Emerging Trends of Sustainability Reporting in the ICT Industry: Insights from Discriminative Topic Mining

Lin Shi
Stanford University
linshi@stanford.edu

Yen Nhi Truong Vu
Stanford University
ntruongv@stanford.edu

Abstract

The Information and Communication Technologies (ICT) industry has a considerable climate change impact and accounts for approximately 3 percent of global carbon emissions. Despite the increasing availability of sustainability reports provided by ICT companies, we still lack a systematic understanding of what impacts have been disclosed at an industry level. In this paper, we make the first major effort to use modern unsupervised learning methods to investigate the sustainability reporting themes and trends of the ICT industry over the past two decades. We build a cross-sector dataset containing 22,534 environmental reports from 1999 to 2019, of which 2,187 are ICT specific. We then apply CatE, a text-embedding-based topic modeling method, to mine specific keywords that ICT companies use to report on climate change and energy. As a result, we identify (1) important shifts in ICT companies' climate change narratives from physical metrics towards climate-related disasters, (2) key organizations with large influence on ICT companies, and (3) ICT companies' increasing focus on data center and server energy efficiency.

1 Introduction

The Information and Communication Technologies (ICT) industry is a sizable industry that accounts for an estimated global economic worth of $11.5 trillion and 15.5 percent of global GDP [1]. The growth of the ICT industry has increased its environmental and social impacts to a level of global concern. In particular, research shows that the ICT industry currently accounts for approximately 3 percent of the global annual carbon footprint. This percentage is expected to grow as the world is increasingly reliant on digital technologies [2].

At an industry level, an increasing but small number of ICT companies have demonstrated understanding of their environmental and social impacts within their supply chains through information disclosure and external verification [3]. Yet, we still lack a systematic understanding of what has been disclosed. We have limited contextual and empirical understanding of how sustainability practices have developed in the ICT sector and how they have evolved over time.

This paper demonstrates our effort to create a text database of companies' sustainability disclosures over the past two decades from the largest sustainability reporting database. Subsequently we use unsupervised natural language processing (NLP) methods to uncover the shifting themes and narratives from companies in the ICT industry over time. In particular, we focus on companies' sustainability disclosures relevant to climate change and energy programs. To extract information relevant to these specific topics, we propose to shift from the standard Latent Dirichlet Allocation [4] topic modeling tool and instead apply CatE, a text-embedding based, category-guided topic mining method [5], to retrieve representative words and phrases that showcase how companies interpret and establish policies regarding climate change over time.

Through our experiments, we found that ICT companies' sustainability reporting related to climate change has shifted from the initial focus on the heat-trapping gases and pollutants contributing to climate change towards climate-related extremes and disasters intensified by climate change. Over time, the climate policy narratives have shifted from mitigation and reduction of greenhouse gas emissions to adaptation and

Tackling Climate Change with Machine Learning workshop at NeurIPS 2020.

preparedness for the impacts of climate change. Over time, nonprofits and international climate conferences have been increasingly and consistently mentioned by ICT companies.

2 Relevant Works

Computational text analysis has been adopted as a beneficial tool in environmental studies only in recent years. Researchers have used it for environmental meta-reviews and organizations' sustainability disclosures, where it helps create "comprehensive, replaceable, interdisciplinary reviews that provide rapid, up-to-date, and policy-relevant reports of existing work" [6]. Early studies used word frequency analysis to understand the topical focus of companies' sustainability programs [7]. More recently, topic modeling has been applied as a tool in reviewing textual data in climate change related areas of environmental science, including identifying emerging themes in life cycle assessment literature as well as companies' sustainability reporting [8, 9]. The most commonly used technique is the Latent Dirichlet Algorithm (LDA) [4]. LDA is an unsupervised method that learns topic-word and document-topic distributions by modeling the generative process of the corpus. In prior analyses using text mining, LDA was used to explore main topics in companies' sustainability reporting between 1995 to 2005 and has revealed a broad picture of companies' sustainability practices and generic reporting themes in economic, social, and environmental dimensions [9].

However, a constraint of applying LDA is that the topical results are generic and non-discriminative. It does not enhance our understanding of specific sustainability issues such as climate change. As a result, LDA is not a suitable tool for more specific discovery-focused questions, such as "How do companies' sustainability reporting and policy shift around climate change?" In the NLP literature, there is a line of research on this issue, where users can provide a list of seed words to guide the topic discovery process [10, 11, 5]. In particular, we choose to employ the CatE model [5], which is a method that learns discriminative category-guided word embeddings and selects category-representative keywords based on embedding similarity and word distributional specificity. As a result, CatE learns an embedding space that best separates the user-interested set of categories and thus retrieve specific words that are specific to each topic.

3 Dataset

We build a large database containing 22,534 environmental reports by 8,088 companies in 118 countries from 1999 to 2019. The data are scraped from the Global Reporting Initiative (GRI) database [12], the largest publicly-available global sustainability reporting database. Companies voluntarily provide their metadata as well as sustainability reports to GRI, and they are also responsible for maintaining and hosting the reports. As a result, in addition to scraping the data using the links provided in GRI database, we conducted additional data collection via the Internet Archive and Google search engine when the url provided by GRI is not valid.

In addition to the original reports, which are mostly PDFs, we also provide extracted texts which are parsed using the Apache Tika engine [13]. We choose Tika because it is the de facto standard technology for textual content and metadata extraction, and it has a robust language detection algorithm. For each text file provided in the dataset, our dataset also includes metadata information such as the company name, size, country, etc. (Appendix A). This information is scraped from GRI database as well as file metadata. In the original GRI database, there is noise in the year label; for example, a report on the year 1999 but published in 2006 may be labeled as 2006. In this case, we manually correct the report year to 1999. In our dataset, we similarly corrected the labels of reports whose published year differs from report year by more than 1 year.

In this study, we concern ourselves only with English reports from the ICT industry. The ICT companies are selected based on their sectors. In this study, we include "Computers", "Technology Hardware," and "Telecommunications" as part of the ICT industry. After filtering, we have a total of 2,187 documents from 429 companies in 77 countries.

4 Method

Preprocessing We remove special characters, common English stopwords as well as company names and standalone numbers from our corpus. Removing company names and standalone numbers is important because the text documents are parsed from PDFs, where company names and meaningless numbers may appear widely in the text due to their presence in every page header and footer. In our experiments, keeping the company names and numbers negatively affects the quality of topic modeling.

Phrasal Segmentation In order to retrieve not only representative words but also quality phrases for topic modeling, we also apply AutoPhrase [14] to segment our corpus. This is critical because without phrase mining, topic modeling can only return single words, while we may be more interested in concepts such as "greenhouse gas", "carbon offset", "energy efficient". AutoPhrase is chosen as it is an automated phrase mining framework that incorporates quality phrases from public knowledge databases such as Wikipedia, and it also allows incorporation of domain knowledge for such phrase mining. We provide seed phrases (Appendix B) that are collected from up-to-date conference summaries from Business for Social Responsibility (BSR) and Electronics Goes Green (EGG) 2020+ conference proceedings [15, 16].

Mining Topic Representative Keywords We apply the CatE algorithm [5], which leverages user provided categories to learn category-distinctive word embeddings. The algorithm then uses the learned word embedding similarity and word distributional specificity to generate relevant keywords that represent topics of user interests. Our procedure is as follows:

1. Identify key topic names of interest provided by domain experts. Here, as a first exploratory pass, we investigate two topics "climate change" and "energy". In general, each topic can be represented by a list of seed words; for example, an expert could guide the model by providing keywords such as "global warming", "mitigation," and "emission" for the "climate change" topic.

2. Bootstrap CatE n times to extract K representative keywords each time for each topic, and choose $s < n$ results with highest topic coherence score. Out of the set of keywords represented in these s runs, extract the R keywords that appear the most times. These steps are implemented to reduce model variance. In our experiments, we choose $n = 15$, $s = 10$, $K = R = 100$.

The CatE algorithm [5] allows initialization from pretrained word embeddings such as Word2Vec [17], GloVe [18], etc. Here, we use pretrained 200-dimensional spherical text embeddings [19] because embedding in spherical space provides state-of-the-art performance on word similarity and document clustering, which is beneficial for the topic mining task.

5 Results

In order to visualize the important subtopics within each of the two topics "climate change" and "energy", we perform k-means clustering [20] on the retrieved topic representative words. Then, principal component analysis is applied to project the 200-dimensional word embeddings onto a visualizable 2-dimensional space spanned by the first two principal components that capture the highest variance in the embeddings.

Climate Change Narratives Figure 1 provides clustering visualization for the top 50 words representing the climate change topic. Our results suggest that in the early years (before 2010) of reporting, ICT companies focus on reporting heat-trapping gases and pollutants contributing to climate change and mitigation as a

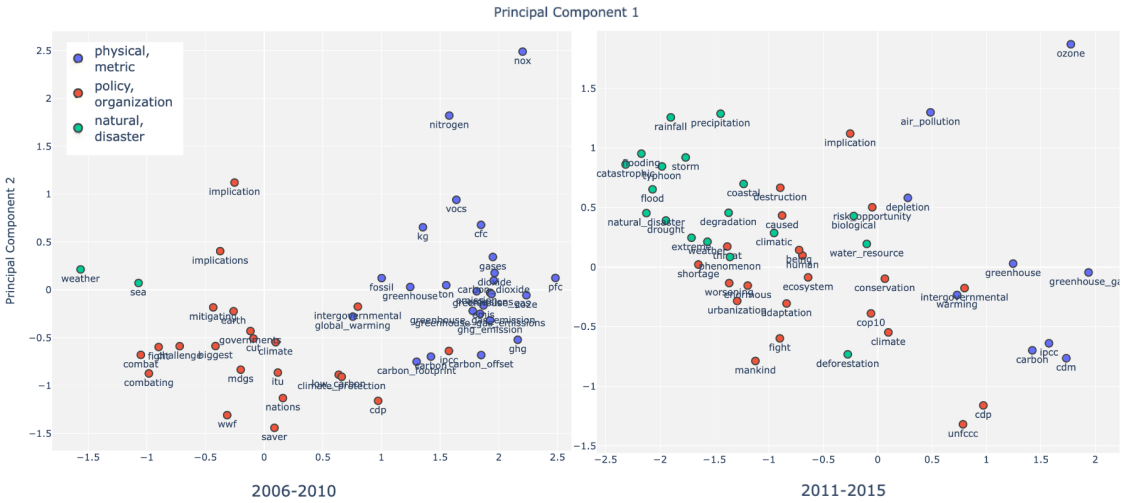

Figure 1: Top 50 words and phrases representative for the topic "climate change"

strategy. Specific greenhouse gases and metrics such as carbon, NOx, CFCs, and PFCs are consistently covered and discussed. Since 2010, ICT companies incorporate more organizations and policy in their sustainability reporting and mention adaptation as a strategy. Additionally, there are increasing number of direct mentions of "disaster" and "catastrophe" as well as descriptions about climate-related extremes and disasters intensified by climate change, such as storm, flood, and drought. Quantitatively, we found that the percentage of documents mentioning specific greenhouse gas substances like CFCs, PFCs steadily drops from 67.6% to 30.0% over 2006-2018, while the percentage of documents mentioning a climate-related disaster steadily increases over the same period.

Institutional Influence Based on the top 100 words and phrases representative for the topic "climate change," we identify a comprehensive list of organizations and policies in companies' reporting (Appendix C). We select 8 primary organizations to highlight their percentage of appearance between 2006 to 2018 in Figure 2. International non-profit organizations, inter-governmental agencies, and industry-wide initiatives are consistently mentioned in ICT companies' reporting over time. The Carbon Disclosure Project (CDP) is the most influential among them on the basis of appearance. Global e-Sustainability Initiative (GeSi) is the only ICT industry specific organization that's been consistently mentioned.

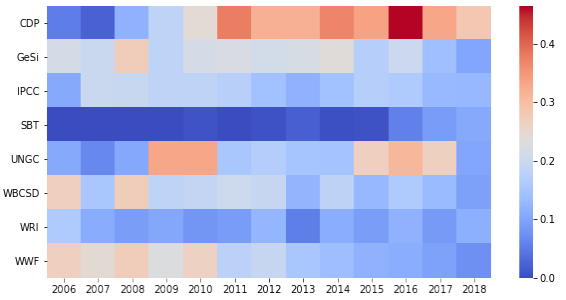

Figure 2: Heatmap showing percentage of documents mentioning organizations over years. Organization acronyms are defined in Appendix C.

Energy Programs Another trend we discover is the shifting emphasis in ICT companies' energy programs. Before 2010, ICT companies' energy programs fall into two major categories: building energy and energy production. In recent years, ICT companies start to emphasize data center and server energy programs in the context of energy savings and efficiency (Figure D.1 in the Appendix).

6 Conclusion and Future Work

Our main discovery is that ICT companies collectively have shifted narratives on climate change from focusing on describing the physical drivers contributing to climate change towards emphasizing climate-related disasters exacerbated by climate change. Given that ICT companies who report to GRI are early adopters of sustainability practices, this finding suggests that a potential opportunity to motivate other companies in the industry to initiate their own climate policy is to educate them on climate-related disasters and consequences.

We identified a comprehensive list of institutions that frequently appear in ICT companies' reporting, suggesting their influential role in ICT companies' climate change decision-making. Studies in corporate environmental management have used institutional theory to analyze stakeholders that encourage firms to adopt environmental management practices beyond regulatory compliance [21]. Although there is evidence of stakeholders imposing pressure on companies to move beyond compliance, the role of specific organizations within a particular industry is often unclear. Our results provide initial insights towards understanding and discovering the amount of relative influence among critical stakeholders in the ICT industry.

In addition, sustainability focus, priorities, and approaches within specific industrial sectors are traditionally studied using methods such as manual coding, survey, and interview. Text analysis provides a less resource intensive and flexible method complimentary to previous studies. It also allows deep-dives into specific topics of interest, uncovering the latest trends in the industry. For example, our results in the energy program suggest that ICT companies are diversifying their climate and energy policies from the emissions implications of buildings, facilities, and energy generation, towards full consideration of climate change impacts affecting data centers and servers.

Given these promising results, we believe that further expansion of this project could help identify meaningful incentives to guide companies towards designing and executing sustainability practices. In particular, we aim to apply this framework to topics such as certification and audits, which are also critical to the ICT industry. In addition, we could compare the ICT sector to other sectors that have longstanding history of sustainability practices (such as food, textile, and forest products) in order to identify influential factors and stakeholders that are both specific to and common across sectors on the topic of climate change and sustainability.

References

[1] Kwadwo Frimpong Sun and Makada Henry-Nickie Hao. Trends in the Information Technology Sector. Technical report, March 2019.

[2] Lotfi Belkhir and Ahmed Elmeligi. Assessing ICT global emissions footprint: Trends to 2040 & recommendations. *Journal of Cleaner Production*, 177:448–463, March 2018.

[3] Tannis Thorlakson, Joann F de Zegher, and Eric F Lambin. Companies' contribution to sustainability through global supply chains. *Proceedings of the National Academy of Sciences of the United States of America*, 115(9):2072–2077, February 2018.

[4] David M Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[5] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*, pages 2121–2132, 2020.

[6] Emily Grubert and Anne Siders. Benefits and applications of interdisciplinary digital tools for environmental meta-reviews and analyses. *Environmental Research Letters*, 11(9):093001, September 2016.

[7] Wan Te Liew, Arief Adhitya, and Rajagopalan Srinivasan. Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*, 65(3):393–400, April 2014.

[8] Emily Grubert. Implicit prioritization in life cycle assessment: text mining and detecting metapatterns in the literature. *The International Journal of Life Cycle Assessment*, 22(2):148–158, February 2017.

[9] Nadine Székely and Jan vom Brocke. What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLOS ONE*, 12(4):e0174807, April 2017.

[10] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, 2012.

[11] David Andrzejewski and Xiaojin Zhu. Latent dirichlet allocation with topic-in-set knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48, 2009.

[12] Global Reporting Initiative. Global reporting initiative database. `https://database.globalreporting.org`.

[13] Apache Software Foundation. Tika. `https://tika.apache.org`.

[14] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837, 2018.

[15] Business for Social Responsibility (BSR). The State of Sustainable Business in 2019. Technical report.

[16] *Electronics Goes Green 2020+ Proceedings*. Fraunhofer IZM and Technical University, 2020.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[19] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. Spherical text embedding. In *Advances in Neural Information Processing Systems*, pages 8208–8217, 2019.

[20] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[21] Magali Delmas and Michael W. Toffel. Stakeholders and environmental management practices: an institutional framework. *Business Strategy and the Environment*, 13(4):209–222.

Acknowledgement

We thank Professor Katharine Mach, Professor Adam Brandt, and Rebecca Miller for reviewing the study draft and providing constructive feedback. We also thank Aroma Mahendru for her valuable mentorship and the Climate Change AI NeurIPS 2020 team for organizing the mentorship program.

A Dataset Metadata

For each report in the dataset, we include metadata that comprises of the reporting company's name, size, type, listing status, sector, country, country status, region, publication year, reporting year, external assurance, reporting language, and whether the company is a part of/reports to/abides to the following organizations: Organisation for Economic Co-operation and Development (OECD), Carbon Disclosure Project (CDP), International Organization for Standardization (ISO), Sustainable Development Goals (SDGs).

B Seed Phrases Provided to AutoPhrase

climate change, sustainable products, economic sustainability, social sustainability, environmental health safety, local supplier, employee benefits, community investment, human rights, health safety, energy efficiency, energy efficient, renewable energy, natural gas, green chemistry, biodiversity, sustainable product, sustainable products, code of conduct, responsible business alliance, regulated substances, supplier risk, supplier training, conflict free, supplier audit, life cycle assessment, materiality analysis, material analysis, carbon offset, carbon footprint, circular economy

C Top Organizations and Protocols in 2006-2018

United Nations (UN), European Union (EU), Intergovernmental Panel on Climate Change (IPCC), United Nations Environment Programme (UNEP), European Telecommunications Network Operators (ETNO), United Nations Climate Change Conference (COP), United Nations Global Compact (UNGC), ICT for Energy Efficiency (ICT4EE), United Nations Framework Convention on Climate Change (UNFCCC), Global e-Sustainability Initiative (GeSi), Reducing Emissions from Deforestation and Forest Degradation (REDD), Carbon Disclosure Project (CDP), Green House Gas Protocol (GHGP), World Business Council for Sustainable Development (WBCSD), Science Based Target Initiatives (SBT), World Wildlife Fund (WWF), United States Environmental Protection Agency (EPA), World Resources Institute (WRI), International Telecommunication Union (ITU), Conference of the Parties 10 (COP10), Paris Agreement, Kyoto Protocol.

D Energy Representative Keywords

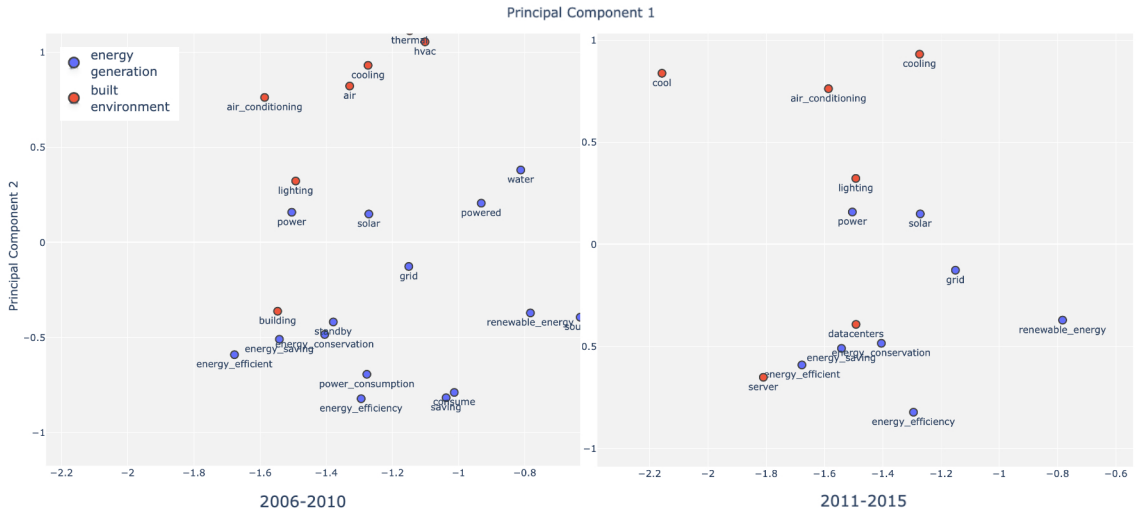

Figure D.1: Subcluster in top 50 words for topic "energy"