Social impacts reflected in CSR reports

Method of extraction and link to firms innovation capacity

Nechaev, Ivan; Hain, Daniel S.

Published in:
Journal of Cleaner Production

DOI (link to publication from Publisher):
10.1016/j.jclepro.2023.139256

Creative Commons License
CC BY-NC 4.0

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Nechaev, I., & Hain, D. S. (2023). Social impacts reflected in CSR reports: Method of extraction and link to firms innovation capacity. *Journal of Cleaner Production*, *429*, Article 139256. https://doi.org/10.1016/j.jclepro.2023.139256

Journal of Cleaner Production 429 (2023) 139256

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Social impacts reflected in CSR reports: Method of extraction and link to firms innovation capacity

Ivan Nechaev [a,b,*], Daniel S. Hain [a]

[a] Aalborg University Business School, Denmark
[b] Sino-Danish College (SDC), University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Handling Editor: Govindan Kannan

Dataset link: https://github.com/ia-nechaev/sbic-method

Keywords:
Social impact
Corporate social responsibility (CSR) report
Global Reporting Initiative (GRI)
Natural language processing
Machine learning
CSR and innovation interconnection

ABSTRACT

Assessing and comprehending the social impact of firms at global and local level is a pressing concern for both researchers and policy-makers. To address this concern, our paper contributes to the stream of literature that studies the content of Corporate Social Responsibility (CSR) reports (which are also referred to as non-financial statements, sustainability reports or parts of annual reports) using text mining methods. We present a novel approach called Standard-based Impact Classification method (SBIC method), which employs natural language processing (NLP) and supervised machine learning techniques to identify the types of social impacts reflected in CSR reports. We deploy a Random Forest model which we train on reports adhering to Global Reporting Initiative (GRI) framework, enabling the identification of social impact in the majority of CSR reports that do not conform to this standard. Our proposed SBIC method serves as a valuable tool for comparing the social impacts generated by firms, industries, or countries. We showcase an application of our approach by examining the relationship between a company's social impact and its innovation capacity. Our findings support the existing literature consensus that CSR activities generally exhibit a positive correlation with a firm's ability to innovate. Furthermore, we reveal that specific types of social impacts have a more pronounced influence on innovation capacity.

1. Introduction

Companies function simultaneously in economic, social, and environmental dimensions. It is vitally important to capture and understand the causes, links and effects a company produces in each dimension (Emerson, 2003; Stiglitz et al., 2009). Companies (as well as their social environment) are aware that results of their functioning stretch beyond the products or services and affect social sphere. Capturing and understanding these social impacts is an actual problem among many branches of contemporary social science.

Various concepts and methods trying to tackle social impacts[1] have emerged through the recent decades. In accordance with the classification made in a report by Clark et al. (2004) we focus on the process methods to capture social impacts. Frameworks for preparing a non-financial statement vary all over the world. In the European Union reporting rules are set by Non-Financial Reporting Directive (Directive 2014/95/EU) and several other legislative acts. Recommended reporting frameworks include Eco-Management and Audit Scheme (EMAS), the United Nations (UN) Global Compact, ISO 26000, the Global Reporting Initiative and some others. The abundance of frameworks in practice turns out to cause serious hindrances to reporting transparency, information credibility and objectiveness, third-party assurance ability, and cross-company/industry/country comparison of social or environmental effects (Olanipekun et al., 2021; Siew, 2015; Asif et al., 2019). No unified format exists for information representation and that complicates the analysis on a scale of more than several companies. There is a distinctive need of easily accessible mechanisms that can evaluate the CSR performance of a company, can assure that the information disclosure reflects CSR activities unambiguously, and can propose and introduce changes to further revisions of standards (Olanipekun et al., 2021). Our study aims to make a step towards developing a tool that will address the issues mentioned above.

We analyze non-financial statements or corporate social responsibility (CSR) reports of companies to retrieve their reflections on the effects they produce in social sphere. CSR reports serve as a source of text data. One of the most fruitful methods to investigate their contents is natural language processing incorporating machine learning tools. Current literature on the topic reviews two distinguished approaches to

* Corresponding author at: Aalborg University Business School, Denmark.
E-mail address: ivann@business.aau.dk (I. Nechaev).

[1] In literature, terms 'social impacts' and 'social effects' are mostly used interchangeably but for clarity purposes we will use social impacts throughout this article.

https://doi.org/10.1016/j.jclepro.2023.139256
Received 1 May 2023; Received in revised form 7 September 2023; Accepted 10 October 2023
Available online 18 October 2023

CSR reports analysis. The first one uses topic modeling tools (namely Latent Dirichlet Allocation — LDA) (Goloshchapova et al., 2019; Lee and Huang, 2020; Ning et al., 2021). The second one utilizes various researcher-predefined thematic dictionaries (Pencle and Mălăescu, 2016; Uyar et al., 2021; Kiriu and Nozaki, 2020; Kumar and Das, 2021; Liu et al., 2017). Both methods produce narrow research-specific results, which complicates study comparison. These methods generally identify broad topics providing little insight into the elements of CSR dimensions. We propose the third type of analysis based on classification techniques of social impacts reflected in reports prepared in any framework. Our Standard-based Impact Classification (SBIC) method of analysis can help get better, more fine-grained insights into the reported effects, check for standards compliance, and compare effects reported in different reporting frameworks.

The main contribution of our research is pre-trained machine learning model for identifying social impact types of a given text (CSR report) in accordance with Global Reporting Initiative (codes 401–419).[2] The research application of the method aims to contribute to the studies of the interconnection that exists between innovation and Corporate Social Responsibility. Our findings support existing theories and provide in-depth overview of social dimension of CSR.

The remainder of the paper is structured as follows. In the Literature background section we frame our study into the existing field and provide a brief overview that serves as a ground for our theory choices. In Research methods section we describe the data acquisition process and the tools we used to develop our classification model. The Results section shows the main characteristics of our predictive model. In the Research application section we demonstrate the model applied to real-world data and replicate the findings of García-Piqueres and García-Ramos (2021) and Broadstock et al. (2020) in more detail using panel data regression. The Discussion and limitations section contains the aspects that needed to be taken into account when applying our proposed classification method. In the Conclusion section we enumerate other possible applications of Standard-based Impact Classification (SBIC) method and future research avenues.

2. Literature background

2.1. Impact assessment

The topic of social impact assessment is extremely large. In this section we provide a brief overview of the main approaches towards the problem of social impact identification that served as theoretical foundation for our research.

Various concepts and frameworks dealing with issues of indented and unintended social consequences have emerged through the recent decades. Social impact assessment finds its way to various social performance standards. An example of methods classification for social impact assessment is compiled in Appendix Table 7 based on Cerioni and Marasca (2021). Clark et al. (2004) identified three categories of social impact assessment by their major function: process methods, impact methods, and methods of monetization. In accordance with this classification we use the results of process method implementation, namely Global Reporting Initiative, as a framework for impact assessment.

Voluminous literature reviews on the social impact topic could be found in works by Jones et al. (2017) and Molecke and Pinkse (2017). Molecke and Pinkse (2017) provide a comprehensive overview upon existing social impact assessment approaches. To name a few: Contingent, stated and revealed preferences (Haab and McConnell, 2002); Estimations of decision utility (Dolan and Kahneman, 2008); Risk assessment methodology (RSIA) (Mahmoudi et al., 2013); Social development needs analysis (SDNA) tool (Esteves and Vanclay, 2009) etc.

Despite the abundance of frameworks and methodologies there are common difficulties of various nature that practitioners meet during assessment, for example, misrepresentation of information, overstating positive impacts (Uyar et al., 2020; Liew et al., 2014; Ruiz-Blanco et al., 2022) and holding back negative ones. Some social issues are ill-considered and undervalued because of a reporting manager's subjectivity (Vanclay and Hanna, 2019). Among up-to-date problems in this field Vanclay (2020) emphasized the need for new methods for assessing impact from projects. Ebrahim and Rangan (2014) directly stated the need for new conceptual framework for social enterprises. Emerson (2003) denoted the need to capture simultaneous impact in all three dimensions, i.e. economic, social, and environmental: 'There is no "trade off" between the three, but rather a concurrent pursuit of value — social, financial, and environmental'. Attempts to solve some of the above mentioned issues could be found in Human Rights, Ethical and Social Impact Assessment (HRESIA) framework (Mantelero, 2018).

All these standards and multiple cases of their implementation provide a solid evidence of a strong need to incorporate social impact in decision making in all spheres of economic activity and at all managerial levels. These social impact assessment methods use data obtained through interviews, surveys and company reports. In the current paper we focus on company reports providing non-financial data in a textual form.

Cerioni and Marasca (2021) gave the following social impact definitions: the ability of an organization to contribute to change; the attribution of the activities of an organization to the overall social results of the longer term; the non-economic change created by business activities and investments; the share of the total outcome obtained as a direct result of the intervention and finally the sustainable change in the long term.

We would argue that definitions of social impact are very method-dependent as every method enumerated captures only some parts of social reality and sees social impact through its own lenses. Global Reporting Initiative (GRI) standards on non-financial reporting is no exception. It is just a particular example of CSR reporting framework and even within this particular example one may come across different perspectives on what should be captured and reflected in companies' reports, hence, on what kind of CSR activity company should focus. The comparison of CSR frameworks is beyond the scope of this paper.[3] Although we can assume that the main criteria when choosing what the frameworks should focus on is the understanding of sustainability, which all of these frameworks have as a common foundation. The differences are in various names of elements and their priorities regarding each other, but the essence of sustainability remains the same across all the frameworks.

The notion of sustainability is changing over time incorporating new social elements and discarding irrelevant ones. We suppose that this change influences all stages of CSR reporting. Under the concept of sustainability we consider parts of social sphere that are related to commercial companies and contribute to and secure personal well-being of stakeholders. If we want to live in a progressive society, we need monitoring tools to check whether standards' developers capture the notion of sustainability correctly, whether companies do things sustainably and what we understand under up-to-date sustainability and whether it conforms to what we observe in reality. We believe that the method we propose here helps develop such types of tools.

[2] Current version of pre-trained model can be found on GitHub: https://github.com/ia-nechaev/sbic-method.

[3] A brief overview can be found in Dimensional Comparison across CSR Frameworks performed by Pencle and Mălăescu (2016).

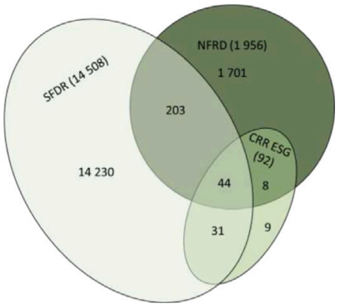

Fig. 1. Companies subject to the NFRD, SFDR and CRR ESG sustainability-related disclosure requirements in the EU27 (de Groen et al., 2021).

2.2. CSR reporting frameworks

From legislative perspective of the EU there are four groups of companies that report on sustainability issues. The first group consists of public interest entities with more than 500 of employees that report according to *Non-Financial Reporting Directive (NFRD, Directive 2014/95/EU)*. The second and third groups are companies that are to report within *Regulation on disclosures relating to sustainable investments and sustainability risks (SFDR)* and *Art. 449(a) of Regulation (EU) No 575/2013 on prudential requirements for credit institutions and investment firms (CRR ESG)* respectively. The fourth group are companies that report on a voluntary basis. As of 2020–2021, first three groups result in 16000 entities total (Fig. 1) and the fourth adds up approximately 9000 entities more resulting overall in 25000 companies that provide information on their corporate responsibility.

2.2.1. CSR reports in general

Well-deployed CSR activities create a good reputation and a nice social and eco-friendly brand image, improve employee commitment and retaining, establish better relations with major stakeholders (consumers, employees, governments, non-governmental organizations (NGOs) and the community), reduce costs and avoid regulatory sanctions, which all in all can provide a company with a competitive advantage and result in higher profitability (Uyar et al., 2021; Wasiuzzaman et al., 2021).

Various reporting tools help a company implement CSR activities systematically, they provide standards and methodology of communicating relevant activities transparently and can serve for the purpose of CSR performance assessment. An extensive overview of social reporting tools is performed by Siew (2015) and a new version by Olanipekun et al. (2021).

According to EU Directive 2014/95/EU companies can choose one out of several reporting frameworks to produce non-financial statement; social impact audit can be both mandatory and voluntary as well as self produced or done by a third party. In practice these frameworks refer to the following basic documents.

Firstly, multinational normative documents: United Nations' *Guiding Principles on Business and Human Rights*; Organization for Economic Co-operation and Development's *Guidelines for Multinational Enterprises* (OECD, 2011) and its continuation in *Due Diligence Guidance for Responsible Business Conduct* (Anon, 2018a); World Bank's *Environmental and Social Framework* (Anon, 2016); Global Sustainability Standards Board's *Global Reporting Initiative Standards* (Anon, 2020a). Secondly, financial guidance: International Financial corporation's *Environmental and Social Performance Standards* (Anon, 2012); *Equator principles* (Anon, 2020b). Thirdly, industry-based standards: Value Reporting Foundation Sustainability's Accounting Standards Board's *Standards Application Guidance* (Anon, 2018b); Morgan Stanley Capital International (MSCI) *ESG Universal Indexes Methodology* (Anon, 2019); International Association for Impact Assessment's *Social Impact Assessment* (Vanclay, Esteves, Aucamp, Research, and Franks, 2015).

Many studies analyze the disclosure of non-financial information for compliance with reporting standards. They reflect upon performance of standards implementation and also link the degree of disclosure of non-financial information with other financial and economic indicators of a company. García-Sánchez (2020) found that a company's incentives and the intrinsic characteristics of assurer have a greater effect on the CSR report assurance quality than the effectiveness of top management or institutional pressures. In the following study García-Sánchez et al. (2022) reflected on the phenomenon of *CSR decoupling* when information is disclosed selectively or does not fully reflect a company's actual performance. The same phenomenon is described as *The bias of the principle of the materiality of the GRI tool* in the study by Olanipekun et al. (2021).

Asif et al. (2019) provides systematic analysis of existing standards. The author differentiates between public and private standards of social compliancy (companies such as IKEA, Nike, Adidas have own developments in terms of disclosure requirements for non-financial information), overviews the decision making process underlying standards adoption, elaborates on how the company employees' perception of standards implementation influences firm performance and argues that multiple standards adoption could be counter productive. There is a need to check the report compliance with reporting standards and with actual performance.

To sum it up, CSR frameworks focus on the effects firms produce on parts of social reality that are considered to be crucial from sustainability standpoint at current moment of time. With certain reservations mentioned in the next subsection we ague that text in CSR reports capture the reflection of firms producing different types of effects.

2.2.2. GRI reports in particular

We used the Global Reporting Initiative (GRI) framework as our reference point for several reasons. Reports prepared in standardized frameworks (such as GRI) tend to be more credible compared with non-standardized (Lock and Seele, 2016). The scope of topics covered within the GRI framework is wide and exceeds the scope of free-format reports (Uyar et al., 2021). The major technical reason for choosing GRI framework is that its every non-financial domain is codified and all reported effects are bound to a certain GRI code. Reports prepared in accordance with GRI framework contain index tables where codes (impact types) and corresponding pages numbers are specified. Therefore, GRI reports can be regarded as a source of labeled data, which allows to solve the classification problem in the field of supervised machine learning: predict the impact type of a given text.

The structure of a GRI report is based on GRI standards Anon (2020a). Current set of standards is in effect since July 2018. It consists of Universal Standards (codes 101–103) and three topic-specific standards: Economic (codes 201–207), Environment (codes 301–308), Social (codes 401–419).[4] Most of standards have additional subcodes that help a more exact disclosure. For example, one of the most reported in our training dataset Social code '401-Employment' contains the following three subcodes: 401-1 'New employee hires and employee turnover'; 401-2 'Benefits provided to full-time employees that are not provided to temporary or part-time employees'; 401-3 'Parental leave'. Every standard code defines reporting requirements, reporting recommendations and guidance. A new version of standards is in act in 2023.

A brief overview of previous studies on the topic of GRI framework application shows the scope of research aims and provides the valuable critique of the framework. Brown et al. (2009) explored how the widespread dissemination and implementation of GRI standards has

[4] Currently, we use Social dimension for training our machine learning model and testing its potential in the Research application section.

influenced the processes of institutionalization taking place in society. Marimon et al. (2012) assessed the degree of distribution of GRI standards in the world from the point of macro- and micro-analysis. At macro level the stages and elements of GRI implementation in various parts of the world are identified and at micro level instability indices are calculated. Fuente et al. (2017) studied the relationship between various board of directors' characteristics (structure, gender composition, etc.) and the implementation of GRI standards in a company.

Fonseca et al. (2012) and Siew (2015) agree on the following flaws in GRI framework: guiding vision overlooks the need to operate within the capacity of the biosphere, conceptual framework is tacit, non-systemic and issues-based, evaluation of trade-offs and synergies across the systems is overlooked, geographical scope is weakly addressed, temporal orientation is predominantly retrospective, types of indicators are non-integrated and disclosures of assumptions and uncertainties are limited.

Olanipekun et al. (2021) analyzed several studies on the issues of GRI framework application and finds six major hindrances of implementation. This critique seems to address the issues that are general to the whole family of CSR frameworks irrespective of the particular toolset used. For example, *Limited empowerment of the civil regulation*[5] and *Lack of external verification are interconnected*[6] both indicate a distinctive need of independent social mechanisms that can evaluate the CSR performance of a company, can assure that the information disclosure reflects CSR activities unambiguously and can propose and introduce changes to further revisions of standards.

Flaws enumerated by Fonseca et al. (2012) and Siew (2015) require further conceptual and theoretical development of the framework. We should be aware of them when applying our method. Hindrances of GRI framework exposed by Olanipekun et al. (2021) are of more technical, applied level and could be addressed by improving the tool accessibility and usability. Our method also aims to reduce the significance of these hindrances.

2.3. Text mining CSR reports

Text mining applied to CSR reports can help to answer different research questions: how has environmental sustainability become embedded in corporate policy and the core business discourse (Castellanos et al., 2015); how an industrial disaster influence the disclosure of economic, social and environmental aspects (Aureli, 2017). Text mining can be used to pursue various research goals: to evaluate the subjectivity and objectivity of corporate social responsibility information disclosure (Duan et al., 2018); identify UK and European CSR main topics (Goloshchapova et al., 2019); evaluate the quality of ESG activities (Kiriu and Nozaki, 2020); determine the benchmark of environment performance (Liu et al., 2017). All studies mentioned provide a retrospective analysis of previously published papers on text mining of CSR reports. Fig. 2 shows a schematic representation of existing text-mining methods and a place of our proposed method.

2.3.1. Latent Dirichlet allocation

Blei et al. (2003) introduced topic modeling techniques based on Latent Dirichlet Allocation (LDA). Several recent studies have utilized natural language processing tools to investigate the contents of CSR reports. Goloshchapova et al. (2019) applied LDA analysis to CSR texts and identified the most common topics among more than 5000 reports (namely: employees safety, employees training support, carbon emission, human right, efficient power, and healthcare medicines). Authors denote that LDA cannot help distinguishing good and bad CSR performance.

Ning et al. (2021) support the idea that the aim of sustainability reporting is to manage firm's reputation with customers. The authors also find that sustainability initiatives on environmental issues can increase firms' financial performance.

Along with regular text mining and topic modeling methods Lee and Huang (2020) used fuzzy rough set theory to identify the important features and constructed a forecasting model by means of extreme learning machine with self-adaptive mechanism. By applying these complex methods and models to reports from Taiwan electronics sector authors found that specific CSR dimensions are highly related to corporate financial performance. Regression result supports their initial hypothesis that including CSR related ratios as explanatory variables in a forecasting model will provide it with a higher forecasting capability.

2.3.2. Thematic dictionary approach

Uyar et al. (2021) with the help of network analysis of word frequencies of 478 reports from 44 countries within Hospitality and Tourism industry find that GRI based reports are more condensed and cover a greater scope of topics than free-format reports. GRI reports better address specific stakeholders. European reports cover more sustainability topics than other regions: they are richer in their keyword usage and they are better at meeting stakeholders expectations.

Kiriu and Nozaki (2020) implemented a comparison of their textual analysis of almost 9000 reports with ESG scores of Japan firms provided by Thomson Reuters Asset4 database. For comparison authors used word embedding, word classification, word structuring (a hierarchical word structure based on the frequency and divergence of words in a tree model) and visualization tools.

Kumar and Das (2021) used text mining tools and a predefined list of 208 keyword for the purpose of scoring ESG topics in 200 reports of firms from top 10 economies by GDP. Through the period of 2008–2017 sustainability performance is found to be improved.

Liu et al. (2017) studied 50 reports from petrochemical industry from 2015 and found that those reports had strong focus on the environmental aspect. This finding is pretty common in all studies devoted to corporate responsibility.

Pencle and Mălăescu (2016) compiled deductive and an inductive wordlist to generate a topic dictionary for further content analysis. Authors used the consensus approach to find the most similar dimensions (topics) among GRI, UN Global Compact, IIRC, MSCI KLD and ESG frameworks. These dimensions are: employee dimension, social and community dimension, environment dimension and human rights dimension. The estimated indicators based on this vocabulary were used to build regression models to predict the size of the IPO (initial public offering) in terms of the offering price and the total number of proposed shares, as well as underpricing on the first day of trading.

Studies using text-mining methods proved that both LDA topic modeling and thematic dictionary approach can be used to investigate the peculiarities of sustainability performance of a company. Most of the studies mentioned above used reports made in accordance to GRI framework. By descriptions provided we conclude that current methods of text-mining CSR reports have certain shortcomings. First, they cannot provide specific details on the types of effects contained in text. These methods are restricted either to broader CSR topics or to narrow research-specific scope that makes studies comparison practically impossible. Second, they both require either initial topic compilation or topic interpretation performed by a researcher that can introduce subjectivity into the process. Third, the thematic dictionary approach is bounded by the predefined dictionary, which weakens the ability to adapt to topic changes in data. Our proposed Standard-based Impact Classification (SBIC) method lacks the shortcomings mentioned above.

[5] There are not enough social mechanisms eligible to check the compliance of the reports to the framework, to check information revealed to actual CSR performance and to influence the development of next generation of standards.

[6] Standards' compliance verification is done mainly by CSR standards developers or such assurers as auditing companies as KPMG, E&Y, Deloitte etc. But not all organizations undergo this process since it is voluntary. Hence, the issue of general credibility of CSR framework still exists.

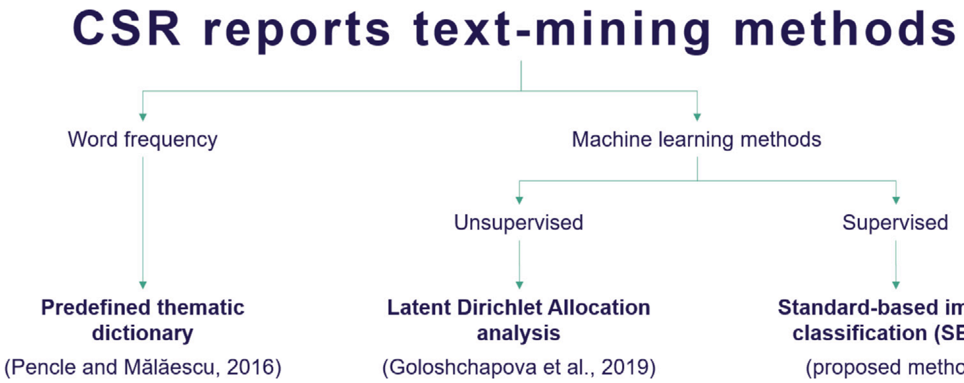
CSR reports text-mining methods
Word frequency
Machine learning methods
Unsupervised
Supervised
Predefined thematic dictionary
(Pencle and Mălăescu, 2016)
Latent Dirichlet Allocation analysis
(Goloshchapova et al., 2019)
Standard-based impact classification (SBIC)
(proposed method)

Fig. 2. Schematic representation of text-mining methods.

3. Research methods

3.1. Classification model dataset

We use reports prepared in accordance with Global Reporting Initiative framework as labeled data to train the model identify social impact types in reports done within other frameworks. We obtained the list of companies that report in GRI framework from globalreporting.org, the official Global Reporting Initiative website. This list contained links to non-financial statements in pdf format that were further downloaded and processed. The time span covers 2017—2020 of published reports. The selection of the time period is dictated by the availability of specifically GRI reports (in 2021 GRI organization changed access rights for the public to their database) and also by the version of the standards remaining stable (2016 version for Social impacts).[7] Descriptive summary is provided in Appendix in Table 5.

To build a training dataset for our classification model, we filtered the reports by the following criteria: the report is done only for one year (avoiding double year reports); the index page contains indication to page numbers, not section names or hyperlinks; the pdf-layout has portrait orientation of pages (avoiding double page layout). Our final dataset contained 151 reports in pdf format from 127 companies. Every GRI-compliant report has an index table indicating impact types (GRI codes) and page numbers where the corresponding information is disclosed. For every report file we manually identified page numbers with GRI codes thus generating a labeled dataset. At the next stage we used text mining package[8] to automatically extract text from the specified pages and binding it to a certain impact type (GRI code). In total we extracted text from 16360 pages.

To train and test the classification model, we used around 15500 observations of unique ‘GRI code-text’ pairs, avoiding multilabel entries. Among basic 19 codes from Social sector of GRI framework (codes and subcodes 401 to 419) we selected nine most reported codes: 401-Employment, 403-Occupational Health and Safety, 404-Training and Education, 405-Diversity and Equal Opportunity, 413-Local Communities, etc. (Table 1). We chose to focus on nine most reported codes due to insufficient observations for proper training the Random forest model. Due to manual process of code-page pair identification and the number of datapoints we did not use GRI subcodes, treating them as code of a higher level. We also had to include the unlabeled class (code 999) to distinguish only the social dimension that we focus on from other topics. The introduction of unlabeled class brings the issue of unbalanced set which can lead to misrepresentation of model performance scores. To overcome this issue in our final set, we left only 300 random observations from the unlabeled class. After train-test split we also upsampled the underrepresented classes for training dataset in order to achieve better model performance.

Table 1
Classification model dataset overview.

GRI code	GRI code name	Number of occurrences (pages)
401	Employment	323
402	Labor-Management Relations	32
403	Occupational Health and Safety	298
404	Training and Education	194
405	Diversity and Equal Opportunity	189
406	Non-discrimination	41
407	Freedom of Association and Collective Bargaining	62
408	Child Labor	28
409	Forced or Compulsory Labor	9
410	Security Practices	10
411	Rights of Indigenous Peoples	8
412	Human Rights Assessment	63
413	Local Communities	202
414	Supplier Social Assessment	106
415	Public Policy	35
416	Customer Health and Safety	84
417	Marketing and Labeling	71
418	Customer Privacy	45
419	Socioeconomic Compliance	35
999	Unlabeled	13721

We propose a pipeline (Fig. 3) to extract text and train the model to predict the impact type of a text page across all possible reporting frameworks. Particularly, we focus on social dimensions, but the pipeline could be applied to cover general, economic and environmental aspects. Fig. 4 shows Natural language processing pipeline.[9]

3.2. Term frequency–inverse document frequency (TF–IDF)

Term frequency — inverse document frequency is a dimensionality reduction technique that allows to represent text documents of variable length as numerical vectors of fixed length (Spark Jones, 1972). The value of the TF–IDF is calculated using the formula:

$$TF - IDF_i = \frac{n_i}{N} \times \log \frac{D}{d_i}$$

where n_i is the number of occurrences of a given term, N is the total number of words in the document, D is the total number of documents, and d_i is the number of documents in which the given term occurs. Words with higher TF–IDF weight considered to be more important.

[7] Except for GRI 403: Occupational Health and Safety which is 2018 version.

[8] R package used: *tabulizer*.

[9] R packages used: *workflows, recipes, textrecipes*.

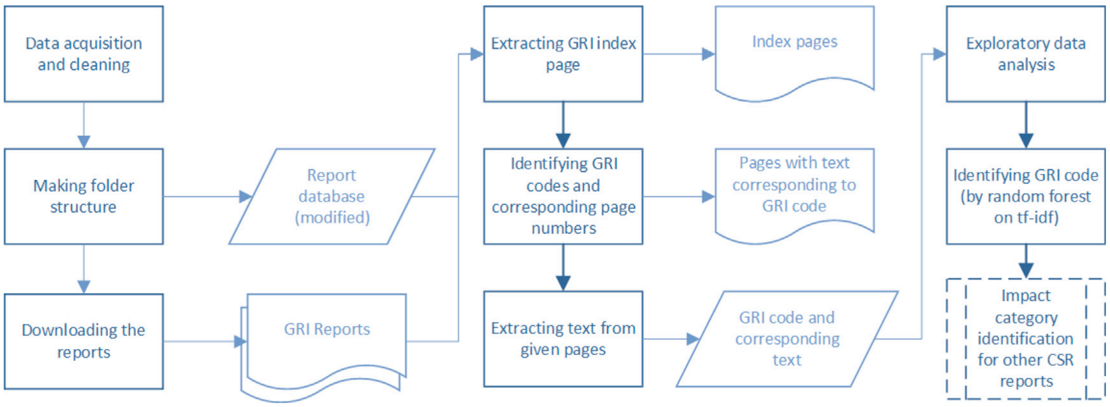
Data acquisition and cleaning
Making folder structure
Downloading the reports
Report database (modified)
GRI Reports
Extracting GRI index page
Identifying GRI codes and corresponding page numbers
Extracting text from given pages
Index pages
Pages with text corresponding to GRI code
GRI code and corresponding text
Exploratory data analysis
Identifying GRI code (by random forest on tf-idf)
Impact category identification for other CSR reports

Fig. 3. GRI reports processing pipeline.

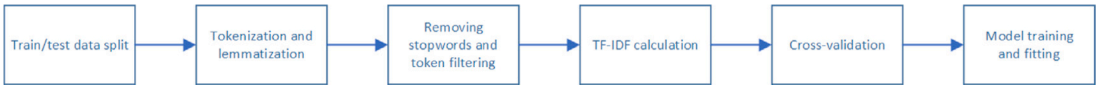
Train/test data split
Tokenization and lemmatization
Removing stopwords and token filtering
TF-IDF calculation
Cross-validation
Model training and fitting

Fig. 4. Natural language processing (NLP) pipeline.

3.3. Machine learning model

One of the most commonly used and effective machine learning methods for classifying is the decision tree. Random Forest (Ho, 1995) is an ensemble learning method for classification that operates by building multiple decision trees. Each tree is given a small set of random elements from a subsample to learn classification. After training, classification occurs by voting: the new data is assigned the class that the majority of trees vote for. A tree alone gives a low quality of classification, but their combination significantly increases the accuracy.

4. Classification model results

For our classification model we chose Random forest as one of the most commonly used and effective machine learning methods. Random Forest operates by building multiple decision trees. As a vectorization technique we utilized ‘term frequency — inverse document frequency’ (TF-IDF). This combination (Random forest on TF-IDF) showed better results than logistic regression models or combination with hashes instead of TF-IDF. The accuracy characteristic that shows the percentage of correctly identified labels was estimated as 0.65 on the test data set. For the multi-class classification model 65% of accuracy is generally considered a decent score. Other characteristics of the model could be found in Table 2 and test run heat map in Fig. 5. Cohen’s Kappa also shows the accuracy but normalized for the imbalance in the dataset; it equals to 59%. Since our dataset could not be considered as perfectly balanced, Kappa measure might be considered a better indicator for accuracy. Precision score shows that the model is 70% correct at identifying true positives among all classes. Recall score represents how complete are model results, meaning that minority classes are also correctly identified. In our case, recall (true positive rate) is 55%. F-score is a synthetic measure that is calculated based on precision and recall and represents a balance between those indicators which equals 56%. For the multi-classification model (we are having 10 classes) these are good results. We conclude that we can apply our model to tackle real-world scientific problems.

So far this paper has focused on the development of supervised machine learning method for text-mining of CSR reports — Standard-based Impact Classification method (SBIC). The following Research application section we provide an example of how our method can be used to tackle real-world scientific problems. We narrow down the general approach to studies of the link between CSR and innovation to investigate underlying mechanisms of this interconnection.

Table 2
Random Forest model train and test run indicators.

	Metric	Train run	Test run
1	Accuracy	0.93	**0.65**
2	Cohen’s Kappa	0.93	0.59
3	Precision	0.93	0.70
4	Recall	0.94	0.55
5	F-measure	0.93	0.56

5. Research application

In this section we show the applicability and potential of our developed Standard-based Impact Classification (SBIC) method. We apply SBIC method to study in more detail the interconnection between corporate social responsibility (CSR) and innovation capacity of a firm, namely how the elements of the social dimension of CSR are affecting innovative capacity. Social dimension of CSR and innovation interaction is crucial for creating a positive impact on society, fostering sustainable business practices, and gaining a competitive advantage in a rapidly changing business environment. This link established properly enables companies to align their goals with societal needs, cultivate stakeholder trust, and contribute to the well-being of local communities they function in.

5.1. Literature review on CSR-innovation interconnection

Much research on the topic of general interrelation between CSR and innovation activities of a company was published during past two decades. According to Resource-based view theory (Hart, 1995) both CSR and innovation capacity represent the intangible resources that can provide competitive advantage (Gallego-Álvarez et al., 2011). The importance of studying CSR-innovation interrelation becomes evident if we provide context for this link. From sustainability perspective, both innovation and CSR influence firms’ well-being and its environment (Khan et al., 2021). From corporate performance standpoint the models that include either CSR or innovation may become upwardly biased (Padgett and Galan, 2010). From a practical managerial perspective it is good to know what CSR activities can increase innovation capacity of a firm (García-Piqueres and García-Ramos, 2021). In other words the necessity of studying innovation and CSR link may come from different sources depending on the topic and the aim of

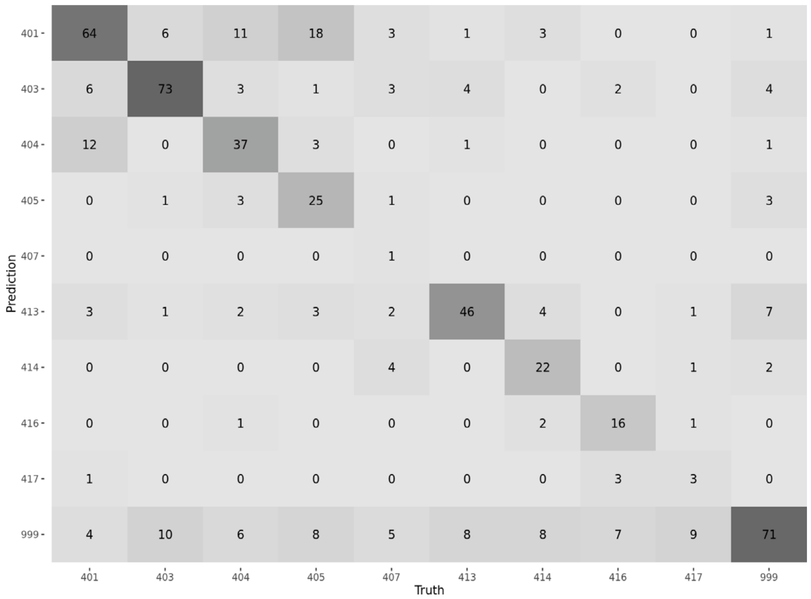

Fig. 5. Heat map of testing set.

particular research, but a literature gap in studies of joint effect of CSR and innovation on company and its stakeholder well-being still exists (Martinez-Conesa et al., 2017).

Ratajczak and Szutowski (2016) found the lack of scientific consensus on aspects of CSR and innovation performance relationship on the basis of a systematic literature review of 24 publications during 2000–2014. Authors show that determinants of this relationship vary greatly from study to study. Among their conclusions we can denote the uncertainty of relationship direction. Authors differentiate 4 main types of relationships present in literature: CSR affecting innovation, innovation affecting CSR, CSR and innovation affect each other and functional (undefined direction) relationship. We provide recent updates in the field focusing on CSR affecting innovation and their mediating role on firm's performance or value.

CSR as a driver for innovation. According to Audretsch et al. (2016) contemporary literature identifies three main ways in which CSR activities influence innovativeness of a company: first, through higher engagement of stakeholders; second, by creating new business opportunities; and third, by modifying organizational structure to become more propitious for innovation. Bocquet et al. (2013) distinguish companies by intensity of CSR practices adoption into strategic and responsive profiles. Authors found that strategic CSR profiles foster both process and product innovation, while responsive profiles can create barriers to innovation. Luo and Du (2015) conclude that CSR can be a catalyst for innovation. Cai et al. (2023) distinguish business and philanthropic CSR activities and proved philanthropic CSR has a positive influence on open innovation. Positive correlation was also found by Mendes et al. (2023). Cook et al. (2019) show that firms with higher CSR performance generate more patents and patent citations. Shahzad et al. (2020) found different CSR dimensions positively influence sustainable development and green innovation.

Mediating role of CSR and innovation on firm's performance or value. MacGregor and Fontrodona (2008) suggest that CSR and innovation should be integrated into one another and should form a virtuous circle. In Spanish context of small and medium-sized businesses Martinez-Conesa et al. (2017) finds partial mediation effect of innovation performance on the relationship between CSR and firm performance, while Becerra-Vicario et al. (2023) argue for the mediating role of CSR in innovation influencing the performance. Ruggiero and Cupertino (2018) argue that investment in innovation activities enables firms to respond to changes in their environment faster and serves as a mediator between financial performance and CSR. Results of Kraus et al. (2020) show that CSR is positively correlated to green innovation and that both increase environmental performance. To similar conclusions come Simmou et al. (2023) also suggesting mediation effect of green innovation on the CSR influencing environmental performance. Results of Gangopadhyay and Homroy (2023) show that country CSR regulations can create indirect incentives for innovation activities. Taking a dynamic perspective on Chinese listed firms Hu and Zhang (2023) found that both CSR and Innovation have influence on firm market value.

In this subsection we show that the topic of CSR-innovation link has a relatively long history of research. However, the mechanisms of this interconnection still require investigation. For the demonstration purposes we focus on investigating the first relationship type: CSR as a driver for innovation. We follow the works of Broadstock et al. (2020), García-Piqueres and García-Ramos (2021) where they use innovation as response (dependent) and CSR as an explanatory (independent) variable. We reckon that applying our method to replicate some findings of previous studies can provide the magnification needed for further theory development.

5.2. Research question

To discover the mechanisms of CSR–innovation interconnection we need to improve the resolution of both concepts. In Global Reporting framework the term impact refers to the effect an organization has or could have on the economy, environment, and people, including effects on their human rights, as a result of the organization's activities or business relationships. Social impacts refer to the impacts on individuals and groups, such as communities, vulnerable groups, company stakeholders or society.[10]

Innovation means finding new ways to use available resources and endow these ways with economic value for the benefit of the organization and its stakeholders, thus creating new resources (Drucker, 2015). The result of the innovative process '...includes both gradual technical change and discrete leaps in technical opportunities' (Lundvall, 2016, p. 20). Innovation can be distinguished by form (product, market, process), sources (closed or open) and scopes of change (radical or incremental) (Halkos and Skouloudis, 2018). Authors of the first study

[10] See Section 2.1 in Anon (2023).

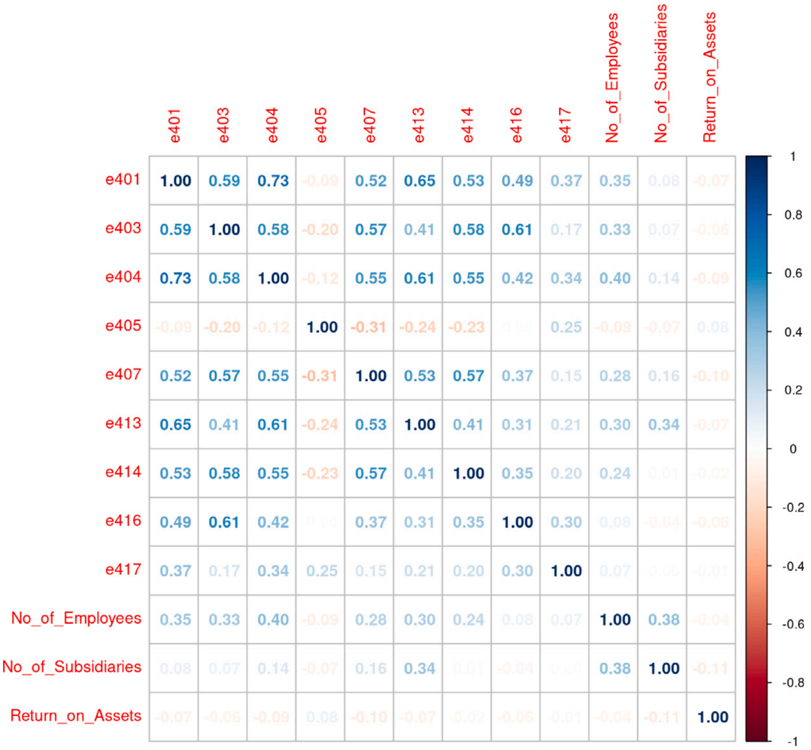

Fig. 6. Correlation matrix of independent variables.

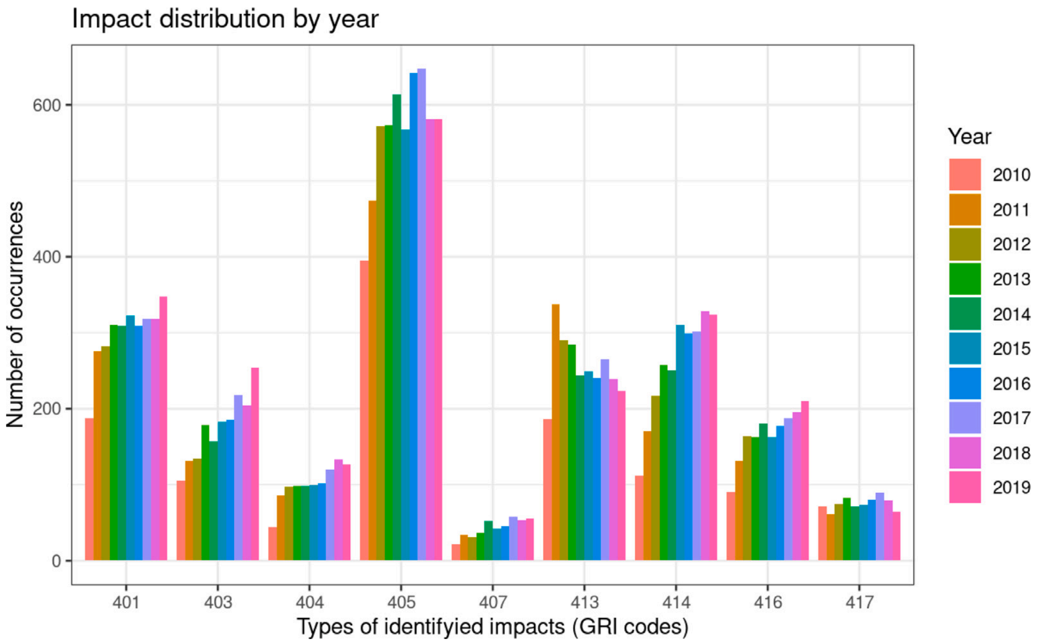

Fig. 7. Total number of pages on social impacts reflecting in German listed companies from 2010–2019.

we are following (García-Piqueres and García-Ramos, 2021) differentiate radical and incremental innovation for their regression models. Radical innovation involves a higher degree of knowledge (Dewar and Dutton, 1986) and can transform a whole industry by significantly changing products, services, or processes. Incremental innovation optimizes operations and performance through relatively small continuous improvements (Freeman and Soete, 2017; Ettlie et al., 1984). Radical innovation can be also split into several categories: organizationally, industry-, user- and technologically radical (Katila, 2000). García-Piqueres and García-Ramos (2021) use firm's turnover as a measure for innovation performance: in case of radical innovation it is a turnover related to products and services that are new to the market, in case of incremental — that are new to the firm.

We use the same definition of innovation capacity (as used by authors of the second study we are replicating the model from) Broadstock et al. (2020): a continuous improvement of the overall capability of firms to generate innovation for developing new products to meet market needs. Authors use technological change levels (i.e. movements of firms's technological production frontier) as an indicator for

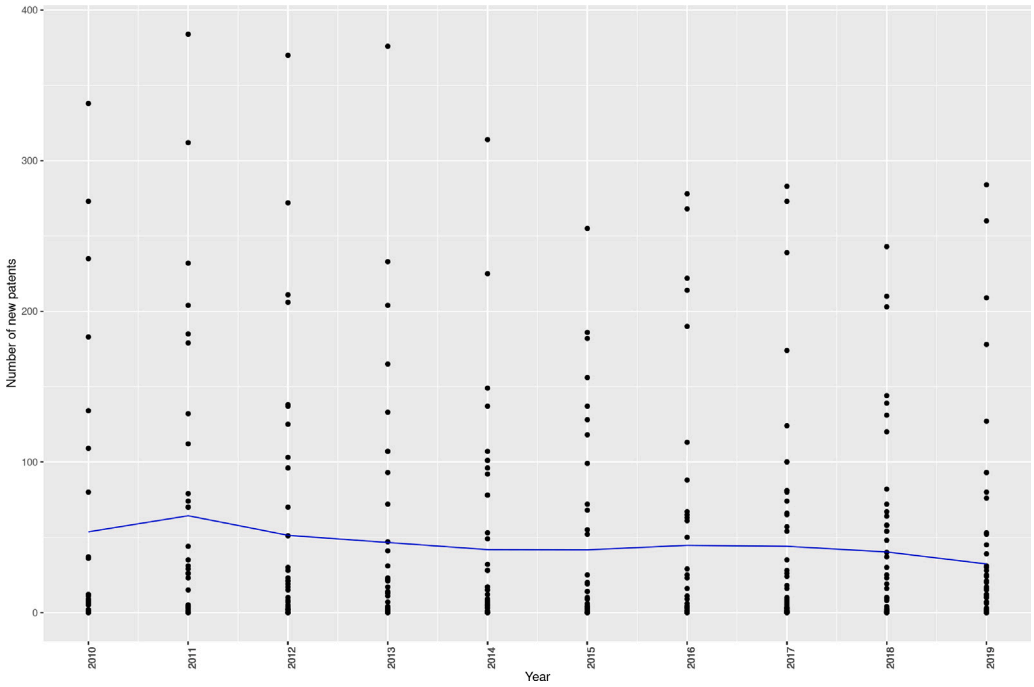

Fig. 8. Patents ownership of 69 German listed companies through 2010–2019.

Table 3
Summary statistics.

Statistic	N	Mean	St. Dev.	Min	Max
No of new Patents a year	510	241.565	577.695	0	3,805
401-Employment	510	5.661	5.918	0	41
403-Occupational Health and Safety	510	3.245	4.893	0	29
404-Training and Education	510	1.914	2.284	0	13
405-Diversity and Equal Opportunity	510	11.000	9.240	0	45
407-Freedom of Association	510	0.831	1.331	0	7
413-Local Communities	510	4.898	8.562	0	90
414-Supplier Social Assessment	510	4.939	6.603	0	53
416-Customer Health and Safety	510	3.047	4.907	0	38
417-Marketing and Labeling	510	1.449	2.048	0	15
No of employees	510	64,030.87	103,854.40	0	580,610
No of subsidiaries	510	537.53	896.65	7	4,922
Return on assets	510	1.871	6.784	−16.370	79.270

innovation capacity and argue that this proxy can also account for non-technological innovation.

Other common proxy measures for innovation and innovation capacity are R&D expenditures and patents or patent applications (Fagerberg et al., 2005; Pavitt, 1985; Acs and Audretsch, 1989; Furman et al., 2002). For our purposes we will use patent measure as a proxy for innovation capacity. This means that the estimates of innovation capacity are of technological nature and inherit all the cons of patent-related measures such as: some types of technology are not patentable, patents do not have direct link to commercially viable products, patents used to prevent a competitor from patenting etc. (Kleinknecht et al., 2002). The significance and value of patents can vary greatly (Jaffe et al., 1993) that is why more in-depth research is needed to establish connection between specific types of innovation and CSR which goes beyond the scope of this method demonstration task.

In this section we look at the CSR part of this interrelation and attempt to push forward the boundaries of current research with fine-grained studies of it. Turban and Greening (1997) marks out 5 dimensions of CSR: Corporate social performance, Community relations, Employee relations, Environment, Product quality and Treatment of women and minorities. But currently the most commonly used approach is Triple Bottom, which splits CSR into social, economic and environmental dimensions (García-Piqueres and García-Ramos, 2021). As a proof of concept in the current study we aim to fine-grain only the *social dimension* by identifying specific types of social impacts according to Global Reporting Initiative standard and at investigating their link to innovation capacity of a firm. We focus only on social dimension for several reasons. First reason is the limitations of the current pre-trained language model. Second, we find social dimension of CSR the least studied (compared to environmental) and one that requires new methods of assessment. Third, the theoretical mechanisms found in literature that explain the CSR–innovation link focus specifically on social aspects. We assume that detailed overview obtained by the means of SBIC method can provide empirical evidence needed to justify these mechanisms. So, we propose the following research question:

What types of social impacts have a stronger effect on innovation capacity of a firm?

To answer this research question, we replicate the CSR–innovation interaction models by García-Piqueres and García-Ramos (2021) and Broadstock et al. (2020). With our method developed we obtain more fine-grained types of social impacts that correlate with innovation capacity of a company.

5.3. Hypothesis

At a very abstract level we can theorize that CSR activities create incentives for technology resources to improve firm innovation capacity (Costa et al., 2015) or regard CSR as a type of investment, which becomes a source for product and service development (Padgett and Galan, 2010). However, very few studies investigate the underlying mechanisms of this influence. Currently, we can single out three main types of such mechanisms found in literature.

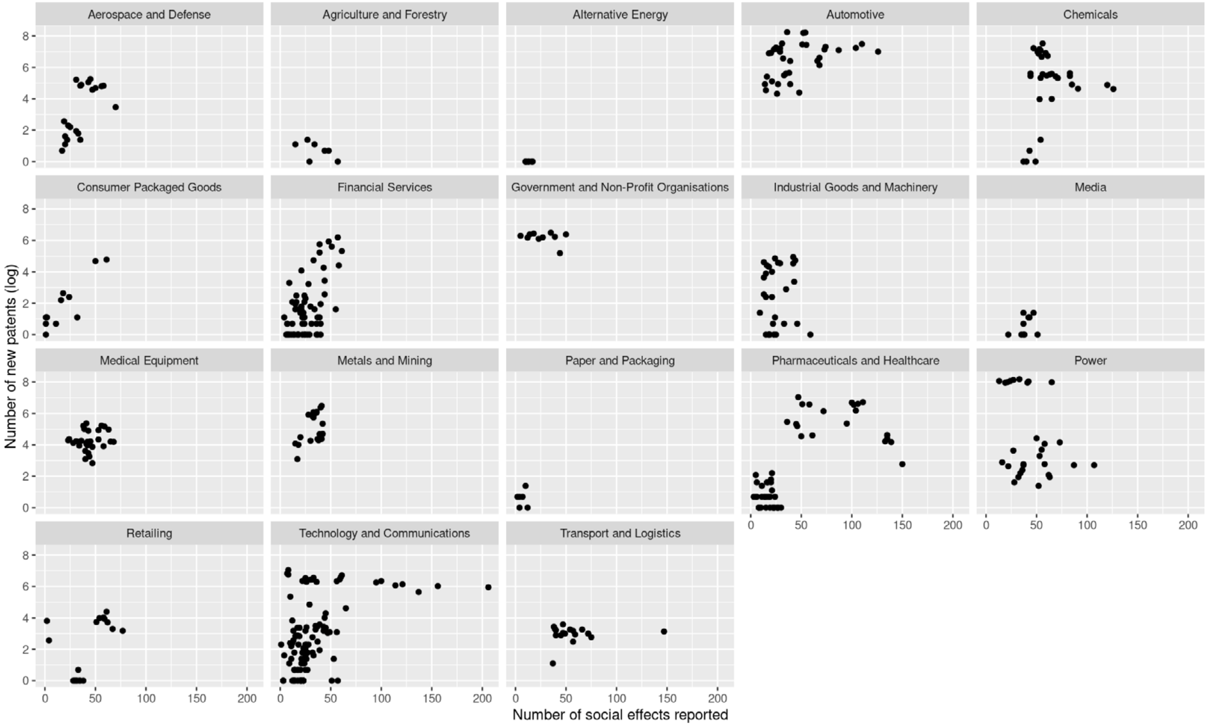

Fig. 9. Number of patents vs number of reflected social impacts grouped by Primary Sector (69 German listed companies through 2010–2019).

Table 4
Panel Poisson regressions results with variable time lag.

	Dependent variable: Number of new patents in possession (by year)				
	Model 1	Model 2	Model 3	Model 4	Model 5
Variable lag	0-year	1-year	2-year	1-year	1-year
Individual effects	Random	Random	Random	Random	Fixed
Data balancing	Unbalanced	Unbalanced	Unbalanced	Balanced	Balanced
401-Employment	−1.53***	−0.96***	−0.16	−2.69***	−2.69***
403-Occupational Health and Safety	−0.78***	−0.77***	−2.46***	0.73*	0.73*
404-Training and Education	1.09***	2.99***	0.71**	7.83***	7.82***
405-Diversity and Equal Opportunity	0.80***	−0.39***	−0.36***	−0.96***	−0.96***
407-Freedom of Association	3.58***	4.85***	−0.78	−2.82***	−2.82***
413-Local Communities	0.52***	0.56***	1.61***	1.81***	1.81***
414-Supplier Social Assessment	−0.84***	−0.51***	0.54***	−3.28***	−3.28***
416-Customer Health and Safety	0.75***	0.61***	−1.43***	0.32	0.31
417-Marketing and Labeling	2.80***	3.25***	−0.18	0.99***	1.00***
No of employees	0.00**	0.00*	0.00*	0.00***	
No of subsidiaries	−0.01	−0.01	−0.02	−0.04	
Return on assets	6.09	13.28*	4.79	−6.54	
Log-Likelihood	−10145.68	−8273.16	−7615.22	−4374.86	−4155.02
Num. obs.	510	423	370	268	268

* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$
Notes: Coefficients are displayed as odds ratio in percentage form.

The first type of CSR–innovation mechanism explains the connection through a change in work engagement. Firms engaged in CSR-activities are more appealing to more innovative candidates (Nazir and Islam, 2020; Turban and Greening, 1997) and foster innovative behavior (Paruzel et al., 2023). It is assumed that ‘social CSR practices seem to be related to the recruitment of the most innovative people, more radical or disruptive innovations can be developed for these kinds of firms’ (García-Piqueres and García-Ramos, 2021). Nazir and Islam (2020) demonstrate through a survey that CSR activities addressing psychological needs of employees enable them to execute innovative approaches in their jobs. Similar methods drive Gaudêncio et al. (2019) to similar findings. Roszkowska-Menkes (2018) marks out the role of openness and collaborative culture as a facilitator for employees' increased absorptive capacity that is directly connected to firm-level absorptive capacity and innovation capacity.

H1. Social impacts disclosure that refers to employee-related practices (i.e. impacts 401–405 in GRI classification) has positive influence on innovation capacity of a firm

The second type of CSR–Innovation mechanism utilizes the networking dimension of innovation (Fagerberg et al., 2005, 151) to explain the link between innovation and CSR. García-Piqueres and García-Ramos (2021) provide the following reasoning in this regard: established CSR activities promote interactions with other agents, that can provide access to knowledge of social responsible practices, which in turn increases the knowledge base of initial firm. This knowledge improvement serves as a resource for the innovation process.

H2. Social impacts disclosure that refers to interaction with external agents (i.e. impact 414 in GRI classification) has positive influence on innovation capacity of a firm

The third type of CSR–Innovation mechanism can occur if we consider that CSR and innovation activities are competing for firm resources: the management needs to make decision which of those to promote. Then we should see a negative correlation between CSR disclosure and innovation as for example in the study by Gallego-Álvarez et al. (2011).

H3. Social impacts disclosure has negative influence on innovation capacity of a firm

In the next subsection we apply our SBIC method to extract social impacts (as a subset of GRI classified impacts) that are reflected in CSR reports and run panel regression to check the hypothesis.

5.4. Panel regression

5.4.1. Dataset

For replicating the findings of García-Piqueres and García-Ramos (2021) and Broadstock et al. (2020) and testing our method of Standard-based Impact Classification (SBIC) we build panel regression using two major datasets. The first one was obtained through Orbis IP database for German firms for 2010:2019 timespan (Fig. 8). To construct the second dataset, we merged the patent dataset with publicly listed German firms (those that are subject to NFRD reporting by legislation) and manually downloaded CSR reports for 2010:2021 time period.[11] The additional two years of CSR reporting is needed to perform a regression model with 2-year lag. Our final version of the second dataset contained 69 German listed firms, 510 reports having 76876 pages in total. We then apply our classification (SBIC) model to identify and count the number of pages containing social impacts of interest (Fig. 7). The number of new patents obtained within a year and total number of social impacts reported within a year (grouped by primary sector) are presented in Fig. 9.

5.4.2. Variables

The dependent variable is the firm's innovation capacity (Szeto, 2000) that in our particular case is represented via proxy measure as the number of patents a firm obtains in a year. This is a simpler version for innovation capacity proxy than the firm's technological change levels used by Broadstock et al. (2020).

Independent variables are the number of social impacts reflected in CSR reports that are identified by our Standard-based Impact Classification (SBIC) method. This type of scoring follows topic scoring of reports done by Pencle and Mălăescu (2016). Each independent variable represents the number of pages that correspond to one of 9 types of social impacts: 401-Employment, 403-Occupational Health and Safety, 404-Training and Education, 405-Diversity and Equal Opportunity, 407-Freedom of Association and Collective Bargaining, 413-Local Communities, 414-Supplier Social Assessment, 416-Customer Health and Safety, 417-Marketing and Labeling. We control for company related effects by including size and financial characteristics (Padgett and Galan, 2010), namely number of employees, number of subsidiaries and return on Assets. Summary statistics is presented in Table 3. Analyzing correlation matrix of independent variables (Fig. 6) we can see that there are some moderate positive relationships, except for variable e405, which shows slight negative relationships. We conclude that no multicollinearity issues should be specifically addressed in this case.

5.4.3. Regression model

For our empirical analysis we use long panel data. We follow the regression model by Broadstock et al. (2020) provided in expression №14 and García-Piqueres and García-Ramos (2021) in expression №4. Each observation corresponds to a firm in a given year. The use of such type of data helps control for unobserved heterogeneity. Our dependent variable is count data (number of patents is non-negative integer discrete type), therefore, we use Poisson regression as our major analysis tool. By comparing mean and variance we can conclude that we are having overdispersion in our models, hence violating the Poisson variance assumption (Wooldridge, 2010, 725). This violation needed to be taken into consideration when interpreting the results as standard errors and significance levels may contain wrong estimates. For control purposes we present OLS regression results in Appendix (Table 6).

We chose random effects model, because the differences between firms might influence their innovation capacity. It also enables to include time-invariant variables as regressors. Our regression equation is represented by the following formula:

$$ln(y_{it}) = \beta_0 + \beta * X_{i,t-lag} + \alpha_i + \mu_{it}$$

where $ln(y_{it})$ is a natural logarithm of new patents obtained per year, β_0 is a constant, β is vector of coefficients for independent variables, $X_{i,t-lag}$ is a vector of explanatory and control variables measured at time t lagged by 0, 1 or 2 years, α_i is unobserved heterogeneity (individual effects) constant over time, μ_{it} is a time-varying idiosyncratic error.

Hausman Test shows that its null hypothesis cannot be rejected, which means that we can use a random effects model. F test and Lagrange Multiplier Test (Breusch–Pagan) for unbalanced panels showed the existence of individual effects. Thus, in random effects model we control for time-invariant size variables to avoid the risk of omitted variable bias.

5.5. Regression results

Table 4 contains results of our Poisson regression modeling. For interpretative purposes we took the exponent from initial coefficients of Poisson regression models and converted results to percentage form. The presented numbers illustrate the influence of factor unit change on the dependent variable in percents. Applied to our analysis this means the influence of change in number of pages on the quantity of patents obtained in a year.

Our findings show that we cannot definitely affirm the positive correlation between the quantity of social impacts reported and number of patents a company obtains, which are used as a proxy for produced social impacts and innovation capacity, respectively. Model 1 is a base line model to compare variations over time. We have two overlapping time periods: one is for patent acquisition and actual technology implementation and the second one is for producing social effect and reporting on it. Patent time period can also be considered not precise: a company firstly uses a technology and then patents it. CSR reports often come as double year reports and impacts have a time-prolonged effect.

[11] Usually reports are published the next year after specified reporting period, meaning that reports for 2021 became available only in 2022.

Table 5
Classification model dataset description.

	Country	Publication year				GRI claim type			Descriptive summary		
		2017	2018	2019	2020	Referenced	Comprehensive	Core	Number of reports	Number of pages	Average report length
1	Argentina	0	1	1	0	0	0	2	2	467	234
2	Australia	0	1	2	1	0	0	4	4	175	44
3	Austria	0	4	1	2	0	0	7	7	930	133
4	Bahrain	0	0	1	1	0	0	2	2	144	72
5	Belgium	0	1	0	1	0	0	2	2	165	82
6	Brazil	0	1	0	1	0	0	2	2	177	88
7	Canada	0	0	1	0	0	0	1	1	48	48
8	Chile	0	1	0	0	0	1	0	1	97	97
9	Cyprus	0	1	1	0	0	0	2	2	95	48
10	Egypt	0	0	0	2	0	0	2	2	378	189
11	Finland	0	1	2	1	1	0	3	4	287	72
12	France	0	0	1	0	0	0	1	1	38	38
13	Germany	0	1	3	8	0	0	12	12	1,214	101
14	Guatemala	0	0	1	0	0	0	1	1	105	105
15	Hong Kong	1	0	1	2	0	0	4	4	280	70
16	Hungary	0	1	2	0	0	0	3	3	360	120
17	India	0	2	3	0	0	0	5	5	462	92
18	Italy	0	1	2	3	0	1	5	6	473	79
19	Japan	0	0	1	0	1	0	0	1	32	32
20	Jordan	1	0	0	0	0	1	0	1	86	86
21	Lebanon	0	0	1	0	0	0	1	1	75	75
22	Malaysia	0	0	0	1	0	0	1	1	45	45
23	Netherlands	0	0	2	0	0	0	2	2	384	192
24	New Zealand	0	0	1	0	0	0	1	1	158	158
25	Norway	0	1	1	2	0	0	4	4	426	106
26	Poland	0	0	1	0	0	0	1	1	93	93
27	Russian Federation	0	1	0	0	0	0	1	1	109	109
28	Singapore	0	5	3	1	2	0	7	9	1,194	133
29	Slovenia	0	1	0	1	0	0	2	2	342	171
30	Spain	0	0	1	2	0	0	3	3	288	96
31	Sri Lanka	0	0	1	0	0	0	1	1	176	176
32	Sweden	1	4	5	4	0	2	12	14	1,670	119
33	Switzerland	0	1	4	3	0	1	7	8	751	94
34	Taiwan	0	1	2	0	0	1	2	3	453	151
35	Thailand	0	0	2	1	0	0	3	3	166	55
36	Turkey	0	2	0	1	0	0	3	3	226	75
37	United Arab Emirates	0	0	1	3	0	0	4	4	341	85
38	United Kingdom	1	1	1	2	1	0	4	5	948	190
39	United States of America	4	4	9	5	1	2	19	22	1,698	77
	Total	8	37	58	48	6	9	136	151	15,556	103 (av.)

At this point of time, we do not possess exact data with timestamps of technology usage and social impact production, but consideration of these time periods is important in our variables lag interpretation. Roughly, 0-year lag (Model 1) should be considered as up to one year lag, 1-year lag (Model 2) as half to two year lag and 2-year lag (Model 3) as one and a half up to three year lag.

Our first hypothesis stating that employee-related practices (i.e. impacts 401–405 in GRI classification) has positive influence on innovation capacity of a firm is mostly rejected. Through *Model 1* to *Model 3* 401-Employment disclosure reduces its negative impact (−1.53% same year, −0.96% next year) on innovation capacity. This code reveals information on new employee hires and employee turnover, benefits provided to full-time employees that are not provided to temporary or part-time employees and parental leave. Code 403-Occupational Health and Safety also has increasing negative impact from −0.78% up to −2.46% after 2 years. Code 404-Training and Education revealing information on programs for upgrading employee skills has a positive impact of 2.99% and 7.83% in case of a balanced dataset. Code 405-Diversity and Equal Opportunity also shows a slight negative impact on innovation capacity, but less than 1% in all models. Results in the *Model 2* speak in favor of existence of mechanism that positively links innovative capacity of a firm only to education and training related practices. While increased attention to other employee-related practices execute negative influence on innovation capacity.

Our second hypothesis which indirectly investigates the influence of networking with socially responsible agents and thus increasing innovation capacity through knowledge spillovers is mostly refuted. Through all regression models (except 2-year lag model) we can see the evidence of code 414-Supplier Social Assessment having negative influence up to −3.28%. The direction of influence for this code is also supported by panel linear regression Individual Random Effects models (Table 6 in Appendix). However, this code may contain information on subcode 414-2: Negative social impacts in the supply chain and actions taken, which in reality may reverse the direction of influence to positive. A more precise classification model or qualitative analysis is needed here.

Third hypothesis is partly supported. We can see that codes 401, 403, 414, 416 display negative influence upon innovation capacity, which indirectly second the conclusions of Gallego-Álvarez et al. (2011) that CSR and innovation activities can compete for managerial resources. This resource competition mechanism can explain why only code 404-Training and education showed positive correlation when testing first hypothesis.

Table 6
Panel linear regression results with 1 year lag.

	Dependent variable: Number of patents in possession (by year)					
	Pooled OLS (unbalanced)	Pooled OLS (balanced)	Time FE (unbalanced)	Time FE (balanced)	Individual RE (unbalanced)	Individual RE (balanced)
401-Employment	−8.187 (7.422)	−5.100 (4.618)	−4.858 (7.865)	−3.869 (4.955)	−4.978 (3.419)	−5.544* (2.937)
403-Occupational Health and Safety	−23.995*** (8.664)	1.427 (4.337)	−8.551 (9.151)	0.00004 (4.751)	−1.390 (5.383)	3.150 (4.207)
404-Training and Education	9.371 (18.326)	5.398 (10.079)	39.144** (19.709)	19.290* (11.275)	10.588 (7.318)	7.326 (6.021)
405-Diversity and Equal Opportunity	1.933 (3.099)	2.096 (1.785)	2.222 (3.351)	3.689* (1.969)	0.102 (1.844)	−0.840 (1.355)
407-Freedom of Association	77.113*** (27.896)	30.760* (16.332)	77.142** (30.166)	45.598** (17.647)	25.371* (13.161)	−3.795 (10.155)
413-Local Communities	−11.438*** (4.405)	0.288 (2.161)	−9.224** (4.445)	−0.101 (2.216)	0.781 (2.050)	3.555** (1.404)
414-Supplier Social Assessment	11.392** (5.317)	−6.122** (2.709)	8.944 (5.685)	−4.389 (2.963)	−0.141 (2.798)	−3.311 (2.063)
416-Customer Health and Safety	26.510*** (7.301)	11.965*** (3.569)	15.094* (7.749)	13.409*** (3.848)	1.509 (4.475)	0.016 (3.258)
417-Marketing and Labeling	−3.282 (14.095)	4.449 (6.674)	−10.141 (15.362)	2.804 (7.423)	8.074 (6.139)	2.487 (4.288)
No of Employees	0.002*** (0.0003)	0.002*** (0.0003)			0.002*** (0.001)	0.002** (0.001)
No of Subsidiaries	0.036 (0.033)	−0.041** (0.016)			0.001 (0.091)	−0.059 (0.046)
Return on Assets	7.655 (5.503)	−15.335*** (4.761)			14.731 (14.420)	−19.025 (14.492)
Constant	11.683 (55.493)	40.675 (34.224)			79.592 (91.923)	118.929 (73.105)
Observations	423	243	423	243	423	243
R^2	0.206	0.348	0.073	0.193	0.048	0.081
Adjusted R^2	0.182	0.314	0.034	0.132	0.020	0.033
F Statistic	8.850*** (df = 12; 410)	10.228 (df = 12; 230)	3.559** (df = 9; 405)	5.995** (df = 9; 225)	19.729*	20.342*

* $p < 0.1$
** $p < 0.05$
*** $p < 0.01$

Other findings that are not covered by our hypothesis also have different influence direction upon innovation capacity of the firm and require theoretical explanation of possible underlying mechanisms. Code 407-Freedom of Association and Collective Bargaining that reveals information on operations and suppliers, in which the right to freedom of association and collective bargaining may be at risk shows up to 4.85% positive influence (Model 2), but has a negative direction if we take balanced data. Focus on code 413-Local communities gives a positive impact on innovation capacity through all models. Code 417-Marketing and Labeling disclosures requirements for product and service information and incidents of non-compliance concerning product and service information and marketing communications. This code also provides relatively significant amount of influence up to 3.25% on innovation capacity of a company. The direction of the influence corresponds to the results in panel linear regressions in Table 6 in Appendix.

In summary, Research application section illustrates the potential of the Standard-based impact classification method for producing in-depth analysis of CSR reports. In combination with regression analysis SBIC method leads us to new findings that can lay the foundation for further theoretical contributions.

6. Method discussion and limitations

Our proposed Standard-based Impact Classification (SBIC) method of analysis of CSR reports incorporates all the advantages and disadvantages of standardization process in general and Global Reporting Initiative in particular. The major aim of this method is to bring comparability of reports irrespective of its framework. By bringing every report to GRI standard our method removes the possible research subjectivity in topic modeling result interpretation. Unlike topic modeling or dictionary approach SBIC method has certain future-proof feature: the method can get adopted to GRI standards change, due to incorporating new reports in learning dataset. The reports made by new standard will reflect it on index pages thus providing updated labeled dataset to train the model on. Our SBIC method provides fine-grained information on types of social impact found in the reports, it helped to replicate the findings of several researches and revealed the new yet theoretically unexplained phenomena.

In comparison with LDA analysis SBIC method lacks the issues of manual topic identification; hence, simplifying the result interpretation stage. A predefined thematic dictionary approach is unperceptive to

Table 7
The methods of social impact assessment compiled from Cerioni and Marasca (2021).

Method category	Method name	Description	References (see Cerioni and Marasca (2021))
Process Methods Ethics, 2016	Best Available Charitable Option (BACO)	Aims to quantify the company's actions that have an impact on the company, related to a given investment (Leverage, technology and efficiency of the company)	Acumen Fund, 2007 Zamagni et al. 2015
	Global Reporting Initiative (GRI)	Proposes guidelines for social and sustainability reports, listing what indicators are considered as best practices to monitor the three dimensions of performance (economic, environmental and social)	Lamberton, 2005
	Impact Reporting and Investment Standards (IRIS)	Provides standard indicators of social, environmental and financial performance for the definition, monitoring and reporting of investment capital performance	International Trade Center, 2011
	Global Impact Investing Rating System (GIIRS)	Aims at assessing the social as well as environmental impact of companies and actives investment funds in emerging or developed markets.	Works, 2014
Impact Methods Bengo et al. 2015	Business Impact Assessment (BIA)	Allows companies to assess themselves in terms of sustainability and transparency through the compilation of a questionnaire.	Grimes et al. 2018
	Theory of change	The aim is to plan and evaluate projects that promote social change through the participation and involvement of stakeholders.	Kail and Lumley, 2012 Taplin and Clark, 2012
	Analysis of the counterfactual	Based on the comparison between two realities, with similar characteristics, that differ in whether or not they have implemented a given project	Bonaga, 2017; Bellucci et al., 2019
	Mixed-method	Involves the administration of qualitative-quantitative research tools for the organization and its stakeholders	Bonaga, 2017; Bellucci et al. 2019 Venturi, 2017
	Measuring Impact Framework	The World Business Council for Sustainable Development (WBCSD) has devised this model with the aim of helping businesses understand the effect of their social contribution.	Ethics, 2016
	Social Impact Assessment (SIA)	Involves three phases: - definition of the objective in terms of the social value of the company, - fundamental to the desired results of the enterprise; - quantification of social value by listing the three main social indicators most closely related to the social results of normal business operations, monetization of the social impact value that the company aims to create over the next 10 years	Esteves et al. 2012
	Ongoing Assessment of Social Impacts (OASIS)	Is a comprehensive and continuous evaluation system	Olsen and Galimidi, 2008; Bengo et al. 2015
Methods of monetization Socialis, 2017	Cost benefit analysis (CBA)	A method of economic analysis in which the social costs and impacts of an investment are expressed in monetary terms and then evaluated and compared by one or more measure: Net present value or Cost-benefit ratio	Clark et al., 2004
	Social Return on Investment (SROI)	The final result of the input-outcome model through which one can communicate how much economic, social and environmental performance exists for every euro invested in a project or activity. The basic idea is to build not only a metric but also a methodology for determining the value of the social, economic and environmental outcomes generated by a company	Manetti, 2014; Buffalo et al. 2017; Bellucci et al., 2019 SROI Network, 2012 Yates and Marra, 2016 Maier et al. 2015 Ethics, 2016

topic changes through time. New topic introduction into a dictionary requires serious effort to level out possible subjectivity issues (Molecke and Pinkse, 2017). Our methodology avoids the disadvantages of LDA or dictionary approaches.

CSR reports are also regarded as another tool of public relations, so companies tend to overstate positive impacts. Meaning that our method also tends to capture only positive side of things. This positive trend of CSR reports and wholesome skepticism may also rise the question of reliability and credibility of such a report. Combinations with other sources of information could be used to benchmark the real activities of a company, which may lead to a push for a better standard of reporting and resolving the issue of CSR decoupling.

Manipulation with different training material will yield different results, which should be taken into account. But we can also use this as an advantage, for example, if we want to fine-tune industry or country specific classification model.

Unfortunately, collecting, processing and verifying a training dataset at this point of time is a semi-automatic process, meaning that many procedures are done manually. It is a very time-consuming process, especially given that although standardized reports are in pdf and still represent highly unstructured data. Currently, our training data is limited only to social impacts, but we look forward to incorporate other CSR dimensions in future research. The quantity and the quality of available training data is the key to high performance machine learning models. The mandatory reporting using eXtensible Business Reporting Language (XBRL) could be one of possible solutions for increasing the quality of classification model and improving the transparency and credibility of CSR reporting practices.

The idea of applying eXtensible Business Reporting Language (XBRL) to non-financial disclosure is not new. The realization of eXtensible Business Reporting Language (XBRL) for financial reporting has been already implemented and used on mandatory basis, for example, in Securities and Exchange Commission (USA). During a recent decade several researches have suggested a logical extension of XBRL taxonomy to cover the non-financial reporting as well (Shahi et al., 2012, 2014; Knebel and Seele, 2015; Seele, 2016). The implementation of XBRL for sustainability reporting brings structure to firms publishing CSR data. No doubt that such standardization and unification can positively influence the scale of reporting and improve transparency, hence credibility and comparability of non-financial statements. It should be taken into account that any standardization has some downsides, such as a possibility to lose some country-, industry- or company-specific, peculiar data. Nevertheless, a theoretical CSR framework should be present inside any particular implementation of XBRL for sustainability. We share the point of view that extending XBRL to CSR will be the next evolutionary step in non-financial disclosure, but currently we need to address flaws and hindrances of existing frameworks.

7. Conclusion

This study contributes to the research field of text-mining Corporate Social Responsibility (CSR) reports. We have developed the Standard-based Impact Classification (SBIC) method for extracting social impact information from non-financial statements, addressing the need for accessible mechanisms to quantitatively evaluate the social impacts generated by companies. Our proposed method combines the flexibility of the LDA approach with the rigidity of a pre-defined topic dictionary approach. In the event of changes to GRI standards, training on new text datasets ensures the model's accuracy remains sufficient. While we focus on the social dimension in this study, the proposed pipeline can be easily expanded for future research to encompass general, economic, and environmental dimensions.

We demonstrated the potential of the SBIC method for investigating the relationship between social dimension of CSR and innovation. Our results partially support previous findings while distinctly highlighting under-theorized connections and revealing gaps for future research. Our fine-grained impact identification reveals that all three previously described mechanisms of CSR activities influencing innovation–work engagement, networking, and resource competing–may coexist.

The proposed method can be utilized for report compliance checks, enhancing transparency and affordability when other dimensions of GRI standards are also included in the dataset. Our approach can also be employed by standard developers to facilitate the creation of next-generation standards, and by reporting companies and assurers to verify report quality and compliance with the GRI framework.

Future research should primarily focus on employing more advanced machine learning techniques to improve the model's predictive capabilities. This includes adopting multi-label classification, leveraging state-of-the-art embeddings techniques, and enhancing text preprocessing accuracy. The underlying principles of our methodology allow for the easy incorporation of other reporting dimensions: general, economic, and environmental. Further improvements in accuracy and scope could be achieved by automating the process of code and page-number extraction or incorporating this information by a standards developer (e.g., by indicating indexing information during report issuance or publication). In-depth analysis of the impacts generated by companies could be conducted by combining the proposed SBIC model with other text mining tools, such as dependency parsing and the subject-object-action approach. This combination enables researchers to obtain fine-grained impacts from reports and can help build an impact knowledge database. Additional avenues for future research include comparing CSR frameworks and standards reporting and developing methods for retrieving quantitative indicators. Our proposed method can serve as a foundation for checking report compliance with the GRI framework, verifying the disclosed information against actual CSR performance, and influencing the development of next-generation standards.

CRediT authorship contribution statement

Ivan Nechaev: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Daniel S. Hain:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data available upon request. Trained model is available at https://github.com/ia-nechaev/sbic-method.

Appendix

See Tables 5–7 and Figs. 7–9.

References

Acs, J.Z., Audretsch, D.B., 1989. Patents as a measure of innovative activity. Kyklos Jahrb. Inst. Gesch. Med. Univ. Leipzig 42 (2), 171–180.

Anon, 2012. Performance Standards on Environmental and Social Sustainability. International Finance Corporation, https://www.ifc.org/wps/wcm/connect/24e6bfc3-5de3-444d-be9b-226188c95454/PS_English_2012_Full-Document.pdf?MOD=AJPERES&CVID=jkV-X6h.

Anon, 2016. World Bank Environmental and Social Framework. World Bank, Washington, DC.

Anon, 2018a. OECD Due Diligence Guidance for Responsible Business Conduct. OECD.

Anon, 2018b. SASB Standards Application Guidance. Sustainability Accounting Standards Board, https://www.sasb.org/standards/download/.

Anon, 2019. MSCI ESG Universal Indexes Methodology. MSCI Inc.

Anon, 2020a. Consolidated Set of GRI Sustainability Reporting Standards. Global Sustainability Standards Board.

Anon, 2020b. The Equator Principles. The Equator Principles Association.

Anon, 2023. GRI 1: Foundation 2021. Global Sustainability Standards Board, https://www.globalreporting.org/.

Asif, M., Jajja, M.S.S., Searcy, C., 2019. Social compliance standards: re-evaluating the buyer and supplier perspectives. J. Clean. Prod. 227, 457–471. http://dx.doi.org/10.1016/j.jclepro.2019.04.157.

Audretsch, D., Lehmann, E., Meoli, M., Vismara, S. (Eds.), 2016. University Evolution, Entrepreneurial Activity and Regional Competitiveness. In: International Studies in Entrepreneurship, vol. 32, Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-17713-7.

Aureli, S., 2017. A comparison of content analysis usage and text mining in CSR corporate disclosure. Int. J. Digit. Account. Res. 17, 1–32. http://dx.doi.org/10.4192/1577-8517-v17_1.

Becerra-Vicario, R., Ruiz-Palomo, D., León-Gómez, A., Santos-Jaén, J., 2023. The relationship between innovation and the performance of small and medium-sized businesses in the industrial sector: The mediating role of CSR. Economies 11 (3), http://dx.doi.org/10.3390/economies11030092.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bocquet, R., Le Bas, C., Mothe, C., Poussing, N., 2013. Are firms with different CSR profiles equally innovative? Empirical analysis with survey data. Eur. Manag. J. 31 (6), 642–654. http://dx.doi.org/10.1016/j.emj.2012.07.001.

Broadstock, D.C., Matousek, R., Meyer, M., Tzeremes, N.G., 2020. Does corporate social responsibility impact firms' innovation capacity? The indirect link between environmental & social governance implementation and innovation performance. J. Bus. Res. 119, 99–110. http://dx.doi.org/10.1016/j.jbusres.2019.07.014.

Brown, H.S., de Jong, M., Levy, D.L., 2009. Building institutions based on information disclosure: Lessons from GRI's sustainability reporting. J. Clean. Prod. 17 (6), 571–580. http://dx.doi.org/10.1016/j.jclepro.2008.12.009.

Cai, W., Gu, J., Wu, J., 2023. The effect of corporate social responsibility on open innovation: The moderating role of firm proactiveness. Manage. Decis. http://dx.doi.org/10.1108/MD-09-2022-1174.

Castellanos, A., Parra, C., Tremblay, M., 2015. Corporate social responsibility reports: Understanding topics via text mining. In: Analyzing CSR Reports via Text Mining Twenty-first Americas Conference on Information Systems. p. 15.

Cerioni, E., Marasca, S., 2021. The methods of social impact assessment: The state of the art and limits of application. Gen. Manag. 22 (183), 9.

Clark, C., Rosenzweig, W., Long, D., Olsen, S., 2004. Double Bottom Line Project Report: Assessing Social Impact In Double Bottom Line Ventures. Working Paper Series, University of California, Berkeley, p. 73.

Cook, K.A., Romi, A.M., Sánchez, D., Sánchez, J.M., 2019. The influence of corporate social responsibility on investment efficiency and innovation. J. Bus. Financ. Account. 46 (3–4), 494–537. http://dx.doi.org/10.1111/jbfa.12360.

Costa, C., Lages, L.F., Hortinha, P., 2015. The bright and dark side of CSR in export markets: Its impact on innovation and performance. Int. Bus. Rev. 24 (5), 749–757.

de Groen, W.P., Alcidi, C., Simonelli, F., Campmas, A., Di Salvo, M., Musmeci, R., Oliinyk, I., Tadi, S., European Commission, Directorate-General for Financial Stability, Financial Services and Capital Markets Union, 2021. Study on the Non-Financial Reporting Directive: Final Report. Publications Office of the European Union.

Dewar, R.D., Dutton, J.E., 1986. The adoption of radical and incremental innovations: An empirical analysis. Manage. Sci. 32 (11), 1422–1433. http://dx.doi.org/10.1287/mnsc.32.11.1422.

Dolan, P., Kahneman, D., 2008. Interpretations of utility and their implications for the valuation of health. Econ. J. 118 (525), 215–234. http://dx.doi.org/10.1111/j.1468-0297.2007.02110.x.

Drucker, P.F., 2015. Innovation and Entrepreneurship: Practice and Principles. In: Routledge Classics, Routledge.

Duan, Z., He, Y., Zhong, Y., 2018. Corporate social responsibility information disclosure objective or not: An empirical research of Chinese listed companies based on text mining. Nankai Bus. Rev. Int. 9 (4), 519–539. http://dx.doi.org/10.1108/NBRI-01-2018-0003.

Ebrahim, A., Rangan, V.K., 2014. What impact? A framework for measuring the scale and scope of social performance. Calif. Manage. Rev. 56 (3), 118–141. http://dx.doi.org/10.1525/cmr.2014.56.3.118.

Emerson, J., 2003. The blended value proposition: Integrating social and financial returns. Calif. Manage. Rev. 45 (4), 35–51. http://dx.doi.org/10.2307/41166187.

Esteves, A.M., Vanclay, F., 2009. Social development needs analysis as a tool for SIA to guide corporate-community investment: Applications in the minerals industry. Environ. Impact Assess. Rev. 29 (2), 137–145. http://dx.doi.org/10.1016/j.eiar.2008.08.004.

Ettlie, J.E., Bridges, W.P., O'Keefe, R.D., 1984. Organization strategy and structural differences for radical versus incremental innovation. Manage. Sci. 30 (6), 682–695. http://dx.doi.org/10.1287/mnsc.30.6.682.

Fagerberg, J., Mowery, D.C., Nelson, R.R. (Eds.), 2005. The Oxford Handbook of Innovation. Oxford University Press, OCLC: ocm57064887.

Fonseca, A., McAllister, M.L., Fitzpatrick, P., 2012. Sustainability reporting among mining corporations: A constructive critique of the GRI approach. J. Clean. Prod. 84, 70–83. http://dx.doi.org/10.1016/j.jclepro.2012.11.050.

Freeman, C., Soete, L., 2017. Economics of Industrial Innovation, third ed. Routledge Taylor & Francis Group, first issued in hardback 2017.

Fuente, J., García-Sánchez, I., Lozano, M., 2017. The role of the board of directors in the adoption of GRI guidelines for the disclosure of CSR information. J. Clean. Prod. 141, 737–750. http://dx.doi.org/10.1016/j.jclepro.2016.09.155.

Furman, J.L., Porter, M.E., Stern, S., 2002. The determinants of national innovative capacity. Res. Policy 31 (6), 899–933.

Gallego-Álvarez, I., Manuel Prado-Lorenzo, J., García-Sánchez, I.-M., 2011. Corporate social responsibility and innovation: A resource-based theory. Manage. Decis. 49 (10), 1709–1727. http://dx.doi.org/10.1108/00251741111183843.

Gangopadhyay, S., Homroy, S., 2023. Do social policies foster innovation? Evidence from India's CSR regulation. Res. Policy 52 (1), http://dx.doi.org/10.1016/j.respol.2022.104654.

García-Piqueres, G., García-Ramos, R., 2021. Complementarity between CSR dimensions and innovation: Behaviour, objective or both? Eur. Manag. J. http://dx.doi.org/10.1016/j.emj.2021.07.010, S0263237321001080.

García-Sánchez, I.-M., 2020. Drivers of the CSR report assurance quality: Credibility and consistency for stakeholder engagement. Corp. Soc. Responsib. Environ. Manag. 27 (6), 2530–2547. http://dx.doi.org/10.1002/csr.1974.

García-Sánchez, I.-M., Hussain, N., Aibar-Guzmán, C., Aibar-Guzmán, B., 2022. Assurance of corporate social responsibility reports: Does it reduce decoupling practices? Bus. Ethics Environ. Responsib. 31 (1), 118–138. http://dx.doi.org/10.1111/beer.12394.

Gaudêncio, P., Coelho, A., Ribeiro, N., 2019. Impact of CSR perceptions on workers' innovative behaviour: Exploring the social exchange process and the role of perceived external prestige. World Rev. Entrepreneurship Manag. Sustain Dev. 15 (1/2), 151–173.

Goloshchapova, I., Poon, S.-H., Pritchard, M., Reed, P., 2019. Corporate social responsibility reports: Topic analysis and big data approach. Eur. J. Finance 25 (17), 1637–1654. http://dx.doi.org/10.1080/1351847X.2019.1572637.

Haab, T.C., McConnell, K.E., 2002. Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation. In: New Horizons in Environmental Economics, E. Elgar Pub.

Halkos, G., Skouloudis, A., 2018. Corporate social responsibility and innovative capacity: Intersection in a macro-level perspective. J. Clean. Prod. 182, 291–300. http://dx.doi.org/10.1016/j.jclepro.2018.02.022.

Hart, S.L., 1995. A natural-resource-based view of the firm. Acad. Manag. Rev. 20 (4), 986–1014.

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1. IEEE, pp. 278–282.

Hu, H., Zhang, J., 2023. How do corporate social responsibility and innovation coevolve with organizational forms? evidence from a transitional economy. J. Bus. Ethics http://dx.doi.org/10.1007/s10551-023-05435-8.

Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.

Jones, N., McGinlay, J., Dimitrakopoulos, P.G., 2017. Improving social impact assessment of protected areas: A review of the literature and directions for future research. Environ. Impact Assess. Rev. 64, 1–7. http://dx.doi.org/10.1016/j.eiar.2016.12.007.

Katila, R., 2000. Using patent data to measure innovation performance. Int. J. Bus. Perform. Manag. 2 (1–3), 180–193.

Khan, A., Chen, C.-C., Suanpong, K., Ruangkanjanases, A., Kittikowit, S., Chen, S.-C., 2021. The impact of CSR on sustainable innovation ambidexterity: The mediating role of sustainable supply chain management and second-order social capital. Sustainability 13 (21), 12160. http://dx.doi.org/10.3390/su132112160.

Kiriu, T., Nozaki, M., 2020. A text mining model to evaluate firms' ESG activities: An application for Japanese firms. Asia-Pac. Financ. Mark. 27 (4), 621–632. http://dx.doi.org/10.1007/s10690-020-09309-1.

Kleinknecht, A., Van Montfort, K., Brouwer, E., 2002. The non-trivial choice between innovation indicators. Econ. Innov. New Technol. 11 (2), 109–121. http://dx.doi.org/10.1080/10438590210899.

Knebel, S., Seele, P., 2015. Quo vadis GRI? A (critical) assessment of GRI 3.1 A+ non-financial reports and implications for credibility and standardization. Corp. Commun. 20 (2), 196–212. http://dx.doi.org/10.1108/CCIJ-11-2013-0101.

Kraus, S., Rehman, S.U., García, F.J.S., 2020. Corporate social responsibility and environmental performance: The mediating role of environmental strategy and green innovation. Technol. Forecast. Soc. Change 160, 120262. http://dx.doi.org/10.1016/j.techfore.2020.120262.

Kumar, A., Das, N., 2021. A text-mining approach to the evaluation of sustainability reporting practices: evidence from a cross-country study. Prob. Ekorozwoju 16 (1), 51–60. http://dx.doi.org/10.35784/pe.2021.1.06.

Lee, M., Huang, Y.-L., 2020. Corporate social responsibility and corporate performance: A hybrid text mining algorithm. Sustainability 12 (8), 3075. http://dx.doi.org/10.3390/su12083075.

Liew, W.T., Adhitya, A., Srinivasan, R., 2014. Sustainability trends in the process industries: A text mining-based analysis. Comput. Ind. 65 (3), 393–400. http://dx.doi.org/10.1016/j.compind.2014.01.004.

Liu, S.-H., Chen, S.-Y., Li, S.-T., 2017. Text-mining application on CSR report analytics: a study of petrochemical industry. In: 2017 6th IIAI International Congress on Advanced Applied Informatics. IIAI-AAI, IEEE, pp. 76–81. http://dx.doi.org/10.1109/IIAI-AAI.2017.164.

Lock, I., Seele, P., 2016. The credibility of CSR (corporate social responsibility) reports in Europe. Evidence from a quantitative content analysis in 11 countries. J. Clean. Prod. 122, 186–200. http://dx.doi.org/10.1016/j.jclepro.2016.02.060.

Lundvall, B.-Å., 2016. The Learning Economy and the Economics of Hope. Anthem Press, http://dx.doi.org/10.26530/OAPEN_626406.

Luo, X., Du, S., 2015. Exploring the relationship between corporate social responsibility and firm innovation. Mark. Lett. 26 (4), 703–714. http://dx.doi.org/10.1007/s11002-014-9302-5.

MacGregor, S.P., Fontrodona, J., 2008. Exploring the Fit between CSR and Innovation. IESE Business School Working Paper.

Mahmoudi, H., Renn, O., Vanclay, F., Hoffmann, V., Karami, E., 2013. A framework for combining social impact assessment and risk assessment. Environ. Impact Assess. Rev. 43, 1–8. http://dx.doi.org/10.1016/j.eiar.2013.05.003.

Mantelero, A., 2018. AI and big data: A blueprint for a human rights, social and ethical impact assessment. Comput. Law Secur. Rev. 34 (4), 754–772. http://dx.doi.org/10.1016/j.clsr.2018.05.017.

Marimon, F., Alonso-Almeida, M.d.M., Rodríguez, M.d.P., Cortez Alejandro, K.A., 2012. The worldwide diffusion of the global reporting initiative: What is the point? J. Clean. Prod. 33, 132–144. http://dx.doi.org/10.1016/j.jclepro.2012.04.017.

Martinez-Conesa, I., Soto-Acosta, P., Palacios-Manzano, M., 2017. Corporate social responsibility and its effect on innovation and firm performance: An empirical research in SMEs. J. Clean. Prod. 142, 2374–2383. http://dx.doi.org/10.1016/j.jclepro.2016.11.038.

Mendes, T., Braga, V., Correia, A., Silva, C., 2023. Linking corporate social responsibility, cooperation and innovation: The triple bottom line perspective. Innov. Manag. Rev. 20 (3), 244–280. http://dx.doi.org/10.1108/INMR-03-2021-0039.

Molecke, G., Pinkse, J., 2017. Accountability for social impact: A bricolage perspective on impact measurement in social enterprises. J. Bus. Ventur. 32 (5), 550–568. http://dx.doi.org/10.1016/j.jbusvent.2017.05.003.

Nazir, O., Islam, J.U., 2020. Influence of CSR-specific activities on work engagement and employees' innovative work behaviour: An empirical investigation. Curr. Issues Tour. 23 (24), 3054–3072. http://dx.doi.org/10.1080/13683500.2019.1678573.

Ning, X., Yim, D., Khuntia, J., 2021. Online sustainability reporting and firm performance: Lessons learned from text mining. Sustainability 13 (3), 1069. http://dx.doi.org/10.3390/su13031069.

OECD, 2011. OECD Guidelines for Multinational Enterprises, 2011 Edition. OECD, http://dx.doi.org/10.1787/9789264115415-en.

Olanipekun, A.O., Omotayo, T., Saka, N., 2021. Review of the use of corporate social responsibility (CSR) tools. Sustain. Prod. Consum. 27, 425–435. http://dx.doi.org/10.1016/j.spc.2020.11.012.

Padgett, R.C., Galan, J.I., 2010. The effect of R&D intensity on corporate social responsibility. J. Bus. Ethics 93 (3), 407–418. http://dx.doi.org/10.1007/s10551-009-0230-x.

Paruzel, A., Schmidt, L., Maier, G., 2023. Corporate social responsibility and employee innovative behaviors: A meta-analysis. J. Clean. Prod. 393, http://dx.doi.org/10.1016/j.jclepro.2023.136189.

Pavitt, K., 1985. Patent statistics as indicators of innovative activities: Possibilities and problems. Scientometrics 7 (1–2), 77–99. http://dx.doi.org/10.1007/BF02020142.

Pencle, N., Mălăescu, I., 2016. What's in the words? development and validation of a multidimensional dictionary for CSR and application using prospectuses. J. Emerg. Technol. Account. 13 (2), 109–127. http://dx.doi.org/10.2308/jeta-51615.

Ratajczak, P., Szutowski, D., 2016. Exploring the relationship between CSR and innovation. Sustain. Account. Manag. Policy J. 7 (2), 295–318.

Roszkowska-Menkes, M.T., 2018. Integrating strategic CSR and open innovation. Towards a conceptual framework. Soc. Responsib. J. 14 (4), 950–966. http://dx.doi.org/10.1108/SRJ-07-2017-0127.

Ruggiero, P., Cupertino, S., 2018. CSR strategic approach, financial resources and corporate social performance: The mediating effect of innovation. Sustainability 10 (10), 3611. http://dx.doi.org/10.3390/su10103611.

Ruiz-Blanco, S., Romero, S., Fernandez-Feijoo, B., 2022. Green, blue or black, but washing–what company characteristics determine greenwashing? Environ. Dev. Sustain. 24 (3), 4024–4045. http://dx.doi.org/10.1007/s10668-021-01602-x.

Seele, P., 2016. Digitally unified reporting: How XBRL-based real-time transparency helps in combining integrated sustainability reporting and performance control. J. Clean. Prod. 136, 65–77. http://dx.doi.org/10.1016/j.jclepro.2016.01.102.

Shahi, A.M., Issac, B., Modapothala, J.R., 2012. Intelligent corporate sustainability report scoring solution using machine learning approach to text categorization. In: 2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology. STUDENT, pp. 227–232. http://dx.doi.org/10.1109/STUDENT.2012.6408409.

Shahi, A.M., Issac, B., Modapothala, J.R., 2014. Automatic analysis of corporate sustainability reports and intelligent scoring. Int. J. Comput. Intell. Appl. 13 (01), 1450006. http://dx.doi.org/10.1142/S1469026814500060.

Shahzad, M., Qu, Y., Javed, S.A., Zafar, A.U., Rehman, S.U., 2020. Relation of environment sustainability to CSR and green innovation: A case of Pakistani manufacturing industry. J. Clean. Prod. 253, 119938. http://dx.doi.org/10.1016/j.jclepro.2019.119938.

Siew, R.Y., 2015. A review of corporate sustainability reporting tools (SRTs). J. Environ. Manag. 164, 180–195. http://dx.doi.org/10.1016/j.jenvman.2015.09.010.

Simmou, W., Govindan, K., Sameer, I., Hussainey, K., Simmou, S., 2023. Doing good to be green and live clean! - Linking corporate social responsibility strategy, green innovation, and environmental performance: Evidence from Maldivian and Moroccan small and medium-sized enterprises. J. Clean. Prod. 384, http://dx.doi.org/10.1016/j.jclepro.2022.135265.

Spark Jones, K., 1972. A statistical interpretation of term importance in automatic indexing. J. Doc. 28 (1), 11–21.

Stiglitz, J.E., Sen, A., Fitoussi, J.-P., et al., 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress. Citeseer.

Szeto, E., 2000. Innovation and strategy innovation capacity: Working towards a mechanism for improving innovation within an inter-organizational network. TQM Mag. 12 (2), 9.

Turban, D.B., Greening, D.W., 1997. Corporate social performance and organizational attractiveness to prospective employees. Acad. Manag. J. 16.

Uyar, A., Karaman, A.S., Kilic, M., 2020. Is corporate social responsibility reporting a tool of signaling or greenwashing? evidence from the worldwide logistics sector. J. Clean. Prod. 253, 119997. http://dx.doi.org/10.1016/j.jclepro.2020.119997.

Uyar, A., Koseoglu, M.A., Kılıç, M., Mehraliyev, F., 2021. Thematic structure of sustainability reports of the hospitality and tourism sector: A periodical, regional, and format-based analysis. Curr. Issues Tour. 24 (18), 2602–2627. http://dx.doi.org/10.1080/13683500.2020.1847050.

Vanclay, F., 2020. Reflections on social impact assessment in the 21st century. Impact Assess. Project Apprais. 38 (2), 126–131. http://dx.doi.org/10.1080/14615517.2019.1685807.

Vanclay, F., Esteves, A.M., Aucamp, I., Research, E., Franks, D.M., 2015. Social impact assessment: Guidance for assessing and managing social impacts of projects. p. 108.

Vanclay, Hanna, 2019. Conceptualizing company response to community protest: Principles to achieve a social license to operate. Land 8 (6), 101. http://dx.doi.org/10.3390/land8060101.

Wasiuzzaman, S., Uyar, A., Kuzey, C., Karaman, A.S., 2021. Corporate social responsibility: Is it a matter of slack financial resources or strategy or both? Manag. Decis. Econ. mde.3537. http://dx.doi.org/10.1002/mde.3537.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, second ed. MIT Press.