Information Extraction: Unstructured to Structured for ESG Reports

Zounachuan Sun*, Ranjan Satapathy†, Daixue Guo*, Bo Li*, Xinyuan Liu*,
Yangchen Zhang*, Cheng-ann Tan‡, Ricardo Shirota Filho†, Rick Siow Mong GOH†
*School of Continuing and Lifelong Education, National University of Singapore, Singapore
†Institute of High-Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
‡ DHI Group, Singapore

***Abstract*—The diverse ESG reporting standards adopted worldwide lead to a significant increase in the volume of unstructured reported information, bringing the need for more efficient processing and standardization of ESG information. Our study reveals a significant enhancement in the comprehensiveness of ESG information disclosed by the real estate industry in Singapore, following the SGX's ESG reporting guidelines. We improved efficiency in collecting and standardizing ESG metrics from reports using an NLP-based automatic extraction algorithm. This project not only streamlines the ESG data extraction process but also contributes to the broader goal of converting unstructured data into structured formats. Furthermore, it sets a valuable precedent for the industry, fostering increased transparency and accountability within Singapore's corporate landscape and potentially influencing global standards. Using the automatic extraction technique, we are paving the way for a more informed and responsible approach to corporate sustainability reporting.**

***Index Terms*—ESG, sustainability, disclosure, NLP**

I. Introduction

In recent years, Environmental, Social, and Governance (ESG) has become increasingly important, driven by a confluence of global and organizational factors. Globally recognized environmental issues, led by climate change, have emphasized companies' green practices to mitigate greenhouse gas emissions [1, 2]. In addition, growing social justice movements and corporate governance scandals have catalyzed a combined requirement for transparency and accountability [3, 4]. Externally, the regulatory environment is becoming more stringent, with bodies such as the US Securities and Exchange Commission (SEC) considering new rules for detailed disclosure of climate-related risks and greenhouse gas emissions [5, 6]. This regulatory momentum is echoed globally, as evidenced by the European Union's evolving ESG disclosure requirements, which force companies to be transparent about their sustainability efforts [7].

A significant increase in the volume of information reported characterizes the current landscape of corporate ESG disclosure. Different mainstream voluntary and mandatory ESG reporting frameworks show varying levels of popularity globally. Voluntary standards, such as the Global Reporting Initiative (GRI) and the Sustainability Accounting Standards Board (SASB), provide frameworks that companies can choose to follow to demonstrate their commitment to transparency and sustainability beyond regulatory requirements [8, 9].

Especially, the GRI, favored for its comprehensive approach to sustainability reporting, remains the most widely used voluntary framework globally, with Singapore, Taiwan and Chile leading the way in its adoption [10, 11]. Companies in the US, Canada, and Brazil predominantly use Sustainability Accounting Standards Board (SASB) standards [11]. The SASB framework is recognized for its focus on financial material sustainability factors relevant to investors [8]. In regions with lower use of the GRI or SASB, there is a notable shift towards the adoption of domestic stock exchange guidelines, which are mandatory disclosure standards requiring companies to report specific ESG information [9]. For example, high adoption rates are observed in Asia-Pacific countries, with Malaysia, India, Singapore, and Taiwan reporting against their respective stock exchange guidelines [11]. This diversity in the adoption of ESG reporting frameworks leads to inconsistencies and reduced comparability and poses challenges to the consistency of ESG information disclosed by companies [12, 13, 14].

Based on the intensive ESG reporting information, there is a growing body of literature revealing ESG disclosure and performance trends across specific industries or market regions. For instance, Arvidsson and Dumay [15] extracted a comprehensive dataset from four years of sustainability reports to analyze the longitudinal trends in ESG information quantity, quality, and their effect on corporate performance among the 30 most-traded Swedish companies. Their study provided valuable insights into how ESG practices have evolved and their impact on corporate success in Sweden [15]. Similarly, another study investigated the influence of financial factors on the disclosure of ESG components in the banking sector in Indonesia, Singapore, and Malaysia. This research calculated ESG scores based on the information extracted from ESG reports [16]. The findings highlighted the varying degrees of ESG disclosure across these countries and underscored the role of financial performance in shaping ESG reporting practices within the banking industry [16]. However, the ESG data collection processes that rely on manual efforts show significant limitations in terms of efficiency, thereby restricting the scale and scope of analysis. ESG disclosure reports often contain unstructured data, including qualitative text descriptions, quantitative figures, and information arranged in tables and graphs [17]. This unstructured nature of the data presents challenges in comparing and assessing ESG performances

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00068

across different companies and regions [18, 19, 20]. To address these challenges, there is a growing need for advanced data analytics and automated tools that can efficiently process and analyze large volumes of unstructured ESG data, facilitating the comparison and evaluation of ESG performance across companies.

Artificial Intelligence (AI) significantly facilitates the extraction of ESG reports through advanced techniques such as deep learning and Natural Language Processing (NLP) [21, 22]. Deep learning models, which consist of multiple processing layers, can learn representations of data with various levels of abstraction [23, 24, 25]. These methods have notably improved areas such as speech recognition, object detection, and emotion recognition [26, 27, 28]. Deep learning models, including Convolutional Neural Networks (CNNs), are particularly effective in learning a hierarchy of features by building high-level features from low-level ones [29, 30, 31, 32]. This capability is crucial for processing the unstructured data in ESG reports, which often include qualitative text descriptions, quantitative figures, and information arranged in tables and graphs. To process intensive ESG information, usually in unstructured reports format, more effectively, the automatic interpretation of ESG information becomes a promising innovation direction that leverages NLP, automatic extraction, and text processing techniques to obtain the relevant disclosure information on interested ESG metrics [33]. Studies have explored various text processing and extraction methods, including tokenization, part-of-speech tagging, and named entity recognition, to structure ESG narratives into a more uniform format for analysis [34, 35, 36, 37]. That transition from unstructured to structured data with the aid of automatic extraction techniques is necessary for accurately analyzing and assessing the reliability and quality of the disclosed ESG information across companies. Nevertheless, despite the progress of developing models with enhanced performance for extracting ESG information from reports, the application of using the models to reveal the ESG disclosure trends within the market is still lacking.

To further inject endeavour on automatic ESG information transition, widening the scope of ESG disclosure trends analyzed, harmonizing ESG information disclosed following various global standards, and enhancing its utility for informed decision-making, this study aims to address the following two main objectives: 1) to develop an algorithm that can facilitate the automatic extraction of the targeted ESG information from the unstructured reports, 2) to apply the structured dataset extracted to evaluate the ESG performances across the chosen industry, assessing the ESG disclosure pattern within the market.

Our results highlight an enhanced comprehensiveness pattern in the ESG information disclosed by the real estate industry of Singapore following the SGX suggested ESG reporting guidance, demonstrating the improved effectiveness with the aid of NLP-based extraction in the process of ESG data collection and standardization from ESG reports. The significance of this project lies not only in its potential to streamline the ESG data extraction process for contributing to the endeavour of transiting the unstructured to structured data but also in setting a precedent for the industry, paving the way for a more transparent and accountable corporate landscape in Singapore and beyond.

II. Related work

Recent advancements in NLP have significantly enhanced the analysis of Environmental, Social, and Governance (ESG) data. One notable technique is the use of Large Language Models (LLMs), such as BERT and GPT-4, which have been employed to classify and extract relevant ESG information from corporate reports [38, 39, 40]. For instance, the ESGReveal framework, developed by Zou et al., utilizes LLMs combined with Retrieval Augmented Generation (RAG) techniques to extract and analyze ESG data from corporate reports systematically [41]. Also, Gupta et al. adopted BERT tokenization and the YAKE (Yet Another Keyword Extractor) technique for keyword extraction from the sustainability reports of Forbes India's top companies in 2021 to reveal their current ESG focus [36]. Those NLP techniques help summarize lengthy ESG reports and categorize content according to ESG criteria so that identifying key information is facilitated with the enhanced precision of ESG assessments.

Current ESG research shows the trend of increased reliance on various industries' bulk ESG performance data, with the sustainability report as a crucial source of data extraction. Specifically, increased studies investigate the linkages between ESG performance and financial performance, thereby facilitating investment decision-making [42, 43, 44]. Specific E, S and G indicators have been widely investigated. For instance, firms that actively reduce their carbon footprint often experience lower operational costs and improved efficiency, leading to higher profit margins [45]. From the social perspective, companies that invest in employee well-being often see reduced turnover rates and higher productivity, positively impacting profitability [46]. Regarding the governance aspect, effective governance can enhance investor confidence, leading to a lower cost of capital and higher stock valuations [47].

However, there is still a knowledge gap in adopting the NLP-based automatic extraction algorithm to obtain the ESG performance datasets used for the ESG research. Those gaps in real applications could be attributed to the relatively high technical algorithm design and implementation, creating barriers to the applications by end users and stakeholders from various backgrounds. Thus, this study aims to develop and examine the NLP-based ESG report extraction algorithm that could reduce the application barrier and serve to enrich bulk ESG performance data across various industries.

III. Data and methods

This study selected 50 companies in the real estate industry listed in the Singapore Exchange (SGX) as the demonstration for analyzing the ESG disclosure pattern by adopting an algorithm extracting the unstructured ESG information from

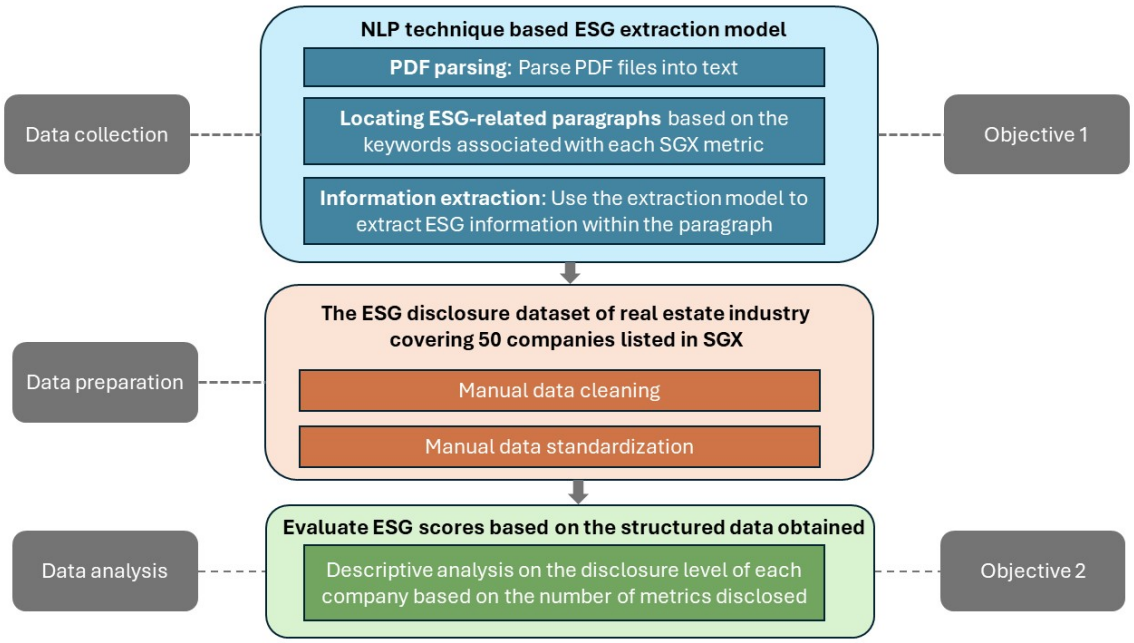

Fig. 1. The workflow of the study

reports. Singapore has been at the forefront of ESG developments, with regulators, investors, and shareholders demanding positive sustainability outcomes and accurate reporting. Singapore's real estate sector is pioneering the use of green technology and sustainable practices, driven by rising global demand for ESG investment and a growing green financing market, which is set to accelerate green property investments [48]. As a densely populated city-state, Singapore's approach to sustainable urban development serves as a model for other cities, highlighting the broader economic impacts and resilience brought about by ESG practices in response to global challenges [49]. The overall workflow of the study is shown in Fig.1.

A. Selection of metrics to be extracted and keyword preparation

To align with the analysis of ESG disclosure patterns in the Singapore market, we refer to the 29 core ESG metrics suggested by the Singapore Exchange (SGX), as detailed in Table 1. In practical operation, we listed the metrics as a data table containing keywords. We used Pandas to read the keywords data table and preprocess these keywords, such as removing spaces and converting cases, to ensure matching accuracy. The preprocessed keywords were then applied to the PDF text analysis and obtain sentences and related information containing these keywords through keyword matching and text extraction methods.

B. Development of NLP-based ESG extraction algorithm

In the algorithm development phase, we first used the PyPDF2 [50] library to read the text content in PDF files. Porter Stemmer in The Natural Language Toolkit (NLTK) was then adopted to extract the stem of keywords and perform keyword matching in the extracted text. By converting text to lowercase and performing stem matching, we could significantly improve the accuracy and coverage of keyword detection. The Pandas library was then adopted to read keywords from Excel files and process the text data extracted from PDF. Finally, the processed results were written into a new Excel file using the OpenPyXL library, ensuring that all matching keywords and related sentences were accurately saved.

To be more specific, PorterStemmer from NLTK (Natural Language Toolkit) was used for stem extraction. Stem extraction is the technique of simplifying words into their root form by removing affixes such as the plural 's', past tense 'ed', and continuous tense 'ing' to normalize different forms of words into the same root, which is particularly important for identifying keywords in text processing. Through stem extraction, we could unify the processing of words such as 'eating' and 'ate', simplifying them all to 'eat'. This technique is of great significance in NLP applications such as search engines and text analysis, as it improves the accuracy and coverage of keyword matching, thereby enhancing the effectiveness of text analysis [51].

The keyword extraction process includes the following steps. Firstly, all text in the report was converted to lowercase to ensure consistency during keyword matching. This step is necessary as the uppercase and lowercase forms in the text may affect the accuracy of matching. For example, 'Environment', 'environment', and 'ENVIRONMENT' become 'environment' after being converted to lowercase, improving matching accu-

TABLE I
SGX SUGGESTED ESG METRICS EXTRACTED IN THIS STUDY

Environmental	Social	Governance
Total GHG absolute emissions	Current employees by gender	Board independence
GHG Emission intensities	New hires by gender	Women on the board
Total energy consumption	Turnover by gender	Women in the management team
Energy intensity consumption	Current employees by age groups	Anti-corruption disclosures
Total water consumption	New hires by age groups	Anti-corruption training for employees
Water consumption intensity	Turnover by age groups	List of relevant certifications
Total waste generated	Total turnover	Alignment with frameworks and disclosure practices
	Total number of employees	Assurance of sustainability report [10]
	Average training hours per employee	
	Average training hours per employee by gender	
	Fatalities	
	High-consequence injuries	
	Recordable injuries	
	Recordable work-related ill health cases	

racy. When matching the keywords within the text, we first attempted to match keywords accurately. If an exact matching keyword is not found, the code will perform a fuzzy match by checking the stem version of the keyword. Stem matching technology helps identify different forms of the same keyword, such as unifying 'writing', 'wrote', and 'writer' as 'write' to ensure a wider range of keyword detection.

The algorithm was run in a testing environment to verify its functionality after the development. This includes verifying the reading of PDF files, extracting keywords, and saving results. Through the validation, we ensure that keywords and related sentences are correctly extracted and saved by checking the generated Excel file.

C. Descriptive analysis of ESG disclosure pattern

To further utilize the information extracted from ESG reports to understand the ESG disclosure pattern within the real estate industry across the years, we determined the disclosure pattern at the SGX-suggested ESG metric level practised by 50 SGX-listed companies from the real estate industry. Specifically, we assessed the number of metrics that a company has disclosed out of the total 29 SGX-suggested metrics included in the study. We present the descriptive analysis in the box plots indicating the maximum, minimum, median and mean of the disclosure score across three years 2021, 2022, and 2023. Besides the industry overall, we specified the evolution of the number of companies making disclosures on each metric across the years. This analysis demonstrates a standardized comparison across different organizations, revealing the industrial patterns of ESG reporting compliance.

IV. RESULTS

A. The effectiveness of algorithm extraction

The code's application for extracting targeted SGX metrics from ESG reports has been evaluated. It successfully identified and extracted 49.5% of the desired data items for SGX metrics from a total of 135 ESG reports (Table 2). When compared with manual validation datasets, our algorithm effectively located approximately 42.2% of the data items present in those 135 ESG reports. Additionally, it accurately generated NA outcomes for 37.7% of the data items not mentioned in the reports (Table 2). These percentages reflect the code's ability to accurately identify and retrieve relevant information that matches the predefined set of keywords. However, the remaining proportions indicate instances where the code either provided extraction results for information not mentioned in the reports or failed to obtain existing information. These situations arose when the code encountered data that did not match the extraction criteria or when keywords appeared multiple times in the reports, making it challenging to obtain the most aligned information. These situations occurred when the code encountered data that did not match the extraction criteria or could not obtain the most aligned information when the keywords appeared multiple times in the reports. These proportions of extraction outcomes are crucial for understanding the scope of the code's applicability and for identifying opportunities for further optimization.

B. The enhanced ESG information disclosed by corporations

The observed trend in overall disclosure scores is a clear indicator of the growing importance that companies are placing on ESG matters. The level of disclosure has increased significantly across the years, signalling a progressive shift towards more transparent and accountable business practices in the real estate industry of Singapore (Fig. 2).

Specifically, the company's disclosed environmental performance metrics show an increasing trend year by year. In particular, significant improvements are observed in metrics related to "greenhouse gas emissions" and "energy consumption" (Fig. 3). However, for metrics regarding "water consumption" and "waste consumption", the data indicates further improvement in management practices in these areas (Fig. 3). In the indicators of the social dimension, It suggests the stability of the gender composition of employees and the improvement of the gender ratio of new employees (Fig. 4). Meanwhile, the number of companies disclosing the age composition and age of new hires is also increasing, which demonstrates the importance companies place on workforce diversity (Fig. 4).

TABLE II
THE VALIDATION OF ALGORITHM EXTRACTION RESULTS

	Actually exists in reports	Not mentioned in reports	Total
Code found	1053	883	1936 (49.5%)
Code not found	1444	535	1979 (50.5%)
Total	2497	1418	3915

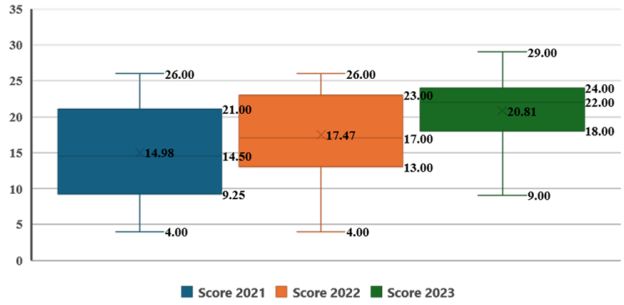

Fig. 2. The overall disclosure pattern of 50 SGX-listed real estate companies

The disclosure of staff turnover also shows an increasing pattern (Fig. 4). Regarding workplace injuries and health and safety, disclosure has improved, with more companies reporting fatalities, serious injuries, and work-related health events (Fig. 4). In the governance dimension, the high disclosure rates are found for metrics of "anti-corruption disclosures" and "consistency with the framework and disclosure practices." Nevertheless, there still needs improvement, particularly in the areas of "anti-corruption training" and "sustainability reporting assurance" (Fig. 5).

V. DISCUSSION AND CONCLUSION

A. The enhancement in ESG disclosure

Over the past three years, the real estate sector has shown a steady improvement in its environmental, social and governance (ESG) reporting. This is evidenced by the improved average scores and the growing number of companies adhering to SGX's recommended core ESG metrics. As for the environmental metrics, companies are increasingly disclosing environmental metrics, with notable progress in reporting on greenhouse gas emissions and energy use. However, there is a recognized need for better management practices in water and waste management. Regarding the social aspect, stability in the gender distribution of the workforce is being maintained, and there's a positive shift in the gender balance of new hires. A steady increase in age diversity disclosure has also been found, signalling a commitment to inclusivity. Disclosure of employee turnover is increasing, with more advanced management still required in this area. There has been an increase in disclosure of workforce size composition, employee training, and work-related injuries. Regarding the governance metrics, there is observed a high level of disclosure on anti-corruption efforts and alignment with established frameworks and practices. Areas identified for further improvement include anti-corruption training and verification of sustainability reports.

That enhanced disclosure pattern is aligned with the more restricted external regulations suggested by the SGX. Since SGX Core ESG Metrics were issued in December 2021, a significant enhancement in the compliance of the ESG disclosure has been witnessed through our analysis of the disclosure score, indicating the progress in standardizing ESG disclosures and providing a common language for ESG communication between companies and investors [19]. Similar trends of ESG disclosure improvement patterns under implemented SGX guidance and requirements are also reported by Phang and Chia [52]. This regulatory requirement ensures that companies disclose relevant environmental, social, and governance factors, with prominent improvement in areas such as climate change actions, social impact programs, diversity and inclusion outcomes, and corporate governance practices [52]. Besides implementing local regulatory requirements, investor influence and the raised public awareness are also recognized as the potential drivers of such improved ESG disclosure in Singapore [2, 52]. Investors increasingly demand comprehensive ESG information to make informed decisions, as sustainability is becoming more prioritized in market competition, pressuring companies to enhance their transparency and disclose relevant ESG data [20, 4]. The enhanced alignment with local SGX guidance also facilitates the ESG disclosure to align with international standards, ensuring ESG reports' compatibility and credibility on an international scale [7, 12].

B. Enhanced efficiency in adopting automatic extraction

Despite the industrial patterns revealed with the validated algorithm extraction datasets, our validation results also indicate two challenging instances for the algorithm to provide accurate matching and extraction outcomes. In the first instance, where the code failed to obtain existing information, it can be further attributed to the data mismatch with the predefined dictionary of keywords. ESG reports are often unstandardized, dense with technical data, and vary significantly in format and terminology for expressing the same metric, resulting in greater requirements of establishing the keywords dictionary that captures as many synonyms used in ESG reports as possible. Traditional NLP techniques may be challenging to extract information accurately due to their limitations in understanding the context [53]. This issue is exacerbated when the reports contain complex or nuanced language that the extraction criteria do not account for [54]. The second

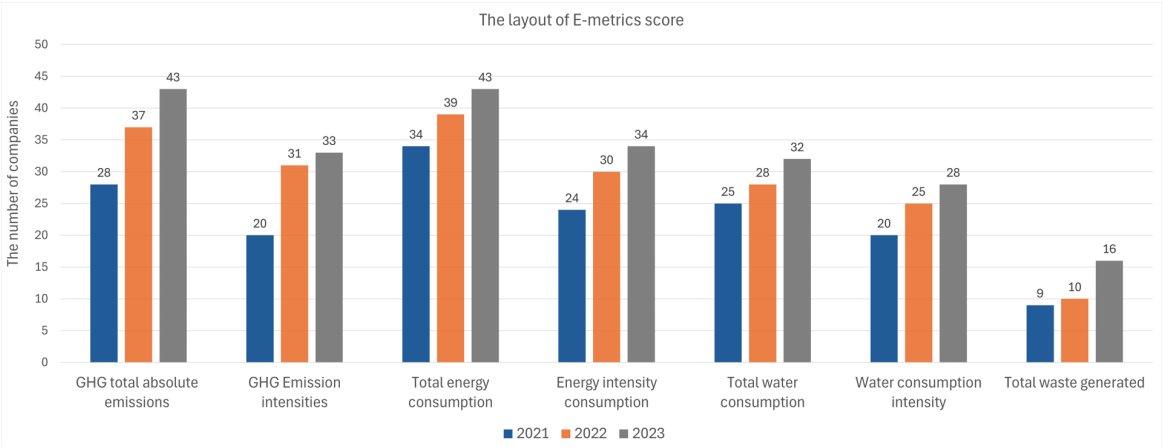

Fig. 3. The number of companies discloses on each environmental metric

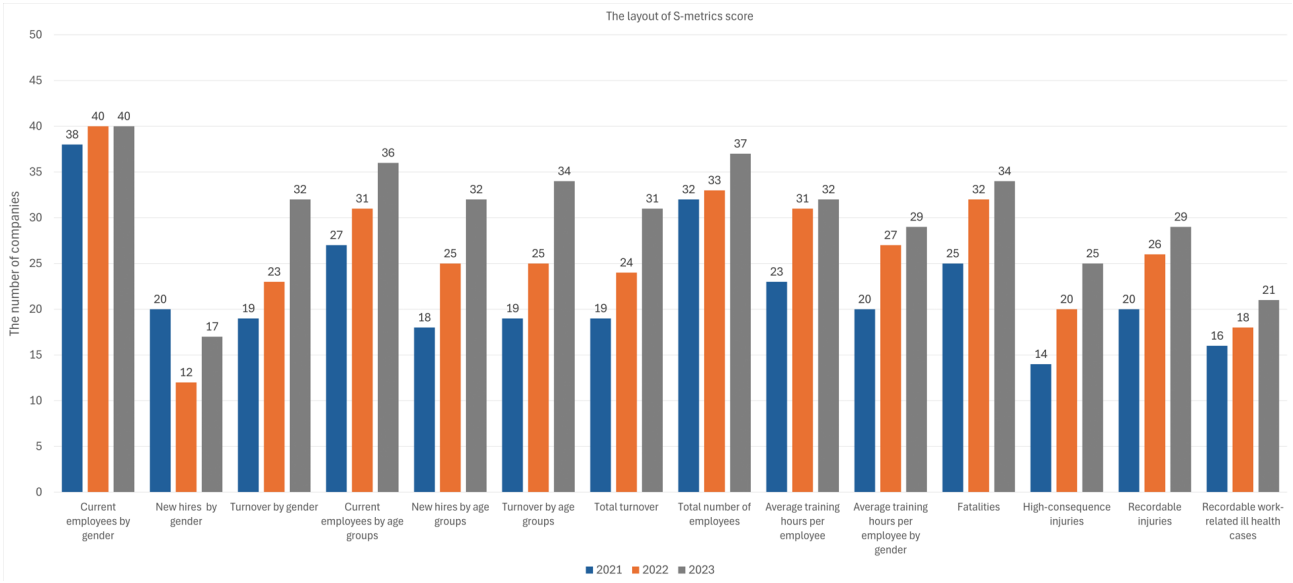

Fig. 4. The number of companies discloses on each social metric

instance hindering the accuracy of our NLP-based algorithm lies in the keywords repetition and ambiguity. When keywords appear multiple times in the reports, it becomes challenging to extract the most relevant information, leading to the inaccuracy that the algorithm provides extraction results for information not mentioned in the reports. This is because keyword-based extraction methods do not effectively differentiate between multiple instances of the same keyword in different contexts. For example, the term "sustainability" might appear in various sections of a report, each with a different focus, making it difficult to determine which matched information is most relevant to the extraction criteria. That brings the further algorithm optimizations on semantics analysis.

To address these limitations, several improvements could be considered. Firstly, utilizing advanced NLP models, such as BERT or GPT-4, can significantly improve the contextual understanding of the paragraphs [55, 36]. These models can understand the nuances and context in which keywords are used, thereby reducing the instances of data mismatch and keyword ambiguity. For example, ESGReveal, an LLM-based approach, has shown higher accuracy rates in data extraction and disclosure analysis by leveraging contextual understanding [41]. Additionally, combining NLP techniques with Retrieval Augmented Generation (RAG) could also be another approach practised [41]. RAG involves using retrieval mechanisms to fetch relevant documents or sections before generating the final output. This approach has the implications of filtering out irrelevant information and focusing on the most pertinent data, facilitating the matching precision [56]. Besides the refinement of NLP-based techniques, techniques

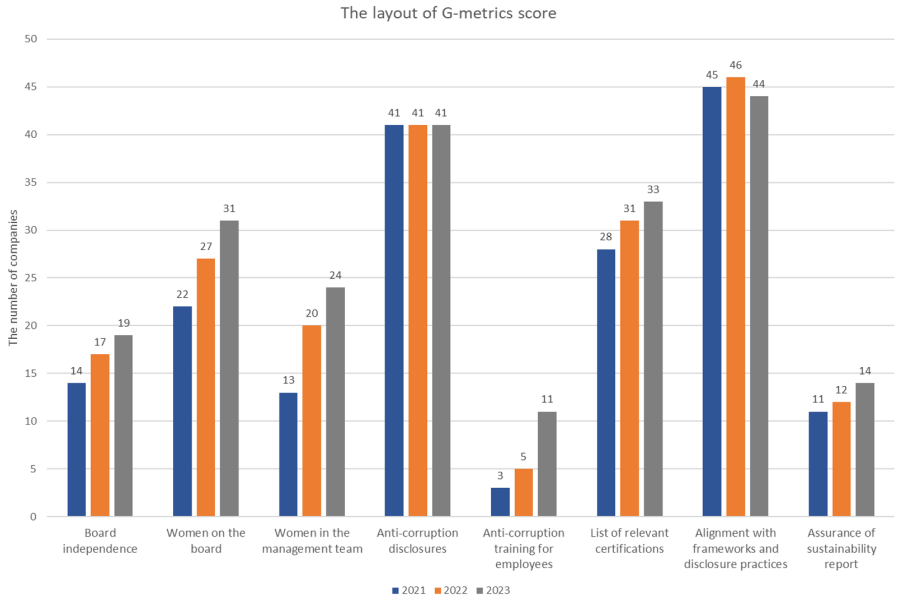

Fig. 5. The number of companies discloses on each governance metric

used for formatting and pre-processing of documents to be extracted are another refinement direction. For example, there has been practice in adopting the XBRL-style (Extensible Business Reporting Language) tagging system to provide a standardized language for the electronic communication of business and financial data and enhance the accuracy and reliability of extracting ESG metrics from reports [55].

C. Outlooks and conclusions

Our findings underscore a notable improvement in the comprehensiveness of ESG information disclosed by the real estate industry in Singapore, with enhanced compliance with the SGX's suggested ESG reporting guidelines. The process of collecting and standardizing ESG metrics information from reports by adopting the NLP-based automatic extraction algorithm in this study demonstrates enhanced efficiency in information gathering. The implications of this study lie in streamlining the ESG data extraction process and contributing to the broader effort of transforming unstructured data into structured formats. Moreover, this study devotes endeavors to promoting greater transparency and accountability within the corporate landscape of Singapore and regions beyond. By leveraging an NLP-based algorithm, we are paving the way for a more comparable and credible approach to corporate sustainability reporting. The implications of those techniques not only facilitate more accurate and comprehensive data collection but also support the development of standardized reporting practices that can be replicated across different sectors and regions. Future work will include comprehensive and multifaceted evaluation of interpretability [57, 58], examining not only faithfulness [59] but also robustness and utility across neurosymbolic AI research [60].

In conclusion, the advancements achieved through this project highlight the critical role of innovative technologies in enhancing ESG reporting. The successful implementation of NLP-based extraction algorithms exemplifies how automation can significantly improve the efficiency and accuracy of data collection. This progress is essential for fostering transparency and accountability in corporate sustainability practices, providing intensive ESG performance datasets for investment decision-making. As the real estate industry in Singapore continues to evolve, the insights gained from this study can serve as a model for other industries and regions, ultimately contributing to the global effort to achieve more sustainable and responsible business practices.

REFERENCES

[1] Luay Jum'a et al. "Towards a sustainability paradigm; the nexus between lean green practices, sustainability-oriented innovation and Triple Bottom Line". In: *International Journal of Production Economics* 245 (2022), p. 108393.

[2] Dan Daugaard and Ashley Ding. "Global drivers for ESG performance: The body of knowledge". In: *Sustainability* 14.4 (2022), p. 2322.

[3] Barbara Novick et al. "Exploring ESG: A practitioner's perspective". In: *Black Rock* (2016), pp. 1–14.

[4] Stuart L Gillan, Andrew Koch, and Laura T Starks. "Firms and social responsibility: A review of ESG and CSR research in corporate finance". In: *Journal of Corporate Finance* 66 (2021), p. 101889.

[5] Wullianallur Raghupathi, Sarah Jinhui Wu, and Viju Raghupathi. "Understanding Corporate Sustainability Disclosures from the Securities Exchange Commission Filings". In: *Sustainability* 15.5 (2023), p. 4134.

[6] George S Georgiev. "The SEC's Climate Disclosure Rule: Critiquing the Critics". In: *Rutgers L. Rec.* 50 (2022), p. 101.

[7] Hammed Afolabi, Ronita Ram, and Gunnar Rimmel. "Harmonization of sustainability reporting regulation: Analysis of a contested arena". In: *Sustainability* 14.9 (2022), p. 5517.

[8] Simone Pizzi, Salvatore Principale, and Elbano De Nuccio. "Material sustainability information and reporting standards. Exploring the differences between GRI and SASB". In: *Meditari Accountancy Research* 31.6 (2023), pp. 1654–1674.

[9] Satyajit Bose. "Evolution of ESG reporting frameworks". In: *Values at work: Sustainable investing and ESG reporting* (2020), pp. 13–33.

[10] Christina WY Wong et al. "Strategies for building environmental transparency and accountability". In: *Sustainability* 13.16 (2021), p. 9116.

[11] Monica Singhania and Neha Saini. "Institutional framework of ESG disclosures: comparative analysis of developed and developing countries". In: *Journal of Sustainable Finance & Investment* 13.1 (2023), pp. 516–559.

[12] Mustapha Ibrahim et al. "Sustainability Reporting Frameworks A Comparative Analysis of Reporting Standards and their Implications for Accounting and Reporting". In: *International Journal of Accounting, Finance and Administrative Research* 1.2 (2024), pp. 32–47.

[13] Henry L Friedman, Mirko Stanislav Heinle, and Irina Luneva. "A theoretical framework for ESG reporting to investors". In: *Available at SSRN 3932689* (2021).

[14] Ting-Ting Li et al. "ESG: Research progress and future prospects". In: *Sustainability* 13.21 (2021), p. 11663.

[15] Susanne Arvidsson and John Dumay. "Corporate ESG reporting quantity, quality and performance: Where to now for environmental policy and practice?" In: *Business strategy and the environment* 31.3 (2022), pp. 1091–1110.

[16] Azwani Aulia, Fiona Febriyanti, and Lita Permata Umi. "Trend analysis of ESG disclosure on green finance performance in Indonesia, Malaysia & Singapore Exchanges". In: *JAK (Jurnal Akuntansi) Kajian Ilmiah Akuntansi* 10.1 (2023), pp. 79–98.

[17] Jiahui Peng et al. "Advanced Unstructured Data Processing for ESG Reports: A Methodology for Structured Transformation and Enhanced Analysis". In: *arXiv preprint arXiv:2401.02992* (2024).

[18] Guochao Wan et al. "Hotspots and trends of environmental, social and governance (ESG) research: A bibliometric analysis". In: *Data Science and Management* 6.2 (2023), pp. 65–75.

[19] Randall E Duran and Peter Tierney. "Fintech data infrastructure for ESG disclosure compliance". In: *Journal of Risk and Financial Management* 16.8 (2023), p. 378.

[20] Bjorg Jonsdottir et al. "Barriers to using ESG data for investment decisions". In: *Sustainability* 14.9 (2022), p. 5157.

[21] Erik Cambria. *Understanding Natural Language Understanding*. Springer, ISBN 978-3-031-73973-6, 2024.

[22] Erik Cambria et al. "Seven Pillars for the Future of Artificial Intelligence". In: *IEEE Intelligent Systems* 38.6 (2023), pp. 62–69.

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[24] Erik Cambria et al. "SenticNet 8: Fusing Emotion AI and Commonsense AI for Interpretable, Trustworthy, and Explainable Affective Computing". In: *Proceedings of the International Conference on Human-Computer Interaction (HCII)*. Washington DC, USA, 2024.

[25] David Vilares et al. "BabelSenticNet: A Commonsense Reasoning Framework for Multilingual Sentiment Analysis". In: *IEEE SSCI*. 2018, pp. 1292–1298.

[26] Jayden Khakurel et al. "The rise of artificial intelligence under the lens of sustainability". In: *Technologies* 6.4 (2018), p. 100.

[27] Qian Liu et al. "PrimeNet: A Framework for Commonsense Knowledge Representation and Reasoning Based on Conceptual Primitives". In: *Cognitive Computation* (2024).

[28] Yosephine Susanto et al. "The Hourglass Model Revisited". In: *IEEE Intelligent Systems* 35.5 (2020), pp. 96–102.

[29] Patrice Y Simard, David Steinkraus, John C Platt, et al. "Best practices for convolutional neural networks applied to visual document analysis." In: *Icdar*. Vol. 3. 2003. Edinburgh. 2003.

[30] Botao Zhong et al. "Convolutional neural network: Deep learning-based classification of building quality problems". In: *Advanced Engineering Informatics* 40 (2019), pp. 46–57.

[31] Iti Chaturvedi et al. "Learning word dependencies in text by means of a deep recurrent belief network". In: *Knowledge-Based Systems* 108 (2016), pp. 144–154.

[32] Erik Cambria et al. "Statistical approaches to concept-level sentiment analysis". In: *IEEE Intelligent Systems* 28.3 (2013), pp. 6–9.

[33] Mohammed Al Qady and Amr Kandil. "Concept relation extraction from construction documents using natural language processing". In: *Journal of construction engineering and management* 136.3 (2010), pp. 294–302.

[34] Emilien Caudron and Frédéric Vrins. "Measuring ESG Performance: A Text Mining Approach". In: *Louvain School of Management, Université catholique de Louvain. CFA Institute* (2022).

[35] Lanxin Jiang, Yu Gu, and Jun Dai. "Environmental, social, and governance taxonomy simplification: A hybrid text mining approach". In: *Journal of Emerging Technologies in Accounting* 20.1 (2023), pp. 305–325.

[36] Akriti Gupta, Aman Chadha, and Vijaishri Tewari. "A Natural Language Processing Model on BERT and YAKE technique for keyword extraction on sustainability reports". In: *IEEE Access* (2024).

[37] Xiaoyun Joy Wang et al. "ESGPDE: An ESG Performance Data Extraction Model." In: *Journal of Financial Data Science* 6.1 (2024).

[38] Abby Yaqing Zhang and Joseph H. Zhang. "Renovation in environmental, social and governance (ESG) research: the application of machine learning". In: *Asian Review of Accounting* 32.4 (2023), pp. 554–572.

[39] Evgeny Burnaev et al. "Practical AI Cases for Solving ESG Challenges". In: *Sustainability* 15.17 (2023). ISSN: 2071-1050. DOI: 10.3390/su151712731. URL: https://www.mdpi.com/2071-1050/15/17/12731.

[40] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. *Bloated Disclosures: Can ChatGPT Help Investors Process Information?* 2024. arXiv: 2306.10224 [econ.GN]. URL: https://arxiv.org/abs/2306.10224.

[41] Yi Zou et al. "ESGReveal: An LLM-based approach for extracting structured data from ESG reports". In: *arXiv preprint arXiv:2312.17264* (2023).

[42] Mahmut Aydoğmuş, Güzhan Gülay, and Korkmaz Ergun. "Impact of ESG performance on firm value and profitability". In: *Borsa Istanbul Review* 22 (2022), S119–S127.

[43] Maria Giuseppina Bruna et al. "Investigating the marginal impact of ESG results on corporate financial performance". In: *Finance Research Letters* 47 (2022), p. 102828.

[44] Simin Chen, Yu Song, and Peng Gao. "Environmental, social, and governance (ESG) performance and financial outcomes: Analyzing the impact of ESG on financial performance". In: *Journal of Environmental Management* 345 (2023), p. 118829.

[45] Tensie Whelan and Carly Fink. "The comprehensive business case for sustainability". In: *Harvard Business Review* 21.2016 (2016).

[46] Tait Shanafelt, Joel Goh, and Christine Sinsky. "The business case for investing in physician well-being". In: *JAMA internal medicine* 177.12 (2017), pp. 1826–1832.

[47] LA Ley, FATHYAH Hashim, and ZAINI Embong. "Board characteristics, investors' confidence and firm value: Malaysian evidence". In: *Asian Journal of Accounting and Governance* 12 (2019), pp. 169–181.

[48] Mansi Jain et al. "Assessing governance of low energy green building innovation in the building sector: Insights from Singapore and Delhi". In: *Energy Policy* 145 (2020), p. 111752.

[49] EUSTON QUAH and JUN RUI TAN. "Pursuing growth and managing the environment: The singapore model". In: *Journal of Business and Economic Analysis* 5.01 (2022), pp. 1–74.

[50] Mathieu Fenniak et al. *The PyPDF2 library*. 2022. URL: https://pypi.org/project/PyPDF2/.

[51] Divya Khyani et al. "An interpretation of lemmatization and stemming in natural language processing". In: *Journal of University of Shanghai for Science and Technology* 22.10 (2021), pp. 350–357.

[52] Rachel Phang and Yaru Chia. "Sustainability and the sunlight of disclosure: ESG disclosure in three Asian financial centres". In: *Review of European, Comparative & International Environmental Law* 33.2 (2024), pp. 209–223.

[53] Marco Bronzini et al. "Glitter or gold? Deriving structured insights from sustainability reports via large language models". In: *EPJ Data Science* 13.1 (2024), p. 41.

[54] Alessandro Del Vitto, Daniele Marazzina, and Davide Stocco. "ESG ratings explainability through machine learning techniques". In: *Annals of Operations Research* (2023), pp. 1–30.

[55] Steven Katz, Yu Gu, and Lanxin Jiang. "Information Extraction from ESG Reports Using Nlp: A Chatgpt Comparison". In: *Available at SSRN 4836432* (2024).

[56] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[57] Wei Jie Yeo et al. *Self-training Large Language Models through Knowledge Detection*. 2024. arXiv: 2406.11275. URL: https://arxiv.org/abs/2406.11275.

[58] Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. "Plausible Extractive Rationalization through Semi-Supervised Entailment Signal". In: *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics, 2024, pp. 5182–5192. URL: https://aclanthology.org/2024.findings-acl.307.

[59] Yeo Wei Jie et al. "How Interpretable are Reasoning Explanations from Prompting Large Language Models?" In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024, pp. 2148–2164.

[60] Wihan van der Heever et al. "Neurosymbolic AI for Mining Key Aspects of Socially Responsible Investing". In: *2024 IEEE International Conference on Data Mining (ICDM)*. 2024.