Handelshøyskolen BI

GRA 19703 Master Thesis

Thesis Master of Science 100% - W

Predefinert informasjon

Startdato:	16-01-2022 09:00	**Termin:**	202210
Sluttdato:	01-07-2022 12:00	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	T		
Flowkode:	202210||10936||IN00||W||T		
Intern sensor:	(Anonymisert)		

Deltaker

Navn:	Hanna Silseth Gautvik og Emma Gammelsrud Thoresen

Informasjon fra deltaker

Tittel *:	Latent Dirichlet Allocation-based method and Cosine Similarity: Aligning Research Publications with the United Nations' Sustainable Development Goals
Navn på veileder *:	Rogelio Andrade Mancisidor

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	63
Andre medlemmer i gruppen:	

BI Norwegian Business School
Oslo, Spring 2022

Latent Dirichlet Allocation-based method and Cosine Similarity: Aligning Research Publications with the United Nations' Sustainable Development Goals

Emma Gammelsrud Thoresen and Hanna Silseth Gautvik

Supervisor: Rogelio Andrade Mancisidor

Master thesis, Business Analytics

BI NORWEGIAN BUSINESS SCHOOL

This thesis was written as a part of the Master of Science in Business Analytics at BI Norwegian Business School. Please note that neither the institution nor the examiners are responsible – through the approval of this thesis – for the theories and methods used, or results and conclusions drawn in this work.

Acknowledgements

This master's thesis marks the end of our MSc in Business Analytics and five years of higher education at BI Norwegian Business School. Writing this thesis has been a challenging, exciting, and educational process. Fortunately, we have received helpful assistance from several people, which we are grateful for.

We want to start by expressing our gratitude to our supervisor, Rogelio Andrade Mancisidor. He assisted us in developing ideas for our thesis from start to finish and in organising the work process. Additionally, he has always been available to answer questions and provided us with valuable guidance and feedback as we wrote the thesis.

Furthermore, we would like to express our gratitude to Marcelo T. LaFleur. His passionate interest in our work, valuable comments, and relevant experience assisted us in evaluating different research approaches for our thesis. Additionally, LaFleur provided us with data and shared a tool he developed for aligning publications with Sustainable Development Goals. Thus, his assistance has been valuable and essential in allowing us to answer our research question.

We will also thank the librarians, research assistants, and BI faculty members who impacted our thought processes and helped us with collecting the data for our master's thesis.

Last but not least, we would like to thank our family and friends for providing us with continuous support throughout our years of study and through the process of writing this thesis. This accomplishment would not have been possible without them. Thank you!

BI Norwegian Business School

Oslo, June 2022

Emma G.Thoresen	Hanna S. Gautvik
Emma Gammelsrud Thoresen	Hanna Silseth Gautvik

Abstract

According to the United Nations Statistics Division (2021) the development of the Sustainable Development Goals is running behind schedule. Higher education plays a significant role in prioritising and implementing research as part of their sustainability agenda. Although there are compelling reasons for research faculty to focus research contributions on advancing the Sustainable Development Goals, current applications for identifying such efforts heavily rely on manual involvement.

In order to assess how the research publications at a Norwegian business school were aligned with the Sustainable Development Goals, we investigated and implemented an SDG classifier based on Latent Dirichlet Allocation developed by LaFleur (2019). In addition, cosine similarity was used as an alternative method for identifying the most similar research publications in the corpus to the Sustainable Development Goals.

Our results show that BI Norwegian Business School, has made clear contributions to the Sustainable Development Goals, when using both the SDG classifier and cosine similarity. As a result, both methods are adequate for identifying Sustainable Development Goals in research publications for our purposes. However, the SDG classifier produces more reliable results than cosine similarity as it is able to capture Sustainable Development Goals-related topics in research publications where they are not always prominent.

***Keywords* –** LDA, Cosine Similarity, SDG

Contents

List of Figures

List of Tables

1 Introduction

The Sustainable Development Goals (SDGs) were adopted by all member states of the United Nations in 2015. A set of solutions to the world's most pressing issues. The SDGs constitute an established framework covering a spectrum of goals related to environmental, social, and economic sustainability (United Nations a, 2022). Research institutions have a critical role to play in the achievement of the SDGs by helping society to transform into pathways of sustainability. Recognising existing contributions is a necessary first step in strengthening universities' engagement with the SDGs (SDSN Northern Europe, 2022).

There has been a growing interest in determining how research aligns with the 17 SDGs, which is beginning to appear in various research tools (Aurora Universities Network, 2022; Clarivate, 2019). While these contributions continue, development on the SDGs is said to be behind schedule, which emphasise the need for universities and research institutions to further develop methods to track and analyse how their research is reflecting or incorporating the SDGs and to increase an understanding that it is a priority to implement a sustainable development agenda (United Nations Statistics Division, 2021).

When commiting to the SDGs, it is likely that research institutions will examine their publishing output to determine how it aligns to the goals. Nevertheless, current applications for identifying such efforts heavily rely on manual involvement. Exploring the field of Natural Language Processing (NLP) to achieve these objectives seem to be crucial given the research faculty's prolific research output.

Algorithms for topic modelling such as Latent Dirichlet Allocation (LDA) open up possibilities for evaluating large amounts of unstructured and unlabeled material (Blei, 2012). Additionally, distance measures allow for measuring similarity between documents. With an LDA-based approach and distance measures, the proposed thesis is to identify how research publications align with the SDGs, which brings us to the research question that has been posed thus far:

Could LDA-based methods and distance measures be useful to assess how well published business school research aligns with the United Nations' 17 Sustainable Development Goals?

2 Related Work

This section provides a review of the current approaches for identifying research publications that align with the SDGs. Additionally, we present the related work in regards to utilising LDA and distance measures for achieving the objectives of the thesis.

2.1 Identifying SDG-related Research Publications

Several researchers have sought out methods of identifying research articles related to the Sustainable Development Goals. Prior research has applied search strings with the help of subject matter experts which are then enhanced through distinct methods to produce a final set of articles (Aurora Universities Network, 2022; Clarivate, 2019). Clarivate (2019) is based on a core set of publications that contain "Sustainable Development Goals", making it relevant to research areas with explicit SDG discourse. As an alternative approach, Aurora Universities Network (2022) bibliometric techniques are focused on searching for strings of keywords, such as those contained in the SDG targets. In contrast to Clarivate (2019), Aurora Universities Network (2022) interpret that publications containing these keywords are the ones that are most aligned with the SDGs (Ràfols, 2020).

An emerging trend is the employment of machine learning methods (Wastl et al., 2020). The citation service Dimensions provides filters for filtering literature by SDGs (Wastl et al., 2020). Dimensions utilises an unsupervised machine learning method. The model compares tagged content, such as the title and source name of the article, with the description of SDGs (LaFleur, 2019).

When comparing the bibliometric corpora acquired with different techniques, Armitage et al. (2020) observed a considerable degree of inconsistency. Ràfols (2020) suggested that the differences were not due to technical issues, but rather were a result of different methodologies employed and subjective interpretations of the goals. As many of the methods are driven by human choices, it is reasonable to observe diverse results. To better understand the methodologies' relevance in discovering SDG-related research, each technique should be investigated further. It is currently too early to rely on any single method to measure progress toward the goals. Due to the exploratory nature of this thesis, two approaches to identifying research publications aligned with the SDGs were assessed: the first used Latent Dirichlet Allocation and the second used text similarity.

2.2 Latent Dirichlet Allocation

Topic modelling methods are powerful and unsupervised machine learning techniques that are widely used in natural language processing (NLP) for uncovering hidden thematic structures and extracting semantic information from unstructured textual data (Blei et al., 2003).

Latent Dirichlet Allocation (LDA) is one of the most commonly used topic models today and was first introduced by Blei et al. (2003). LDA is a generative probabilistic model of a corpus where the core idea is that documents are represented as random mixtures over latent topics, where each topic is expressed by a distribution over words (Blei et al., 2003). In statistical terms, this means that the words with the highest probabilities in each topic usually give a good indication of what the document's content is about (Jelodar et al., 2019).

Topic modelling algorithms based on LDA have been shown to be applicable to a wide range of knowledge domains, including software engineering, political science, and cognitive science, e.g. Consequently, the literature on LDA is extensive.

In software engineering, Linstead et al. (2007) utilised LDA to extract topics from software data and visualised software similarity for the first time. The effectiveness of topic modelling was demonstrated *"on 1,555 projects from SourceForge and Apache consisting of 19 million source lines of code (SLOC)"* (Linstead et al., 2007). The researchers demonstrated that LDA is an intuitive solution for computing similarity between files by examining their distributions over topics, which we also aim to do using LDA (Linstead et al., 2007). Additionally, their findings show that topic modelling is useful in software engineering because it can effectively extract functional and meaningful topics from source code.

Furthermore, LDA has been applied to analyse trends over a period in Cognition, a journal that publishes important publications in the field of cognitive science. As a result of applying LDA to a corpus of abstracts, several significant historical trends in the journal's paper topics were discovered. For instance, the researchers found that publications containing Moral topics such as social and emotional aspects of cognition increased *"from producing around 0.5% of the words in the mid-2000s* *to producing about 4% in 2014"* (Priva and Austerweil, 2015). Analysing trends is relevant to our thesis because we want to investigate if more SDG-related research publications have been published after the adoption of the SDGs in 2015.

Another group of researchers applied LDA to estimate scientific topics in a corpus consisting of abstracts from the Proceedings of the National Academy of Sciences of the United States of America (PANS) from 1991 to 2001 (Griffiths and Steyvers, 2004). A total of 300 meaningful topics were discovered by the researchers. For instance, topic 2 was related to Climate Change (Griffiths and Steyvers, 2004). By assessing trends and "hot topics" in the publications, the researchers use LDA in a transferable and appropriate way for this thesis, which is to identify topics related to SDGs in research publications.

LaFleur (2019) describes a proof-of-concept process for developing a classification system to assess how work for The United Nations system aligns with the SDGs. For instance, the results show that SDG 13 (climate action) and SDG 17 (partnership goal) are the most prominent in several publications from The United Nations Department of Economic and Social Affairs (DESA). The working paper serves as a great example of how specific topics, such as the SDGs, can be mapped in a consistent, scalable, and objective manner using LDA-based methods (LaFleur, 2019). LaFleur's methodology is particularly interesting for our thesis since LaFleur (2019) demonstrates how LDA can be used to identify SDGs in publications rather than just identifying "hot topics" in the corpus.

2.3 Distance Measures

Distance measures allow for computationally analysing how similar or dissimilar data objects are to one another. The distance between two points in Cartesian space can be calculated in a variety of ways, and various distance metrics have distinct applications.

Euclidean distance is referred to as a measure of dissimilarity and is commonly used when dealing with continuous data, since it generalises to any number of dimensions. Euclidean distance is calculated by the length of the straight line between two data objects (Friedman, 1997).

Another method, known as the *Manhattan distance*, measures the distance traveled if a grid-like path is taken, to get from one data object to another (Krause, 1986). If the variables being studied are not of the same type, the distance measure is a suitable choice. Additionally, the Manhattan distance metric consistently outperforms the conventional Euclidean distance metric in applications involving high-dimensional data (Aggarwal et al., 2001).

Whereas the distance measure *Cosine similarity*, in contrast, differs greatly from the other two in that it places more emphasis on the orientation of the data objects in the space than on their precise distance from one another (Bhattacharyya, 1946). The most common use of cosine similarity is in the context of documents, as we will see in the literature review that follows. This suggests that cosine similarity may be a distance metric that is appropriate for our objective.

Cosine similarity is often combined with *Term Frequency – Inverse Document Frequency* ($TF - IDF$), since the method considers documents as vectors in a vector space. Moreover, $TF - IDF$ is used to convert text into numbers so that each document can be represented as a vector (Schütze et al., 2008). Cosine similarity determines how similar two non-zero vectors are by calculating the cosine of the angle between them. It returns a similarity value between the vectors that ranges from 0 to 1, with 0 indicating "no similarity" and 1 indicating that the vectors are identical. When the magnitude of the vectors is irrelevant, cosine similarity is commonly utilised as a metric for determining distance since the metric corrects for documents of uneven lengths (Ristanti et al., 2019).

Measuring the similarity between documents using cosine similarity has been applied in a variety of domains including engineering, information technology and the educational field (Singh et al., 2020; Jain et al., 2017; Triwijoyo and Kartarina, 2019).

In the field of engineering, Singh et al. (2020) used cosine similarity to determine the similarity between movies when building a movie recommendation system. The recommendation system would recommend movies to the user based on the cosine similarity between a new movie and movies that the user had seen and rated. The researchers conclude that using cosine similarity gave more accurate recommendations than other distance metrics such as Euclidian distance mentioned in their study (Singh et al., 2020).

Jain et al. (2017) are looking into some of the most well-known algorithms and methods for retrieving desired information from large amounts of data. They conclude that cosine similarity is an effective similarity measures used in conjunction with $TF - IDF$. They have also come to the same conclusion as Singh et al. (2020) that Euclidean distance should not be used because it separates identical documents by a significant distance (Jain et al., 2017).

Furthermore, cosine similarity has been used to cluster documents based on their cosine similarity score. Triwijoyo and Kartarina (2019) clustered 83 scientific documents based on their cosine similarity score using the K-means algorithm for minimising the distance within each cluster. According to the researchers, clustering documents based on cosine similarity scores is appropriate because accurate grouping necessitates a precise definition of closeness between the objects being compared (Triwijoyo and Kartarina, 2019). Their method had an accuracy of 84.3%, implying that cosine similarity scores are useful for identifying and grouping documents that are similar.

Since there are several distance measures, studies such as Singh et al. (2020), Triwijoyo and Kartarina (2019) and Jain et al. (2017) demonstrate the applicability and advantages of cosine similarity. Ristanti et al. (2019) also concluded that cosine similarity is appropriate for evaluating the similarity between documents and for classifying them based on their cosine similarity score. The results from Triwijoyo and Kartarina (2019) and Ristanti et al. (2019) are pivotal, since our ambition is to measure how similar research publications by BI-affiliated researchers are to each of the 17 SDGs.

In the education field, cosine similarity has been used to develop an Automated Essay Scoring (AES) for assisting lecturers to score essays handed in by students efficiently and effectively (Lahitani et al., 2016). The essays are scored by AES based on how similar they are to an expert answer. After the essays have been scored, they are ranked according to how closely they match the expert's answer. The researchers concluded that using cosine similarity would improve the objectiveness of essay evaluations and speed up the correction process for lecturers (Lahitani et al., 2016).

In addition to confirming that cosine similarity is a good method for measuring textual similarity, Lahitani et al. (2016) show how documents can be ranked based on their cosine similarity score, which we also intend to do in our analysis of research publications.

3 Methodology

In this section we will present the proposed LDA-based approach for identifying SDGs in research publications. Then, we will discuss cosine similarity as an method for measuring text similarity between research publications an the SDGs.

3.1 LDA

LDA is a statistical model of document collections that attempts to capture the intuition that documents are composed of a variety of topics (Blei et al., 2010). One research publication, for instance, might discuss topics such as gender diversity, politics, and the economy. The LDA algorithm, which is the simplest topic model, creates semantic meaningful clusters from collections of textual data by considering documents as the result of probabilistic sampling over the topics describing the corpus and over the words that generate each topic (Hsu et al., 2022; Blei et al., 2003). The Plate Notation illustrates the overall generative process of how LDA assumes that documents are generated:

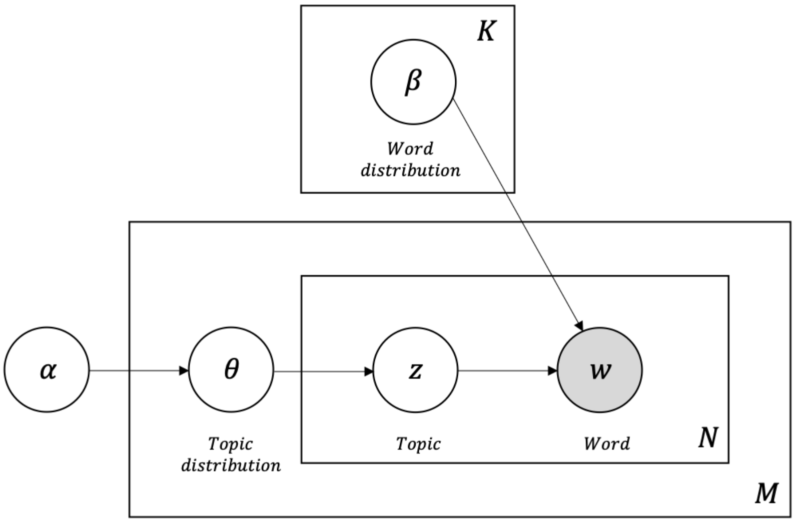

Figure 3.1: LDA Plate Notation

The outer rectangular box represents the number of documents *M* while the inner rectangular box represents the repeated choice of topics and words within a document *N*. The squared box represents the number of topics *K*, which must be defined in advance. α and β are hyperparameters, where is the Dirichlet prior parameter per-document topic distributions and β is the Dirichlet prior parameter for the per-topic word distribution (Blei et al., 2003). A high α value will make documents appear more similar to one another and a high β value will make the topics appear similar to one another. These parameters are assumed to be sampled once in the process of generating a corpus and influence the smoothing over word distributions in the topics (Vayansky and Kumar, 2020). θ_m represents the multinominal distribution of topics within a document m. Z is the topic variable used to denote each topic assigned to each word Z_{mn}. w is the word variable used to denote observed words, the *n-th* word in the *m-th* document. These are word-level variables sampled once for each word in each document (Blei et al., 2003).

As mentioned, LDA assumes that each document is a mixture of topics and that each topic is a mixture of word (Blei et al., 2003). Additionally, LDA makes the "bag of words" assumption, meaning that the order of the words in the documents does not matter, only the words' frequency of occurrence is considered when determining the likelihood that a word belongs to a particular topic. Each topic is composed of a list of words with differing probabilities of belonging to a topic. Thus, a topic can formally be defined as a distribution over a fixed vocabulary (Blei et al., 2010).

To identify the topics, LDA extrapolates backwards from the collection of documents to determine which topics have generated the documents and which words have generated the topics (LaFleur, 2019). More precisely, LDA assumes that there are k topics for the entire collection of documents and generates the words making up each topic in a two-step process. First, LDA randomly selects a distribution over topics. Then, for each word in each document, the words are randomly chosen from the distribution over the vocabulary. Thereafter, a topic that best describes the collection of words is "activated" from the distribution over topics. All documents share the same topics, but each document exhibits the topics with different proportions (Blei et al., 2010). For instance, the model can determine that research publication, *m*, contains 30% of topic 1, 50% of topic 2, and 20% of topic 3.

For our corpus of research publications, it is more likely that topics associated with politics, economics, and leadership i.e., will be chosen as the topics that best describe the collection of documents rather than SDG-related topics. Thus, we investigated how LDA could be transformed from unsupervised into a more supervised method allowing us to identify specific topics such as the SDGs.

In the literature on LDA, we discovered LaFleur (2019) who developed an "SDG classifier" to understand how DESA publications were aligned with the SDGs.

The SDG classifier is based on topics estimated using LDA. LaFleur (2019) used a pre-selected collection of texts representing each of the SDGs to estimate an 18-topic model, one topic for each SDG and one general topic. The extra topic serves as a filter, capturing the words that appear in the texts representing each SDG (LaFleur, 2019). Table 3 in LaFleur (2019) displays the 20 most important words making up each topic. For instance, the words "poverty", "social_protection" and "disaster" make up the SDG 1-topic, the words "climate_change", "paris_agreement" and "emissions" make up the SDG 13-topic. Additionally, the words "sustabinable_development", "people" and "economic" make up the filter-topic. This demonstrates that the LDA algorithm successfully generated a probabilistic model capable of distinguishing between the 17 "specific" topics representing each of the SDGs and the one "general" topic (LaFleur, 2019).

The 18-topic model was trained on the texts that represented each of the SDGs and then applied to out-of-sample data, the DESA publications, in order to understand how the DESA publications were related to the SDGs (LaFleur, 2019).

LaFleur's approach demonstrates how LDA can be used to identify specific topics in a corpus. In order to help us answer our research question, LaFleur was willing to share his work with us. Hence, the next section will go into more detail on how the SDG classifier works.

3.2 SDG classifier

LaFleur developed the SDG classifier based on an increasing interest in measuring how the work of the UN systems aligned with the SDGs. In addition, he saw a need for having *"a scalable, objective, and consistent way to measure how similar any given publication is to each of the 17 SDGs"* (LaFleur, 2019).

LaFleur (2019) carefully selected 17 texts to represent each SDG in order to estimate the 18-topic model using LDA, which forms the basis of the classifier. As a result of the texts being sufficiently unique, the LDA algorithm successfully generated a probabilistic model capable of differentiating the 17 SDG topics and the general topic, meaning that it also could be used to classify other collections of documents in accordance with the 17 SDGs (LaFleur, 2019). In order to make the model capable of classifying other datasets, LaFleur trained the model on the texts representing each SDG using the open-source tool Mallet (LaFleur, 2019; Mallet, 2021).

When training a classifier, it can be a challenge to have enough labelled data to train the classifier to correctly classify the data (LaFleur, 2019). Additionally, creating high-quality labelled training data for SDG classification is difficult since each SDG consists of multiple concepts and themes (Hsu et al., 2022). By relying on the probabilistic nature of how LDA assigns topics and using them as true labels for the training data, LaFleur (2019) avoids the need for having large, labelled training data when training the classifier (LaFleur, 2019; Hsu et al., 2022). The SDG classifier is trained on data designed to maximise its ability to distinguish the 17 SDG topics, making it a semi-supervised method for assessing how publications align with the SDGs (LaFleur, 2019).

LaFleur (2019) The SDG classifier makes it possible to compute “SDG scores” for individual texts and for larger collections of texts. When the SDG classifier is applied to a collection of documents, each document is assigned a score indicating how similar each document is to each of the 17 SDGs. Each document is assigned 18 scores, which add up to 1 when the filter topic is also considered. The scores are interpretable as percentages.

To validate the classifier, LaFleur (2019) first applied the SDG classifier to the training data before applying it on out-of-sample data. Table 2 in LaFleur (2019) shows that each of the 17 topics is strongly associated with only one of the texts representing each SDG (LaFleur, 2019). Additionally, since this method is aimed to assist us in answering a part of our research question, we also wanted to run the SDG classifier on the training data ourselves. The results are displayed in Table 3.1 and are consistent with those obtained when LaFleur (2019) validated the classifier.

The findings imply that the SDG classifier will be able to identify content associated with each sustainability goal in other texts as well, which gives us the confidence to utilise this approach in our thesis.

Texts	**SDG 1**	**SDG 2**	**SDG 3**	**SDG 4**	**SDG 5**	**SDG 6**	**SDG 7**	**SDG 8**	**SDG 9**	**SDG 10**	**SDG 11**	**SDG 12**	**SDG 13**	**SDG 14**	**SDG 15**	**SDG 16**	**SDG 17**	**Filter**
sdg1.txt	**0.72**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.28**
sdg2.txt	0.00	**0.75**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.25**
sdg3.txt	0.00	0.00	**0.83**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.17**
sdg4.txt	0.00	0.00	0.00	**0.81**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.19**
sdg5.txt	0.00	0.00	0.00	0.00	**0.84**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.16**
sdg6.txt	0.00	0.00	0.00	0.00	0.00	**0.76**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.24**
sdg7.txt	0.00	0.00	0.00	0.00	0.00	0.00	**0.73**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.27**
sdg8.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.78**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.22**
sdg9.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.74**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.26**
sdg10.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.68**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.32**
sdg11.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.76**	0.00	0.00	0.00	0.00	0.00	0.00	**0.24**
sdg12.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	0.00	0.00	**0.17**
sdg13.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	0.00	**0.16**
sdg14.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.83**	0.00	0.00	0.00	**0.17**
sdg15.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.84**	0.00	0.00	**0.16**
sdg16.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.74**	0.00	**0.26**
sdg17.txt	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.77**	**0.23**

Table 3.1: SDG Scores per SDG text

3.3 Cosine Similarity

We also used cosine similarity from Scikit-learn to compute the similarity between research papers and the SDGs. Cosine similarity allowed us to measure the same as when using the SDG classifier, but in a different way. As a result, we could compare the SDG classifier's results with another method's results.

Mathematically, the cosine similarity is defined as the dot product of the vectors A and B divided by their magnitude. The similarity between vector A and vector B is calculated as:

$$cos(\theta) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{||\boldsymbol{A}|| \cdot ||\boldsymbol{B}||} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \tag{3.1}$$

θ is the cosine angle between the two vectors, which determines whether the vectors are pointing in roughly the same direction (Xia et al., 2015). $A{\cdot}B$ is the dot product of the vectors, which is calculated as shown in the equation's numerator. The magnitude of the vectors, ||A|| and ||B||, is calculated as shown in the denominator of the equation. If the angle between the vectors is 0 degrees, the cosine similarity is 1 and the documents are identical to each other. Cosine similarity varies between 0 and 1, the higher the similarity values are the more similar the documents are (Beysolow et al., 2018).

In order to compute the cosine similarity, we first needed to compute $TF - IDF$ vectors for each document within the corpus. We decided to use $TF - IDF$ because it makes computing the cosine similarity in Python relatively simple and quick, requiring only a few lines of code. We used Scikit-learns tfidfVectorizer, which both counts each word's occurrence word and weights them according to how frequently they appear in the corpus (Scikit-learn, 2022). In other words, the words are scored according to their relative importance within the corpus (Ozsoy et al., 2011). Mathematically, the $TF - IDF$ is calculated in the following way:

$$TF - IDF = tf_{ij} * \log(\frac{N}{df_i + 1}) \tag{3.2}$$

tf_{ij} is defined as the number of times a word i appears in document j, while df_i is the number of documents in which word i appears. N is the total number of documents in the corpus (Kim et al., 2019). For instance, rare words will have low $TF-IDF$ values because of TF being low and common words appearing in most or all documents will have low $TF-IDF$ values because of IDF being low (Apeltsin, 2021).

In addition to the standard data pre-processing steps required when working with textual data, tfidfVectorizer makes it possible to define some parameters for reducing the vocabulary by ignoring words that may be considered noise. We defined max_{df} to be 0.80 and min_{df} to be 0.01. Words appearing in more than 80% of the documents are most likely common English words and thus are assumed to be general. Moreover, words appearing in less than 1% of the documents are considered less important (Scikit-learn, 2022).

TfidfVectorizer returns a $TF-IDF$ matrix, which we converted into a dataframe where each column represented a word in the corpus's vocabulary and each row represented a document. Thereafter, the $TF-IDF$ dataframe was used as input into scikit-learns cosine similarity to compute the cosine similarity between the research paper and the 17 texts representing each SDG. Cosine similarity returns a matrix, showing the pairwise similarities between all documents within the corpus.

4 Data

The datasets used in our thesis consist of research publications published by BI-affiliated researchers and 17 unique texts describing each of the SDGs. In the following section we review the data and also present the necessary pre-processing steps for textual data before analysing the results.

4.1 Research Publications

The research publications used in this thesis was made available in collaboration with our supervisor and the faculty staff at BI Norwegian Business School. The research publications are collected during the time period from 2006 to 2019 and includes 1847 documents. The documents were collected in.txt format and include the abstracts as well as the body of the publication. The documents are composed of research publications to which BI researchers have contributed, thus not explicitly published by BI researchers.

4.2 Sustainable Development Goals

After reading *"Art is long, life is short: An SDG Classification System for DESA Publications"*, we asked LaFleur if he would share the data that was used to develop and train the SDG classifier described in the working paper. He was willing to share the training data and sent us 17 different .txt files, describing each of the 17 SDGs.

The texts describing each of the SDGs were retrieved from two sources, the UNs webpage which describes the SDGs and their associated sup-targets, and from the Secretary-General's annual report *"Progress towards the Sustainable Development Goals"* for the years 2016 to 2018 (LaFleur, 2019). The texts are well-balanced in length and designed to maximise each SDG's uniqueness, enabling a model to distinguish between the 17 SDGs (LaFleur, 2019). As a result, research papers that match how the goals and their sub-targets are described will be considered relevant to the SDGs.

Further in this thesis, we will refer to the 17 texts describing each SDG as the SDG texts.

4.3 Data Pre-processing

We went through all the research papers as a first step in the data pre-processing process to familiarise ourselves with the data we had. We discovered several duplicates and files with no content while reviewing the data, and these were removed. Additionally, we removed all research papers written in Norwegian because we would be unable to capture content related to any of the SDGs in Norwegian papers since the SDG texts were written in English. We ended up with 1799 research papers published between 2006 and 2019 after removing duplicates and research papers written in languages other than English.

Furthermore, we cleaned the data using the Natural Language Toolkit, which is a common NLP pre-processing method (NLTK, 2022). It was crucial to remove words that provided no information as well as noise, such as signs and different word spellings because the words used in the corpus hav a significant impact on how well both of the methods perform. Therefore, we used NLTK (2022) to remove stopwords, tokenize, and stem the corpus, which included both research papers and SDG texts.

4.3.1 Stopword removal

We removed stopwords because they are commonly used English words that are unlikely to be relevant to the content of research papers or the SDGs. For instance, stopwords can be words like “I”, “the” and “it”. We used NLTK’s built-in stopword list to remove stopwords from the English dictionary because all the text we were going to analyse is written in English (Accessing Text Corpora and Lexical Resources, 2019).

When removing stopwords, it is necessary to double-check which words are removed, as some deletions may result in the loss of vital information. A sentence’s meaning can be significantly altered by removing some stopwords. As an example, “not like” will become “like” when stopwords are removed since “not” is defined as a stopword. Therefore, we checked the stopwords that were removed to ensure that this did not have a significant impact on the content of our corpus.

4.3.2 Tokenization

We used NLTK's tokenizer to split the corpus into lists of substrings (NLTK, 2022). As a result of the tokenization, every continuous string of letters in the alphabet will be recognised as a single token. For instance, "Studying stable and sustainable organizations" was converted into a list of tokens: ["Studying", "stable", "and", "sustainable", "organisations"].

Furthermore, we made sure to make all tokens lowercase. This is an important step since "Sustainable" and "sustainable" will be identified as two different words if "Sustainable" was not made lowercase. These procedures are carried out to reduce the complexity of the corpus and increase the likelihood of obtaining more meaningful results.

4.3.3 Stemming

To stem the words in our corpus, we used NLTKs PorterStemmer. The goal of stemming is to eliminate multiple forms of a single word by reducing it to its root, removing derivational suffixes such as -ed, -ize, and -de (The Porter Stemming Algorithm, 2006). For instance, "economic" and "economy" was reduced to "econom". The PorterStemmer also changes words like "policy" to "polici" in some cases.

One issue related to the PorterStemmer is that it frequently generates stems that are not valid English words because it does not keep a lookup table for actual stems, but instead uses algorithmic rules to generate stems (The Porter Stemming Algorithm, 2006). Consequentially, there is a risk of under- and over stemming words since they are not always valid English. For instance, we discovered that the word "organisation" was stemmed to "organ" which can also refer to an organ such as the heart or kidney. Additionally, the word "identify" was stemmed to "indentifi". Even though several words had multiple meanings, or the English was incorrect, we used the PorterStemmer because of its simplicity and speed. We could still derive meaning from the data by conducting tests, examining the context of the words, and generally interpreting the meaning of the words.

5 Analysis

We have analysed the corpus using the SDG classifier and cosine similarity. We reveal the top three research publications associated with each of the most addressed SDGs for both methodologies, along with the most pertinent SDGs found in the corpus.

5.1 Basis for the analysis

After the pre-proseccing process the corpus consists of 1799 research publications published by BI-affiliated researchers between 2006 and 2019.

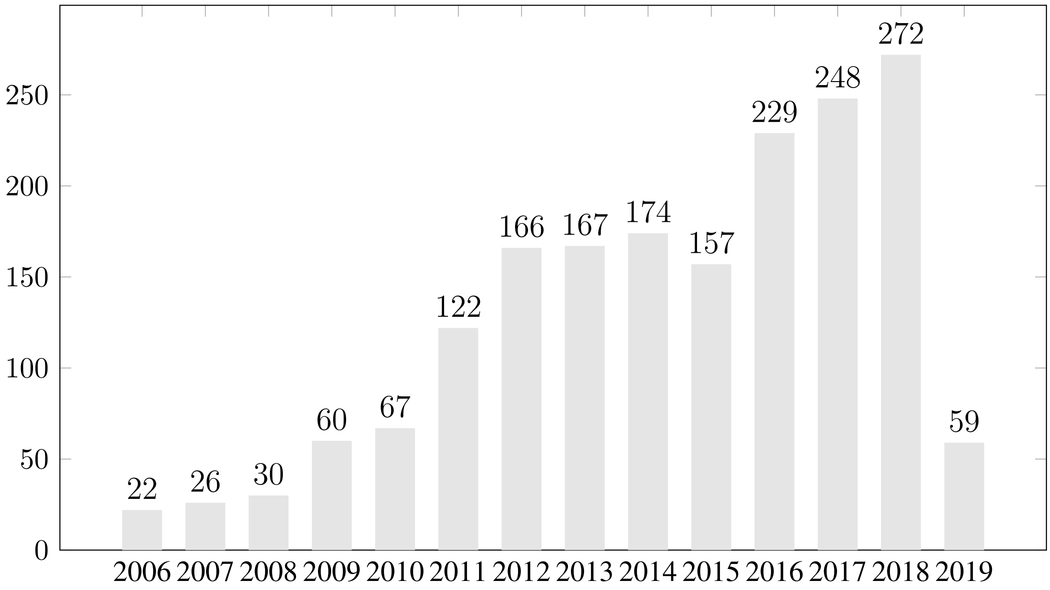

Figure 5.1: Research Publications by Year

The research publications are analysed using two different approaches, thus, there is discrepancies in how the results are interpreted, particularly in this section, as the two approaches assign scores differently. The SDG classifier assigns a score to each of the SDGs, which adds up to 1 when all the scores per research publication are added together. Cosine similarity, on the other hand, assigns each publication a score between 0 and 1, depending on how similar the publications is to one of the SDGs.

5.2 SDG classifier

The SDG classifier assigned scores to the research publications based on how closely they corresponded to each of the SDGs. As a result, each research publication was assigned 17 scores, which add up to 1. These scores can be interpreted as probabilities.

All the research publications were found to be relevant to at least one of the SDGs. Even though all research publications received some SDG scores above 0, many received scores of less than 0.005, indicating a similarity of less than 5% between the research publications and the SDG texts. For instance, a research publication was classified as 0.13% similar to SDG 1, 43% similar to SDG 3, 4% similar to SDG 17 and 0% similar to the rest of the SDGs.

Consequently, we read several papers with SDG scores indicating less than 10% similarity. We found it challenging to determine with certainty whether these research publications were relevant to any of the SDGs. As a result, we decided to set 10% similarity as a lower threshold for research publications that we wanted to investigate further. After applying the threshold, 91.78 % of the research publications were classified as relevant to the SDGs. The bar chart below shows the number of research publications with content that was classified as similar to the SDG texts.

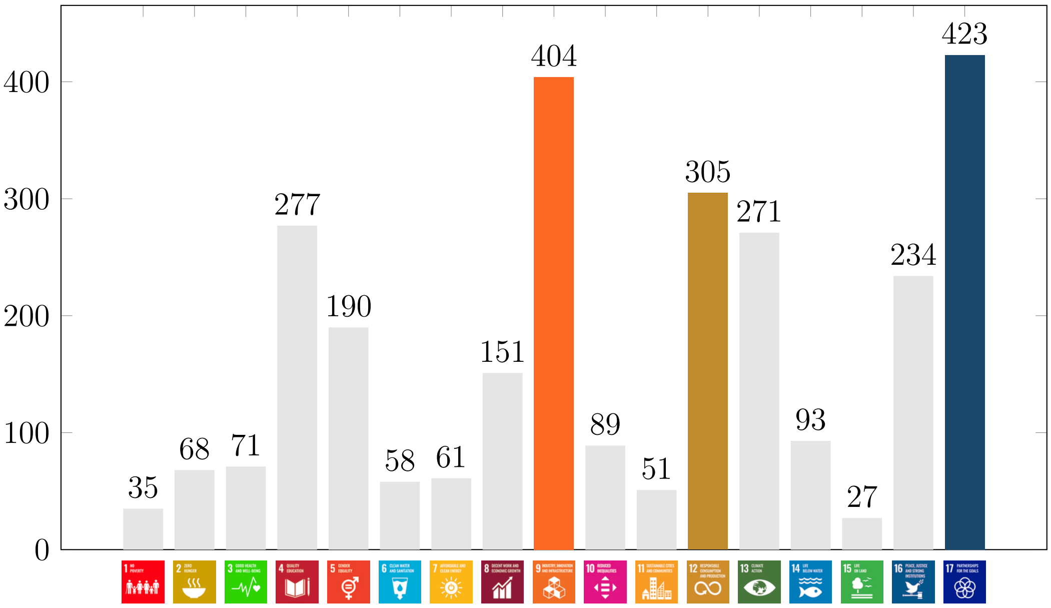

Figure 5.2: Research Publications with SDG Score ≥ 0.1

According to our results with the SDG classifier, the most frequently addressed SDGs by BI-affiliated researchers appear to be *SDG 17 – Partnership for the goals, SDG 9 – Industry, innovation, and infrastructure, and SDG 12 – Responsible consumption and production*. We ranked the research publications classified as relevant for SDG 17, SDG 9, and SDG 12 and read the three with the highest SDG scores for each

SDG to see if the highest-scoring papers actually contained topics relevant to the goals they were classified as similar to.

Rank	Research Publication	*SDG Score*
1	Tax holidays in a BEPS perspective (Bjerkestuen and Willie, 2015)	0.3193
2	Leaning against the credit cycle (Gelain et al., 2018)	0.3054
3	Partial fiscal descentralization and sub-national government fiscal discipline: empirical evidence from OECD countries (Asatryan et al., 2015)	0.2935

Table 5.1: SDG Score - Top three Research Publications related to SDG 17

SDG 17 – *Partnership for the goals* encompasses a wide range of topics, including civilising business enterprises, socially responsible investment, and monetary policy (Bjørnland et al., 2019). SDG 17 is among the three goals to which BI claims that their academic work contributes the most, together with SDG 16, and SDG 3 (BI Norwegian Business School, 2022). Thus, we also expected that at least one of these SDGs would appear as significant in research conducted by BI-affiliated researchers. The three research publications with the highest SDG scores for SDG 17 are about tax policies, monetary policy, and government expenditure financing, all related to topics covered by SDG 17.

Rank	Research Publication	*SDG Score*
1	Dynamic capabilities and innovation capabilities: The case of the ‘Innovation Clinic’ (Strønen et al., 2017)	0.3070
2	Geographic versus Industry Diversification: Constraints Matter (Ehling and Ramos, 2006)	0.2878
3	Public Policy and Industry Views on Innovation in Construction (Bygballe and Ingemansson, 2011)	0.2785

Table 5.2: SDG Score - Top Three Research Publications related to SDG 9

SDG 9 – *Industry, innovation, and infrastructure* cover topics like collaboration, digital inequalities, strengthening developing countries, sustainable industrialisation, and fostering innovation (United Nations b, 2022). To ensure continuous research development within selected sectors, BI has established research centres relevant to reaching specific SDGs. The research Centre for Construction Industry is concerned with issues related to sustainability and climate change in the construction industry and is thus relevant to reaching SDG 9 (BI Norwegian Business School, 2022). The top-ranked research publications align with SDG 9, discussing the development of capabilities and innovation, whether geographic diversification outperforms industry diversification, and how public policies affect construction innovation.

SDG 12 – *Responsible consumption and production* necessitate that everyone takes a comprehensive set of actions to adapt to sustainable practises. SDG 12 covers topics such as recycling, procurement practises, and making supply chains more efficient and sustainable (United Nations c, 2022). The global economy relies on both consumption and production. Simultaneously, anthropogenic harm to the environment and human health is almost entirely caused by production and consumption activities. As a result, SDG 12 is strongly linked to many of the SDGs, if not all of them (Hoballah and Averous, 2015).

Even though SDG 12 is not one of the goals BI is focusing the most on through its research centres i.e., publishing research that addresses a goal as broad as SDG 12 is appealing because it implies that BI's research is relevant to a wide range of topics that may contribute to reaching several of the SDGs. The three research publications with the highest SDG scores for this goal are all relevant to this goal because they address issues like managing construction supply chains and how to make them more efficient, reusing bottles, environmentally friendly product packaging, and identifying business networks to ensure sustainable collaborations.

Rank	**Research Publication**	***SDG Score***
1	Interdependence in Supply Chains and Projects in Construction (Bankvall et al., 2010)	0.4417
2	Environmental impact of refillable vs. non-refillable plastic beverage bottles in Norway (Bø et al., 2013)	0.3683
3	Conceptualising, delineating and analysing business networks (Prenkert and Hallén, 2006)	0.3517

Table 5.3: SDG Score - Top three Research Publications related to SDG 12

Although the SDG scores are not particularly high for the top-ranked research publications, they all contain content clearly relevant to the sustainability goals to which they are classified as most similar. The findings demonstrate that the SDG classifier is capable of correctly classifying research publications that contain content relevant to the 17 SDGs. The table below shows the highest SDG score assigned to a research publication for each SDG.

SDGs	***max. SDG Score***	$d_{SDGScore} \geq 0.1$
1 - No Poverty	0.1882	35
2 - Zero Hunger	0.2244	68
3 - Good Health and Well-being	0.4334	71
4 - Quality Education	0.5157	277
5 - Gender Equality	0.3677	190
6 - Clean Water and Sanitation	0.2166	58
7 - Affordable and Clean Energy	0.2878	61
8 - Decent Work and Economic Growth	0.3617	151
9 - Industry, Innovation and Infrastructure	0.3070	404
10 - Reduced Inequalities	0.2321	89
11 - Sustainable Cities and Communities	0.2059	51
12 - Responsible Consumption and Production	0.4417	305
13 - Climate Action	0.2766	272
14 - Life Below Water	0.1895	93
15 - Life on Land	0.1746	27
16 - Peace, Justice and Strong Institutions	0.3986	234
17 - Partnerships for the Goals	0.3193	423

Table 5.4: SDG Score Distribution

SDG 4 has the highest SDG score of 0.5157, meaning that a research publication is 51.57% similar to the text describing this goal. Also, none of the research publications is more than 51.57% similar to any other SDG. This can be interpreted as several research publications containing content relevant to each SDG, but the SDGs' topics are not prominent in the majority of them. Despite this, the research publication with the lowest maximum SDG score is well above the threshold we set, implying that if we read the research publication, we would most likely be able to say that it is relevant to the SDG.

Research publications may also contribute more to the SDGs in some years than in others. As a result, we visualised the average SDG scores over time. Each SDG is listed on the vertical axis and the years on the horizontal axis. The size of the bubbles represents the average SDG score for each SDG over the years, indicating how closely the text of the research publications matches the SDG text.

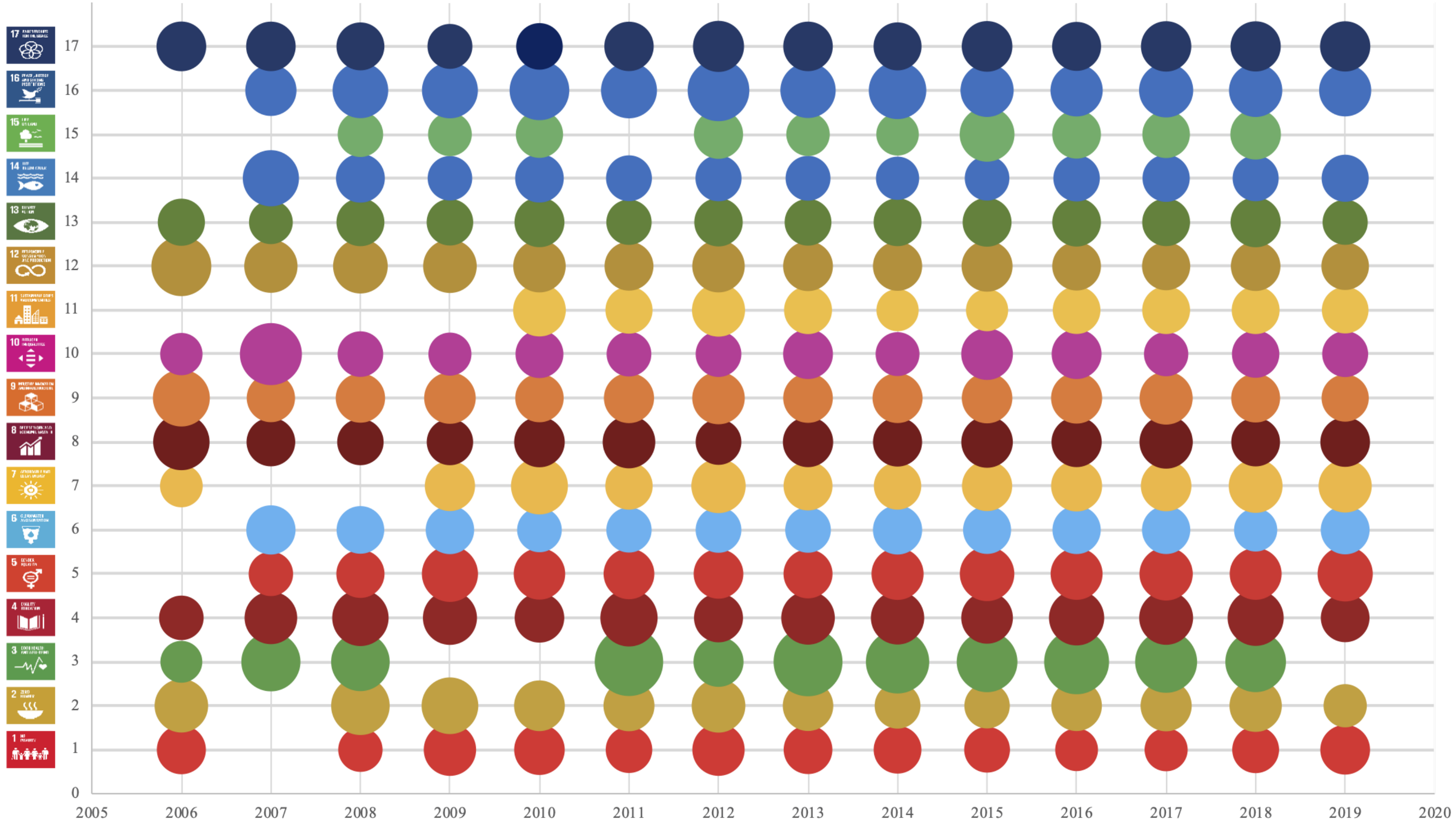

Figure 5.3: Average contributions to the SDGs over time - SDG Scores

At first glance, the contributions to the SDGs appear to be consistent across all years, but closer examination reveals that some of the SDGs have no bubbles in certain years. For instance, it appears that research was not relevant to SDG 11 until 2010 with the similarity threshold we have defined. Additionally, the number of research publications considered relevant to the SDGs does not appear to have increased significantly as a result of the adoption of SDGs in 2015.

SDG 4, SDG 9, SDG 12, SDG 13, SDG 16, and SDG 17 appear to have the largest bubbles and most consistent bubbles over time, indicating that they have the highest average SDG score over time. These observations are also consistent with Figure 5.2, which shows that SDG 17, SDG 9, and SDG 12 are the most frequently addressed SDGs in the research publications.

Following the top three goals, 5.2 show that SDGs 4 – *Quality education*, 13 – *Climate action*, and SDG 16 – *Peace, justice, and strong institutions* are also highly addressed goals.

Table 5.4 show that SDG 4 are linked to the research publication with the highest SDG score in our corpus. Consequently, we wanted to see how well the publication corresponded to SDG 4.

SDG 4 – *Quality Education* is concerned with issues such as educational policies, gender equality in the classroom, and the development of skills such as learning to read and write (United Nations d, 2022). BI is committed to reaching SDG 4 by providing quality education to its students as a part of its sustainability strategy towards 2025, implying that publishing research related to these topics is pertinent and essential for BI (BI Norwegian Business School, 2022).

The research publication *"Developmental dynamics of early reading skill, literacy interest and readers' self-concept within the first year of formal schooling"* (Walgermo et al., 2018) received an SDG score of 0.5157, which is the highest SDG score given to a research publication in our analysis. The score implies that the research publication is 51.57% similar to SDG 4. Walgermo et al. (2018) are about factors influencing children's early reading skills, a topic highly relevant to this goal.

The SDG classifier produces reliable results because the SDG scores correspond well with how closely the content of the research publications matches the content of the SDG texts. After reading the research publications with the highest SDG scores for each of the SDGs discussed above, we have seen that all the top-ranked research publications contain topics relevant to all the goals.

Additionally, the majority of the goals that were relevant to investigate because many research publications were classified as relevant to those goals are also goals that BI claims to be actively working on.

5.3 Cosine Similarity

All research publications received a cosine similarity score indicating that the content was similar to at least one of the SDGs. However, many research publications also received low scores when using cosine similarity. For instance, a research publication only received scores below 0.03, indicating that the research publication was less than 3% similar to any of the SDGs.

As a result, we followed the same steps as with the SDG classifier and read several research publications with scores below 0.1. Predictably, determining whether a research publication with such a low score is relevant to one or more of the SDGs was difficult. We had to be a bit stricter with the lower threshold than with SDG classifier because the content was a little more indefinable on a threshold of 0.1, so we decided to set it to 0.2.

After applying 20 % as the lower threshold for cosine similarity as well, 10.11% of the research publications were relevant for further analysis. The bar chart in the following page shows the number of research publications considered relevant to the SDG texts after applying the threshold. According to our results using cosine similarity, the most frequently addressed SDGs by BI-affiliated researchers appear to be SDG 13 – *Climate action*, SDG 5 – *Gender equality* and SDG 10 – *Reduce inequality*. These results did not align with the results we obtained using the SDG classifier. Therefore, we also ranked and read the top three research publications for these SDGs.

SDG 13 – *Climate Action* seeks to combat climate changes and addresses topics like energy efficiency, the meat industry, waste recycling, and carbon footprint reduction (United Nations e, 2022). BI is engaged in reaching SDG 13 in several ways, including through BI's research Centre for Construction Industry and as research partners in Klima 2050, a Centre for Research-based Innovation (BI Norwegian Business School, 2022). Consequently, publishing research relevant to achieving SDG 13 is prioritised, as evidenced by our findings. The top three research

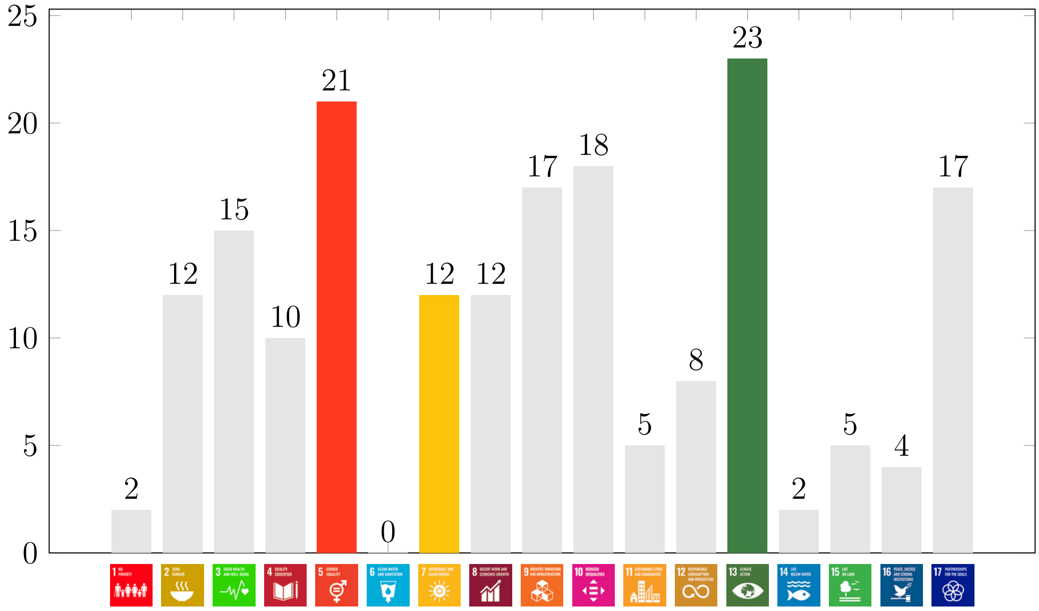

Figure 5.4: Research Publications with Cosine Similarity Score ≥ 0.2

publications discuss climate communication strategies, preparing society for future climate challenges, and benchmarking national climate strategy.

Rank	Research Publication	*CSS*
1	Rethinking climate communications and the "psychological climate paradox" (Stoknes, 2014)	0.6228
2	User guides for the climate adaptation of buildings and infrastructure in Norway – Characteristics and impact (Hauge et al., 2017)	0.4959
3	A Kantian approach to sustainable development indicators for climate change (Greaker et al., 2013)	0.4855

Table 5.5: CSS - Top three Research Publications related to SDG 13

SDG 5 – *Gender equality* addresses a wide range of topics, including sexual violence, workplace harassment, childcare services, and women in leadership (United Nations f, 2022).

According to BI's sustainability strategy towards 2025, SDG 5 is one of three goals that BI will prioritise. As a result, SDG 5 will receive a special focus in BI's operations, including employment, running facilities and supply chain management in the coming years (Bjørnland et al., 2019).

The top three research publications are about cultural factors that influence women serving on boards of directors, discrimination in the workplace and how economic and non-economic factors affect the empowerment of women in self-help groups.

Rank	Research Publication	*CSS*
1	Understanding Cultural Factors Which Affect Women Serving on Boards of Directors (Nguyen et al., 2017)	0.4362
2	Career equality: Inclusion and opportunities in a professional service firm in Norway (Traavik, 2018)	0.4285
3	Factors empowering women in Indian self-help group programs (Bali Swain and Wallentin, 2012)	0.4223

Table 5.6: CSS - Top three Research Publications related to SDG 6

SGD 10 – *Reducing inequality* aims to reduce inequalities within and among countries, making relevant topics about digital inequality, refugees, income inequality, discriminatory laws, and politics (United Nations g, 2022).

BI is contributing to reaching this goal through its research Centre for Internet and Society, which focuses on digital inequality as well as the social and labour characteristics of the sharing economy.

The articles with the highest cosine equality scores discuss topics connected to infrastructure investments, migration, and the factors that drive equity returns in the Euro-zone.

Rank	Research Publication	*CSS*
1	Complementing clusters: a competitiveness rationale for infrastructure investments (Sasson and Reve, 2015)	0.3331
2	Brain drain or brain gain? (Maurseth, 2019)	0.3315
3	Euro-zone equity returns: country versus industry effects (Eiling et al., 2012)	0.2610

Table 5.7: CSS - Top three Research Publications related to SDG 10

The cosine similarity scores for the top-ranked research publications range from 0.2610 to 0.6228, implying that the research papers are considered to be between 26% and 62% similar to the SDG texts. After reading the top-ranked publications, we observed that they cover topics that are relevant to each of the SDGs, indicating that cosine similarity is also an effective approach for identifying similar documents. The highest cosine similarity scores assigned to a research publication for each SDG are shown in the table below.

SDGs	***max. CSS***	$d_{CSS} \geq 0.2$
1 - No Poverty	0.2846	2
2 - Zero Hunger	0.4183	12
3 - Good Health and Well-being	0.3378	15
4 - Quality Education	0.4031	10
5 - Gender Equality	0.4362	21
6 - Clean Water and Sanitation	0.1462	0
7 - Affordable and Clean Energy	0.5824	13
8 - Decent Work and Economic Growth	0.3537	12
9 - Industry, Innovation and Infrastructure	0.4530	17
10 - Reduced Inequalities	0.3332	18
11 - Sustainable Cities and Communities	0.3496	5
12 - Responsible Consumption and Production	0.2787	8
13 - Climate Action	0.6228	23
14 - Life Below Water	0.2824	3
15 - Life on Land	0.2767	5
16 - Peace, Justice and Strong Institutions	0.2379	4
17 - Partnerships for the Goals	0.3341	17

Table 5.8: Cosine Similarity Score Distribution

With a cosine similarity score of 0.6228, SDG 13 has the highest cosine similarity score (CSS), indicating that a research publication is 62.28% similar to the SDG text that describes SDG 13. Additionally, no other research publication is more similar to one of the SDG texts than that one. This research publication which received this score has already been mentioned: *"Rethinking climate communications and the "psychological climate paradox"* (Stoknes, 2014).

SDG 6 has the lowest maximum CSS, which may indicate that the research publication's relevance to SDG 6 is not immediately apparent to the reader. The lower threshold of 0.2 was chosen for a reason; it varied how confident we could be that a research publication was relevant to a specific goal when CSS was less than 0.2.

Furthermore, we also had to look at the contributions to the SDGs over the years based on the cosine similarity score. We averaged the CSS per year. As previously stated, each SDG is listed in order on the vertical axis, with the years on the horizontal axis. The size of the bubbles represents the average CSS for each goal over time.

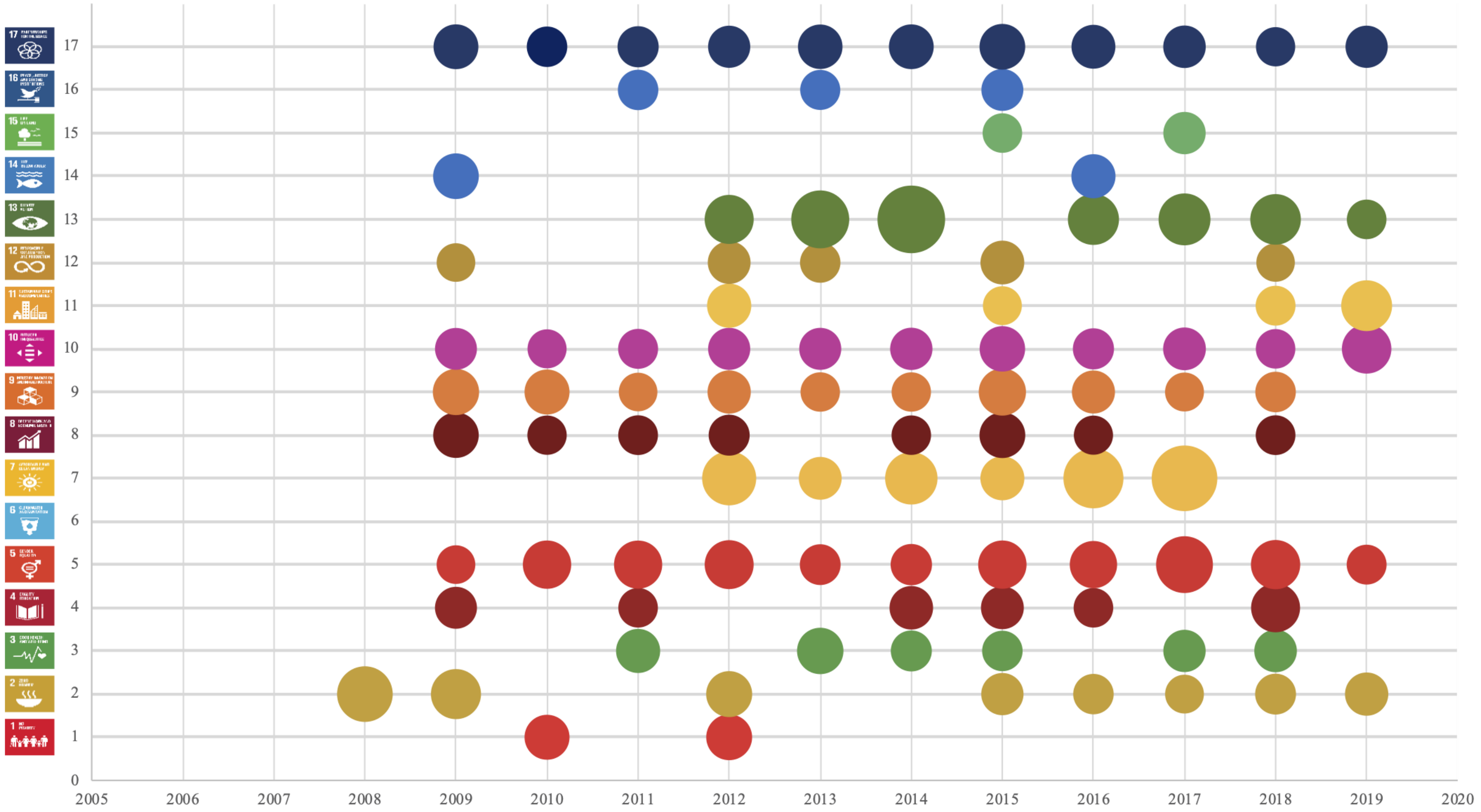

Figure 5.5: Average contributions to the SDGs over time - Cosine Similarity Scores

The contributions appear to be inconsistent based on the average cosine similarity score per year. Nevertheless, contributions to SDG 17, SDG 10, SDG 9 and SDG 5 appear to be stable over time. The figure also shows that there are more bubbles after 2015, which implies that the number of research publications considered to be relevant to the goals may have increased slightly after the adoption of the SDGs in 2015.

There appear to be no research publications relevant to the sustainability goals for several years, while others have larger bubbles. Larger bubbles can indicate one of two things: either that there were a lot of relevant research publications published in those years, or that there were fewer but more relevant research publications published in those years. For instance, SDG 13 has a large bubble in 2014, but this can be somewhat misleading because only one research publication was above the threshold that year. This is also the research publication with the highest CSS. Therefore, it is important to keep in mind that this graph depicts average scores, which reveal more about the importance or relevance of the research publications than the number of relevant research publications.

Apart from SDG 13, the diagram shows that continuous contributions have been made to the sustainability goals that we identified as the most discussed in Figure 5.2.

Because the lowest threshold was set at 0.2, we ended up with very few research publications that we could say with certainty were relevant to specific sustainability goals using cosine similarity. Despite the small number of research publications, we have found that cosine similarity produces results that adequately measure text-similarity for our purposes. Furthermore, we see that SDG 13, SDG 5, and SDG 9 are the most addressed SDGs, which is consistent with the fact that BI says they contribute to these goals through research centres and are focusing on in their sustainability strategy.

6 Conclusion

We can conclude that LDA-based methods and distance measures are useful to assess the alignment of published business school research with the United Nations' 17 Sustainable Development Goals.

We assessed the alignment of research publications published by BI-affiliated researchers with the 17 SDGs using an LDA-based methodology, the SDG classifier, and cosine similarity. Both methods successfully identified research publications relevant to the SDGs. For instance, we can confidently state that the top three research publications related to each of the sustainability goals we have discussed are indeed relevant to the associated goals. Thus, both methods produced reliable results in terms of identifying SDGs in the research publications.

However, compared to cosine similarity, the results showed that the SDG classifier was the most successful approach since it more accurately represented how the research publications were aligned with the SDGs. The SDG classifier assigned scores across a wide range. We observed that research publications were pertinent to the SDG when they earned a score of greater than 0.1. The results were not as obviously related to the assigned SDG when we used cosine similarity, where we established a threshold of 0.2 to ensure that the research paper was relevant to a goal. Therefore, it was concluded that cosine similarity could only be utilized to detect SDGs in research publications with more specific SDG material.

The results obtained with the SDG classifier indicate that the SDG classifier is capable of capturing SDG-related topics in research publications where they are not always prominent and may not even contain a plurality of words explicitly related to the goal for which the publication was classified as relevant.

Our results show that research at BI has made clear contributions to the SDGs, using both methods. 91.73% of the research publications were classified as relevant for the SDGs using the SDG scores, whereas 10.11% of the research publications were considered relevant for the goals using the cosine similarity.

Despite the substantial variations in the number of research publications considered relevant to the SDGs using the two methods, it is evident that BI contributes to the sustainability goals throughout the period we have analysed. Figure 5.3 showing SDG scores indicates that the contributions are consistent over time, while Figure 5.5 showing CSS indicates that the contributions are somewhat inconsistent. Nevertheless, Figure 5.5 also implies that the contributions might have increased after the adoption of the goals in 2005.

BI claims to contribute particularly to SDG 3, SDG 16, and SDG 17, which aligns with our findings to some extent (BI Norwegian Business School, 2022). According to the SDG classifier, SDG 17 is the most addressed goal. However, it is only the fourth most addressed goal according to the results using cosine similarity.

SDG 3 and SDG 16 are not among the top three results for any of the methods. The fact that BI's statements contradict our findings could be due to a variety of reasons. We believe that one significant reason is that we did not analyse all research publications from 2006 to 2019. Furthermore, there were only a few research publications from 2019 in our data, and none from 2020 to 2022. As a result, we were unable to analyse the most recent research publications and thus cannot claim that our findings are representative of the current situation.

Nevertheless, we can see that several of the same SDGs emerge as highly addressed when we look beyond the top three sustainability goals identified by both methods. For instance, when comparing the results, is it just one research publication that keeps SDG 17 from being among the top three addressed goals using both methods. SDG 3 and SDG 16 are only represented if we look at the top six goals that have been addressed.

Rank	**SDG Classifier**	**Cosine Similarity**
1	**SDG 17**	**SDG 13**
2	**SDG 9**	SDG 5
3	SDG 12	SDG 10
4	SDG 4	**SDG 17/SDG 9**
5	**SDG 13**	SDG 3
6	SDG 16	SDG 7

Table 6.1: Top six addressed goals using both methods

7 Research Limitations

Due to the lack of currently available NLP methodologies to identify SDG-related research publications, we were required to have an exploratory approach. While we applied LaFleur (2019) LDA-based method, we were evaluating a domain other than the one for which the model was designed. Furthermore, it was revealed throughout the discussion that the results of the two alternative methodologies used in the thesis produced very different conclusions. This could be related to the fact that LDA analyses topics and cosine similarity evaluates text similarity. They were, however, both used to assist us in better understanding how to identify SDG-related research publications. We note that, at this stage in development, the results should be interpreted with caution due to the differences in results.

Another limitation of this study's methodology is the thresholds used to determine which research publications are clearly relevant to an SDG. The thresholds were determined entirely subjectively by reading research publications that received scores ranging from 0 to 0.1. Given that a threshold strongly depends on the domain one is examining, creating a threshold based on a sample of documents could be a more satisfactory approach.

The study's third limitation is that the corpus of research publications does not include all research publications that have been published; just a sample was available for us to analyse. If we had looked at the complete corpus of research at BI Norwegian Business School, the outcomes might have been different. We have only analysed research publications published between 2006 and 2019. Additionally, our data only included a few research publications from 2019 and none from 2020 to 2022. As a result, we were unable to analyse the most recent research publications and thus cannot argue that our findings reflect the current situation.

8 Future Research

There is no commonly agreed approach for identifying SDG-related research publications, despite the fact that the SDG framework serves as a blueprint for aiding institutions in framing agendas and engagement (Armitage et al., 2020). Given that higher education has been recognised as crucial to the UN's 2030 agenda's success, this research imbalance is significant because it could work against the SDGs' success (United Nations a, 2022). We believe and hope that the findings in this thesis will assist universities and others interested in determining how their work or research aligns with the SDGs in developing new ideas for how LDA-based methods and text similarity measures like cosine similarity can be used to identify SDGs in research and provide insight into the challenges of implementing the SDGs.

We propose that the progress of research toward the SDGs be assessed in future studies, using approaches that analyse efforts in terms of their underlying themes and concepts. Each methodology could capture a unique understanding or translation of the SDGs. In the future, it will be necessary to considerably extend the corpus or to thoroughly examine various business schools in a given region. Then, to identify the direction of future study, we advocate conducting research that provides a deeper insight into the publishing culture. Recommendations are made to better integrate academic research with the Sustainable Development Goals, impacting how business school faculty and institutions prioritise research.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer.

Apeltsin, L. (2021). *Data Science bookcamp five python projects*. Manning.

Armitage, C. S., Lorenz, M., and Mikki, S. (2020). Mapping scholarly publications related to the sustainable development goals: Do independent bibliometric approaches get the same results? *Quantitative Science Studies*, 1(3):1092–1108.

Asatryan, Z., Feld, L. P., and Geys, B. (2015). Partial fiscal decentralization and sub-national government fiscal discipline: empirical evidence from oecd countries. *Public Choice*, 163(3):307–320.

Aurora Universities Network (2022). Sdg analysis: Bibliometrics of relevance.

Bali Swain, R. and Wallentin, F. Y. (2012). Factors empowering women in indian self-help group programs. *International review of applied economics*, 26(4):425–444.

Bankvall, L., Bygballe, L. E., Dubois, A., and Jahre, M. (2010). Interdependence in supply chains and projects in construction. *Supply chain management: an international journal*.

Beysolow, I. et al. (2018). Topic modeling and word embeddings. In *Applied Natural Language Processing with Python*, pages 77–119. Springer.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406.

BI Norwegian Business School (2022). Sustainability research.

Bjerkestuen, H. M. and Willie, H. G. (2015). Tax holidays in a beps-perspective. *Intertax*, 43:106.

Bjørnland, H. C., Aaen-Stockdale, C., Næss, K. M., and Zhulanova, J. (2019). Sustainability research at bi norwegian business school.

Blei, D., Carin, L., and Dunson, D. (2010). Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. ieee signal process mag 27 (6): 55–65.

Blei, D. M. (2012). Introduction to probabilistic topic models.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bø, E., Hammervoll, T., and Tvedt, K. (2013). Environmental impact of refillable vs. non-refillable plastic beverage bottles in norway. *International journal of environment and sustainable development*, 12(4):379–395.

Bygballe, L. E. and Ingemansson, M. (2011). Public policy and industry views on innovation in construction.

Clarivate (2019). Navigating the structure of research on sustainable development goals.

Ehling, P. and Ramos, S. B. (2006). Geographic versus industry diversification: Constraints matter. *Journal of Empirical Finance*, 13(4-5):396–416.

Eiling, E., Gerard, B., and De Roon, F. A. (2012). Euro-zone equity returns: country versus industry effects. *Review of Finance*, 16(3):755–798.

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77.

Gelain, P., Lansing, K. J., and Natvik, G. J. (2018). Leaning against the credit cycle. *Journal of the European Economic Association*, 16(5):1350–1393.

Greaker, M., Stoknes, P. E., Alfsen, K. H., and Ericson, T. (2013). A kantian approach to sustainable development indicators for climate change. *Ecological Economics*, 91:10–18.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Hauge, Å. L., Almås, A.-J., Flyen, C., Stoknes, P. E., and Lohne, J. (2017). User guides for the climate adaptation of buildings and infrastructure in norway–characteristics and impact. *Climate Services*, 6:23–33.

Hoballah, A. and Averous, S. (2015). Ensure sustainable consumption and production patterns. *UN Chronicle*, 51(4):28–29.

Hsu, D. F., LaFleur, M. T., and Orazbek, I. (2022). Improving sdg classification precision using combinatorial fusion. *Sensors*, 22(3):1067.

Jain, A., Jain, A., Chauhan, N., Singh, V., and Thakur, N. (2017). Information retrieval using cosine and jaccard similarity measures in vector space model. *Int. J. Comput. Appl*, 164(6):28–30.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Kim, D., Seo, D., Cho, S., and Kang, P. (2019). Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information Sciences*, 477:15–29.

Krause, E. F. (1986). *Taxicab geometry: An adventure in non-Euclidean geometry*. Courier Corporation.

LaFleur, M. (2019). Art is long, life is short: An sdg classification system for desa publications.

Lahitani, A. R., Permanasari, A. E., and Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6. IEEE.

Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., and Baldi, P. (2007). Mining concepts from code with probabilistic topic models. In *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*, pages 461–464.

Mallet (2021). Machine learning for language toolkit.

Maurseth, P. B. (2019). Brain drain or brain gain? In *Forum for Development Studies*, volume 46, pages 195–202. Taylor & Francis.

Nguyen, H. T. H., Bertsch, A., Warner-Søderholm, G., and Ondracek, J. (2017). Understanding cultural factors which affect women serving on boards of directors.

NLTK (2022). Natural language toolkit.

Ozsoy, M. G., Alpaslan, F. N., and Cicekli, I. (2011). Text summarization using latent semantic analysis. *Journal of Information Science*, 37(4):405–417.

Prenkert, F. and Hallén, L. (2006). Conceptualising, delineating and analysing business networks. *European Journal of Marketing*.

Priva, U. C. and Austerweil, J. L. (2015). Analyzing the history of cognition using topic models. *Cognition*, 135:4–9.

Ristanti, P. Y., Wibawa, A. P., and Pujianto, U. (2019). Cosine similarity for title and abstract of economic journal classification. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 123–127. IEEE.

Ràfols, I. (2020). Consensus and dissensus in 'mappings' of science for sustainable development goals (sdgs).

Sasson, A. and Reve, T. (2015). Complementing clusters: a competitiveness rationale for infrastructure investments. *Competitiveness Review*.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Scikit-learn (2022). Tfidfvectorizer.

SDSN Northern Europe (2022). Sdgs in universities.

Singh, R. H., Maurya, S., Tripathi, T., Narula, T., and Srivastav, G. (2020). Movie recommendation system using cosine similarity and knn. *International Journal of Engineering and Advanced Technology*, 9(5):556–559.

Stoknes, P. E. (2014). Rethinking climate communications and the "psychological climate paradox". *Energy Research & Social Science*, 1:161–170.

Strønen, F., Hoholm, T., Kværner, K. J., and Støme, L. N. (2017). Dynamic capabilities and innovation capabilities: The case of the 'innovation clinic'. *Journal of Entrepreneurship, Management and Innovation*, 13(1):89–116.

Traavik, L. E. (2018). Career equality: Inclusion and opportunities in a professional service firm in norway. *Gender in Management: An International Journal*.

Triwijoyo, B. K. and Kartarina, K. (2019). Analysis of document clustering based on cosine similarity and k-main algorithms. *Journal of Information Systems and Informatics*, 1(2):164–177.

United Nations a (2022). The 17 goals | sustainable development.

United Nations b (2022). Goal 9: Build resilient infrastructure, promote sustainable industrialization and foster innovation.

United Nations c (2022). Goal 12: Ensure sustainable consumption and production patterns.

United Nations d (2022). Goal 4: Quality education.

United Nations e (2022). Goal 13: Take urgent action to combat climate change and its impacts.

United Nations f (2022). Goal 5: Achieve gender equality and empower all women and girls.

United Nations g (2022). Goal 10: Reduce inequality within and among countries.

United Nations Statistics Division (2021). The sustainable development goals report 2021.

Vayansky, I. and Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94:101582.

Walgermo, B. R., Foldnes, N., Uppstad, P. H., and Solheim, O. J. (2018). Developmental dynamics of early reading skill, literacy interest and readers' self-concept within the first year of formal schooling. *Reading and writing*, 31(6):1379–1399.

Wastl, J., Porter, S., Draux, H., Fane, B., and Hook, D. (2020). Contextualizing sustainable development research. *Digit. Sci.*

Xia, P., Zhang, L., and Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, 307:39–52.

Appendix

A1 SDG over time figures - Numbers

Year	SDG 1	SDG 2	SDG 3	SDG 4	SDG 5	SDG 6	SDG 7	SDG 8	SDG 9	SDG 10	SDG 11	SDG 12	SDG 13	SDG 14	SDG 15	SDG 16	SDG 17
2006	0.14	0.17	0.10	0.12	0.00	0.00	0.11	0.18	0.19	0.10	0.00	0.21	0.13	0.00	0.00	0.00	0.14
2007	0.00	0.00	0.20	0.16	0.11	0.14	0.00	0.14	0.14	0.22	0.00	0.16	0.11	0.18	0.00	0.15	0.14
2008	0.11	0.20	0.20	0.18	0.13	0.13	0.00	0.13	0.14	0.12	0.00	0.17	0.13	0.14	0.12	0.18	0.13
2009	0.16	0.19	0.00	0.17	0.18	0.14	0.15	0.13	0.15	0.11	0.00	0.17	0.13	0.12	0.11	0.18	0.12
2010	0.15	0.15	0.00	0.14	0.15	0.12	0.19	0.15	0.14	0.13	0.16	0.16	0.14	0.14	0.13	0.21	0.13
2011	0.13	0.15	0.27	0.19	0.15	0.12	0.13	0.16	0.15	0.12	0.13	0.14	0.12	0.12	0.00	0.18	0.14
2012	0.16	0.17	0.15	0.14	0.14	0.12	0.17	0.12	0.16	0.12	0.16	0.16	0.14	0.12	0.14	0.22	0.15
2013	0.14	0.15	0.28	0.17	0.14	0.12	0.14	0.15	0.14	0.14	0.13	0.14	0.13	0.12	0.11	0.18	0.14
2014	0.13	0.12	0.23	0.17	0.16	0.14	0.13	0.14	0.15	0.11	0.10	0.14	0.13	0.11	0.11	0.19	0.13
2015	0.12	0.12	0.21	0.16	0.17	0.13	0.15	0.15	0.15	0.16	0.10	0.15	0.14	0.12	0.17	0.16	0.15
2016	0.11	0.15	0.24	0.18	0.16	0.14	0.15	0.15	0.15	0.14	0.13	0.14	0.14	0.13	0.14	0.16	0.14
2017	0.11	0.15	0.22	0.17	0.17	0.14	0.15	0.16	0.16	0.11	0.13	0.13	0.13	0.13	0.13	0.17	0.15
2018	0.13	0.16	0.21	0.18	0.16	0.11	0.17	0.14	0.14	0.13	0.13	0.14	0.15	0.12	0.15	0.16	0.15
2019	0.14	0.11	0.00	0.14	0.18	0.14	0.16	0.14	0.13	0.12	0.13	0.13	0.12	0.13	0.00	0.16	0.15

Table 8.1.1: Appendix - Figure 5.3: Average contributions to the SDGs over time, SDG Scores

Year	SDG 1	SDG 2	SDG 3	SDG 4	SDG 5	SDG 6	SDG 7	SDG 8	SDG 9	SDG 10	SDG 11	SDG 12	SDG 13	SDG 14	SDG 15	SDG 16	SDG 17
2006	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2007	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2008	0.00	0.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2009	0.00	0.34	0.00	0.24	0.20	0.00	0.00	0.28	0.29	0.24	0.00	0.20	0.00	0.28	0.00	0.00	0.27
2010	0.27	0.00	0.00	0.00	0.32	0.00	0.00	0.21	0.27	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.22
2011	0.00	0.00	0.26	0.21	0.31	0.00	0.00	0.21	0.20	0.22	0.00	0.00	0.00	0.00	0.00	0.22	0.23
2012	0.28	0.29	0.00	0.00	0.32	0.00	0.40	0.23	0.25	0.24	0.26	0.25	0.33	0.00	0.00	0.00	0.24
2013	0.00	0.00	0.29	0.00	0.23	0.00	0.25	0.00	0.21	0.24	0.00	0.22	0.45	0.00	0.00	0.22	0.27
2014	0.00	0.00	0.23	0.26	0.23	0.00	0.37	0.21	0.21	0.24	0.00	0.00	0.62	0.00	0.00	0.00	0.27
2015	0.00	0.24	0.22	0.25	0.32	0.00	0.26	0.28	0.30	0.28	0.20	0.26	0.00	0.00	0.21	0.24	0.28
2016	0.00	0.22	0.00	0.21	0.30	0.00	0.49	0.21	0.25	0.23	0.00	0.00	0.35	0.26	0.00	0.00	0.26
2017	0.00	0.20	0.24	0.00	0.44	0.00	0.58	0.00	0.20	0.25	0.00	0.00	0.36	0.00	0.24	0.00	0.24
2018	0.00	0.23	0.25	0.32	0.33	0.00	0.00	0.21	0.22	0.21	0.21	0.20	0.35	0.00	0.00	0.00	0.21
2019	0.00	0.25	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.33	0.35	0.00	0.21	0.00	0.00	0.00	0.24

Table 8.1.2: Appendix - Figure 5.5: Average contributions to the SDGs over time, CSS