Topic Taxonomy Construction from ESG Reports

Saif Alnajjar[1], **Xinyu Wang**[2], **Yulan He**[1,3]
[1]King's College London [2]University of Warwick [3]The Alan Turing Institute
saif.alnajjar@kcl.ac.uk, Xinyu.Wang.11@warwick.ac.uk, yulan.he@kcl.ac.uk

Abstract

The surge in Environmental, Societal, and Governance (ESG) reports, essential for corporate transparency and modern investments, presents a challenge for investors due to their varying lengths and sheer volume. We present a novel methodology, called MultiTaxoGen, for creating topic taxonomies designed specifically for analysing the ESG reports. Topic taxonomies serve to illustrate topics covered in a corpus of ESG reports while also highlighting the hierarchical relationships between them. Unfortunately, current state-of-the-art approaches for constructing topic taxonomies are designed for more general datasets, resulting in ambiguous topics and the omission of many latent topics presented in ESG-focused corpora. This makes them unsuitable for the specificity required by investors. Our method instead adapts topic modelling techniques by employing them recursively on each topic's local neighbourhood, the subcorpus of documents assigned to that topic. This iterative approach allows us to identify the children topics and offers a better understanding of topic hierarchies in a fine-grained paradigm. Our findings reveal that our method captures more latent topics in our ESG report corpus than the leading method and provides more coherent topics with comparable relational accuracy.

Keywords: Text Mining, Text Analytics, Document Classification, Text categorisation, Knowledge Discovery/Representation, Topic Detection and Tracking

1. Introduction

Environmental, Societal, and Governance (ESG) reports are a type of report that companies release to discuss their plans and performance in, as the name suggests, *environmental*, *societal*, and *governance* issues. As the world shifts towards transparency and accountability, ESG reports serve as an indispensable resource for stakeholders, especially given the dramatic 27-fold increase in socially responsible investing (SRI) assets over 25 years (Christiansen et al., 2023).

However, with the rising importance of ESG reporting, as well as a recent EU directive that mandates larger European companies to publish ESG reports, there has been a significant upswing of companies issuing such reports, which can vary in lengths, spanning from a few pages to several hundred pages. The proliferation of ESG reports poses a challenge for investors who need to review them when making investment decisions.

As such, one analytical approach that can help investors and consumers is the creation of a topic taxonomy for a collection of ESG reports. A topic taxonomy is a hierarchical structure that displays the relationship between topics within a corpus. Each topic could serve as a parent to one or more subtopics, forming a structured hierarchy. Figure 1 shows an example of what a topic taxonomy looks like. Within each topic, a list of relevant terms represents the overarching concept, and a primary term is selected from that list to represent the topic in the taxonomy.

However, current state-of-the-art taxonomy methods, namely TaxoCom (Lee et al., 2022a), are often tailored for more general datasets, and as such falter with the distinct nuances of ESG reports. As depicted in Table 4, their extracted topics often emerge ambiguous or overly broad, missing many of the latent topics in the corpus, making the result barely usable for investors, who usually prefer a much deeper level of information. Not only that, these methods also use a phrase mining tool, like AutoPhrase (Shang et al., 2017), to get a list of potential terms, and thus some terms that are relevant but in lower frequency are missed, while at the same time non-ESG terms are also mined, creating some noise and worsening the results.

Recognising these limitations, we propose a novel method, called MultiTaxoGen, that leverages topic modeling techniques to better capture the intricacies of ESG reports, and heavily adapt and optimise them for building a topic taxonomy for our corpus. The main idea is to recursively run the topic modeling technique on every topic's local neighbourhoods based on the idea of local embeddings used in previous topic taxonomy works (Lee et al., 2022a; Shang et al., 2020; Zhang et al., 2018), to find its subtopics. Local neighbourhoods refer to the subcorpus of documents that were assigned to the current topic.

We modify the topic modeling technique to suit each level of the taxonomy to find more generalised topics in the second-level, and more specific and focused topics in the bottom-level. Unfortunately, these topic modeling techniques, in general, have no hierarchical understanding of our topics, so we create embeddings for the topics and compare them in the taxonomy and remove any deemed as outliers or redundant. We also improve the efficacy of assigning documents at the top-level by using an ESG classifier, giving better results downstream

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 178–187

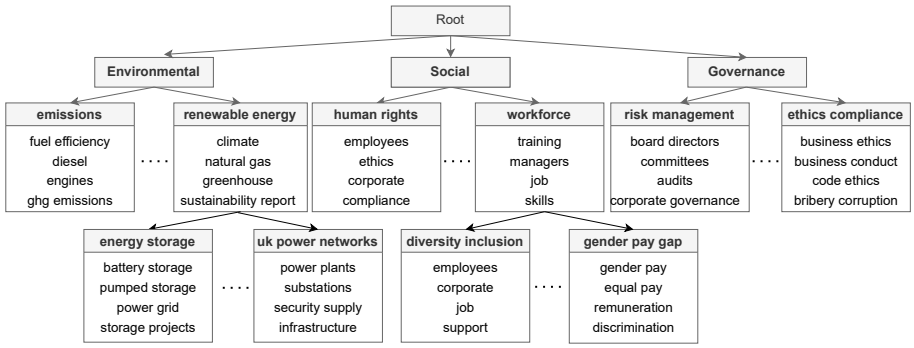

Figure 1: A sample from our constructed three-level topic taxonomy, featuring the top-level topics along with their respective subtopics, and the terms associated with those subtopics.

due to less documents being misassigned to the incorrect local neighbourhood.

Our main contributions are two-fold:

- We introduce a three-level framework for ESG reporting taxonomy. At each level, we employ tailored strategies adapted to the specific text and topic granularity.
- We conduct comprehensive experiments and evaluations of our method, including human assessments. The experimental results show that our method captures more latent topics than the leading method and provides more coherent topics.

2. Related Work

Topic Modeling Topic taxonomy construction and topic modeling are, naturally, very similar, and so a lot could be learned from topic modeling, especially since it is a widely studied field with many methods being researched (Blei et al., 2003; Angelov, 2020; Grootendorst, 2022; Bianchi et al., 2021). The two most common types of methods are statistical and neural topic models. The most widely used method, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), is one such example of a statistical method. Early hierarchical topic modelling approaches built on LDA (Blei et al., 2010; Kim et al., 2012) were proposed for the discovery of topical hierarchies within the abstracts of scientific papers. In such models, each document is presumed to be linked to a path where each level represents a topic. The assignment of paths adheres to an nested Chinese Restaurant Process (nCRP) or recurrent CRP prior. Additionally, Paisley et al. (2014) proposed a nonparametric model called the nested hierarchical Dirichlet process, enabling the incorporation of shared groups among clusters, thus extending the capabilities of the nCRP model through the incorporation of a hierarchical Dirichlet process. The primary challenges associated with hierarchical topic modeling methods include the complexity of incorporating prior knowledge about topics and their dependency on having access to the complete vocabulary of the corpus.

More recent literature, however, suggests that neural topic models, namely those that use embedding techniques like BERT (Devlin et al., 2019) or Word2Vec (Mikolov et al., 2013), can outperform the classic topic modeling techniques, with BERTopic (Grootendorst, 2022) and CTM (Bianchi et al., 2021) being some examples.

Topic Taxonomy Generation Topic taxonomy generation primarily follows two approaches: from scratch and seed-guided. The former constructs taxonomies without any prior knowledge of the taxonomy and just relying on the corpus. Seed-guided, a more weakly supervised approach, uses an initial seed taxonomy in addition to the corpus to nudge the generated topics towards that seed. Currently, most of the highest-performing methods in either approach rely on what they call "local embeddings" (Lee et al., 2022a; Shang et al., 2020; Zhang et al., 2018). To improve the granularity of the embedding space when adding children topics to a parent topic, we create a subcorpus of documents that are related to that parent topic, and train an embedding, like Word2Vec (Mikolov et al., 2013), on that subcorpus, instead of using a global embedding that was trained on the entire corpus for all the children to be added (Lee et al., 2022a). Since the documents in the subcorpus are clustered to find the new subtopics, having different subcorpora for each of the topics can make the embeddings more discriminative and ultimately improve results. One other promising seed-guided approach is TopicExpan (Lee et al., 2022b), which out-performs all the other taxonomy generation method, but is a supervised method that requires all the documents in the corpus to be labelled with a term and topic related to that document.

ESG Baier et al. (2020) develops a word list for ESG topics and a corresponding taxonomy, then analyzes the distribution of these topics in ESG reports to determine their prevalence. This expert-curated taxonomy is valuable as it gives us a good

starting point for the seed we will be using. Meanwhile, FinBERT (Huang et al., 2023) further pretrains BERT on a corpus of financial documents, improving its performance in the financial domain. Most relevant to us though, the authors fine-tune FinBERT for classifying a document as Environmental, Social, and Governance, achieving state-of-the-art performance for ESG classification.

3. Preliminary - BERTopic

Our proposed approach is built on BERTopic (Grootendorst, 2022). Throughout this paper, we choose the topic modeling method of BERTopic (Grootendorst, 2022) as our primary focus, alongside corresponding experiments. In this section, we give an overview of the BERTopic method, which consists of three steps: document embedding generation, document clustering, and topic term extraction.

Document Embedding Generation First, document embeddings are generated using a language model. A common choice for this task is the Sentence Transformers (Reimers and Gurevych, 2019), which have been fine-tuned for document embedding generation. Following this, dimensionality reduction is performed on the embeddings using techniques such as Principal Component Analysis (PCA) or Uniform Manifold Approximation Projection (UMAP) (McInnes et al., 2020). Dimensionality reduction accelerates the model and also mitigates the curse of dimensionality (Keogh and Mueen, 2017), prior to proceeding with the subsequent step of the pipeline, clustering.

Document Clustering BERTopic then clusters the reduced document embeddings, and each cluster would thus count as a topic. Clustering is of particular importance for the topic taxonomy generation, as changes in the cluster size and clustering algorithm can allow for either more specific or more general topics.

Topic Term Extraction The final step in the pipeline is to extract the top terms of each topic based on the class-specific TF-IDF scores, or c-TF-IDF. To do this, all the documents in a cluster are combined to form a single document and a term-document matrix is formed for all the newly created documents. Then, the c-TF-IDF score of a term **w** in a cluster **c** is calculated using Equation (1).

$$score_{w,c} = tf_{w,c} \times \log(1 + \frac{n\_avg\_words}{tf_w}) \quad (1)$$

The highest scoring words/terms, usually the top 10, are thus used to represent the topic. The scoring mechanism naturally favour terms that appear frequently in a certain cluster while being less common in others. Thus, in the case of larger clusters that encompass more documents, the scoring tends to emphasise more general or overarching terms, as one would anticipate in higher levels of a taxonomy. On the other hand, when the clustering algorithm is forced to generate as many clusters as possible, leading to smaller clusters that ultimately represent all potential topics, the highest scored terms tend to be more specific and focused.

4. Methodology

We propose a multi-level topic taxonomy generation approach, named as MultiTaxoGen, as shown in Figure 2. At the first level, documents are segregated into three main topics: *Environmental*, *Social*, and *Governance*. We deploy a classifier to partition all documents to each of those topics and split the corpus into three distinct subcorpora. Next, on each subcorpus, we utilise a topic modeling technique, specifically BERTopic (Grootendorst, 2022), to search for a small number of topics. However, it is pertinent to note that alternative neural topic modelling techniques are also available, such as Top2Vec (Angelov, 2020).

This methodology echoes the ideas propounded by preceding studies (Zhang et al., 2018; Shang et al., 2020; Lee et al., 2022a), where a local embedding is trained on a topic-specific subcorpus. Our approach, instead, involves operating BERTopic on what can instead be called a local neighbourhood of documents rather than training a local embedding. The topics found from each of the subcorpora would thus constitute the second level of our taxonomy. Then, BERTopic is rerun on the documents under each newly discovered topic, allowing BERTopic to find as many topics as possible. Finally, redundant or unrelated topics are then merged or removed respectively. This would thus establish the third and bottom level of the topic taxonomy.

4.1. Local Neighbourhoods

Previous works (Zhang et al., 2018; Shang et al., 2020; Lee et al., 2022a) on creating topic taxonomies have found great success in training a local embedding for each sub-corpus of documents in a topic, rather than using one global embedding training on the entire corpus, allowing for better granularity and discriminativeness between embeddings, and ultimately improved performance when finding subtopics.

Rather than training our own embeddings, which would require a massive corpus for training transformers, we simply run BERTopic separately for each topic's subcorpus, and the child topics found would be more tuned towards the parent topic with

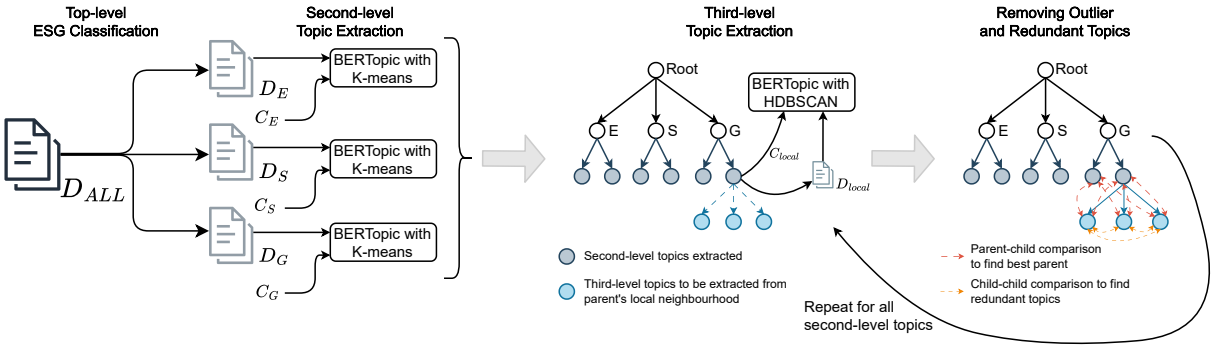

Figure 2: An illustration of MultiTaxoGen, where D refers to a corpus of documents and C is a set of topics. In the context of this illustration, C is the seed topics. Initially, an ESG Classifier split documents in our full corpus into three topics. Subsequently, the second-level topics are extracted on the corpus of each of them. Finally, we use BERTopic on each second-level topic with HDBSCAN to find their children topic. We further remove outlier topics with the parent-child comparisons to find the correct parent, and merge redundant topics by comparing them with the other children and then merging them.

the c-TF-IDF scoring, along with some modifications and optimizations.

4.2. First Level

The top-level topics of any ESG report will, naturally, be *Environmental*, *Social*, and *Governance*. Splitting our corpus into three separate subcorpora gives us the advantage of having more focus on subcorpus for each of the topics when we find the subtopics in the subsequent steps. To facilitate this division, we employ the FinBERT-ESG (Huang et al., 2023) classifier to assign all the documents into either one of the three subtopics or a"none" class if they lack relevance to any of the primary topics. It was reported in (Huang et al., 2023) that the classifier achieves an accuracy of 89.5% on a small set of ESG-related discussions. We then filter documents which have a probability of less than a threshold $\tau_c = 0.7$. The majority of the filtered documents consist of tables and numerical data found in report appendices, which fall outside the scope of our primary focus, or irrelevant documents that can lead to non-ESG related topics being extracted.

4.3. Second Level

To find the second-level topics, i.e. the children of the top-level topics, *Environmental*, *Social*, and *Governance*, we run BERTopic on each of the top-level topic's subcorpus, while using k-means as the clustering algorithm to guide BERTopic in identifying a limited set of clusters by setting the number of clusters k to a small value.

The goal of this is to create large clusters with many documents that discuss many different topics, but all share a certain high-level topic in each cluster. Thus, the highest scoring terms will be those that match that high-level topic and will typically be more general and less focused, while also being inherently related to the parent topic, since they are derived from documents assigned to their parent.

To nudge the generated topics towards our seed, BERTopic takes in a seed of topics with their potential terms and then steers the c-TF-IDF scoring of the terms in the clusters towards those seed topics by applying a multiplier to the score if a term is related any of the seed topics. As a seed, we use the curated ESG topic taxonomy created by Baier et al. (2020), and also remove any topics that rarely appear in their corpus. This approach ensures that the top terms chosen have a strong relevance to ESG topics.

A main term will also need to be selected to represent that topic in the taxonomy, and for our case, we simply select the highest scoring term in the cluster as the main term, as that term typically represents the topic.Table 1 shows some second-level extracted topics, showcasing how the topics clearly relate to the parent topic, *Environmental*.

	Topic's Top 3 Terms	Main Term
T1	sustainability report, environmental management, ghg emissions	sustainability report
T2	water consumption, wastewater, groundwater	water consumption
T3	waste management, recycling, hazardous waste	waste management
T4	greenhouse gas emissions, scope emissions, energy consumption	greenhouse gas emissions

Table 1: Sample second-level topics, their terms, and their main terms for the parent and top-level topic *Environmental*

4.4. Third Level

Topics in the third or bottom level are expected to be more specific compared to their parent's. As such, rather than forcing it to find a certain arbitrary number of clusters like in the previous level, finding as many topics as possible is a more optimal alternative, as all the latent topics need to captured at this level. Therefore, we employ Hierarchical DBSCAN (HDBSCAN) (Malzer and Baum, 2020), which in itself is an extension of the popular DBSCAN (Ester et al., 1996), to find all potential clusters of all different sizes. We set the minimum cluster size as the minimum document count required for a topic to be formed in order to modify the number of topics that are found, and this number is based on the size of the second-level topic's local neighbourhood.

Compared to the topics in the previous level, we cannot predetermine a list of potential seed topics for the third level due to the variability inherent in the second-level topics. Instead, we take the top 5 scoring terms of the parent topic as the seed topics when identifying the subtopics of that parent. The highest scoring terms of each topic's top terms are then again selected as the main term. However, due to the high number of topics in the bottom level, many topics may share the same highest-scoring term. In this case, the next highest-scoring term that is not the main term of any other topic is selected as the main term.

One issue in extracting a large number of topics, especially when using HDBSCAN and considering misassigned documents, is the emergence of irrelevant and redundant (where two or more topics can be very similar or even exactly the same as each other) topics. Additionally, topic modeling methods such as BERTopic do not take any hierarchy into account, other than us applying it to a local neighbourhood of documents of a parent topic. It consequently cannot discern when an irrelevant topic is extracted. To address these issues, further optimisations are required to remove or merge the unnecessary topics.

As a final note, we could, in theory, repeat this same process again with the third-level to get a fourth-level, however we opted to stop at three levels for several reasons. At the third level, topics become exceedingly specific, making it challenging to extract meaningful latent topics from the documents associated with third-level topics, as they often revolve around very similar subject matter. Additionally, for the sake of consistency in comparisons with other methods, a three-level taxonomy appears to be more suitable, as the majority of related papers on topic taxonomy construction primarily employ two or three-level hierarchies

Topic Embeddings Generation To determine the necessity of a topic, we convert the topics into embeddings to properly compare different topics. We represent the top 5 terms of a topic as its embedding by using any word embeddings methods. However, using context-independent word embeddings such as GloVe (Pennington et al., 2014), leads to the out-of-vocabulary problem, and multi-word terms would need to be found using the less-than-ideal workaround of calculating the average embedding of their words. Therefore, we instead employ the Sentence Transformers (Reimers and Gurevych, 2019), the same embedding model used to represent our documents. Even though not explicitly trained for this task, the embeddings generated by it are still fairly good and better than the ones found using GloVe embeddings. It also brings the added benefit of being able to directly compare documents with the topic embedding. when assigning relevant topics to documents. We average the embeddings of the topic's top 5 terms, generated by the embedding model, based on the term's score, where the highest-scoring terms hold a higher weight.

Removing Outlier Topics To minimize the number of unrelated topics, we initially check the generated topic embeddings for the third-level topic, as well as embeddings for the second-level topics, including its parent, by comparing the cosine similarity of the third-level topic with each of the second-level topics. If the second-level topic most similar to a third-level topic does not correspond to its parent, the third-level topic is removed, and its associated documents are temporarily marked as outliers. This procedure is reiterated for all the third-level topics.

Merging Redundant Topics Similarly, the redundant topics should be removed as well by merging all the redundant topics into one topic. We first generate the topic embeddings with the embedding model for each of the third-level topics of one of the second-level topics. Next, each third-level topic embedding is compared with all the other third-level topic embeddings by their cosine similarity, creating a similarity matrix. If the similarity between one topic embedding and another is greater than a threshold τ_r, then those topics are merged, meaning the topics are combined by putting the documents of each of the two topics into one topic. In our case, we set $\tau_r = 0.8$ as that was found to be the optimal value in the experiments. The threshold in our case was set to a high value because most of the topics found at the bottom level will have a high similarity score between them since those topics are inherently similar as they share the same parent topic. Finally, we repeat this process for the third-level topics of all the second-level topics separately.

4.5. Assigning Documents to Multiple Topics

Normally, with BERTopic, documents are assigned to only one single topic, rather than all relevant topics. As such, after finding all the topics and creating the topic taxonomy, we attempt to identify each topic within the taxonomy that holds relevance to a specific document.

The first step is the identification of which top-level topic a document is associated with via the ESG classifier, where we assume that only one top-level topic is in each document, as almost documents are focused on only one of the top-level topics. For documents categorized under top-level topics, we identify their corresponding second-level topics using the BERTopic models that have been previously trained for the respective top-level topics.

Then, the remaining potential topics are found by calculating the cosine similarity between the embeddings of documents in the top-level topics and their children (second-level) topic embeddings. Each document's topic assignment is determined by a threshold τ_t. However, a single universal threshold may not be optimal for all topics. Therefore, we designate distinct thresholds τ_t tailored to each specific topic by calculating:

$$\tau_t = \mu_s + (1.5 \times \sigma_s) \tag{2}$$

where μ_s and σ_s denotes the mean of similarities of the second-level topics and is their standard deviation, respectively. Finally, we classify the documents to the third-level topics. For the documents assigned to the parent (second-level) topics, we again compute the cosine similarity with the third-level topic embeddings and find the relative thresholds as described above. To be noticed that a document can also be assigned to more than one second-level topic and thus be compared and checked multiple times. Third-level topics differ slightly from second-level topics in that a document could potentially not be assigned any third-level topics, as HDBSCAN may mark a document as an outlier if it does not match with any topics.

5. Experiments

5.1. Experimental Setup

Dataset We collect 10,645 publicly available ESG reports released from 1992 to 2022 across 2,001 companies from ResponsibilityReports.com[1]. The reports are in PDF format, and we extract text content from ESG reports using PyMuPDF[2]. Next, we split, as best as we can, the reports into paragraphs of a maximum length of 256 words to constitute a total of 1,208,546 documents after splitting, and these would be considered the documents of our corpus.

Splitting them into paragraphs shorter than 256 words is necessary, as BERT models typically have a maximum length of 512 tokens. Also, while a document may encompass various subjects, when we assign a document to a specific topic (as a part of its local neighborhood), we assign it to only the topic most relevant to it, as was done in previous works (Lee et al., 2022a). Naturally, shorter documents will end up having less topics being discussed in them.

Baselines We will be comparing our method with the current state-of-the-art weakly-supervised method, TaxoCom (Lee et al., 2022a). We split our corpus into three subcorpora for each of the top-level topics, and run TaxoCom separately for each of those.

Hyperparameter Setup The document embedding model we use is MiniLM (Wang et al., 2020) comprising 6 layers trained in accordance with the Sentence Transformers (Reimers and Gurevych, 2019) paradigm. The number of clusters k topics, when extracting the second-level topics, for the three top-level topics are $k_{environmental} = 8$, $k_{social} = 8$, and $k_{governance} = 5$. For TaxoCom (Lee et al., 2022a), we set $\beta_1 = 3.5$ and $\beta_2 = 6.0$ for the second and third levels respectively, which controls how many novel topics are found, and keep the other parameters the same as in the paper.

5.2. Evaluation

Considerable research has been done to try and automatically evaluate topic coherence. Some measures have been shown to correlate with humans quite well (Lau et al., 2014) and are commonly used when evaluating topic models, namely NPMI (Bouma, 2009) and C_v (Röder et al., 2015). Other works, however, suggest that although classical topic models like LDA (Blei et al., 2003) do correlate, they may not do so with neural topic models (Hoyle et al., 2021). Ultimately, we opted for human evaluation to get the most accurate results, but have included the C_v scores in the results as well. In particular, we employ two metrics to compare the methods, *topic coherence* and *relation accuracy*. The evaluations of these metrics has been carried out by 3 computer science graduates, who were paid hourly rate of £20 for their evaluation of the methods. We then average out their results to minimize human bias.

Topic Coherence The first metric tries to measure how "coherent" a topic is by how clearly the

[1] https://www.responsibilityreports.com/

[2] https://pymupdf.readthedocs.io/

	Total Number of Topics		Topic Coherence		C_v		Relation Accuracy	
	MultiTaxoGen	TaxoCom	MultiTaxoGen	TaxoCom	MultiTaxoGen	TaxoCom	MultiTaxoGen	TaxoCom
Environmental	174	12	**0.902**	0.731	0.521	**0.526**	0.865	**0.910**
Social	272	23	**0.910**	0.732	0.533	**0.715**	0.831	**0.917**
Governance	51	22	**0.917**	0.790	0.407	**0.711**	**0.895**	0.883
All	500	60	**0.908**	0.754	0.487	**0.651**	0.850	**0.899**

Table 2: Results of our method, MultiTaxoGen, and TaxoCom. Best results in each taxonomy are in bold.

.

Parent Topic	Outlier Topic's Top 3 Terms	Most Similar Parent
renewable energy	**air filters**, indoor air quality, air filtration	greenhouse gas emissions
charity	**consumer credit**, credit history, experian	local communities
corporate governance	**auditing standards**, auditor report, statutory sustainability	audit committee

Table 3: Examples of outlier topics found, their parent topic, and the parent topic that found to be most similar to them. Main term of the outlier topic is highlighted in bold.

.

set of terms in a topic represents a recognisable overarching topic or category (Lund et al., 2019). By definition, this is inherently a subjective measure, as one person may see a certain set of terms as more coherent compared to another person that may see those set of terms to have a different meaning. For the human evaluation, the topic coherence score of each topic is calculated by counting the number of terms in the topic that do belong in that topic, and are then averaged. Next, all the topic coherence scores are averaged as well to get the average topic coherence of the method.

Relation accuracy Relation accuracy tries to evaluate the accuracy of the relationships among the child, parent, and grandparent topics. This is also human evaluated. To find the relation accuracy of a topic, the topic is compared to its parent. If the parent-child relationship is correct, it is given a score of 1. However, if they do not match but the child matches the grandparent, it is given a score of 0.25 instead. If it does not match any of them, then the relation accuracy of that topic is 0. The final relation accuracy of the method is found by averaging the accuracy of all the topics.

5.3. Quantitative Results

The results have been split into the three taxonomies for each of the top-level topics so we provide a deeper look into the results. Table 2 shows a comparison of the results. We can observe that our method found a much larger number of topics, almost ten-fold, compared to TaxoCom, which managed to only find a total of 60 topics. Considering our corpus of more than a million documents, 60 topics do seem to be considerably lower than expected. As we show later, the topics extracted by TaxoCom also are more vague, even in the lower levels. In contrast, our method manages to extract more specific topics at the bottom level.

Our method gives significantly more coherent topic across the board, achieving an average topic coherence of above 0.9. Conversely, TaxoCom achieves a higher relation accuracy, though in all taxonomies, the results are still close, with only a small difference between the two methods. TaxoCom also achieves a higher C_v score, though it is important to be noticed the limitations of automatic topic coherence measures, as described in Section 5.2

5.4. Case Study

We present examples to explore the effects of different parts of our method, as well as the results from both our final taxonomy and TaxoCom's.

Outlier Topics One can anticipate the emergence of irrelevant topics when employing a topic modeling technique, particularly due to their lack of capability to identify hierarchical relationships within a taxonomy. Consequently, if documents unrelated to the parent topic are found in its subcorpus, it could result in the formation of a cluster for those documents, thus generating an outlier topic. Our approach, described in Section 4.4, enhances the model's understanding of the hierarchy, enabling it to detect and remove any outlier topics that do not align with their parent topic. Approximately 54% the initial topics were subsequently removed. Table 3 showcases some of these outlier topics, which were subsequently flagged and removed.

Parent Topic	Sub-topics Generated by MultiTaxoGen	Sub-topics generated by TaxoCom
Renewable Energy	natural gas, greenhouse gas (⊗), sustainable development, red electrica, uk power networks, oil sands, mining (⊗), cenovus, demand response programs, energy storage systems...	vehicles, environmental (⊗)
labor	diversity inclusion, benefits, social responsibility (⊗), innovation, employability, gender paygap, employees diversity board (✂), skills (⊗), best employers...	sony group (⊗), diversity, rights, forced labor
ethics	business ethics, anti money laundering, business integrity, concerning marketing communications, financial industry laws, anti competitive behavior, ownes curning sustainability (⊗), modern slavery act	corruption

Table 4: Comparison between MultiTaxoGen and TaxoCom. Redundant topics are marked with (✂) and incorrect topics are marked with (⊗)

.

Redundant Topics When delving deeper into a taxonomy, topics become more specific when using a subcorpus derived from their parent topic. Due to the shared parent topic, documents within this subcorpus are closely related, posing challenges in distinguishing between child topics and resulting in the emergence of many similar child topics that may need merging. Table 5 highlights some redundant topics that are merged once they exceed the threshold τ_r. This led to a further reduction of 48% in the number of topics, thereby indicating that after removing the redundant and outlier topics, the total number of initial topics found by our method was reduced by approximately 75%.

Parent Topic		Child Topic's Top 3 Terms	Similarity
greenhouse gas emissions	T1	**emission reductions**, emissions kt, kyoto protocol	0.882
	T2	**emissions reducing**, reduce carbon, ghg emissions	
human rights	T1	**human rights assessment**, training human rights, rights policies	0.850
	T2	**rights assessment**, supplier human rights, grievance	

Table 5: Examples of redundant topics found (denoted as T1 and T2), their parent topic, and the similarities between the topics to be merged. Main term of child topic is highlighted in bold.

.

Generated Topics Comparison A major issue that we've observed with TaxoCom and other state-of-the-art topic taxonomy construction methods is that they have been designed to work on much more general datasets and are hard to optimise for specific tasks, such as analyzing ESG reports. Also, TaxoCom struggles to identify the majority of topics within larger corpora, and those it does identify often prove ambiguous or overly broad, even after tuning the novelty parameter β to enhance topic discovery. In contrast, our approach, as demonstrated in Table 4, notably discovers more topics in our corpus. Figure 1 shows a small sample of a taxonomy constructed by our method.

5.5. Discussion

We have introduced a novel method called MultiTaxoGen for creating a topic taxonomy from ESG reports. Our compiled corpus consists of ESG reports with varying styles, formats, and lengths, ranging from a few pages to several hundred pages, all originally in PDF format. The use of an existing PDF parser introduced some text inaccuracies, adding complexity to the data. To enhance the effectiveness of constructing a topic taxonomy, we divided the text documents into 256-word segments. Ideally, the segmentation of ESG reports should be based on their actual content. Future research could explore discourse relations and topic transitions to improve document segmentation. Our preprocessed corpus comprises more than 1 million documents, making it challenging to directly apply existing hierarchical topic models for taxonomy construction due to the extensive computational time required and the difficulty in controlling the topic quality. The current leading approach, TaxoCom, only managed to identify a total of 60 topics, missing many salient ones. In contrast, our approach has the capability to uncover more nuanced topics.

6. Conclusions

We have proposed a novel method optimized for analyzing ESG reports though topic taxonomy construction, addressing limitations in existing methods for complex ESG reports. For future work, we can explore the use of Large Language Models (LLMs) to generate more suitable main term for the topics. Although we tested ChatGPT (Brown et al., 2020) for this purpose, the results were unsatisfactory and the generated main term were generally incorrect. However, it is worth noting that LLMs are rapidly improving, with new models constantly being produced. We anticipate that new LLMs equipped with enhanced capabilities may yield more accurate and contextually relevant main terms for topics, making them valuable tools for future research in topic taxonomy construction.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council through a Turing AI Fellowship (EP/V020579/1, EP/V020579/2).

7. Bibliographical References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Philipp Baier, Marc Berninger, and Florian Kiesel. 2020. Environmental, social and governance reporting in annual reports: A textual analysis. *Financial Markets, Institutions & Instruments*, 29(3):93–118.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M Blei, Thomas L Griffiths, and Michael I Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Charlotte Christiansen, Thomas Jansson, Malene Kallestrup-Lamb, and Vicke Noren. 2023. Households' investments in socially responsible mutual funds. *The Quarterly Review of Economics and Finance*, 87:46–67.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Allen H. Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text*. *Contemporary Accounting Research*, 40(2):806–841.

Eamonn Keogh and Abdullah Mueen. 2017. *Curse of Dimensionality*, pages 314–315. Springer US, Boston, MA.

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 783–792.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022a. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. *CoRR*, abs/2201.06771.

Dongha Lee, Jiaming Shen, Seonghyeon Lee, Susik Yoon, Hwanjo Yu, and Jiawei Han. 2022b. Topic taxonomy expansion via hierarchy-aware topic phrase generation.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. IEEE.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

John Paisley, Chong Wang, David M Blei, and Michael I Jordan. 2014. Nested hierarchical dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):256–270.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora.

Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of The Web Conference 2020*, WWW '20, page 1908–1919, New York, NY, USA. Association for Computing Machinery.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *ArXiv*, abs/2002.10957.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. *CoRR*, abs/1812.09551.