TCFD-NLP: Assessing alignment of climate disclosures using NLP for the financial markets

Rylen Sampson
Manifest Climate
rylen.sampson

Aysha Cotterill
Manifest Climate
aysha.cotterill

Quoc Tien Au
Manifest Climate
quoc-tien.au

Abstract

Climate-related disclosure is increasing in importance as companies and investors alike aim to reduce their environmental impact and exposure to climate-induced risk. Companies primarily disclose this information in annual reports or other lengthy documents where climate information is not the sole focus. To assess the quality of a company's climate-related disclosure, these documents, often hundreds of pages long, must be reviewed manually by climate experts. We propose a more efficient approach to assessing climate-related information. We construct a model leveraging TF-IDF, sentence transformers and multi-label k nearest neighbors (kNN). The developed model is capable of assessing alignment of climate disclosures at scale, with a level of granularity and transparency that will support decision-making in the financial markets with relevant climate information. In this paper, we discuss the data that enabled this project, the methodology, and how the resulting model can drive climate impact.

1 Introduction

The effects of climate change are accelerating globally, and it has been established that corporations are at risk of major financial loss due to climate impacts. A 2019 Carbon Disclosure Project report found that 215 of the world's largest companies reported nearly USD $1 trillion at risk from climate change over the following five years[1]. This has motivated stakeholders to request transparency from companies regarding climate change induced risk and their strategy to address these risks - also known as climate-related financial information.

In recent years, new regulatory standards and frameworks have been introduced to provide guidance to companies on what climate information they should be disclosing. Most notable among these are the 11 recommendations introduced in 2017 by the Task Force on Climate-related Financial Disclosure (TCFD). Each recommendation falls under one of four pillars (Governance, Strategy, Risk Management, and Metrics and Targets), and outlines information that a company should include in its reports.

Despite the existence of climate disclosure regulations and frameworks such as the TCFD, there is a high degree of inconsistency in how companies communicate climate-related financial information. This information may be spread across discretionary reports (e.g. sustainability reports, climate reports, ESG reports) as well as regulatory reports (e.g. proxy filings, annual reports), and is often a small portion of the report in question. Manually parsing through reports to extract TCFD-relevant information remains a time-consuming task, impeding large-scale analysis of climate reporting.

Previous attempts to solve this problem employ natural language processing (NLP) methods and have seen varying levels of success. A critical oversight of past work citing the TCFD is that their datasets do not contain all of the TCFD recommendations, or that there is a lack of transparency. Our preliminary work and results show how we could assess TCFD-alignment across all 11 recommendations.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

Additionally, we benchmark three models leveraging recent research in assessing TCFD-alignment and NLP on our dataset. Applied at the recommendation level, these models would enable automatic assessment of TCFD-alignment of companies across geographies and sectors. It would also create a path to drive climate action in the financial markets by benchmarking climate disclosures at scale and providing best practices in terms of climate risk management.

2 Related work

Introduction of the TCFD recommendations prompted researchers to analyze trends and identify generalizations from existing climate-related financial disclosure patterns. While one-off analyses explore at a small-scale whether companies are TCFD-aligned, there is a push for market-wide assessment to compare companies against each other and drive climate action. The first organization to conduct such a review is the TCFD, in their status reports from 2020 and 2021. In these reports, they mention the use of artificial intelligence (AI) to review reports from 1,701 and 1,651 companies respectively [2, 3]. The organization only briefly reveals how AI is used to assess TCFD-alignment, meaning that there is no way to evaluate their methodology or benchmark new models.

Other research has looked at a variety of methods for analyzing climate-related content in reports. One approach queried reports using climate-relevant keywords [4] - further models can then be applied to the queried text [5]. Despite reducing the time required to assess company reports, this approach has a high likelihood of missing relevant information given that a keyword list cannot fully encapsulate climate disclosure language. This issue is compounded as climate language evolves and further information is missed.

Language models (LMs) also show promise for identifying climate content in corporate reports. BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. in 2018, is a transformer-based language model that provides state-of-the-art performance for most NLP tasks. Climate-related financial information is no exception, and BERT has shown the capability to detect climate risk-related content in company reports [7, 8]. However, risk management is only one of the four TCFD pillars as such this is not comprehensive of a company's TCFD alignment.

Most relevant is the work of **climateBERT**, a DistilRoBERTa [9] model fine-tuned on climate-related financial disclosures to better incorporate context [10]. ClimateBERT has been applied to analyze whether companies who declared support for the TCFD are actually TCFD-aligned. This uncovered the reality that support for the TCFD is primarily *cheap talk* by companies [11]. Although their analysis is extensive in companies covered, it is limited to the four TCFD pillars and does not analyze coverage of the 11 recommendations.

3 Methodology

We collect and label more than 1000 regulatory and discretionary reports across numerous companies. Each report is reviewed by a team of climate experts using our organization's TCFD recommendation-aligned disclosure review methodology. Each review goes through a QA process to ensure label agreement. The resulting dataset contains approximately 2 million sentences labelled with the TCFD recommendations. These labels can be laddered up to the page and file levels (see table 1 for further details). Given the nature of the data, we approach this task as a multi-label classification problem. The dataset is split on a per-company basis into train, validation, and test sets with the split of companies in each set being 81%, 9%, and 10% respectively. All hyperparameters are chosen on the validation set, and all reported metrics are computed on the test set.

3.1 TF-IDF and Random Forest

TF-IDF (Term Frequency-Inverse Document Frequency) transformation is used to represent page text numerically. TF-IDF is a statistic that attempts to measure the *importance* of a term within a set of documents. Basic text processing steps involving stop word removal, lemmatization and bigram extraction are applied prior to the TF-IDF vectorization. The TF-IDF is built only using pages that contain TCFD-relevant information (positive pages), with a vocabulary length of 5000. Random Forests (RF) offer a robust classification model suited for large input dimensions. The multi-label RF

model is trained using downsampled training data with a ratio of 2 negative pages for every positive page.

3.2 Language models

We compare DistilRoBERTa [9] and ClimateBERT [10] for our multi-label classification task. Both are fine-tuned (FT) with a classification head with the optimal hyperparameters. We build paragraphs for the training data with a rolling window of 10 sentences. Each page will then have overlapping paragraphs containing up to 10 sentences. The goal is to add more context to each sentence, as the language models need it to perform better. The training data is also downsampled so that only negative paragraphs appearing next to positive paragraphs are used for training.

3.3 Sentence-transformers with multi-label kNN

Sentence-transformers [12] provide sentence embeddings trained specifically on pairs of sentences. The resulting embedding space better translates the semantic similarity between sentences compared to the BERT encoder. We also use the pre-trained sentence-transformers DistilRoBERTa model. Multi-label kNN [13] using the cosine distance is then applied for our multi-label classification task.

3.4 Stacked models

Given the nature and imbalance of the dataset, we want to filter out pages that are not climate-related at all, before applying any language model. The TF-IDF + RF model (1) offers a robust keyword-based approach. In the stacked versions of the models, we first apply the model (1), then the language models on the positive pages found by model (1). The probability threshold used for model (1) is a hyperparameter that is tuned.

4 Preliminary results

We use the ROC-AUC score to evaluate our models on the test set. Table 2 shows the test metrics at the file level for every TCFD recommendation, as well as the average score across all recommendations. The predictions at the file level come from heuristics defined on the scores generated at the paragraph or the page level. The score of a file is the average of the top N scores found in this file. N is a hyperparameter that is tuned.

Table 2 shows that stacking keyword and embedding-based models give the best results. First finding relevant climate texts, and then labelling them seems to be the optimal approach. Model (1) is a strong keyword-based baseline that filters pages of text efficiently. It predicts Metrics and Targets recommendations better than language models. One reason could be that the information for those recommendations is often displayed as tables, without proper sentences. Single keywords such as units (e.g. CO2/kg) may contain enough signal.

The stacked model using TF-IDF (1) and multi-label kNN (2) has the best ROC-AUC scores both overall and across the majority of recommendations. It combines the simplicity of keyword detection, and the semantic ability of sentence transformers. Additionally, by design, the model enables better transparency by giving access to the most similar climate disclosures for a given climate text excerpt. Not only will companies be able to compare themselves to others in terms of climate risk management, but they will also know what good disclosure looks like. We want to push for transparency in the financial markets, using an approach of surfacing peer examples and positive reinforcement to encourage climate action.

Future improvements include better table and text extraction from PDF files so that we can capture Metrics and Targets information contained in tables. We also want to experiment with different stacking methods and heuristics for each recommendation to adapt our methods to each recommendation specifically. Finally, given the boost in performance using fine-tuning on relevant financial and climate disclosures, we want to fine-tune the sentence transformers on pairs of similar climate texts.

References

[1] Major risk or rosy opportunity: Are companies ready for climate change? Technical report, Carbon Disclosure Project, 2019.

[2] Task Force on Climate-related Financial Disclosures. Task force on climate-related financial disclosures. 2020 status report. Technical report, Financial Stability Board, 2020.

[3] Task Force on Climate-related Financial Disclosures. Task force on climate-related financial disclosures. 2021 status report. Technical report, Financial Stability Board, 2021.

[4] Kevin L Doran and Elias L Quinn. Climate change risk disclosure: a sector by sector analysis of sec 10-k filings from 1995-2008. *NCJ Int'l L. & Com. Reg.*, 34:721, 2008.

[5] Alexandra Luccioni and Hector Palacios. Using natural language processing to analyze financial climate disclosures. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*, 2019.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. doi: 10.48550/ARXIV.1810.04805. URL `https://arxiv.org/abs/1810.04805`.

[7] David Friederich, Lynn H Kaack, Alexandra Luccioni, and Bjarne Steffen. Automated identification of climate risk disclosures in annual corporate reports. *arXiv preprint arXiv:2108.01415*, 2021.

[8] Elham Kheradmand, Didier Serre, Manuel Morales, and Cedric B Robert. A nlp-based analysis of alignment of organizations' climate-related risk disclosures with material risks and metrics. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. URL `https://www.climatechange.ai/papers/neurips2021/69`.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019. doi: 10.48550/ARXIV.1907.11692. URL `https://arxiv.org/abs/1907.11692`.

[10] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

[11] Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, page 102776, 2022.

[12] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[13] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2006.12.019. URL `https://www.sciencedirect.com/science/article/pii/S0031320307000027`.

Appendix

Table 1: Number of labeled sentences, pages, and files by TCFD recommendation. Totals with text having no TCFD recommendations present are included for reference in the column headers

TCFD Recommendation	**Sentences (Total = 1,912,424)**	**Pages (104,829)**	**Files (1,090)**
Governance A	6,614	1,226	501
Governance B	7,694	1,287	454
Strategy A	37,679	5,119	769
Strategy B	22,074	2,904	619
Strategy C	8,795	1,342	477
Metrics and Targets A	15,860	2,965	630
Metrics and Targets B	5,039	974	400
Metrics and Targets C	19,237	3,419	701
Risk Management A	22,433	2,897	648
Risk Management B	13,103	1,627	496
Risk Management C	3,845	555	259

Table 2: ROC-AUC scores at the file level for every TCFD recommendation.

File level ROC-AUC	TF-IDF (1)	kNN (2)	FT ClimateBERT (3)	FT distilRoBERTa (4)	(1) + (2)	(1) + (3)	(1) + (4)
Average AUC	0.867	0.847	0.852	0.819	**0.884**	0.865	0.857
Governance A	0.891	0.854	0.865	0.854	**0.91**	0.887	0.875
Governance B	**0.835**	0.812	0.809	0.795	0.832	0.823	0.823
Strategy A	0.844	0.875	0.834	0.766	**0.888**	0.841	0.838
Strategy B	0.877	0.874	0.853	0.824	**0.896**	0.864	0.852
Strategy C	**0.881**	0.845	0.856	0.805	0.866	0.86	0.854
Risk Management A	0.874	0.919	0.89	0.845	**0.935**	0.893	0.874
Risk Management B	0.839	0.882	0.871	0.836	**0.882**	0.88	0.862
Risk Management C	0.864	**0.941**	0.862	0.836	0.931	0.892	0.882
Metrics and Targets A	0.859	0.758	0.869	0.79	**0.895**	0.888	0.876
Metrics and Targets B	**0.932**	0.8	0.87	0.864	0.875	0.877	0.882
Metrics and Targets C	**0.843**	0.773	0.795	0.796	0.814	0.816	0.813