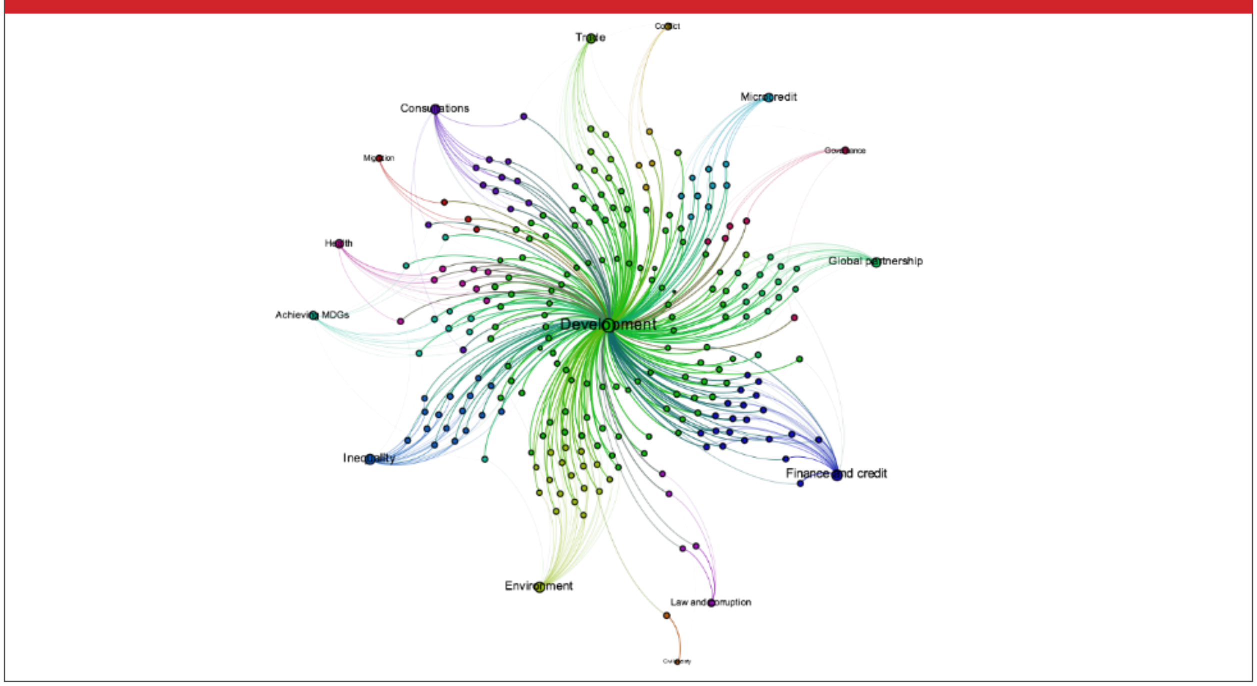
As the number of publications being analyzed increases, presenting the results of the topic model becomes increasingly challenging. The simplest solution is to present average results or the results of individual publications of interest. A table of average results is informative but it is often useful to have a full view of the corpus to answer questions like: which issues are well covered, or which publications are more closely associated with an issue.

A network diagram of the publications and their connections with each theme gives immediate insights without the need to present multiple tables of results. The diagram is created by an algorithm that allows each point to "push and pull" against all SDG nodes, like a spring, until the entire network is balanced and settles on a shape. The strength of each connection is the SDG Score computed from the model above.

The figure below shows the results of an unsupervised 15-topic model of over 200 of DESA's publications (Figure 6). Just as described in the main text above, an algorithm first built a classifier by identifying the 15 topics that best represent the entire collection.[12] Then, using the trained classifier (or topic model), the algorithm computed the strength of the connection of each publication to each of the 15 topics.

Using a network diagram to interpret the results is straightforward. As the figure shows, all publications share a strong connection to a central topic (labelled "development" in the diagram). This makes sense as it represents

Figure 6
Network diagram showing the results of a 15-topic model of 233 DESA publications



12 The number of topics is chosen by the user, based on technical parameters or research priorities.

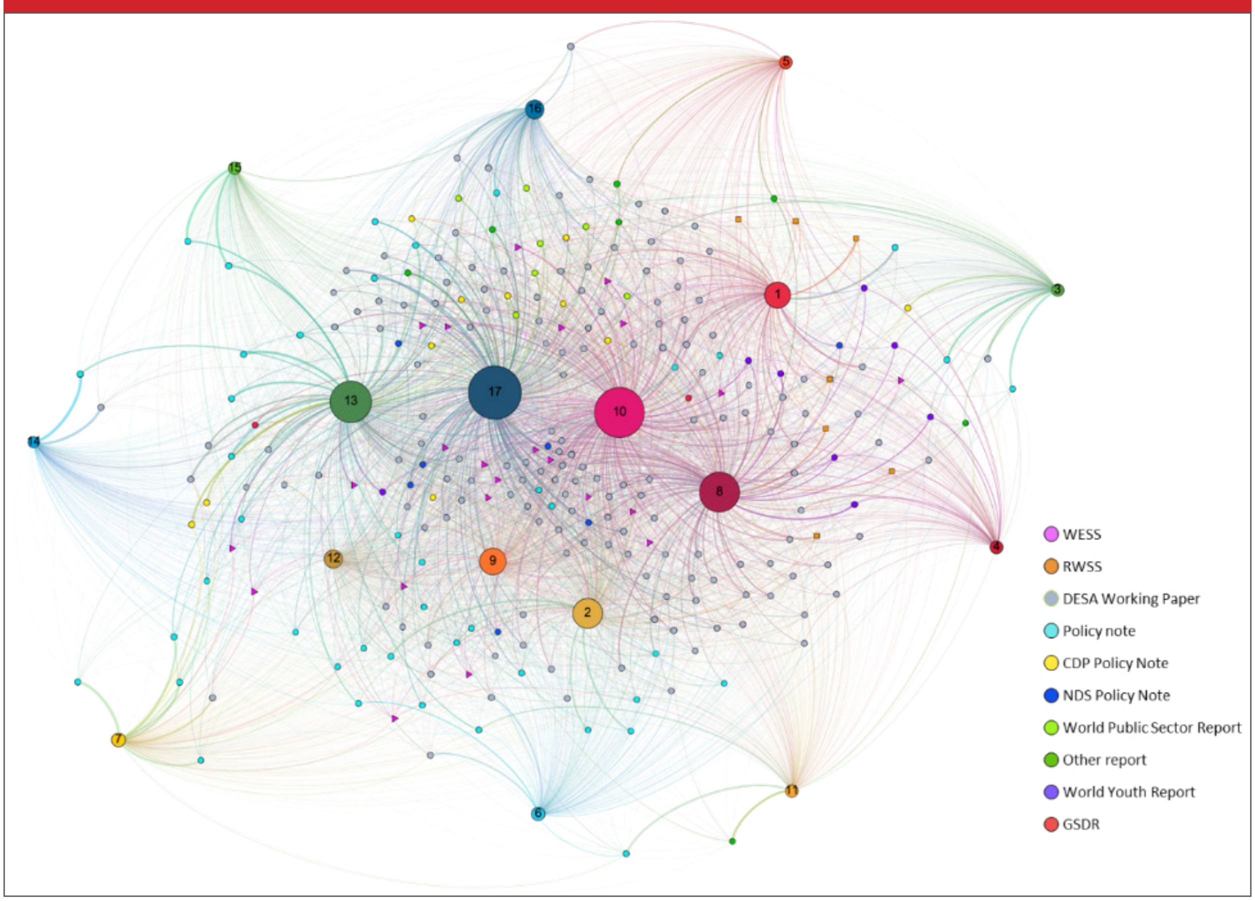
a common, department-wide topic. More interesting are the 14 topics that surround the publications. Some topics, like "finance and credit" are closer to the center, indicating a stronger connection to some documents. Others like "conflict", "governance" and "civil society" show much weaker links to each publication.

The results for the semi-supervised model, trained to classify according to each of the 17 SDGs is very different. The network diagram below clearly shows how well each publication fits with each of the SDGs (Figure 7). Each dot represents an individual publication (publication types are colored according to the legend) and the SDGs are the 17 larger numbered circles. Each publication is linked to each of the 17 SDGs.

Interpreting the network diagram is intuitive. The distance between the publication and an SDG is a function of the relative importance of that SDG to the publication. It is easy to quickly "place" the publication in the space of SDGs by simple visual inspection. For example, in the figure below, the SDGs most important for the entire collection of DESA publications are those pulled into the center due to the stronger connections with more publications: SDGs 8, 10, 13 and 17. The SDGs and the publications at the periphery of the space are those with weaker links, and therefore weaker importance, to the corpus. SDG 14, for example, shows only two publications with strong connections. SDG 7 shows five nearby publications. In the diagram below, the SDG nodes are also sized according to how important they are to the overall network, giving an additional indication of the importance of each SDG.

These SDGs found in the periphery may well represent areas of opportunity for more specialized and focused policy research and advice.

Figure 7
A topological view of the link between DESA publications and the SDGs



Appendix II

A conceptual explanation of the SDG classification system

It may be difficult for those who first encounter topic models and machine learning to gain a conceptual understanding and a mental image of the process. This appendix aims to help in this regard.

Imagine there are 17 "buckets" that each contain hundreds of words. Bucket 1 may contain words like "poverty", "extreme", "income", "cash", etc. Bucket 2 may contain the words "food", "hunger", "agriculture", etc.

Some of the words in each bucket may appear more than once. In bucket 1, for example, the word "poverty" may appear many more times than the word "fish". Words can also be part of multiple buckets. "Poverty" appears many times in buckets 1 and 10, but only rarely in bucket 14.

In this set up, each bucket represents the vocabulary of each of the SDGs, and the frequency of the words in each bucket represents the importance of each word in that SDG. The contents of the buckets are created using a machine learning algorithm that tries to find the best fit to the 17 texts that represent the SDGs.

To classify a publication we want to know how well its vocabulary matches each of the 17 SDG buckets. Put another way, how much should we use from each bucket to get as close as possible to the words that make up the publication?

A publication about poverty will likely use more words from bucket 1, some from bucket 10, and almost none from bucket 14. A publication about hunger will use more from bucket 2 but some from other buckets also.

The proportion of words from each bucket represents how important each SDG is in the publication. This is the SDG Score.

Appendix III

Possible contribution to and alignment with other initiatives

There are ongoing initiatives in the UN System designed to bring documents and data into the Semantic Web[13] and which would be complemented by the methodology described in this paper.

One approach is to create metadata structures to support the automatic retrieval, processing, and integration of information related to the SDGs. For example, the Akoma Ntoso initiative[14] is an XML schema that exposes the structure and semantic components of parliamentary, legislative and judiciary documents including those from the UN. This enables efforts to increase the efficiency and accountability of processes using analytics.

There are also efforts to formalize and normalize the definitions of terms to support interconnections. The UN Environment SDG Interface Ontology (SDGIO) at the request of the IAEG-SDG in its 2nd meeting (Bangkok, October 2015) is developing such an ontology of the SDG terms. They aim to organize information logically in a coherent network of meaning to guide information, data mobilization and analysis. A common SDG ontology will formalize the structure of connections between goals, targets, and indicators. Once complete, individual terms like "tourism" in SDG Indicator 8.9.1 will have the information needed for machines to understand what the goal/target/indicator is about. This term will then be linked to the corresponding terminology in other platforms such as the UN System Data Catalogue and the SDG Innovation platform.[15]

As these efforts mature and new interconnections are created using semantic terminology, new interconnections and analytics will become possible. An SDG Classification will be useful in this regard, helping to inform the degree of interconnectivity and add an additional dimension to this analysis. It will also be useful to further inform existing document retrieval systems like the UN Bibliographic Information System Thesaurus (UNBISNET).[16]

13 The Semantic Web extends the World Wide Web by using common data formats and exchange protocols to facilitate the sharing of data across applications, enterprises, and communities.

14 www.akomantoso.org

15 http://aims.fao.org/activity/blog/sustainable-development-goals-interface-ontology-sdgio-support-united-nations

16 http://unbisnet.un.org/