Scientometrics (2022) 127:6661–6693
https://doi.org/10.1007/s11192-022-04358-x

# Mapping the sustainable development goals (SDGs) in science, technology and innovation: application of machine learning in SDG-oriented artefact detection

**Arash Hajikhani**[1] · **Arho Suominen**[1,2]

Received: 15 August 2021 / Accepted: 17 March 2022 / Published online: 5 April 2022

## Abstract

The sustainable development goals (SDGs) are a blueprint for achieving a better and more sustainable future for all by defining priorities and aspirations for 2030. This paper attempts to expand on the United Nations SDGs definition by leveraging the interrelationship between science and technology. We utilize SDG classification of scientific publications to compile a machine learning (ML) model to classify the SDG relevancy in patent documents, used as a proxy of technology development. The ML model was used to classify a sample of patent families registered in the European Patent Office (EPO). The analysis revealed the extent to which SDGs were addressed in patents. We also performed a case study to identify the offered extension of ML model detection regarding the SDG orientation of patents. In response to global goals and sustainable development initiatives, the findings can advance the identification challenges of science and technology artefacts. Furthermore, we offer input towards the alignment of R&D efforts and patenting strategies as well as measurement and management of their contribution to the realization of SDGs.

**Keywords** Sustainable development goals · Patents · Publications · Natural language processing · Machine learning model

✉ Arash Hajikhani
arash.hajikhani@vtt.fi

[1] Quantitative Science and Technology Studies, VTT Technical Research Centre of Finland, Tekniikantie 21, 02044 Espoo, Finland

[2] Industrial Engineering, Tampere University, Korkeakoulunkatu 8. PL 541, 33014 Tampere, Finland

## Introduction

Our planet faces massive economic, social and environmental challenges. To combat these, the United Nations established a universal initiative known as the sustainable development goals (SDGs) that define global priorities and aspirations for 2030 in the form of 17 goals and 169 targets.[1] The goals represent a pathway to eliminate extreme poverty and aims to put the world on a sustainable path. Achieving the SDGs is not just a political commitment but an opportunity to establish a technology transition (Nations, 2021). The post-2015 sustainable development agenda calls on all countries to enhance research, upgrade technological capabilities, encourage innovation and increase public and private investment (Giovannini et al. 2015).

Science, technology and innovation (STI, as referred to in the UN and OECD contexts) have been recognized as one of the main drivers behind productivity increases and a key long-term lever for economic growth and prosperity (Daejeon Declaration, 2015). STI is a fundamental tool to implement the new agenda, as it allows improving efficiency in both the economy and society, developing new and more sustainable ways to satisfy human needs, and empowering people to drive their own future (Giovannini et al. 2015). In the SDGs framework, STI features strongly in Goal 17, as well as a cross-cutting one to achieve several sectoral Goals and Targets. Fostering innovation is part of Goal 9 related to resilient infrastructure and inclusive, sustainable industrialization, while Target 9.5 elevates the role of research and innovation policy well beyond STI as one of the means of implementation.

In the United Nations general assembly briefing materials, the importance of science, technology and innovation (STI) for the SDGs has been numerously mentioned in yearly forums.[2] A direct quote from Marie Chatardová, President of the Economic and Social Council, at the 2018 New York STI Forum "No one can ignore the vital role of science, technology and innovation (STI) in advancing the transformative impact".[3] Similarly, the Technology Adviser to the US Secretary of State said that the integrated nature of the SDGs requires multi-disciplinary and holistic science, technology and innovation approaches that break silos and take into account different sources of knowledge at the concluding session of the Forum.

Innovation in general and innovation in the context of sustainable development affects many parts of human life and should thus be treated with concern. The SDG framework is built on the expectation of technological development and innovation. For example, Goal 9 needs to be highlighted in the context of innovation. This goal states that the objective by the United Nations is to "promote inclusive and sustainable industrialization and foster innovation". STI is a democratizing tool for transferring science to society and can be instrumental in introducing solutions for achieving SDGs. The importance of innovation in reaching sustainable development is also recognized by previous research, for example by Ashford and Hall (2011).

[1] Sustainable Development Goals and targets descriptions can be accessed at United Nation website: https://sdgs.un.org/goals.

[2] UN, Science, Technology and Innovation for the SDGs 2018 Forum: https://www.un.org/development/desa/indigenouspeoples/science-technology-and-innovation-for-the-sdgs.html.

[3] Science, technology and innovation crucial to 'transformative impact' of Development Goals, UN 2018 forum hearing: https://www.un.org/development/desa/en/news/sustainable/sti-forum-2018-opening.html.

One of the main ways how STI's oriented efforts are manifested is through scholarly literature and intellectual property protected in the form of patents. The ability to analyse science and technology output has increased tremendously in the past decade due to the increasing degree of digitalization of research articles and intellectual property databases (e.g., Web of Science, Scopus, PATSTAT, Google Patents). The deep technical analysis of patents' and publications' textual content and bibliographic metadata provides valuable interpretation for the most complex and technology-oriented artefacts. To a significant extent, previous literature still focuses on descriptive values rather than creating in-depth data that creates additional vantage points to evaluate STI's orientation towards SDGs in overall innovation systems. This research aims to extend the advancements in the quantitative analysis of STI to enhance detection and instances of SDGs in documented scientific and patented materials. Advance text analytics tools combined with artificial intelligence techniques offer the opportunity to analyse science and technology development through publications on a large scale, with easy implementation, and in a harmonized manner with long time horizons.

This research focuses on analysing science and technology literature contribution to SDGs using machine learning methods. Using a limited training data incorporated from a lexical based search query in a publication database, a machine learning model is compiled to identify the relevance of patents towards SDGs on a larger scale. Furthermore, the machine learning model will be evaluated to estimate its extension of SDG oriented artifact coverage in comparison to a standard lexical based query-based search.

The paper will continue to present a background study and outline the research design. After that, the methodological approach is described, followed by a presentation of the results. Finally, the results will be discussed in the last section, elaborating on further implications and future research.

## Background

### The science, technology and innovation perspective towards SDGs

STI and the interaction between different actors are the core indicators for economic growth (Dosi et al., 2006; Freeman, 2004; Nelson & Sidney, 1982). Increase in scientific and technological knowledge production acts as the key source of innovation and competitive advantage (Pavitt, 1991). This has been central to our understanding of the competitiveness of nations and the competitiveness at the firm level. The centrality of the concept of productivity and its increase to sustain the long-term competitiveness of nations has been a vital paradigm of economic policy for decades. This also explains that much of the literature on the STI process focuses on innovation outcomes (Kahn, 2018). Much of the focus is centred on scientific work and research and development within the innovation system within this literature. These are seen as vehicles to enable job creation, firm performance and ultimately increases in the gross domestic product (Fukuda, 2020; Goos et al., 2015; Klomp & Van Leeuwen, 2001).

However, an ongoing debate is focused on extending our focus beyond productivity or gross domestic product to other impact measures (Stiglitz et al., 2018). Global challenges like the climate crisis have strengthened these developments, which has led governments and businesses to reassess the role of pure productivity as a goal. The climate crisis has strengthened the call for additional outcome measures for innovation system activities in the public sector, often referred to as the Beyond GDP measurement framework (Hayden, 2021; Malay, 2019, 2021; Schreyer, 2021). This has been clear in broadening public impact assessment (Nieminen & Hyytinen, 2015). Policy has also actively discussed the role of grand challenges and the role of governments to take an active role in facilitating transitions unlikely to happen through other means but creating significant overall benefits (e.g. Mazzucato, 2011). We have also seen significant transitions in company leaders' positions to the role of companies in grand challenges. Large companies' CEOs' call to extend the firm's objectives beyond shareholder value (Gelles & Yaffe-Bellany, 2019) can be considered a major transition beyond the current paradigm towards a sustainable economy.

In this transition, the work of the United Nations on the creation of SDGs has been central. SDGs offer one of the first holistic taxonomies of grand challenges. With the advent of the SDG framework, as well as a shift in thinking, the innovation system has had to undergo transformative changes (Schot & Steinmueller, 2018). This again is mainly discussed in the policy domain, but the literature also looks at the role of industry and innovation activities concerning the SDGs. There is a need to adjust all aspects of economic, governance, and public policy at all levels if science, technology, and innovation to reorient to the SDG agenda (Walsh et al., 2020).

It is clear that the SDG framework calls for broad changes in technologies, policies and innovation (Leach et al., 2012). The sustainability transition expects that the socio-technical regime will realign to produce a totally different type of value as compared to the current regime (Schot & Steinmueller, 2018). Hajer et al. (2015) call for deeper integration of "planetary boundaries", "safe and just operating space", "energetic society" and "green competition" to realize the sustainability transition described through the SDGs. Many argue that it is the private sector that is apt to respond to the transformation (Scheyvens et al., 2016) by taking advantage of new technological solutions (Sinha et al., 2020), such as Industry 4.0 (Bonilla et al., 2018). We have seen evidence that the SDG framework has been used as a policymaking tool and frame for corporate innovation, partnerships, and strategy development (Sullivan et al., 2018).

The relevance of mapping and monitoring STI interaction and development is crucial for understanding the dynamics behind the innovative performance, growth and competitiveness of nations and even firms. Indicators signalling interactions between scientific and technological activities are highly relevant in this respect (Ranaei et al., 2017). Researchers and policymakers have for several decades recognized patents as valid and reliable indicators of technology development and innovation (Callaert et al., 2014). Patent documents contain essential research results that are valuable to the industry, business, and policymaking communities. If carefully analysed, they can show technological details and relations, reveal business trends, inspire novel industrial solutions, or help make investment policy (Campbell, 1983). In addition, patent data is an essential source of information for the company's STI policymakers and other stakeholders. Analysis of patent data could essentially enrich the scope and depth of strategic technology policymaking, affect the alignment of a company's innovation strategies, as well as the evaluation of R&D proposals and the assessment of technology competitiveness (Ena, 2021).

Considering patent documents, we estimate the impacts of industrial activity on the SDGs. This requires the development of practical proxy measures to establish a measurement of corporate activities' impact on the goals. For example, approaches towards measuring the societal impact of innovation have been done in the context of frugal innovation (Altgilbers et al., 2020). In the SDG context, van der Waal et al. (2021) used patents to explore the impact of firms on the SDG goals. Similar to Xie and Miyazaki (2013), the authors extend the established patent classification based analysis of Migotto and Haščič

(2015), e.g. green technologies, into patent applications' content. This requires building a taxonomy of terms to be identified from the patent text, subsequently informing on the SDG relevance of the document.

However, the literature has not shown practical approaches for creating a proxy measure for large scale analysis of STI impact on the SDGs. Our study attempts to measure SDG relevancy to intellectual property (IP) types of documents such as patents. The approach used is based on utilizing the classification of scientific documents based on their relevance to SDGs. While patent documents differ from scientific documents, previous research has shown that using natural language processing and machine learning can model the interaction between science and technology documents (Ranaei et al., 2017). The approach used relies on an existing classification of scientific publications' relevance to SDGs as a gold standard, trains a model with scientific publications and transfers the model to patent data to create a classification of patent relevance. We demonstrate the approach on the European Patent Office patent families of 2020 and explore the classification on SDG 7 (clean energy) in depth.

## Machine learning-based classification of science and patent documents

Machine learning-based classification of science and patent documents is an important avenue to complement existing human assigned classifications (Suominen & Toivanen, 2015). It has been shown that the modeling of topics can be an effective approach to identifying important scientific and technological (S&T) documents (Yau et al., 2014). Overall, automated text classification is essential for knowledge management and S&T studies (Nedjah et al., 2009). Classifications have been seen to enable the creation of novel ways of mapping science (Suominen & Toivanen, 2015) and businesses capabilities (Suominen et al., 2016). Text classification is not merely an application of an algorithm across a text corpus but rather a process. Classifying text includes several phases, including preprocessing, document modeling, e.g., vector-space, feature selections, and algorithm utilization for model building and evaluation (Mirończuk & Protasiewicz, 2018). An important aspect of the process is the selection of the algorithms implemented, namely the selection between a supervised and unsupervised approach (Samira Ranaei et al., 2019).

Focusing particularly on the classification methods in automated classification of S&T documents, Ranaei et al. (2019) review the use of supervised and unsupervised classification methods in S&T literature classification. Supervised learning, an approach reliant on the use of training data, lists six major approaches used in literature. The most commonly used supervised method to classify S&T related text is the Support Vector Machine (SVM). For example, Kreuchauff and Korzinov (2017) use an SVM approach to patent data and found that the approach allowed them to reduce expert bias and inherent issues in using a citation based approach. In Kenekayoro (2018), SVM was used to accurately identify name entities from academic biographies, enabling a model for classifying information on researchers. The second most often used approach is a Naïve Bayes approach. This is used by Lee and Lee (2019) to identify technological opportunities using patent forward citations. Wang et al. (2019) used the Naïve Bayes approach to predict the success of academic articles. Among the other methods employed, namely KNN and Random Forest, the Naïve Bayes approach has shown consistent performance.

In addition to the two most commonly used approaches, S&T literature classification has also been done using Neural networks, K-Nearest Neighbor, Logistic Regression and Supervised Fuzzy Algorithms (Samira Ranaei et al., 2019). Similar to the ones previously mentioned, these methods are employed to answer S&T-relevant questions. For example, Neural networks have been used to predict academic authors' success (Mistele et al., 2019) and the impact to patent citations of firms (Chen & Chang, 2010). But in many cases multiple classifiers are used to tackle the research question. For example, this is the case in Wang et al. (2019), where the study considered a Naïve Bayes, K-Nearest Neighbors, SVM, and Logistic Regression approach.

For unsupervised methods, which rely on a formal framework, Ranaei et al. (2019) highlighted eight approaches commonly used. These are Principale Components Analysis (PCA), Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), Kmeans, Unsupervised Fuzzy Algorithm, Artificial Neural Network (ANN), Hierarchical clustering, and Probabilistic LSI. For example, Kenekayoro et al. (2015) used PCA on research groups' website key phrase vectors to cluster similar research. Zhou et al. (2021) used PCA on patent data to identify structures in patent data that could indicate clusters of emerging technology. Their research particularly focuses on the identification of outlier patents. The work of Yau et al. (2014) extensively covers different topic modeling approaches, including LDA. In the work, Yau et al. (2014) tested several approaches of topic modeling to scientific publication clustering and show the potential of methods such as LDA in producing meaningful research clusters. LDA has been since used on science mapping (Suominen & Toivanen, 2015), clustering themes in emerging technologies (Ranaei et al., 2020) and analysing firms' behaviour through patents (Suominen et al., 2016). ANN is also a reoccurring method in S&T classification task. For example Lu et al. (2020) used a Multi-layer perceptron as a layer in their approach to classify and find similar patents. In addition, Yoon and Phaal (2013) used a K-means approach to classify patent documents in the context of technology road mapping. Yoon et al. (2010) used hierarchical clustering to develop meaningful research and development clusters.

The performance of text classifiers is not simply determined by the number of training examples or the quality of the training set, but may also be influenced by the intrinsic characteristics of the texts being analysed, which can include the size of the vocabulary and the lengthy nature of the document (Figueroa & Zeng-Treitler, 2013). In a comparative study on classification methods' performances on social media data, authors have foreseen that results may differ for other sources, languages, and classification objectives (Hartmann et al., 2019). The study of Ruijie et al. (2021) on patent text modeling strategy and its classification based on structural features, sample size, and imbalance dataset was vital in exploiting the full advantages of Word2vec and deep learning algorithms. As a result of such sensitivity of context, data type, sample size and its diversity, quality of labelled data, and notably the classification application, the majority of studies have benchmarked various methods for finding the best performer in their niche application.

Meanwhile, specific methods are frequently used and benchmarked in text classification challenges, and reportedly although some perform better than others, the differences among methods' performances are not significant. On a text classification experiment with BBC news, Shah et al. (2020) revealed a Logistic Regression classifier with the TF-IDF vectorizer feature attains the highest accuracy in a Comparative analysis among Random Forest and KNN Models. The findings of another study also indicated that the Logistic Regression multi-class classification method for product-reviews has achieved the highest classification accuracy in comparison with the Naïve Bayes, Random Forest, Decision Tree, and Support Vector Machines classification methods (Pranckevičius & Marcinkevičius, 2017).

Considering the context of our study, the objectives, data types, availability of labelled data and its quality, we proceed with a research design to experiment with a combination of methods. We acquire a set of methods both on feature designing and classification

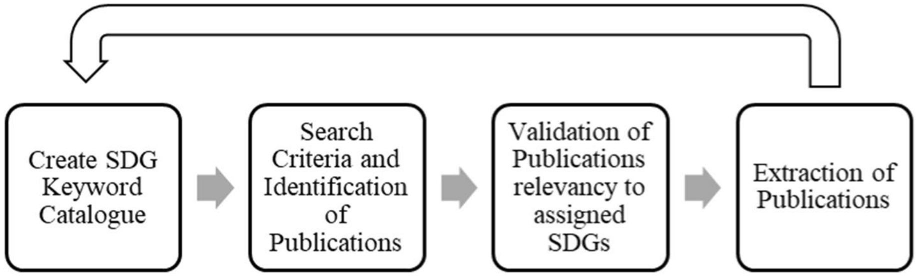

Create SDG Keyword Catalogue
Search Criteria and Identification of Publications
Validation of Publications relevancy to assigned SDGs
Extraction of Publications

**Fig. 1** Workflow process of identifying and retrieving SDG related publications

approaches to allow us to observe the performance across methods. The following section will elaborate on our methodological choices, starting from data collection to machine learning classification and their evaluation.

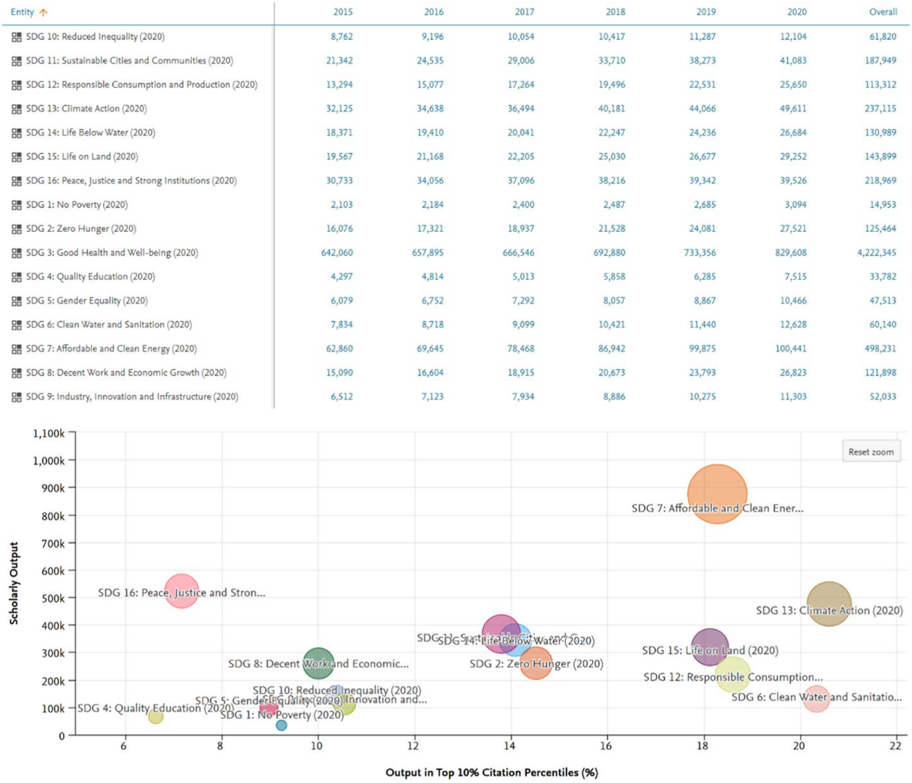
## Methodology

### Query crafting and data collection

With the overall aim to identify SDG related science and technology and innovation relevance, the study design is grounded on creating a lexical query by utilizing an iteratively developed database of SDG terminology (Fig. 1). A keyword search was applied to map scientific publications during the period of 2015–2020. Our decision to include papers from 2015 is motivated by the fact that in September 2015, the General Assembly adopted the 2030 Agenda for SDGs (United Nations Department of Public Information, 2015). Therefore, this period has been chosen as it encompasses the most recent activities, at least partly impacted by the SDG agenda. The outcome of this process results in research publications in relation to the SDG focus. Meanwhile, patent documents due to the nature of their content and descriptive, have not given representative results in regard to their relevancy to SDGs by use of SDG lexical search queries. While we have received representative results from publication data, we are able to extrapolate the machine-learning model based on publication data to detect relevant patents to SDGs.

A detailed taxonomy has been developed for mapping SDG relevant publications. To start we utilized the taxonomy curated by Scopus SciVal[4] and its effort for compiling SDG queries in "Identifying research supporting the United Nations Sustainable Development Goals". Furthermore, the process of curating the lexical keywords is complemented by involving the analyses of the UN Sustainable Development Goals documents (UN, 2015). From the semantics perspective, each word or concept has been expanded to lexically similar words. In addition, the extracted list of keywords was matched with existing taxonomies (Elsevier, 2015; Jia et al., 2019; UNSDG, 2019; Vatananan-Thesenvitz et al., 2019). Using the keywords, gathered queries are compiled for each SDG and then searched from the SCOPUS publication database. SCOPUS is the largest abstract and citation database

[4] The Scopus methodology regarding Identifying research supporting the United Nations Sustainable Development Goals can be found from this publication: https://data.mendeley.com/datasets/87txkw7khs/1.



| Entity ↑ | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | Overall |
|---|---|---|---|---|---|---|---|
| SDG 10: Reduced Inequality (2020) | 8,762 | 9,196 | 10,054 | 10,417 | 11,287 | 12,104 | 61,820 |
| SDG 11: Sustainable Cities and Communities (2020) | 21,342 | 24,535 | 29,006 | 33,710 | 38,273 | 41,083 | 187,949 |
| SDG 12: Responsible Consumption and Production (2020) | 13,294 | 15,077 | 17,264 | 19,496 | 22,531 | 25,650 | 113,312 |
| SDG 13: Climate Action (2020) | 32,125 | 34,638 | 36,494 | 40,181 | 44,066 | 49,611 | 237,115 |
| SDG 14: Life Below Water (2020) | 18,371 | 19,410 | 20,041 | 22,247 | 24,236 | 26,684 | 130,989 |
| SDG 15: Life on Land (2020) | 19,567 | 21,168 | 22,205 | 25,030 | 26,677 | 29,252 | 143,899 |
| SDG 16: Peace, Justice and Strong Institutions (2020) | 30,733 | 34,056 | 37,096 | 38,216 | 39,342 | 39,526 | 218,969 |
| SDG 1: No Poverty (2020) | 2,103 | 2,184 | 2,400 | 2,487 | 2,685 | 3,094 | 14,953 |
| SDG 2: Zero Hunger (2020) | 16,076 | 17,321 | 18,937 | 21,528 | 24,081 | 27,521 | 125,464 |
| SDG 3: Good Health and Well-being (2020) | 642,060 | 657,895 | 666,546 | 692,880 | 733,356 | 829,608 | 4,222,345 |
| SDG 4: Quality Education (2020) | 4,297 | 4,814 | 5,013 | 5,858 | 6,285 | 7,515 | 33,782 |
| SDG 5: Gender Equality (2020) | 6,079 | 6,752 | 7,292 | 8,057 | 8,867 | 10,466 | 47,513 |
| SDG 6: Clean Water and Sanitation (2020) | 7,834 | 8,718 | 9,099 | 10,421 | 11,440 | 12,628 | 60,140 |
| SDG 7: Affordable and Clean Energy (2020) | 62,860 | 69,645 | 78,468 | 86,942 | 99,875 | 100,441 | 498,231 |
| SDG 8: Decent Work and Economic Growth (2020) | 15,090 | 16,604 | 18,915 | 20,673 | 23,793 | 26,823 | 121,898 |
| SDG 9: Industry, Innovation and Infrastructure (2020) | 6,512 | 7,123 | 7,934 | 8,886 | 10,275 | 11,303 | 52,033 |

Reset zoom
Scholarly Output
1,100k
1,000k
900k
800k
700k
600k
500k
400k
300k
200k
100k
0
SDG 7: Affordable and Clean Ener...
SDG 16: Peace, Justice and Stron...
SDG 13: Climate Action (2020)
SDG 11: Sustainable Cities and C...
SDG 14: Life Below Water (2020)
SDG 15: Life on Land (2020)
SDG 8: Decent Work and Economic...
SDG 2: Zero Hunger (2020)
SDG 12: Responsible Consumption...
SDG 10: Reduced Inequality (2020)
SDG 5: Gender Equality (2020)
SDG 9: Industry, Innovation and...
SDG 6: Clean Water and Sanitatio...
SDG 4: Quality Education (2020)
SDG 1: No Poverty (2020)
6
8
10
12
14
16
18
20
22
Output in Top 10% Citation Percentiles (%)

**Fig. 2** Top: count of publication per each SDG category for 2015–2020 based on lexical query searched on Scopus. Bottom: SDG category publication volume on top 10% citation percental spectrum

of peer reviewed literature. It includes books, scientific journals, and conference proceedings. Compared to other scientific databases such as the Web of Science, SCOPUS has a broader coverage and it is a widely used database to create datasets for systematic reviews of research (Mongeon & Paul-Hus, 2016). The resulting publications were manually validated for the relevancy of the resulted records to the corresponding SDG. Figure 2 illustrates the number of publications within each SDG category based on Scopus/Scival query per year (2015–2020).

The bibliometric data for the resulting publications were extracted for each SDG category. For each category, we download 2000 top publications based on Scopus "Relevance Score".[5] Within the bibliometric data, we extracted the textual content of the publications, such as title, abstract and keywords, used to train a model for automating unseen SDG-related document detection. Appendix 1 offers descriptive statistics of the publication labelled data.

[5] Scopus's Relevance Score is a statistical feature that calculates how well text in the documents returned as search results reflect the terms and criteria executed in a search query. The score implements itself in the order of the search result meaning the higher the relevance the lower the search rank would be. More detail on Relevance score,coputation can be read at: https://service.elsevier.com/app/answers/detail/a_id/14182/supporthub/scopus/.

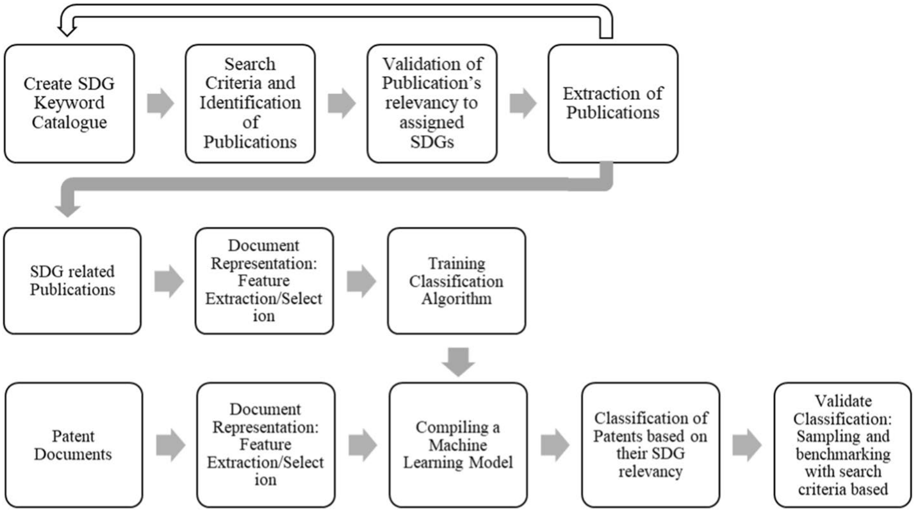

Create SDG Keyword Catalogue
Search Criteria and Identification of Publications
Validation of Publication's relevancy to assigned SDGs
Extraction of Publications
SDG related Publications
Document Representation: Feature Extraction/Selection
Training Classification Algorithm
Patent Documents
Document Representation: Feature Extraction/Selection
Compiling a Machine Learning Model
Classification of Patents based on their SDG relevancy
Validate Classification: Sampling and benchmarking with search criteria based

**Fig. 3** Workflow for the sustainable development goals (SDG) mapping of patents

## Text modeling and machine learning classification

This research has benefited from the development of machine learning classification algorithms on texts to facilitate the construction of the SDG detecting model. Text classification is one of the research hotspots in the field of Natural Language Processing (NLP). Originated from computer science and evolved from pattern recognition, the automated process of categorization (or classification) of objects such as text has become a growing interest in utilizing machine learning. Additionally, the increasing availability of documents in digital form has led to the development of methods to facilitate comprehending the essence and insight of documents in new ways (Sebastiani, 2001). Text classification by machine learning algorithms allows processing a large amount of text automatically and creating better insights on the documents. Approaches for algorithmic classification of documents can be based on several methods, mainly supervised, unsupervised or reinforced learning. The difference in the approaches can generally be made on either the use of human tagged data, reliance on the algorithms' framework, or learning by rewards from correct actions. Within this study, the SDG related publications' content identified in the previous step will be utilized as a training set for the classification algorithm.

The Python programming language was used to handle the data structuring and ML model building.[6] There are several classification methods within supervised learning for performing a multi-class text classification. We perform a comparison among the top frequent used methods to select the best performing model for our classification task. In the process of validating the accuracy and reliability of the model, we use a test set (new publications) to confirm the usability of the ML model. The adjustment for the size of the training was set to be 70% and the test set 30%. Figure 3 illustrates the workflow of the methodological steps.

[6] Python version 3.8.5 running in anaconda's Jupyter-notebook version 6.1.4.

Before applying the classification algorithms on the SDG labelled publications, preprocessing was applied on the texts. The preprocessing involved cleaning procedures (e.g. stop words removal, non-alphanumeric characters removal, stemming and lowercase transformation) applied to harmonize and increase the consistency of the text.[7] We considered Naive Bayes Classifier for Multinomial Models, Linear Support Vector Machine, Logistic Regression as classifiers and TF-IDF, Word2vec and Doc2vec as text modeling strategies.[8]

## Text modeling

**TF-IDF** For Natural Language Processing (NLP) to work, it requires transforming natural language (text) into a numerical vector form. Text vectorization techniques, namely TF-IDF, Bag of Words and vectorization, are very popular choices for machine learning algorithms and can help convert text to numeric feature vectors. Therefore, to quantify and convert text into numerical representation in documents, (1) we assign a fixed integer id to each word occurring in any document of the training set and (2) For each document *i*, count the number of occurrences of each word *w* and store it in *X[i, j]* as the value of feature *j* where *j* is the index of word *w*. This method of representation of text implies that *n* features is the number of distinct words in the corpus that can be handled with today's computational capacity. This high-dimensional sparse datasets will be built with scikit-learn built in function "CountVectorizer" which supports counts of N-grams of words or consecutive characters. Text preprocessing, tokenizing and filtering of stop words are all included in CountVectorizer, which builds a bundle of features and transforms documents to feature vectors.

Meanwhile occurrence count is a good start but the issue would be longer documents will have higher average count values than shorter documents, even though they might talk about the same topics. This downscaling of features is treated by computing a weight to each phrase that signifies the importance of the phrase in the document and corpus. The TF-IDF method is a widely used technique in Information Retrieval and Text Mining (Manning et al., 2012). Term Frequency Inverse Document Frequency (TF-IDF) is a weighting procedure that tries to evaluate the relevance of terms in the document corpus. As the term implies, TF-IDF calculates values for each phrase in a document through an inverse proportion of the frequency of the phrase in a particular document to the percentage of documents the phrase appears in. The formula below was used to compute the TF-IDF:

$$w_{jk} = tf_{jk} \times idf_j \tag{1}$$

$w_{jk}$ = phrase weight of phrase $j$ in document $k$.

$$idf_j = \log_2\left(\frac{n}{df_j}\right) \tag{2}$$

$tf_{jk}$ = the number of phrases $j$ that occur in document $k$;$n$ = the total number of documents in the document set; $df_j$ = the number of documents containing the phrase $j$ in the document set.

[7] For text cleaning, harmonizing and structuring, the following python packages was utilized: NLTK 3.5, Pandas 1.1.3, Numpy 1.19.2 and re 2.2.1.

[8] For utilizing classification algorithms and machine learning tasks, scikit-learn version 0.23.2 and Genism 3.8.3 packages was employed.

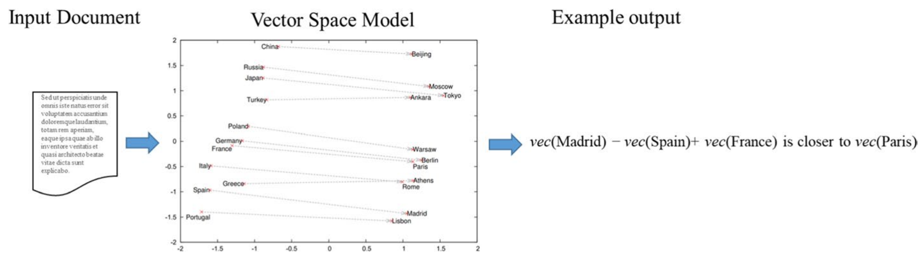

Input Document
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.
Vector Space Model
China
Beijing
Russia
Japan
Moscow
Tokyo
Turkey
Ankara
Poland
Germany
France
Warsaw
Berlin
Paris
Italy
Greece
Athens
Rome
Spain
Portugal
Madrid
Lisbon
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
Example output
vec(Madrid) − vec(Spain)+ vec(France) is closer to vec(Paris)

**Fig. 4** Illustrative representation of Word2vec with a sample of skip-gram model projected for country and capital vectors by Mikolov et al. (2013)

Both tf and tf–idf can be computed as follows using TfidfTransformer from scikit-learn 1.0.2. With the feature implementation explained here, we will perform the classifier task with Naive Bayes Classifier for Multinomial Models, Linear Support Vector Machine and Logistic regression. We utilized the scikit-learn 1.0.2 pipeline class that behaves like a compound classifier, including CountVectorizer and TfidfTransformer.

**Word2vec** The Word2vec model and application by Mikolov et al. (2013) have attracted significant attention in recent years. The vector representations of words learned by Word2vec models have been shown to carry semantic meanings and are helpful in various NLP tasks. In order to make a learning algorithm more efficient and eliminate the risk of over-fitting, we are seeking for a way to decrease dimensionality. This model is based on a neural network that is designed to reconstruct the linguistic contexts of words. Specifically, Word2vec uses a large corpus of text and produces a vector space model where words that share common contexts in the corpus (i.e., being nearby words) are positioned close to each other (Mikolov et al., 2013).

There are two different approaches to word2vec: continuous bag-of-words (CBOW) and continuous skip-gram (Rong, 2014). The CBOW model architecture predicts the current word based on the window of surrounding context words, while the skip-gram model uses the current word to predict the surrounding window of context words. Previous studies show that although slower, the skip-gram is more accurate and reliable for infrequent words versus the CBOW (Goldberg & Levy, 2014). This study adopts the skip-gram model considering that the distribution of products in the integrated patent-product database is highly skewed, and infrequent products in the database can provide valuable insights into potential technology opportunities. Figure 4 is a simple illustration of a Word2vec model showing the transformation of a document where a vector is built for each word in the document. Having all the word vectors and portraying them in a vector space, a distance measure can identify similar words to each other.

Figure 4 presents part of the results of the skip-gram model trained on country-capital city data (Mikolov et al., 2013). The model organizes concepts (i.e., country and capital city) and examines their relationships without any supplementary information. For instance, countries such as France and Germany are located nearby in the resulting vector space, and the distance between France and Paris corresponds with that of Germany and Berlin.

Doc2vec is a generalization of Word2vec with the distinction that it summarizes text contained within multiple documents. It is a straightforward extension of the Word2vec model which was developed to represent words meaningfully in a vector space (provide "word embeddings") (Demeester et al., 2016; Mikolov et al., 2013). The objective of Word2vec is to situate words that have similar meanings close to one another. Similarly, Doc2vec aims to situate similar documents close to one another by placing document vectors (Docvec) close to each other in vector space. To do this, the algorithm uses the "context" around each term in the document to derive a vector representation that maximizes the probability of its appearance.

For operationalization, we benefited from Python's Gensim Word2vec feature as it provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. The tool takes a text corpus as input and produces word vectors as output. It first constructs a vocabulary from the training text data and then learns vector representations of words. We used the Google News corpus (3 billion running words) word vector model as a pre-trained model (3 million 300-dimension English word vectors).

## Machine learning classification

**Naive Bayes classifier for multinomial models** Naive Bayes is one of the algorithms used in our research to train a machine learning classification model. Naive Bayes is a probabilistic model which works well on text categorization (Weikum, 2002). Naive Bayes classifiers are built on Bayesian classification methods. These rely on the Bayes's theorem, which is an equation describing the relationship of conditional probabilities of statistical quantities. Naive Bayes classifier is a general term that refers to the conditional independence of each of the features in the model. In contrast, multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier that uses a multinomial distribution for each feature. Manning et al. (2012) described that the multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as TF-IDF may also work. The probability calculation is described in Eq. 3:

$$P(c \mid d) \propto P(c) \prod_{1 \leq k \leq n_d} P\left(t_k \mid c\right) \tag{3}$$

where $P(tk \mid c)$ is the conditional probability of the word $tk$ that appears in a document having class $c$. In the equation $P(tk \mid c)$ is the likelihood probability of $tk$ in class $c$. While $P(c)$ is the prior probability of the document appearing in class $c$. Class determination compares the posterior probability results obtained, then the class with the largest posterior probability is the class chosen as the predicted result. The prior probability formula can be seen in Eq. 4:

$$P(c) = \frac{N_c}{N} \tag{4}$$

where $Nc$ is the sum of category $c$, while $N$ is the sum of all categories. The formula for likelihood probability can be seen in Eq. 5:

$$P(t_k|c) = \frac{T_{tc}}{\sum_{t' \in V} T_{ct''}} \tag{5}$$

where $T_{tc}$ is the number of occurrences of the word t in the document having class $c$, and $\sum_{t' \in V} T_{ct''}$ is the total number of occurrences of all words in class$c$.

For operationalizing this method, we used the sklearn 1.0.2 “sklearn.naive_bayes.MultinomialNB” package featured in Python 3.8.5.

**Linear support vector machine** Support vector machines are supervised machine learning algorithms that are very effective for classification problems (Awad & Khanna, 2015). Support vector machines (SVM) are primarily classifiers that can classify by constructing hyperplanes that separate cases that belong to different categories (Ingo & Andreas, 2008). In our application, the SVM classifier should first be trained on a set of tagged publications with their SDG category relevancy so that the machine gains knowledge for effective categorization.

In the following, we briefly describe its core concept and discuss some advantages that are relevant to the problem at hand. Simply put, the core idea of the method is to create a unique discrimination profile (represented by a linear function) between samples from different classes. We have an iterative training algorithm for the application to define a hyperplane that effectively separates two or multiple classes. During the training phase, a support vector machine partitions a high-dimensional space based on the points it contains that belong to known classes.

The ‘‘touching’’ data points are termed support vectors. In fact, the resulting separation plane is shaped only by these constraining (= supporting) points. Below, we provide the mathematical notation of a support vector machine following Hsu et al. (2010), an article which is a comprehensive introduction to the method for purposes such as ours. Formally defined, we have a training set $\left(x_i, y_i\right)$ of $i = 1, \ldots, 1$ sample points (here: our publications), where every $x_i \in R^n$ is an attribute vector (consisting of our normalized word and n-gram frequencies) and $y_i \in \{-1, 1\}^l$ is a decision for that specific data point which thus defines its class. The SVM then yields the solution to the following optimization problem (see as well Boser et al. 1992; Guyon et al. 1993):

$$\begin{gathered}\min_{w,b,\xi} \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{l} \xi_i \\ \text{s.t. } y_i\left(w^{\mathrm{T}}\Phi\left(x_i\right) + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0\end{gathered} \tag{6}$$

in which $W$ is the normal vector between the separating hyperplane and the parallel planes spanned by the support vectors. The mapping $\Phi$ is related to so-called Kernel functions, such that $K\left(x_i, x_j\right) \equiv \Phi\left(x_i\right)^T \Phi\left(x_j\right)$. For problems in which the data under consideration are not linearly separable, $\Phi$ maps the training attributes into a higher dimensional space where a hyperplane may be found.

The above version of the classification procedure also incorporates the so-called Soft-Margin method (Cortes and Vapnik 1995) that allows for mislabelled training sample points. The approach introduces $\xi_i$ as nonnegative slack variables which measure the extent of incorrectly classified items in the training set. $\sum_{i=1}^{l} \xi_i$ is thus a penalty term, and C a penalty parameter, on which we will comment later. For operationalizing this method, we used the sklearn 1.0.2 “sklearn.linear_model.SGDClassifier” package featured in Python 3.8.5.

**Logistic regression** Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable (Cramer, 2005). In Logistic Regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y = 1) as a function of X.

Multinomial Logistic Regression (MLR) studies the probability of the D-dimensional point $xi \in \{x1, x2, x3 \ldots, xN\}$ pertinent to a class $k \in \{1, 2, \ldots, K\}$. The data of the shape $(xi, yi)i = 1, \ldots, N$ wherever $xi \in R$, d is a d-dimensional representation of the feature vector and $yi \in \{1, 2, 3, \ldots, K\}$ is a label connected with it; K is the number of class labels. Let $yik = I(yi = k)$ indicate the membership of data point x i to class k.W $= \{w1, w2, \ldots, wK\}$ indicates the parameter vector for all K classes. The probability that xi is pertinent to class k is given by:

$$p(y = k \mid xi) = \frac{\exp\left(w_K^T xi\right)}{\sum_{j=1}^{K} \exp\left(W_K^T xi\right)} \tag{7}$$

For model building we used the built-in function in Python 3.8.5 Scikit-learn 1.0.2 “sklearn.linear_model.LogisticRegression”.

## Model assessments metrics

We measure the performance of the models by precision (also called positive predictive value) which is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. In an imbalanced classification problem with more than two classes, precision is calculated as the sum of true positives across all classes divided by the sum of true positives and false positives across all classes (Sokolova & Lapalme, 2009). As a result, the generalization from a two-class confusion matrix when extrapolated to a multiclass, would be to sum over rows / columns of the confusion matrix. Given that the matrix is oriented as a two-class case, i.e., that a given row of the matrix corresponds to specific value for the "truth", we have:

$$\text{Precision}_i = \frac{M_{ii}}{\sum_j M_{ji}}$$

$$\text{Recall}_i = \frac{M_{ii}}{\sum_j M_{ij}}$$

$$F = 2 \cdot \frac{\text{precision} \bullet \text{recall}}{\text{precision} + \text{recall}}$$

That is, precision is the fraction of events (M) where we correctly declared $i$ out of all instances where the algorithm declared $i$. Conversely, recall is the fraction of events where we correctly declared $i$ out of all of the cases where the true of state of the world is $i$. Both precision and recall are therefore based on relevance and F1 Score is the weighted average of precision and recall. For reporting on the above-mentioned measures, we used the sklearn 1.0.2, “classification_report” feature in Python 3.8.5.

# Analysis and results

## Machine learning model design and deployment

The described approach to identify relatedness of patents to SDG goals is implemented in Python programming language. At the initial stage, the textual content in the patents and publications was preprocessed with python functions to clean irregular characters, harmonize and tokenize the input texts. We performed a strict cleaning process regarding stop words and meaningless words, which were facilitated by visualization of top word clusters to identify such words. Text preprocessing is traditionally essential for natural language processing (NLP) tasks. It transforms text into numerical values so that machine learning algorithms can perform better. The phases for text preprocessing include, stop word removal, stemming, lemmatization which at the end aims to normalize all the text on a level playing field. The text will also be tokenized, which is about splitting text strings into smaller pieces, or "tokens". We also remove noise from the initial text, such as cleaning up text from extra whitespaces, lowercase all text, and removing special characters. Finally, we will convert our text documents into a numerical representation with either TF-IDF (term frequency-inverse document frequency) or other vectorizing techniques depending on the classification approach. After that, we train several classifiers using of Python's Scikit-Learn library and Gensim library.

Once the features are generated from the text, the machine learning classifier can be trained with the labelled publication data we had collected earlier. Five different classification strategies have been tested, and the performance of each model is reported in Table 1. The overall accuracy of the models did not reach above 60% (based on the first prediction of the model). This benchmark is not a high standard of accuracy for accepting the model for good classification of all the classes. Nevertheless, our model experimentations deliver acceptable accuracy (above 60%) for most of the SDG classes such as: SDG 1, 2, 3, 4, 5, 6, 7, 9, 10, 13 and 16. This highlights that some of the SDGs seems to be highly difficult to identify from text, while many of the classifiers operate with significant accuracy. This ultimately leads to some limitations in the ability to identify some of the SDGs. Overall, based on the model comparison in Table 1, the highest overall accuracy (f-score) is achieved by "Word2vec and logistic regression" models. This model is selected for further analysis while keeping in mind the limitations of identifying e.g. SDG 8, 14 and 15.

The classifier performances reflected in Table 1 is varied among the SDG classes. Over the iterative training, testing, and validating process, we tried to adopt the recommended measures to enhance the classification performances, yet some classes suffer from very low accuracy. This has multiple reasons which we tried to elaborate on the nature of such an issue in the paper. Other reports and early studies also admit the unbalanced performance between SDG classes (Performance of the SDG-BAI algorithm developed by OECD (2021a)). One initial cause is the SDG classes definitions. While some classes have a certain focus (i.e. SDG7), the others are described more broadly, making it difficult to isolate a specific artefact that has addressed that particular class. For example, looking at SDG classification of Scopus publication based on publication volume and being in citation percentile indicates categories are over-represented, and some have minor representation. Figure 2 shows the publication volume of SDG relevant publications in Scopus with the cohort positioning in the top citation percentiles. The figure indicates the disparity in each SDG category where SDG7 is dominant in size and citation percentile.

**Table 1** Classification models performance comparison

| | Naive Bayes classifier for multinomial models | | | Linear support vector machine | | | Logistic regression | | | Word2vec and logistic regression | | | Doc2vec and logistic regression | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score |
| SDG1 | 0.53 | 0.75 | 0.62* | 0.67 | 0.72 | 0.69* | 0.60 | 0.56 | 0.58 | 0.65 | 0.68 | 0.66* | 0.58 | 0.62 | 0.60* |
| SDG2 | 0.57 | 0.60 | 0.58 | 0.57 | 0.67 | 0.62* | 0.52 | 0.61 | 0.56 | 0.61 | 0.62 | 0.62* | 0.57 | 0.60 | 0.58 |
| SDG3 | 0.82 | 0.88 | 0.85* | 0.69 | 0.93 | 0.79* | 0.87 | 0.89 | 0.88* | 0.86 | 0.87 | 0.86* | 0.81 | 0.89 | 0.85* |
| SDG4 | 0.75 | 0.74 | 0.74* | 0.71 | 0.87 | 0.78* | 0.78 | 0.77 | 0.77* | 0.78 | 0.75 | 0.76* | 0.76 | 0.77 | 0.77* |
| SDG5 | 0.55 | 0.72 | 0.63* | 0.59 | 0.79 | 0.68* | 0.60 | 0.57 | 0.58 | 0.65 | 0.61 | 0.63* | 0.62 | 0.61 | 0.61* |
| SDG6 | 0.57 | 0.63 | 0.60* | 0.58 | 0.76 | 0.66* | 0.62 | 0.59 | 0.61* | 0.62 | 0.66 | 0.64* | 0.64 | 0.63 | 0.63* |
| SDG7 | 0.83 | 0.78 | 0.80* | 0.69 | 0.86 | 0.76* | 0.80 | 0.81 | 0.81* | 0.82 | 0.84 | 0.83* | 0.83 | 0.82 | 0.82* |
| SDG8 | 0.50 | 0.46 | 0.48 | 0.59 | 0.40 | 0.48 | 0.41 | 0.41 | 0.41 | 0.52 | 0.46 | 0.49 | 0.50 | 0.48 | 0.49 |
| SDG9 | 0.49 | 0.66 | 0.56 | 0.61 | 0.70 | 0.65* | 0.59 | 0.56 | 0.57 | 0.59 | 0.66 | 0.62* | 0.59 | 0.54 | 0.56 |
| SDG10 | 0.69 | 0.45 | 0.54 | 0.76 | 0.56 | 0.64* | 0.62 | 0.57 | 0.59 | 0.63 | 0.56 | 0.60* | 0.52 | 0.49 | 0.50 |
| SDG11 | 0.51 | 0.44 | 0.47 | 0.55 | 0.54 | 0.55 | 0.42 | 0.48 | 0.45 | 0.53 | 0.54 | 0.53 | 0.53 | 0.55 | 0.54 |
| SDG12 | 0.56 | 0.40 | 0.47 | 0.65 | 0.36 | 0.46 | 0.45 | 0.43 | 0.44 | 0.54 | 0.49 | 0.51 | 0.47 | 0.42 | 0.44 |
| SDG13 | 0.50 | 0.64 | 0.56 | 0.55 | 0.60 | 0.58 | 0.55 | 0.50 | 0.53 | 0.56 | 0.59 | 0.58 | 0.60 | 0.66 | 0.63* |
| SDG14 | 0.16 | 0.16 | 0.16 | 0.13 | 0.10 | 0.11 | 0.10 | 0.11 | 0.11 | 0.31 | 0.37 | 0.34 | 0.30 | 0.29 | 0.29 |
| SDG15 | 0.10 | 0.07 | 0.09 | 0.09 | 0.06 | 0.07 | 0.11 | 0.11 | 0.11 | 0.25 | 0.19 | 0.22 | 0.31 | 0.31 | 0.31 |
| SDG16 | 0.73 | 0.49 | 0.58 | 0.74 | 0.51 | 0.61* | 0.60 | 0.64 | 0.62* | 0.60 | 0.69 | 0.65* | 0.56 | 0.57 | 0.57 |

*Marks performances which were above 0.60 or 60% of accuracy

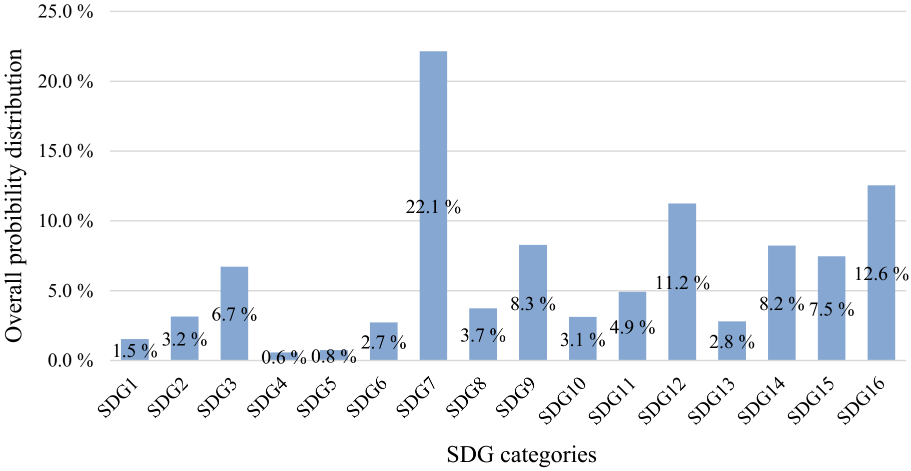


**Fig. 5** 2020 EPO patents and their orientation towards SDG categories

There have been experiments indicating the similarity between categories, which reduces the classification task in distinguishing each class separately. Based on Patentsight[9] metrics, we also realized a very low identification of some of the SDG classes, which we also had difficulties isolating, such as 15, 14 and 8. Nevertheless, our focus has been to expand the identification capability of SDG from the classes where we had better performance and therefore made the case analysis on SDG7, which was among our highest performing classes.

The classifier is used on the patent data to evaluate patents' relatedness to the SDGs. Patent data is collected from the Patbase service. The data is limited to the year 2020 and all the granted patent families, and for these we retrieved textual content (title and abstract), patent number, assignee and patent country code. This resulted in a dataset of 132,226 patent families which is produced by over 37 thousand assignees. Appendix 1 offers some descriptive statistics regarding the patent data.

The patent data was processed using the same preprocessing steps used in publication data. After passing the preprocessed texts into the ML model, we were able to return the relevancy of each patent document to any of the SDGs. The result is a probability distribution for each document to be relevant to each SDG. Keeping in mind the ability of the classifier to predict relatedness to specific SDGs, we have aggregated the probability score of each record regarding each SDG category. Figure 5 presents the accumulative highest to lowest relevant SDGs which was addressed in the patent application texts.

Figure 5 indicates that more than 20% of patent families address SDG 7 in their textual content. The ML model by design can give us the affiliation of the patent text to any of the SDGs. This in practice results in knowing to what other SDG classes a patent text is close to. In Appendix 4, we have illustrated the top three ML guesses of patents.

In order to benchmark the ML model ability to identify SDG relatedness, a lexical query was retrieved and evaluated. Based on the ML model results in terms of precision, recall and overall f-score performance, the Word2vec classifier was the most accurate in

[9] Patentsight (https://www.patentsight.com/en/) is a provider of analytics solutions on Patent related data.

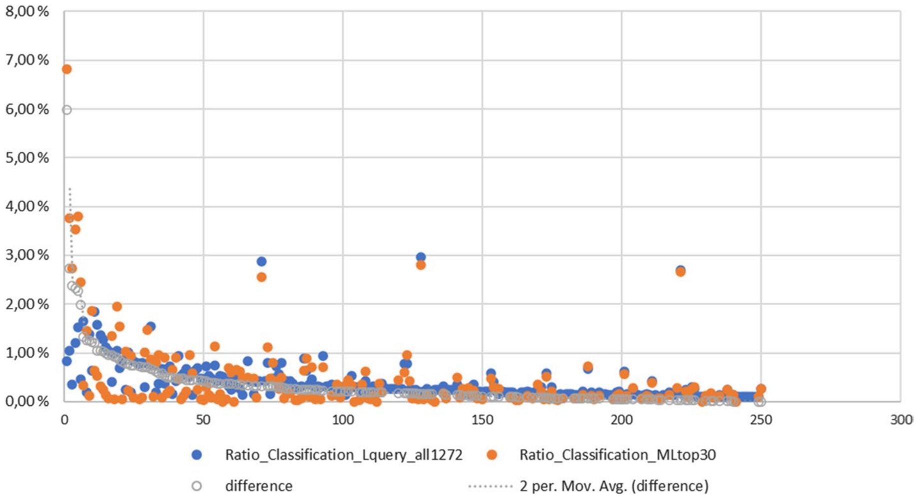


**Fig. 6** Difference of family sizes in IPC classes between lexical query and ML resulted query

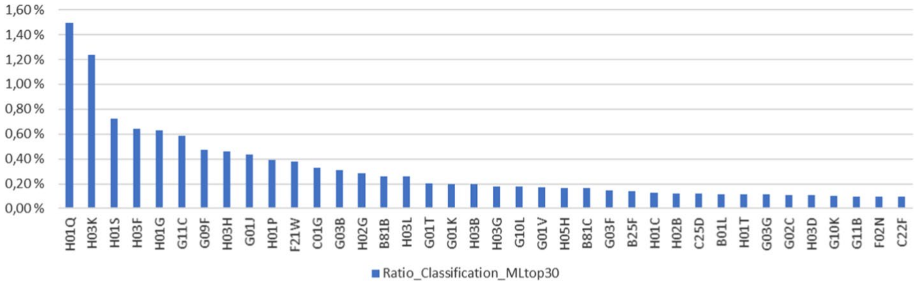


**Fig. 7** Top IPC classes within 2020 EPO patent families classified as SDG 7 by the ML model

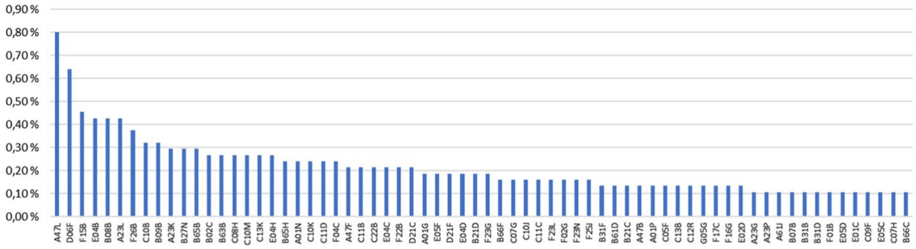


**Fig. 8** Top IPC classes within SDG 7 lexical query identified patent families

detecting SDG7 relevance. Therefore, we continue to detect the SDG 7 categorized patents in our EP families set as a good candidate subset for benchmarking. We adopted the lexical query (Appendix 1) used for SDG 7 publications and used it to retrieve granted patents for 2020 from Patbase. This resulted in a dataset of 1272 patent families.

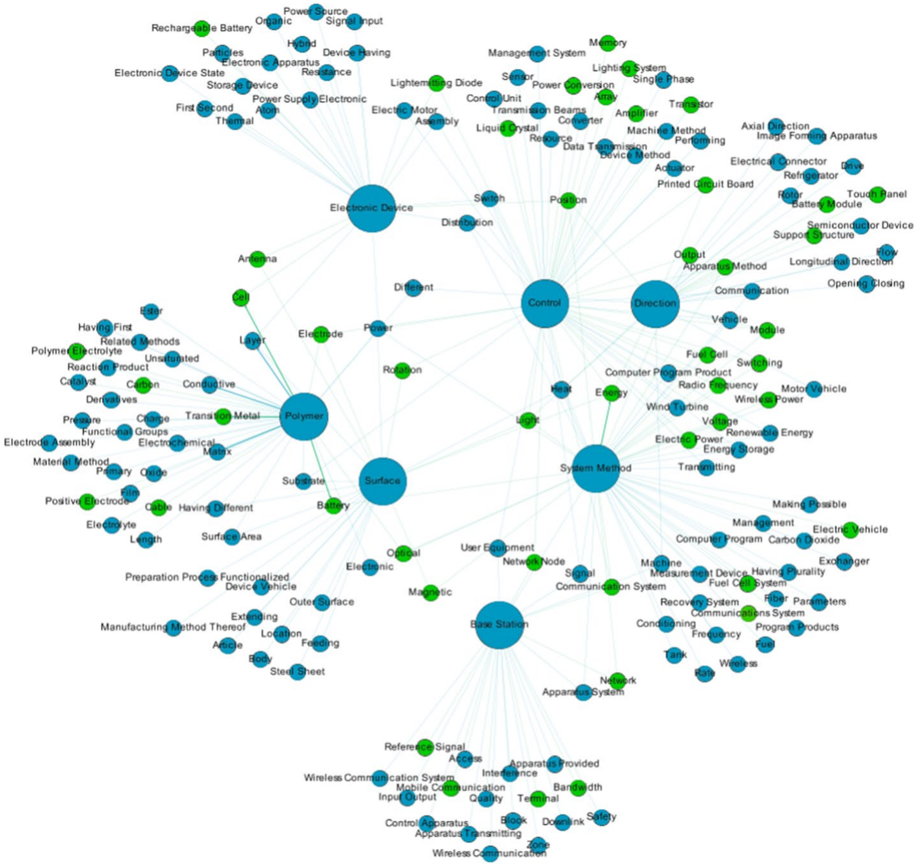


**Fig. 9** Network visualization of clusters affiliated to IPC classes

Applying the ML model for detecting the SDG 7 categories in 132 k patent sample results in almost 10 thousand patent families, which have the probability of above 85% for SDG 7. Lowering the probability percentage to over 70% results in 15,500 patent families. Cross-referencing (1) SDG7 query-driven records, (2) ML model-driven records for SDG7 at a probability level 85% or above, we used the International Patent Classification (IPC) to check the number of patent families included in each patent class. Figure 6 illustrates the difference in the number of patent families in top similar IPC classes between lexical query and ML detected patents.

The similar IPC classes between both sets go over 175 (out of 380 classes). While the proportional volume of patents in each IPC class differs between the two sets, the proportional difference in similar classes will not exceed more than 4%. This analysis suggests that both lexical query resulted patent families and ML decided patent families for SDG7 overlap overall, but there is some difference of coverage in terms of IPC classes as well.

Figure 7 illustrates the distribution of patent families over IPC classes for the ML model.

The ML model expanded the scope of IPC classes by 14%, where 75 new classes were identified to be relevant patent families for SDG7. On the contrary, 68 (14%) classes identified with the lexical search query were not identified by the ML model top 30% results

(Fig. 8 is a visual illustration of top IPC classes). However, our further checks on the ML results for 50% to 70% probability levels highlighted that most of the IPC classes are covered in that cohort. Relaxing the requirement on which a patent is associated with a specific SDG seems to have an impact on the results.

For learning the thematic of the ML model-driven patent families for SDG 7, Patbase keyword landscaping analysis, which uses the most commonly occurring keywords (found within titles and abstracts), was used. In contrast to IPC class comparison, this view could add an angle on the thematical coverage of the patent families with the families clustered by either the relevant keywords or the concepts. We shortlisted the keywords based on the occurrence of both SDG 7 lexical query-driven results and ML model detected results. The keywords' occurrences are in a two-level hierarchy, so we picked the seven top clusters and their affiliated sub-cluster for visualization purposes. Figure 9 illustrates the keyword clusters in a network visualization with nodes as clusters and edges as subclusters to the higher-level clusters.

The colour of the new keyword clusters identified by the ML model is marked as green while the overlapping clusters within two modes of SDG 7 identification patent families remained blue. It is noticeable that the ML model was able to detect patent families that are thematically aligned within the first level cluster with the lexical query-based method. However, at the same time, new sub-clusters were introduced. For example, from the figure, we can identify technology clusters like a rechargeable battery in sub class electronic devices that ML model has identified as SDG7 related. This identification is an extension for SDG 7, which has been learned from SDG 7 related publications and is now able to be transformed and performed on different artefacts (in our case, patents). The data file offers the patent families identified as SDG7 category based on our ML method with a probability of over 80%. In addition, we have included the keyword clusters generated from the patents IPC classes to give more information on the content of the patents and furthermore to be used as a lexical query for keyword searches.

## Discussion and conclusion

The United Nations Sustainable Development Goals (SDGs) offer a framework to achieve a better and more sustainable future. Governments worldwide have already agreed to these goals. Now the emphasis is to take action on a global as well as on local levels. Scientific and technological innovations are necessary but enabling them to make an impact requires an understanding of their utility to the sustainable positive economy. Our study contributes to systematically comprehending sustainability-oriented science, technology and innovation. It offers a systematic identification approach for sustainable development goals, requirements and objectives. Second, based on the publications with the highest relatedness to SDGs, the study trains and creates a machine learning model to detect the relatedness of scientific publications and, finally, uses the approach to analyse patent documents. This approach extends previous work on novel ways of identifying science and technology linkages (Ranaei et al., 2017) by overlaying the SDG context. While we can question the linearity of innovation (Suominen & Seppänen, 2014), understanding the co-evolution of science and technology is vital for the economy and society overall (Pavitt, 1991). Development intersects with IP policies as creativity and innovation are either fostered or frustrated by an economy's chosen development policy. Therefore, including consideration of the SDGs in IP policy could lead to more significant and more lasting success.

Based on the UN's SDG definition, we queried for relevant publications which address the 16 SDGs. This approach resulted in a vocabulary to capture the breadth and depth of SDGs within scientific publications. The comprehensive taxonomy created for SDG identification was used to create labelled data to train a machine learning model. Benchmarking with five classification methods resulted in identifying Word2vec with Logistic Regression as the top performer in the multiclassification task. The performance metrics were satisfactory for ten SDG categories. The highest performing model was then extended to unseen documents, in this case patents. The patents were assigned a SDG relevance at different levels of probability. In order to test the validity of the ML model in detecting SDG contribution in patents, the classification was compared to the already existing lexical queries using patent classifications. The analysis illustrated that the ML model increased the identification of SDG-oriented patent families in the year 2020 for EPO patents by 14.5% on the IPC class level.

Our study showed the application of machine learning for expanding lexical query and information retrieval. In response to such complex and multidimensional topics as SDGs, comprehending their definitions of breadth and depth often becomes a challenge. Importantly not all scientific and technological artefacts share the same features, and for example, patent documents, due to their technical nature, do not directly correspond to an SDG definition. While the linkage between science and technology and publication and patents is evident, we could utilize this connection and expand the identification of SDG oriented artefacts from publications to patents.

The implication of our study is multi-faceted. Moving towards further contribution aligned with SDGs requires an accurate understanding of historical STI efforts. The opportunities to identify STI artefacts towards SDGs could give an accurate image of progressed, challenged and undermined domains. Therefore, the implicit implications of the study can provide an overview of the STI contributions to SDGs on a macro level so far. On a micro level, it guides companies on how they can align their strategies as well as measure and manage their contribution to the realization of the SDGs when it comes to IP strategies.

The policy relevance of our results comes from the call for measurement frameworks beyond productivity and gross domestic product (Hayden, 2021; Malay, 2021; Schreyer, 2021). There have been calls from policymakers to create a measurement framework enabling analysis of the sustainable transition. The ability of currently available Beyond GDP measures seem to have low potential in capturing the transformation (Malay, 2019). Malay (2021) particularly called for the creation of cohesive measurement frameworks built around a shared goal. SDG is a good representation of this type of shared goal. A good practical example of the policy relevance of the Beyond GDP agenda is the European Commission initiative to develop indicators that would be clear, transformative and more inclusive to environmental and social aspects.[10] This has led, for example, for the European Commission to call for the development of a Beyond GDP measurement framework.[11]

There are of course limitations to this research. First, we have analysed an exhaustive set of publication labelled data and the applicability to transform it via an ML model to detect patent data. Despite this exercise, the results may differ for other types of text sources, languages, and classification objectives. For example, the OECD's algorithm for detecting SDG relevancy in text has been trained with 6,000 labelled descriptions of firms' actions (OECD, 2021b). The difference in performance of their SDG identification algorithm estimation is noticeable as the training data nature is different compared to our work. Second, our work

[10] Refer to the European Commission website on the Beyond GDP initiative https://ec.europa.eu/environment/beyond_gdp/index_en.html.

[11] Refer to the call for proposals at the European Commission website https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl2-2022-transformations-01-01.

departed from the patent and publication merging approach by (Ranaei et al., 2017), which also framed the approach in our study. This being said, employing algorithms and methods beyond the ones we study such as deep learning approaches could have offered an additional vantage point. However, we believe we cover a representative set of both machine learning and lexical methods and demonstrated their functionality in our classification task.

Third, we also follow prior comparisons by applying standard procedures in terms of preprocessing and document representation. There are many ways of extending this to the specific task at hand, which can improve performance. Overall, our results can be viewed as a lower performance boundary which further emphasizes the potential of automated text classification in discovering STI and SDG connections. Still, an even more extensive optimization may produce different results. We hope these findings make sound automated text classification more approachable to STI researchers and encourage future research to integrate novel approaches for complex classification tasks such as SDGs.

## Appendix 1

### Query in Scopus: Results, 322,088

TITLE-ABS-KEY ( ( {energy efficiency} OR {energy consumption} OR {energy transition} OR {clean energy technology} OR {energy equity} OR {energy justice} OR {energy poverty} OR {energy policy} OR renewable* OR {2000 Watt society} OR {smart micro-grid} OR {smart grid} OR {smart microgrid} OR {smart micro-grids} OR {smart grids} OR {smart microgrids} OR {smart meter} OR {smart meters} OR {affordable electricity} OR {electricity consumption} OR {reliable electricity} OR {clean fuel} OR {clean cooking fuel} OR {fuel poverty} OR energiewende OR {life-cycle assessment} OR {life cycle assessment} OR {life-cycle assessments} OR {life cycle assessments} OR ( {photochemistry} AND {renewable energy}) OR photovoltaic OR {photocatalytic water splitting} OR {hydrogen production} OR {water splitting} OR {lithium-ion batteries} OR {lithium-ion battery} OR {heat network} OR {district heat} OR {district heating} OR {residential energy consumption} OR {domestic energy consumption} OR {energy security} OR {rural electrification} OR {energy ladder} OR {energy access} OR {energy conservation} OR {low-carbon society} OR {hybrid renewable energy system} OR {hybrid renewable energy systems} OR {fuel switching} OR ( {foreign development aid} AND {renewable energy}) OR {energy governance} OR ( {official development assistance} AND {electricity}) OR ( {energy development} AND {developing countries})) AND NOT ( {wireless sensor network} OR {wireless sensor networks})) AND PUBYEAR < 2018 AND PUBYEAR > 2012.

### Query in Patbase: For year 2020, Granted, Alive, is 347

‘energy efficiency’ OR’energy consumption’ OR’energy transition’ OR’clean energy technology’ OR’energy equity’ OR’energy justice’ OR’energy poverty’ OR’energy policy’ OR renewable* OR’2000 Watt society’ OR’smart micro-grid’ OR’smart grid’ OR’smart microgrid’ OR’smart micro-grids’ OR’smart grids’ OR’smart microgrids’ OR’smart meter’ OR’smart meters’ OR’affordable electricity’ OR’electricity consumption’ OR’reliable electricity’ OR’clean fuel’ OR’clean cooking fuel’ OR’fuel poverty’ OR energiewende OR’life-cycle assessment’ OR’life cycle assessment’ OR’life-cycle assessments’ OR’life cycle assessments’

OR ('photochemistry' AND'renewable energy') OR photovoltaic OR'photocatalytic water splitting' OR'hydrogen production' OR'water splitting' OR'lithium-ion batteries' OR'lithium-ion battery' OR'heat network' OR'district heat' OR'district heating' OR'residential energy consumption' OR'domestic energy consumption' OR'energy security' OR'rural electrification' OR'energy ladder' OR'energy access' OR'energy conservation' OR'low-carbon society' OR'hybrid renewable energy system' OR'hybrid renewable energy systems' OR'fuel switching' OR ('foreign development aid' AND'renewable energy') OR'energy governance' OR ('official development assistance' AND'electricity') OR ('energy development' AND'developing countries') AND NOT ('wireless sensor network' OR'wireless sensor networks').

## Appendix 2: Patent data descriptive

See Figs. 10, 11 and Table 2.

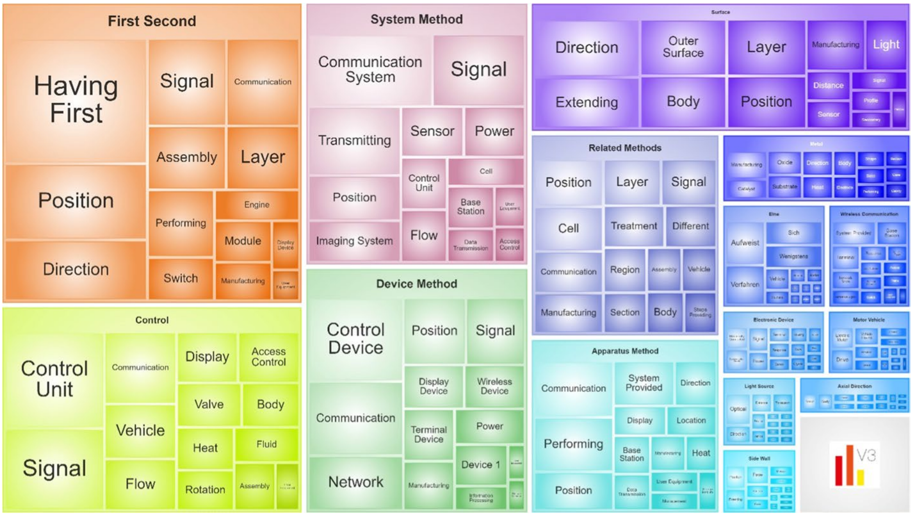


**Fig. 10** Patents top keyword distribution

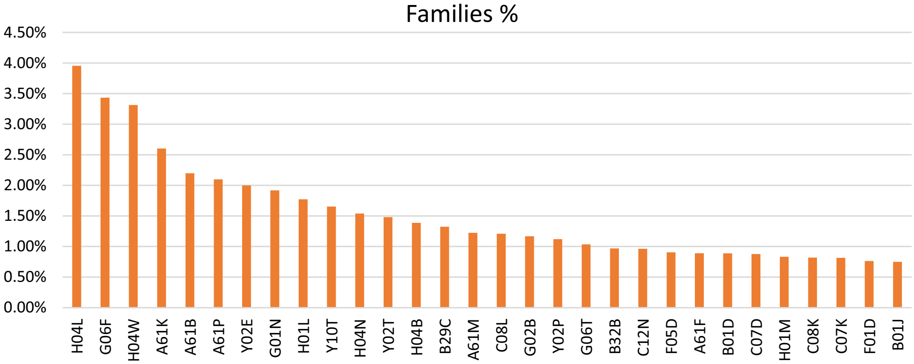
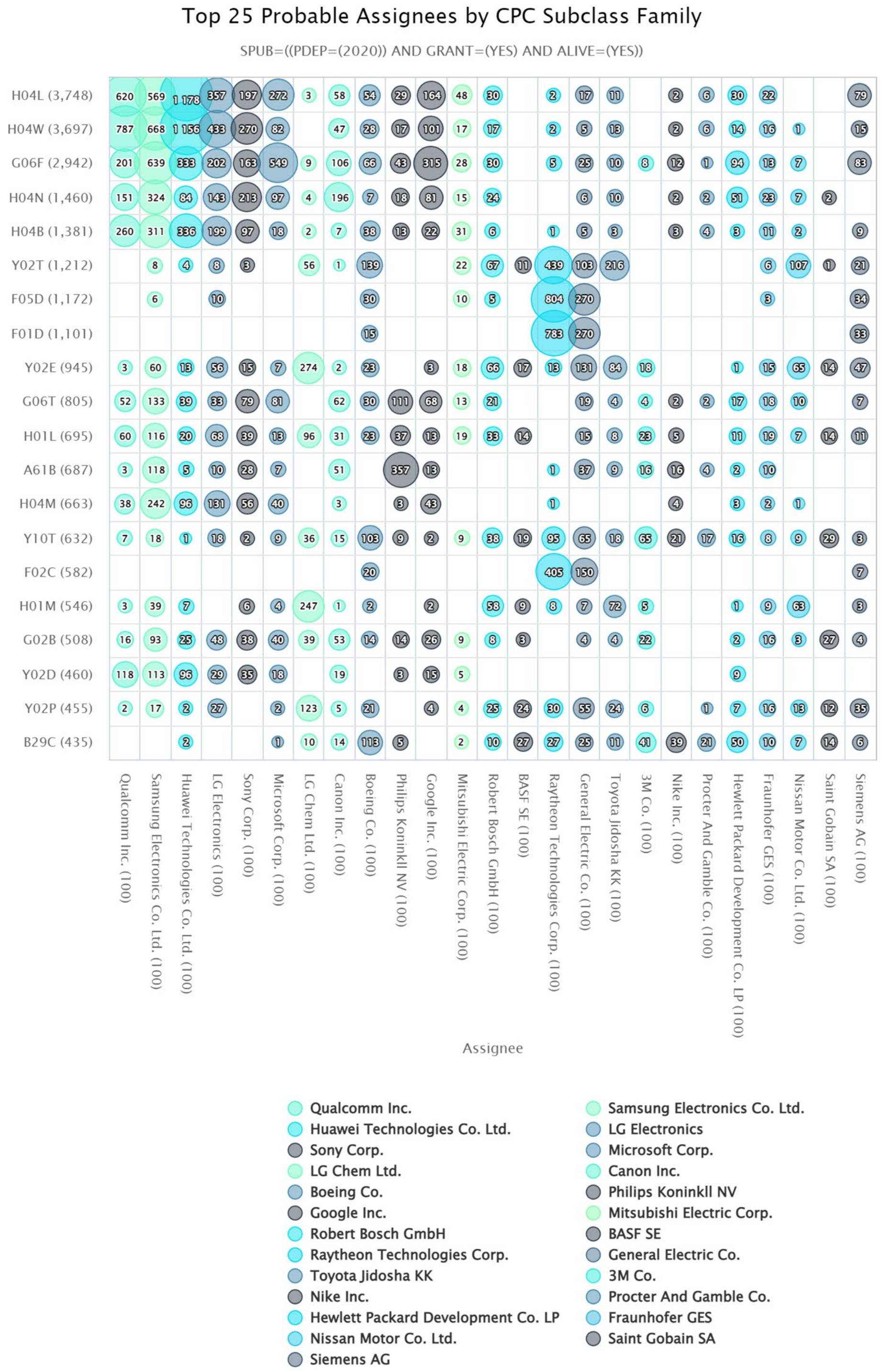


**Fig. 11** Patent family distribution on CPC classes

**Table 2** Patent topical clusters distribution by family number count

| Cluster | Families | Cluster | Families | Cluster | Families |
|---|---|---|---|---|---|
| Communication system | 131 | Cell module | 13 | Web application | 8 |
| Heat resistance | 59 | Blood vessel | 13 | Injection molding | 7 |
| Display device | 48 | Fuel cell system | 13 | Haptic feedback | 7 |
| Combustion engine | 43 | Battery module | 13 | Heat exchanger | 7 |
| Protein kinase | 36 | Heat sealing | 13 | Golf ball | 7 |
| Light source | 31 | Dosage form | 12 | Polyester film | 7 |
| Turbine section | 27 | Steel sheet | 12 | Touch input | 7 |
| Combination therapy | 23 | Nonwoven fabric | 11 | Colorectal cancer | 7 |
| Lithium ion | 21 | Stratum corneum | 11 | Acid residue | 7 |
| Additive manufacturing | 19 | Antigen binding fragment | 11 | Power consumption | 7 |
| Wind energy | 19 | Transdermal delivery | 11 | Injection device | 7 |
| Circuit board | 18 | Atrial fibrillation | 11 | Density polyethylene | 7 |
| Shelf LIFE | 18 | Mesh network | 10 | Energy consumption | 6 |
| Blood vessel formation | 18 | Nucleotide polymorphisms Snp | 9 | Access control | 6 |
| Communication device | 17 | Power Source | 9 | Parenteral administration | 6 |
| Power supply | 17 | Projection Data | 9 | Blood pressure | 6 |
| Gesture recognition | 16 | Water vapor | 8 | Principles briefly | 6 |
| Expression level | 15 | Image quality | 8 | Computer data system | 6 |
| Semiconductor device | 15 | Carbon dioxide | 8 | Video coding | 6 |
| Access point (AP) | 14 | Turbine engine | 8 | Speech recognition | 6 |
| Ground state | 14 | Light chain | 8 | Electromagnetic signal | 6 |



Top 25 Probable Assignees by CPC Subclass Family

SPUB=((PDEP=(2020)) AND GRANT=(YES) AND ALIVE=(YES))

| | Qualcomm Inc. (100) | Samsung Electronics Co. Ltd. (100) | Huawei Technologies Co. Ltd. (100) | LG Electronics (100) | Sony Corp. (100) | Microsoft Corp. (100) | LG Chem Ltd. (100) | Canon Inc. (100) | Boeing Co. (100) | Philips Koninkll NV (100) | Google Inc. (100) | Mitsubishi Electric Corp. (100) | Robert Bosch GmbH (100) | BASF SE (100) | Raytheon Technologies Corp. (100) | General Electric Co. (100) | Toyota Jidosha KK (100) | 3M Co. (100) | Nike Inc. (100) | Procter And Gamble Co. (100) | Hewlett Packard Development Co. LP (100) | Fraunhofer GES (100) | Nissan Motor Co. Ltd. (100) | Saint Gobain SA (100) | Siemens AG (100) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H04L (3,748) | 620 | 569 | 1 178 | 357 | 197 | 272 | 3 | 58 | 54 | 29 | 164 | 48 | 30 | | 2 | 17 | 11 | | 2 | 6 | 30 | 22 | | | 79 |
| H04W (3,697) | 787 | 668 | 1 156 | 433 | 270 | 82 | | 47 | 28 | 17 | 101 | 17 | 17 | | 2 | 5 | 13 | | 2 | 6 | 14 | 16 | 1 | | 15 |
| G06F (2,942) | 201 | 639 | 333 | 202 | 163 | 549 | 9 | 106 | 66 | 43 | 315 | 28 | 30 | | 5 | 25 | 10 | 8 | 12 | 1 | 94 | 13 | 7 | | 83 |
| H04N (1,460) | 151 | 324 | 84 | 143 | 213 | 97 | 4 | 196 | 7 | 18 | 81 | 15 | 24 | | | 6 | 10 | | 2 | 2 | 51 | 23 | 7 | 2 | |
| H04B (1,381) | 260 | 311 | 336 | 199 | 97 | 18 | 2 | 7 | 38 | 13 | 22 | 31 | 6 | | 1 | 5 | 3 | | 3 | 4 | 3 | 11 | 2 | | 9 |
| Y02T (1,212) | | 8 | 4 | 8 | 3 | | 56 | 1 | 139 | | | 22 | 67 | 11 | 439 | 103 | 216 | | | | | 6 | 107 | 1 | 21 |
| F05D (1,172) | | 6 | | 10 | | | | | 30 | | | 10 | 5 | | 804 | 270 | | | | | | 3 | | | 34 |
| F01D (1,101) | | | | | | | | | 15 | | | | | | 783 | 270 | | | | | | | | | 33 |
| Y02E (945) | 3 | 60 | 13 | 56 | 15 | 7 | 274 | 2 | 23 | | 3 | 18 | 66 | 17 | 13 | 131 | 84 | 18 | | | 1 | 15 | 65 | 14 | 47 |
| G06T (805) | 52 | 133 | 39 | 33 | 79 | 81 | | 62 | 30 | 111 | 68 | 13 | 21 | | | 19 | 4 | 4 | 2 | 2 | 17 | 18 | 10 | | 7 |
| H01L (695) | 60 | 116 | 20 | 68 | 39 | 13 | 96 | 31 | 23 | 37 | 13 | 19 | 33 | 14 | | 15 | 8 | 23 | 5 | | 11 | 19 | 7 | 14 | 11 |
| A61B (687) | 3 | 118 | 5 | 10 | 28 | 7 | | 51 | | 357 | 13 | | | | 1 | 37 | 9 | 16 | 16 | 4 | 2 | 10 | | | |
| H04M (663) | 38 | 242 | 96 | 131 | 56 | 40 | | 3 | | 3 | 43 | | | | 1 | | | | 4 | | 3 | 2 | 1 | | |
| Y10T (632) | 7 | 18 | 1 | 18 | 2 | 9 | 36 | 15 | 103 | 9 | 2 | 9 | 38 | 19 | 95 | 65 | 18 | 65 | 21 | 17 | 16 | 8 | 9 | 29 | 3 |
| F02C (582) | | | | | | | | | 20 | | | | | | 405 | 150 | | | | | | | | | 7 |
| H01M (546) | 3 | 39 | 7 | | 6 | 4 | 247 | 1 | 2 | | 2 | | 58 | 9 | 8 | 7 | 72 | 5 | | | 1 | 9 | 63 | | 3 |
| G02B (508) | 16 | 93 | 25 | 48 | 38 | 40 | 39 | 53 | 14 | 14 | 26 | 9 | 8 | 3 | | 4 | 4 | 22 | | | 2 | 16 | 3 | 27 | 4 |
| Y02D (460) | 118 | 113 | 96 | 29 | 35 | 18 | | 19 | | 3 | 15 | 5 | | | | | | | | | 9 | | | | |
| Y02P (455) | 2 | 17 | 2 | 27 | | 2 | 123 | 5 | 21 | | 4 | 4 | 25 | 24 | 30 | 55 | 24 | 6 | | 1 | 7 | 16 | 13 | 12 | 35 |
| B29C (435) | | | 2 | | | 1 | 10 | 14 | 113 | 5 | | 2 | 10 | 27 | 27 | 25 | 11 | 41 | 39 | 21 | 50 | 10 | 7 | 14 | 6 |

Assignee

- ● Qualcomm Inc.
- ● Samsung Electronics Co. Ltd.
- ● Huawei Technologies Co. Ltd.
- ● LG Electronics
- ● Sony Corp.
- ● Microsoft Corp.
- ● LG Chem Ltd.
- ● Canon Inc.
- ● Boeing Co.
- ● Philips Koninkll NV
- ● Google Inc.
- ● Mitsubishi Electric Corp.
- ● Robert Bosch GmbH
- ● BASF SE
- ● Raytheon Technologies Corp.
- ● General Electric Co.
- ● Toyota Jidosha KK
- ● 3M Co.
- ● Nike Inc.
- ● Procter And Gamble Co.
- ● Hewlett Packard Development Co. LP
- ● Fraunhofer GES
- ● Nissan Motor Co. Ltd.
- ● Saint Gobain SA
- ● Siemens AG

## Appendix 3: Publication data description

See Tables 3 and 4.

**Table 3** Top publication's subject area and Publication's country

| Subject area | Documents (%) | Country/territory | Documents (%) |
|---|---|---|---|
| Environmental science | 15.25 | United States | 23.54 |
| Social sciences | 15.21 | United Kingdom | 9.50 |
| Medicine | 8.83 | China | 6.39 |
| Agricultural and biological sciences | 5.97 | Germany | 4.81 |
| Engineering | 5.90 | Australia | 4.60 |
| Business, management and accounting | 5.19 | Canada | 4.36 |
| Economics, econometrics and finance | 5.11 | Netherlands | 3.64 |
| Earth and planetary sciences | 4.35 | France | 3.19 |
| Energy | 4.17 | Spain | 2.95 |
| Chemistry | 4.05 | Italy | 2.84 |
| Multidisciplinary | 3.87 | Switzerland | 2.31 |
| Biochemistry, genetics and molecular biology | 3.55 | Sweden | 2.22 |
| Psychology | 3.14 | Japan | 1.61 |
| Materials science | 2.34 | Belgium | 1.46 |
| Chemical engineering | 2.30 | Denmark | 1.46 |
| Computer science | 2.25 | Norway | 1.37 |
| Arts and humanities | 2.24 | India | 1.28 |
| Physics and astronomy | 1.61 | South Korea | 1.10 |
| Decision sciences | 1.00 | Austria | 1.05 |
| Immunology and microbiology | 0.96 | Finland | 0.96 |
| Mathematics | 0.76 | South Africa | 0.90 |
| Nursing | 0.58 | Brazil | 0.90 |
| Neuroscience | 0.56 | Hong Kong | 0.85 |
| Pharmacology, toxicology and pharmaceutics | 0.46 | Singapore | 0.77 |
| Health professions | 0.26 | New Zealand | 0.77 |
| Veterinary | 0.06 | Taiwan | 0.69 |
| Dentistry | 0.04 | Portugal | 0.63 |

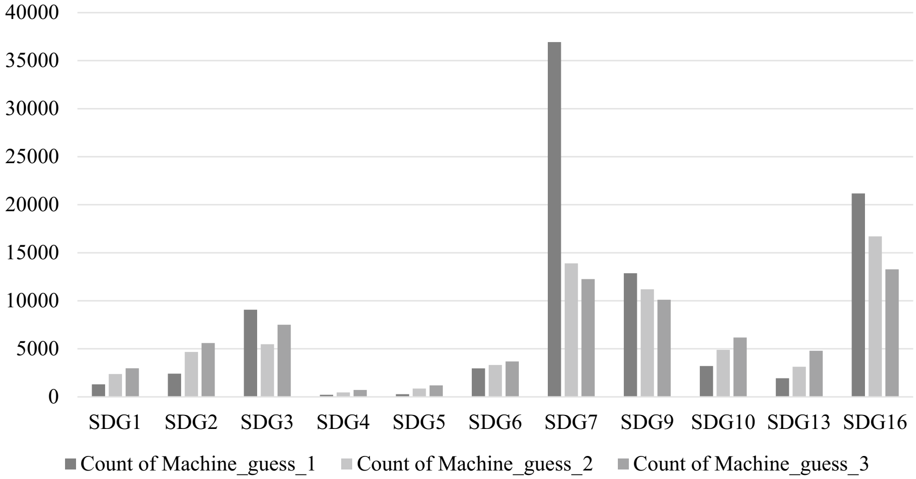
**Table 4** Top publication funders and Publication affiliations

| Funding sponsor | Documents (%) | Affiliation | Documents (%) |
|---|---|---|---|
| National science foundation | 7.88 | Chinese Academy of Sciences | 2.83 |
| National Natural Science Foundation of China | 6.00 | University of Washington | 1.74 |
| Seventh Framework Programme | 4.99 | Stanford University | 1.66 |
| European Commission | 3.63 | University of Oxford | 1.56 |
| National Institutes of Health | 3.42 | University of California. Berkeley | 1.35 |
| Economic and Social Research Council | 2.83 | Massachusetts Institute of Technology | 1.26 |
| Natural Environment Research Council | 2.65 | Harvard Medical School | 1.26 |
| National Cancer Institute | 2.56 | Columbia University | 1.24 |
| UK Research and Innovation | 2.49 | Harvard University | 1.23 |
| Eunice Kennedy Shriver National Institute of Child Health and Human Development | 1.97 | CNRS Centre National de la Recherche Scientifique | 1.18 |
| National Institute of Mental Health | 1.95 | Wageningen University & Research | 1.17 |
| Medical Research Council | 1.79 | University of Toronto | 1.14 |
| Japan Society for the Promotion of Science | 1.41 | University of Cambridge | 1.12 |
| U.S. Department of Energy | 1.39 | University College London | 1.08 |
| Engineering and Physical Sciences Research Council | 1.37 | The University of Queensland | 1.06 |
| Ministry of Science and Technology of the People's Republic of China | 1.27 | University of Michigan, Ann Arbor | 1.04 |
| Australian Research Council | 1.25 | University of California, Los Angeles | 1.04 |
| National Heart, Lung, and Blood Institute | 1.25 | The University of British Columbia | 1.01 |
| Ministry of Education of the People's Republic of China | 1.21 | The University of North Carolina at Chapel Hill | 0.97 |
| Chinese Academy of Sciences | 1.13 | University of California, San Diego | 0.95 |
| National Institute on Drug Abuse | 1.13 | University of Minnesota Twin Cities | 0.93 |
| National Institute of Diabetes and Digestive and Kidney Diseases | 1.03 | Yale University | 0.93 |
| Deutsche Forschungsgemeinschaft | 1.03 | University of Melbourne | 0.89 |
| National Institute of General Medical Sciences | 0.99 | University of Pennsylvania | 0.88 |
| National Institute on Aging | 0.96 | ETH Zürich | 0.87 |
| National Institute of Allergy and Infectious Diseases | 0.89 | University of California, San Francisco | 0.86 |

**Table 4** (continued)

| Funding sponsor | Documents (%) | Affiliation | Documents (%) |
|---|---|---|---|
| National Center for Research Resources | 0.87 | University of Colorado Boulder | 0.82 |

## Appendix 4: ML top three guesses of patents affiliation to SDG categories



**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11192-022-04358-x.

**Acknowledgements** This work is supported by Business Finland under the project "Mapping Sustainable Development Activity; Its Evolution and Impact in Science, Technology, Innovation and Businesses (INNOSDG)" and VTT Technical Research Centre of Finland project 132376. The earlier version of this paper was presented in the AII2021 workshop.

**Funding** Open Access funding provided by Technical Research Centre of Finland (VTT).

## References

Altgilbers, N., Walter, Lo., & Moehrle, M. G. (2020). Frugal invention candidates as antecedents of frugal patents—The role of frugal attributes analysed in the medical engineering technology. *International Journal of Innovation Management*. https://doi.org/10.1142/S1363919620500826

Ashford, N. A., & Hall, R. P. (2011). The importance of regulation-induced innovation for sustainable development. *Sustainability, 3*(1), 270–292. https://doi.org/10.3390/su3010270

Awad, M., & Khanna, R. (2015). Support Vector Machines for Classification. *Efficient Learning Machines*. https://doi.org/10.1007/978-1-4302-5990-9_3

Bonilla, S. H., Silva, H. R. O., da Silva, M. T., Gonçalves, R. F., & Sacomano, J. B. (2018). Industry 4.0 and sustainability implications: A scenario-based analysis of the impacts and challenges. *Sustainability (Switzerland)*. https://doi.org/10.3390/su10103740

Callaert, J., Vervenne, J.-B., Looy, B., Magerman, T., Song, X., & Jeuris, W. (2014). Patterns of science-technology linkage. *Directorate-General for Research and Innovation (European Commission)*. https://doi.org/10.2777/55249

Campbell, R. S. (1983). Patent trends as a technological forecasting tool. *World Patent Information, 5*(3), 137–143. https://doi.org/10.1016/0172-2190(83)90134-5

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Mach Learn, 20*, 273–297. https://doi.org/10.1007/BF00994018

Chen, Y. S., & Chang, K. C. (2010). Analyzing the nonlinear effects of firm size, profitability, and employee productivity on patent citations of the US pharmaceutical companies by using artificial neural network. *Scientometrics, 82*(1), 75–82. https://doi.org/10.1007/s11192-009-0034-x

Cramer, J. S. (2005). The origins of logistic regression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.360300

Daejeon Declaration. (2015). *Daejeon Declaration on Science, Technology and Innovation Policies for the Global and Digital Age and Innovation Policies for the Global and Digital Age*. https://legalinstruments.oecd.org/Instruments/PrintInstrumentView.aspx?InstrumentID=335&amp;InstrumentPID=389&InstrumentHID=0&amp;Lang=en

Demeester, T., Sutskever, I., Chen, K., Dean, J., & Corado, G. (2016). Distributed Representations of Words and Phrases and their Compositionality. In *EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1389–1399

Dosi, G., Llerena, P., & Labini, M. S. (2006). The relationships between science, technologies and their industrial exploitation: An illustration through the myths and realities of the so-called "European Paradox." *Research Policy, 35*(10), 1450–1464. https://doi.org/10.1016/j.respol.2006.09.012

Elsevier. (2015). *Sustainability Science in a Global Landscape*. https://www.elsevier.com/__data/assets/pdf_file/0018/119061/SustainabilityScienceReport-Web.pdf

Ena, O. (2021). 'Domain-specific' patent analytics: Focus on company's technology priorities. *World Patent Information*. https://doi.org/10.1016/j.wpi.2021.102037

Figueroa, R. L., & Zeng-Treitler, Q. (2013). Text classification performance: Is the sample size the only factor to be considered? *Studies in Health Technology and Informatics, 192*(1–2), 1193. https://doi.org/10.3233/978-1-61499-289-9-1193

Freeman, C. (2004). Technological infrastructure and international competitiveness. *Industrial and Corporate Change, 13*(3), 541–569. https://doi.org/10.1093/icc/13.3.541

Fukuda, K. (2020). Science, technology and innovation ecosystem transformation toward society 5.0. *International Journal of Production Economics*. https://doi.org/10.1016/j.ijpe.2019.07.033

Gelles, D., & Yaffe-Bellany, D. (2019). Shareholder Value Is No Longer Everything, Top C.E.O.s Say. *The New York Times*. Retrieved May 20, 2020, from https://www.nytimes.com/2019/08/19/business/business-roundtable-ceos-corporations.html

Giovannini, E., Niestroy, I., Nilsson, M., Roure, F., & Spanos, M. (2015). The role of science, technology and innovation policies to foster the implementation of the sustainable development goals (SDGs) report of the expert group " follow-up to Rio + 20, notably the SDGs." *European Commission*. https://doi.org/10.2777/485757

Goldberg, Y., & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. http://arxiv.org/abs/1402.3722

Goos, M., Konings, J., & Vandeweyer, M. (2015). Employment growth in Europe: The roles of innovation, local job multipliers and institutions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2671765

Hajer, M., Nilsson, M., Raworth, K., Bakker, P., Berkhout, F., de Boer, Y., et al. (2015). Beyond cockpit-ism: Four insights to enhance the transformative potential of the sustainable development goals. *Sustainability (switzerland), 7*(2), 1651–1660. https://doi.org/10.3390/su7021651

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing, 36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

Hayden, A. (2021). From fantasy to transformation: Steps in the policy use of "Beyond-GDP" indicators. *The Well-Being Transition*. https://doi.org/10.1007/978-3-030-67860-9_7

Ingo, S., & Andreas, C. (2008). *Support vector machines*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-77242-4

Jia, Q., Wei, L., & Li, X. (2019). Visualizing sustainability research in business and management (1990–2019) and emerging topics: A large-scale bibliometric analysis. *Sustainability (Switzerland)*. https://doi.org/10.3390/su11205596

Kahn, K. B. (2018). Understanding innovation. *Business Horizons, 61*(3), 453–460. https://doi.org/10.1016/j.bushor.2018.01.011

Kenekayoro, P. (2018). Identifying named entities in academic biographies with supervised learning. *Scientometrics, 116*(2), 751–765. https://doi.org/10.1007/s11192-018-2797-4

Kenekayoro, P., Buckley, K., & Thelwall, M. (2015). Clustering research group website homepages. *Scientometrics, 102*(3), 2023–2039. https://doi.org/10.1007/s11192-014-1497-y

Klomp, L., & Van Leeuwen, G. (2001). Linking innovation and firm performance: A new approach. *International Journal of the Economics of Business, 8*(3), 343–364. https://doi.org/10.1080/13571510110079612

Kreuchauff, F., & Korzinov, V. (2017). A patent search strategy based on machine learning for the emerging field of service robotics. *Scientometrics, 111*(2), 743–772. https://doi.org/10.1007/s11192-017-2268-3

Leach, M., Rockström, J., Raskin, P., Scoones, I., Stirling, A. C., Smith, A., et al. (2012). Transforming innovation for sustainability. *Ecology and Society*. https://doi.org/10.5751/ES-04933-170211

Lee, C., & Lee, G. (2019). Technology opportunity analysis based on recombinant search: Patent landscape analysis for idea generation. *Scientometrics, 121*(2), 603–632. https://doi.org/10.1007/s11192-019-03224-7

Lu, Y., Xiong, X., Zhang, W., Liu, J., & Zhao, R. (2020). Research on classification and similarity of patent citation based on deep learning. *Scientometrics, 123*(2), 813–839. https://doi.org/10.1007/s11192-020-03385-w

Malay, O. E. (2019). Do Beyond GDP indicators initiated by powerful stakeholders have a transformative potential? *Ecological Economics, 162*, 100–107. https://doi.org/10.1016/j.ecolecon.2019.04.023

Malay, O. E. (2021). How to articulate beyond GDP and businesses' social and environmental indicators? *Social Indicators Research*. https://doi.org/10.1007/s11205-020-02583-6

Manning, C. D., Raghavan, P., & Schutze, H. (2012). Scoring, term weighting, and the vector space model. *Introduction to Information Retrieval*. https://doi.org/10.1017/cbo9780511809071.007

Mazzucato, M. (2011). The entrepreneurial state. *Soundings, 49*(49), 131–142. https://doi.org/10.3898/136266211798411183

Migotto, M., & Haščič, I. (2015). Measuring environmental innovation using patent data. *OECD Environment Working Papers, 89*(89), 1–59. https://doi.org/10.1787/5js009kf48xw-en

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013—Workshop Track Proceedings*

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications, 106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Mistele, T., Price, T., & Hossenfelder, S. (2019). Predicting authors' citation counts and h-indices with a neural network. *Scientometrics, 120*(1), 87–104. https://doi.org/10.1007/s11192-019-03110-2

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Nations, U. (2021). Science, Technology and Innovation for the SDGs. *United Nations*. https://www.un.org/development/desa/indigenouspeoples/science-technology-and-innovation-for-the-sdgs.html

Nedjah, N., Mourelle, L. D. M., Kacprzyk, J., Fran, F. M. G., Tsihrintzis, V. G. a, Virvou, M., & Howlett, R. J. (2009). *Intelligent Text Categorization and Clustering Studies in Computational Intelligence, Volume 164*. *Pattern Recognition*. Springer; Softcover reprint of hardcover 1st ed. 2009 edition (October 28, 2010)

Nelson, R. R., & Sidney, G. (1982). *An Evolutionary Theory of Economic Change*. Harvard Business School Press

Nieminen, M., & Hyytinen, K. (2015). Future-oriented impact assessment: Supporting strategic decision-making in complex socio-technical environments. *Evaluation, 21*(4), 448–461. https://doi.org/10.1177/1356389015606540

OECD. (2021a). *Industrial Policy for the Sustainable Development Goals Increasing the Private Sector's Contribution: Increasing the Private Sector's Contribution*. OECD Publishing

OECD. (2021b). *Industrial policy for the Sustainable Development Goals: How to increase the private sector's contribution to the SDGs*. *Forthcoming*. https://one.oecd.org/document/DSTI/CIIE(2021b)10/en/pdf

Pavitt, K. (1991). What makes basic research economically useful? *Research Policy, 20*(2), 109–119. https://doi.org/10.1016/0048-7333(91)90074-Z

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing, 4*, 5. https://doi.org/10.22364/bjmc.2017.5.2.05

Ranaei, S., Suominen, A., & Dedehayir, O. (2017). A topic model analysis of science and technology linkages: A case study in pharmaceutical industry. In *2017 IEEE Technology and Engineering Management Society Conference, TEMSCON 2017*. https://doi.org/10.1109/TEMSCON.2017.7998353

Ranaei, S., Suominen, A., Porter, A., & Kässi, T. (2019). Application of text-analytics in quantitative study of science and technology. In *Springer Handbooks* (pp. 957–982). https://doi.org/10.1007/978-3-030-02511-3_39

Ranaei, S., Suominen, A., Porter, A., & Carley, S. (2020). Evaluating technological emergence using text analytics: Two case technologies and three approaches. *Scientometrics, 122*(1), 215–247. https://doi.org/10.1007/s11192-019-03275-w

Rong, X. (2014). word2vec Parameter Learning Explained. http://arxiv.org/abs/1411.2738

Ruijie, Z., Ying, X., Shuaichen, J., & Yonghe, L. (2021). Patent text modeling strategy and its classification based on structural features. *World Patent Information*. https://doi.org/10.1016/j.wpi.2021.102084

Scheyvens, R., Banks, G., & Hughes, E. (2016). The private sector and the SDGs: The need to move beyond 'business as usual.' *Sustainable Development, 24*(6), 371–382. https://doi.org/10.1002/sd.1623

Schot, J., & Steinmueller, W. E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. *Research Policy, 47*(9), 1554–1567. https://doi.org/10.1016/j.respol.2018.08.011

Schreyer, P. (2021). *Framing measurement beyond GDP*

Sebastiani, F. (2001). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*. https://arxiv.org/abs/cs/0110053v1

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*. https://doi.org/10.1007/s41133-020-00032-0

Sinha, A., Sengupta, T., & Alvarado, R. (2020). Interplay between technological innovation and environmental quality: Formulating the SDG policies for next 11 economies. *Journal of Cleaner Production, 242*, 118549. https://doi.org/10.1016/j.jclepro.2019.118549

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Stiglitz, J. E., Fitoussi, J.-P., & Durand, M. (2018). *Beyond GDP: Measuring what counts for economic and social performance*. (OECD, Ed.). OECD. https://doi.org/10.1787/9789264307292-en

Sullivan, K., Thomas, S., & Rosano, M. (2018). Using industrial ecology and strategic management concepts to pursue the Sustainable Development Goals. *Journal of Cleaner Production, 174*, 237–246. https://doi.org/10.1016/j.jclepro.2017.10.201

Suominen, A., & Seppänen, M. (2014). Bibliometric data and actual development in technology life cycles: flaws in assumptions. *Foresight*, *16*(1), 37–53. http://www.emeraldinsight.com/10.1108/FS-03-2013-0007

Suominen, A., & Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.23596

Suominen, A., Toivanen, H., & Seppänen, M. (2016). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2016.09.028

UN. (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development; United Nations: New York, NY, USA, 2015*. https://sustainabledevelopment.un.org/post2015/transformingourworld

United Nations Department of Public Information. (2015). 2030 Agenda for Sustainable Development—Sustainable Development Goals. United Nations, 1–24. https://www.undp.org/content/undp/en/home/librarypage/corporate/sustainable-development-goals-booklet.html

UNSDG. (2019). *Sustainable development report*. http://sustainabledevelopment.report

van der Waal, J. W. H., Thijssens, T., & Maas, K. (2021). The innovative contribution of multinational enterprises to the Sustainable Development Goals. *Journal of Cleaner Production*. https://doi.org/10.1016/j.jclepro.2020.125319

Vatananan-Thesenvitz, R., Schaller, A. A., & Shannon, R. (2019). A bibliometric review of the knowledge base for innovation in sustainable development. *Sustainability (switzerland), 11*(20), 1–22. https://doi.org/10.3390/su11205783

Walsh, P. P., Murphy, E., & Horan, D. (2020). The role of science, technology and innovation in the UN 2030 agenda. *Technological Forecasting and Social Change, 154*, 119957. https://doi.org/10.1016/j.techfore.2020.119957

Wang, M., Wang, Z., & Chen, G. (2019). Which can better predict the future success of articles? bibliometric indices or alternative metrics. *Scientometrics, 119*(3), 1575–1595. https://doi.org/10.1007/s11192-019-03052-9

Weikum, G. (2002). Foundations of statistical natural language processing. *ACM SIGMOD Record*. https://doi.org/10.1145/601858.601867

Xie, Z., & Miyazaki, K. (2013). Evaluating the effectiveness of keyword search strategy for patent identification. *World Patent Information, 35*(1), 20–30. https://doi.org/10.1016/j.wpi.2012.10.005

Yau, C. K. C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786. http://link.springer.com/article/10.1007/s11192-014-1321-8

Yoon, B., Lee, S., & Lee, G. (2010). Development and application of a keyword-based knowledge map for effective R&D planning. *Scientometrics, 85*(3), 803–820. https://doi.org/10.1007/s11192-010-0294-5

Yoon, B., & Phaal, R. (2013). Structuring technological information for technology roadmapping: Data mining approach. *Technology Analysis and Strategic Management, 25*(9), 1119–1137. https://doi.org/10.1080/09537325.2013.832744

Zhou, Y., Dong, F., Liu, Y., & Ran, L. (2021). A deep learning framework to early identify emerging technologies in large-scale outlier patents: An empirical study of CNC machine tool. *Scientometrics, 126*(2), 969–994. https://doi.org/10.1007/s11192-020-03797-8