INF2209: Human-Centered Topic Models Juliette Zaccour
Fall 2022 1007943834

Part 2 (thematic analysis) starts page 19

"We're as committed as ever": Exploring Topics in Corporate Social Responsibility Reports from the Tech Industry

Introduction

Corporate Social Responsibility (CSR) can be defined as designating corporate practices and discourses that integrate the social and environmental responsibilities to which they are connected (Wickert & Risi, 2019). As stakeholders' expectations of corporations evolve, so does the conception of corporate responsibility. CSR in its modern definition was first captured in the 1970s-1980s, with a growing sense of awareness in the international community concerning sustainable development, and, by extension, corporate behavior. It is only in the 1990s, however, that CSR gained "popularity", with some industries being under scrutiny much earlier than others (e.g. oil companies). Yet, CSR was still described as a "niche strategy" in the late 1990s (Vogel, 2005), and only exploded as a generic and mainstream practice in the past two decades. The launch of the United Nations' Global Compact initiative in 2000 can be considered a springboard, as it directly aimed to promote responsible practices in businesses, as well as the Sustainable Development Goals (SDGs), created in 2015, which reframed corporations as main actors of change, and established concrete objectives and metrics of progress. More specifically, companies have started to disclose CSR information as a means of increasing transparency and dialogue with the public, and report on this aspect of the company's values, culture, and societal engagement. As a PR practice, then, CSR has gained popularity more or less recently depending on the sector, and through different channels, including re-branding, advertising, or through their annual reports—the medium of interest in this project. Initially, many corporations started incorporating "CR" sections in their yearly financial reports, but have since expanded as their own dedicated type of reports.

While an increasing number of corporations have joined the movement in releasing CSR reports, their content varies across individual corporations, industries, and from year to year. Indeed, it is expected to observe key differences in tone, themes and terminology when reading CSR reports from the early 2000s compared to the last couple of years. Mainly, the purpose of these reports can be resumed as attempts to "control the narrative", gain or reinforce the trust of stakeholders, and more broadly avoid legislation by showcasing their ability to self-regulate—a more convenient substitute. Indeed, Jackson and Apostolakou (2009) report that CSR practices have a strong inverse correlation with the strength of institutional coordination, regulatory standards and aggregate measures of social and ecological performance; in other words, the more proactive companies are with CSR, the less regulations are imposed on them. Additionally, CSR spendings have been shown to financially benefit companies[1]. As a result of these emerging company-issued discourse patterns, a number of issues have also reached new scales (e.g. greenwashing and social-washing); thereby, despite the development and formalization of corporate

[1] The Economist. (2015, June 25). "The halo effect". *The Economist*. https://www.economist.com/business/2015/06/25/the-halo-effect

responsibility, the movement has been continuously criticized as not only pointless, but also misleading, hypocrite and counter-productive (e.g. Musings, 2002). Discussing the ins and outs of CSR practices is beyond the scope of this project, but this allows to establish a key distinction: studying CSR output and CSR outcomes are two very distinct endeavors.

This project aims to explore CSR discourses from tech companies, comparing the performance of various topic modeling methods. The tech industry has increasingly been in the spotlight over the past decade, with issues related to privacy, biases, etc. being brought to the attention of the public (whether through legal, media or academic scrutiny), hence propelling tech companies into heated debates about their responsibilities, both in public and in regulatory discussions (Wickert & Risi, 2019). This context makes it particularly interesting to look into the CSR discourses within that industry, and how they have evolved throughout the years in response to shifts in public opinion and expectations. CSR has become a major topic in the field of business ethics (Kumar & Srivastava, 2021), and concepts like AI ethics and algorithmic fairness and accountability have similarly risen to the forefront of Machine Learning literature, but both perspectives/frameworks have rarely crossed paths. Yet, if we expect the industry to do better ethically, it is relevant to look into their habits/patterns of voluntary CR disclosure. Topic modeling has been used to analyze academic papers about CSR (Kumar & Srivastava, 2022), tweets related to CSR (Chae & Park, 2018), clustered reports from top CSR award-winning companies in Malaysia (Mangsor et al., 2022), CSR reports from oil companies (Jaworska & Nanda, 2018), and CSR reports from UK versus other European major companies (Goloshchapova et al., 2019), among a few other examples. However, this technique has not yet been applied to reports from the technology sector. Thereby, I am interested in exploring common trends in how tech companies self-present their engagement with social and environmental responsibilities, and what topics they choose to highlight, by comparing results from four topic modeling methods: latent Dirichlet allocation (LDA), Non-negative Matrix Factorization (NMF), Biterm topic modeling (BTM), and Correlation Explanation (CorEx).

Methods

Data collection and preprocessing

Most of the reports in the dataset were collected from the ResponsibilityReports database[2], which contains over 12,000 such documents from global companies. All reports from all companies listed under the “Technology” industry were downloaded. A few reports were downloaded directly from the company’s website (for instance, Netflix was missing from the ResponsibilityReports database, and was therefore retrieved so that the dataset contains all FAANG companies). Reports were not all released under the “CSR” label; other titles included “Impact report”, “Sustainability report”, “ESG report” (Environment, Social and Governance), “Diversity report”, “Digital impact & sustainability report”, “Global Citizenship report”, etc., all of which fall under the CSR umbrella term. The final dataset includes 1,096 reports from 209 companies (where most companies have around 3 reports in the corpus, the range being 1-24), spanning over 24 years (1997-2021).

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

Since these reports are only released as PDFs, a Python script was coded and used to extract all text as *.txt* files, allowing to automate the process and bypass the PDF encryptions that a number of the documents have. The *.txt* files were then imported into a Python notebook for analysis. While being the only efficient and free option, this conversion process did introduce significant noise in the corpus, with a number of missing or misplaced spaces within/between words and flawed character recognition, some documents being heavily affected and others remaining fairly clean. Unsurprisingly, the noisiness of the corpus did affect the topic modeling results, hence the number of pre-processing steps that were taken in order to minimize its impact.

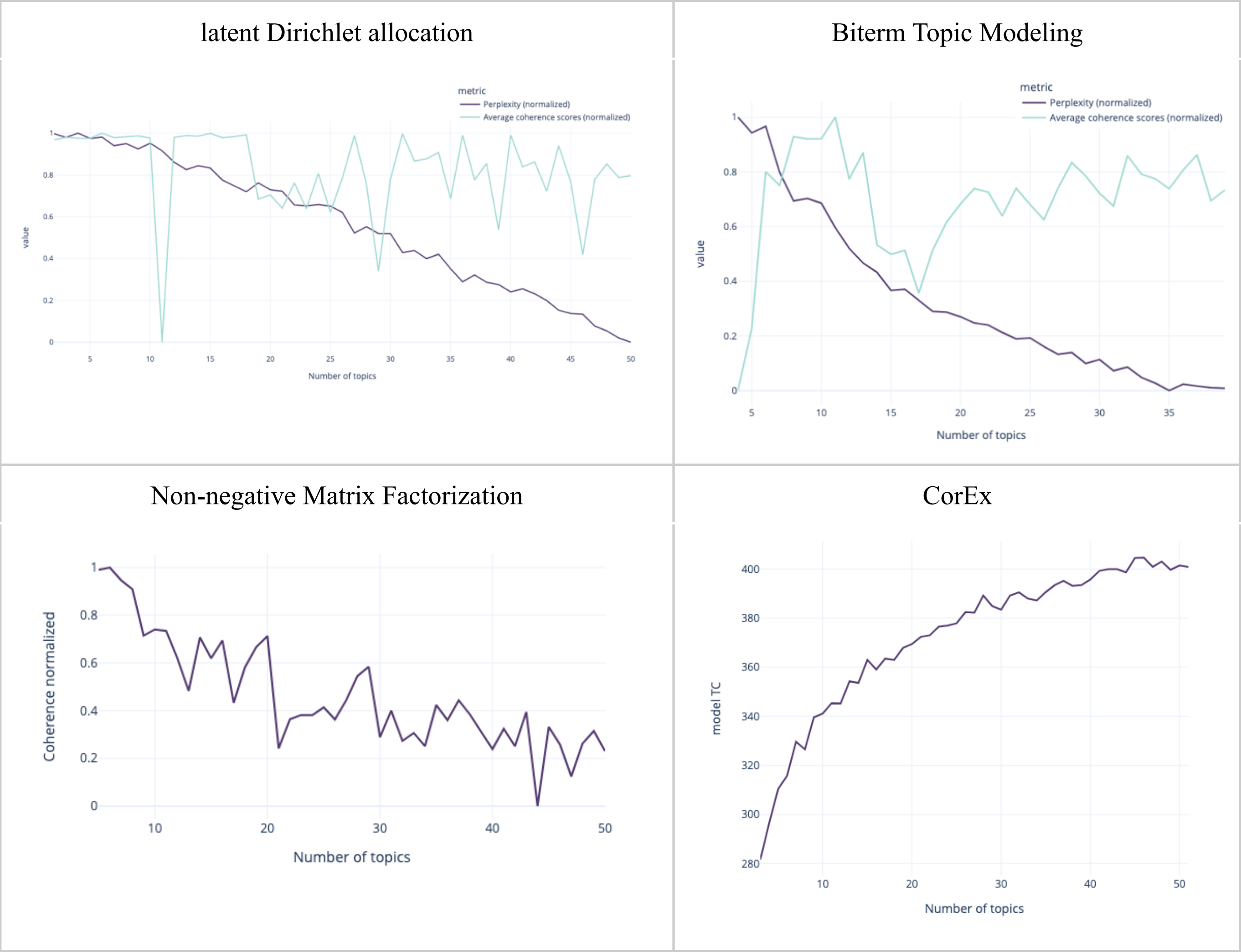
Once imported, the text files were structured into a dataframe, and company names and year of release were extracted from the filenames into separate columns. Using a list of the Global 500 companies, each company in the corpus was labeled as “Global 500” or not, in preparation for comparative analysis. The Global 500 list was selected in order to obtain a significant enough sample on both sides, which would not have been the case with the Fortune 100/500 list (the sample being worldwide and not solely US-based). 342 (31%) reports are from companies listed in the 2022 Global 500, 132 (12%) come from companies listed in the Global 100, and 25 (2.3%) come from FAANG.

As for the text pre-processing, all text was set to lowercase, numeric characters, punctuation and words shorter than 3 characters were removed, and generic stop-words were removed as well using the *little mallet wrapper* package predefined list. Words were tokenized, and then lemmatized.

From the term-frequency dictionary, I decided on a threshold for inclusion; all tokens that appear less than 100 times over the whole corpus were removed, allowing to eliminate 229,183 unique tokens from the corpus (the overwhelming majority of which were noise), and bringing the corpus vocabulary down to a more realistic size of 7,367 tokens. Still, a significant amount of noise passed this filter, and as visible in results, several topics are just combinations of non-words. However, using a stricter filter meant losing a considerable amount of potentially meaningful words, and I therefore preferred going for a higher number of topics and filtering out the non-meaningful ones, rather than taking the risk of impacting the quality of all topics.

A custom stopwords list was iteratively refined through multiple runs of the topic modeling algorithms. From looking at the first outputs, many topics were found to be specific to one company or country, due to the presence of company, country and currency names, as well as non-english words. Therefore, all countries, currencies and company names were removed, in addition to structure-related words (“introduction”, “contents”, “appendix”, “table”, “figure”, “www”, “page”, etc.).

Analysis



latent Dirichlet allocation

Biterm Topic Modeling

Non-negative Matrix Factorization

CorEx

Figure 1. Selecting optimal number of topics for each topic modeling method

Starting with LDA, for every number of topics ranging from 2 to 50, two metrics were computed and plotted (see Figure 1); the perplexity score of the model, and the average of every topic's coherence score. Both metrics were normalized. The aim is to minimize the model's perplexity while maximizing the coherence within topics; since the shape of the lines do not really allow for a straightforward elbow method, a few different numbers of topics seem to offer a good compromise. After testing the 18, 27, 37 and 40 topics solutions, I decided on keeping 27 as the final model.

The same two metrics were computed, normalized and plotted for BTM, with several numbers of topics appearing as good solutions. Values that show a spike in coherence were tested, and after several iterations I landed on 28 topics, offering a good compromise between interpretability and level of detail in the topics' top terms.

For NMF, coherence was the only metric used, with several spikes indicating potential solutions. After examining the top words per topic for 10, 20 and 29 topics, the 29 topics solution was retained as it offers a more complete picture and some more precise topics.

Total correlations in CorEx topics overall keep going up as more topics are added to the model, with most gains occurring between 2 to 15 topics. The model was tested with 12-16 topics, among which the 15 topics solution appeared best. However, similarly to the other three methods, a larger number of topics allows for more diversity in topics, and after further examination, the 25 topics model appears as a good compromise.

Although the metrics did not indicate a clear "best choice" for the number of topics on any of the methods, I did land on a similar number (25-29) for all four of them, after multiple iterations in different ranges. In each case, there may have been satisfactory solutions with a lower number of topics (10-15), but they did seem reductive given the wide range of thematics addressed by these reports, especially given that several spots are systematically given to noise topics.

Results

a. Exploratory Data Analysis

The corpus includes 1,096 reports from 209 companies within the technology industry. Due to the important amount of noise in the text, I will spare reporting descriptors such as vocabulary size, as they are not meaningful. As illustrated by Figure 2, the corpus reflects the exponential growth in CSR reporting in the industry over the past 5 years, with 61% of documents being from 2016-2021. Figure 3 shows the top 15 companies with most reports in the corpus, basically listing the long-runners of the industry, but also the CSR early birds.

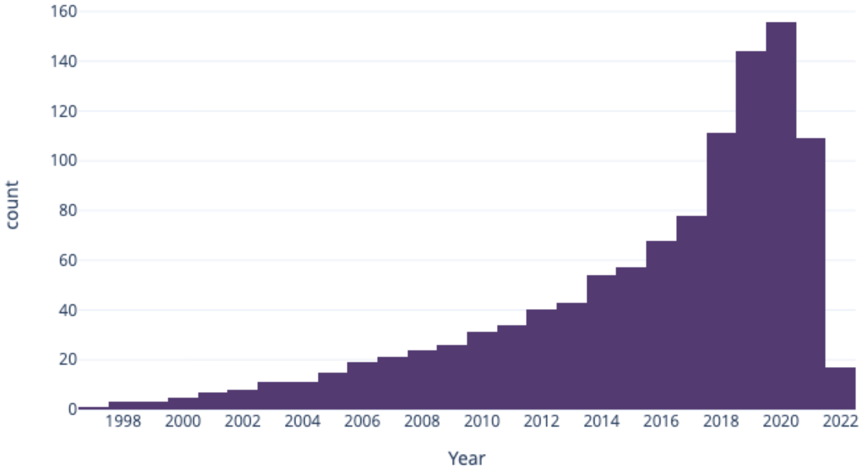


Fig. 2: Number of documents per year

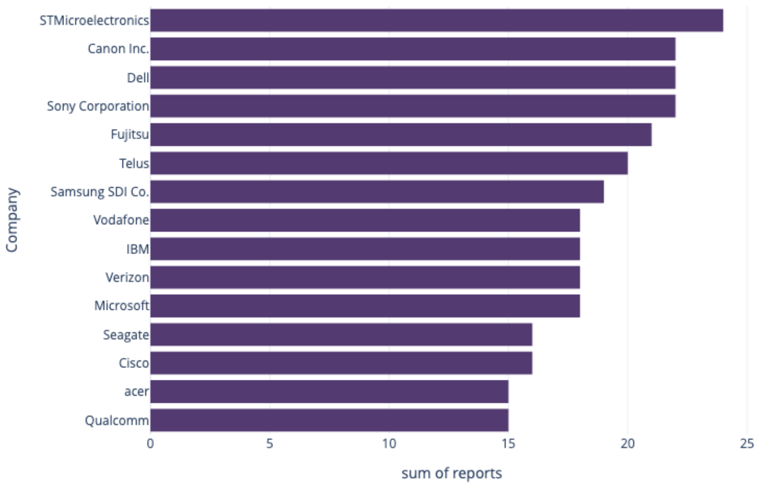


Fig. 3: Top 15 companies with most reports in the corpus

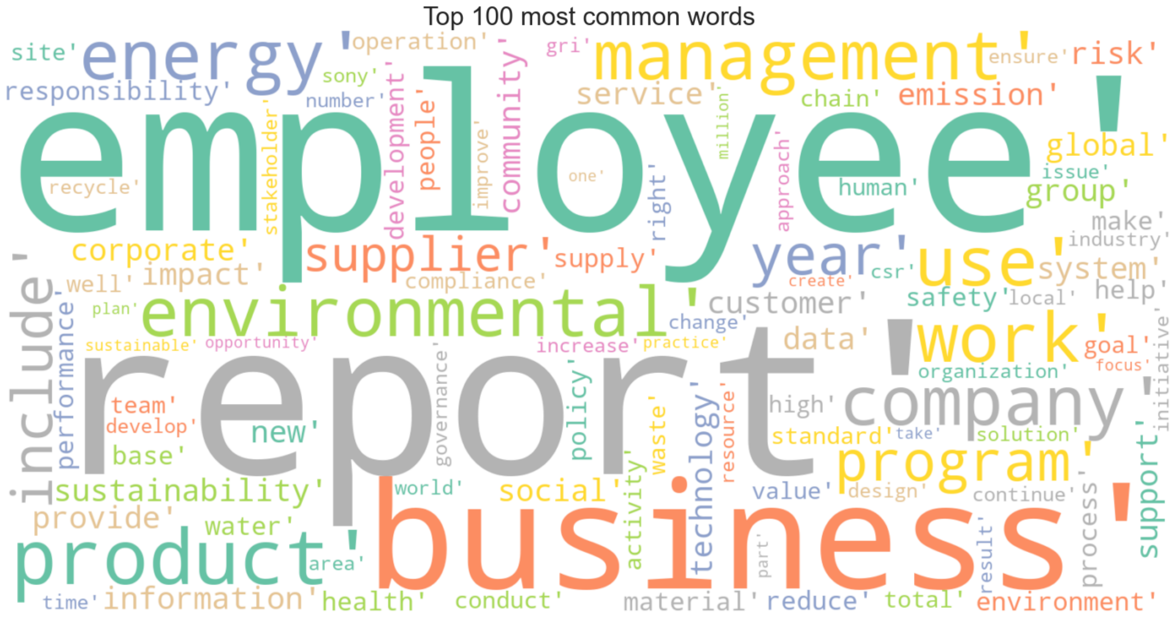


Fig. 4: Word cloud for the top 100 most frequent words in the corpus

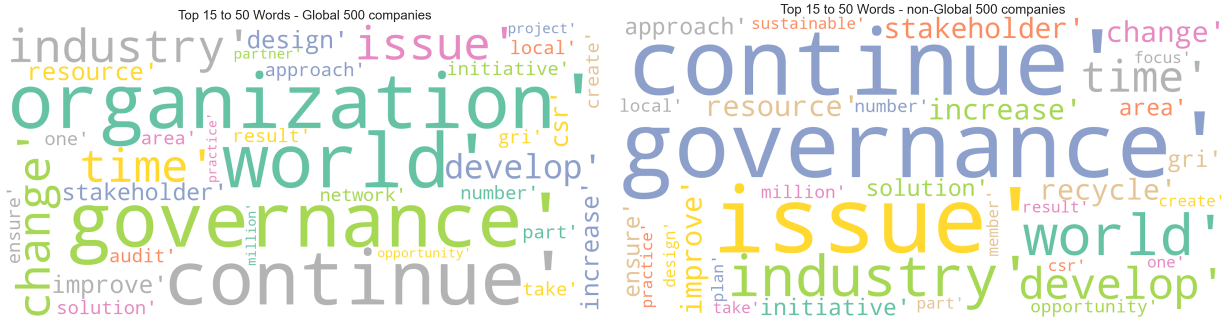


Fig. 5: World clouds for the top 15 to 50 words for Global 500 companies (left) vs others (right)

Figure 4 visualizes the relative frequencies of the hundred most common words in the corpus, offering a taste of this type of corporate discourse. A number of words relate to the corporate world (employee, business, product, company, customer, …) along with industry-specific terms (technology, system, data) , while others represent the main two topics addressed by these reports; environmental and social. The list also includes a number of action verbs, such as “include”, “make”, “improve”, “reduce” and “provide”, all conveying the idea of “doing better”.

Two word clouds were also generated, separating Global 500 companies from others. The top 14 most frequent words were removed on both sides, as the ranking was the exact same. The word frequencies for the top 15 onwards do have variations between the two corpuses, but the list of words in itself remains very similar, which is expected. Performing topic modeling on each subset of documents, however, might reveal some differences in how Global 500 companies perform CSR disclosure. The same exploratory process was applied on FAANG reports versus others, and similarly revealed no key differences in their most common words.

b. *Final models*

b.1 LDA

Topic	Label	Top 10 terms
1	Renewable energies	energy emission data use carbon renewable water environmental report power
2	Internal initiatives	group programme employee digital business work year support emission environmental
3	ICT & Services	mobile service management network information development customer system high employee
4	Performance review?	site product performance year new company program end management process
5	General	product year supplier team report sustainability global use continue work
6	Supply chain	product supplier use report data impact chain supply emission include
7	General	report sustainability employee management year business total emission aspect client
8	General	employee company management business group work sustainability report customer safety
9	General - supply chain	management corporate product social responsibility report supplier service risk supply
10	Innovation in airports/airplanes?	airport sea passenger linate total service source management area value
11	Managing waste?	environmental product csr use fiscal group management recycle site substance

12	Customer service?	team member customer service report community program cent wireless include
13	General	company ing year tion work ment energy product new business
14	IT hardware standards?	sdi management display company safety battery system business employee sustainability
15	Emissions related to supply chains	csr product use report emission supplier chain ghg employee people
16	Energy & waste	water solar energy semiconductor first product manufacturing use material waste
17	General CSR	employee company corporate environmental program responsibility energy business provide system
18	General - supply chain	employee report supplier management sustainability gri risk chain business supply
19	General	corporate business report technology program supplier employee global percent use
20	General	data employee report esg include business governance global impact risk
21	Corporate structure	service board director company year adi share market committee executive
22	Noise	–
23	Waste management	environmental product use recycle system activity management material environment inc
24	Community support	community people program help work technology support employee business make
25	Noise	–
26	General	group activity business management fiscal environmental system society csr company
27	Customer service?	report mobile use group customer network service business company work

Table 1: LDA topics

There is an important overlap between topics (as illustrated by the intertopic distance map in Figure 6), with many being very difficult to interpret or label due to how generic they are. Keywords that appear in the word cloud of most common terms (Figure 4) strongly dominate topics, which I hypothesize can partially explain that level of overlap between topics. However, even when running the model without words that appear in more than 95% of documents and without words that appear in less than 30 documents (as the nth attempt at filtering out noise), topics are still quite vague and with important overlap. Among the better defined topics, several concern supply chains, and several concern waste management and energy use. There is overall a lack of diversity and distinction among the topics.

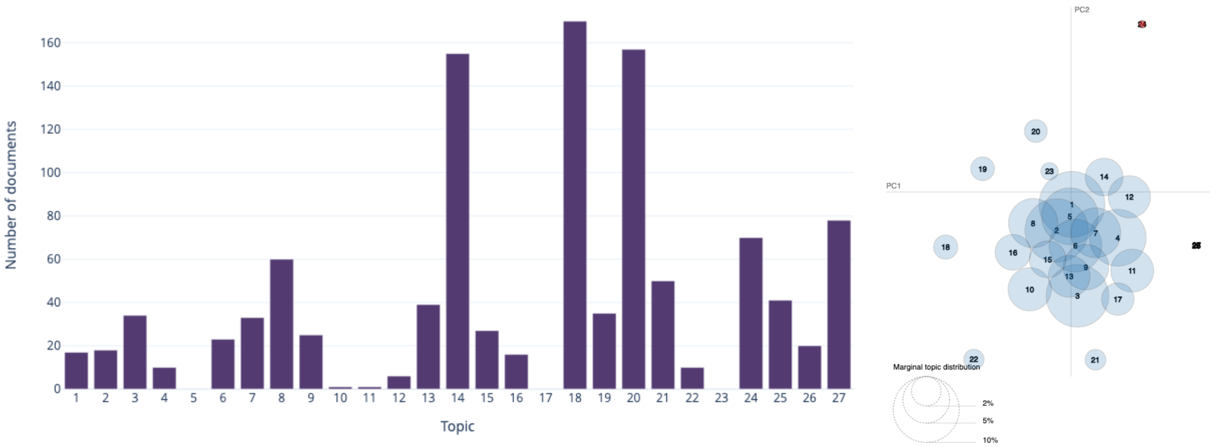


Fig. 6: Number of documents per topic (left) and intertopic distance map (right)
(note: topic labels range from 1-27 on the left, but 0-26 on the right…)

b.2 NMF

Topic	Label	Top 10 terms
1	Responsibility in supply chains	supplier chain supply risk audit material safety mineral rba assessment
2	Environmental impact of printing/ink	printer inc substance recycle cartridge production inkjet material site chemical
3	Education and EDI (equity, diversity & inclusion) (women in STEM programs?)	diversity student woman learn organization digital privacy education inclusion school
4	UK spelling	programme plc organisation recognise centre digital emission staff target key
5	South Korea IT development? (i.e. Samsung topic)	sdi krw battery pdp overseas rechargeable material issue cheonan sustainability
6	Quantifying energy consumption	water ton metric emission scope gallon gas fab facility consumption
7	Renewable energy, carbon footprint	renewable emission carbon center scope electricity power ghg water project
8	General CSR	csr ict human society governance supplier promote issue stakeholder network
9	Telecom service?	cent member wireless million target tax site footnote wireline labour
10	Management?	solution market branch sale project sustainable installers offer growth logistics
11	General?	sustainability risk sustainable emission compliance stakeholder optical network future ict
12	Bang & Olufsen topic	bang csr sound corruption labour consumption ing accident anti supplier

13	Clean energy	solar first module power recycle water grid manufacturing clean manufacture
14	Noise	–
15	Measuring / tracking progress	gri topic index approach emission disclosure material risk ghg scope
16	Samsung 2?	display oled safety krw overseas lcd joyful risk growth dream
17	Manufacturing & environmental impact	site semiconductor sts chemical water decalogue eicc award safety manufacturing
18	Regulatory entities?	board director committee code compliance governance ethic member executive financial
19	Telecom access	mobile network telecommunication phone communication million rural station alleviation poverty
20	Education	citizenship percent education human volunteer student software learn nonprofit summary
21	Waste management & fiscality?	fiscal recycle green material substance ton corporation plant chemical reduction
22	Electronic components & regulations	nordic semiconductor wireless rba shall code mineral eicc cfsp smelter
23	Aerial transportation	airport sea passenger milan aviation cargo airline terminal source flight
24	?	percent associate supplier million robert location emission metric project member
25	Stakeholders?	client associate member fiscal skill office organization financial across consumption
26	Telecom	bce network wireless mental member centre fibre broadband inc million
27	Regulation compliance?	boe uif gps compliance corporation safety site strategy efforts
28	Main topics in recent years?	esg governance covid risk sasb pandemic security privacy diversity inclusion
29	Micron	micron sustainability member water memory idaho risk supply semiconductor supplier

Table 2. NMF results: top 10 terms per topic

My first observation is that NMF appears to be efficient in handling noise, with only one topic being a combination of non-words, in contrast with the other methods where there were always a couple. It also produces more distinct topics compared to LDA. Still, the labeling task was not self-evident.

Topic 5 is company-specific (Samsung), but given the number of documents the company has in the corpus (spanning over 20 years), it still has some value. Topic 12 and 29 are also company specific, but only because the names of the companies were not successfully removed from documents; the topics themselves do not hold much meaning.

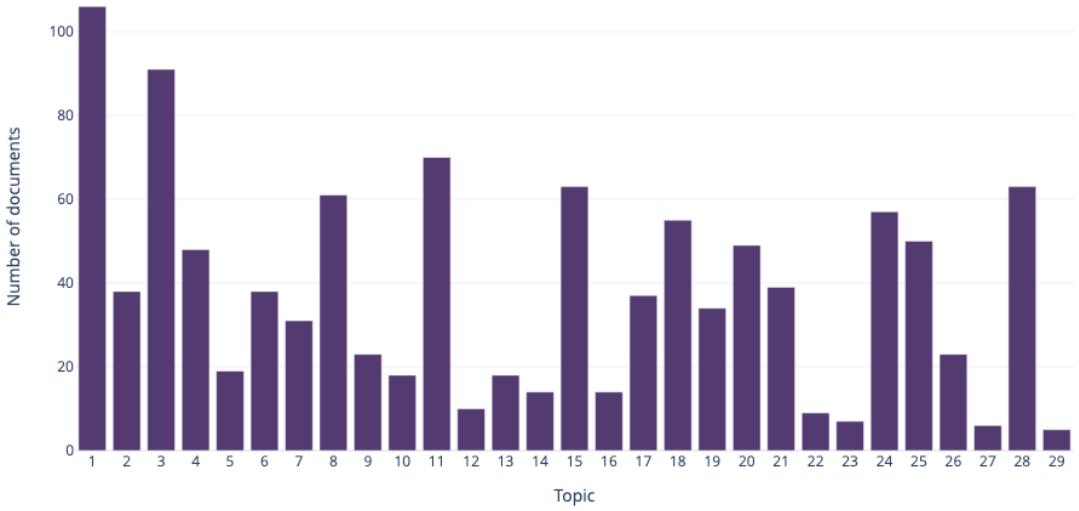


Fig 7. Number of documents per topic (main topic per document)

Figure 7 visualizes how many documents in the corpus have each topic as their main one, allowing to distinguish between more “important” and widespread topics versus more “niche” ones that may be more specific to a company, sub-sector, or geographical area. By that metric, the five most important topics are topic 1 [Supply chain], 3 [Education], 11 [General CSR], 28 [Hot topics] and 15 [Tracking progress]. While the topics are fairly general (one of them literally being the CSR topic, and one being a combination of recent thematics), this provides interesting information about which topics companies tend to prioritize in CSR reports.

b.3 BTM

I did not expect BTM to perform well on this corpus, given it was developed specifically for short text—the opposite of the documents at hand here. However, it did yield interesting results, and revealed topics that do not appear in any of the other models.

Topic	Label	Top 10 terms
1	Energy	electricity, intensity, tco, gallon, mwh, combustion, nox, normalize, gwh, mtco
2	Plastic components	pvc, flame, retardant, mercury, resin, cadmium, phthalate, chloride, chromium, polystyrene
3	Environment protection	conservation, cycle, eco, environmentally, biodiversity, lifecycle, conscious, burden, lca, stev, recyclability
4	Regulation compliance	compliance, regulation, iso, comply, ohsas, authorization
5	Employee satisfaction	survey, feedback, score, satisfaction, quarterly, scorecard, suggestion, analyst, matrix

6	Privacy & confidentiality	privacy, breach, enforcement, attack, criminal, consent, misuse, substantiate, spam, lawful
7	Company values	philosophy, mutual, kyosei, jeong, harmony, mutually, qfinity, realization, disseminate, harmonious
8	Leadership	board, committee, director, executive, officer, chief, oversight, supervisory, oversight, chair, nominate
9	Geographical-related terms (cities and country-specific business suffixes)	ltd, trademark, fax, gmbh, tel, oita, tianjin, ibaraki, kanagawa, bhd
10	Accounting?	tax, income, cash, expense, dividend, earnings, debt, gaap, krw, liability
11	Employee benefits / perks	fitness, instructor, tuition, reimbursement, webinars, videos, referral, dental, eap, nursing, counseling, daycare
12	HVAC (heating, ventilation & air conditioning) and transportation…?	cool, truck, chiller, hvac, rack, lighting, fixture, cold, ventilation, cooling, roof
13	“Responsible mineral program”	mineral, conflict, eicc, diligence, smelter, vap, tantalum, rmi, tungsten, vap, mine
14	Community & creativity	feel, passion, creativity, talented, mindset, teamwork, something, teammate, courage, authentic
15	Occupational safety and health	rate, occupational, accident, injury, lose, disease, illness, severity, fatality, recordable, overtime
16	Regulatory bodies & consultancy	boundary, explanation, disclosures, tcfd, kpmg, omission, restatement, materially, dnv, completeness
17	Noise	–
18	Specific project of technological innovation in an Italian airport	airport, sea, passenger, linate, milan, cargo, flight, airline, aviation, terminal
19	The climate crisis & the UN Sustainable Development Goals	climate, sdgs, sdg, tackle, neutral, ambitious, ambition, resilient, aggressive, urgent
20	Hardware & innovation	iot, oled, processor, memory, lte, wearable, supercomputer, ethernet, transistor, gpu
21	Equity & diversity in hiring	gender, female, hire, male, men, race, ethnicity, native, asian, sts
22	Human rights	anti, corruption, discrimination, bribery, freedom, bargaining, harassment, slavery, compulsory, uphold
23	Health & wellness	healthy, healthcare, wellness, pandemic, wellbeing, productivity, resilient, flexibility, needs, productive
24	Noise	–

25	Noise	–
26	Awards & public recognition	award, consecutive, magazine, winner, ftse, chamber, fortune, jones, forbes, prestigious
27	Charity (specifically related to children & poverty) (and german words)	donate, die, elementary, und, homeless, underprivileged, toy, clothes, fundraiser, christmas
28	Water contamination	wastewater, discharge, fire, soil, remediation, groundwater, contamination, drill, contaminate, sanitary

Table 3. BTM topics

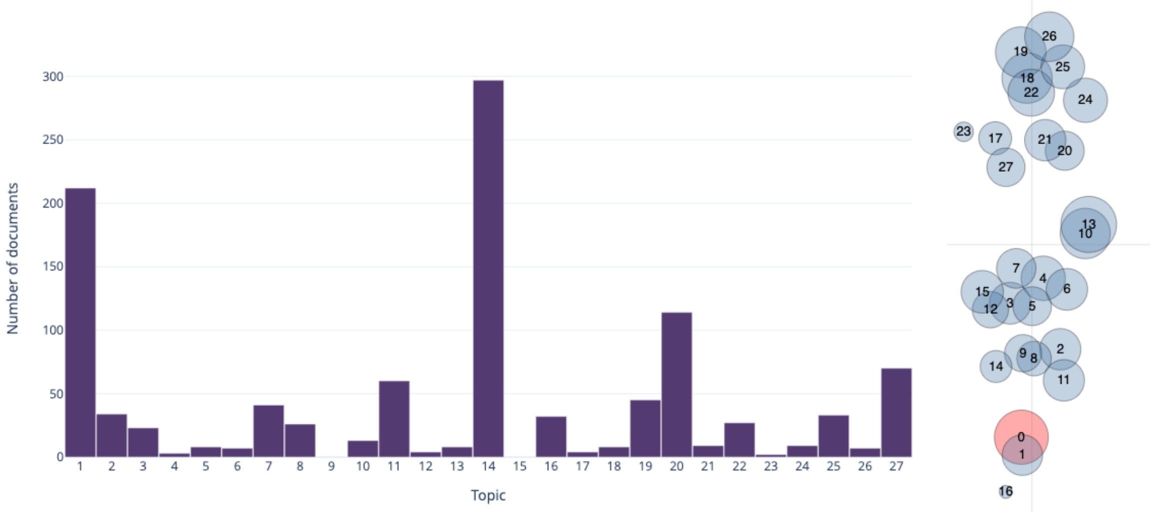


Fig 8. Number of documents with t as their dominant topic (left) and inter-topic distance plot (TSNE method) *(note: topic labels range from 0-27 in the bubble plot, but from 1-28 in the bar graph)*

Figure 8 hints at the relative weight of each topic in the corpus, with close to 300 reports having topic 14 [Community & creativity], and over 200 reports having topic 1 [Energy] as their dominant one. Obviously, these documents cover a wide range of topics and cannot be reduced to one, but it does indicate which ones tend to dominate the text.

5 out of the 28 topics relate to different aspects of engagement for the environment, and 5 relate to different facets of employee well-being. 4 topics concern the corporation as an entity, with topics such as company values, awards, and leadership.

It is relevant to report that running the model with different parameters revealed other topics of interest covered by an important number of documents, such as AI innovation, health, long-term goals, recycling and waste management, access to education and humanitarian aid, or promoting transparency and fighting corruption, which do not appear in the results presented here, or at least not with the same salience. This might indicate that there is a BTM solution with a larger number of topics. Overall, BTM is surprisingly efficient on this corpus, producing mostly coherent and interpretable topics. It is also interesting to note the highly different levels of abstraction from one topic to another; topics 10 and 11 seem to concern a specific project, while others like 12 and 18 are fairly general (making them more difficult to interpret and label).

b.4 CorEx

Topic	TC	Label	Top 10 terms
1	36	?	index, function, expectation, feedback, response, contract, infrastructure, position, face, concern
2	32	Noise	–
3	31	Collective bargaining, and responsible sourcing?	assurance, gri, complaint, identification, screen, bargaining, precautionary, collective, disclosure, criterion
4	28	?	joint, trend, basic, regional, second, would, conference, depend, study, apply
5	23	Collaborative projects?	jointly, mainly, label, scheme, proposal, preparation, division, moreover, ratio, opinion
6	17	?	image, billion, medium, particular, special, mail, equip, parent, practical, cut
7	16.5	?	consultation, divide, domestic, attention, frequency, maintenance, express, serious, not, reliability
8	16	Inspections in supply chains	ministry, inspection, planning, logistics, liquid, transfer, warm, groundwater, damage, burden
9	14.2	Compounds & safety (specifically monitors?)	flame, retardant, lcd, pollutant, chloride, vocs, display, dust, mold, thin
10	14	Collaborative projects	collaborate, explore, deploy, big, otherwise, collaboration, advocate, inform, real, embed
11	13	Transportation	volume, weight, phase, indicate, transport, package, original, occur, packaging, reasonable
12	12.4	Events / meetings?	attend, cultural, young, teach, hear, middle, language, talk, session, participant
13	12.3	Managing chemical hazards/waste (environment)	chemical, discharge, wastewater, substance, raw, compound, factory, assembly, manufacturing, rohs
14	12	Academia & research	academic, shift, realize, intellectual, host, university, organize, announce, series, submit
15	11	Inclusion & diversity	inclusive, representation, mitigate, bias, inclusion, align, cybersecurity, transition, disruption, esg
16	10.9	Ozone depletion	sox, ozone, nox, deplete, effluent, specify, grievance, fine, spill, absenteeism
17	10	Cultural events & traveling	overseas, exhibition, sport, theme, lecture, load, seminar, abroad, pass, federation

18	9.4	Regulations?	propose, quantity, govern, format, acquire, vary, legislation, variation, segment, separate
19	9	Education access / support programs (“bridging the gap”)	teacher, urban, video, intervention, spending, rural, accessibility, bridge, notice, enforcement
20	8.7	Minorities	hispanic, black, veteran, islander, underrepresented, transgender, lesbian, african, lgbtq, latino
21	8	?	nearly, summit, earn, literacy, decade, learner, ambitious, desktop, white, story
22	7	UK spelling	centre, labour, tonne, recognise, behaviour, programme, organisation, programmes, analyse, organise
23	6	Responsible Minerals Initiative (or similar program/organization)	tantalum, tungsten, tin, smelter, mineral, democratic, drc, gold, republic, diligence
24	5.7	Silicon valley 1	nonprofit, percent, california, metric, stewardship, gallon, volunteerism, proud, valley, landfill
25	5.2	Silicon valley 2	labor, behavior, preparedness, forth, periodically, periodic, multinational, center, silicon, recognize

Table 4. Unsupervised CorEx topics

The unsupervised CorEx model yields some very well-defined topics (e.g. 12 [Academia & research] and 20 [Minorities]), but a majority of topics that are less evident. For instance, topic 3 seems to mix several topics, including both collective bargaining and responsible sourcing principles and precautions together.

I therefore performed some semi-supervised experiments, using lists of anchor words I was expecting to retrieve in topics and based on what other models revealed. Table 5 presents results. The approach was to provide keywords relating to somewhat broad thematics that are commonly addressed in CSR reports (with some being more specific to the industry), and see which topics form around that. Different anchor strengths were tested, and a value of 3 was selected as it was sufficient in providing relatively coherent topics. Overall, anchoring allows to observe in which contexts specific keywords tend to be used; for instance, inclusion tends to be addressed through the “empowerment” and “collaboration” perspective, thereby being presented through a positive and encouraging tone.

Anchor words	Top 10 words
united, nation	nation, united, study, billion, would, medium, half, late, clear, effect
sustainable, development, goal	goal, sustainable, capability, feedback, external, expectation, position, development, response, assess
political, conflict	political, conflict, bargaining, precautionary, corruption, collective, compulsory, mitigate, anti, freedom

ngo, nonprofit	ngo, institute, inquiry, forum, cumulative, analyze, initial, realize, organize, economy
supply, chain	chain, consultation, criterion, discrimination, disclosure, screen, supply, disclose, accordance, subject
human, rights	serious, planning, satisfaction, resolve, check, ngos, implementation, human, database, provision
equity, diversity, inclusion	diversity, inclusion, equity, hire, talent, competitive, career, skill, woman, align
gender, discrimination	discrimination, gender, occupational, consultation, corrective, violation, promotion, compensation, minimum, wage
gender, women	gender, boundary, bargaining, precautionary, restatement, gri, disclosures, complaint, substantiate, withdrawal
gender, sexual, violence	sexual, gender, orientation, religion, harassment, race, marital, violence, sex, ancestry
minority, ethnicity	minority, ethnicity, empower, lesbian, gay, learning, collaborate, transgender, leverage, peer
accessibility, disability	accessibility, disability, text, parent, young, computer, teacher, web, telephone, card
covid	covid, pandemic, sasb, esg, distancing, tcfd, coronavirus, racial, lgbtq, cybersecurity
climate, change	climate, assurance, calculation, specify, calculate, analysis, protocol, agreement, consume, previous
waste, recycle	tin, rohs, substance, manufacturing, electronics, smelter, chemical, directive, tungsten, tantalum
waste, management	function, determine, trend, planning, depend, exercise, handle, representative, method, trade
biodiversity, environment	biodiversity, joint, label, ratio, preparation, extent, nature, special, price, selection
data, privacy	privacy, discussion, resolve, participant, discuss, cultural, hear, care, real, listen
artificial, intelligence	intelligence, artificial, collaborate, cloud, deploy, transition, embed, virtual, accelerate, outcome
innovation, technology	innovation, technology, expectation, competitive, governance, skill, foster, talent, security, leadership

Table 5. Semi-supervised CorEx topics

Discussion

Environment and supply chains

Across the four modeling methods, several common topics appear. Energy consumption and discussions around renewable energy appear in all four models. Waste management, and specifically related to dangerous and polluting chemicals, also makes its appearance as a topic in each model, although with a seemingly different angle, sometimes focusing more on the safety aspects and sometimes on sustainability issues. On a wider scale, responsibility in supply chains is a recurring topic, joined with both environmental concerns and occupational safety. In fact, several topics (topic 1 in NMF, topic 13 in BTM, topic 23 in CorEx) point distinctively at the Responsible Minerals Initiative, launched by the Responsible Business Alliance in 2008 and providing resources for the responsible sourcing of minerals in supply chains[3], which shows it has become an important reference point in the technology industry.

Other environmental-related topics vary from one model to the other, with ozone depletion appearing through CorEx [topic 16], responding to the climate crisis [topics 3 & 19], water and soil contamination [topic 28] and plastic pollution [topic 2] appearing through BTM, reducing carbon emissions through NMF [topic 7], and general waste management (unrelated to supply chains and manufacturing chemicals) through LDA [topic 11].

Social issues

Regarding social issues, inclusion and diversity makes an expected appearance in NMF, BTM and CorEx. In NMF, the topic [3] incorporates the terms “education”, “learn”, “woman”, “digital” and “privacy”, which leads me to hypothesize as referring to women in STEM or similar programs, although it might just be a poorly defined topic. The BTM topic [21] is clearly about diversity in hiring, specifically focusing on gender and ethnicity. CorEx has two related topics: topic 20 combines minority labels, while topic 15 is more general with words such as “representation” and “bias”, although combined with a surprising set of lower-probability words: “cybersecurity”, “transition” and “disruption”, which is not an obvious association to me.

Accessibility only appears in a CorEx topic [19], despite my impression that this theme took the forefront in technology company discourses in recent years. Similarly, health-related topics were scarce, only appearing in BTM through topic 15 [Occupational safety & health] and topic 23 [Employee health & wellness]. Even more surprising was the absence of covid-specific topics, which I expected to see due to the proportion of documents being from 2020 and 2021 (24%) (especially considering how the pandemic has taken over corporate discourses overall). To a lesser extent, human rights-related topics were also within my expectations, but only one BTM topic [22] distinctively belonged to that theme.

Industry-specific topics

Topics related to telecommunications were seen in all four methods, although they only appear in the final models for LDA and NMF. This might be due to the number of telecommunication companies present in the corpus (e.g. one of these topics was fitted for Bell, which has 13 reports in the corpus). Since these companies tend to have been in activity for a relatively long time and are public-facing, they may have started producing CSR reports earlier than other sub-sectors, which would explain why this topic appears over other sectors.

Few other industry-specific topics make an appearance. Privacy protection surprisingly only appears in BTM [6], along with a hardware innovation topic [20] and a different hardware-related topic in LDA [14].

Company

Several topics concern the company culture, including corporate values and philosophy (topics 7 and 14 in BTM), corporate structure (topic 21 in LDA), community support (topic 24 in LDA), leadership (topic 8 in BTM) and employee benefits and perks (topic 11 in BTM, and potentially topic 17 in CorEx).

[3] Responsible Minerals Initiative. https://www.responsiblemineralsinitiative.org/

Finally, a few topics describe broad CSR practices and strategies (several LDA and NMF topics) and efforts to track and quantify progress (topics 6 and 15 in NMF). Regulations, whether self-imposed policies or initiatives or legislated ones, take an important place in each model's topics, with a number of acronyms for regulatory bodies and organizations. Only one BTM topic [26] describes awards and recognition, which I did expect to see more consistently since many companies include such sections in their reports.

Conclusion

For each topic modeling method, there appeared to be a solution in the 10-15 range and another in the 25-30 range; after examining and comparing results from all four models, however, I suspect there may be a larger solution, at least for BTM which was the only one producing relatively easily interpretable and well-defined topics. Indeed, the labeling process was quite tenuous for the most part; this can definitely be partially accounted for by my lack of domain knowledge, but I would also say is due to the nature of the analyzed documents. Indeed, it could be the case that CSR encompasses a much wider array of complex topics within a much less clear-cut space than I assumed. This could also mean that CSR reports have less commonalities than expected, at least when searching beyond the obvious general lines. Again, the lack of cleanliness in the data surely blurs the lines for the algorithms, making it more difficult to extract coherent topics. It might be worth removing the documents in the worst condition from the corpus altogether in a future run, assuming the noise is centralized in some specific ones. Finally, the specificity or level of abstraction of the topics, within a same model, widely varies, some being very general (where I could not label them with anything else than "CSR") and others strongly overfitted (being specific to one company).

One final takeaway is that the industry appears to be moving towards standardization in self-assessments and disclosure practices, through the systematic reference to CSR organizations and initiatives such as the RBA (Responsible Business Alliance) and GRI (Global Reporting Initiative). This reflects the impressive increase in CSR reporting in recent years, whereby corporate responsibility has gone from marginal to mainstream.

PART 2: Thematic Analysis

Method

In order to get a better understanding of the meaning behind topics, further qualitative analysis was performed on one of the models. Since BTM seems to produce relatively well-defined topics compared to the other three methods, it was re-trained with slightly different parameters in order to improve the initial results, and retained for this in-depth analysis. The number of topics was increased from 28 to 32, and the lambda value used to sort the top terms per topic was set to 1, in order to prevent company or project-specific terms from appearing in the list and facilitate the analysis and interpretation. This final model yields significantly more interpretable and meaningful topics compared to the previous four.

I proceeded to read each topic's top 5 documents, in order to put them into context before labeling them. The length of reports is highly variable (from 3-4 pages to over 400 pages), and in the case of more consequent documents, keywords searches were used to directly locate potential topic-relevant sections (where the top terms co-occur). The meta-data was useful in providing initial contextual information. Document titles sometimes directly indicate a subtype of CSR and can help explain a topic. For instance, some topics' top documents are titled "sustainability metrics". The companies who originated the documents can also hint at a topic's context (e.g. when the 5 top documents are from Japanese electronics companies [Topic 4]), and so are the release years (e.g. when all 5 are from the early 2000s [Topic 2] versus the late 2010s [Topic 13]). All topics were tentatively labeled and included in the overall thematic analysis (apart from the 3 noise topics), and I report the uncovered themes, as well as closer analysis on a selection of topics. Where example sentences from the top documents are reported to illustrate a topic's context, words that appear in the topic's top 20 terms are signaled in bold.

Results and Discussion

After examining the context and meaning of each individual topic, and selecting labels for them, the analysis moved to a higher level, by examining the connections between topics. The model generates topics representing a wide range of subjects that are common to CSR reports in the tech industry, namely: environmental responsibility, social responsibility, supply chain monitoring, company-level initiatives and discourses, and industry-specific topics. These five identified themes each contain 2-9 topics. Topic to theme mapping is illustrated in Figure 9 (page 25).

0. Special cases

While some topics are more specific and easily definable, others can be fairly general or appear quite vague. This is to be expected from topic modeling, and I tried not to extrapolate the meaning of labels when unclear. Hence, a few topics were voluntarily given generic labels, such as Topic 27 [Generic CSR reporting], and Topic 8 [General / Introductory]. Topic 8 was found to be more characteristic of introductions in its top documents, but was also mostly represented by a single company (Fujitsu), and could therefore be a company-specific language pattern. The model contains several such company-specific topics (including Topic 18, where all 5 top documents are from the same company), due to the presence of multiple documents from each company, which use the same template, introductions and overall language from one year to the next.

Topic 27 is quite general in that it relates to the Global Reporting Initiative (GRI), which is a set of principles meant as a framework for companies to produce CSR reports, and are often directly referred to in these documents. Since the GRI's mission is to standardize CR reporting, its presence in the model makes sense. These two topics were loosely placed in relation with the company-level theme, in that they relate to CSR presentation and strategies, but pertain to a different "level" compared to others within that theme, and are hence presented separately.

1. Environmental responsibility

Two sub-themes can be identified within the environmental theme; energy efficiency and waste management, either related to product design and development, infrastructures, or supply chains. Topic 1 is separate, as it concerns the reporting of metrics for environmental impact. The presence of terms such as "total", "year", "ton", "data" and "million" in the topic's top terms hint at the use of metrics, which is confirmed by the closer analysis. The top document for that topic is, in fact, mostly composed of tables reporting environmental impact metrics (the document being titled "Sustainability Metrics"). The 4th top document, which is from a different company, also includes several such tables, and the topic is also retrievable in unstructured text: *"By 2018, NETGEAR headquarters office has realized 49%* ***water consumption reduction*** *compared to 2013", "NETGEAR has converted Natural* ***Gas****, Diesel,* ***Electricity consumption data*** *to CO2e (****Carbon*** *dioxide equivalent) for the purpose of quantifying our* ***greenhouse gas emissions*** *for a* ***year*** *over* ***year*** *comparison"*. The topic therefore points to quantified progress-tracking on environmental impact.

Topics 2 and 4 have significant overlap, as visible in the intertopic distance plot (see Appendix A) and by the fact that they share 2 of their 5 top documents. They represent an earlier stage of CSR discourses, as confirmed by their top documents being from the early 2000s. In both cases, these documents are from electronics companies, therefore addressing environmental concerns that differ from service companies. Electronic waste management is a prominent issue for these corporations, as visible through Topic 2 ("*Canon will* ***recycle****,* ***reduce*** *and* ***reuse*** *more than 90% of its* ***used products***"), although the topic encompasses other aspects of waste management as well, such as printing and packaging. Overall, companies report their strategies to reduce their product's impact, from production and shipment to end-of-life recycling. Topic 4 is broader than Topic 2, in that the former concerns sustainable product development in general, while the latter is more focused on waste management throughout a product's lifecycle.

Topics 13 and 23 address energy efficiency, but within a slightly different context. Topic 13 keywords appear in sections and sentences relating to energy efficiency in data centers [*"More than two-thirds of this advantage is attributable to the combination of a more* ***energy efficient server*** *population and much higher* ***server*** *utilization"*, *"When possible, AWS incorporates direct evaporative* ***technology*** *for cooling our* ***data centers****, significantly* ***reducing energy*** *and water consumption."* (Amazon, 2019), *"****Power Usage Effectiveness*** *(PUE) is* ***used*** *to measure how* ***efficiently*** *our* ***data centers*** *convert* ***energy*** *into compute* ***power****."* (Facebook, 2019)], as well as other types of infrastructures [*"We use an automated software platform to track and manage* ***environmental data*** *about actual electricity, fuel and water* ***usage*** *at our properties. This platform allows us to pull* ***data*** *on our sites'* ***energy*** *and water* ***usage*** *from utility* *bills by synchronizing directly with the sites' utility providers."* (ESCO, 2020)]. These sections report how the company measures energy efficiency, and what is in place to improve it. Other topic-relevant sections from a solar panel company indicate how their systems and products are participating to transitioning towards a more energy-efficient world: *"SolarEdge addresses a broad range of **energy** market segments through its PV, storage, EV charging, **batteries**, UPS, and grid **services** solutions and is leading the **energy** transition from large centralized **power** stations to an interconnected network of distributed **energy** networks based on smart solar **energy systems**."* (SolarEdge, 2018). In contrast, Topic 23 seems to reflect a different type of statements, referring to specific projects, as hinted by the geographic vocabulary appearing in the lower probability terms ("river", "garden", "angeles", "burlington", "orlando", …), while Topic 13's lower probability terms are equipment ("server", "hybrid", "processor", "ventilation", …).

2. Social responsibility

The social responsibility theme captures three sub-themes: diversity and inclusion in the workplace (internal initiatives), solidarity and social outreach (external initiatives), and human rights (both internal and external).

a) Diversity and inclusion

Topic 24 is particularly distinctive, which may be due to its pervasiveness in reports from recent years. Indeed, "diversity and inclusion" have become buzzwords in organizations across all industries, as illustrated by the topic's top reports being from 2018-2022. It encompasses both diversity/inclusion in hiring (e.g. reporting demographic proportions among new hires; "*We had the largest percentages ever of **Black**+ and **Latinx**+ **new hires** in the U.S. and **women new hires globally***"), as well as within the company's hierarchy (e.g. reporting on leadership programs; "*we identified areas of opportunity including increasing the percentage of **women** in **leadership** and technology roles, and increasing representation from different races/ethnicities across Shopify and in **leadership** roles*"). Although gender equality is at the top of the topic in terms of token probability, ethnic and other underrepresented minorities appear as well, amounting to a discourse around diversity, broadly: "*Google gathers **global** data on race, **gender**, and other identities to help give us a more complete picture of our **workforce**. This data is helping to power our **diversity**, equity, and **inclusion** (DEI) efforts **globally**, and helps to make everyone at Google more visible - so that we can create an even more **inclusive workplace***".

b) Solidarity and social outreach

Topic 30 illustrates the "education for all" discourse, and includes reporting on both STEM-specific programs (e.g. "*Jobs in **science**, **technology**, engineering, and **math** (STEM) have grown substantially in the last few decades, but the number of women **working** in those jobs has not. [...] In 2017 the group launched a five-year program that combines intensive summer courses, enrichment activities throughout the **school** year, paid summer internships, and a trip abroad focused on STEM knowledge in a global setting*") and on initiatives helping underprivileged children access education, both locally and abroad (e.g. "*Allscripts **volunteers** visited six after-**school** houses throughout the city of Be'er Sheva, and proudly donated 30 laptops to **children** and families without the means to purchase their own computer*").

Topic 12 pertains more largely to employee volunteering and company donations to charities, with a focus on supporting local communities. For instance: "*To maximize the impact of* ***employee charitable giving****, we provide matching funds to the* ***charitable*** *organizations* ***employees*** *personally* ***support****.*" (Onto, 2020), "*we care deeply about the* ***health*** *of our* ***community*** *and are proud to* ***support*** *the American Heart Association (AHA). Our vision is to build connected* ***communities*** *of* ***health****, improving the* ***health*** *of our associates as well as the* ***communities*** *we* ***serve****.*" (Allscripts, 2020).

c) Human rights

Topic 15 was labeled "Labor Rights", as its top terms appear in multiple contexts: policies against discrimination, diversity in recruitment, access to healthcare, right to parental leave, etc. Some sentences describing the company's values capture the essence of the topic, for instance: "*In keeping with the value 'One team: Good collaborative people treating each other well', the* ***health****, safety and wellbeing of* ***employees*** *and contractors are Isentia's highest priorities and the company is mindful of* ***human rights*** *[...]*" (Isentia, 2018).

3. Supply chain monitoring

Supply chain monitoring is in-between the social and environmental responsibility themes, as it encompasses issues pertaining to both. Topic 20 concerns a human rights issue within the supply chain, while Topic 7 addresses responsible sourcing from an environmental standpoint.

Topic 20, labeled "Responsible Mineral Sourcing", relates to the Responsible Minerals Initiative (RMI), created by the Responsible Business Alliance (RBA), which provides resources for companies to make informed choices and monitor the sourcing of minerals in their supply chain. The organization is directly referred to in all 5 top documents, along with terms such as "conflict-free supply chain" (e.g. "*Certain areas of the world rich in 3TG are affected by violent* ***conflict*** *and rampant human rights abuses funded by the exploitation and trade of the 3TG mined in those regions. Skyworks intends to maintain a "****conflict-free****"* ***supply chain****.*") and "conflict minerals" (e.g. "*Based on risk assessment by Nordic* ***Semiconductor****, the main focus area relating to Human Rights is* ***Conflict Minerals****.*"). In short, companies report their practices in tracing the source of minerals used for manufacturing, and ensuring their stakeholders that they are not connected to zones of conflict and of known human rights violations. Specifically, references to the Democratic Republic of Congo, where militias violate human rights to profit from the mining and trading of said resources, appear repeatedly.

In addition, "occupational health and safety" was identified as a sub-theme, connected to both human rights and supply chain monitoring, and including two topics.

Topic 3 gravitates around reports of injury rates (usually as bullet points or in tables), auditing facilities for respect of health and safety regulations ("*Meeting or exceeding all applicable environmental,* ***health*** *and* ***safety*** *requirements and verifying our performance through* ***audits***"), providing regular training for employees ("*At each* ***site****, regular* ***health*** *and* ***safety****-related trainings are planned and executed for all levels of* ***employees*** *to raise awareness and refresh knowledge*"), and other practices aimed at reducing incidents. Topic 17 centers on metrics reporting injury rates, work-related illnesses and incidents, etc. In most documents, that information is presented within tables and infographics, comparing these metrics across geographical areas and across the years.

4. Company-level initiatives and discourses

Nine topics gravitate around the company's values, corporate governance and CSR strategies, and represent a different category of discourse within CSR reports. While topics in previous themes relate directly to a type of sustainability or social issue, or their responsibility within the larger system in which they operate (supply chains), these topics center on the company. For instance, companies write about received awards that are relevant to CSR (Topic 29), or relevant financial data (Topics 11 and 6). General discourses around the company's culture (Topic 16) were also placed within this theme, being detached from social concerns (in contrast with Topic 24 [Diversity & Inclusion in the workplace] placed in the "Social responsibility" theme). Two clusters of topics center around corporate CSR strategies (Topics 18 and 22) and corporate governance (Topics 9, 10 and 28).

a) CSR Strategies

Topic 18 concerns target-setting, planning and performance evaluation for CSR activities. It may be less meaningful, in that all top 5 documents are from the same company (Bang & Olufsen), suggesting it reflects a single company rather than an industry pattern. Topic 22 is more representative of a subtype of reports, since it captures shorter documents (from multiple companies). In addition to being short, all top 5 documents are very general, in that they don't report specific activities, metrics or achievements, but rather just present a CSR "framework", specifically by referring to the UN's SDGs; "*IHS Markit uses the United Nations* ***Sustainability Development Goals (SDGs)*** *as a framework to understand our* ***corporate sustainability*** *efforts, providing us a unifying approach to guide, identify, assess and measure the* ***impact*** *of our actions*"). In some of these documents, the topic is concentrated within the "Strategy" section, hence its label: "CSR Strategy & Framework". If this was a critical analysis, this topic would have been labeled as corporate blathering, which is present throughout most topics but particularly in this one, along with Topics 8 and 27 (which are therefore loosely connected in Figure 9).

b) Business Ethics and Corporate Governance

Topic 9 appears to mostly represent semiconductor company discourses, and relates to codes of conduct, CSR principles, and business ethics in general. Topics 10 and 28 more distinctively address corporate governance, each from a different perspective. Anti-corruption and whistleblower protection measures appear through Topic 28 ("*In support of this effort, Keysight's* ***Standards*** *of* ***Business Conduct*** *(SBC) and global* ***anti-corruption policy*** *are guideposts in* ***conducting business*** *with honesty and integrity*"). Topic 10 captures the actors of corporate governance, namely, boards of directors and audit committees ("*The* ***business*** *and affairs of the* ***Company*** *are managed under the direction of the* ***Company's Board*** *of* ***Directors****, and the* ***Company*** *conducts its* ***business*** *through meetings of the* ***Board*** *and its* ***Audit Committee***") and addresses subjects such as board diversity ("*Our* ***board chairman*** *is an African American male. In addition, three of our* ***independent directors*** *are women and another is of Asian ethnicity. This makes Cedar a leader among REITs in terms of* ***Board*** *diversity*") or process of nomination/election ("*The* ***Board*** *currently consists of five* ***members elected*** *by the shareholders and two members* ***elected*** *by employees*"). In other words, this cluster of topics pertains to companies' self-regulation.

5. Industry-specific topics

Two topics apply distinctively to the technology sector. Topic 31 concerns discourses around technology and innovation, loosely related to CSR. Topic 5 captures industry-specific ethical considerations around security, privacy and confidentiality. Terms such as cybersecurity, information security, data protection, and respecting user privacy are recurring in these sections, for instance: "*All **data security** attacks (both successful and failed) and **privacy** breaches are reported to our Board **Risk** Committee*" (Tyro, 2021), "*We will maintain **systems** to **provide** the **security** of people and all corporate assets, and will be prudent in handling **information**.*" (Epson, 2016) and "*In order to **provide** reliable, high **quality service** to subscribers, we keep regular **business** records containing **information** about them that may constitute personally identifiable **information***" (Consolidated, 2021). This topic in itself encompasses a broad and important area for the industry, and it is somewhat surprising that data security (preventing data losses and malicious attacks) is captured in the same topic as companies' ethical principles and practices in respecting their users' privacy and confidentiality. While it makes sense that reports would present both aspects under the same section, the fact that they are so intertwined/blurred could reflect that many companies have not yet fully developed their discourses around these emerging topics.

Conclusion

A wide range of topics are covered in the corpus, and even a 30-topics model can only capture broad associations and patterns in corporate discourses. In most cases, the topics refer to aspects that are themselves dense and diverse, and could be broken down to a lower level. While this reflects the complexity behind corporate responsibility issues, it also conceals much granularity in how companies tend to address them. From this level of analysis, I can only conclude about the broad thematics companies choose to highlight in their reports, which I identified as being their environmental responsibility, actions in supporting (/following) social change, awareness of their role within supply chains, and addressing industry-specific stakeholder concerns, in addition to emphasizing their corporate identity, governance and other company-specific language. However, perhaps what is most telling is to consider what fails to appear in the corpus. Observing how the model's topics have progressed throughout the years could also be a thread for future exploration.

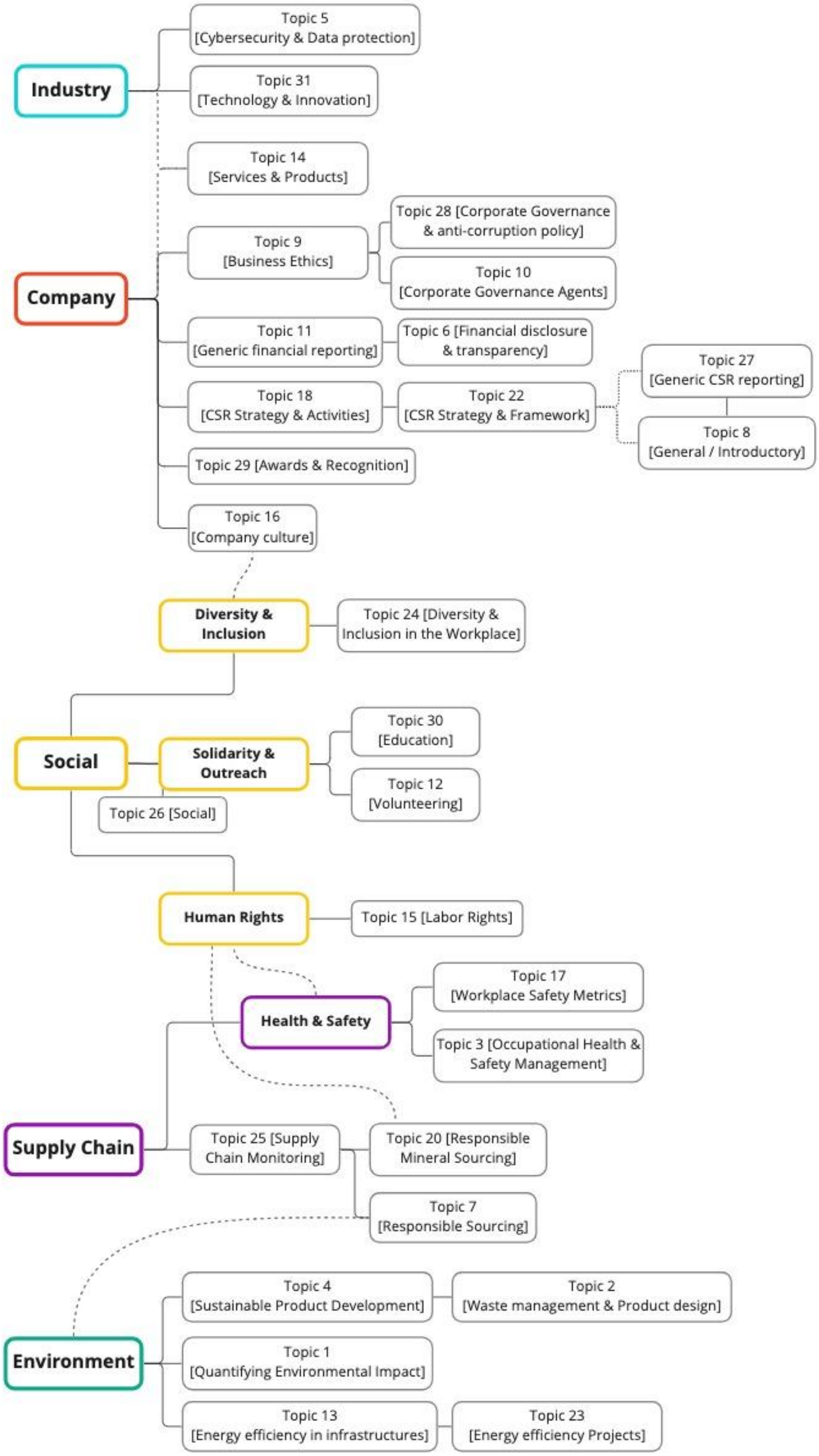

Industry
Topic 5 [Cybersecurity & Data protection]
Topic 31 [Technology & Innovation]
Topic 14 [Services & Products]
Company
Topic 9 [Business Ethics]
Topic 28 [Corporate Governance & anti-corruption policy]
Topic 10 [Corporate Governance Agents]
Topic 11 [Generic financial reporting]
Topic 6 [Financial disclosure & transparency]
Topic 18 [CSR Strategy & Activities]
Topic 22 [CSR Strategy & Framework]
Topic 27 [Generic CSR reporting]
Topic 8 [General / Introductory]
Topic 29 [Awards & Recognition]
Topic 16 [Company culture]
Diversity & Inclusion
Topic 24 [Diversity & Inclusion in the Workplace]
Social
Solidarity & Outreach
Topic 30 [Education]
Topic 12 [Volunteering]
Topic 26 [Social]
Human Rights
Topic 15 [Labor Rights]
Health & Safety
Topic 17 [Workplace Safety Metrics]
Topic 3 [Occupational Health & Safety Management]
Supply Chain
Topic 25 [Supply Chain Monitoring]
Topic 20 [Responsible Mineral Sourcing]
Topic 7 [Responsible Sourcing]
Environment
Topic 4 [Sustainable Product Development]
Topic 2 [Waste management & Product design]
Topic 1 [Quantifying Environmental Impact]
Topic 13 [Energy efficiency in infrastructures]
Topic 23 [Energy efficiency Projects]

Figure 9: Thematic analysis diagram

Topic	Label	Top 20 terms (sorted by descending conditional probability)	Top 5 documents (in descending order)
1	**Quantifying Environmental Impact**	emission, energy, water, use, scope, waste, total, reduce, consumption, ghg, gas, reduction, year, electricity, ton, carbon, renewable, data, greenhouse, million	Sunpower 2017, Sunpower 2019, Sunpower 2014, Netgear 2019, Sunpower 2015
2	**Waste management & Product design**	product, recycle, use, material, waste, substance, plastic, chemical, paper, packaging, reuse, design, equipment, hazardous, process, part, include, reduce, electronic, printer	Canon 2000, Canon 1999, Logitech 2010, Canon 2001, Dell 2002
3	**Occupational Health & Safety Management**	safety, health, risk, management, employee, work, occupational, incident, site, system, include, supplier, program, report, compliance, worker, injury, related, audit, assessment	Commscope 2014, Logitech 2008, Logitech 2009, Kulicke&Soffa 2018, Cree 2019
4	**Sustainable Product Development**	environmental, product, management, material, impact, use, waste, energy, system, report, approach, environment, process, water, reduce, standard, emission, supplier, green, iso	Sony 2012, Fujitsu 2002, Fujitsu 2001, Canon 2000, Canon 1999
5	**Cybersecurity & Data protection**	customer, security, information, product, data, service, business, company, privacy, system, quality, management, use, group, provide, include, risk, process, protection, environmental	Epson 2016, Consolidated comm. 2021, Tyro 2021, ACI 2019, América Movil 2020
6	**Financial disclosure & transparency**	employee, total, non, financial, tax, include, company, business, management, number, risk, report, annual, plan, performance, pay, year, base, share, compensation	NortonLifeLock 2021, Rexel 2006, Cohu 2019, Rexel 2007, Rexel 2008
7	**Responsible sourcing**	human, right, supply, chain, sustainability, business, principle, global, report, policy, standard, management, responsibility, environmental, corporate, social, environment, governance, approach, labor	Sony 2019, Sony 2021, Sony 2020, Flex 2020, Microsoft 2016
8	**General / Introductory**	business, company, management, environmental, group, activity, global, work, employee, sustainability, report, issue, risk, climate, program, include, change, plan, development, system	Fujitsu 2013, Fujitsu 2014, Fujitsu 2012, Trimble 2021, Fujitsu 2009
9	**Business ethics and Code of conduct**	management, report, business, governance, risk, corporate, stakeholder, conduct, responsibility, social, compliance, ethic, policy, code, sustainability, approach, issue, environmental, supplier, high	Nordic Semiconductor 2018, 2013, 2015, 2016, Lattice Semiconductor 2021
10	**Corporate Governance Agents**	board, director, committee, corporate, executive, governance, company, management, member, report, audit, officer, group, business, president, independent, chief, sustainability, responsibility, senior	Televisa 2015, Cedar Realty Trust 2020, SimCorp 2015, NortonLife 2021, SimCorp 2014
11	**Financial reporting**	million, year, service, billion, mobile, customer, network, data, sale, increase, revenue, market, cost, use, wireless, business, time, net, include, new	Rexel 2006, 2007, 2008, 2009, NortonLife2021

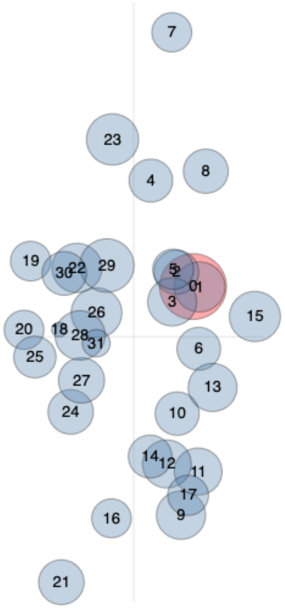
12	**Volunteering and local community support**	employee, community, support, volunteer, program, local, activity, provide, work, health, event, organization, company, group, give, team, donation, include, report, service	Allscripts 2020, Onto 2020, Bottomline 2018, Bottomline 2019, Bottomline 2021
13	**Energy efficiency in infrastructures**	energy, system, power, use, product, data, reduce, efficiency, center, technology, design, new, equipment, efficient, environmental, customer, solution, improve, service, battery	Amazon 2019, SolarEdge 2018, Trimble 2019, ESCO 2020, Facebook 2019
14	**Services & Products**	customer, product, service, technology, use, solution, help, business, improve, provide, network, new, work, reduce, enable, design, protect, system, energy, need	Trimble 2019, Maxar 2020, AT&T 2014, AMD 2019, CalAmp2020
15	**Labor Rights**	employee, work, labor, right, employment, child, human, health, worker, leave, gender, include, workplace, provide, hour, policy, train, discrimination, diversity, take	isentia 2018, SimCorp 2018, Jabil 2019, PTC 2019, Nordic 2016
16	**Company culture**	employee, community, team, work, business, people, program, support, company, culture, help, development, create, value, corporate, opportunity, make, leadership, provide, global	Insight 2021, Insight 2022, Comcast 2016, Fiserv 2020, Insight 2020
17	**Workplace safety metrics**	rate, employee, year, work, time, day, number, asia, europe, hour, region, total, america, injury, case, lose, global, company, africa, average	Netgear 2016, SunPower 2016, Xerox 2018, CommScope 2014, TE 2019
18	**CSR strategy and activities**	year, company, employee, business, target, work, report, group, plan, goal, base, result, data, include, performance, new, program, management, global, product	Bang & Olufsen 2018, 2012, 2015, 2014, 2019
19	Noise	uif, boe, poz, gps, xjui, tdbm, uibu, jut, psqpsbujpoo, bsf, gspn, spvq, qspevdut, fqpsujoh, ibt, pqzsjhiu, xijdi, cvtjoftt, fnqmpzfft, owjsponfoubm	
20	**Responsible Mineral Sourcing**	supplier, mineral, conflict, supply, chain, source, responsible, use, report, country, due, smelter, program, free, product, industry, business, work, policy, include	Nordic 2013, Qorvo 2019, Nordic 2014, Skyworks 2018, Kemet 2018
21	Noise *(airport management company that shouldn't be in the corpus)*	airport, service, year, sea, business, company, area, group, base, new, energy, network, passenger, data, use, market, increase, system, source, time	SEA 2013, 2012, 2015, 2016, 2017
22	**CSR Strategy & Framework**	business, sustainability, goal, sustainable, report, impact, environmental, stakeholder, global, value, corporate, responsibility, product, social, environment, strategy, change, development, company, customer	IHS Markit 2017, IHS Markit 2016, Bang & Olufsen 2019, Viavi 2020, Aveva 2021
23	**Energy-efficiency Projects**	center, project, new, plant, technology, use, office, solar, city, first, award, site, program, system, include, state, support, power, provide, area	Bottomline 2018, BottomLine 2019, ComCast 2012, Maxar 2020, SunPower 2016
24	**Diversity and Inclusion in the workplace**	employee, woman, program, year, diversity, female, workforce, career, total, development, people, new, gender, hire, leadership, group, talent, work, diverse, global	Google 2022, 2U 2018, ComCast 2020, Shopify 2019, ComCast 2019

25	**Supply chain monitoring**	supplier, audit, risk, conduct, code, assessment, supply, business, chain, process, management, compliance, rba, eicc, standard, requirement, include, program, ensure, system	Dell 2020, Qorvo 2019, Motorola 2016, Motorola 2015, Nordic 2013
26	**Social - Generic**	employee, work, business, company, customer, community, health, social, development, management, program, value, support, safety, provide, corporate, global, environment, team, practice	KLA-Tencor 2017, dun & bradstreet 2020, dun & bradstreet 2021, Akamai 2021, Data#3 2021
27	**CSR reporting - Generic**	report, gri, sustainability, csr, index, corporate, responsibility, performance, information, social, data, global, disclosure, environmental, statement, assurance, reporting, management, year, governance	Adva 2020, Adva 2019, Logitech 2011, Flex 2020, Flex 2021
28	**Corporate Governance**: standards of business and anti-corruption policy	report, employee, compliance, policy, ethic, management, conduct, stakeholder, anti, business, information, corporate, code, corruption, concern, include, gri, standard, company, governance	Kitron 2018, Keysight 2017, Kitron 2017, Bang & Olufsen 2016, Kitron 2019
29	**Awards & Recognition**	community, award, company, world, business, corporate, people, development, technology, work, innovation, support, employee, global, program, social, best, help, create, sustainable	ComCast 2016, ComCast 2018, ComCast 2015, Blackbaud 2018, Fiserv 2020
30	**Education Support**	program, student, school, education, support, help, provide, community, child, technology, learn, employee, foundation, skill, work, million, volunteer, science, university, people	Accenture 2020, Allscripts 2020, ComCast 2018, América Movil 2013, Oracle 2020
31	**Technology & Innovation**	technology, solution, customer, service, business, help, product, new, network, innovation, world, work, digital, industry, energy, develop, use, development, support, make	Bechtle 2015, RedHat 2019, Québecor 2020, RedHat 2020, AMD 2019
32	Noise	tc, business, management, emplo, tion, par, int, sustainabilit, thr, include, ough, ing, ener, ear, ompan, ser, global, yees, fully, base	

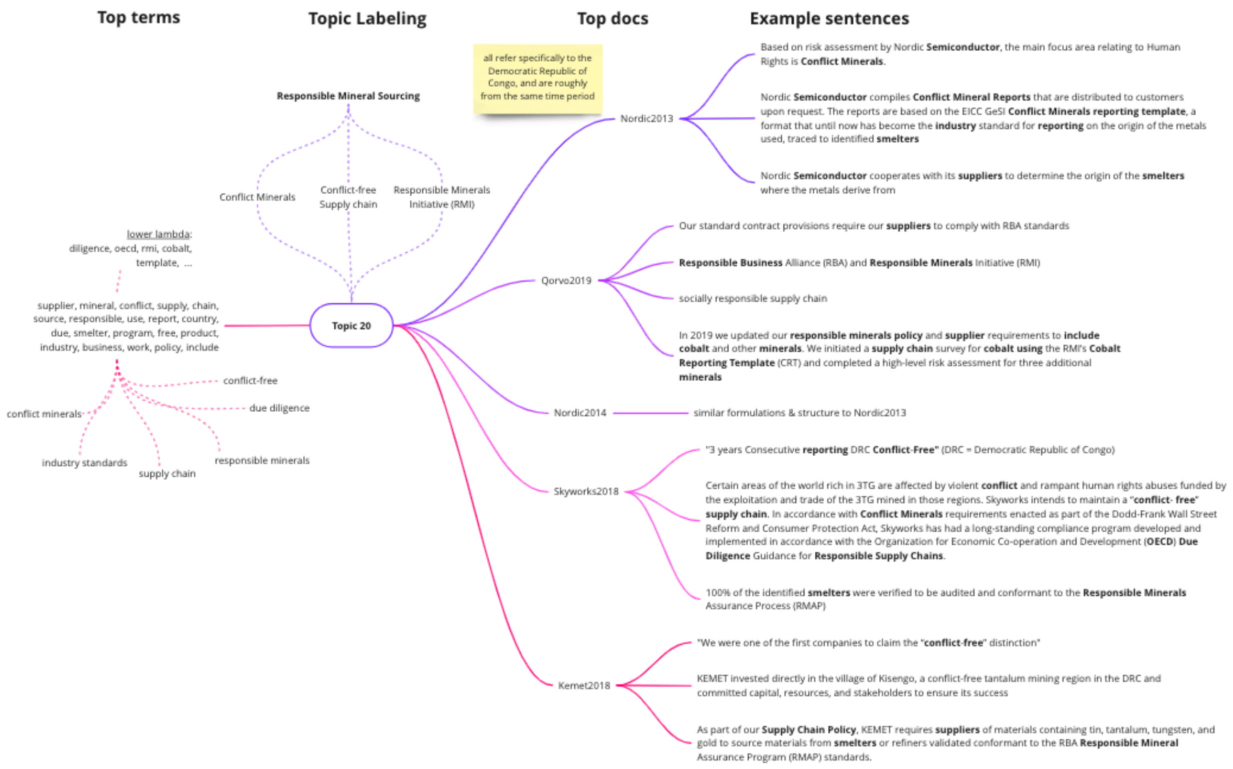
Table 6. Biterm Topic Model summary

Appendix

A. Intertopic distance plot (spectral embedding) (0-31)



B. Thematic analysis process: example of a topic mind map


Top terms
Topic Labeling
Top docs
Example sentences
Responsible Mineral Sourcing
Conflict Minerals
Conflict-free Supply chain
Responsible Minerals Initiative (RMI)
lower lambda:
diligence, oecd, rmi, cobalt, template, ...
supplier, mineral, conflict, supply, chain, source, responsible, use, report, country, due, smelter, program, free, product, industry, business, work, policy, include
conflict-free
due diligence
conflict minerals
industry standards
supply chain
responsible minerals
Topic 20
Nordic2013
Qorvo2019
Nordic2014
Skyworks2018
Kemet2018

all refer specifically to the Democratic Republic of Congo, and are roughly from the same time period

Based on risk assessment by Nordic **Semiconductor**, the main focus area relating to Human Rights is **Conflict Minerals**.

Nordic **Semiconductor** compiles **Conflict Mineral Reports** that are distributed to customers upon request. The reports are based on the EICC GeSI **Conflict Minerals reporting template**, a format that until now has become the **industry** standard for **reporting** on the origin of the metals used, traced to identified **smelters**

Nordic **Semiconductor** cooperates with its **suppliers** to determine the origin of the **smelters** where the metals derive from

Our standard contract provisions require our **suppliers** to comply with RBA standards

Responsible Business Alliance (RBA) and **Responsible Minerals** Initiative (RMI)

socially responsible supply chain

In 2019 we updated our **responsible minerals policy** and **supplier** requirements to **include cobalt** and other **minerals**. We initiated a **supply chain** survey for **cobalt using** the RMI's **Cobalt Reporting Template** (CRT) and completed a high-level risk assessment for three additional **minerals**

similar formulations & structure to Nordic2013

"3 years Consecutive **reporting** DRC **Conflict-Free"** (DRC = Democratic Republic of Congo)

Certain areas of the world rich in 3TG are affected by violent **conflict** and rampant human rights abuses funded by the exploitation and trade of the 3TG mined in those regions. Skyworks intends to maintain a "**conflict- free**" **supply chain**. In accordance with **Conflict Minerals** requirements enacted as part of the Dodd-Frank Wall Street Reform and Consumer Protection Act, Skyworks has had a long-standing compliance program developed and implemented in accordance with the Organization for Economic Co-operation and Development (**OECD**) **Due Diligence** Guidance for **Responsible Supply Chains**.

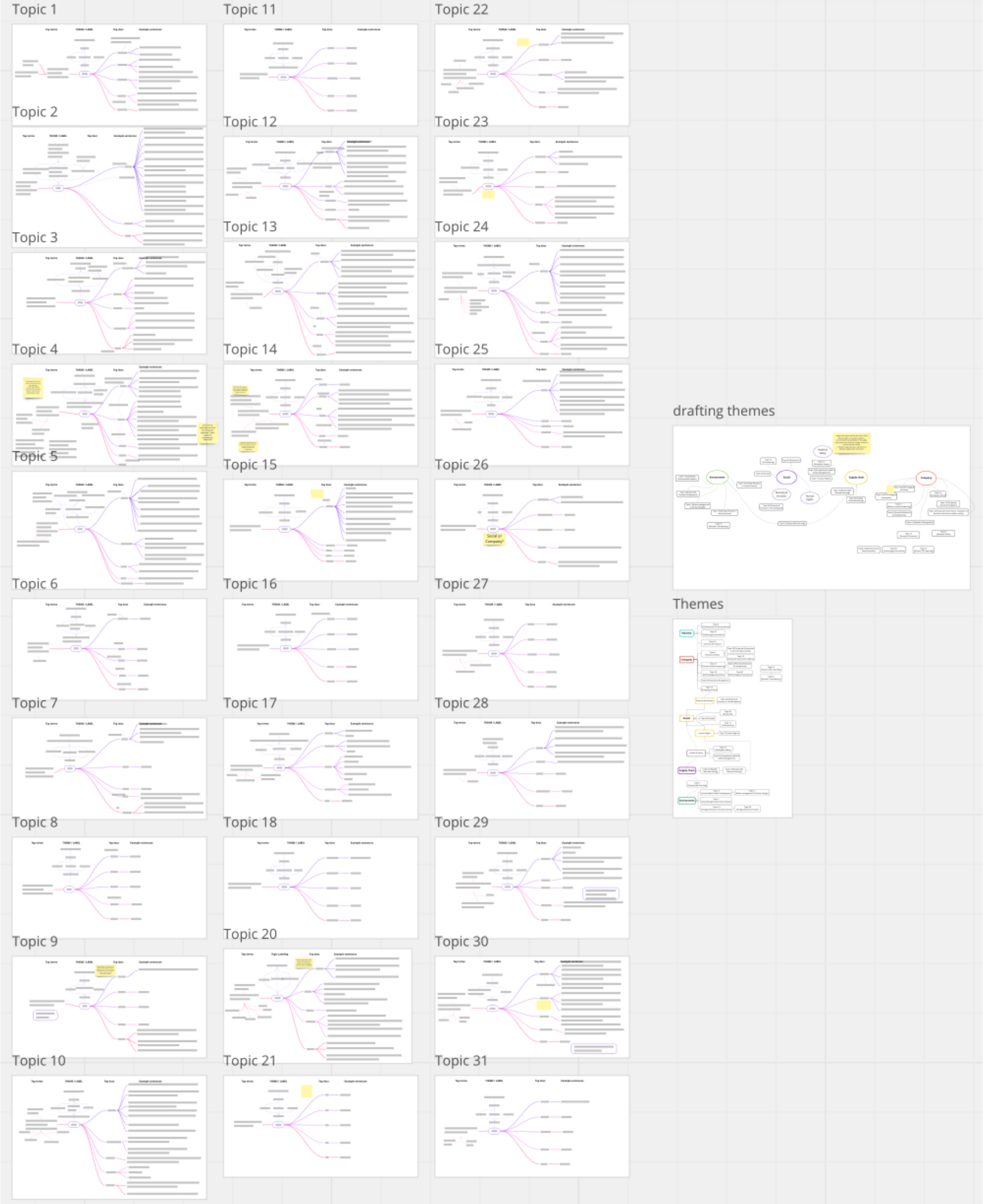
100% of the identified **smelters** were verified to be audited and conformant to the **Responsible Minerals** Assurance Process (RMAP)

"We were one of the first companies to claim the "**conflict-free**" distinction"

KEMET invested directly in the village of Kisengo, a conflict-free tantalum mining region in the DRC and committed capital, resources, and stakeholders to ensure its success

As part of our **Supply Chain Policy**, KEMET requires **suppliers** of materials containing tin, tantalum, tungsten, and gold to source materials from **smelters** or refiners validated conformant to the RBA **Responsible Mineral** Assurance Program (RMAP) standards.

C. Thematic analysis process – overview



References

Chae, B., & Park, E. (2018). Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. Sustainability, 10(7), 2231–. https://doi.org/10.3390/su10072231

Goloshchapova, I., Poon, S.-H., Pritchard, M., & Reed, P. (2019). Corporate social responsibility reports: topic analysis and big data approach. The European Journal of Finance, 25(17), 1637–1654. https://doi.org/10.1080/1351847X.2019.1572637

Jaworska, S., & Nanda, A. (2018). Doing Well by Talking Good: A Topic Modelling-Assisted Discourse Study of Corporate Social Responsibility. Applied Linguistics, 39(3), 373–399. https://doi.org/10.1093/applin/amw014

Kumar, V., & Srivastava, A. (2021). Mapping the evolution of research themes in business ethics: a co-word network analysis. *VINE Journal of Information and Knowledge Management Systems*, *ahead-of-print*(ahead-of-print). https://doi-org.myaccess.library.utoronto.ca/10.1108/VJIKMS-10-2020-0199

Kumar, V., & Srivastava, A. (2022). Trends in the thematic landscape of corporate social responsibility research: A structural topic modeling approach. Journal of Business Research, 150, 26–37. https://doi.org/10.1016/j.jbusres.2022.05.075

Latapí Agudelo, M. A., Jóhannsdóttir, L., & Davídsdóttir, B. (2019). A literature review of the history and evolution of corporate social responsibility. International Journal of Corporate Social Responsibility, 4(1), 1–23. https://doi.org/10.1186/s40991-018-0039-y

Mangsor, N.S.M.N., Nasir, S.A.M, Yaacob, W.F.W., Ismail, Z. & Rahman, S.A. (2022). Analysing corporate social responsibility reports using document clustering and topic modeling techniques. Indonesian Journal of Electrical Engineering and Computer Science, 26(3), 1546–1555. 10.11591/ijeecs.v26.i3.pp1546-1555

Moskowitz, M. (2002). Musings: What Has CSR Really Accomplished? Business Ethics (Madison, Wis.), 16(4), 4–4. https://doi.org/10.5840/bemag2002163/419

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python,. JMLR 12, pp. 2825-2830.

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Székely, N., & Vom Brocke, J. (2017). 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–e0174807. https://doi.org/10.1371/journal.pone.0174807

Vogel, D. (2005). The market for virtue: the potential and limits of corporate social responsibility. Brookings Institution Press.

Wickert, C., & Risi, D. (2019). Corporate social responsibility. Cambridge University Press.