The Impact of TCFD Reporting - A New Application of Zero-Shot Analysis to Climate-Related Financial Disclosures

Alix Auzepy, [1] **David Lenz,** [1] **Elena Tönjes,** [1] **Christoph Funk** [2]
[1]Justus Liebig University Giessen
Faculty of Economics and Business Studies
Licher Strasse 64
35394 Giessen, Germany
[2]Justus Liebig University Giessen
Centre for International Development and Environmental Research (ZEU)
Senckenbergstrasse 3
35390 Giessen, Germany
Alix.Auzepy@wi.jlug.de, David.Lenz@wi.jlug.de, Elena.Tönjes@wi.jlug.de, Christoph.Funks@wi.jlug.de

Abstract

We examine climate-related disclosures in 3,335 reports based on a sample of 188 banks that officially endorsed the recommendations of the Task Force for Climate-related Financial Disclosures (TCFD). In doing so, we introduce a new application of the zero-shot text classification based on the BART model and a Multi-Natural Language Inference (MNLI) task. By developing a set of robust and fine-grained labels, we show that zero-shot analysis provides high accuracy in classifying companies' climate-related disclosures without further model training. Overall, our findings show that TCFD-supporting banks increase their level of disclosure after the launch of the TCFD recommendations and following their individual declaration of support. However, we also find significant variation in the extent of reporting by topic, suggesting that some recommendations have not yet been fully met. Our findings yield important conclusions for the design of climate-related disclosure frameworks.

1 Introduction

Published in 2017, the recommendations of the Financial Stability Board's (FSB) Task Force on Climate-related Financial Disclosures (TCFD) have been described by the Government of the United Kingdom (UK) as "one of the most effective frameworks for companies to analyse, understand and ultimately disclose climate-related financial information" (FSB 2015; TCFD 2017a).

The TCFD recommendations, which have been officially endorsed by more than 3400 companies worldwide to date, have become one of the most established standards for climate-related disclosures (Beyene, Ongena, and Delis 2022b). Such standards, while often non-binding, are intended to help companies as well as investors and regulators find common ground on the types of non-financial data to be disclosed, analyzed and applied in investment decisions. The main focus of the TCFD recommendations lays on the integration of climate-related risks within companies. In this context, the TCFD relies on four disclosure categories (Governance, Strategy, Risk Management, Metrics and Targets), which all deal with the integration and management of climate-related issues within companies.[1]

A growing body of empirical evidence emphasizes the importance of "non-financial" disclosures for a comprehensive assessment and management of climate-related risks by companies (Matsumura, Prakash, and Vera-Muñoz 2013; Ilhan, Sautner, and Vilkov 2020; Krueger, Sautner, and Starks 2020). Hence, an increasing number of investors have been exerting pressure on companies to issue reports that comply with TCFD recommendations.[2] Recently, several countries, including the UK, Switzerland and New Zealand have taken steps to make TCFD reporting mandatory for large companies in their jurisdictions.

On the other hand, a new strand of academic research emphasizes the current weaknesses of climate-related disclosures (e.g., greenwashing, lack of standardization, lack of quantitative data, and lack of transparency) and argues that such disclosures often fail to deliver on their promise of providing relevant information for decision-making (Bingler et al. 2022; Beyene, Ongena, and Delis 2022b). Most recently, a study investigated a sample of 818 TCFD-supporting firms[3] from 2015 until 2020 and presented evidence that climate-related reporting is associated with selective disclosure, implying that firms tend to disclose information primarily on non-material TCFD categories (Bingler et al. 2022). In particular, the authors argue that the ineffectiveness of TCFD reporting is highlighted by the fact that reporting on "Strategy" as well as "Metrics and Targets" is generally quite limited, even though these two categories are the most material in the authors' eyes (Bingler et al. 2022).

As of today, research on climate-related disclosures remains very sparse. A possible reason for this could be that the manual analysis of companies' sustainability reports is very time-consuming and difficult to perform systematically.

[1]Figure 1 provides an overview of the TCFD categories.

[2]See e.g., Larry Flink's 2022 Letter to CEOs: https://www.blackrock.com/corporate/investor-relations/larry-fink-ceo-letter.

[3]A list of the official TCFD supporters can be found on the TCFD's website: https://www.fsb-tcfd.org/supporters/.

One way to address this issue is to use natural language processing (NLP), which enables systematic and automated extraction of textual information from large amounts of reports. A recently introduced language model for the analysis of climate-related disclosures is ClimateBERT (Webersinke et al. 2021), an algorithm trained to automatically identify and classify climate-related content.[4]

However, a weakness of ClimateBERT and, more generally, of the models underlying the algorithm is that such models require an extensive training set of human-labeled sentences. Manual labeling of sentences is not only time-consuming, but can also be error-prone. This is particularly the case when disclosures are multi-dimensional, i.e. address multiple issues in a single sentence.[5] Therefore, for quality and consistency reasons, highly-trained and specialized "labelers" are required, which can also make the labeling process costly. Furthermore, the more classes (or categories of labels) to be included into the classification scheme of the model, the more labeled data is needed to ensure that each class comes with a reasonable amount of examples attached to it. For example, ClimateBERT (Webersinke et al. 2021) only distinguishes between 4 classes of climate-related disclosures, which can be a limiting factor.

To overcome this weakness, we introduce in this paper a new NLP method for analyzing climate-related disclosures: the zero-shot text classification. We apply this method by using BART as our base model and fine-tune it with a Multi-Natural Language Inference (MNLI) task (Davison 2020; Yin, Hay, and Roth 2019).[6] As it is the case for most language models, the BART model itself is already pre-trained on approximately 160 GB of text from the English Wikipedia and BookCorpus dataset and is therefore particularly well suited to "understand" the semantics of climate-related reporting, especially since a large part of the TCFD recommendations is related to non-financial information, such as governance and strategy issues. The zero-shot method has the advantage of being able to classify sentences using labels for which it has received no prior training. In the context of this analysis, the zero-shot returns probabilities that a text sequence deals with a specific topic.

We apply the zero-shot to a sample of 3,335 hand-collected reports published by 188 TCFD-supporting banks between 2010 and 2021. Furthermore, we develop 14 fine-grained labels, which are designed to capture key aspects of the TCFD recommendations for the financial sector. Overall, the TCFD recommendations appear to be particularly suited for a zero-shot analysis, as they provide us with an already-existing framework and semantics. [7] More importantly, this paper explores whether climate-related disclosures in the "Strategy" and "Metrics and Targets" categories remain at low levels even after the launch of the TCFD recommendations until 2020, as postulated in existing research (Bingler et al. 2022)), and examine whether this result holds for all recommended disclosures within each TCFD category.

We focus on banks for several reasons: First, the TCFD recommendations are largely aimed at financial institutions. As financial intermediaries, banks are exposed to climate-related risks through various channels, including their borrowers and counterparties (TCFD 2017b). As climate change affects the credit risk of different types of assets and poses the risk of stranded assets, high exposure to climate change by financial institutions could also increase the risk of financial instability (Beyene et al. 2021). Second, we deliberately focus on banks to ensure that our labels are well-suited and reflect material disclosure topics.[8] Highlighting the need for sector-specific disclosure guidelines, the TCFD itself published supplemental guidance for the financial sector. Compared to other sustainability reporting standards that focus primarily on GHG emissions as key metrics, the TCFD recommendations place great emphasis on disclosure of fossil fuel exposure and on quantifying the concentrations of carbon-related assets in the financial sector (FSB 2015; Beyene, Ongena, and Delis 2022a). As a result, our interpretation of the TCFD recommendations contrasts with that of (Bingler et al. 2022) since we argue that considering "Strategy" and "Metrics and Targets" disclosures to more material may prove to be inaccurate and overlooks some industry specificities.

In a first step, we analyze the probabilities of climate-related disclosures based on the four overall TCFD categories (Governance, Strategy, Risk Management, Metrics and Targets). We observe an increase in reporting for all four climate-related overall categories starting in fiscal year 2017. Compared to (Bingler et al. 2022) who measure a slight increase of 2.2 percentage points in information disclosed after the launch of the TCFD recommendations, we find an average increase of approximately 4 percentage points across the four categories, as captured by our labels GO.1, ST.1, RM.1 and MT.1.

In a second step, and due to the weaknesses of considering solely the overall TCFD-categories, we combine the precise recommended disclosures and the additional financial sector guidance into a set of fine-grained labels. In doing so, we find considerable variation in reporting regarding the recommended disclosures within each TCFD category. In particular, across all four categories, our results show the lowest mean probabilities for the following labels: "Climate-related physical risks such as acute weather events and chronic shifts in weather patterns", "Financing and investment for carbon-intensive industries such as fossil fuel industry", "Use of climate-related scenario models to analyse the impact of climate-related risks", "Incorporation of climate-related performance metrics into remuneration policies", "Emissions reduction and carbon neutrality targets" and "Board's responsibility for overseeing climate-related issues", "Executive management's strategic role related to

[4]BERT stands for Bidirectional Encoder Representations from Transformers (Devlin et al. 2018).

[5]For example, in the context of the TCFD recommendations, a sentence describing climate goals adopted and monitored by the board could be labeled "Governance" and "Metrics and Targets".

[6]The BART model is a hybrid of the sequence-to-sequence transformer architecture used in BERT (Devlin et al. 2018) and a left-to-right autoregressive decoder architecture (Radford et al. 2018). For a more detailed description of MNLI, see section 3.

[7]See Table 5 for an overview of our labels.

[8]See also the SASB Materiality Map (SASB 2018).

the assessment and management of climate-related issues". Since most of these labels correspond to recommended disclosures under the Strategy and Metrics and Targets categories, our results partly corroborate and complement the findings in (Bingler et al. 2022), who argue that disclosures in the Strategy and Metrics and Targets categories are more limited. These results suggest that the TCFD-supporting banks in our sample have not yet implemented all the recommendations to the same extent or are performing selective disclosure.

Third, we find that the individual support of the TCFD recommendations goes along with an increase in climate-related reporting, which is statistically significant for all disclosure topics, although we also observe some variation in the extent of disclosure. Finally, we analyze whether our results corroborate existing literature on the relationship between company size and CSR activities (Jackson et al. 2019) and show that larger banks display larger disclosure probabilities as measured by the zero-shot text classification compared to medium or small banks.

Our findings complement the existing literature on climate-related disclosures. Our main contributions to the literature are threefold: First, we introduce the zero-shot text classification as a new method for automated content analysis of climate-related disclosures. We demonstrate the reliability of the method by creating and evaluating a new dataset based on 3,335 financial reports from 188 TCFD-supporting banks. Second, we develop a set of robust and fine-grained labels that allow us to leverage information on the TCFD recommendations beyond the four pillars and can be used for further NLP research on climate-related disclosures. Third, we offer new insights into TCFD recommendations and their limitations by showing that several recommended disclosures are based on abstract concepts (e.g., resilience) that cannot be fully measured. Without concrete standards, this could potentially lead to misleading information.[9]

The remainder of this paper is organized as follows. In Section 2, we describe our data by providing descriptive statistics on our text corpus. Section 3 briefly explains the underlying methods for parsing the PDFs and the zero-shot text classification, followed by Section 4, which presents an evaluation of the classification performance of our zero-shot model. Section 5 provides our main findings. Our concluding remarks can be found in Section 6.

2 Data

We apply the zero-shot classification to a sample of 3,335 hand-collected reports between 2010 and 2021. A detailed overview of the data collection steps, can be found in Figure 8. As a first step, we retrieved all the names of TCFD-supporting banks from the TCFD website and filtered them by the industry categories of "banks", "central banks" and "capital markets". After removing the banks that we could not identify or for which reports are not available online, we are left with 188 TCFD-supporting banks.[10] Interestingly, almost half of our sample consists of banks from the Asia-Pacific region. European banks account for one-third and North American banks for about 10% of our sample. The majority of banks in our sample are mid-sized banks with total assets between USD 50 billion and USD 500 billion.

Region	Large	Medium	Small	$\sum$
Asia Pacific	15	51	24	90
Europe	23	26	17	66
Latin America	0	4	3	7
Middle East & Africa	0	3	2	5
North America	9	8	3	20
$\sum$	47	92	49	188

Table 1: Size and region of TCFD-supporting banks

As a next step, we collect the banks' available reports for the period 2010 to 2021 in order to include data well before and well after the release of the TCFD recommendations. The reports are classified according to the following categories: Annual reports, CDP reports, corporate governance reports, integrated reports, remuneration reports, sustainability reports, TCFD reports. We do not rely solely on TCFD reports, as we have found that most TCFD supporters do not publish a stand-alone report for their climate-related disclosures, but rather tend to integrate key information into their annual and sustainability reports. This is mostly consistent with the TCFD guidelines, which stress that climate-related disclosures should be included in "mainstream (i.e., public) annual financial filings" (TCFD 2017b). Finally, we parse the reports to ensure they are in a format suitable for the zero-shot classification. We are left with a total sample of 3,335 bank reports. The majority of reports in our sample consists of annual and sustainability reports.

Report Category	# of reports	Average pages	Average # sentences
Annual Report	1869	207.98	2711.21
CDP Report	75	63.43	699.79
Corporate Governance Report	148	69.44	1014.25
Integrated Report	183	163.98	2354.95
Remuneration Report	83	36.88	494.42
Sustainability Report	896	81.01	1158.54
TCFD Report	81	36.68	544.37

Table 2: Descriptive Statistics - Sample composition

[9](Beyene, Ongena, and Delis 2022a) argue that even more specific recommended disclosure topics such as fossil fuel exposure are not concretely defined by the TCFD. Thus, the key metrics used to measure this exposure are at discretion of the reporting firms. Furthermore, reporting practice does not allow for an evaluation over time and against peers, and complicates accurate monitoring of stranded assets risk.

[10]We categorize the banks in our sample according to the region of their headquarters and their total asset size. Banks are also classified by asset size, with banks with total assets greater than USD 500 billion considered "large", banks with total assets between USD 50 billion and USD 500 billion considered "medium", and banks with less than USD 50 billion considered "small".

3 Methodology

Parsing PDFs

All of the reports used in our analysis are in PDF format. Converting textual information contained in the PDFs into a suitable format for further NLP analysis is not as trivial as analyzing textual information stored in CSV or TXT files. We use a layout parsing model based on Visual-Layout (VILA) groups (text lines or blocks) introduced in previous research (Shen et al. 2021). One can choose between two model variants, both based on a BERT-based model. In a nutshell, the idea behind VILA is that texts consist of groups of tokens (lines or blocks). These tokens then can be extracted by rule-based parsing or layout detection models (Shen et al. 2021). The authors introduce the variants for injecting the group structure, called H-VILA (Visual Layout-guided Hierarchical Model) and I-VILA (Injecting Visual Layout Indicators).

As the H-VILA block variant trained on the grotoap2 training using the layoutLM model (Xu et al. 2020) as a base model gave better results, we chose this combination based on a personal assessment using a randomly selected sample. The output of the model consists of the extracted text together with the corresponding layout information. Depending on the training set, the layout tags are different. For the grotoap2 data set, the tags which are of interest are, for example, figure, body content, abstract and title. For our analysis we disregard most of the tagged parts except for the body content and abstract.

Zero-Shot Classification

A widely used and important NLP task is text classification (Belinkov and Glass 2019). Text classification is used to organize and understand very large amounts of textual data by assigning so-called "labels" based on the topic of individual text sequences. Depending on the analysis to be performed, the text sequences can be either sentences, paragraphs, or entire pages. Thus, in the context of text classification, labels are used to capture key topics and predict the likelihood that a text sequence addresses one or more previously labeled topics.[11]

We employ a zero-shot text classification model introduced by Davison (2020), using BART as a base model and fine-tuned with the MNLI task. The model classifies text sequences by making use of the semantics of the input sequences and the labels. A simplified structure of our model architecture is shown in Figure 2.[12]

As a base model for the zero-shot classification, we use BART, which is pre-trained on roughly 160GB of text from the English Wikipedia and BookCorpus dataset in order to "understand' the semantics of texts. Since the TCFD recommendations are not about highly complex and specialized financial language, but about more general semantics, we consider such a training data set appropriate. In contrast to BERT, the BART model (Lewis et al. 2019) not only makes use of the sequence-to-sequence translation architecture with bidirectional encoders (BERT), but also uses a left-to-right autoregressive decoder (GPT model) and is therefore a mixture of both. In combination with the zero-shot text classification, BART as a base model demonstrates high performance results (Davison 2020).

The method introduced in previous research uses a BART model for the pre-trained MNLI base model (Davison 2020). The specific NLP task used for zero-shot classification is Natural Language Inference (NLI); more specifically Multi-Natural Language Inference (MNLI).The zero-Shot text classification is based on embedding the sentences of a text (a sequence of words) and the labels themselves into the same latent space. In such a latent space, the distance between the sentence and the label can be computed. The closer the label is to the sentence, the higher the probability that the label matches the sentence. A refinement of this technique consists in NLI.

When using NLI, the model treats text sequences as premises and labels as hypotheses. It then tokenizes them and uses the underlying language model to embed both, the sentence and the label. It then runs both through the pre-trained MNLI layer. The MNLI end-layer is a simple fully connected neural network where the output is a vector of logit scores for three outcomes: "neutral", "contradiction", and "entailment" (Lewis et al. 2019; Davison 2020). Hence, the hypothesis is tested against the premise and the result can be a classification as entailment, a contradiction, or neutral. The score for "neutral" is discarded and a softmax function is applied to the "contradiction" and "entailment" scores in order to be able to interpret them as a probability. In our analysis, the scores shown are for the "entailment" only. They can be interpreted as the probability that the corresponding sequence matches the label, or in other words, the probability that the label is true.

The fine-tuned end layer then can be used for zero-shot classification without any further training (Davison 2020). In some cases, a zero-shot model is able to predict the class of a text with a higher accuracy than supervised models that have been trained on hundreds of labeled training items. However, the success of such a model depends on a careful selection of labels, which we further discuss in our section 4.

4 Label Evaluation

The zero-shot text classification does not require any pre-training by us. Thus, we cannot validate the model in the usual way where we would split a data set into a training set and a test set and validate our model using the test set. However, we can still manually label a set of sentences and then compare the labels assigned by the model to the labels assigned by us or by (Webersinke et al. 2021). Hence, we perform a battery of robustness tests to evaluate the classification performance of our model.

First, we perform an evaluation by applying our model to the training repository used in (Webersinke et al. 2021),

[11] Please refer to Table 5 for an overview of our labels.

[12] In general, language models are pre-trained on an extensive amount of text data. They can afterwards be trained on specific tasks in the fine-tuning stage. One can therefore use different combinations of several pre-trained language models (the base model) and task-specific end-layers.

available on GitHub.[13] This data set includes about 50.000 text sequences annotated with 5 classes (4 TCFD categories as well as "none" for non-climate-related text) For this test, we run our zero-shot text classification by using the same classes and add "climaterelated" in front of the four TCFD labels (i.e. climaterelated governance, climate-related strategy, climate-related risks management and climate-related metrics and targets) to make sure that our model captures climate-related text sequences. We also include sentences which where labeled as not belonging to one of the four groups, as a negative example. The results can be seen in Figure 3.

The sentences annotated by (Webersinke et al. 2021) are represented on the x-axis, and the results from our model for our climate-related labels are on the y-axis and represent the mean probability returned by the zero-shot model that the sentences in the column deal with the topic of the label in the row. The darker the entries, the higher the likelihood classified by the model. In the case of a perfectly working zero-shot text classification, the entries on the diagonal would be the darkest. When looking at figure 3, we can see that our zero-shot classification model comes to very similar result compared to the annotation made by (Webersinke et al. 2021). The sentences labeled by (Webersinke et al. 2021) as governance-related sentences do also, according to our model, have a higher probability to deal with climate-related governance topics compared to the other sets of sentences. The same goes for "Metrics and Targets", although "Metrics and Targets" has lower probabilities in general for all labels.

For the sentences annotated by (Webersinke et al. 2021) as strategy-related and risk management-related, the results are less clear. The zero-shot model still assigns the highest probabilities to the ST.1 and RM.1 labels for the corresponding groups of sentences, but other groups also achieve high probabilities. For example, according to our model, the strategy sentences labeled by (Webersinke et al. 2021) have a 51% probability of addressing climate-related strategy issues. However, sentences labeled by (Webersinke et al. 2021) as metrics and targets also have a relatively high probability (49%) of addressing climate-related strategy topics. In the area of risk management, several sentences annotated by (Webersinke et al. 2021) as governance-related also appear to address risk management according to our model.

These results do not necessarily mean that the zero-shot model has a poor understanding of the semantics of text sequences. The zero-shot classification model still assigns the highest probabilities to the corresponding sentence categories annotated by (Webersinke et al. 2021). Instead, these results highlight the difficulty of classifying sentences according to one main broad category, even though these categories are defined by the TCFD. For example, the TCFD recommends that climate-related risks, such as physical and transition risks, be described under the Strategy pillar. This description could also be closely related to risk management processes used to identify, assess, and manage such risks. However, such processes would fall under the Risk Management category, and are at the same time closely related to the Strategy category. Hence, this observation stresses the need for fine-grained labels as well as a multi-label approach, which is applied in the following robustness checks.

As a second step, we create precise labels with the help of a financial and ESG expert. Each label attempts to capture the main idea behind a TCFD recommendation and applies similar wording. As highlighted in Section 3, one of the challenges of the zero-shot approach is label selection. Therefore, we performed several classification tasks to find the most fitting labels. In applying the zero-shot classification, we choose a multi-label approach since the TCFD recommendations are interrelated, as explained earlier. Therefore, we do not force the model to return probabilities that add up to one (single label approach). Instead, we choose an approach where the model is able to assign probabilities from 0 to 1 for each label (multi-label approach). Consequently, the results for all labels for each sentence do not add to one.

Next, we build a larger test data set by collecting sentences that are consistent with the TCFD recommendations. The data set consists of sentences from the TCFD good practice handbook (CDSB 2019, 2021), which contains examples of best practice disclosures selected by the TCFD. Furthermore, we extract sentences from TCFD reports of companies that are not in the banking sector. Finally, we also use the aforementioned training repository used in (Webersinke et al. 2021), available on GitHub.

We repeat the labeling process by labeling the sentences collected by us as well as 1500 sentences from the training repository provided by (Webersinke et al. 2021). We include the label "none" to the classification task. The purpose of this label is to capture the non-climate-related text, i.e. the text sequences that do not fit any of our fine-grained labels. It also ensures that the labels are not randomly assigned by the zero-shot, but that the assignment is really based on the semantics of the text sequences. In order to test the performance of our model with regards to the "none" label, we use sentences labeled as "none" in the training repository from (Webersinke et al. 2021) mentioned above, as well as additional sentences labeled by us as "none".[14]

The results of the zero-shot text classification using our fine-grained labels are presented in Figure 4. On the one hand, the results in Figure 4 show that our model provides satisfactory results and assigns high probabilities to the relevant labels. On the other hand, the results also illustrate the challenge of capturing abstract concepts using labels as well as the fact that the topics of the TCFD recommendations are often closely intertwined. Looking more closely at the results, we find that labels ST.1.1, ST.1.3, ST.1.7, RM.1.1, and RM 1.2 tend to perform worse than other labels in terms of text sequence recognition. For example, although the sentences in the ST.1.1 column were assigned the ST.1.1 label with a probability of 86%, the zero-shot classification also

[13]The data can be downloaded at: https://github.com/ClimateBert/training-example/blob/main/training_data.json.

[14]During our labeling process, we changed some of the labeled sentences from (Webersinke et al. 2021) to "none", as they were, in our opinion, not addressing any of our fine-grained labels.

assigns a relatively high probability to this label for most of the other sentence categories (on the y-axis). As for the labels RM.1.1 and RM1.2, not only those we labeled as risk management text sequences, but also several sentence groups (on the y-axis), were assigned these labels with a probability of 60% or more.

There could be several reasons for these results. First, it could be that these particular labels were poorly chosen. Second, another reason could be that the topics addressed are too abstract for zero-shot text classification. For example, concepts such as "resilience" or "risk management processes" can be described in many different ways and could therefore be partially reflected in many text sequences. Third, the topics of the TCFD recommendations are often closely connected, so that text sequences can often fit several labels at once. For example, the only label which does not have the highest probability for its own sentences is label MT.1.1 (Carbon footprint, direct and indirect greenhouse gas emissions), where a higher probability was assigned by the zero-shot to "emissions reduction and carbon neutrality targets" (MT.1.3). However, since these topics are closely related, a high value for both sets of sentences is not surprising.

When looking at Figure 4, the probabilities for the "none" label are very low across all sentences. This suggests that the model does not only label correctly by chance, but actually incorporates the semantics of the labels. Moreover, sentences tagged as "none" were classified with a relatively low probability by the zero-shot text classification for nearly all labels. The label "climate-related transition risks" (ST.1.1) has a slightly higher probability of 37% for the "none" sentences. This may be linked to the fact that this label encompasses many different types of risks, such as political and legal risks, technological risks, market risks, and reputation risks, all of which belong to the "transition risks" category (TCFD 2017a). For some sentences describing these risks, the zero-shot classification may not directly identify the link to climate change. More abstract labels such as RM.1.1 ("Processes to identify, assess and manage climate-related risks and integrate them into overall risk management") and RM.1.2 (Relationship between climate-related risks and financial risks such as credit risk, market risk, liquidity risk and operational risk") have higher values for "none" sentences as well for the reasons explained above.

In sum, the zero-shot text classification does not appear to assign probabilities purely by chance. The overlap between some labels is not necessarily due to a poor model, but rather to the fact that the topics of the TCFD recommendations are very much intertwined. This is an argument for the multi-label approach we use.

5 Results

Disclosures measured by overall TCFD categories

We apply our zero-shot text classification to a sample of 3,335 reports from 188 TCFD-supporting banks across various regions for the period 2010 to 2021. As a first step, we analyze the evolution of climate-related disclosures by looking at the 4 overall TCFD categories, as shown in Figure 5. We deliberately include two types of labels representing the TCFD categories. The blue label represents the category names without the adjective "climate-related" (i.e. "Governance", "Strategy", "Risk Management", and "Metrics and Targets"), while the orange label represents the category names with the explicit mention of "climate-related". Hence, Figure 5 presents the probability that the text sequences extracted from the reports in our analysis deals with the labels "Governance", "Strategy", "Risk Management", and "Metrics and Targets", as well as GO.1., ST.1, RM.1 and MT.1, respectively.

Several observations can be made based on our overall sample: First, the zero-shot text classification appears to make a good distinction between climate-related and non-climate-related textual data. For example, reporting on "Governance" is higher than reporting on "Climate-related Governance" (GO.1), which is not surprising. This observation holds true for all 4 TCFD categories. Second, reporting on the overall categories represented by the blue lines remains largely constant over time, with the exception of a slight increase in "Metrics and Targets". When looking at the climate-related TCFD categories in orange, however, we observe an increase in reporting, often starting in 2016 and increasing after 2017. Third, compared to existing literature (Bingler et al. 2022), we do not find that disclosures in the "Strategy" and "Metrics and Targets" areas remain at comparatively lower levels even after the launch of the TCFD recommendations, despite being the most relevant for stakeholders. Rather, we find that disclosures in the Risk Management category are more limited than in the other categories.

Figure 6 represents the sum of label probabilities (for the labels GO.1, ST.1, RM.1, MT.1) for the full sample between 2010 and 2021. By comparing the periods 2010 to 2016 with 2017 to 2021, we find an increase in disclosures, which is statistically significant (Table 3) across all TCFD categories. Our results reveal the largest increase in disclosures in the Strategy category (4.8 percentage points), followed by Metrics and Targets (3.90 percentage points), and Governance (3.59 percentage points). We record the lowest increase in the Risk Management category (3.39 percentage points).

This observation could be due to several reasons. First, risk management processes in the banking sector are highly regulated and can be particularly difficult to change. In addition, while banking regulators in Europe have started to require banks to integrate climate-related risks into their risk management processes, this is not necessarily the case for Asian or North American banks. Second, we note that the TCFD recommendations under the Risk Management pillar, in particular the supplemental guidance for the financial sector, are less concrete and detailed than the other reporting categories. For instance, recommended disclosures (a), (b) and (c) strongly overlap, and recommended disclosure (c) is very broadly formulated ("Organizations should describe how their processes for identifying, assessing, and managing climate-related risks are integrated into their overall risk management")(TCFD 2017a). Therefore, the quality (and quantity) of disclosures may also depend heavily on the precision of the recommendations themselves.

Disclosures measured by fine-grained TCFD labels

One way to assess the quality of climate-related disclosures might be to look not only at the quantity of reporting for each TCFD category, but rather to look more closely at reporting on the material issues within each category. Hence, we further investigate climate-related corporate disclosures by looking at the underlying recommendations, and in particular the specific guidance for banks, which we combine into precise labels described summarized in Table 5. This approach is also particularly useful because many sentences can often be assigned to more than one TCFD category. For example, a sentence describing climate goals adopted and monitored by the board could be labeled "Governance" and "Metrics and Targets".

Figure 7 provides an overview of the climate-related disclosures for the full sample of banks and reports. The boxplots illustrate that there is considerable variation in the extent of disclosure even within each TCFD category. The Strategy category is the most comprehensive and includes several specific recommended disclosures for banks, which explains why we have more labels in this area. Here we find that, on average, banks report less on climate-related physical risks (ST.1.2), financing of carbon-intensive industries (ST.1.5), and use of climate-related scenario models (ST.1.6). However, it is noticeable that the reporting on the impact of climate-related issues seems to be much higher (ST.1.3). Several aspects could explain these results: On the one hand, banks might report less because they do not yet perform comprehensive scenario analyses and do not yet have the technical expertise to identify physical risks in their portfolios. On the other hand, these banks may also perform selective disclosure and omit material information for stakeholders. For example, existing research shows that the banking sector has been providing significant financing to the fossil fuel industry (Beyene, Ongena, and Delis 2022b). Thus, banks may be intentionally not disclosing all information, particularly with respect to ST.1.5. Under Metrics and Targets, we find that reporting on the incorporation of climate-related performance metrics into remuneration policies (MT.1.2) is much lower than reporting on metrics related to GHG emissions (MT.1.1). This could also be due to the fact that most banks have not yet aligned their compensation policies with climate-related performance metrics and therefore report very little information. Finally, in the Governance area, the TCFD-supporting banks seem to report less on the role of management in assessing and managing climate-related issues (GO.1.2) than on the board's responsibility for overseeing climate-related issues (GO.1.1), which is consistent with the observation made under Section 5 that disclosure on processes for the assessment and management of climate-related risks appear to be lagging.

Evolution of climate-related reporting after TCFD support

As a next step, we examine whether climate-related reporting increases after banks individually support the TCFD recommendations. Table 3 presents the results of a paired t-test, in which we compare the mean difference of label probabilities before and after the TCFD introduction in 2017 for the whole sample as well as for the year of individual TCFD support. In fact, several banks did not become official TCFD supporters immediately after the TCFD recommendations were published, but joined in subsequent years.

		TCFD support since				
	All n = 188	2017 n = 38	2018 n = 26	2019 n = 25	2020 n = 30	2021 n = 53
GO.1	3.59***	5.69***	4.72***	4.25***	4.51***	3.24***
GO.1.1	3.29***	5.02***	4.43***	3.74***	4.34***	3.12***
GO.1.2	2.21***	3.34***	2.88***	2.61***	2.98***	2.25***
ST.1	4.80***	7.70***	6.51***	5.61***	6.16***	4.23***
ST.1.1	2.41***	4.15***	3.41***	2.59***	3.22***	2.54***
ST.1.2	1.47***	2.57***	1.89***	1.80***	2.10***	1.75**
ST.1.3	2.74***	4.84***	5.36***	4.05***	4.52***	3.28***
ST.1.4	2.45***	4.60***	3.98***	2.53***	3.25***	3.10***
ST.1.5	0.50***	1.19***	0.96***	0.28	0.65***	0.71**
ST.1.6	1.59***	3.24***	2.19***	1.77***	1.75***	1.93***
ST.1.7	2.44***	4.13***	3.54***	3.02***	3.24***	2.45***
RM.1	3.39***	5.76***	5.02***	3.42***	4.25***	3.09***
RM.1.1	2.82***	4.72***	3.84***	3.11***	3.56***	2.74***
RM.1.2	0.33*	1.12**	2.20***	0.20	1.40**	1.29**
MT.1	3.90***	6.22***	5.03***	4.57***	4.78***	3.62***
MT.1.1	4.06***	6.43***	5.44***	5.28***	5.46***	3.75***
MT.1.2	0.63***	1.06***	0.37*	1.03***	0.55**	0.86**
MT.1.3	3.74***	5.34***	4.19***	4.94***	4.56***	3.38***

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 3: Paired t-test of climate-related disclosures

For the individual TCFD support, we compare the mean difference before and after the year in which banks decided to officially become TCFD supporters. For example, if banks became TCFD supporters in 2018, we compare the mean of each label probability for all of these banks (26 in total) by taking the mean per bank from 2010 to 2017 and comparing it to the mean from 2018 to 2021 after the banks became supporters. For the whole sample, we compare the mean difference up to the publication of the official TCFD recommendations (mean of years 2010 to 2016) and after they were published (mean of years 2017 to 2021). Table 3 shows the results of a paired t-test for the mean difference (in percentage points) for our label probabilities [15].

We find that the official TCFD introduction leads to a significant increase in climate-related reporting across all TCFD recommendations for our full sample. Thereby, the results vary widely, ranging from an increase of 0.63 percentage points before and after TCFD introduction for MT.1.1 to 4.8 percentage points for ST.1. In general, these results appear to be robust across all sub-samples, although they vary in size. The groups of banks that became supporters in 2017, 2018, 2020, and 2021 reported more on the issues addressed in the TCFD from the year they became TCFD supporters. Similarly to (Bingler et al. 2022), we report the largest nominal effects are found for banks that already became TCFD supporters in 2017 and 2018. This

[15] As a robustness check, we also looked at permutation p-values from a monte-carlo-simulation test and bootstrap confidence intervals with qualitatively very similar results.

shows an increase of up to 7.7 percentage points for ST.1. after the start of TCFD support for banks in 2017. Only in the group of TCFD supporters who became official supporters in 2019 do we find no significant impact on reporting on ST.1.5 and RM.1.2., which could be linked to observations made earlier (i.e. potential lack of risk management processes and selective disclosure).

All in all, our results suggest that the introduction of the TCFD recommendations does indeed seem to increase reporting on climate-related reporting. At this stage, however, we cannot say whether banks are actually looking at these issues more intensively or whether it is merely the number of pages reporting on them that has increased.

Size effects

	Large - Medium	Large - Small	Medium - Small
GO.1	2.06***	3.46***	1.40***
GO.1.1	1.24***	2.23***	0.99***
GO.1.2	0.93***	1.58***	0.65***
ST.1	2.55***	4.68***	2.13***
ST.1.1	1.64***	2.35***	0.71**
ST.1.2	0.95***	0.96***	0.01
ST.1.3	3.06***	4.15***	1.09**
ST.1.4	2.67***	2.83***	0.16
ST.1.5	0.73***	0.71***	0.01
ST.1.6	1.76***	1.69***	0.07
ST.1.7	1.48***	2.30***	0.83***
RM.1	2.03***	2.91***	0.88***
RM.1.1	1.87***	2.40***	0.54
RM.1.2	1.40***	1.93***	0.54*
MT.1	2.38***	3.83***	1.46***
MT.1.1	2.39***	3.86***	1.48***
MT.1.2	0.46***	0.67***	0.21
MT.1.3	1.72***	3.46***	1.74***

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 4: Tukey difference-in-mean test of climate-related disclosures

Finally, we also analyze the relationship between bank asset size and TCFD reporting. To provide an overview of the differences in TCFD reporting by bank size, we use a Tukey test for the difference in means to assess the overall change in mean between large, medium, and small banks. Here, the Tukey test compares the means of all bank sizes to the mean of every other mean by pairwise comparison using a student range statistic that corrects for multiple comparisons.

Table 4 summarizes the results of the Tukey test for the differences between the means of the different bank sizes and their adjusted p-values. We report that the mean values increase significantly between large and medium banks and between large and small banks for all TCFD recommendations. However, for 5 of the 18 labels, we find no significant increase between medium and small banks. Moreover, the effects are nominally smaller for this group comparison than for the other groups. These results corroborate existing literature on corporate sustainability disclosures, which highlight that company size has an effect on company disclosures.

6 Conclusion

This paper investigates the climate-related disclosures in 3,335 reports based on a sample of 188 TCFD-supporting banks for the period 2010 to 2021. Our findings are threefold: First, we find that reporting on broad categories that do not explicitly relate to "climate" remains largely constant over time, while we observe an increase in reporting for all four climate-related overall categories beginning in fiscal year 2017. Second, when combining the precise recommended disclosures and the additional financial sector guidance into fine-grained labels, we find variation in reporting regarding the recommended disclosures within each TCFD category, suggesting that TCFD-supporting banks are not reporting to the same extent on all recommended disclosure topics. In particular, we find that banks have a lower probability to report on topics such climate-related physical risks, financing and investment for carbon-intensive industries, the use of climate-related scenario models and the incorporation of climate-related performance metrics into remuneration policies. These results may suggest that the TCFD-supporting banks have not yet implemented all the recommendations to the same extent or are disclosing information selectively. Third, we find that the individual support of the TCFD recommendations goes along with an increase in climate-related reporting, with some variation between the disclosure topics. Furthermore, we also report that larger banks are more likely to disclose climate-related information than medium or small banks. Overall, our findings complement previous research on climate-related disclosures frameworks (Bingler et al. 2022).

Our study certainly entails some limitations, which warrant careful consideration but also may highlight the potential for further research. First, we recognize that a higher quantity of disclosure does not necessarily goes along with better quality. Even though we find an increase in climate-related reporting following the release of the TCFD recommendations, we cannot say whether banks are actually addressing theses issues in question more intensively. Another limitation regarding the variations in disclosures is that we cannot say whether these differences are related to selective disclosure (i.e. greenwashing) or simply due to the fact that the recommendations have not been implemented.

Despite these limitations, our findings highlight the need for concrete and clearly delineated disclosure guidelines in the context of climate-related reporting frameworks. The more specific the guidelines are, the easier it is for companies to assess what kind of information is expected from them, and the easier it is for stakeholders to assess whether these guidelines are actually followed. Another important aspect besides specificity is materiality, which means that that guidelines should be relevant for the industry in question. An assessment of climate-related disclosures should not be based on the quantity of reporting for each overall TCFD category, but rather on the "sub-recommendations" that are the most material for stakeholders. In addition to the topics suggest above, future research could further investigate the link between quality of disclosure guidelines and implementation by companies using NLP.

Acknowledgements

Christoph Funk acknowledges the funding by the German Academic Exchange Service (DAAD) from funds of Federal Ministry for Economic Cooperation (BMZ), SDGnexus Network (Grant No. 57526248), program "exceed - Hochschulexzellenz in der Entwicklungszusammenarbeit."

References

Belinkov, Y.; and Glass, J. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7: 49–72.

Beyene, W.; De Greiff, K.; Delis, M. D.; and Ongena, S. R. G. 2021. Too-Big-To-Strand: Bond to Bank Substitution in the Transition to a Low-carbon Economy. *CEPR Discussion Paper No. DP16692*.

Beyene, W.; Ongena, S.; and Delis, M. D. 2022a. Disclosure of Bank Fossil Fuel Exposures. *European Economy - Banks, Regulation and the Real Sector*, 4(2): 89–105.

Beyene, W.; Ongena, S.; and Delis, M. D. 2022b. Financial institutions' exposures to fossil fuel assets: An assessment of financial stability concerns in the short term and in the long run, and possible solution.

Bingler, J. A.; Kraus, M.; Leippold, M.; and Webersinke, N. 2022. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47: 102776.

CDSB. 2019. TCFD Good Practice Handbook.

CDSB. 2021. TCFD Good Practice Handbook 2nd Edition.

Davison, J. 2020. Zero-Shot Learning in Modern NLP. Last accessed Feb. 28, 2022.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

FSB. 2015. Proposal for a Disclosure Task Force on Climate-Related Risks.

Ilhan, E.; Sautner, Z.; and Vilkov, G. 2020. Carbon Tail Risk. *The Review of Financial Studies*, 34(3): 1540–1571.

Jackson, G.; Bartosch, J.; Avetisyan, E.; Kinderman, D.; and Knudsen, J. S. 2019. Mandatory Non-financial Disclosure and Its Influence on CSR: An International Comparison. *Journal of Business Ethics*, 162(2): 323–342.

Krueger, P.; Sautner, Z.; and Starks, L. T. 2020. The Importance of Climate Risks for Institutional Investors. *The Review of Financial Studies*, 33(3): 1067–1111.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *CoRR*, abs/1910.13461. Last accessed Feb. 28, 2022.

Matsumura, E. M.; Prakash, R.; and Vera-Muñoz, S. C. 2013. Firm-Value Effects of Carbon Emissions and Carbon Disclosures. *The Accounting Review*, 89(2): 695–724.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

SASB. 2018. SASB Materiality Map.

Shen, Z.; Lo, K.; Wang, L. L.; Kuehl, B.; Weld, D. S.; and Downey, D. 2021. VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups. Last accessed Feb. 28, 2022.

TCFD. 2017a. Final report: Recommendations of the task force on climate-related financial disclosures.

TCFD. 2017b. Implementing the Recommendations of.

Webersinke, N.; Kraus, M.; Bingler, J. A.; and Leippold, M. 2021. ClimateBert: A Pretrained Language Model for Climate-Related Text. *CoRR*, abs/2110.12010.

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1192–1200.

Yin, W.; Hay, J.; and Roth, D. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. Last accessed Feb. 28, 2022.

7 Appendix

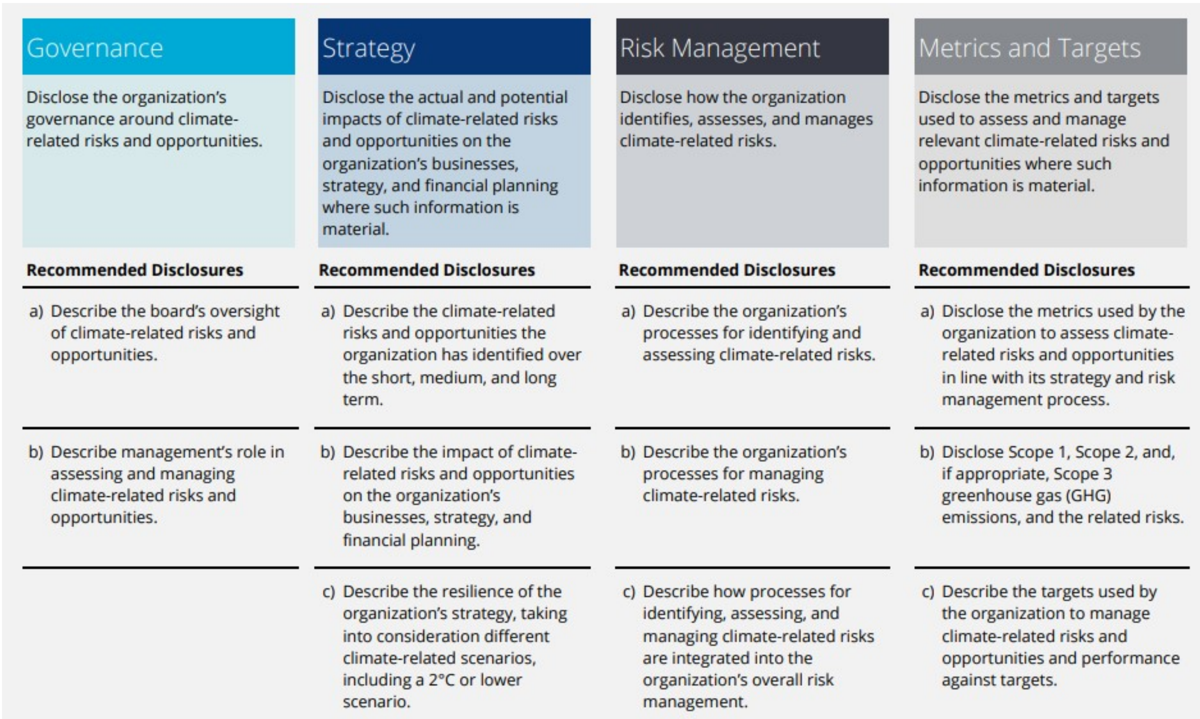

Figure 1: The TCFD recommendations and disclosure guidance (TCFD 2017a)

TCFD Category	Label name	Label Description
Governance	GO.1.	Climate-related Governance
	GO.1.1	Board's responsibility for overseeing climate-related issues
	GO.1.2	Executive management's strategic role related to the assessment and management of climate-related issues
Strategy	ST.1.	Climate-related Strategy
	ST.1.1	Climate-related transition risks such as policy, legal, technology, market and reputation risks emerging from climate change
	ST.1.2	Climate-related physical risks such as acute weather events and chronic shifts in weather patterns
	ST.1.3	Material financial impact of climate-related issues
	ST.1.4	Credit exposure to carbon-related sectors such as oil, gas, coal and electric utilities
	ST.1.5	Financing and investment for carbon-intensive industries such as fossil fuel industry
	ST.1.6	Use of climate-related scenario models to analyse the impact of climate-related risks
	ST.1.7	Resilience of the bank's strategy under different climate-related scenarios
Risk Management	RM.1.	Climate-related Risk Management
	RM.1.1	Processes to identify, assess and manage climate-related risks and integrate them into overall risk management
	RM.1.2	Relationship between climate-related risks and financial risks such as credit risk, market risk, liquidity risk and operational risk
Metrics & Targets	MT.1.	Climate-related metrics and targets
	MT.1.1	Carbon footprint, direct and indirect greenhouse gas emissions
	MT.1.2	Incorporation of climate-related performance metrics into remuneration policies
	MT.1.3	Emissions reduction and carbon neutrality targets

Table 5: Overview of TCFD Labels

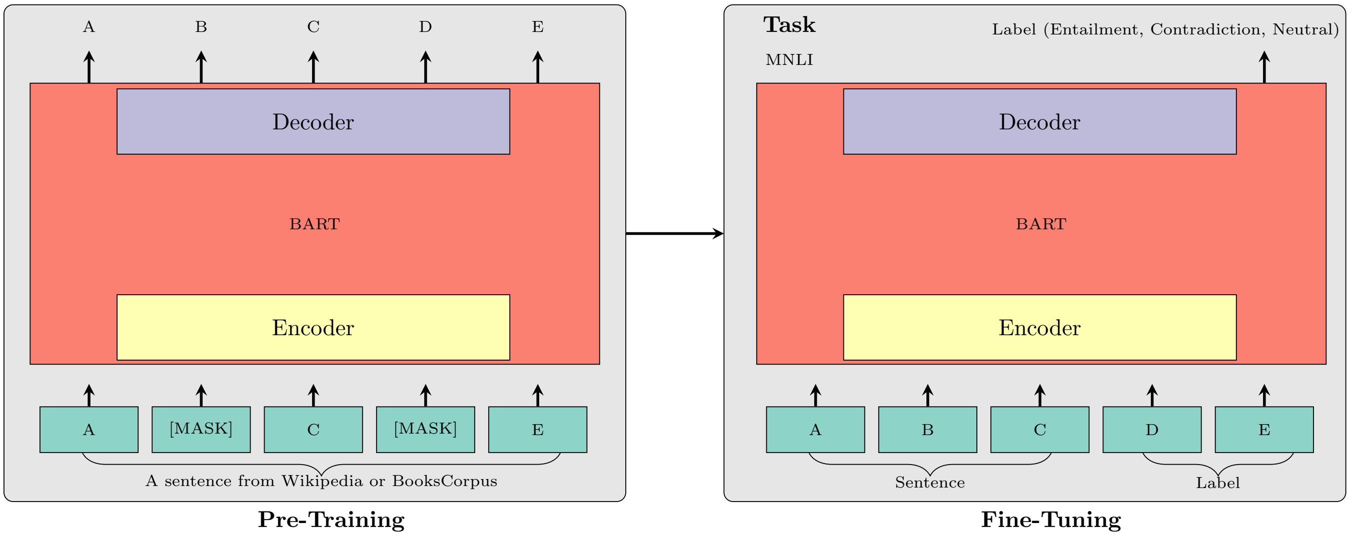

Figure 2: First, the BART model is pre-trained on all English Wikipedia articles and the BooksCorpus data set. This is shown on the left hand side. By masking parts of sentences (`[MASK]`), the model predicts the missing parts and learns the semantics. The process is done fore all sentences of the pre-training data set. On the right hand side the model is fine-tuned on the MNLI task. The already fine-tuned then can be used for zero-shot text classification.

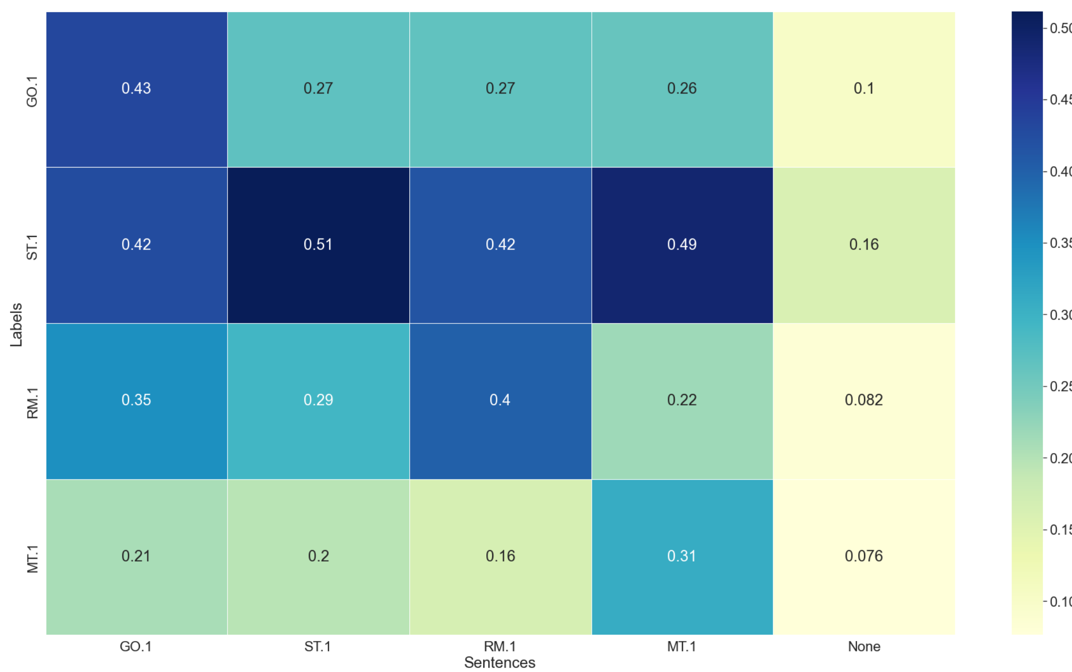

Figure 3: This matrix presents the results of the zero-shot classification applied to the training data set from (Webersinke et al. 2021).

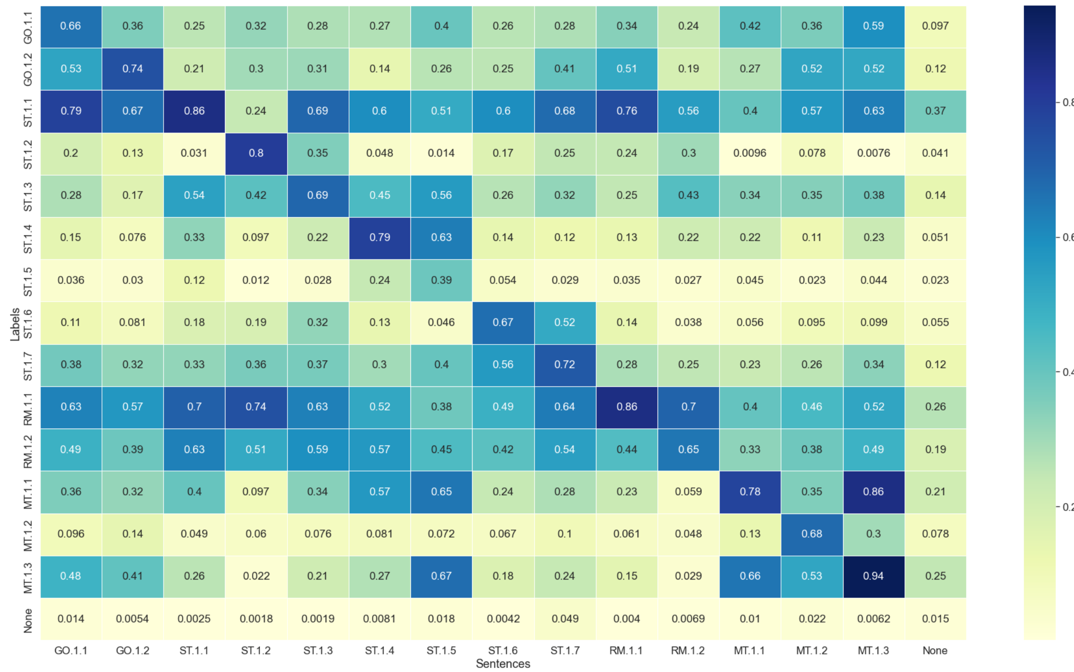

Figure 4: This matrix presents the results of the zero-shot text classification based on shorter and fine-grained labels.

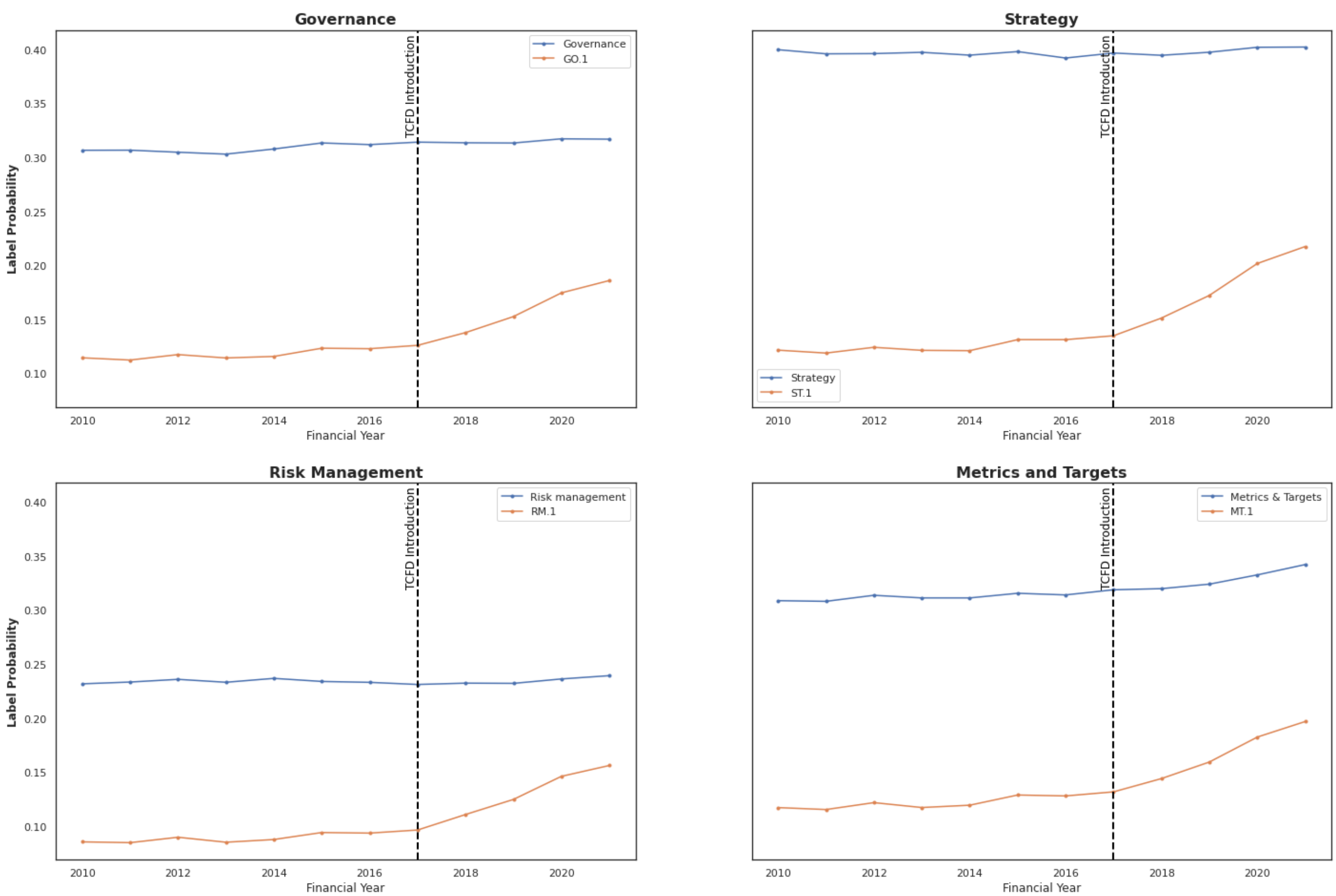

Figure 5: Climate-related disclosures by TCFD categories

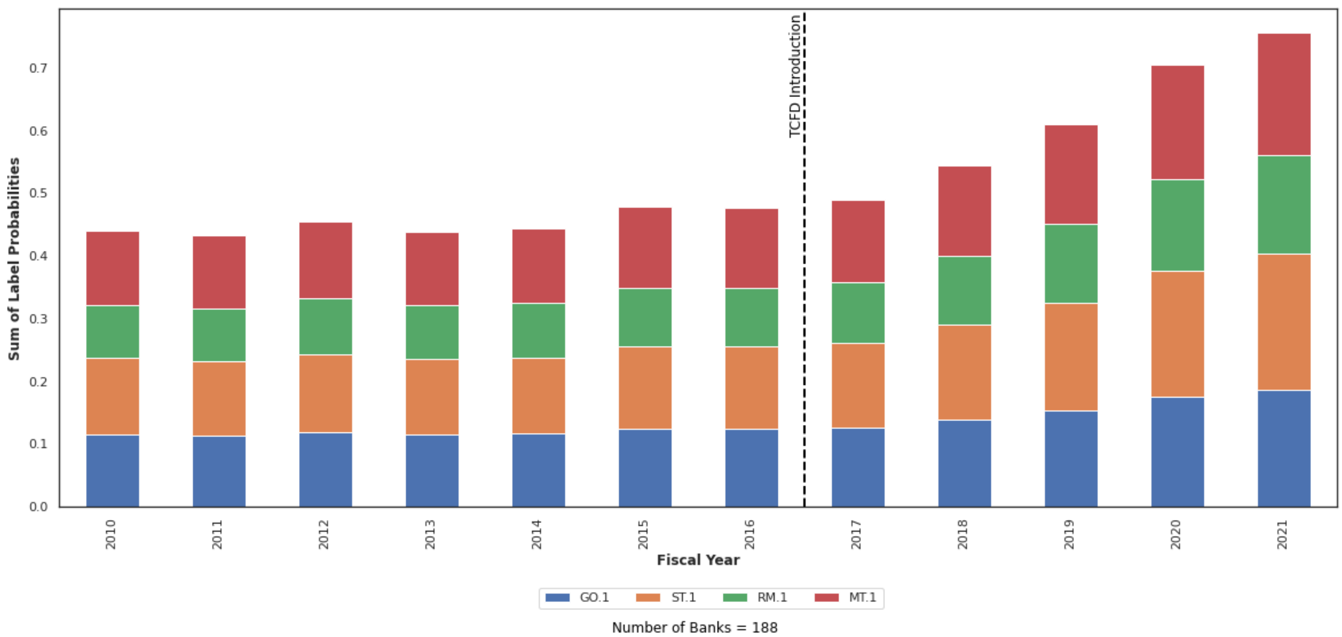

Figure 6: Average relevance by TCFD category for the full sample

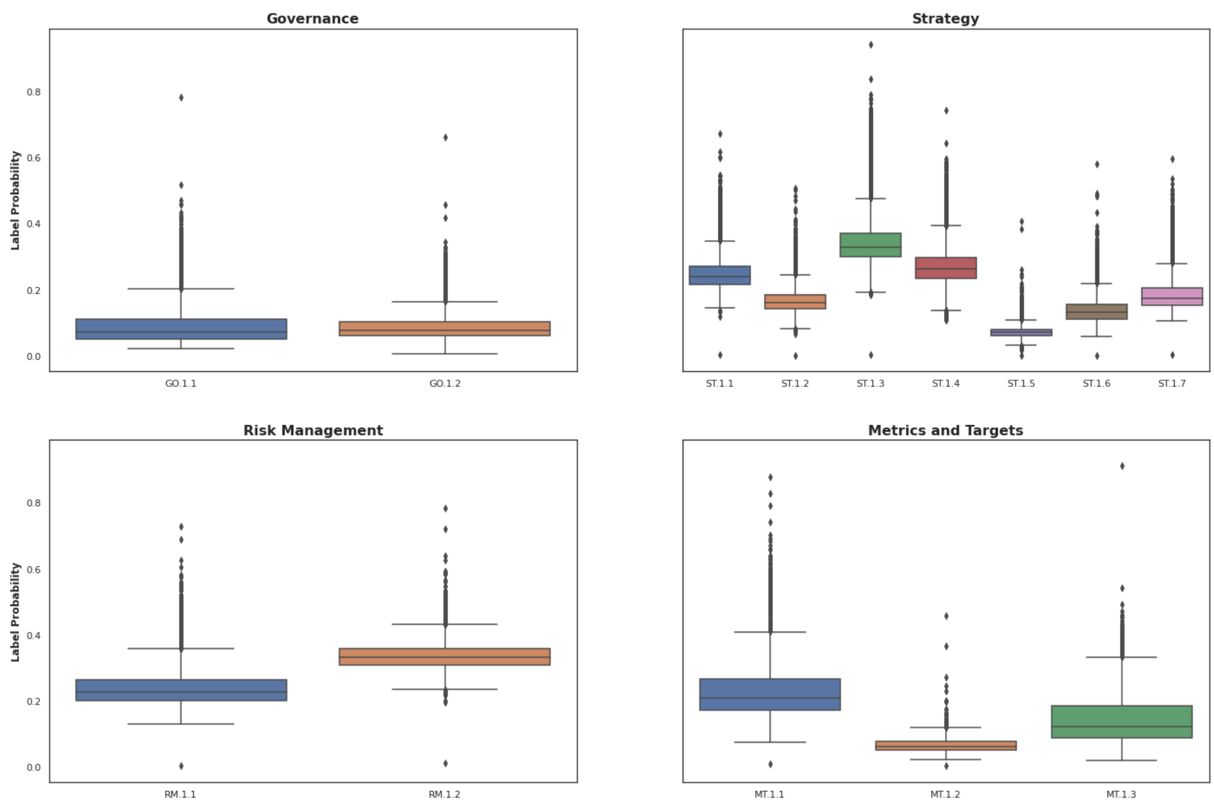

Figure 7: Climate-related disclosures by fine-grained labels

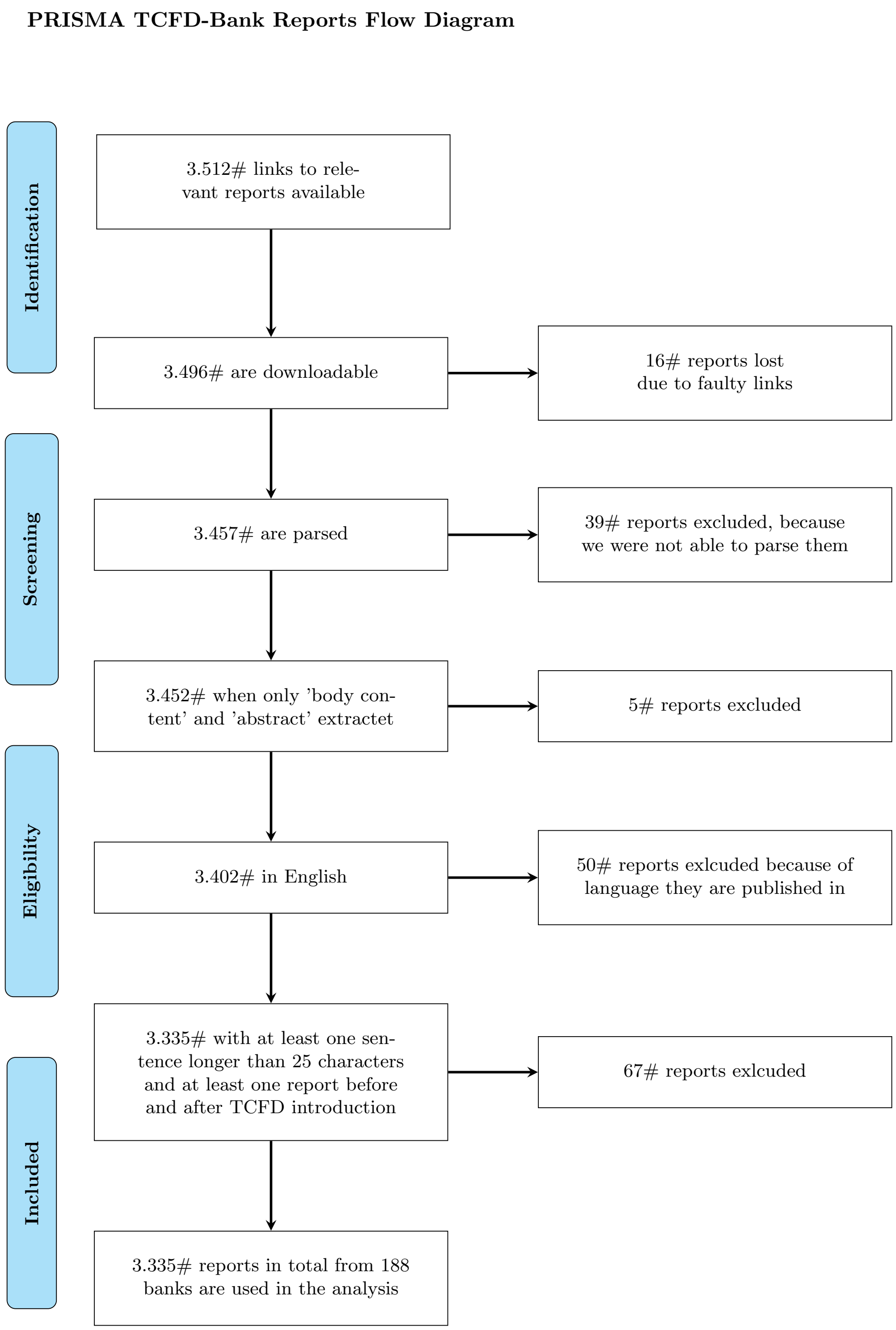

Figure 8: Data collection steps