Analyzing Sustainability Reports Using Natural Language Processing

Alexandra (Sasha) Luccioni
Université de Montréal + Mila

Emily (Emi) Baylor
McGill University

Nicolas Duchene
Université de Montréal

Abstract

Climate change is a far-reaching, global phenomenon that will impact many aspects of our society, including the global stock market [1]. In recent years, companies have increasingly been aiming to both mitigate their environmental impact and adapt their practices the changing climate context. This is reported via increasingly exhaustive reports, which cover many types of sustainability measures, often under the umbrella of Environmental, Social, and Governance (ESG) disclosures. However, given this abundance of data, sustainability analysts are obliged to comb through hundreds of pages of reports in order to find relevant information. We have leveraged recent progress in Natural Language Processing (NLP) to create a custom model, ClimateQA, which allows the analysis of financial reports in order to identify climate-relevant sections using a question answering approach. We present this tool and the methodology that we used to develop it in the present article.

1 Introduction

In the coming years and decades, climate change will have a major impact on many global systems and structures, from agriculture to transportation and urban planning, affecting individual and collective behavior [2]. Its impact on the global stock market is predicted to be extensive, with damages estimated in the trillions of dollars [1]. However, it is difficult to predict exactly how and where climate change will impact financial assets, largely due to the lack of quantitative data on the subject. Nonetheless, gathering data regarding the risks and exposure that climate change poses to specific companies, as well as organizations' efforts to address and mitigate this risk, is a key part of the effort of predicting the extent of climate change impacts on the stock market. This data is often in textual format, buried in hundreds of pages of financial documents which must be manually analyzed, requiring significant time and effort. In fact, disclosing climate change risks and liabilities currently consists of a mix of mandatory and voluntary initiatives, resulting in disclosures that are often very heterogeneous and lack structure with regards to the subjects they cover, the metrics they use and the extent to which climate risk is quantified.

The Task Force on Climate-related Financial Disclosures (TCFD) was founded in 2015, with the goal of improving the state of climate disclosing and encouraging companies to increase their climate transparency. In 2017, the TCFD released a set of recommendations to help structure and formalize companies' sustainability and climate risk reporting [3]. One of the key proposals that they made was a set of 14 questions to guide sustainability reporting, covering many different topics, from quantifying greenhouse gas emissions to ensuring the proper identification of climate-related risks and/or opportunities by the board of directors. In recent years, the TCFD questions are increasingly

used to guide the analysis of climate risk disclosures, with analysts using them to assess the extent and type of climate exposure of companies and the risk this poses to their business.

The TCFD itself also uses the 14 questions to publish their yearly status reports on the progress of sustainability reporting. In 2019, they estimated that out of the 1000 companies whose reports they analyzed, only 29% made relevant climate disclosures, stating that they were *"concerned that not enough companies are disclosing decision-useful climate-related financial information"* [4]. However, this figure is hard to calculate since it is based on a subset of the total number of companies and since the current approach used to analyse the reports is based on keyword search for analyst-defined terms. This suggests that much of the available data regarding financial climate disclosures is under-utilized or simply ignored. In recent years, a few proposals have also been made regarding using NLP for analyzing sustainability reporting, signifying that this field is gathering momentum [5, 6]. The goal of our research project is to use Natural Language Processing (NLP) to create a tool allowing more efficient analysis of financial reports, reducing the time and effort required to identify climate-relevant disclosures.

2 Model and Approach

Transformer models [7] have had substantial success in many NLP tasks, from question answering [8] to machine translation [9] and natural language inference [10]. In recent years, different variations of architectures such as BERT [11] have also been used for domain-specific applications, from biomedical text mining [12] to sentiment analysis [13], as well as finance [14]. We have adopted this Transformer-based approach to develop ClimateQA, our tool for extracting climate-relevant passages in financial documents; we will describe our approach in more detail below.

2.1 Pretraining on Unlabelled Data

Progress in NLP applications in finance has proven to be challenging notably because of the specialized language used: terms such as 'bull' and 'short' do not have the same meanings in finance as in general discourse [1] , whereas technical terms such as 'liquidity' or 'Keynesian' may not even be present in training corpora. In fact, research in financial NLP has found that using general-purpose NLP models trained on corpora such as Wikipedia and the Common Crawl fail to capture domain-specific terms and concepts which are critical for a coherent representation of the financial lexicon, and are therefore difficult to use out-of-the-box for financial tasks [14].

To address this issue, we have scraped 2,249 publicly available financial and sustainability reports from sources such as EDGAR and the Global Reporting Initiative databases. These documents covered over 10 years of reports from a variety of publicly-traded companies, in sectors ranging from banking to agriculture. We extracted the raw text from the PDFs of the reports using the Tika package and used the raw text output to pre-train a word embedding model on the documents. Our hope in doing so was to reflect the context-specific nature of financial discourse, and to better represent the vocabulary used in sustainability reports.

2.2 Fine-tuning on Labeled Data

Given the prevalence of the TCFD questions and the diversity of subjects that they cover, we endeavored to use them to guide the analysis of sustainability reports. More specifically, we framed our task as one of question answering: given one of the 14 TCFD questions and a set of candidate sentences extracted from a financial report, we trained a model to determine whether or not a given sentence is a potential answer to one of the questions (see Table 1 for examples of TCFD questions and answers). In order to gather labeled data, we reached out to a team of sustainability analysts, who were able to provide us with a small set of financial reports from previous years, hand-labeled using the 14 TCFD questions. Based on these reports, we constructed our training set for the question answering task: positive examples consisted of pairs of questions and sentences which contained the answers to the questions, whereas negative examples were generated by pairing the remaining sentences with the questions that they did not answer. We split our dataset into three sets (training, development and test) and took stratified random samples of each set, separating the documents on a

---

[1] A 'bull' market is one that is on the rise; to 'short' a stock means investing in such a way that the investor will profit if the value of the stock falls.

per-company basis. Our ClimateQA model was therefore trained on 15,000 negative examples and 1,500 positive examples, with the development set comprised of 7,500 negative examples and 750 positive examples, and the test set had 1,200 negative and 400 positive examples.

Table 1: Examples of Question-Answer pairs from our corpus

TCFD Question	Answer Passage
Does the organization describe the board's (or board committee's) oversight of climate-related risks and/or opportunities?	*The Company's Audit Committee has the delegated risk management oversight responsibility and receives updates on the risk management processes and key risk factors on a quarterly basis.*
Does the organization describe the climate-related risks or opportunities the organization has identified?	*The availability and price of these commodities are subject to factors such as changes in weather conditions, plantings, and government policies*

3 Results

3.1 Comparing Large and Base Models

To ensure the best performance of our model, we decided to use the RoBERTa (Robustly optimized BERT approach) architecture, whose performance was found to be either matching or exceeding that of the original BERT architecture on a variety of tasks [15].However, there are in fact two different versions of the RoBERTa architecture: RoBERTa-Large, which has 355M parameters, and RoBERTa-Base, which has only 125M. In traditional NLP applications, RoBERTa-Large is often the preferred version of BERT due to it being considered larger and more accurate [15]. However, more parameters comes with significantly higher memory requirements and a longer training time – in our case, RoBERTa-Large took almost 12 hours to train on a 12GB GPU, whereas RoBERTa-Base took less than 5 hours. In order to choose between the two, we compared their performance. A metric we were particularly interested in was the difference between the validation and test F1 scores, since we see this as a good indicator of the model's ability to generalize. As can be seen in Table 2 while RoBERTa-Large does slightly better than RoBERTa-Base, the differences are minor, between 0.5 and 2.5%. Given the major difference between the two models in terms of memory constraints and training time, and the importance of energy efficiency in choosing neural network architectures [16], we decided to use RoBERTa-Base for further hyper-parameter tuning and online deployment.

Table 2: Comparing RoBERTa-Large and RoBERTa-Base

	Train F1	Validation F1	Test F1	Val-Test Difference
RoBERTa-Large	99.9%	92.2%	85.5%	-6.7%
RoBERTa-Base	99.9%	91.7%	82%	-9.7%

3.2 Analyzing Results by Sector and by Question

Both the labelled and unlabelled datasets that we trained ClimateQA on were from a variety of sectors, which made them differ slightly in the terminology used and the types of disclosures made. For instance, the climate change-related risks identified by insurance companies was mostly due to physical risk (e.g. due to coastal properties being damaged by repeated flooding), whereas those identified in the energy sector can be linked to market or legislative risk (e.g. for oil and gas companies impacted by higher taxation rates or customers switching to renewable energy). When analyzing our results by sector, we found that the Energy sector had the best results, most likely due to the homogeneity of the companies in the labeled data we received, whereas Materials and Buildings had the biggest validation-test difference, given the large variety of companies (and disclosure types) that it encompassed. Overall, the generalization capacity of ClimateQA in respect to the different sectors was good across the board, with an average 13.3% difference between validation and test scores.

We also looked at ClimateQA's performance on each of the 14 TCFD questions (a detailed per-question analysis is presented in Appendix 1). This comparison is difficult to do systematically,

Table 3: ClimateQA Results by Sector

	Validation F1 Score	Test F1 Score	Val - Test Difference
Agriculture, Food & Forests	89.4%	72.1%	-17.2%
Energy	94.2%	89.8%	-4.4%
Banks	91.9%	86.6%	-5.3%
Transportation	86.9%	72.5%	-14.4%
Insurance	92.9%	78.7%	-14.2%
Materials & Buildings	91.8%	67.6%	-24.2%
Average across sectors	**91.7%**	**82.0%**	**-9.7%**

because some questions are answered much more frequently than others. For instance, Question 1 (*"Does the organization describe the board's oversight of climate-related risks and/or opportunities?"*) is answered by all companies in our dataset, whereas Question 12 (*"Does the organization disclose Scope 1 and Scope 2, and, if appropriate Scope 3 GHG emissions?"*) is answered much less diligently, with only 8% of companies providing answers. Nonetheless, there are some observations can be made, for instance the fact that Question 4 (*"Does the organization describe time frames associated with its climate-related risks or opportunities?"*) shows the worst performance; we believe that this is due to its genericity, since time frames can be anything from numbers, e.g. 'by 2025', to time horizons, e.g. 'within the next five years. Furthermore, Question 10 (*"Does the organization describe how processes for identifying, assessing, and managing climate-related risks are integrated into the organization's overall risk management?"*) has the most significant difference between validation and test data, up to 51%. We believe that this is due to the diversity of answers, since we found that these processes that may vary very much depending on the size and type of company that is involved. Nonetheless, it is interesting to observe which questions were easier to generalize for the ClimateQA model, and which needed more data to attain better performance.

4 The Final Tool and Future Work

While the model training and fine-tuning involved in our project was an interesting and worthwhile endeavor in itself, our end goal was always to create a user-friendly sustainability report tool that could easily be used by analysts. This is why we have spent a significant amount of time and effort deploying our ClimateQA model. To this end, the model is hosted on the Microsoft Azure cloud solution (see Appendix 2 for high-level architecture), allowing users to interact with a web application without needing ML expertise. Via the website, a user is able to upload PDF files to be analyzed and receive an output file containing the relevant passages, the TCFD questions they pertain to, and the model confidence in its prediction. In terms of the processing pipeline, once the PDF files are uploaded by the user, they are processed using an Azure ML pipeline composed of three steps: (1) Extraction of the raw text from the PDF, (2) Parsing and splitting of the text, and (3) Inference using the ClimateQA model, which identifies sections of the text that answer the TCFD question(s) submitted. The final output is sent to a cloud storage from which the user is able to download the results of the inference both in terms of the questions and the relevant answers extracted from the report they submitted. The process takes 5 to 15 minutes per report, which saves considerable time compared to a manual analysis of the data, which can take several hours.

We are working on two future directions for our project: improving the ClimateQA model itself and improving the user experience of the website. On the one hand, error analysis of our results has brought to light that sentences that were part of a table were often not identified by our approach. We are therefore currently working on improving our text extraction approach, notably by exploring commercial PDF extraction documents to identify rows and cells within tables in the reports themselves. We are also looking into ways of better utilizing the domain-specific word embedding models in our approach, since we believe that our model's generalization capacity can be improved by leveraging the finance-specific vocabulary that is present in the reports. On the other hand, we are working on improving the ClimateQA website so that it can display results interactively, allowing users to visualize the passages identified by the model within the documents. This visualization tool will help democratize ClimateQA by bridging the gap between end-users and a specialized climate finance model, all the while keeping a simple user interface abstracting away all implementation details.

References

[1] Simon Dietz, Alex Bowen, Charlie Dixon, and Philip Gradwell. 'climate value at risk'of global financial assets. *Nature Climate Change*, 6(7):676–679, 2016.

[2] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.

[3] FSB TCFD. Final report: recommendations of the task force on climate-related financial disclosures. *Financial Stability Board Task Force on Climate-related Financial Disclosures, available at: www. fsb-tcfd. org/wp-content/uploads/2017/06/FINAL-TCFD-Report-062817. pdf (accessed 15 January 2018)*, 2017.

[4] TCFD. 2019 status report. `https://www.fsb-tcfd.org/publications/tcfd-2019-status`, 2019.

[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] Jonas Becker. Recognition and assessment of climate disclosure in annual business reports with natural language processing. *Master's thesis, University of Frankfurt*, 2019.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[8] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[9] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. *arXiv preprint arXiv:1806.00187*, 2018.

[10] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[13] Saif M Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier, 2016.

[14] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

[15] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[16] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Appendix 1

Table 4: ClimateQA results by question

| TCFD Question	Validation F1 Score	Testing F1 Score	Val - Test Difference
1) Does the organization describe the board's oversight of climate-related risks and / or opportunities?	97.78%	84.85%	-12.93%
2) Does the organization describe management's role in assessing and managing climate-related risks and/or opportunities?	96.60%	84.75%	-11.85%
3) Does the organization describe the climate-related risks or opportunities the organization has identified?	90.61%	89.50%	-1.11%
4) Does the organization describe time frames (short, medium, or long term) associated with its climate-related risks or opportunities?	75.00%	N/A	N/A
5) Does the organization describe the impact of climate-related risks and opportunities on the organization?	90.91%	86.59%	-4.32%
6) Does the organization describe the resilience of its strategy, taking into consideration different climate-related scenarios, including a potential future state aligned with the Paris Agreement?	94.12%	N/A	N/A
7) Does the organization disclose the use of a 2C scenario in evaluating strategy or financial planning, or for other business purposes?	100.00%	100.00%	0.00%
8) Does the organization describe the organization's processes for identifying and/or assessing climate-related risks?	89.87%	81.08%	-8.79%
9) Does the organization describe the organization's processes for managing climate-related risks?	92.54%	60.00%	-32.54%
10) Does the organization describe how processes for identifying, assessing, and managing climate-related risks are integrated into the organization's overall risk management?	96.15%	44.44%	-51.71%
11) Does the organization disclose the metrics it uses to assess climate-related risks and/or opportunities?"	88.48%	90.67%	2.18%
12) Does the organization disclose Scope 1 and Scope 2, and, if appropriate Scope 3 GHG emissions?	97.50%	94.74%	-2.76%
13) Does the organization describe the targets it uses to manage climate-related risks and/or opportunities?	90.20%	98.31%	8.11%
14) Does the organization describe its performance related to those targets (referenced in question 13)?	93.33%	73.08%	-20.26%
Average	92.35%	82.24%	-10.98%

Appendix 2
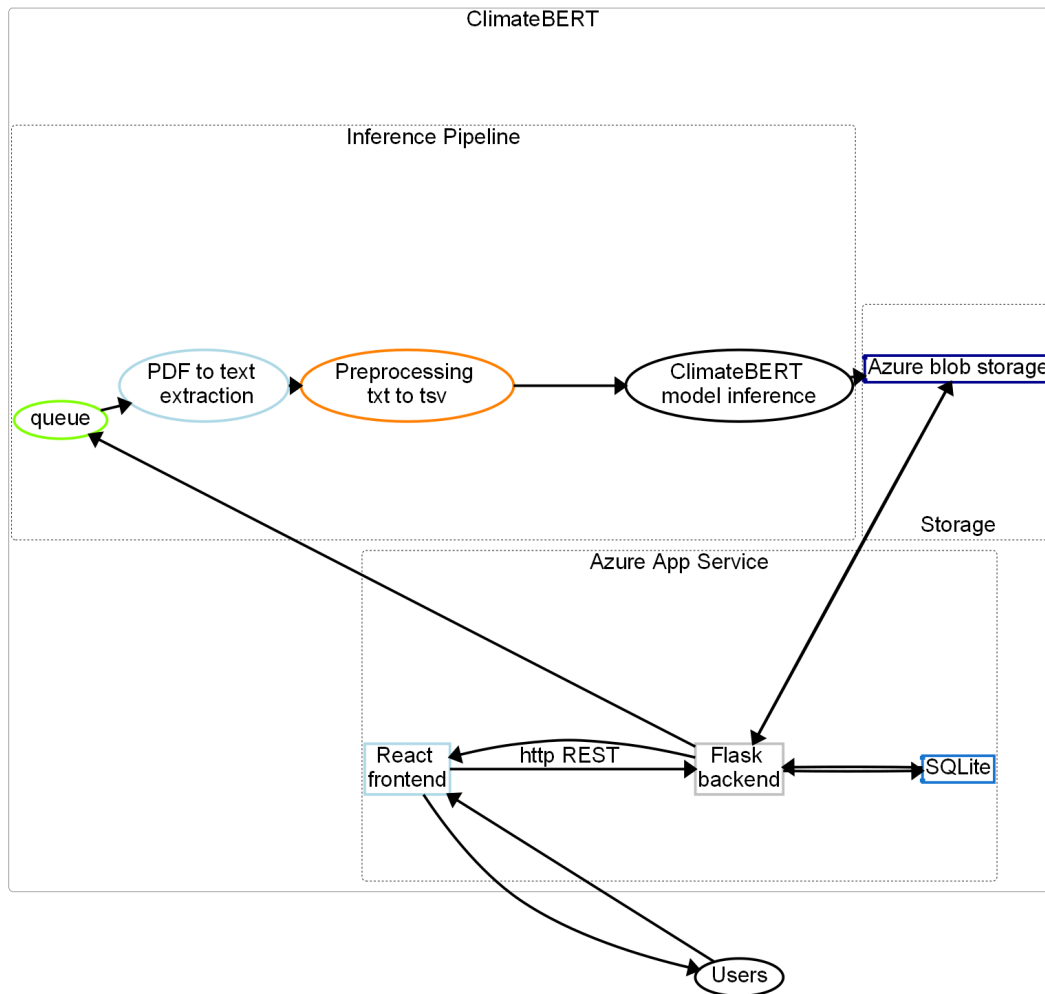
Figure 1: Architecture of the ClimateQA web application