IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 February 2024, accepted 6 May 2024, date of publication 22 May 2024, date of current version 6 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3404368

RESEARCH ARTICLE

Topic Trends in Sustainability Disclosure of German DAX 40 Companies–A Text Mining-Based Analysis

TOBIAS CONTALA, ALEXANDER-MICHAEL GERK, JOHANNES HOETTLER, AND RICARDO BUETTNER, (Senior Member, IEEE)
Chair of Information Systems and Data Science, University of Bayreuth, 95447 Bayreuth, Germany
Corresponding author: Ricardo Buettner (ricardo.buettner@uni-bayreuth.de)

This work was supported by the Open Access Publishing Fund of the University of Bayreuth.

ABSTRACT Given the increasing importance of sustainability in the business world, there has been growing interest in using topic modeling approaches to identify reported topics in corporate sustainability reports (CSR). Due to the inconsistent legal foundation and different sustainability standards, the content of individual reports can vary greatly. In this paper, the corporate social responsibility reports of DAX 40 companies from 2017 to 2021 are therefore analyzed using Latent Dirichlet Allocation (LDA). In particular, we attempt to identify topics that are suggested by the Global Reporting Initiative (GRI) Sustainability Standard for large public companies. In addition, a comparison is made throughout the years. The study shows that specific guidelines of the GRI can only be identified to a certain degree using LDA. Although some topics that partly reflect the content of the GRI can be found by the model, the overall structure of the GRI can't be replicated. Overall, this study shows that an evaluation of the content of sustainability reports can be successful in terms of the relevance of the reported topics, although the results depend heavily on the respective pre-processing steps.

INDEX TERMS Corporate social responsibility reporting (CSR), global reporting initiative (GRI), latent dirichlet allocation (LDA), topic modeling.

I. INTRODUCTION

The disclosure of corporate social responsibility (CSR) activities has become increasingly important over the years. In 2014, the European Union adopted the CSR directive 2014/95/EU, which obliges the member states to implement this directive in national law. The CSR directive was implemented in national law of the EU members. For example, this directive is applied in Germany through the national CSR-RUG. However, it is problematic that it gives corporations extensive leeway and leaves the content requirements vaguely formulated. As a result, the reporting corporations can largely decide individually on the information to be included in their reports, which makes it difficult to compare these reports [1]. In general, this has led to the development of several standards for sustainability reports, which is why the content of individual CSR reports can vary widely [2]. The most widespread standards are the Global Reporting Initiative (GRI), the ISO 26 000 or the Integrated Report (IR) [3], [4], [5].

The theoretical concept of the study is based on the signaling theory. Due to the asymmetrical distribution of information between management and equity investors, CSR reporting should provide information about the activities of sustainable and social value creation [6]. However, this disclosure also triggers reactions among stakeholders. Various studies have examined the effects of positive CSR reporting. For example, Saeidi et al. [7] found a positive influence of CSR reporting on the reputation of a company. Furthermore, Dhaliwal et al. [8] were able to show that positive CSR reporting can lead to a decrease in the cost of capital.

The associate editor coordinating the review of this manuscript and approving it for publication was Ajit Khosla.

Due to the emerging signal effect of CSR reports and the problem of inconsistent sustainability reports, the interest of this work consists in an overview of currently reported sustainability topics. This paper will therefore examine the contents of CSR reports of German DAX (Deutscher Aktien Index) companies over a period of five years by means of textual analysis. The aim of this work is (1.) to classify the main topics of the individual CSR reports based on predefined GRI Topic standards and (2.) to compare them over time. We then use these results to provide a comparison between the topics captured in the sustainability reports and the topics prescribed in the standard. In this paper, we analyze the CSR reporting practice among German companies. Using Text Mining algorithms like Latent Dirichlet Allocation (LDA) topic Analysis, we examine the current practice among the corporations included in the DAX 40. We process their CSR reports since the introduction of mandatory disclosure in 2017. Due to the volume of data, the systematical analysis of CSR reports can't be done manually. We use LDA textual analysis to detect leading topics included in the reports. The usage of text mining techniques on corporate data isn't new. It has been used in various research fields and disciplines, including for instance, finance [9], and marketing [10], [11].

Text Mining is a vast discipline, and its techniques vary in terms of complexity and scope [12]. Topic Modeling approaches, such as LDA provide a comprehensive overview of unstructured, textual data [13]. Text Mining is therefore well suited for the analysis of CSR reports to identify deviation between obligatory information and actual information included in the reports as well as understanding the development of issues in time and in context of regulatory changes [14], [15], [16].

To the best of our knowledge existing literature combining non-financial reporting and text mining is insufficient. Especially literature on CSR analysis using text mining techniques is rare or due to major changes in the reporting directives outdated. This paper contributes to the field by combining recent data with state-of-the-art text mining methods. With our analysis, we gain insights into the structure and content of German CSR reports and assess the relevance of the topics contained, measured against the respective reporting standards. This not only allows a comparison between the standard requirements and the actual reporting, but due to the periodic view, changes in the structure and practice of CSR reporting can also be examined for each individual year. Finally, our results confirm the applicability of LDA topic modeling in the field of sustainability disclosure research.

II. THEORETICAL BACKGROUND

A. CSR REPORTING

In the European Union, all large capital market-oriented companies and groups with more than 500 employees have been obliged to publish a non-financial report since 2017. This was decided by the Non-Financial Reporting Directive 2014/95/EU, which obliges member states to implement this in their national law [1].

To this end, CSR-RUG was passed in Germany. The CSR-RUG applies from 01.01.2017 and obliges all large capital market-oriented companies with more than 500 employees to publish a non-financial report, also called a sustainability report, in the management report or on the website of the company or group. The companies are to provide information on environmental issues, employee issues, social issues, respect for human rights and the fight against corruption and bribery. Precise specifications on the content of these topics are not given, but only examples of topics, such as greenhouse gas emissions and water consumption for environmental issues. Companies can also omit adverse disclosures on topic areas, which is regulated in.

Furthermore, the CSR-RUG does not provide any precise regulations on the structure or preparation of the sustainability report and leaves this up to the companies. However, frameworks can be used for the preparation, which should then be mentioned by name in the report. This means that companies are free to choose between the frameworks. Frameworks for the preparation of sustainability reports include the GRI, ISO 26 000 or the IR [3], [4], [5].

In addition to the frameworks mentioned here, there are other frameworks for the preparation of sustainability reports. An overview of three of the above-mentioned frameworks is given in Table 1. All three frameworks have a common purpose. It becomes clear that all three frameworks are internationally applicable, but only the GRI provides topic-specific guidelines.

An analysis of the underlying frameworks of CSR reports shows, that most companies prepare their sustainability reports based on the GRI standard [20].

B. GRI STANDARD

Table 1 shows that the GRI is a very comprehensive framework that provides guidance on specific ESG (Environment, Social, and Governance) topics. In addition, studies such as Guo and Yang [21] show that the GRI is most frequently used for the preparation of sustainability reports. Guo and Yang show this by means of a survey of 30 Dow Jonas companies. In this survey, 66 per cent of the companies used the GRI as a standard and reported according to it [21].

The GRI framework is divided into three standard parts. The first part is the Universal Standard, which is a general part that provides information on the requirements and principles of the GRI and clarifies the requirements and guidelines for the report. The second part refers to Sector Standards. The Topic Standard or topic-specific standard, form the third part of the GRI framework. The topic-specific standards are subdivided into the ESG topic areas (E = Environment, S = Social, G = Governance). The topic area Environment is contained in the GRI 300, Social aspects are contained in the GRI 400 and Governance related topics are listed in the economic section (GRI 200). For these topic areas the GRI defines specified subtopics. For example, GRI 300 includes the subtopics 301 materials or 308 supplier environmental

TABLE 1. Overview of reporting frameworks [17], [18], [19].

	GRI	ISO 26 000	IR
publisher	Global Reporting Initiative	ISO (International Organization for Standardization)	IIRC (International Intregrated Reporting Concil)
volume	870 sides	160 sides	58 sides
scope of application	international	international	international
prescribing specific ESG topics	yes	no	no

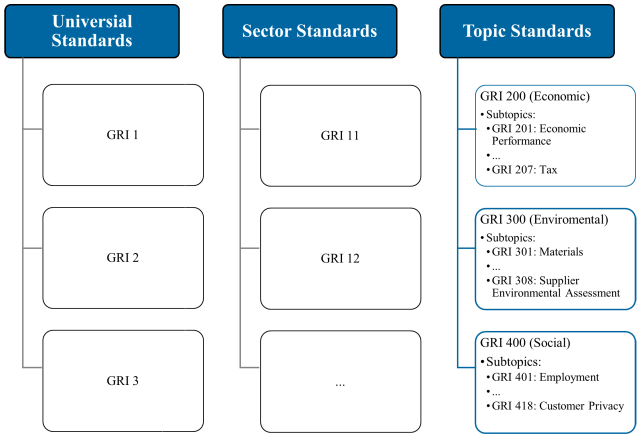

FIGURE 1. Structure of GRI framework in accordance with [17].

assessment. In some cases, these subtopics are subdivided again. The structure of the GRI is shown in Figure 1. A complete overview of the GRI Topic Standards 200 to 300 is depicted in Table 22 in the appendix.

III. RELATED WORK

A. TEXTUAL ANALYSIS AND ESG REPORTING

The use of textual analysis and text mining techniques is becoming increasingly important in the financial literature [22]. This need lies in growing proportions of textual passages in financial disclosure and knowledge of potentially important insights within this unstructured data [23].

An example of such an approach is provided by Lang and Stice-Lawrence [24], who examined the length of annual financial statements of more than 15,000 companies. The term textual analysis includes a set of methods [22]. For example, conducting an analysis of the information content of 10-K reports, Li [25] followed a Naive Bayes classification approach to determine the information content and tone of financial-statements. Loughran and McDonald [22] on the other hand, opted for a ''bag of words'' approach to determine the tone of 10-K financial statements. Sai et al. [23] investigated how the textual component of annual reports can be used to deduct meaningful information about company's performance by employing textual analysis. They performed sentiment analysis as well as statistical calculations to explore the relation between the emotions in annual reports and the future performance of the firm.

In the era of Big Data, however, another method has been gaining interest from financial and non-financial disclosure researchers due to its promising application on large amounts of unstructured textual data. Seemakurthi et al. [26] applied various advanced supervised machine learning and natural language processing (NLP) techniques, including LDA to the problem of detecting fraud in financial reporting documents. They used LDA to a collection of type 10-K financial reports to generate document-topic frequency matrix, and then submit these data to advanced classification algorithms. Zhu et al. [27] analyzed corporate risk disclosures as part of U.S. public companies' financial reports to identify risk types and examined their potential implications on stock returns. They applied a Sentence LDA (Sent-LDA) model to infer risk types. They then quantified the impact of these risk factors on stock returns. Hoberg and Lewis [28] used text-based analysis of 10-K disclosures to examine verbal disclosure of fraudulent firms. They employed LDA to identify interpretive verbal topics that firms emphasize compared to peers. This is based on the central hypothesis that the verbal factor structure of these text sections can be interpreted to reveal likely mechanisms that surround fraudulent behavior. Dyer et al. [29] documented trends in 10-K disclosure over the period 1996-2013, with increases in length, boilerplate, stickiness, and redundancy and decreases in specificity, readability, and the relative amount of hard information, using LDA to examine the specific topics. Huang et al. [30] applied topic modeling (LDA) to compare the content of reports from financial analysts after the annual general meeting with the actual topics of the general meeting to investigate how financial analysts serve their information intermediary role.

In the context of non-financial disclosure like the analysis of companies ESG/CSR reports, text mining applications are considered to have a lot of potential. With more corporations publishing corporate sustainability reports due to stricter regulations, the manual process of analyzing the reports is becoming inefficient and tedious [31], [32], [33]. In recent years several publications performing textual analysis and text mining methods on corporate reports like ESG/CSR disclosure were published. Following the joint goal of gaining insights in a vast amount of unstructured ESG data and to extract hidden information in corporate disclosure the researchers target various research questions. Modapothala and Issac [34] examined and analyzed corporate environmental reports in line with GRI disclosures to identify differences founded in the nature of the business using a data

TABLE 2. Output search strings.

Search string	Science Direct	Web of Science	IEEEXplore	Jstor
"csr" AND "text mining"	55	15	3	1
"sustainablility reporting" AND "text mining"	7	1	1	1
"csr" AND "machine learning"	132	6	22	5
"sustainability reporting" AND "machine learning"	24	1	1	2
("non" AND "financial reporting") AND "text mining"	56	1	0	2
("non" AND "financial reporting") AND "machine learning"	150	2	1	8
Additional information: - All searches from 2018 onwards, as CSR RUG only applies by law in Germany from then onwards - Selected Categories: Business, Management, Accounting, Economics and Finance				

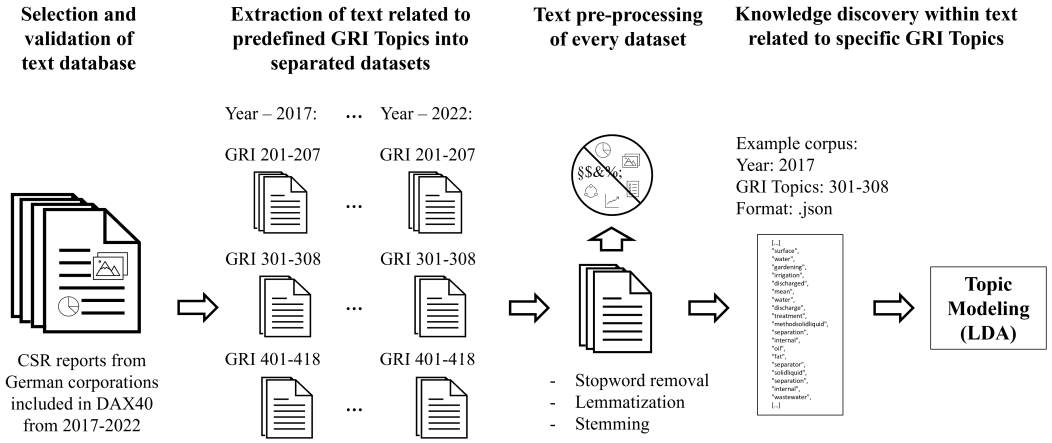

FIGURE 2. Construction of database and methodology of intratopic-topic modeling.

TABLE 3. Top 25 bigrams included in GRI Topics 200-400.

Bigrams	All	200	300	400
supply chain	16	4	6	6
board management	15	4	5	6
financial statement	12	3	3	6
corporate governance	11	2	4	5
consolidated financial	10	3	2	5
human right	10	3	4	3
climate change	9	2	3	4
combined management	9	3	2	4
health safety	9	1	4	4
management system	9	2	4	3
climate protection	8	2	4	2
environmental social	8	0	2	6
per cent	8	2	3	3
raw material	8	1	3	4
additional information	7	1	2	4
due diligence	7	2	3	2
environmental protection	7	2	2	3
ifrs information	7	1	2	4
management consolidated	7	1	2	4
value chain	7	1	2	4
co2 emission	6	2	2	2
code conduct	6	1	3	2
executive board	6	1	2	3
key figure	6	1	3	2
risk management	6	1	1	4

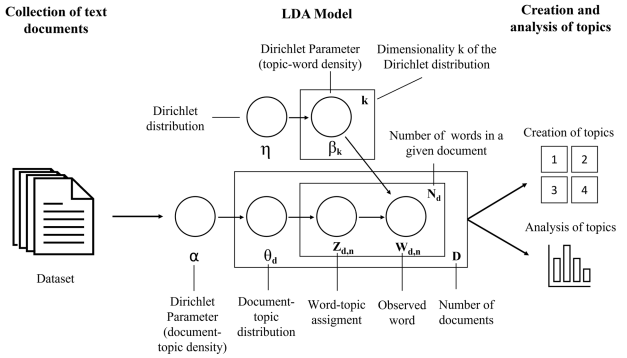

FIGURE 3. LDA model structure in accordance with [40] and [46].

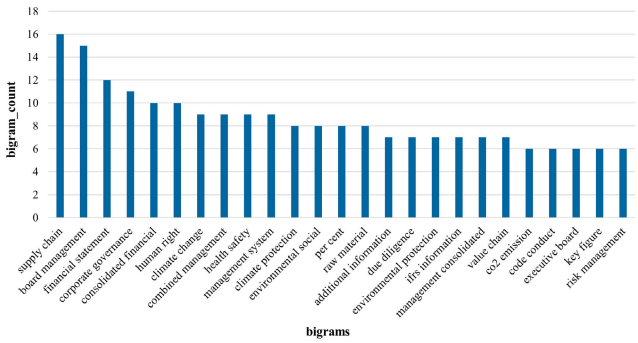

FIGURE 4. Top 25 bigrams represented in entirety of topics.

mining approach. Freundlieb and Teuteberg [35] analyzed CSR reports published by 97 market-listed companies from the USA, Germany and the rest of Europe to point out regional differences as well as changes over time.

TABLE 4. GRI 200 - Weight and word tount of topic 0-2 (all reports).

	Topic: 0			Topic: 1			Topic: 2		
	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count
GRI 200s	**financial statement**	0,00483	8975	**supervisory board**	0,01402	9796	**management approach**	0,00453	1807
	consolidated financial	0,00410	7459	**executive board**	0,00935	6189	**assurance engagement**	0,00275	1129
Number of	**fair value**	0,00291	3322	**board member**	0,00369	2463	**material topic**	0,00241	723
bigrams:	**cash flow**	0,00278	2755	**financial statement**	0,00355	8975	**financial statement**	0,00218	8975
14940	**management consolidated**	0,00172	2017	**corporate governance**	0,00326	3857	**consolidated financial**	0,00202	7459
	combined management	0,00167	2514	consolidated financial	0,00217	7459	limited assurance	0,00175	803
	statement ifrs	0,00149	1259	board management	0,00191	3588	explanation material	0,00154	364
	financial asset	0,00146	1654	board executive	0,00155	1773	evaluation management	0,00153	355
	interest rate	0,00141	1351	board supervisory	0,00153	746	human right	0,00143	5338
	ifrs information	0,00129	1117	annual general	0,00148	983	gri standard	0,00140	597
	additional information	0,00127	1947	general meeting	0,00148	978	topic boundary	0,00129	252
	stakeholder combined	0,00127	1071	member executive	0,00144	949	gri content	0,00119	492
	intangible asset	0,00125	1413	audit committee	0,00138	985	approach component	0,00107	230
	income tax	0,00113	1370	member supervisory	0,00113	915	combined management	0,00106	2514
	financial liability	0,00107	954	executive director	0,00109	1648	boundary management	0,00101	122
	profit loss	0,00104	1107	combined management	0,00105	2514	component evaluation	0,00101	181
	property plant	0,00103	1251	target achievement	0,00098	834	supply chain	0,00094	3867

TABLE 5. GRI 300 - Weight and word tount of topic 0-2 of (all reports).

	Topic: 0			Topic: 1			Topic: 2		
	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count
GRI 300s	**risk management**	0,00282	1626	**real estate**	0,00119	368	**supply chain**	0,00382	4193
	data protection	0,00251	920	**raw material**	0,00107	3111	**human right**	0,00346	4751
Number of	**management system**	0,00237	2230	**environmental protection**	0,00100	1449	**code conduct**	0,00200	1351
bigrams:	**human right**	0,00186	4751	**customer satisfaction**	0,00095	287	**work council**	0,00147	238
9587	**health safety**	0,00163	1496	**business model**	0,00080	891	**parental leave**	0,00141	85
	management approach	0,00126	1931	community development	0,00077	98	year year	0,00140	50
	climate change	0,00121	1455	climate change	0,00075	1455	business partner	0,00132	889
	corporate governance	0,00116	2135	action area	0,00074	201	proportion woman	0,00128	89
	risk assessment	0,00110	700	passenger car	0,00074	383	number employee	0,00122	239
	business partner	0,00107	889	climate protection	0,00068	1505	equal opportunity	0,00122	257
	code conduct	0,00104	1351	key figure	0,00066	742	raw material	0,00116	3111
	compliance management	0,00097	488	life cycle	0,00065	793	corporate culture	0,00105	237
	sustainability strategy	0,00097	901	residential environment	0,00064	78	human resource	0,00102	409
	board management	0,00096	1424	management system	0,00061	2230	management system	0,00094	2230
	occupational health	0,00096	594	development project	0,00058	144	key figure	0,00092	742
	risk opportunity	0,00089	1122	supply chain	0,00058	4193	working hour	0,00092	205
	corporate responsibility	0,00085	682	circular economy	0,00056	655	training course	0,00088	403
	occupational safety	0,00083	426	society customer	0,00054	61	working environment	0,00081	145
	business activity	0,00079	940	corporate governance	0,00053	2135	due diligence	0,00078	971
	supply chain	0,00077	4193	sustainability society	0,00053	54	health safety	0,00076	1496
	compliance risk	0,00076	273	environmental impact	0,00052	847	board management	0,00075	1424
	product service	0,00075	926	customer service	0,00051	119	employee representative	0,00070	242
	environmental management	0,00070	717	capital market	0,00050	342	woman management	0,00068	218
	environmental protection	0,00069	1449	greenhouse gas	0,00049	1837	attractive employer	0,00068	102
	compliance officer	0,00069	301	modernization measure	0,00046	75	total number	0,00065	195
	environmental social	0,00064	1045	construction project	0,00046	89	female employee	0,00063	36
	integrity compliance	0,00063	136	sustainable development	0,00046	969	vocational training	0,00062	68
	sustainable development	0,00061	969	opportunity risk	0,00045	714	diversity equal	0,00062	160
	global compact	0,00060	979	sustainable construction	0,00044	85	management position	0,00061	176
	information security	0,00058	205	health safety	0,00043	1496	level management	0,00060	48
	core business	0,00055	329	long term	0,00042	366	occupational safety	0,00059	426
	environment society	0,00053	338	gas emission	0,00042	1132	business human	0,00058	271
	climate risk	0,00053	323	renewable energy	0,00041	302	management level	0,00058	163
	due diligence	0,00052	971	performance indicator	0,00041	515	reporting period	0,00057	844
	sustainable finance	0,00051	285	steering committee	0,00040	193	working condition	0,00057	390
	climate protection	0,00047	1505	sustainability environment	0,00040	85	management board	0,00056	326

By conducting a text mining supported quantitative content analysis on the occurrence frequency of positively and negatively connoted bigrams/keywords within CSR reports, they identified shifts in regard of the CSR reports' focus, and

TABLE 6. GRI 400 - Weight and word tount of topic 0-2 (all reports).

	Topic: 0			Topic: 1			Topic: 2		
	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count
GRI 400s	**real estate**	0,00300	295	**health safety**	0,00574	2158	**co2 emission**	0,00386	984
	customer satisfaction	0,00161	306	**occupational safety**	0,00384	813	**raw material**	0,00373	1672
Number of	**society customer**	0,00119	92	**occupational health**	0,00355	1005	**co₂ emission**	0,00344	331
bigrams:	**development project**	0,00115	139	**health management**	0,00201	390	**commercial vehicle**	0,00330	173
9858	**capital market**	0,00112	340	**management system**	0,00141	2504	**passenger car**	0,00324	153
	residential environment	0,00110	110	sustainable development	0,00123	929	life cycle	0,00306	428
	sustainable development	0,00108	929	aero engine	0,00095	390	light commercial	0,00226	81
	climate change	0,00103	1148	development goal	0,00087	479	greenhouse gas	0,00198	882
	residential real	0,00085	44	corporate citizenship	0,00086	370	car light	0,00196	79
	business model	0,00084	932	safety management	0,00079	294	environmental impact	0,00192	536
	urban development	0,00083	65	employee retention	0,00077	161	electric vehicle	0,00177	226
	customer service	0,00080	142	zero harm	0,00077	107	carbon emission	0,00176	523
	living space	0,00075	61	decent work	0,00076	176	environmental protection	0,00161	1269
	corporate governance	0,00074	2814	work economic	0,00075	129	fuel consumption	0,00147	282
	construction project	0,00073	56	employee health	0,00074	168	climate protection	0,00145	955
	see chapter	0,00071	306	social responsibility	0,00072	611	per vehicle	0,00138	113
	neighborhood development	0,00071	51	per employee	0,00069	332	gas emission	0,00135	597
	action area	0,00069	195	working environment	0,00067	396	vehicle fleet	0,00122	120
	esg rating	0,00066	107	health well-being	0,00066	150	environmentally friendly	0,00121	264
	long term	0,00064	451	economic growth	0,00063	226	renewable source	0,00119	204
	residential unit	0,00063	37	sustainability strategy	0,00063	895	entire life	0,00111	123
	green bond	0,00063	105	engine sustainability	0,00061	196	water consumption	0,00110	213
	sustainability sustainability	0,00062	138	work council	0,00060	655	production site	0,00109	584
	affordable housing	0,00061	55	hour worked	0,00059	132	use phase	0,00105	96
	estate company	0,00060	24	sdg decent	0,00056	94	per cent	0,00103	1001
	contribution urban	0,00059	42	safety health	0,00054	230	green electricity	0,00102	151
	mobility concept	0,00057	92	offer employee	0,00054	285	carbon footprint	0,00097	253
	housing industry	0,00056	29	responsible consumption	0,00054	94	supply chain	0,00096	3168
	sustainability sustainable	0,00055	162	safety employee	0,00054	175	value chain	0,00081	1180
	home customer	0,00054	31	technical service	0,00054	103	power plant	0,00081	695
	residential property	0,00054	38	production site	0,00052	584	figure fuel	0,00079	128
	parking space	0,00052	56	consumption production	0,00050	149	natural gas	0,00078	152
	increasingly important	0,00051	124	working hour	0,00050	458	cycle assessment	0,00077	91
	modernization work	0,00050	34	risk assessment	0,00049	813	environmental management	0,00074	588
	society contribution	0,00049	36	industry innovation	0,00049	86	resource efficiency	0,00073	162

TABLE 7. GRI 200 - Weight and word tount of topic 0 (2017-2021).

	2017			2018			2019			2020			2021		
	Number of bigrams: 2128			Number of bigrams: 2519			Number of bigrams: 2756			Number of bigrams: 3212			Number of bigrams: 3969		
	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count
Topic: 0	**human right**	0,003100	658	**raw material**	0,001287	417	**per cent**	0,003069	246	**executive board**	0,002155	1637	**human right**	0,001857	1491
	code conduct	0,001265	283	**human right**	0,000943	682	**supply chain**	0,000890	803	**supervisory board**	0,002152	2477	**board management**	0,001265	1303
	risk management	0,001249	369	**financial statement**	0,000875	1783	**climate protection**	0,000698	408	**financial statement**	0,001632	2309	**per cent**	0,000931	599
	work council	0,001076	127	**interest rate**	0,000874	290	**environmental portfolio**	0,000557	75	**consolidated financial**	0,001268	1809	**supply chain**	0,000703	984
	commercial vehicle	0,000989	55	**executive board**	0,000814	977	**human right**	0,000517	1085	**board member**	0,001214	634	**supervisory board**	0,000701	2836
	technical service	0,000963	48	supply chain	0,000768	512	greenhouse gas	0,000503	298	board executive	0,001075	429	risk management	0,000693	764
	supply chain	0,000863	468	per cent	0,000657	655	employee engagement	0,000501	132	executive director	0,000977	364	financial statement	0,000667	2700
	internal control	0,000839	170	board management	0,000655	716	corporate governance	0,000500	896	environmental portfolio	0,000974	134	management system	0,000647	762
	business partner	0,000761	173	consolidated financial	0,000590	1307	sport car	0,000462	101	supply chain	0,000888	903	due diligence	0,000630	373
	control system	0,000734	53	supervisory board	0,000589	1843	vehicle delivered	0,000430	29	board management	0,000779	878	united state	0,000614	454
	fair value	0,000676	450	constant currency	0,000524	130	carbon dioxide	0,000422	59	corporate governance	0,000730	976	corporate governance	0,000580	791
	risk opportunity	0,000675	138	eurex clearing	0,000512	152	increase per	0,000414	27	co2 emission	0,000581	437	environmental social	0,000541	310
	management system	0,000645	425	united state	0,000508	323	co2 emission	0,000408	311	target achievement	0,000577	222	health safety	0,000533	557
	risk assessment	0,000621	119	fair value	0,000487	615	board management	0,000405	553	per cent	0,000523	513	climate risk	0,000514	244
	employee representative	0,000601	84	key figure	0,000469	201	carbon emission	0,000397	172	member board	0,000519	257	sustainable finance	0,000486	250
	strategic objective	0,000598	33	management system	0,000399	482	raw material	0,000388	533	greenhouse gas	0,000517	441	climate change	0,000470	598
	highest governance	0,000588	42	börse annual	0,000387	363	health safety	0,000378	400	performance bonus	0,000488	188	executive board	0,000441	1648
	sustainability requirement	0,000535	30	executive supervisory	0,000385	365	consolidated financial	0,000365	1591	human right	0,000461	1253	net revenue	0,000438	190
	natural gas	0,000527	68	billion billion	0,000385	115	special item	0,000361	171	per share	0,000446	175	code conduct	0,000437	447
	whistleblower system	0,000517	21	annual executive	0,000384	353	climate change	0,000358	278	carbon dioxide	0,000429	76	environmental management	0,000416	298

also whether CSR reports were being (mis)used as marketing instruments. Chang and Cheng [36] applied text mining as a tool to explore annual corporate sustainability reports in Taiwan to investigate the industrial context of enterprises on corporate sustainability reporting and to identify the differences in viewpoints in corporate sustainability reporting.

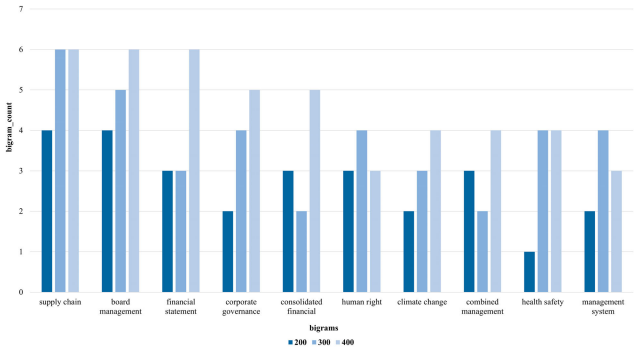

FIGURE 5. Occurrence of top 10 bigrams represented in GRI Topics 200-400 (all years accumulated).

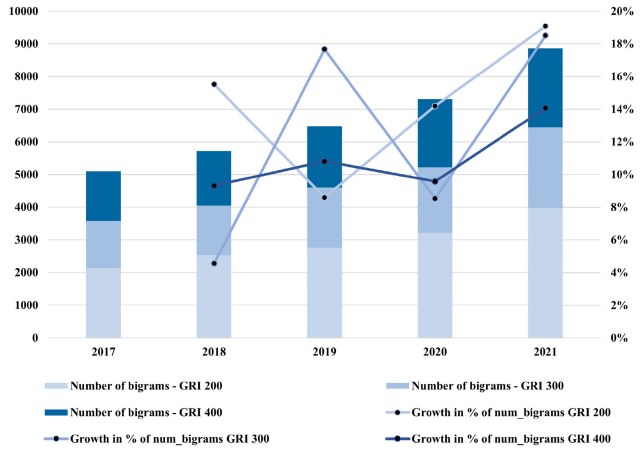

FIGURE 6. Absolute and relative development of number of bigrams per GRI Topic and year.

Shahi et al. [31] analyzed a corpus of corporate sustainability reports using a supervised learning based text mining software to build a classifier and a feature selector, to investigate the possibility of measuring the completeness of CSR reports. By assessing them based on GRI content index, they identified sections, which fulfill particular performance indicators to figure out whether or not a CSR report complies with the GRI guidelines. Te Liew et al. [37] used text mining techniques to identify frequent use of sustainability-related terms in sustainability reports in the four main sectors of the process industry. These terms presumably reflect the main concerns and emphasis of the reports and allow insights in the most prominent sustainability topics and sector-specific sustainability issues. Liu et al. [32] conducted text-mining to determine the benchmark of environment performance indicators defined in the GRI and advise an example model for the petrochemical industry's CSR reports to gain valuable information on decision making and strategy planning.

In the context of ESG reporting, LDA has also found its first applications. Benites-Lazaro et al. [38] analyzed a large amount of data from public corporate documents using LDA to identify companies' commitment to sustainability

TABLE 8. Top 20 bigrams included in GRI 200 (2017-2021).

Bigrams	All	2017	2018	2019	2020	2021
consolidated financial	30	3	7	7	7	6
financial statement	30	3	7	6	7	7
human right	27	3	6	6	6	6
supervisory board	22	1	3	6	5	7
corporate governance	21	2	3	5	7	4
executive board	21	1	2	6	5	7
supply chain	21	3	2	5	5	6
combined management	16	0	5	4	3	4
environmental social	16	0	5	5	5	1
board management	15	1	1	2	5	6
management consolidated	15	0	5	6	1	3
management system	15	1	2	4	5	3
risk management	13	3	2	2	2	4
board member	12	1	1	3	2	5
per cent	12	0	3	1	4	4
additional information	11	0	0	5	3	3
economic environmental	11	0	5	5	1	0
raw material	11	1	3	4	1	2
fair value	10	2	4	1	2	1
health safety	10	0	1	2	3	4

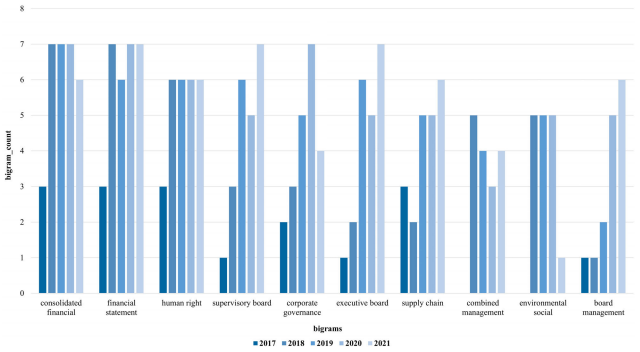

FIGURE 7. Occurrence of top 10 bigrams represented in GRI 200 (2017-2021).

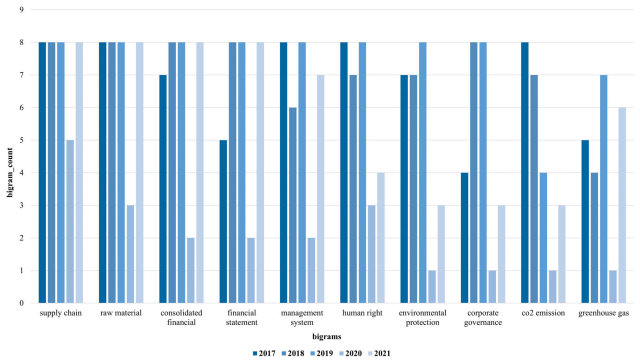

FIGURE 8. Occurrence of top 10 bigrams represented in GRI 300 (2017-2021).

and business-led governance by identifying main themes that demonstrate the rule-setting power of the sugarcane ethanol industry. Goloshchapova et al. [14] performed topic modeling using text mining and LDA analysis on CSR reports for all constituent firms of the major stock market indices of 15 countries included in MSCI Europe. Their

TABLE 9. GRI 300 - Weight and word tount of topic 0 (2017-2021).

	2017			2018			2019			2020			2021		
	Number of bigrams:	1450		Number of bigrams:	1519		Number of bigrams:	1845		Number of bigrams:	2017		Number of bigrams:	2475	
	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count
Topic: 0	assurance engagement	0,00115	170	management approach	0,00249	353	cash flow	0,00097	248	human right	0,00274	1138	human right	0,00434	1525
	raw material	0,00107	439	material topic	0,00134	185	consolidated financial	0,00096	861	financial statement	0,00250	1147	supply chain	0,00174	1059
	co2 emission	0,00090	424	evaluation management	0,00120	130	fair value	0,00093	177	consolidated financial	0,00230	1009	financial statement	0,00149	1350
	supply chain	0,00079	553	explanation material	0,00118	130	raw material	0,00092	644	management approach	0,00149	416	consolidated financial	0,00118	1042
	per cent	0,00074	139	supply chain	0,00095	568	financial statement	0,00081	820	supply chain	0,00142	946	code conduct	0,00118	337
	environmental portfolio	0,00067	85	power plant	0,00086	232	management system	0,00081	468	corporate governance	0,00103	591	supervisory board	0,00079	1084
	management system	0,00066	334	human right	0,00078	568	interest rate	0,00074	155	health safety	0,00096	321	due diligence	0,00076	328
	limited assurance	0,00062	156	raw material	0,00076	529	supply chain	0,00066	852	combined management	0,00090	368	executive board	0,00075	730
	climate change	0,00056	184	topic boundary	0,00060	84	foreign currency	0,00062	87	value chain	0,00078	382	management system	0,00073	567
	co2 emission	0,00054	111	approach component	0,00059	61	climate protection	0,00061	386	code conduct	0,00076	284	raw material	0,00072	804
	human right	0,00053	593	approach evaluation	0,00059	64	corporate governance	0,00056	518	material topic	0,00074	162	board management	0,00071	518
	risk management	0,00052	209	boundary management	0,00058	59	employee engagement	0,00053	103	global compact	0,00072	207	corporate governance	0,00069	445
	greenhouse gas	0,00052	256	health safety	0,00054	212	value chain	0,00051	309	management system	0,00066	477	risk management	0,00063	443
	fuel consumption	0,00049	123	financial statement	0,00052	777	carbon emission	0,00050	172	greenhouse gas	0,00059	477	value chain	0,00063	461
	value chain	0,00047	180	interest rate	0,00050	124	management consdidated	0,00050	229	raw material	0,00059	682	management approach	0,00058	366
	sustainable development	0,00045	160	consolidated financial	0,00046	655	health safety	0,00046	279	gri content	0,00058	178	climate protection	0,00052	364
	code conduct	0,00042	223	fair value	0,00043	174	human right	0,00046	95	co2 emission	0,00055	481	circular economy	0,00052	265
	quality control	0,00041	38	component evaluation	0,00042	47	business activity	0,00044	192	sustainable development	0,00053	204	business human	0,00052	77
	environmental protection	0,00040	235	climate change	0,00041	157	environmental protection	0,00044	340	environmental social	0,00052	270	global compact	0,00047	177
	environmental impact	0,00037	131	special item	0,00038	92	environmental social	0,00043	248	sustainability information	0,00052	155	business partner	0,00047	204
	executive director	0,00035	129	analysis management	0,00038	31	property plant	0,00043	100	gri standard	0,00049	106	health safety	0,00045	375
	carbon dioxide	0,00035	58	data protection	0,00037	173	greenhouse gas	0,00043	323	environmental protection	0,00047	291	combined management	0,00040	425
	health safety	0,00035	230	sustainable development	0,00037	175	code conduct	0,00041	233	see page	0,00047	175	climate change	0,00040	517
	metric ton	0,00034	145	corporate governance	0,00036	402	management approach	0,00041	386	executive board	0,00045	597	environmental protection	0,00038	301
	business activity	0,00034	139	climate protection	0,00036	214	additional information	0,00039	188	statement management	0,00045	101	social enterprise	0,00038	23
	corporate responsibility	0,00034	162	environmental protection	0,00036	239	power plant	0,00037	188	management consolidated	0,00045	253	additional information	0,00037	336
	risk opportunity	0,00034	113	gri standard	0,00033	131	board management	0,00037	261	management corporate	0,00045	219	management consolidated	0,00037	346
	climate protection	0,00033	118	fuel consumption	0,00033	135	exchange rate	0,00037	72	metric ton	0,00044	229	sustainability strategy	0,00037	321
	executive board	0,00032	260	corporate responsibility	0,00030	187	sustainable development	0,00036	197	shareholder management	0,00043	293	sustainable development	0,00037	200
	gas emission	0,00032	153	environmental impact	0,00030	128	plant equipment	0,00036	89	board management	0,00043	330	climate risk	0,00036	182
	product service	0,00030	176	co2 emission	0,00029	360	profit loss	0,00036	64	gri management	0,00043	108	data protection	0,00034	188
	product safety	0,00029	84	occupational health	0,00029	88	target measure	0,00035	56	business activity	0,00042	207	united nation	0,00033	128
	sport car	0,00029	49	ebit special	0,00029	70	economic environmental	0,00034	170	research development	0,00041	177	healthcare innovation	0,00033	64
	materiality analysis	0,00029	118	research development	0,00028	170	co2 emission	0,00033	247	climate change	0,00041	264	compliance management	0,00031	125
	environmental management	0,00027	120	development goal	0,00027	87	air quality	0,00033	105	due diligence	0,00040	262	greenhouse gas	0,00031	463
	north america	0,00027	119	oil gas	0,00027	297	gas emission	0,00033	197	responsible business	0,00040	118	environmental social	0,00031	206

TABLE 10. Top 20 bigrams included in GRI 300 (2017-2021).

Bigrams	All	2017	2018	2019	2020	2021
supply chain	37	8	8	8	5	8
raw material	35	8	8	8	3	8
consolidated financial	33	7	8	8	2	8
financial statement	31	5	8	8	2	8
management system	31	8	6	8	2	7
human right	30	8	7	8	3	4
environmental protection	26	7	7	8	1	3
corporate governance	24	4	8	8	1	3
co2 emission	23	8	7	4	1	3
greenhouse gas	23	5	4	7	1	6
value chain	23	5	2	6	2	8
executive board	21	5	4	4	2	6
climate protection	20	2	5	8	0	5
health safety	20	6	4	5	1	4
supervisory board	19	3	6	4	1	5
risk management	18	4	4	6	0	4
climate change	17	5	1	2	1	8
code conduct	17	6	4	4	1	2
board management	16	0	5	5	1	5
management approach	15	4	3	4	1	3

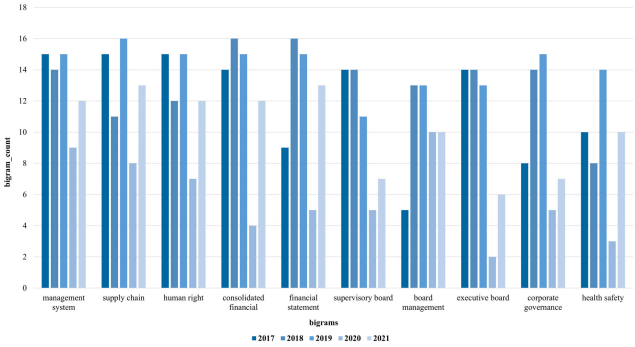

FIGURE 9. Occurrence of top 10 bigrams represented in GRI 400 (2017-2021).

goal was to identify common topics reported and sector bias with industrial firms and consumer discretionary and consumer staples. Nakagawa et al. [33] analyzed non-financial information in integrated reports by using NLP, LDA and Word2Vec to identify the differences between good and less good integrated reports. Niveditha et al. [39] applied NLP as a technique towards CSR research. By combining LDA to identify the topics of primary focus in the reports and a classifier for labeling the text data, they created a supervised machine learning model and applied it on annual reports of various Indian companies to extract dominant topics and thus the most common CSR concerns and practices.

B. MOTIVATION AND LITERATURE GAP

From an economic point of view, the analysis of unstructured textual sections of financial disclosure using text mining applications such as LDA has evident benefits. Financial analysts can extend conventional methods based on numerical information for analyzing the performance of the organization by analyzing textual data in annual reports, which was inaccessible information for a huge number of novice investors [23]. With blurring lines between financial and non-financial disclosure, stricter GRI reporting guidelines and dependencies between CSR performance, CSR reporting and business performance, the textual analysis of sustainability reports attracting more and more attention.

TABLE 11. GRI 400 - Weight and word tount of topic 0 (2017-2021).

	2017			2018			2019			2020			2021		
	Number of bigrams:	1519		Number of bigrams:	1675		Number of bigrams:	1878		Number of bigrams:	2077		Number of bigrams:	2417	
	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count	Word	Weights	Word Count
Topic: 0	**executive board**	0,00180	658	**human right**	0,00485	670	**supervisory board**	0,00237	1209	**combined non-financial**	0,00431	56	**supervisory board**	0,00928	1929
	supervisory board	0,00127	853	**supply chain**	0,00210	467	**human right**	0,00116	972	**assurance engagement**	0,00397	106	**executive board**	0,00570	1093
	climate change	0,00072	157	**united state**	0,00108	178	**supply chain**	0,00110	638	**limited assurance**	0,00209	75	**board member**	0,00280	436
	corporate governance	0,00060	277	**fair value**	0,00095	284	**corporate governance**	0,00109	648	**quality control**	0,00142	17	**financial statement**	0,00210	1079
	board member	0,00060	275	**financial statement**	0,00093	805	**member supervisory**	0,00103	156	**assurance procedure**	0,00121	27	**consolidated financial**	0,00136	716
	power plant	0,00059	169	consolidated financial	0,00087	576	consolidated financial	0,00099	679	highest governance	0,00118	24	board management	0,00125	852
	consolidated financial	0,00058	344	raw material	0,00069	287	raw material	0,00067	329	governance body	0,00118	42	member supervisory	0,00122	202
	corporate responsibility	0,00057	167	code conduct	0,00058	293	board management	0,00059	400	reporting period	0,00116	141	general meeting	0,00104	170
	raw material	0,00055	282	per cent	0,00056	324	financial statement	0,00057	603	financial statement	0,00113	1020	annual general	0,00104	170
	climate protection	0,00050	83	cash flow	0,00052	223	code conduct	0,00057	296	environmental social	0,00102	329	member executive	0,00091	204
	code conduct	0,00046	277	global compact	0,00052	198	member board	0,00053	170	stakeholder management	0,00102	13	management financial	0,00090	267
	supply chain	0,00045	427	due diligence	0,00051	110	defined benefit	0,00052	145	preparation combined	0,00102	5	raw material	0,00079	385
	financial statement	0,00041	281	management system	0,00050	377	executive board	0,00051	723	consolidated financial	0,00101	773	member board	0,00078	223
	product service	0,00040	156	respect human	0,00049	49	board since	0,00050	48	economic environmental	0,00094	193	audit committee	0,00074	149
	member executive	0,00037	111	working condition	0,00048	95	board director	0,00046	31	german commercial	0,00091	89	corporate governance	0,00069	615
	management system	0,00035	340	guiding principle	0,00047	68	management consolidated	0,00046	258	sustainability reporting	0,00089	44	product service	0,00063	257
	united state	0,00035	112	corporate governance	0,00045	555	management system	0,00045	566	collective bargaining	0,00086	95	board supervisory	0,00059	152
	environmental social	0,00034	127	business partner	0,00039	190	ebit special	0,00044	40	gri content	0,00085	144	employee representative	0,00058	188
	human right	0,00033	648	action plan	0,00037	63	special item	0,00044	66	bargaining agreement	0,00082	38	additional information	0,00058	354
	greenhouse gas	0,00032	139	united nation	0,00037	116	shareholder management	0,00044	236	material topic	0,00082	113	information annual	0,00057	145
	value chain	0,00030	155	value chain	0,00037	162	child labor	0,00041	77	internal control	0,00080	103	shareholder management	0,00057	142
	financial reporting	0,00028	74	respect system	0,00037	22	pension plan	0,00040	103	separate non-financial	0,00080	17	financial review	0,00057	149
	around world	0,00027	125	right respect	0,00037	24	health safety	0,00040	405	stakeholder group	0,00080	44	shareholder representative	0,00057	85
	annual shareholder	0,00027	164	supervisory board	0,00037	1228	board member	0,00039	312	content index	0,00079	59	review consolidated	0,00057	144
	pension plan	0,00026	80	environmental social	0,00036	237	environmental protection	0,00037	299	sustainability management	0,00079	57	share-based payment	0,00056	62
	internal control	0,00025	90	business human	0,00036	40	global compact	0,00035	202	sustainability strategy	0,00077	193	performance bonus	0,00055	82
	vocational training	0,00025	63	oil gas	0,00035	206	research development	0,00032	191	legal representative	0,00076	13	statement additional	0,00055	150
	share-based payment	0,00025	59	sustainable development	0,00032	181	cash equivalent	0,00032	73	material misstatement	0,00075	11	management management	0,00054	158
	management management	0,00025	118	health safety	0,00032	298	additional information	0,00030	227	data capture	0,00074	11	börse annual	0,00054	177
	business activity	0,00025	131	combined management	0,00032	248	risk management	0,00030	348	representative responsible	0,00073	9	management consolidated	0,00054	305
	social responsibility	0,00025	90	business model	0,00032	179	united nation	0,00030	160	social topic	0,00072	10	management system	0,00053	598
	risk management	0,00025	244	corporate citizenship	0,00032	78	sustainable finance	0,00030	93	audit firm	0,00072	13	statement note	0,00052	256
	health safety	0,00023	311	sustainability management	0,00031	97	cash cash	0,00030	69	315c conjunction	0,00071	11	executive supervisory	0,00051	172
	see page	0,00023	91	corporate responsibility	0,00031	241	risk assessment	0,00030	185	managerial employee	0,00071	18	per cent	0,00051	288
	human resource	0,00023	109	financial liability	0,00031	76	cost capital	0,00028	35	stakeholder dialog	0,00071	10	middle east	0,00049	51

TABLE 12. Top 20 bigrams included in GRI 400 (2017-2021).

Bigrams	All	2017	2018	2019	2020	2021
management system	65	15	14	15	9	12
supply chain	63	15	11	16	8	13
human right	61	15	12	15	7	12
consolidated financial	61	14	16	15	4	12
financial statement	58	9	16	15	5	13
supervisory board	51	14	14	11	5	7
board management	51	5	13	13	10	10
executive board	49	14	14	13	2	6
corporate governance	49	8	14	15	5	7
health safety	45	10	8	14	3	10
code conduct	41	14	10	9	2	6
risk management	37	10	5	10	5	7
raw material	34	8	8	8	1	9
environmental social	31	5	12	10	2	2
environmental protection	27	9	6	8	1	3
business partner	27	7	6	4	2	8
corporate responsibility	26	6	10	10	0	0
data protection	24	3	5	9	1	6
combined management	23	2	10	3	1	7
board member	22	7	6	5	1	3

TABLE 13. Comparison keywords with modeled topics (full match).

GRI 200	GRI 300	GRI 400
climate change	fuel consumption water consumption ghg emissions supply chain	parental leave

TABLE 14. Comparison keywords with modeled topics (partial match).

Partial match	
GRI 300	
co2	emission consumption
material	raw topic
electricity	green
GRI 400	
health	safety occupational
diversity	equity inclusion

TABLE 15. Comparison keywords with modeled topics (synonymous word combinations).

Synonymous word combinations	
GRI 300	
co2	co2 emission co_2 emission carbon emission carbon dioxide
ghg emission	greenhouse gas
environment	environmental impact environmental protection

Corporate non-financial information has become more important. Within the guidelines for CSR reporting the content may vary due to sector or company specific practices. Since it is difficult to manually read and compare multiple reports consisting of hundreds of pages, the need to read unstructured textual data efficiently and objectively is increasing [33]. Despite growing interest in ESG disclosure and early promising research in this respective field as introduced in subsection A. TEXTUAL ANALYSIS AND ESG REPORTING, we noticed a lack of publications keeping up with both the evolving practices and guidelines for ESG disclosure as well as the state-of-the-art text mining and textual analysis techniques. Our search strategy was structured as follows. To search specifically for literature on the topic of sustainability reporting, the databases

Science Direct, Web of Science, IEEE-Xplore and Jstor were examined. These databases were searched for literature using the search strings ''csr'' AND ''text mining'', ''sustainability reporting'' AND ''text mining'', ''csr'' AND ''machine learning'', ''sustainability reporting'' AND ''machine learning'', (''non'' AND ''financial reporting'') AND ''text mining'', and (''non'' AND ''financial reporting'') AND ''machine learning''. The results of the literature search are presented in Table 2.

Based on this literature research, it appears that little literature is available in the area of sustainability reporting, and its analysis using text mining techniques, which consequently highlights the literature gap in this topic. While topic modeling is a well-established technique for textual data analysis, the application of topic modeling to ESG reports is a relatively new area of research. There are only very limited scientific publications that have used topic modeling approach to extract topics from ESG reports and compare them to GRI reporting requirements. Regarding CSR reports of German DAX40 companies from 2017 till 2021, we built a unique corpus of up-to-date ESG disclosure, which include changes in GRI guidelines as well as the listing of respective companies. By applying LDA Topic Modelling algorithm, building on existing promising research on CSR disclosure while addressing the lack of up-to-date publications combining recent data and state-of-the-art machine learning algorithms.

IV. METHODOLOGY

The following subsections contain a detailed description of our research approach and the methodology used to apply a textual analysis.

A. DATA COLLECTION AND TEXTUAL ANALYSIS PROCEDURE

The sustainability reports of the DAX 40 companies in the period from 2017 to 2021 serve as the data basis for the following analysis. An overview of the DAX 40 companies with a year of admission to the DAX is provided in Table 16 in the appendix. The period was chosen from 2017, as the reporting obligation for large capital market-oriented companies came into force from this year. The law stipulates that the companies either add their reports to the management report or publish them on their website. The reports are thus publicly accessible and can be downloaded from the companies' websites. Only a few reports have been published for the reporting year 2022. An overview of the reports used with sources is shown in Tables 17 to 20 in the appendix.

Our aim is to analyze the subtopics included in corporate reports targeting predefined GRI Topics. Figure 2 visualizes our approach as follows:

First we collect textual data in form of CSR reports of German corporations.

TABLE 16. Overview of companies included in DAX 40 [49].

company/group	year of recording
Adidas	1998
Airbus	2021
Allianz	1988
BASF	1988
Bayer	1988
Beiersdorf	2022
BMW	1988
Brenntag	2021
Continental	2012
Covestro	2018
Daimler Truck	2022
Deutsche Bank	1988
Deutsche Börse	2002
Deutsche Post	2001
Deutsche Telekom	1996
E.ON	2000
Fresenius	2009
Fresinus Medical Care	1999
Hannover Rück	2022
HeidelbergCement	2010
Henkel vz.	1988
Infineon	2009
Linde	1988
Mercedes-Benz Gruppe (Daimler)	1998
Merck	2007
MTU Aero Engines	2019
Müncher Rückversicherungs-Gesellschaft	1996
Porsche AG	2022
Porsche SE	2021
QIAGEN	2021
RWE	1988
SAP	1995
Sartorius vz.	2021
Siemens	1988
Siemens Energy	2022
Siemens Heathineers	2021
Symrise	2021
Volkswagen	1988
Vonovia	2015
Zalando	2021

Second the textual data we extract parts of the text related to a specific GRI Topic (201-207, 301-308, and 401-418) and accumulate them into individual data sets for every respective year. The keywords used to segregate the text passages are listed in Table 21. Those keywords were derived from the GRI Topics and subtopics as depicted in Table 22 in the appendix.

Third we pre-processed the text in several sequential steps as described in subsection B. PRE-PROCESSING and finally applied LDA Topic Modeling on the resulting corpora. In doing so we want to identify patterns, content tendencies and new information within the disclosure given by corporations on predefined GRI Topics. The model identifies influential words by measuring their co-

TABLE 17. Internet source of the sustainability report [Adidas-Covestro].

DAX company/group	year	Internet source of the sustainability report
Adidas	2017	https://report.adidas-group.com/2017/index.html#25
	2018	https://report.adidas-group.com/2018/#downloadcenter
	2019	https://report.adidas-group.com/2019/en/servicepages/downloads.html
	2020	https://report.adidas-group.com/2020/en/servicepages/downloads.html
	2021	https://report.adidas-group.com/2020/en/servicepages/downloads.html
Airbus	2017	not found
	2018	not found
	2019	not found
	2020	not found
	2021	https://www.airbus.com/en/sustainability/reporting-and-performance-data/our-approach-to-sustainability-reporting
Allianz	2017	https://www.allianz.com/en/press/news/commitment/community/180412-allianz-sustainability-report-2017.html
	2018	not found
	2019	not found
	2020	not found
	2021	https://www.allianz.com/en/sustainability/publications/sustainability-report.html
BASF	2017	https://report.basf.com/2017/en/servicepages/downloads.html
	2018	https://report.basf.com/2018/en/servicepages/downloads.html
	2019	https://report.basf.com/2019/en/servicepages/downloads.html
	2020	https://report.basf.com/2020/en/servicepages/downloads.html
	2021	https://report.basf.com/2021/en/services/downloads.html
Bayer	2017	not found
	2018	https://www.bayer.com/de/nachhaltigkeit/nachhaltigkeitsberichte
	2019	https://www.bayer.com/de/nachhaltigkeit/nachhaltigkeitsberichte
	2020	https://www.bayer.com/de/nachhaltigkeit/nachhaltigkeitsberichte
	2021	https://www.bayer.com/de/nachhaltigkeit/nachhaltigkeitsberichte
Beiersdorf	2017	https://www.beiersdorf.com/sustainability/reporting/downloads
	2018	https://www.beiersdorf.com/sustainability/reporting/downloads
	2019	https://www.beiersdorf.com/sustainability/reporting/downloads
	2020	https://www.beiersdorf.com/sustainability/reporting/downloads
	2021	https://www.beiersdorf.com/sustainability/reporting/downloads
BMW	2017	https://www.bmwgroup.com/de/search.html?q=sustainable%20value%20report
	2018	https://www.bmwgroup.com/de/search.html?q=sustainable%20value%20report
	2019	https://www.bmwgroup.com/de/search.html?q=sustainable%20value%20report
	2020	not found
	2021	https://www.bmwgroup.com/en/download-centre.html?area=investor
Brenntag	2017	https://corporate.brenntag.com/en/sustainability/downloads-and-contact/
	2018	https://corporate.brenntag.com/en/sustainability/downloads-and-contact/
	2019	https://corporate.brenntag.com/en/sustainability/downloads-and-contact/
	2020	https://corporate.brenntag.com/en/sustainability/downloads-and-contact/
	2021	https://corporate.brenntag.com/en/sustainability/downloads-and-contact/
Continental	2017	https://www.continental.com/de/nachhaltigkeit/downloads
	2018	https://www.continental.com/de/nachhaltigkeit/downloads
	2019	https://www.continental.com/de/nachhaltigkeit/downloads
	2020	https://www.continental.com/de/nachhaltigkeit/downloads
	2021	https://www.continental.com/de/nachhaltigkeit/nachhaltige-unternehmensfuehrung/nachhaltigkeitsberichte/
Covestro	2017	https://www.covestro.com/de/investors/reports-and-presentations/
	2018	https://report.covestro.com/annual-report-2018/servicepages/downloads.html
	2019	https://bericht.covestro.com/geschaeftsbericht-2019/serviceseiten/downloads.html
	2020	https://report.covestro.com/annual-report-2020/servicepages/downloads.html
	2021	https://report.covestro.com/annual-report-2021/servicepages/downloads.html

occurrence [40]. By applying it on pre-selected text related to a specific GRI Topic, the resulting clusters represent the main subtopics within a main-topic and reveal interesting insights in reporting practices and tendencies. In subsection C. LATENT DIRICHLET ALLOCATION we provide a brief introduction to LDA and describe our Topic Modeling approach.

B. PRE-PROCESSING

This section describes the procedure of collecting the data and all the pre-processing steps carried out, necessary for further data analysis. Beginning with the handling of the downloaded CSR reports to the point of applying text mining methods on the prepared data set. In a first, manual step, we examined the structure of the CSR reports. This provided an overview of the technical requirements for further processing. We then saved the reports of the respective companies as a PDF file in English language. To the best of our knowledge, there is no uniform database available at the current time through which sustainability reports are made available. Therefore, the reports were downloaded from the respective homepage. To be able to perform subsequent processing steps with the Python programming language, the PDF files were converted into txt-files. For this the Python package ''PyMuPDF'' was used, which filtered the PDF files according to their text passages. To be able to analyze the content and the topics included in the reports we conducted textual feature reduction based on multiple raw pre-processing steps. These steps are performed to detect, identify, and remove unwanted elements in the data set [41].

Thus, the text contents of the reports were further filtered by regular expressions in a second step. We removed non-language elements, non-alphanumeric symbols, num-

TABLE 18. Internet source of the sustainability report [Daimler Truck-HeidelbergCement].

DAX company/group	year	Internet source of the sustainability report
Daimler Truck	2017	not found
	2018	https://www.daimlertruck.com/en/sustainability/reportings
	2019	https://www.daimlertruck.com/en/sustainability/reportings
	2020	https://www.daimlertruck.com/en/sustainability/reportings
	2021	https://www.daimlertruck.com/en/sustainability/reportings
Deutsche Bank	2017	https://www.db.com/what-we-do/responsibility/reports/reports?language_id=1
	2018	https://www.db.com/what-we-do/responsibility/reports/reports?language_id=2
	2019	https://www.db.com/what-we-do/responsibility/reports/reports?language_id=3
	2020	https://www.db.com/what-we-do/responsibility/reports/reports?language_id=4
	2021	https://www.db.com/what-we-do/responsibility/reports/reports?language_id=5
Deutsche Börse	2017	https://www.deutsche-boerse.com/dbg-en/investor-relations/financial-reports/annual-reports/archive
	2018	https://www.deutsche-boerse.com/dbg-en/investor-relations/financial-reports/annual-reports/archive
	2019	https://www.deutsche-boerse.com/dbg-en/investor-relations/financial-reports/annual-reports/archive
	2020	https://www.deutsche-boerse.com/dbg-en/investor-relations/financial-reports/annual-reports/archive
	2021	https://www.deutsche-boerse.com/dbg-de/investor-relations/finanzberichte/geschaeftsberichte/geschaeftsbericht-2021
Deutsche Post	2017	https://www.dpdhl.com/de/nachhaltigkeit/unser-ansatz/nachhaltigkeitsberichte.html
	2018	https://www.dpdhl.com/de/nachhaltigkeit/unser-ansatz/nachhaltigkeitsberichte.html
	2019	https://www.dpdhl.com/de/nachhaltigkeit/unser-ansatz/nachhaltigkeitsberichte.html
	2020	https://www.dpdhl.com/de/nachhaltigkeit/unser-ansatz/nachhaltigkeitsberichte.html
	2021	https://www.dpdhl.com/de/nachhaltigkeit/unser-ansatz/nachhaltigkeitsberichte.html
Deutsche Telekom	2017	https://www.cr-bericht.telekom.com/2021/download-center
	2018	https://www.cr-bericht.telekom.com/2021/download-center
	2019	https://www.cr-bericht.telekom.com/2021/download-center
	2020	https://www.cr-bericht.telekom.com/2021/download-center
	2021	https://www.cr-bericht.telekom.com/2021/download-center
E.ON	2017	https://www.eon.com/de/ueber-uns/nachhaltigkeit/nachhaltigkeitsbericht.html
	2018	https://www.eon.com/de/ueber-uns/nachhaltigkeit/nachhaltigkeitsbericht.html
	2019	https://www.eon.com/de/ueber-uns/nachhaltigkeit/nachhaltigkeitsbericht.html
	2020	https://www.eon.com/de/ueber-uns/nachhaltigkeit/nachhaltigkeitsbericht.html
	2021	https://www.eon.com/de/ueber-uns/nachhaltigkeit/nachhaltigkeitsbericht.html
Fresenius	2017	not found
	2018	not found
	2019	https://annualreport.fresenius.com/2019/downloads/
	2020	https://annualreport.fresenius.com/2020/downloads/
	2021	https://annualreport.fresenius.com/2021/downloads/
Fresinus Medical Care	2017	https://www.freseniusmedicalcare.com/en/agm/archive
	2018	https://www.freseniusmedicalcare.com/en/agm/archive
	2019	https://www.freseniusmedicalcare.com/en/agm/archive
	2020	not found
	2021	https://www.freseniusmedicalcare.com/en/sustainability
Hannover Rück	2017	not found
	2018	not found
	2019	not found
	2020	https://www.hannover-rueck.de/1766617/nachhaltigkeitsbericht-2020.pdf
	2021	https://www.hannover-re.com/171507/csr-publications
HeidelbergCement	2017	https://www.heidelbergmaterials.com/en/sustainability-reports
	2018	https://www.heidelbergmaterials.com/en/sustainability-reports
	2019	https://www.heidelbergmaterials.com/en/sustainability-reports
	2020	https://www.heidelbergmaterials.com/en/sustainability-reports
	2021	https://www.heidelbergmaterials.com/en/sustainability-reports

bers, dots, repeat space character, one letter words, letters attached with special characters and converted the letters into lower case. Because text cannot be processed by computer programs in its raw form, text indexing is carried out to convert the texts into a list of words. This step is called *Tokenization* because it segments texts into tokens by punctuation or white space marks [42]. Important was to keep words such as o2 and co2 included in the text, for them being expected to have explanatory content. For *Tokenization* the package "*tokenize*" from nltk was used. To obtain meaningful results, the text was divided into bigrams. Thereby a list is generated from the combinations of the predecessor and descendant word.

The subsequent step of *Stemming/Lemmatization* is the process of mapping each token to its root form, which is usually applicable to nouns, verbs, and adjectives [42]. The implementation requires stemming rules, which are rules for associating tokens with their own root form. Nouns for example, are converted into their singular form or verbs given in their third-person singular form need to have the postfix "s" or "es" removed. For Stemming we used the Python package "*Snowballer*" from nltk.

Also, we have conducted *stopword* removal as recommended for long data sets [43]. Stopwords are words that are unrelated to the content of the text and should be eliminated to improve efficiency. Any words that are already on the pre-defined list are eliminated. Words without relevance for the content of the Text are for example prepositions, such as "in," "on," and "to", conjunctions such as "and," "or," "but," and "however" or definite article "the," and the indefinite articles, "a" and "an". Apart from the list for of stopwords for the English language we included months, company names, various country names etc. To increase the

TABLE 19. Internet source of the sustainability report [Henkel vz.-QIAGEN].

DAX company/group	year	Internet source of the sustainability report
Henkel vz.	2017	https://www.henkel.de/nachhaltigkeit/nachhaltigkeitsbericht
	2018	https://www.henkel.de/nachhaltigkeit/nachhaltigkeitsbericht
	2019	https://www.henkel.de/nachhaltigkeit/nachhaltigkeitsbericht
	2020	https://www.henkel.de/nachhaltigkeit/nachhaltigkeitsbericht
	2021	https://www.henkel.de/nachhaltigkeit/nachhaltigkeitsbericht
Infineon	2017	https://www.infineon.com/cms/en/about-infineon/sustainability/csr-reporting/
	2018	https://www.infineon.com/cms/en/about-infineon/sustainability/csr-reporting/
	2019	https://www.infineon.com/cms/en/about-infineon/sustainability/csr-reporting/
	2020	https://www.infineon.com/cms/en/about-infineon/sustainability/csr-reporting/
	2021	https://www.infineon.com/cms/en/about-infineon/sustainability/csr-reporting/
Linde	2017	https://www.linde.com/sustainable-development/reporting-center
	2018	https://www.linde.com/sustainable-development/reporting-center
	2019	https://www.linde.com/sustainable-development/reporting-center
	2020	https://www.linde.com/sustainable-development/reporting-center
	2021	https://www.linde.com/sustainable-development
Mercedes-Benz Gruppe (Daimler)	2017	https://group.mercedes-benz.com/sustainability/archive-sustainability-reports.html
	2018	https://group.mercedes-benz.com/sustainability/archive-sustainability-reports.html
	2019	https://sustainabilityreport.daimler.com/2019/servicepages/downloads.html
	2020	https://sustainabilityreport.daimler.com/2020/servicepages/downloads.html
	2021	https://sustainabilityreport.mercedes-benz.com/2021/servicepages/downloads.html
Merck	2017	https://merck.online-report.eu/2017/cr-report/servicepages/downloads.html
	2018	https://www.merckgroup.com/en/cr-report/2018/servicepages/downloads.html
	2019	https://www.merckgroup.com/en/cr-report/2019/servicepages/downloads.html
	2020	https://www.merckgroup.com/en/sustainability-report/2020/servicepages/downloads.html
	2021	https://www.merckgroup.com/en/sustainability-report/2021/services/downloads.html
MTU Aero Engines	2017	https://sustainability.mtu.de/en/downloads/
	2018	https://sustainability.mtu.de/en/downloads/
	2019	https://sustainability.mtu.de/en/downloads/
	2020	https://sustainability.mtu.de/en/downloads/
	2021	https://sustainability.mtu.de/en/downloads/
Müncher Rückversicherungs-Gesellschaft	2017	https://www.munichre.com/en/company/sustainability/download-center.html
	2018	https://www.munichre.com/en/company/sustainability/download-center.html
	2019	https://www.munichre.com/en/company/sustainability/download-center.html
	2020	https://www.munichre.com/en/company/sustainability/download-center.html
	2021	https://www.munichre.com/de/unternehmen/sustainability.html
Porsche AG	2017	https://newsroom.porsche.com/de/unternehmen/porsche-geschaefts-und-nachhaltigkeitsbericht-2021/download-center.html
	2018	https://newsroom.porsche.com/de/unternehmen/porsche-geschaefts-und-nachhaltigkeitsbericht-2021/download-center.html
	2019	https://newsroom.porsche.com/de/unternehmen/porsche-geschaefts-und-nachhaltigkeitsbericht-2021/download-center.html
	2020	https://newsroom.porsche.com/de/unternehmen/porsche-geschaefts-und-nachhaltigkeitsbericht-2021/download-center.html
	2021	https://newsroom.porsche.com/de/unternehmen/porsche-geschaefts-und-nachhaltigkeitsbericht-2021/download-center.html
Porsche SE	2017	not found
	2018	not found
	2019	https://www.porsche-se.com/en/company/corporate-governance
	2020	https://www.porsche-se.com/en/company/corporate-governance
	2021	https://www.porsche-se.com/en/company/corporate-governance
QIAGEN	2017	not found
	2018	not found
	2019	not found
	2020	not found
	2021	not found

result, the stemming step was also applied to the stopword list. Countries with precarious labor conditions or environmental issues are not removed because they could be essential for understanding some topics. In an iterative process common but non-recognizable letters and words are also removed.

In a final step, to align the topics with the GRI standards, the existing text corpus was filtered on the basis of the existing GRI subtopics and divided into three parts. For each of the GRI Topics 200, 300 and 400, a separate text corpus was created to which the LDA model was applied. Therefore, the sustainability reports were filtered based on the GRI Topics. Thus, only those text passages were included in the corpus that could be related to the respective GRI standard. This procedure served as a basis for comparing the identified topics with the standard. In a further step, this procedure was additionally applied on an annual basis to enable the development over time.

C. LATENT DIRICHLET ALLOCATION

In this section, we provide a brief introduction to LDA and describe our Topic Modeling approach, with which we want to gain insights on main keywords and clusters within the analyzed CSR reports. Generally, text represents unstructured data which consists of strings that are called words. Text mining can be defined as the process of distilling actionable insights from text [44]. Topic Modeling is ideal for organizing, summarizing, and understanding vast amount of textual data by providing a probabilistic framework. Topic Modeling captures the latent semantic structure of a document collection. This is accomplished by expressing documents as distributions over a topic and topic rankings

TABLE 20. Internet source of the sustainability report [RWE-Zalando].

DAX company/group	year	Internet source of the sustainability report
RWE	2017	https://www.rwe.com/en/responsibility-and-sustainability/cr-report-archive/
	2018	https://www.rwe.com/en/responsibility-and-sustainability/cr-report-archive/
	2019	https://www.rwe.com/en/responsibility-and-sustainability/cr-report-archive/
	2020	https://www.rwe.com/en/responsibility-and-sustainability/cr-report-archive/
	2021	https://www.rwe.com/en/responsibility-and-sustainability/cr-report-archive/
SAP	2017	https://www.sap.com/investors/en/reports.html?sort=latest_desc&tab=reports&tag=investor-relations:financial-reports/integrated-report
	2018	https://www.sap.com/investors/en/reports.html?sort=latest_desc&tab=reports&tag=investor-relations:financial-reports/integrated-report
	2019	https://www.sap.com/integrated-reports/2019/en/environmental-performance.html
	2020	https://www.sap.com/integrated-reports/2020/en.html
	2021	https://www.sap.com/integrated-reports/2021/en.html
Sartorius vz.	2017	not found
	2018	not found
	2019	not found
	2020	https://www.sartorius.com/download/722398/sag-gri-bericht-2020-data.pdf
	2021	https://www.sartorius.com/download/1273406/sag-gri-bericht-de-2021-pdf-data.pdf
Siemens	2017	https://new.siemens.com/global/en/company/sustainability/sustainability-figures.html
	2018	https://new.siemens.com/global/en/company/sustainability/sustainability-figures.html
	2019	https://new.siemens.com/global/en/company/sustainability/sustainability-figures.html
	2020	https://new.siemens.com/global/en/company/sustainability/sustainability-figures.html
	2021	not found
Siemens Energy	2017	not found
	2018	not found
	2019	not found
	2020	https://www.siemens-energy.com/global/en/company/sustainability.html?gclid=EAIaIQobChMIxZuj-Pj-_AIVxeN3Ch0_ygJTEAAYAiAAEgJw3fD_BwE
	2021	https://www.siemens-energy.com/global/en/company/sustainability.html?gclid=EAIaIQobChMIxZuj-Pj-_AIVxeN3Ch0_ygJTEAAYAiAAEgJw3fD_BwE
Siemens Heathineers	2017	not found
	2018	not found
	2019	not found
	2020	not found
	2021	https://www.siemens-healthineers.com/deu/investor-relations/presentations-financial-publications
Symrise	2017	https://www.symrise.com/sustainability/reports-policies-standards-audits/#our-corporate-reports
	2018	https://www.symrise.com/sustainability/reports-policies-standards-audits/#our-corporate-reports
	2019	https://www.symrise.com/sustainability/reports-policies-standards-audits/#our-corporate-reports
	2020	https://www.symrise.com/sustainability/reports-policies-standards-audits/#our-corporate-reports
	2021	https://www.symrise.com/sustainability/reports-policies-standards-audits/#our-corporate-reports
Volkswagen	2017	https://www.volkswagenag.com/en/sustainability/reporting-and-esg-performance/sustainability-report.html
	2018	https://www.volkswagenag.com/en/sustainability/reporting-and-esg-performance/sustainability-report.html
	2019	https://www.volkswagenag.com/en/sustainability/reporting-and-esg-performance/sustainability-report.html
	2020	https://www.volkswagenag.com/en/sustainability/reporting-and-esg-performance/sustainability-report.html
	2021	https://www.volkswagenag.com/en/sustainability/reporting-and-esg-performance/sustainability-report.html
Vonovia	2017	https://investoren.vonovia.de/en/news-and-publications/reports-publications/
	2018	https://reports.vonovia.de/2018/sustainability-report/servicepages/downloads.html?lp_campaign=downloads
	2019	https://reports.vonovia.de/2019/sustainability-report/services/downloads.html
	2020	https://reports.vonovia.de/2020/sustainability-report/services/downloads.html
	2021	https://report.vonovia.de/2021/nachhaltigkeitsbericht/en/downloads/
Zalando	2017	not found
	2018	not found
	2019	not found
	2020	not found
	2021	not found

of words based on their relevance to themes. Hereby variable relations between words and their occurrence in the data set are extracted as topics [45]. LDA is a three-stage Bayesian model first described by Blei et al. [46]. It is used to differentiate groups of words as topics within textual data. Each topic is thus modeled as an infinite mixture over a collection of topic probabilities. The topic probabilities give an explicit representation of a document in the context of text modeling [40].

To gain a better understanding of LDA, it is important to clarify the terms word, document, and corpus. According to Blei et al. [46], a word is the fundamental unit of discrete data and an element of a vocabulary indexed by $\{1, \ldots, V\}$. A sequence of N words is adding up to a document W and is defined as $W = (w_1, w_2, \ldots, w_N)$, with w_n as the *nth* word of the sequence [46]. Multiple documents form a corpus D and are defined by $D = \{W_1, W_2, \ldots, W_M\}$ [46]. The variable α represents a Dirichlet prior weight of topic by document and Z the assignment of a word to a given topic. According to the structure as depicted in Figure 3, documents contain a random mixture of topics characterized by the distribution of words [40], [46].

The variable θ describes the extent to which a topic is represented in a document. Therefore, each document has a probability of belonging to each topic [40]. The variable η represents Dirichlet distribution. A high value for η indicates that the topics are likely to cover most of the words, lower value implies that the topics contain a fewer number of words. The parameters α and β are corpus level parameters, assumed to be sampled once in the process of generating the corpus [46]. The Python package "models.ldamodel" from gensim was selected for the application of the LDA model to the sustainability reports. The number of topics to be modelled was selected based on the topics prescribed in

TABLE 21. Keywords derived from each topic.

200	300	400
economic value	materials	employee hires
operating costs	non-renewable materials	employee full-time
employee wage	renewable materials	employee part-time
employee benefits	recycled materials	employee benefits
payments government	materials inputs	parental leave
community investments	reclaimed products	health
climate change	materials packaging	hazard
plan obligations	energy	worker training
pension liabilities	energy consumption	work-related hazards
financial assistance	fuel	work-related injuries
tax relief	fuel consumption	ill health
tax credits	electricity	education
investment grants	heating	career development
royalty holidays	cooling	diversity
wage	energy intensity	gender diversity
senior management	reduction energy	salary woman
infrastructure investments	energy requirements	salary men
economic impacts	water	discrimination
local suppliers	water withdrawn	child labor
corruption	water consumed	labor forced
anti-corruption	water discharged	security training
anti-competitive behavior	effluent discharge	indigenous people
monopoly	water withdrawal	communities
	groundwater	labeling
	seawater	
	surface water	
	water discharge	
	water consumption	
	biodiversity	
	habitat areas	
	species	
	emissions ghg	
	co2	
	energy indirect	
	ghg emissions	
	ozone	
	nitrogen oxides	
	oxides	
	waste generation	
	recycling waste	
	reuse waste	
	landfilling	
	incineration	
	suppliers	
	environment	
	supply chain	
	suppliers impact	

the respective GRI standards. For example, the GRI Topic standard 200 contains a total of 7 subtopics.

V. ANALYSIS AND RESULTS

In the following section, we present our results from the Topic Modeling approach applied on the entirety of text belonging to the GRI Topics 200, 300 and 400, as well as individually for every year from 2017 to 2021.

A. MODEL APPLIED ON ENTIRETY OF TEXT

The model parameters were adapted for each of the three corpora including all reports 2017-2021 belonging to GRI

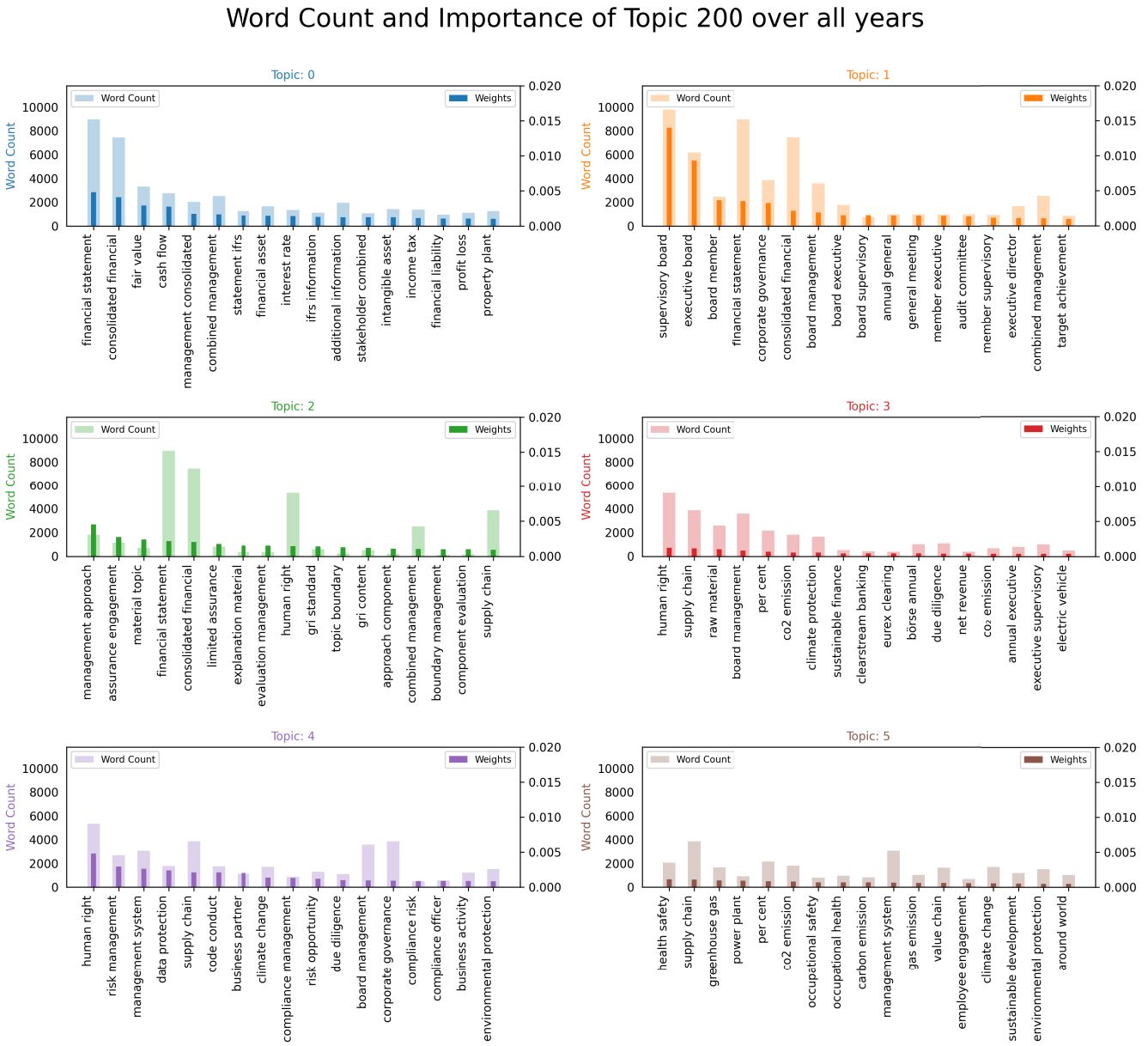

FIGURE 10. GRI 200 - Weight and Word count of topic 0-5 (all reports).

200, GRI 300 and GRI 400. For GRI 200, the model was configured to build 7 topics consisting of 17 words in accordance with GRI 200 topic structure (7 topics and 17 subtopics). For GRI 300, the model was configured to build 8 topics consisting of 36 words in accordance with GRI 300 topic structure (8 topics and 36 subtopics). For GRI 400, the model was configured to build 16 topics consisting of 35 words in accordance with GRI 400 topic structure (16 topics and 35 subtopics). The Number of Bigrams extracted from the input data was for the GRI 200: 14940, GRI 300: 9587, GRI 400: 9858.

Figure 4 displays the 25 most represented bigrams in all topics founded by the model. Those are "supply chain", "board management", "financial statement", "corporate governance", "consolidated financial", "human right", "climate change", "combined management", "health safety", "management system", "climate protection", "environmental social", "per cent", "raw material", "additional information", "due diligence", "environmental protection", "ifrs information", "management consolidated", "value chain", "co2 emission", "code conduct", "executive board", "key figure", and "risk management".

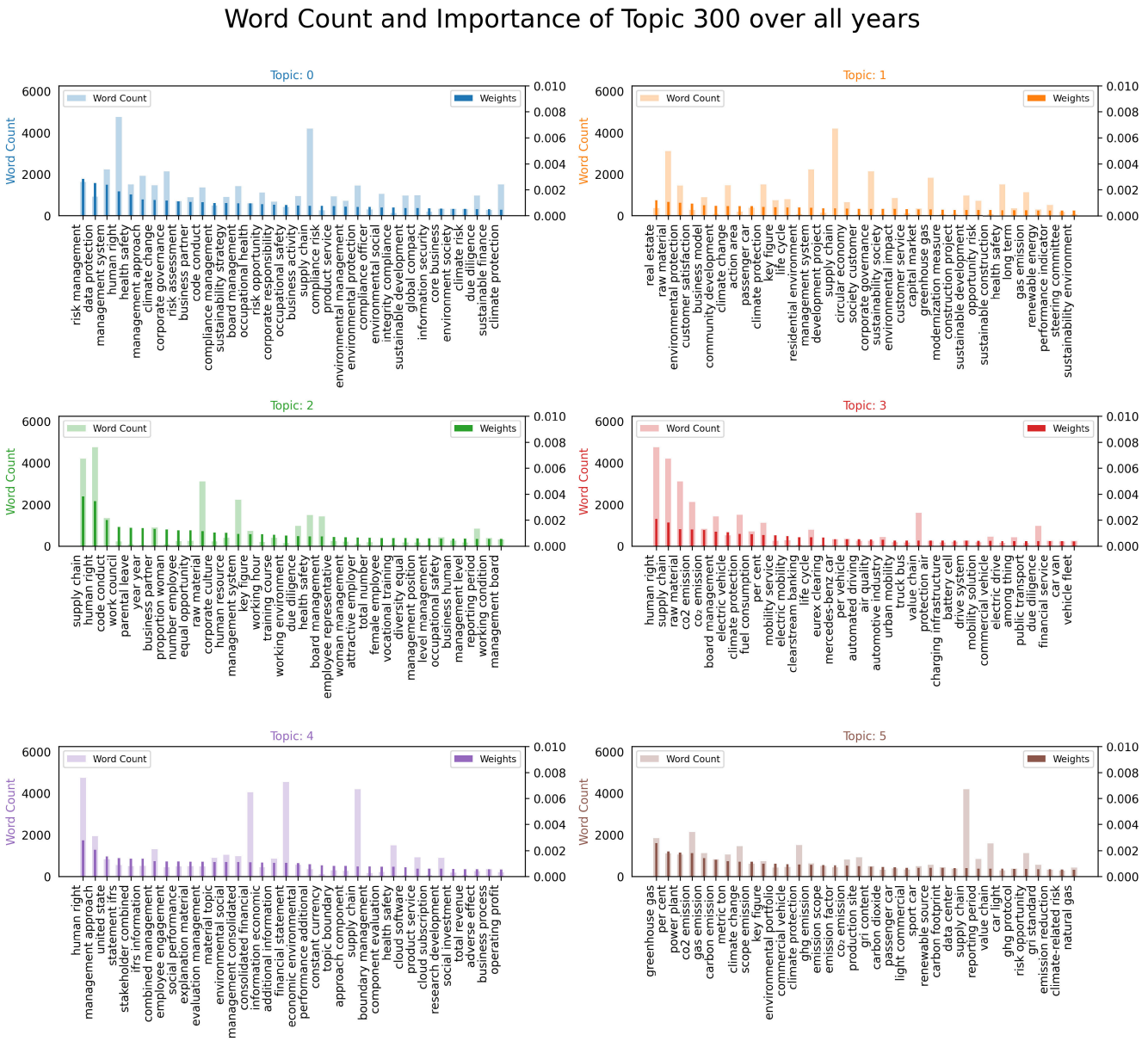

FIGURE 11. GRI 300 - Weight and word count of topic 0-5 (all reports).

Table 3 breaks down this list into their occurrence in the individual GRI Topics 200-400 (all years accumulated).

Figure 5 visualizes the distribution of the top 10 bigrams individually for GRI 200 to GRI 400.

The results of every topic found by the LDA model were combined in a table so that the words or word combinations of a topic are arranged in descending order of importance (weight) and evaluated graphically. The 5 words with the highest weight for the topic are high-lighted in bold letters and the column Word Count is color-coded. High values are shaded with darker blue, lower values with lighter shading. The following Tables and Figures contain the results for 3 of the discovered topics (Topic 0, 1 and 2) for each of the 3 text corpora.

Table 4 contains the LDA results for Topic 0, 1 and 2 included in the GRI 200 corpus. The most important words for Topic 0 are "financial statement", "consolidated financial", "fair value", "cash flow", and "management consolidated". The most important words for Topic 1 are "supervisory board", "executive board", "board member" "financial statement", and "corporate governance". The most important words for Topic 2 are "management approach", "assurance engagement", "material topic", "financial statement", and "consolidated financial". The

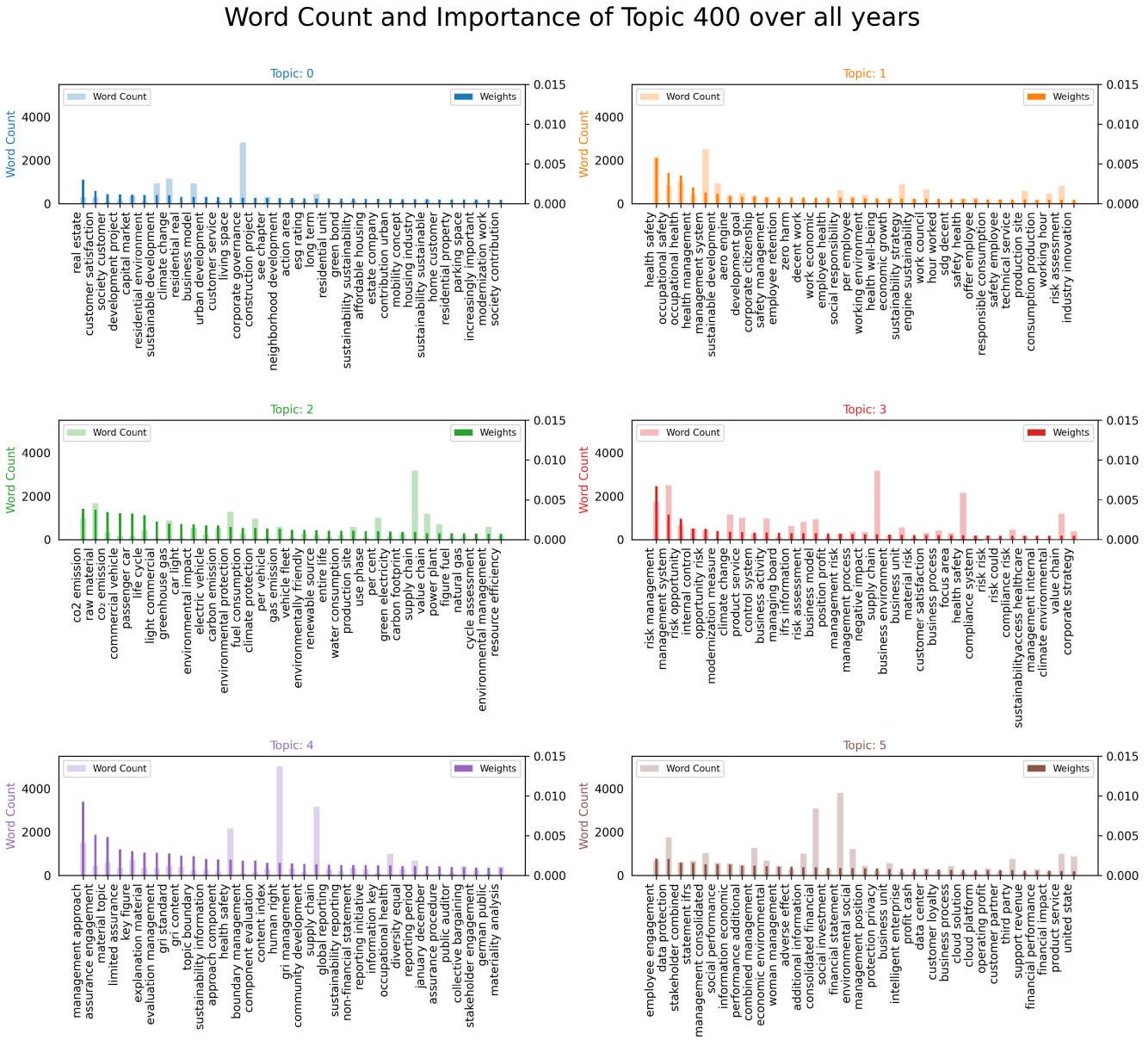

FIGURE 12. GRI 400 - Weight and word count of topic 0-5 (all reports).

charts of the Topics 0-5 found by the model are included in the appendix (Figure 10).

Table 5 contains the LDA results for Topic 0, 1 and 2 included in the GRI 300 corpus. The most important words for Topic 0 are "risk management", "data protection", "management system", "human right", and "health safety". The most important words for Topic 1 are "real estate", "raw material", "environmental protection", "customer satisfaction", and "business model". The most important words for Topic 2 are "supply chain", "human right", "code conduct", "work council", and "parental leave". The charts of the Topics 0-5 found by the model are included in the appendix (Figure 11).

Table 6 contains the LDA results for Topic 0, 1 and 2 included in the GRI 400 corpus. The most important words for Topic 0 are "real estate", "customer satisfaction", "society customer", "development project", and "capital market". The most important words for Topic 1 are "health safety", "occupational safety", "occupational health", "health management", and "management system". The most important words for Topic 2 are "co2 emission", "raw material", "co_2 emission", "commercial vehicle", and "passenger car". The charts of the Topics 0-5 found by the model are included in the appendix (Figure 12).

B. MODEL APPLIED INDIVIDUALLY FOR EVERY YEAR

After evaluating the report sections divided according to their GRI Topic affiliation, the model was adapted to

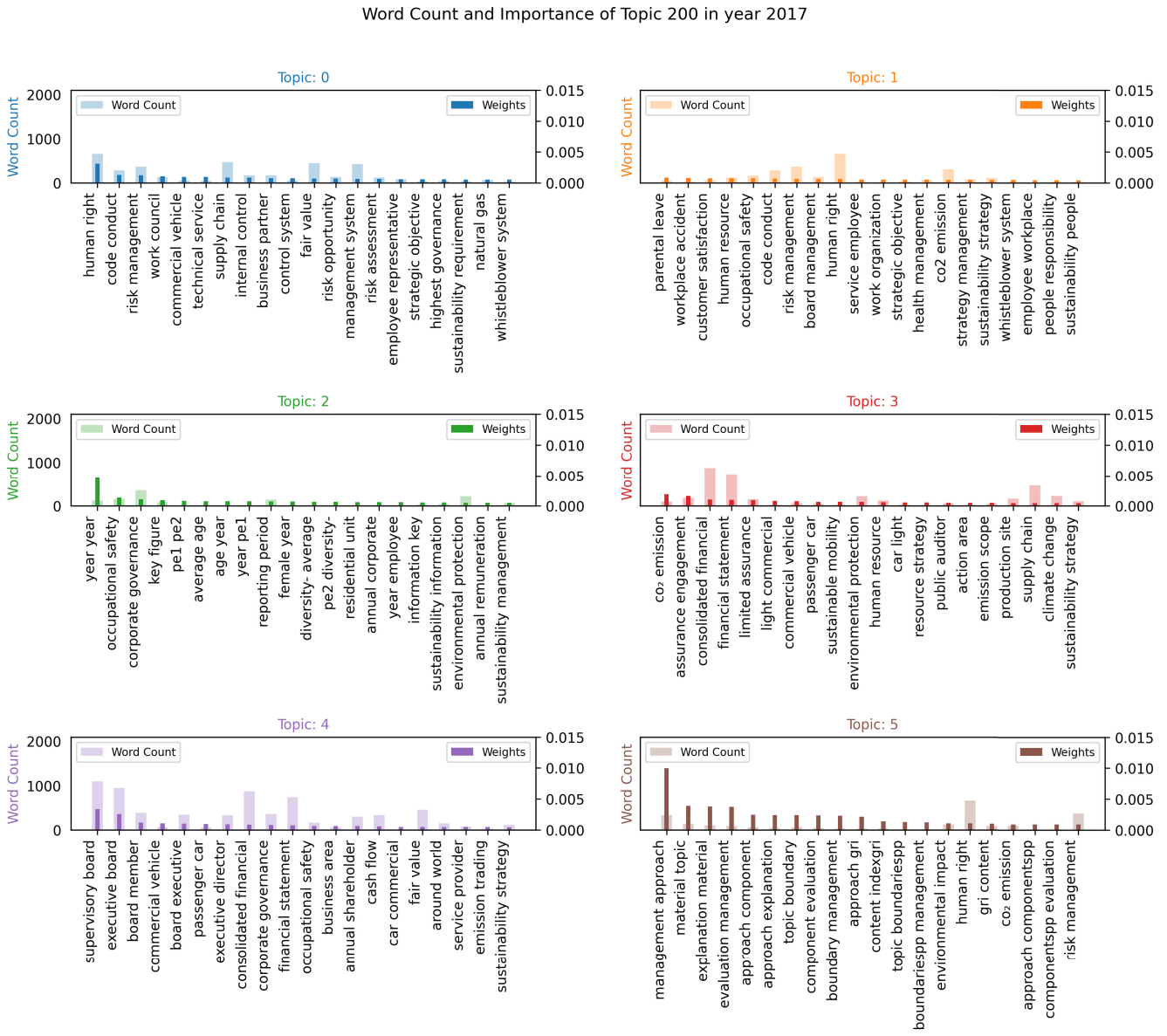

FIGURE 13. GRI 200 - Weight and word count of topic 0-5 2017.

evaluate the data additionally according to individual years. Consequently, the respective input data consists of text belonging to GRI 200, GRI 300, and GRI 400 as well as a specific year between 2017 and 2021. The Number of bigrams extracted by the input data grows year by year for all three GRI Topics. GRI 200 grew by 14.35 per cent on average, GRI 300 by 12.31 per cent and GRI 400 by 10.94 per cent. The absolute and relative development of the number of bigrams per GRI Topic and year is visualized in Figure 6.

The model parameters were adapted for each of the corpora belonging to GRI 200, GRI 300 and GRI 400 and a specific year (2017-2021). For GRI 200, the model was configured to build 7 topics consisting of 17 words for every year. For GRI 300, the model was configured to build 8 topics consisting of 36 words for every year. For GRI 400, the model was configured to build 16 topics consisting of 35 words for every year.

The results were combined in tables so that the words or word combinations of a topic are arranged in descending order of importance (weight) and evaluated graphically. The 5 words with the highest weight for the topic are high-lighted in bold letters and the column Word Count is color-coded. High values are shaded with darker blue, lower values with lighter shading. The following Tables and Figures contain the results of one exemplary topic discovered by the model for

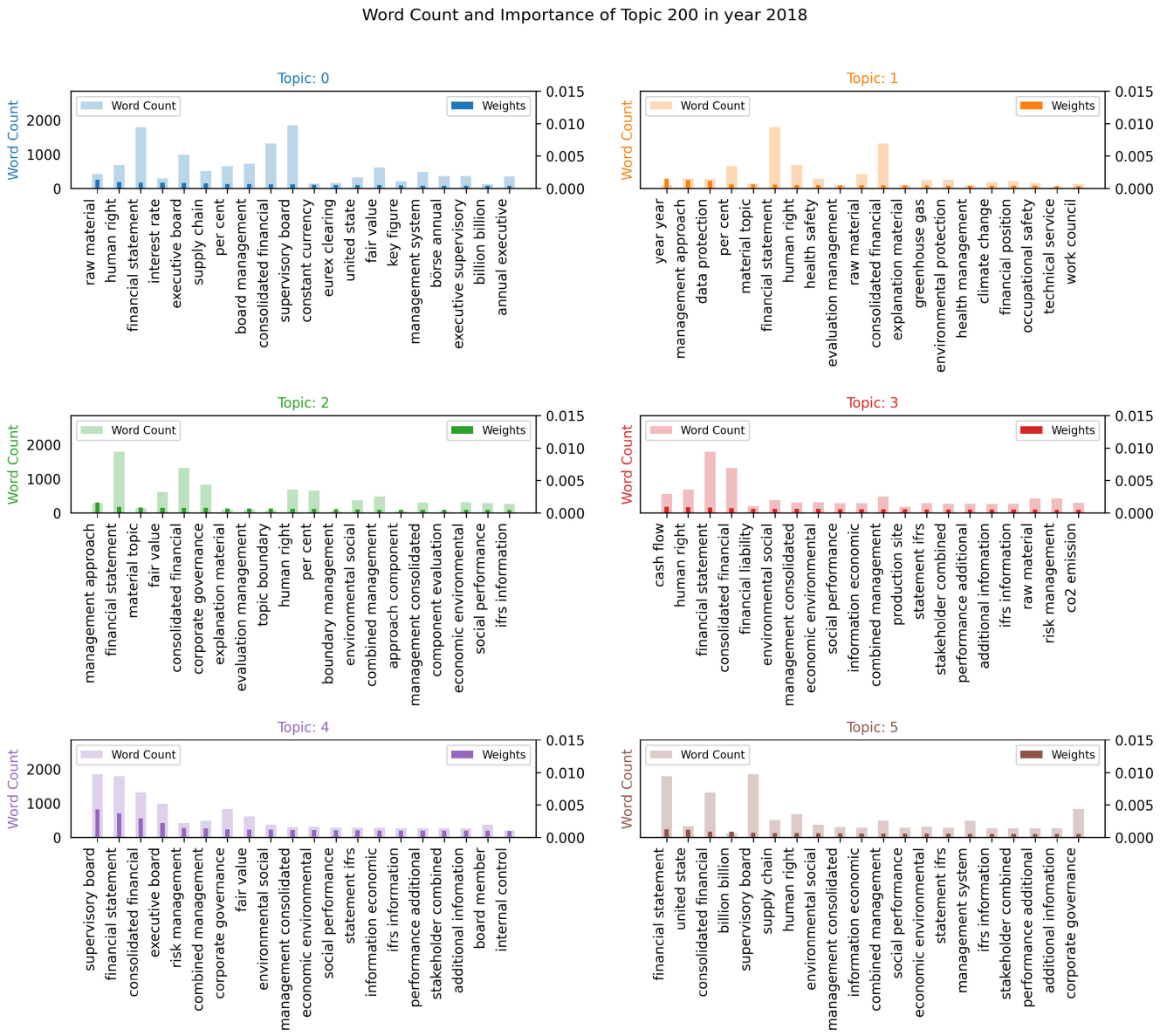

FIGURE 14. GRI 200 - Weight and word count of topic 0-5 2018.

each of the 3 GRI Topics (200, 300 and 400) and year of the respective time span.

Table 7 contains the LDA results for Topic 0 included in the GRI 200 corpus for every year in the period 2017 to 2021. The most important words for Topic 0 in 2017 are "human right", "code conduct", "risk management", "work council", and "commercial vehicle". The most important words for Topic 0 in 2018 are "raw material", "human right", "financial statement", "interest rate", and "executive board". The most important words for Topic 0 in 2019 are "per cent", "supply chain", "climate protection", "environmental portfolio", and "human right". The most important words for Topic 0 in 2020 are "executive board", "supervisory board", "financial statement", "consolidated financial", and "board member". The most important words for Topic 0 in 2021 are "human right", "board management", "per cent", "supply chain", and "supervisory board". The charts of the Topics 0-5 found by the model in 2017 are included in the appendix (Figures 13 to 17).

Table 8 contains the 20 bigrams occurring most frequently over the years in the Topics belonging to GRI 200 formed by the LDA model. Furthermore, it breaks down their occurrence individually for every year between 2017 and 2021. Figure 7 visualizes the information contained in Table 8 for the 10 most frequent bigrams of the topics found in GRI 200 for every year.

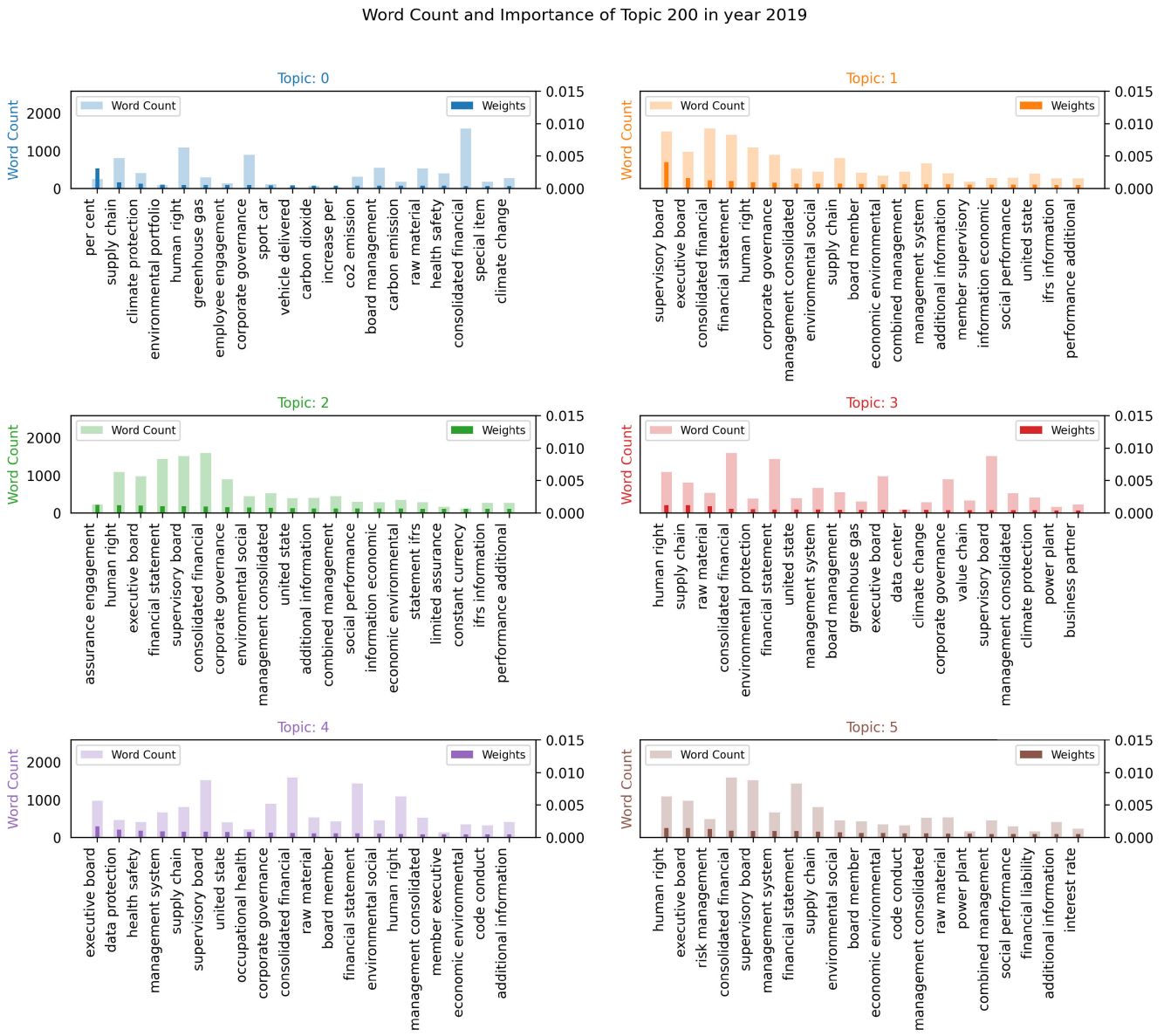

FIGURE 15. GRI 200 - Weight and word count of topic 0-5 2019.

Table 9 contains the LDA results for Topic 0 included in the GRI 300 corpus for every year in the period 2017 to 2021. The most important words for Topic 0 in 2017 are "assurance engagement", "raw material", "co2 emission", "supply chain", and "per cent". The most important words for Topic 0 in 2018 are "management approach", "material topic", "evaluation management", "explanation material", and "supply chain". The most important words for Topic 0 in 2019 are "cash flow", "consolidated financial", "fair value", "raw material", and "financial statement". The most important words for Topic 0 in 2020 are "human right", "financial statement", "consolidated financial", "management approach", and "supply chain". The most important words for Topic 0 in 2021 are "human right", "supply chain", "financial statement", "consolidated financial", and "code conduct". The charts of the Topics 0-5 found by the model in 2017 are included in the appendix (Figures 18 to 22).

Table 10 contains the 20 bigrams occurring most frequently over the years in the Topics belonging to GRI 300 formed by the LDA model. Furthermore, it breaks down their occurrence individually for every year between 2017 and 2021. Figure 8 visualizes the information contained in Table 10 for the 10 most frequent bigrams of the Topics found in GRI 300 for every year.

Table 11 contains the LDA results for Topic 0 included in the GRI 400 corpus for every year in the period 2017 to

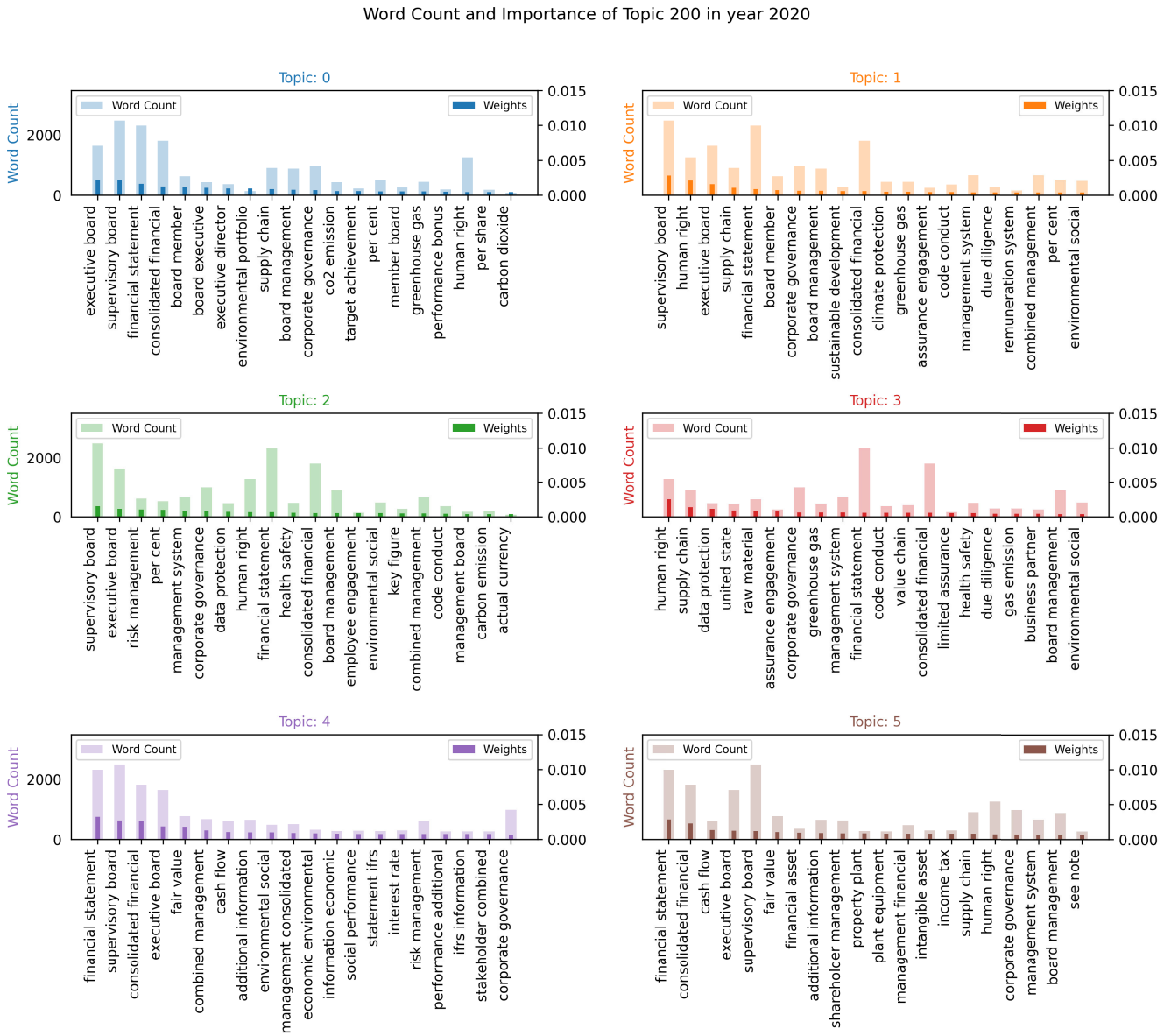

FIGURE 16. GRI 200 - Weight and word count of topic 0-5 2020.

2021. The most important words for Topic 0 in 2017 are "executive board", "supervisory board", "climate change", "corporate governance", and "board member". The most important words for Topic 0 in 2018 are "human right", "supply chain", "united state", "fair value", and "financial statement". The most important words for Topic 0 in 2019 are "supervisory board", "human right", "supply chain", "corporate governance", and "member supervisory". The most important words for Topic 0 in 2020 are "combined non-financial", "assurance engagement", "limited assurance", "quality control", and "assurance procedure". The most important words for Topic 0 in 2021 are "supervisory board", "executive board", "board member", "financial statement", and "consolidated financial". The charts of the Topics 0-5 found by the model in 2017 are included in the appendix (Figures 23 to 27).

Table 12 contains the 20 bigrams occurring most frequently over the years in the Topics belonging to GRI 400 formed by the LDA model. Furthermore, it breaks down their occurrence individually for every year between 2017 and 2021. Figure 9 visualizes the information contained in Table 12 for the 10 most frequent bigrams of the Topics found in GRI 400 for every year.

VI. DISCUSSION AND CONTRIBUTIONS

The first part of this study deals with the possibility of identifying topics in sustainability reports using an LDA model. To identify topics in sustainability reports,

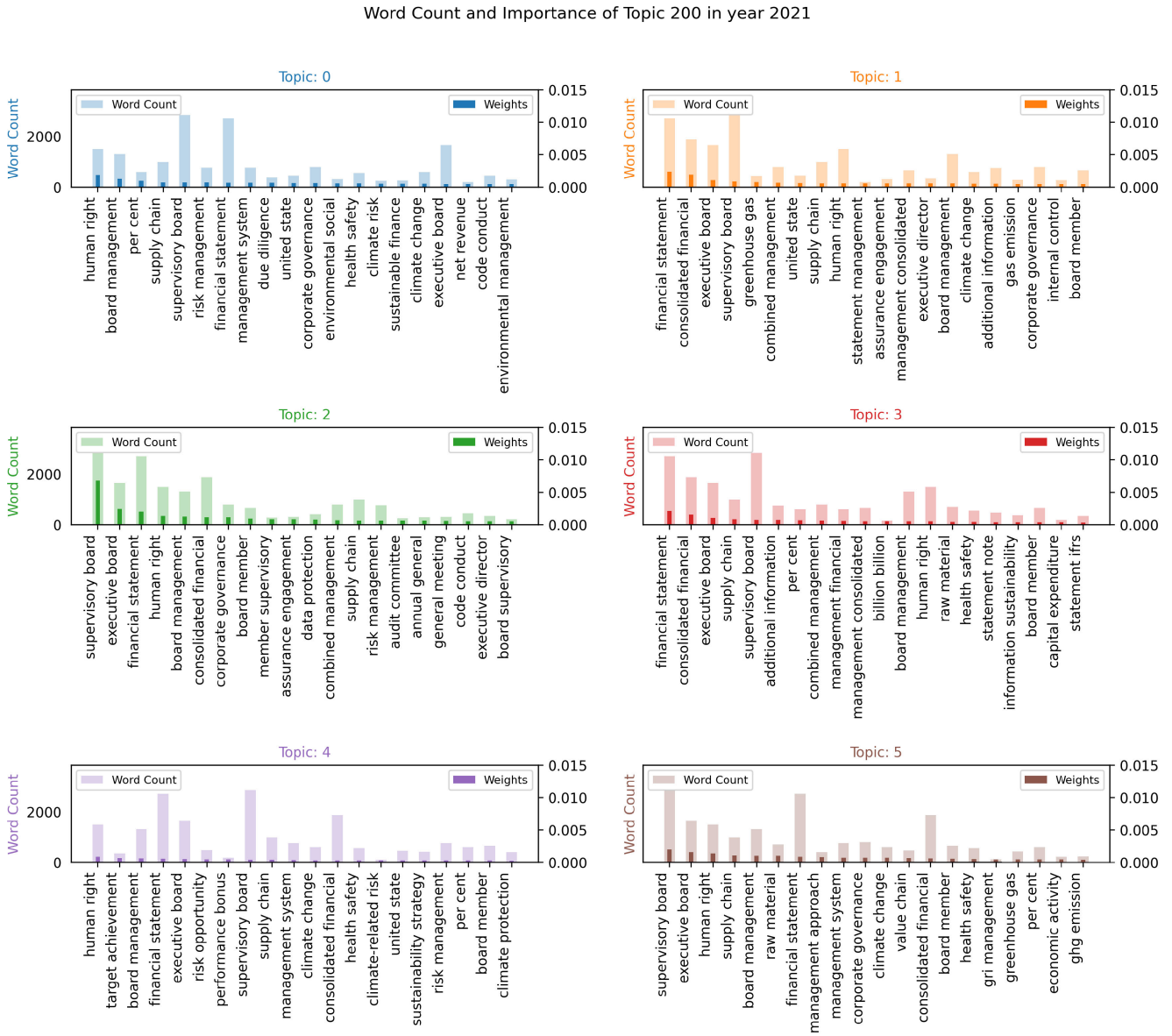

FIGURE 17. GRI 200 - Weight and word count of topic 0-5 2021.

keywords from the GRI framework were searched for in the sustainability reports of the DAX40 companies. Table 13 shows the keywords found for GRI Topics by means of the model used.

Only the keyword ''climate change'' was found for the GRI 200. For GRI 300, the keywords ''fuel consumption'', ''water consumption'', ''ghg emissions'' and ''supply chain'' were identified. In the GRI 400 Topic area, only the keyword ''parental leave'' was identified. This means that the highest hit rate was found for the keywords of the GRI 300 Topics. However, it can be noted that only very few keywords could be identified using LDA.

If partial matches for individual keywords are also considered, no matches were visible for the GRI 200. For the GRI 300, the combinations ''co2 emissions'' and ''co2 consumption'' were evident for the keyword ''co2'', the combinations ''material raw'' and ''material topic'' for the keyword ''material'', and the combination ''green electricity'' for the keyword ''electricity''. For the GRI 400, the combinations ''health safety'' and ''health occupational'' were identified for the keyword ''health'', and the combinations ''diversity equity'' and ''diversity inclusion'' for ''diversity''. These partial matches are listed in Table 14 and only word combinations that were recognized at least twice are shown. In addition, a number of synonyms or word combinations were identified for keywords of the GRI 300, which are listed in Table 15. The synonyms ''co2 emission'', ''co_2 emission'', ''carbon emission'' and

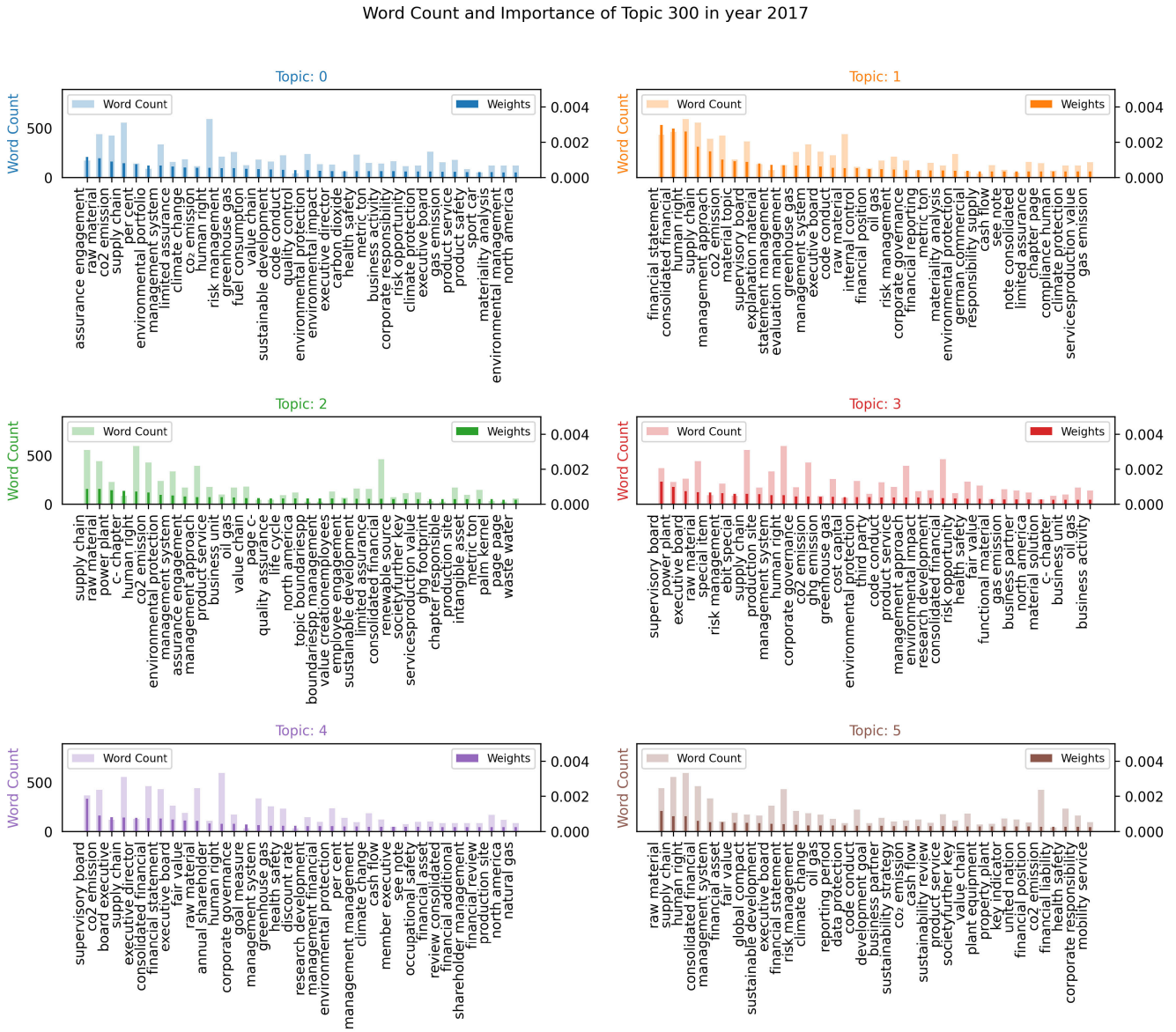

FIGURE 18. GRI 300 - Weight and word count of topic 0-5 2017.

"carbon dioxide" were identified for "co2". For the keyword "ghg gas", "greenhouse gas" was extracted, and for "environment", "environmental impact" and "environmental protection".

Despite adding the partial hits and synonyms to the fully identified keyword hits, only very few keywords could still be identified using LDA. This suggests that the LDA model is only sparsely suitable for identifying GRI Topics in sustainability reports. Moreover, this study investigates whether the LDA model can be used to reflect the structure of the GRI Sustainability Reporting Framework.

Table 8 shows the top 20 bigrams for the GRI 200 over the period under review. Keywords such as "supply chain", "raw material" or "health safety" can be identified, but these do not fall within the topic area of the GRI 200. It can therefore be concluded that the GRI 200 could not be retrieved using the LDA model. Table 10 shows the top 20 bigrams for the GRI 300 in the period from 2017 to 2021. The keywords "supply chain", "raw material", "environmental protection", "co2 emission" and "greenhouse gas" listed for the GRI 300 can be identified. This suggests that the structure of the GRI 300 is partially visible. Looking at the keyword hits for the GRI 300, where a total of four full hits were detected and two of these were reflected in the top 20 bigrams for this GRI, it could be said that it was recovered. In addition, many of the partial hits or synonyms for the GRI 300 were also visible, which consolidates this result. For the GRI 400, the keyword "health safety" was found in the top 20 bigrams of the GRI

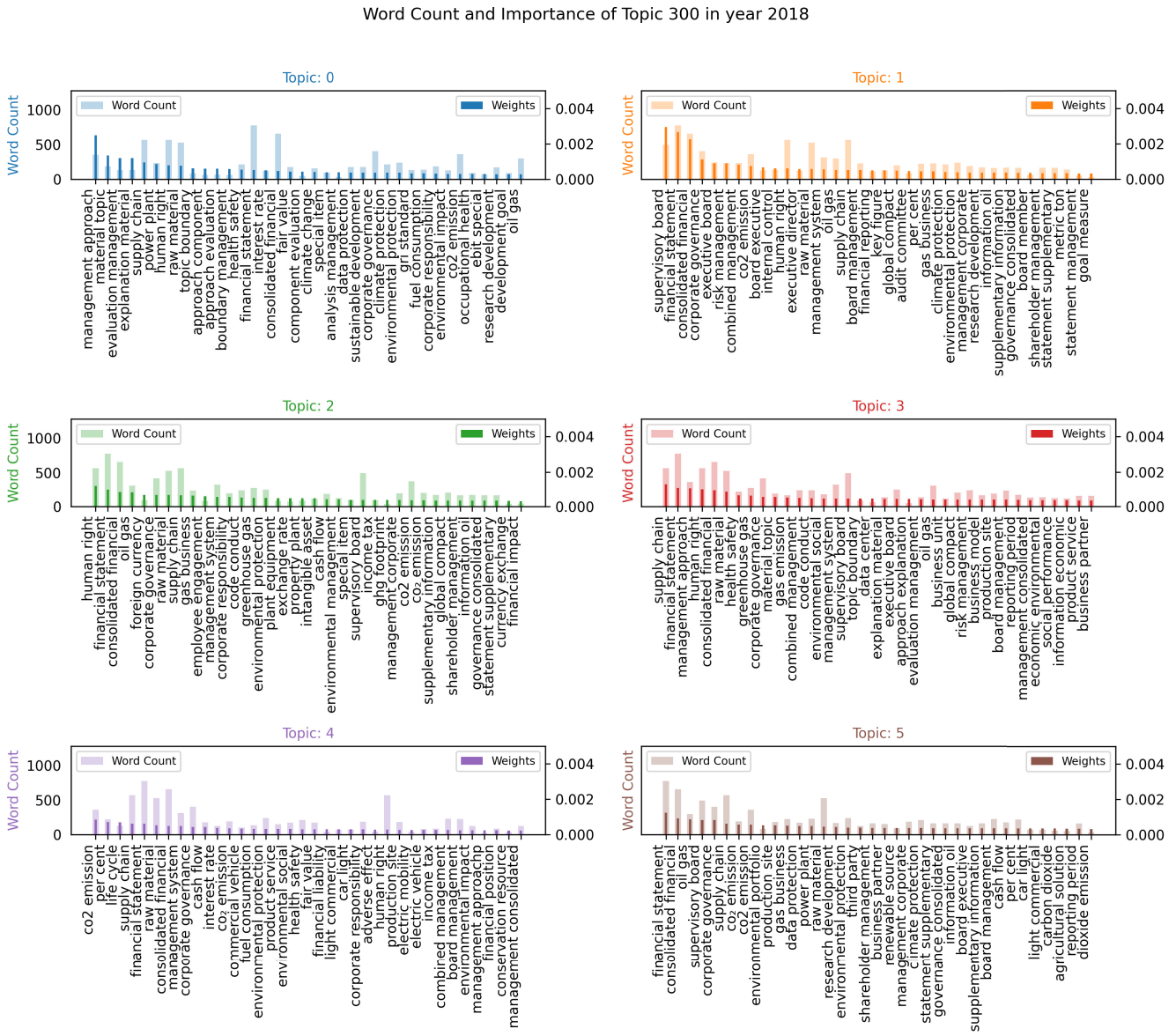

FIGURE 19. GRI 300 - Weight and word count of topic 0-5 2018.

400 (Table 12). It can therefore be assumed that the structure of the GRI 400 is not reflected by means of LDA.

It can thus be stated that the structure of the GRI cannot be found using LDA. Although the GRI 300 could be partially reflected, the GRI 200 and GRI 400 could not. One reason for the retrieval of the GRI 300 could be that most of the keywords for the GRI 300 were recognized. However, when comparing the top 20 bigrams of GRI 200-400, it becomes visible that the hits ''supply chain'', ''health safety'', ''raw materials'' and ''human rights'' appear in all three of the top 20 bigrams, whereby ''supply chain'', ''health safety'' and ''raw materials'' can clearly be assigned as keywords to GRI Topics.

Although the model could not replicate the Topic structure of the GRI, the results of the model reveal insightful information in the input data. Analyzing the frequency of bigrams included in the Topics built by the model for every year, it turns out that there are some content trends that appear consistently over the years and some that change in their occurrence. Figure 6 shows that the number of bigrams extracted from the LDA model is growing over the years. This growth is distributed approximately equally among all GRI Topics, although GRI 200 has the highest average in growth. Figure 5 shows that overall frequent bigrams are included in all 3 GRI Topics and indicates that GRI 200, 300 and 400 interfere in terms of content.

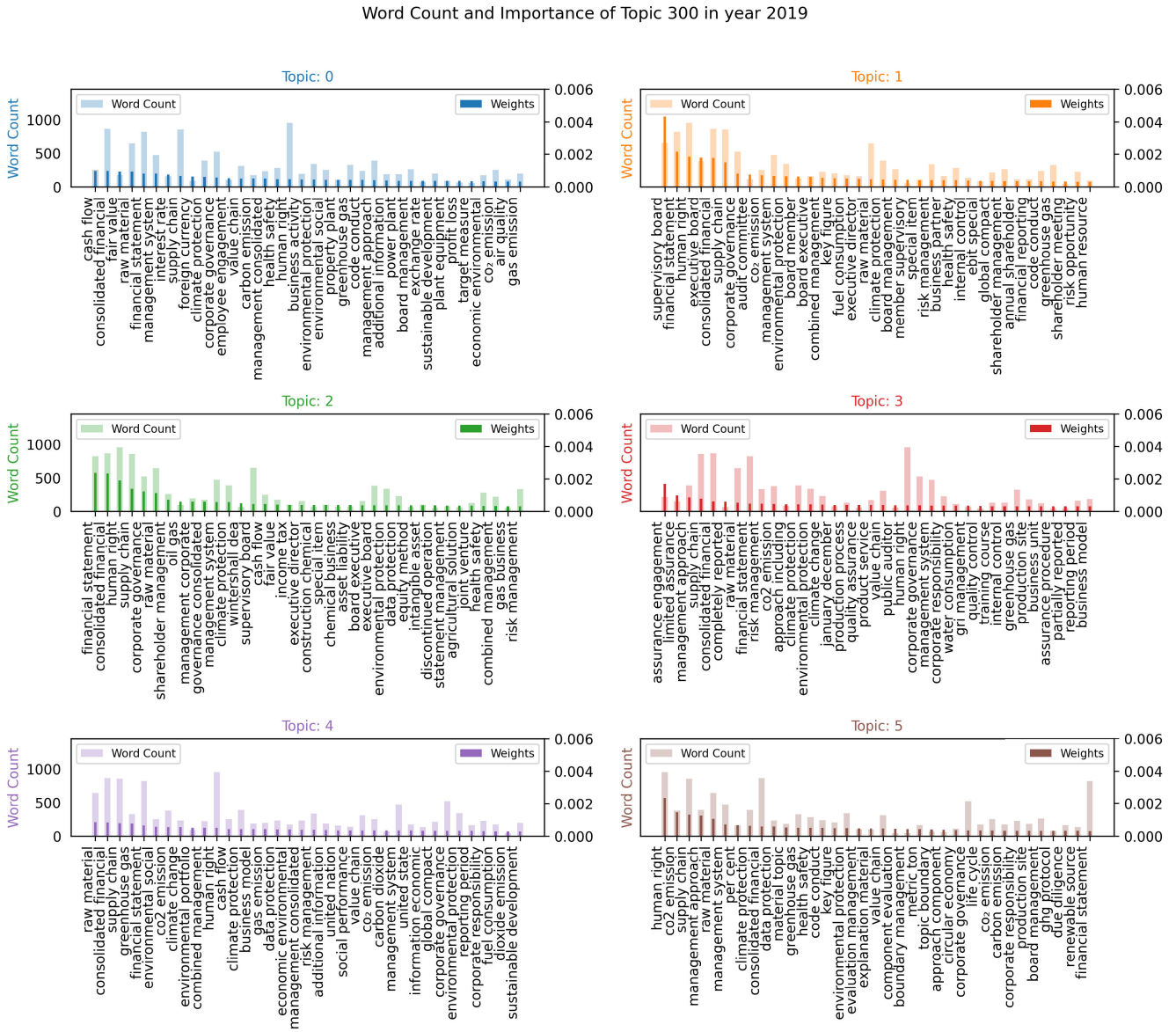

FIGURE 20. **GRI 300 - Weight and word count of topic 0-5 2019.**

For the top 20 bigrams included in GRI 200 (2017-2021) as shown in Table 8 and partially in Figure 7 it appears, that with the exception of 2017 bigrams like ''consolidated financial'', ''financial statement'', and ''human right'' appear regularly and in high numbers in the topics found by the LDA model. Bigrams like ''executive board'', ''supply chain'', ''board management'', ''board member'', and ''health safety'' seem to be more frequent keywords for topics in recent years. Decreasing trends of certain bigrams are not evident, although the remarkably many bigrams of the top 20 bigrams included in topics modeled for 2017 are solely represented this frequently in this specific year. ''co_2 emission'', ''commercial vehicle'', ''occupational safety'', ''sustainability strategy'', ''action area'', ''co2 emission'', ''environmental impact'', ''environmental protection'', ''human resource'', ''key figure'', ''parental leave'', and ''passenger car'' are not included in the 20 most frequent bigrams of the other years. Compared with maximum 3 unique bigrams in the other years, these 11 unique bigrams in the top 20 of 2017 indicate a thematic shift in CSR reporting in GRI 200 after 2017.

For the top 20 bigrams included in GRI 300 (2017-2021) as shown in Table 10 and partially in Figure 8 it appears, that with the exception of 2020 bigrams like ''supply chain'', ''raw material'', ''consolidated financial'', and ''financial statement'' appear regularly and in high numbers in the topics found by the LDA model. Increasing or decreasing trends of certain bigrams are not evident. Noticeable is that many bigrams of the top 20 bigrams

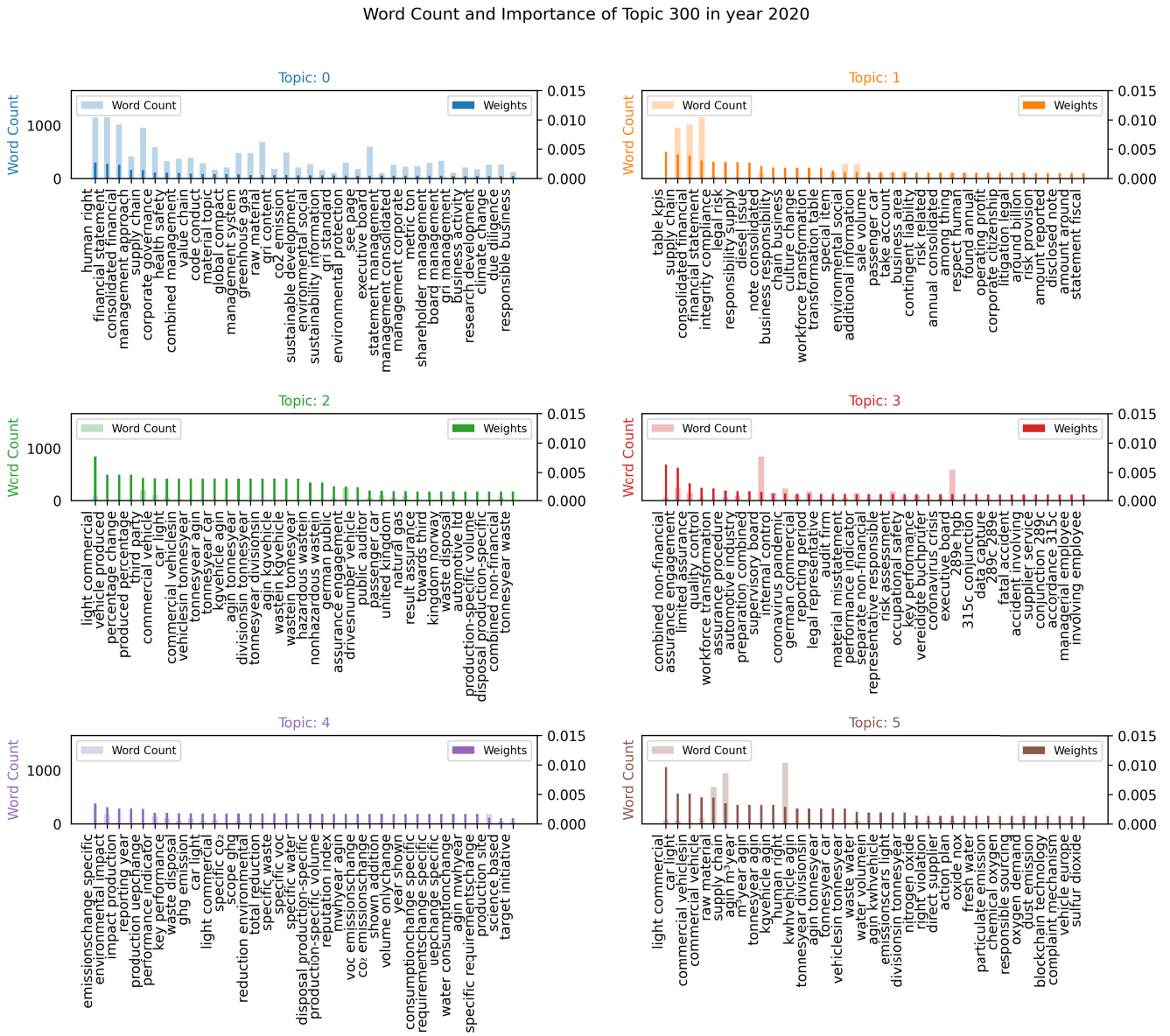

FIGURE 21. GRI 300 - Weight and word count of topic 0-5 2020.

included in topics modeled for 2020 are solely represented this frequently in this specific year. Bigrams like for example "car light", "light commercial", "commercial vehicle", "passenger car", "workforce transformation", "assurance engagement", "chain business", "combined non-financial", "disposal production-specific", "due diligence", and "environmental social" are not included in the 20 most frequent bigrams of the other years. Compared with maximum 3 unique bigrams in the other years, 14 unique bigrams in the top 20 of 2020 indicates, that the model found a thematic focus in GRI 300 in 2020, apparently with a focus on reports from the automotive industry.

For the top 20 bigrams included in GRI 400 (2017-2021) as shown in Table 12 and partially in Figure 9 it appears, that with some exceptions in 2020, were the frequency tends to be lower, bigrams like "management system", "supply chain", "consolidated financial", "human right", financial statement", and "board management" appear regularly and in high numbers in the topics found by the LDA model. Bigrams like "supervisory board" and "executive board" seem to be more frequent keywords for topics in recent years. Increasing trends of certain bigrams are not evident, although many bigrams of the top 20 bigrams included in topics modeled for 2020 are solely represented this frequently in this specific year. Bigrams like for example "key figure", "around world", "business human", "compliance management", "first time", "sustainability strategy", "co2 emission", "commercial vehicle", "coronavirus pandemic",

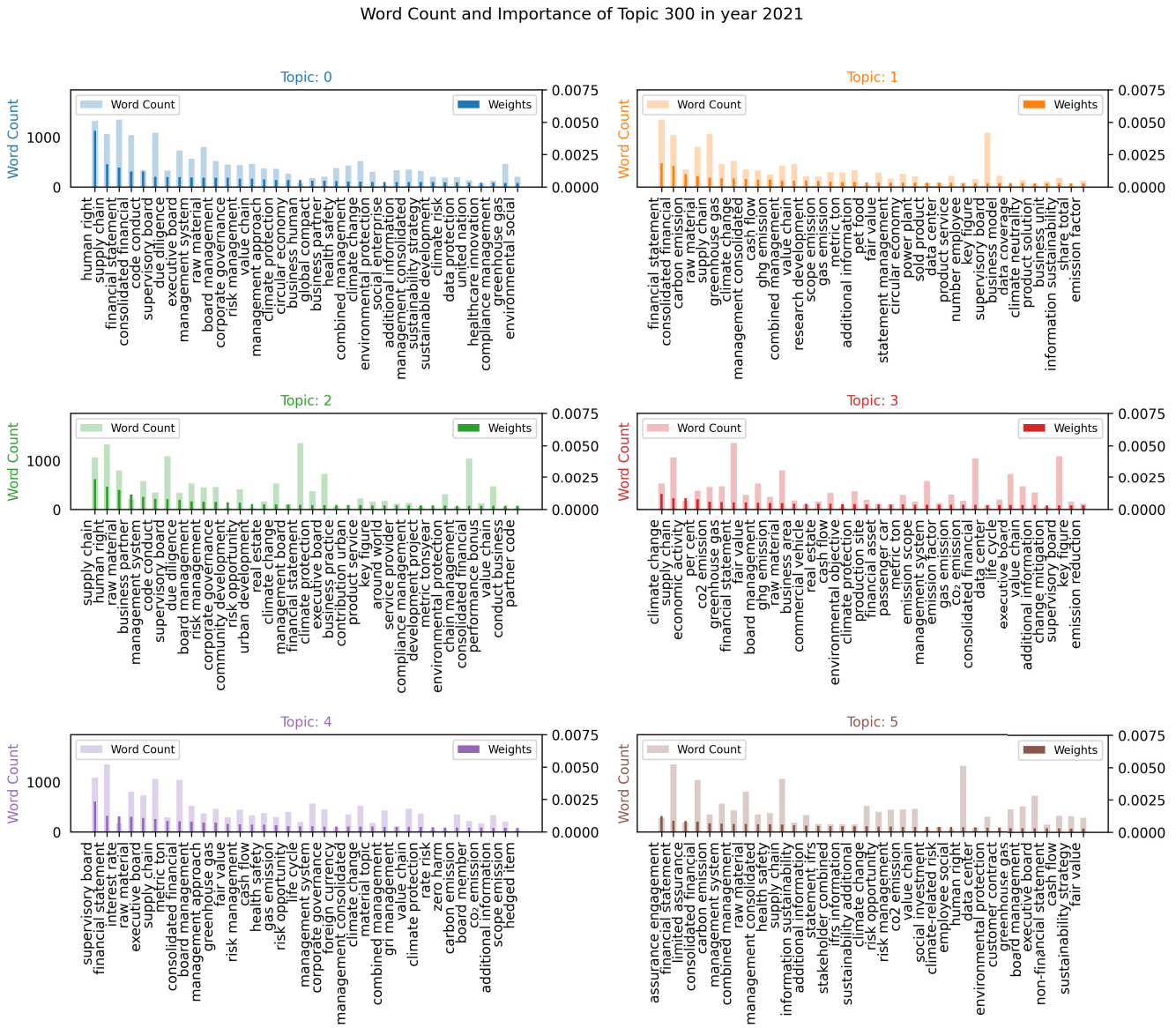

FIGURE 22. GRI 300 - Weight and word count of topic 0-5 2021.

and "corporate culture" are not included in the 20 most frequent bigrams of the other years. Compared with maximum 4 unique bigrams in the other years, 11 unique bigrams in the top 20 of 2020 indicates, that the model found a thematic focus in CSR reporting in GRI 400 in 2020, apparently related to the appearance of the coronavirus pandemic.

VII. CONCLUSION

From the perspective of signaling theory, the work demonstrates valuable insights by using text-mining (here: LDA) on CSR reports. While CSR-relevant information is asymmetrically distributed between companies, customers and the public [47], [48], text mining helps to make this asymmetry visible. The phenomena typical of CSR such as hidden characteristics, intention, action, and information can thus be addressed, and self-selection and sincerity can be promoted.

A. SUMMARY

To do justice to the emerging signal effect of CSR reports and the problem of inconsistent sustainability reports, the interest of this work was to give an overview of currently reported sustainability topics. This paper examined the contents of CSR reports of German DAX companies over a period of five years using LDA Topic Modeling. We examined the current reporting practice among the corporations included in the DAX 40.

The first objective of this work was to classify the main topics of the individual CSR reports based on predefined GRI

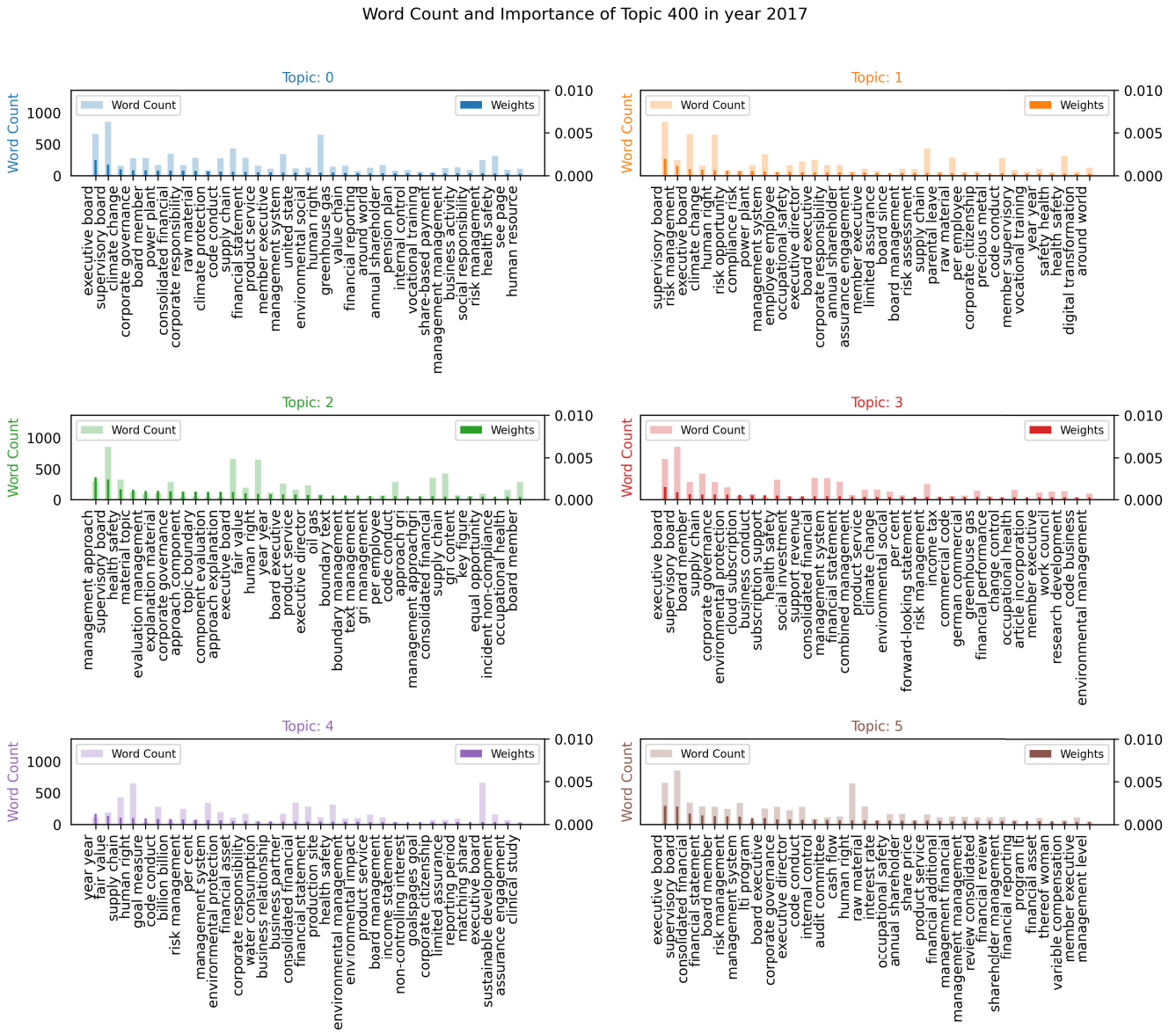

FIGURE 23. GRI 400 - Weight and word count of topic 0-5 2017.

Topic standards. By setting the parameters number of topics and number of words per topic according the overall GRI reporting standards for the respective topic (GRI 200, GRI 300, GRI 400) we tried to replicate the formal structure of the GRI with the Topic Modeling results. The absence of precise matches comparing the manually extracted meaningful Keywords for each GRI Topic with the weighty bigrams included in the topics by the model indicates that this undertaking was not entirely successful. The best result was achieved for the GRI 300 with 4 full matches, 5 partial matches and numerous synonymous word combinations. Although our approach did not succeed in extracting the GRI reporting structure from the input data we are convinced, that if you can successfully tackle the issues as mentioned in section B. LIMITATIONS, Topic Modeling is a promising technique for the analysis of CSR reports and able to identify deviation between obligatory information and actual information included in the reports.

Our second objective was to gain a general understanding of the development of issues in time and in context of external effects. LDA proved to be insightful comparing the occurrence of specific bigrams in the topics configured by the model. We were able to identify bigrams that appear regularly and in high numbers, and therefor are consistent over time. Noteworthy is that some of the detected bigrams appear in all 3 GRI Topics (200-400). This shows that the GRI Topics are not free of crossover in terms of content. Here, bigrams occurring in high quantities in the input data are to

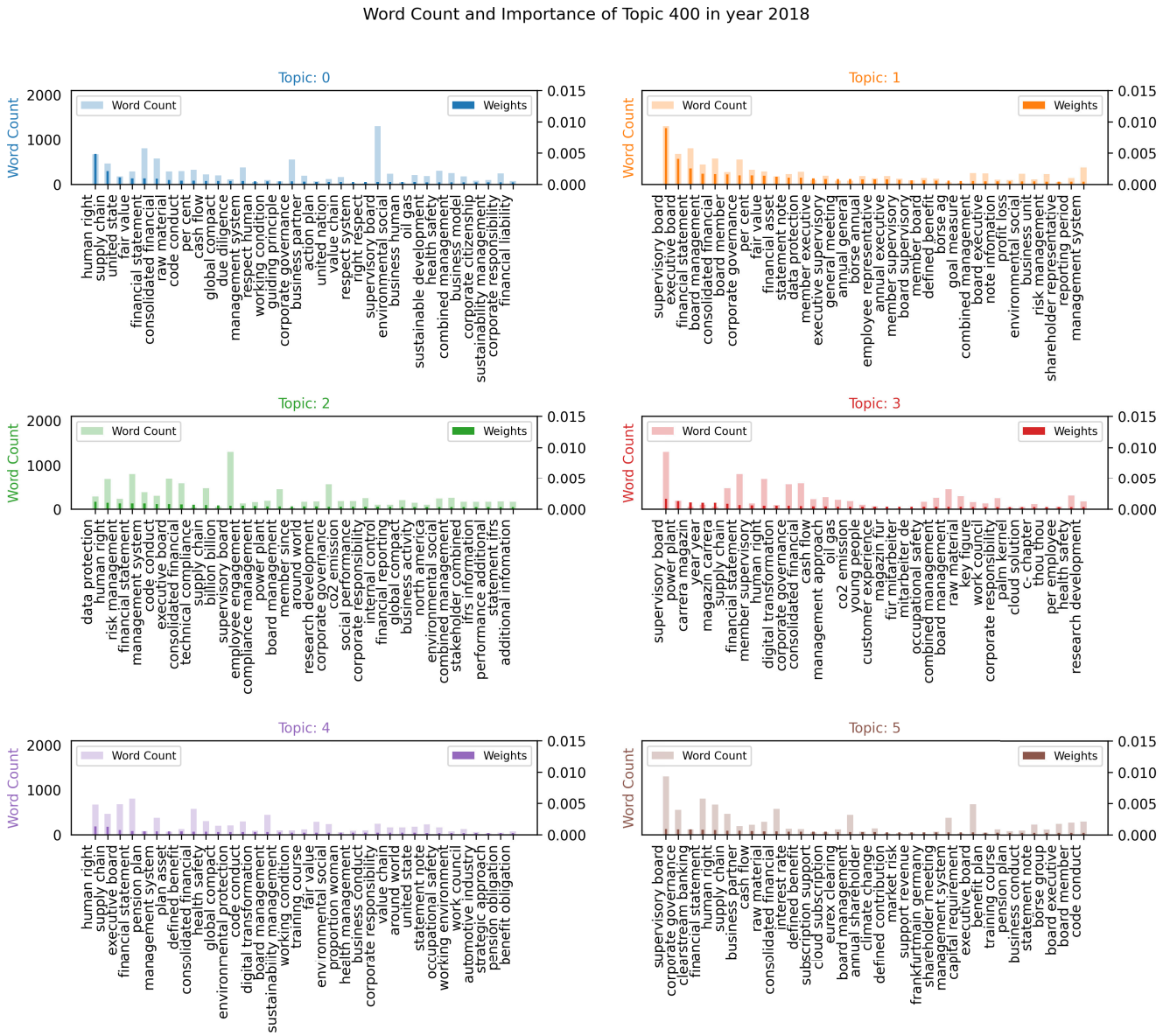

FIGURE 24. GRI 400 - Weight and word count of topic 0-5 2018.

be critically questioned and might be due to "noise" in the data. Furthermore, we compared the occurrence of important bigrams over time and identified bigrams with increasing and decreasing occurrence tendencies. Especially, the comparison year by year for frequent bigrams that are unique to one year allowed to extract shifts in the reporting practices due to external effects like for example the coronavirus pandemic in 2020.

B. LIMITATIONS

The LDA approach has been widely used for topic modeling in various research fields, including sustainability reporting. However, like any other method, it has some limitations that should be taken into account when applying it to sustainability reports. One of the primary limitations of LDA for sustainability reporting is the pre-processing of the texts. Sustainability reports often contain complex and technical language, making it challenging to pre-process them accurately. For example, stop words removal or stemming may remove important information from the texts, and tokenization may split some important words into meaningless parts. Therefore, researchers should carefully choose the pre-processing steps to avoid misleading results.

Furthermore, LDA is based on the bag-of-words approach, where word order and context are ignored. This can lead to a loss of semantic meaning, especially in languages with complex syntax and semantics.

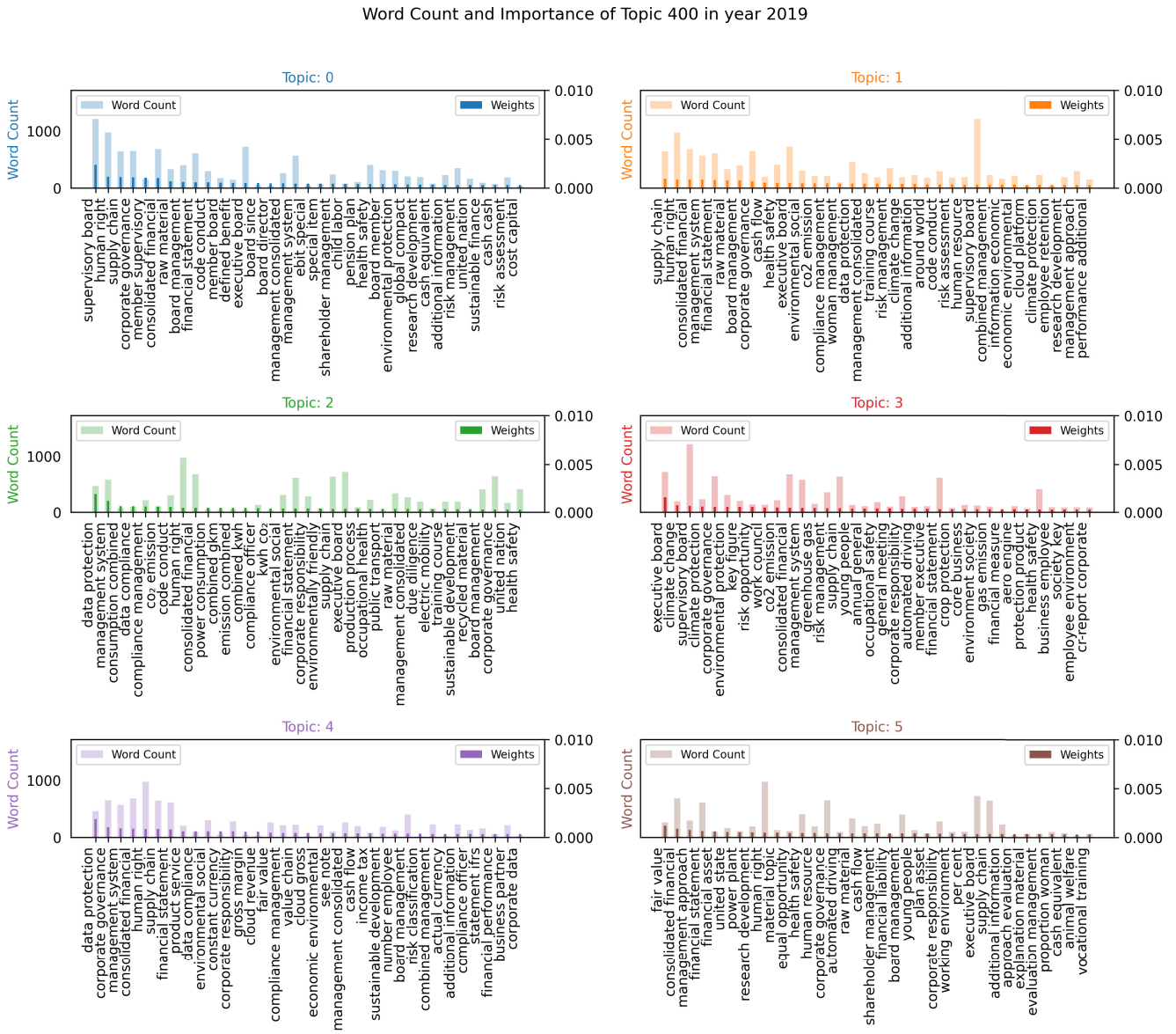

FIGURE 25. GRI 400 - Weight and word count of topic 0-5 2019.

In addition, the number of selected topics and parameter settings of the LDA model are critical factors that can affect the results of the analysis. If the number of topics is too high, the model may overfit the data, and if it is too low, the model may not capture all the important information in the data. Similarly, the choice of the hyperparameters of the model can also affect the results significantly. The use of bigrams is another factor that can affect the results of the LDA model. Bigrams can provide more contextual information, but they can also increase the sparsity of the data, which can affect the quality of the results. Therefore, it is important to experiment with different n-gram sizes and choose the optimal setting based on the performance of the model. Furthermore, company reports may contain biased data (more favorable representations of the company).

In addition to the above limitations, LDA also has some general limitations that should be considered when using it for sustainability reporting. For example, it assumes that the topics are generated from a Dirichlet distribution, which may not be the best choice for all data sets. Finally, LDA is an unsupervised method, which means that it may not capture all the important information in the data, especially if the data set is complex or noisy. Given the above limitations of LDA for sustainability reporting, researchers may consider using other machine learning approaches.

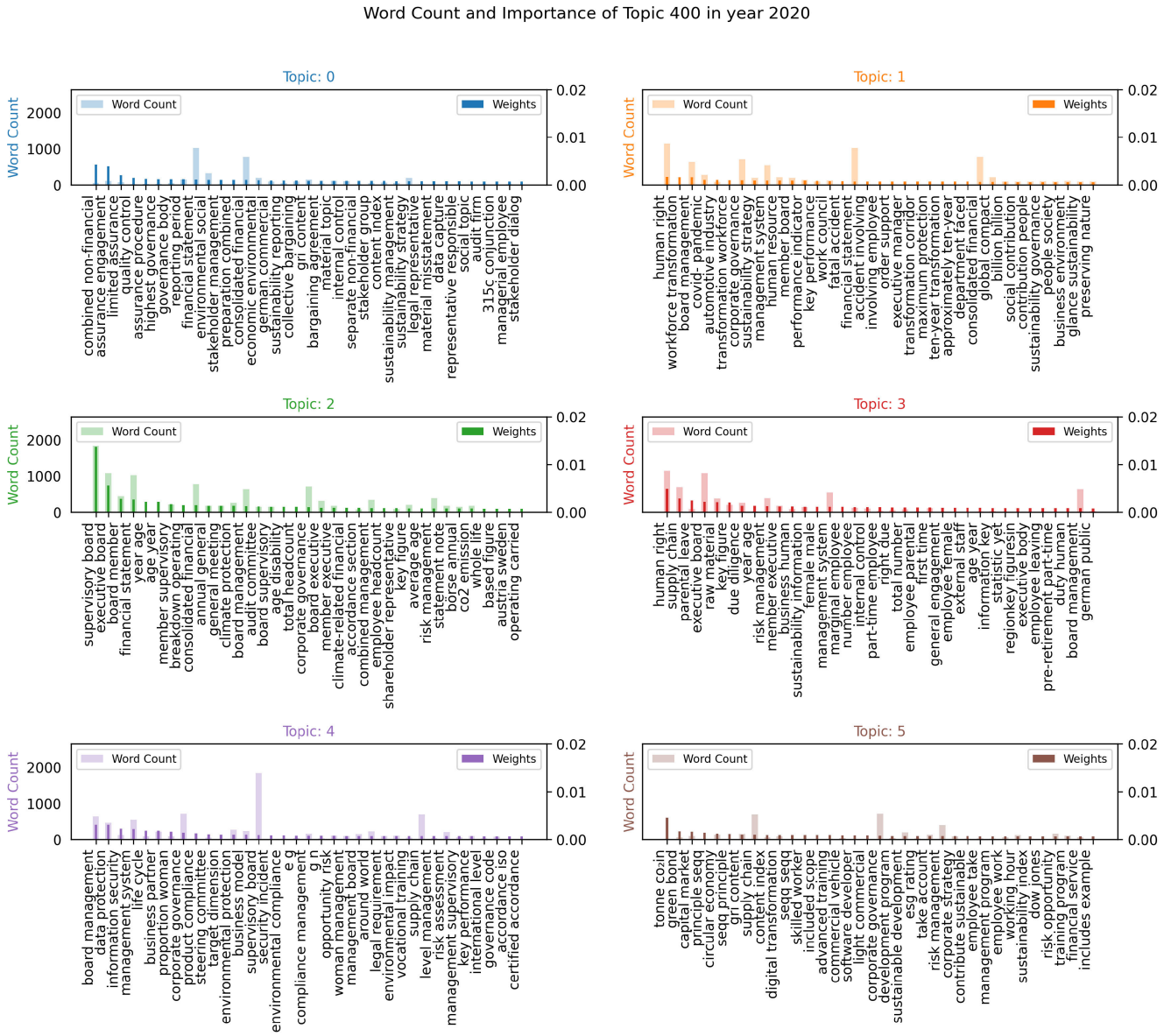

FIGURE 26. GRI 400 - Weight and word count of topic 0-5 2020.

C. FUTURE WORK

The importance of sustainability reporting and the need for quick and automated information retrieval from the reports will continue to grow in the upcoming years. Especially in the context of analyzing and comparing sustainability reports machine learning approaches provide promising tools to gain a quick understanding of reporting trends and insights into large quantities of unstructured textual data. This isn't only beneficial for the recipients of CSR Disclosure but also helps companies to improve their reporting practices.

For the analysis of topics in sustainability reports, a masked-language model like BERT might be an interesting alternative for further research than the basic LDA model used in this paper. In terms of meeting the GRI's content requirements, a possible solution might be the usage of text classification techniques. For example, a machine learning model could be trained to classify each section of a sustainability report based on the requirements of the GRI standard which later provide the input data for a Topic Model. Furthermore, NLP can be used to analyze the content of sustainability reports. In contrast to previous work, NLP could be used in sustainability reports, for example, to identify cases where sustainability reports do not provide enough detail on certain topics or to identify deviations from a specific reporting standard or legal requirements for sustainability reports.

TABLE 22. Structure and title of the GRI standards.

GRI Standard	GRI Topic (year of last modification)	GRI Subtopic
GRI 200s	GRI 201_ Economic Performance (2016)	Disclosure 201-1 Direct economic value generated and distributed
		Disclosure 201-2 Financial implications and other risks and opportunities due to climate change
		Disclosure 201-3 Defined benefit plan obligations and other retirement plans
		Disclosure 201-4 Financial assistance received from government
	GRI 202_ Market Presence (2016)	Disclosure 202-1 Ratios of standard entry level wage by gender compared to local minimum wage
		Disclosure 202-2 Proportion of senior management hired from the local community
	GRI 203_ Indirect Economic Impacts (2016)	Disclosure 203-1 Infrastructure investments and services supported
		Disclosure 203-2 Significant indirect economic impacts
	GRI 204_ Procurement Practices (2016)	Disclosure 204-1 Proportion of spending on local suppliers
	GRI 205_ Anti-corruption (2016)	Disclosure 205-1 Operations assessed for risks related to corruption
		Disclosure 205-2 Communication and training about anti-corruption policies and procedures
		Disclosure 205-3 Confirmed incidents of corruption and actions taken
	GRI 206_ Anti-competitive Behavior (2016)	Disclosure 206-1 Legal actions for anti-competitive behavior, anti-trust, and monopoly practices
	GRI 207_ Tax (2019)	Disclosure 207-1 Approach to tax
		Disclosure 207-2 Tax governance, control, and risk management
		Disclosure 207-3 Stakeholder engagement and management of concerns related to tax 11 2. Topic disclosures
		Disclosure 207-4 Country-by-country reporting
GRI 300s	GRI 301_ Materials (2016)	Disclosure 301-1 Materials used by weight or volume Disclosure
		Disclosure 301-2 Recycled input materials used
		Disclosure 301-3 Reclaimed products and their packaging materials
	GRI 302_ Energy (2016)	Disclosure 302-1 Energy consumption within the organization
		Disclosure 302-2 Energy consumption outside of the organization
		Disclosure 302-3 Energy intensity
		Disclosure 302-4 Reduction of energy consumption
		Disclosure 302-5 Reductions in energy requirements of products and services
	GRI 303_ Water and Effluents (2018)	Disclosure 303-1 Interactions with water as a shared resource
		Disclosure 303-2 Management of water discharge-related impacts
		Disclosure 303-3 Water withdrawal
		Disclosure 303-4 Water discharge
		Disclosure 303-5 Water consumption
	GRI 304_ Biodiversity (2016)	Disclosure 304-1 Operational sites owned, leased, managed in, or adjacent to, protected areas and areas of high biodiversity value outside protected areas
		Disclosure 304-2 Significant impacts of activities, products and services on biodiversity
		Disclosure 304-3 Habitats protected or restored operations
		Disclosure 304-4 IUCN Red List species and national conservation list species with habitats in areas affected by
	GRI 305_ Emissions (2016)	Disclosure 305-1 Direct (Scope 1) GHG emissions
		Disclosure 305-2 Energy indirect (Scope 2) GHG emissions
		Disclosure 305-3 Other indirect (Scope 3) GHG emissions
		Disclosure 305-4 GHG emissions intensity
		Disclosure 305-5 Reduction of GHG emissions
		Disclosure 305-6 Emissions of ozone-depleting substances (ODS)
		Disclosure 305-7 Nitrogen oxides (NOx), sulfur oxides (SOx), and other significant air emissions
	GRI 306_ Effluents and Waste (2016)	Disclosure 306-1 Water discharge by quality and destination
		Disclosure 306-2 Waste by type and disposal method
		Disclosure 306-3 Significant spills
		Disclosure 306-4 Transport of hazardous waste
		Disclosure 306-5 Water bodies affected by water discharges and/or runoff
	GRI 306_ Waste (2020)	Disclosure 306-1 Waste generation and significant waste-related impacts
		Disclosure 306-2 Management of significant waste-related impacts 10 2. Topic disclosures
		Disclosure 306-3 Waste generated
		Disclosure 306-4 Waste diverted from disposal
		Disclosure 306-5 Waste directed to disposal
	GRI 308_ Supplier Environmental Assessment (2016)	Disclosure 308-1 New suppliers that were screened using environmental criteria
		Disclosure 308-2 Negative environmental impacts in the supply chain and actions taken
GRI 400s	GRI 401_ Employment (2016)	Disclosure 401-1 New employee hires and employee turnover
		Disclosure 401-2 Benefits provided to full-time employees that are not provided to temporary or part-time employees
		Disclosure 401-3 Parental leave
	GRI 403_ Occupational Health and Safety (2018)	Disclosure 403-1 Occupational health and safety management system
		Disclosure 403-2 Hazard identification, risk assessment, and incident investigation
		Disclosure 403-3 Occupational health services
		Disclosure 403-4 Worker participation, consultation, and communication on occupational health and safety
		Disclosure 403-5 Worker training on occupational health and safety
		Disclosure 403-6 Promotion of worker health
		Disclosure 403-7 Prevention and mitigation of occupational health and safety impacts directly linked by business relationships 18 2. Topic disclosures
		Disclosure 403-8 Workers covered by an occupational health and safety management system
		Disclosure 403-9 Work-related injuries
		Disclosure 403-10 Work-related ill health
	GRI 404_ Training and Education (2016)	Disclosure 404-1 Average hours of training per year per employee
		Disclosure 404-2 Programs for upgrading employee skills and transition assistance programs
		Disclosure 404-3 Percentage of employees receiving regular performance and career development reviews
	GRI 405_ Diversity and Equal Opportunity (2016)	Disclosure 405-1 Diversity of governance bodies and employees
		Disclosure 405-2 Ratio of basic salary and remuneration of women to men
	GRI 406_ Non-discrimination (2016)	Disclosure 406-1 Incidents of discrimination and corrective actions taken
	GRI 407_ Freedom of Association and Collective Bargaining (2016)	Disclosure 407-1 Operations and suppliers in which the right to freedom of association and collective bargaining may be at risk
	GRI 408_ Child Labor (2016)	Disclosure 408-1 Operations and suppliers at significant risk for incidents of child labor
	GRI 409_ Forced or Compulsory Labor (2016)	Disclosure 409-1 Operations and suppliers at significant risk for incidents of forced or compulsory labor
	GRI 410_ Security Practices (2016)	Disclosure 410-1 Security personnel trained in human rights policies or procedures
	GRI 411_ Rights of Indigenous Peoples (2016)	Disclosure 411-1 Incidents of violations involving rights of indigenous peoples
	GRI 413_ Local Communities (2016)	Disclosure 413-1 Operations with local community engagement, impact assessments, and development programs
		Disclosure 413-2 Operations with significant actual and potential negative impacts on local communities
	GRI 414_ Supplier Social Assessment (2016)	Disclosure 414-1 New suppliers that were screened using social criteria
		Disclosure 414-2 Negative social impacts in the supply chain and actions taken
	GRI 415_ Public Policy (2016)	Disclosure 415-1 Political contributions
	GRI 416_ Customer Health and Safety (2016)	Disclosure 416-1 Assessment of the health and safety impacts of product and service categories
		Disclosure 416-2 Incidents of non-compliance concerning the health and safety impacts of products and services
	GRI 417_ Marketing and Labeling (2016)	Disclosure 417-1 Requirements for product and service information and labeling
		Disclosure 417-2 Incidents of non-compliance concerning product and service information and labeling
		Disclosure 417-3 Incidents of non-compliance concerning marketing communications
	GRI 418_ Customer Privacy (2016)	Disclosure 418-1 Substantiated complaints concerning breaches of customer privacy and losses of customer data

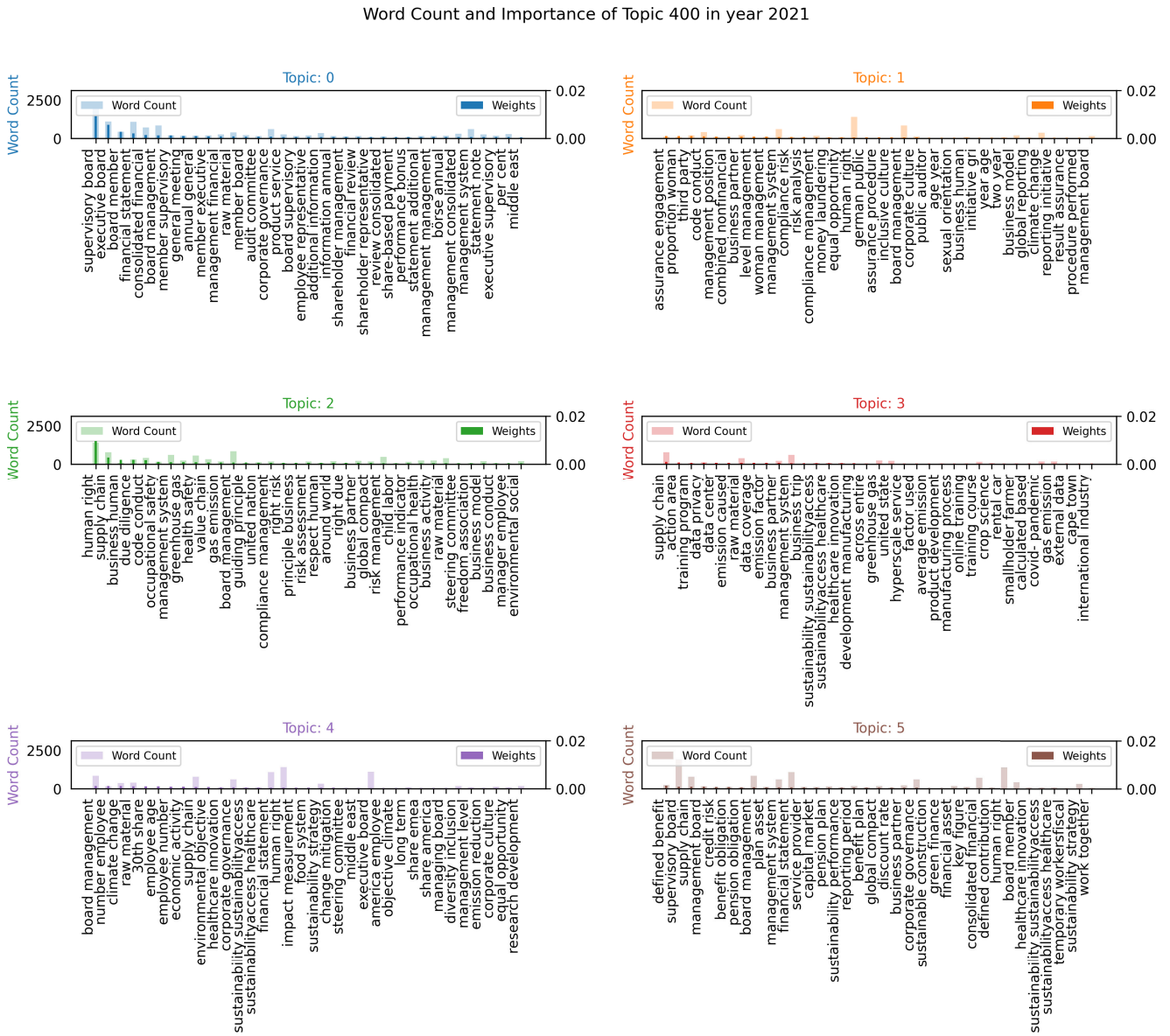

FIGURE 27. **GRI 400 - Weight and word count of topic 0-5 2021.**

APPENDIX

See Tables 16–22 and Figure 10–27.

REFERENCES

[1] H. B. Christensen, L. Hail, and C. Leuz, "Mandatory CSR and sustainability reporting: Economic analysis and literature review," *Rev. Accounting Stud.*, vol. 26, no. 3, pp. 1176–1248, Sep. 2021.

[2] S. Mittelbach-Hörmanseder, K. Hummel, and M. Rammerstorfer, "The information content of corporate social responsibility disclosure in Europe: An institutional perspective," *Eur. Accounting Rev.*, vol. 30, no. 2, pp. 309–348, Mar. 2021.

[3] S. P. Sethi, J. L. Rovenpor, and M. Demir, "Enhancing the quality of reporting in corporate social responsibility guidance documents: The roles of ISO 26000, global reporting initiative and CSR-sustainability monitor," *Bus. Soc. Rev.*, vol. 122, no. 2, pp. 139–163, Jun. 2017.

[4] F. Doni, S. Bianchi Martini, A. Corvino, and M. Mazzoni, "Voluntary versus mandatory non-financial disclosure: EU directive 95/2014 and sustainability reporting practices based on empirical evidence from Italy," *Meditari Accountancy Res.*, vol. 28, no. 5, pp. 781–802, Aug. 2020.

[5] G. Nicolò, G. Zanellato, and A. Tiron-Tudor, "Integrated reporting and European state-owned enterprises: A disclosure analysis pre and post 2014/95/EU," *Sustainability*, vol. 12, no. 5, p. 1908, Mar. 2020.

[6] N. Sun, A. Salama, K. Hussainey, and M. Habbash, "Corporate environmental disclosure, corporate governance and earnings management," *Managerial Auditing J.*, vol. 25, no. 7, pp. 679–700, Jul. 2010.

[7] S. P. Saeidi, S. Sofian, P. Saeidi, S. P. Saeidi, and S. A. Saaeidi, "How does corporate social responsibility contribute to firm financial performance? The mediating role of competitive advantage, reputation, and customer satisfaction," *J. Bus. Res.*, vol. 68, no. 2, pp. 341–350, Feb. 2015.

[8] D. S. Dhaliwal, O. Z. Li, A. Tsang, and Y. G. Yang, "Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting," *Accounting Rev.*, vol. 86, no. 1, pp. 59–100, Jan. 2011.

[9] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," *Financial Innov.*, vol. 6, no. 1, pp. 1–25, Dec. 2020.

[10] M. Reisenbichler and T. Reutterer, "Topic modeling in marketing: Recent advances and research opportunities," *J. Bus. Econ.*, vol. 89, no. 3, pp. 327–356, Apr. 2019.

[11] M. Mustak, J. Salminen, L. Plé, and J. Wirtz, "Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda," *J. Bus. Res.*, vol. 124, pp. 389–404, Jan. 2021.
[12] C. M. Lewis and F. Grossetti, "A statistical approach for optimal topic model identification," *J. Mach. Learn. Res.*, vol. 23, no. 58, pp. 1–20, 2022.
[13] A. Lesnikowski, E. Belfer, E. Rodman, J. Smith, R. Biesbroek, J. D. Wilkerson, J. D. Ford, and L. Berrang-Ford, "Frontiers in data analytics for adaptation research: Topic modeling," *Wiley Interdiscipl. Rev., Climate Change*, vol. 10, no. 3, p. e576, 2019.
[14] I. Goloshchapova, S.-H. Poon, M. Pritchard, and P. Reed, "Corporate social responsibility reports: Topic analysis and big data approach," *Eur. J. Finance*, vol. 25, no. 17, pp. 1637–1654, Nov. 2019.
[15] C. M. Parra, M. C. Tremblay, and A. Castellanos, "Prominent voices and prevalent discourses: A corporate social responsibility application," in *Proc. 11th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2016, pp. 74–78.
[16] J. R. Modapothala and B. Issac, "Study of economic, environmental and social factors in sustainability reports using text mining and Bayesian analysis," in *Proc. IEEE Symp. Ind. Electron. Appl.*, vol. 1, 2009, pp. 209–214.
[17] Global Reporting Initiative. (2022). *Consolidated Set of the Gri Standards*. [Online]. Available: https://www.globalreporting.org/how-to-use-the-gri-standards/gri-standards-english-language/
[18] International Intregrated Reporting Concil. (2021). *Integrated Reporting Framework*. Accessed: Jan. 19. 2023. [Online]. Available: https://www.integratedreporting.org/
[19] International Organization for Standardization. *Social Responsibility—Discovering ISO 26000*. Accessed: Jan. 23, 2023. [Online]. Available: https://www.iso.org/
[20] A. O. Olanipekun, T. Omotayo, and N. Saka, "Review of the use of corporate social responsibility (CSR) tools," *Sustain. Prod. Consumption*, vol. 27, pp. 425–435, Jul. 2021.
[21] Y. Guo and D. C. Yang, "Sustainability accounting reporting: A survey on 30 U.S. Dow-Jones companies," *Int. J. Accounting Taxation*, vol. 2, no. 3, pp. 1–15, 2014.
[22] T. Loughran and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-ks," *J. Finance*, vol. 66, no. 1, pp. 35–65, 2011.
[23] P. K. Sai, P. Gupta, and S. F. Fernandes, "Analysing performance of company through annual reports using text analytics," in *Proc. Int. Conf. Digitization (ICD)*, Nov. 2019, pp. 21–31.
[24] M. Lang and L. Stice-Lawrence, "Textual analysis and international financial reporting: Large sample evidence," *J. Accounting Econ.*, vol. 60, nos. 2–3, pp. 110–135, Nov. 2015.
[25] F. Li, "The determinants and information content of the forward-looking statements in corporate filings—A Naive Bayesian machine learning approach," in *Proc. AAA Financial Accounting Reporting Section (FARS) Paper*, Sep. 2008, doi: 10.2139/ssrn.1267235.
[26] P. Seemakurthi, S. Zhang, and Y. Qi, "Detection of fraudulent financial reports with machine learning techniques," in *Proc. Syst. Inf. Eng. Design Symp.*, Apr. 2015, pp. 358–361.
[27] X. Zhu, S. Y. Yang, and S. Moazeni, "Firm risk identification through topic analysis of textual financial disclosures," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2016, pp. 1–8.
[28] G. Hoberg and C. Lewis, "Do fraudulent firms produce abnormal disclosure?" *J. Corporate Finance*, vol. 43, pp. 58–85, Apr. 2017.
[29] T. Dyer, M. Lang, and L. Stice-Lawrence, "The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation," *J. Accounting Econ.*, vol. 64, nos. 2–3, pp. 221–245, Nov. 2017.
[30] A. H. Huang, R. Lehavy, A. Y. Zang, and R. Zheng, "Analyst information discovery and interpretation roles: A topic modeling approach," *Manage. Sci.*, vol. 64, no. 6, pp. 2833–2855, Jun. 2018.
[31] A. M. Shahi, B. Issac, and J. R. Modapothala, "Automatic analysis of corporate sustainability reports and intelligent scoring," *Int. J. Comput. Intell. Appl.*, vol. 13, no. 1, Mar. 2014, Art. no. 1450006.
[32] S.-H. Liu, S.-Y. Chen, and S.-T. Li, "Text-mining application on CSR report analytics: A study of petrochemical industry," in *Proc. 6th IIAI Int. Congr. Adv. Appl. Informat. (IIAI-AAI)*, Jul. 2017, pp. 76–81.
[33] K. Nakagawa, S. Sashida, R. Kitajima, and H. Sakai, "What do good integrated reports tell us?: An empirical study of Japanese companies using text-mining," in *Proc. 9th Int. Congr. Adv. Appl. Informat. (IIAI-AAI)*, Sep. 2020, pp. 516–521.
[34] J. R. Modapothala and B. Issac, "Evaluation of corporate environmental reports using data mining approach," in *Proc. Int. Conf. Comput. Eng. Technol.*, Jan. 2009, pp. 543–547.
[35] M. Freundlieb and F. Teuteberg, "Corporate social responsibility reporting—A transnational analysis of online corporate social responsibility reports by market–listed companies: Contents and their evolution," *Int. J. Innov. Sustain. Develop.*, vol. 7, no. 1, pp. 1–26, 2013.
[36] D.-S. Chang and Y.-W. Cheng, "Explore the effects of industrial context and leaders' viewpoints on corporate sustainability in Taiwan by text mining," in *Proc. 10th Int. Conf. Service Syst. Service Manage.*, Jul. 2013, pp. 670–673.
[37] W. T. Liew, A. Adhitya, and R. Srinivasan, "Sustainability trends in the process industries: A text mining-based analysis," *Comput. Ind.*, vol. 65, no. 3, pp. 393–400, Apr. 2014.
[38] L. L. Benites-Lazaro, L. Giatti, and A. Giarolla, "Sustainability and governance of sugarcane ethanol companies in Brazil: Topic modeling analysis of CSR reporting," *J. Cleaner Prod.*, vol. 197, pp. 583–591, Oct. 2018.
[39] R. Niveditha, M. R. Parimi, and S. Babu, "Develop CSR themes using text-mining and topic modelling techniques," in *Proc. IEEE Int. Conf. Cloud Comput. Emerg. Markets (CCEM)*, Nov. 2020, pp. 67–71.
[40] D. Buenaño-Fernandez, M. González, D. Gil, and S. Luján-Mora, "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach," *IEEE Access*, vol. 8, pp. 35318–35330, 2020.
[41] S. SAKTHI VEL, "Pre-processing techniques of text mining using computational linguistics and Python libraries," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 879–884.
[42] A. I. Kadhim, Y.-N. Cheah, and N. H. Ahamed, "Text document preprocessing and dimension reduction techniques for text document clustering," in *Proc. 4th Int. Conf. Artif. Intell. with Appl. Eng. Technol.*, Dec. 2014, pp. 69–73.
[43] G. Orellana, B. Arias, M. Orellana, V. Saquicela, F. Baculima, and N. Piedra, "A study on the impact of pre-processing techniques in Spanish and English text classification over short and large text documents," in *Proc. Int. Conf. Inf. Syst. Comput. Sci. (INCISCOS)*, Nov. 2018, pp. 277–283.
[44] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[45] Z. Tong and H. Zhang, "A text mining research based on LDA topic modelling," in *Proc. Comput. Sci. Inf. Technol. (CS IT)*, May 2016, pp. 201–210.
[46] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[47] M. Spence, "Job market signaling," *Quart. J. Econ.*, vol. 87, no. 3, p. 355, Aug. 1973.
[48] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *J. Financial Econ.*, vol. 3, no. 4, pp. 305–360, Oct. 1976.
[49] STOXX Ltd. (2023). *Index Composition Report*. [Online]. Available: https://www.dax-indices.com/documents/dax-indices/Documents/Resources/WeightingFiles/Composition/2023/January/DAX_ICR.20230106.xls

TOBIAS CONTALA received the bachelor's degree in business administration and business informatics from the University of Augsburg, Germany, in 2020 and 2021, respectively. He is currently pursuing the master's degree in business administration with the University of Bayreuth, Germany, with a focus on information systems and controlling. His current research interests include machine learning, natural language processing, and process mining, in the field of finance and controlling.

ALEXANDER-MICHAEL GERK received the bachelor's degree in economics from the University of Bayreuth, Germany, in 2022, where he is currently pursuing the master's degree in business administration, with a focus on finance, accounting, controlling, and taxation. His current research interests include sustainability reporting and the digitalization of SMEs.

JOHANNES HOETTLER received the bachelor's degree in business administration from the University of Bayreuth, Bavaria, Germany, in 2021, where he is currently pursuing the master's degree in business administration, with a focus on technology, operations, and processes. His current research interests include machine learning and natural language processing.

RICARDO BUETTNER (Senior Member, IEEE) received the Dipl.-Inf. degree in computer science and the Dipl.-Wirtsch.-Ing. degree in industrial engineering and management from the Technical University of Ilmenau, Germany, the Dipl.-Kfm. degree in business administration from the University of Hagen, Germany, the Ph.D. degree in information systems from the University of Hohenheim, Germany, and the Habilitation (venia legendi) degree in information systems from the University of Trier, Germany. He is currently a Chaired Professor of information systems and data science with the University of Bayreuth, Germany. He has published over 140 peer-reviewed articles, including articles in *Electronic Markets*, *AIS Transactions on Human–Computer Interaction*, *Personality and Individual Differences*, *European Journal of Psychological Assessment*, *PLOS One*, and IEEE Access. He has received 17 international best papers, the best reviewer, and the service awards and award nominations, including the Best Paper Awards by *AIS Transactions on Human–Computer Interaction*, *Electronic Markets* journal, and HICSS, for his work.