European Journal of Sustainable Development (2023), **12**, 4, 319-329 ISSN: 2239-5938
Doi: 10.14207/ejsd.2023.v12n4p319

A Conceptual Framework for Subdomain Specific Pre-Training of Large Language Models for Green Claim Detection

By Wayne Moodaley[1], Arnesh Telukdarie[2]

Abstract

Detection of false or misleading green claims (referred to as "greenwashing") within company sustainability disclosures is challenging for a number of reasons, which include the textual and qualitative nature, volume, and complexity of such disclosures. In recent years, notable progress made in the fields of artificial intelligence and specifically, large language models (LLMs), has showcased the capacity of these tools to effectively analyse extensive and intricate textual data, including the contents of sustainability disclosures. Transformer-based LLMs, such as Google's BERT architecture, were trained on general domain text corpora. Subsequent research has shown that further pre-training of such LLMs on specific domains, such as the climate or sustainability domains, may improve performance. However, previous research often uses text corpora that exhibit significant variation across topics and language and which often consist of heterogeneous subdomains. We therefore propose a conceptual framework for further pre-training of transformer based LLMs using text corpora relating to specific sustainability subdomains i.e. subdomain specific pre-training. We do so as a basis for the improved performance of such models in analysing sustainability disclosures. The main contribution is a conceptual framework to advance the use of LLMs for the reliable identification of green claims and ultimately, greenwashing.

Keywords: greenwashing, artificial intelligence, sustainability, sustainability reporting, sustainability disclosures.

1. Introduction

Companies are facing growing pressure from investors, customers, and other stakeholders around the world to provide transparency regarding the impact of their operations on environmental, social, and governance (ESG) factors, and their efforts and performance in dealing with ESG and climate challenges (JSE Limited, 2021).

The increased stakeholder focus on sustainability has resulted in increased pressure on companies to present an image of sustainability and corporate citizenship. This has manifested in a trend of deceptive disclosures regarding company sustainability practices and performance – referred to as "greenwashing" – has emerged (Siano et al., 2017).

Greenwashing, as the European Commission states, "misleads market actors and does not give due advantage to those companies that are making the effort to green their products and activities. It ultimately leads to a less green economy" (European Commission, 2020).

The implication is that greenwashing threatens the objectives of both company

|[1]Senior Lecturer in Accounting and Financial Management at the Johannesburg Business School, University of Johannesburg.
[2]Professor of Digital Business at the Johannesburg Business School, University of Johannesburg.

sustainability initiatives, as well as the objectives of sustainability reporting, given the incongruence of the practice of greenwashing with those objectives.

Detection of greenwashing within company sustainability disclosures is challenging, for a number of reasons which include the textual and qualitative nature of company disclosures and the volume and complexity of such disclosures (In & Schumacher, 2021) (Macpherson et al., 2021).

Rapid and sustained progress made in the field of artificial intelligence (AI), and LLMs which use natural language processing (NLP), has demonstrated the ability of these technologies to effectively analyse large volumes of frequently complex text corpora, including those containing or consisting of sustainability disclosures.

More recently, the emergence of pre-trained transformer-based large language models (LLMs), such as Google's BERT architecture, has fundamentally transformed the practice and use of NLP techniques. The innovative aspect of the architecture lies in its utilization of a bi-layer transformer architecture, which incorporates an attention-based mechanism to effectively capture contextual relationships among words. This enables the achievement of unsupervised learning by integrating both text input and output within a decoder-encoder framework.

Transformer models have brought about a new era in NLP, where these models undergo a "pre-training" phase on semi-supervised tasks prior to fine-tuning for downstream tasks – an approach that enables the models to acquire a broader understanding of language patterns and structures, thereby enhancing their performance on specific tasks (Gururangan et al., 2020). The work of Devlin et al (2019) and Huang et al (2023) have highlighted the ability of these LLMs in evaluating analyse extensive and intricate textual information, including sustainability disclosures.

While LLMs trained on domain-specific corpora have been shown to provide superior performance, there is no literature which explores subdomain specific pretraining of LLMs – this despite the fact that numerous studies, such as that conducted by Trewartha et al (2022) and Zheng (2022), demonstrate that fine-tuning models with more specific pre-training leads to enhanced performance outcomes. We therefore develop a subdomain-specific conceptual framework for LLMs pretrained on subdomain-specific text corpora, specifically in relation to the analysis of sustainability disclosures for greenwashing detection.

2. Overview of related work

2.1. Sustainability reporting and greenwashing

Sustainability reporting is a term used to describe various forms of reporting on sustainability factors. According to Jestratijevic (2022) et al. sustainability reporting refers to a company's "voluntary, non-financial disclosure of the social and environmental impacts of their business." In contrast, greenwashing is "an umbrella term for a variety of misleading communications and practices that intentionally or not, induce false positive perceptions of an organization's environmental performance" (Nemes et al., 2022). Lyon and Maxwell (2011) define greenwashing as "selective disclosure of positive information about a company's environmental or social performance, without full disclosure of negative information on these dimensions, so as to create an overly positive corporate image".

The detrimental effects of greenwashing on sustainability efforts have been widely acknowledged by numerous industry and regulatory agencies and stakeholders. Testa et al. (2015) affirm this when stating that that greenwashing undermines both "corporate accountability" and "the credibility of environmental initiatives". Greenwashing poses a threat to the precision, dependability, and openness of company sustainability disclosures due to the fundamental difference between a company's disclosure (and signalling) to stakeholders about its ESG performance, and its practices. This discrepancy is described by Steiner et al. (2018) as the "incongruence between the reputational intention and the actual, real sustainability performance of the company." The practice of greenwashing poses a threat to a diverse range of stakeholders, from public entities, to consumers, regulators, shareholders, and potential investors.

2.2. Greenwashing

Greenwashing encompasses a broad spectrum of misleading practices, ranging from outright lies or false claims, to vague statements, selective or omitted disclosures, unsubtantiated claims, empty or exaggerated claims, and even use of environmental jargon that is difficult to understand. (de Freitas Netto et al., 2020; Nemes et al., 2022).

This diversity is acknowledged in the academic literature relating to the field. de Freitas Netto et al (2020) differentiate between firm-level and product level environmental or green claims, whilst also situating greenwashing within extant literature which defines greenwashing in terms of selective disclosures or decoupling behaviours. Nemes et al (2022) develop an integrated greenwashing taxonomy, which include, inter alia, unsubstantiated claims, claims that are vague, broad or poorly understood, claims that exaggerate achievements, and claims that use jargon that consumers are unable to parse or verify. In light of the widely publicised Volkswagen emissions scandal, in which Volkswagen deliberately manipulated vehicle emissions tests results using in-house software, Siano et al (2017) propose an extension of the greenwashing taxonomy to include a new category, "deceptive manipulation".

Examples of greenwashing span the spectrum from the "deceptive manipulation" by Volkswagen, involving information relating to software that was not initially publicly available, to the European Commission's study of product and advertising green claims by companies, which found that that 53.3% of the items analysed "provide vague, misleading or unfounded information on products' environmental characteristics"(Pimonenko et al., 2020)(European Commission, 2023). Other examples include Li et al's (2022)examination of the gap between fossil fuel companies' decarbonisation communications and actions, using keyword analysis, and categorisation of pledges and actions found in company sustainability disclosures.

2.3. Greenwashing detection

Despite the significant negative impact that greenwashing may have, its detection in company sustainability disclosures remains challenging. The increase in company sustainability information makes it more challenging for stakeholders to manually analyse sustainability reports and detect greenwashing (Pimonenko et al., 2020). There is both more information for stakeholders to analyse, and that information has become more complex (Macpherson et al., 2021). The challenge is compounded when such analyses consider multiple companies' reports across multiple time-periods (Ning et al., 2021). (2021)Lastly, the textual and qualitative nature of sustainability reporting has in the past made the analysis of such reports to identify greenwashing more challenging (Luccioni et al., 2020).

However, dramatic advancements in AI tools such as machine learning and natural language processing (NLP) techniques present the means and opportunity to analyse large volumes of complex company sustainability disclosures (Luccioni & Palacios, 2019). The value of NLP in assessing large quantities of text-based sustainability disclosures is reflected in the research of Smeuninx et al (2016) , who applied NLP to a "2.75-million-word corpus" to characterize the language of sustainability reports across specific dimensions. A significant contribution by Kotzian involves the proposition of methodologies for employing AI integrated with machine-learning techniques within the area of Corporate Social Responsibility (CSR) to identify instances of CSR non-compliance.

While greenwashing may take various forms, a crucial first step in the detection of greenwashing in company sustainability disclosures is the identification of specific sustainability-related disclosures; and the second is the identification of environmental claims, commonly referred to "green claims", made by companies within those disclosures (Kobti et al., 2021) (Stammbach et al., 2022). While the identification of sustainability-related disclosures may seem trivial in theory, in practice, such identification is challenging and time consuming due to the volume and complexity of information disclosed by companies across time periods.

2.4. Large language models

The BERT model and its successors, such as ROBERTA, were trained using large general text corpora, such as Wikipedia or other Web sources, which contain information from variety of domains, hereinafter referred to as "general domain" (Zheng et al., 2022). These models provide superior performance in relation to earlier word embedding models, such as Word2Vec or GLOVE, for downstream tasks such as text classification or sentiment analysis (Devlin et al., 2019) (Webersinke et al., 2022) An important limitation of these models is, however, their use of general domain text corpora, which, from a semantic and vocabulary perspective, may differ to the semantics and vocabulary of the dataset assessed for downstream tasks.

Building on these models, Gururangan et al (2020) showed that further pre-training on a specific domain produced better performance on downstream tasks than LLMS trained on general domain text corpora, such as BERT. This in turn has led to a trend of domain-adaptive or domain-specific pre-training within the literature, where BERT architecture based models are subjected to further pre-training using text corpora drawn from a specific domain, hereinafter referred to as the "main-domain". Huang et al. (2023) use the BERT model created by Devlin et al (2019) as a base for further unsupervised pre-training on a text corpus created from analysts' reports in the financial domain. The resulting model, Finbert, is tailored to the finance domain. The model, after being fine-tuned on the downstream task of sentiment analysis, outperforms other deep learning algorithms such as long short-term memory and convolutional neural network LLMs. Other examples include MatBERT, pretrained on the materials science domain; ClimateBert, pre-trained on the climate-related domain, and PubmedBert pre-trained on biomedical domain (Huang et al., 2023) (Gu et al., 2021) (Webersinke et al., 2022).

These domain-specific pre-trained models have been shown to provide superior performance, when compared to general domain models fine-tuned only for downstream tasks, when tackling downstream tasks such as text classification or sentiment analysis (Huang et al., 2023). Sanchez et al (2022) find that pre-training BERT models on specific domain text data or corpora yields superior performance to BERT models trained on general corpora and fine-tuned for downstream tasks.

2.5. Pre-trained LLMs in the sustainability domain

Luccioni et al (2020) researched the use of NLP to quickly identify sustainability disclosures using the Task Force on Climate-Related Financial Disclosures (TCFD) framework, and in so doing developed an NLP tool, Climate QA to meet that objective. To do so, a RoBERTa model was pretrained on a text corpus that included both financial and sustainability information, sourced from Global Reporting Initiative and Edgar databases. Bingler et al (2021) developed and used a pre-trained RoBERTa based language model, ClimateBert, to identify TCFD disclosures in more than eight-hundred companies applying TCFD. The model used a text corpus consisting of company webpages, TCFD reports, sustainability reports, and annual reports. Webersinke et al developed a different ClimateBert model, pre-trained on specific and general climate-related text. Cojoianu et al (2020) as part of their research developed Greenwatch.ai to monitor the authenticity of climate-friendly green claims relating to greenhouse gas (GHG) emissions.

The variation in text corpora used for each of these studies illustrates the diversity of the sustainability-domain from a language perspective, and similarly, how large and diverse even main domain corpora used to train these models can be.

2.6. Subdomains

Text corpora used for pre-training LLMs, even if related to a main domain, may exhibit significant internal heterogeneity or consist of subdomains, which in essence represent a different data distribution to that of the main domain (Zheng et al., 2022). As noted by Aharoni et al (2020), language exhibits substantial variability across classes, categories, and topics.

Subdomains may arise based on the source of the text. For example, text corpora from sustainability-related news articles differ to academic research articles, which in turn differ to company-created sustainability reports and analysts' reports. Another source of subdomains relates to taxonomies, typologies, and topics. For example, text corpora drawn from sustainability-related disclosures created in terms of the TCFD disclosures may exhibit differences to those drawn from Global Reporting Initiative Standards. Each of these specific elements, such as text source or typology, represent a subdomain within the broader climate domain. Another example relates to sub-categories within the climate disclosures domain, such as biodiversity, renewable energy, water and emissions.

Despite these linguistic nuances and their possible impact on the performance of LLMs within a specific domain, current literature does not provide sufficient resolution on the potential to extend the pre-training of pre-trained LLMs using subdomain-specific corpora. To address this, we develop a conceptual framework for subdomain-specific pretraining of BERT based LLMs using sustainability subdomain-specific text corpora.

2.7. Summary

Extant literature indicates that AI tools such as LLMS provide the potential for the rapid, automatic identification of specific green claims within company sustainability disclosures. This potential exists within the confines of the current capabilities of LLMs, as well the data available, such as publicly available company sustainability disclosures. In addition, because of the range of types and forms of greenwashing, including deceptive manipulation examples such as the Volkswagen emissions scandal, LLMs do not provide a panacea for greenwashing detection. However, a crucial first step on the road to reliable and efficient greenwashing detection is automated identification of green claims within large volumes of sustainability disclosures, for which a conceptual framework utilising LLMS for targeted, sub-domain specific identification of green claims in this regard is developed.

3. Research methodology: Conceptual Framework

The conceptual framework extends the previous attempts discovered in the literature relating to domain-specific pretraining. The goal is the creation of a pre-trained LLM, further pretrained on a subdomain specific text corpus – hereinafter referred to as an SPM – capable of identifying sustainability-related statements that are specific to the subdomain. The sequencing workflow for the model is shown in Figure 1 below:

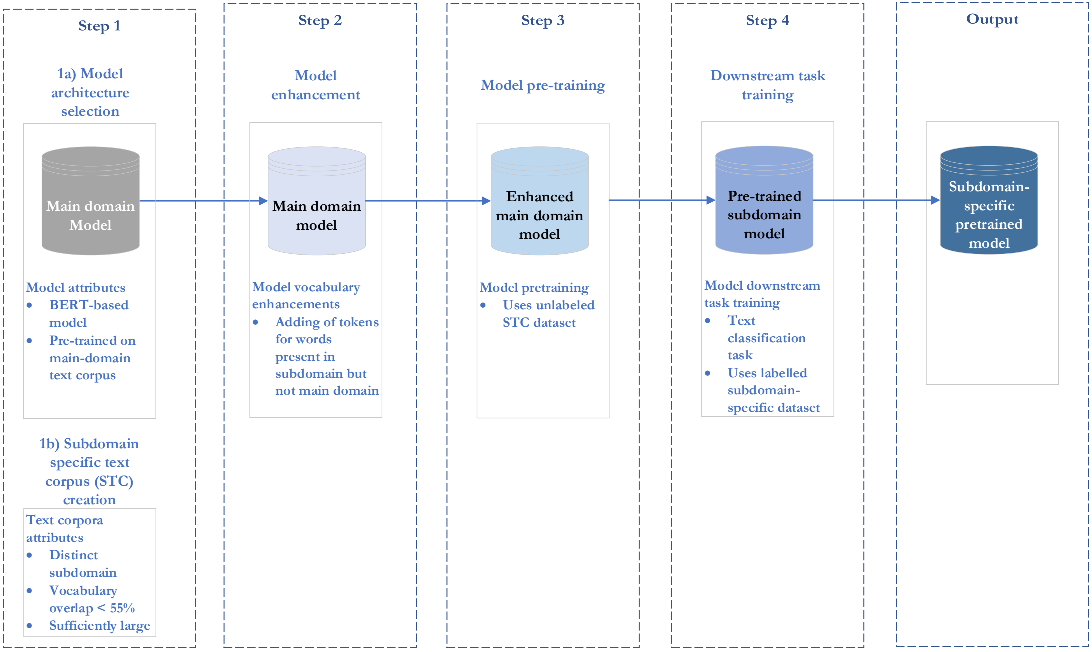
Step 1
1a) Model architecture selection
Main domain Model
Model attributes
• BERT-based model
• Pre-trained on main-domain text corpus
1b) Subdomain specific text corpus (STC) creation
Text corpora attributes
• Distinct subdomain
• Vocabulary overlap < 55%
• Sufficiently large
Step 2
Model enhancement
Main domain model
Model vocabulary enhancements
• Adding of tokens for words present in subdomain but not main domain
Step 3
Model pre-training
Enhanced main domain model
Model pretraining
• Uses unlabeled STC dataset
Step 4
Downstream task training
Pre-trained subdomain model
Model downstream task training
• Text classification task
• Uses labelled subdomain-specific dataset
Output
Subdomain-specific pretrained model

Figure 1: Sequencing workflow for model

3.1. Text corpora creation

The first step in the conceptual model is the selection of an appropriate subdomain text corpora (STC). Previous literature reflects models trained on text corpora from specific main domains (e.g. materials science domain, financial domain etc.) rather than subdomains. For example, Finbert is trained on the general financial domain, rather than on a specific subdomain of that domain, such as IFRS disclosures.

Given the significant effect of the choice of text corpora on the training and performance of a selected model demonstrated by other studies, the definition of the characteristics of the subdomain is critical. Therefor for this step, subdomains within sustainability disclosures are constructed. The STC should reflect a distinct subdomain of the specific domain. For example, text corpora drawn from sustainability disclosures created in terms of Global Reporting Initiative (GRI) standards would represent a subdomain within the broader sustainability disclosures main domain. The text corpora should also be sufficiently large to enable better model performance and to make full use of the selected model architecture's capabilities.

The degree of vocabulary overlap of the subdomain text corpora, and the main domain text corpora previously used to train the selected model should be assessed. Consistent with the approaches of previous domain-specific pretraining, a vocabulary overlap of no more than 55% is recommended to ensure sufficient differences in vocabulary between the subdomain and main domain texts (Sanchez & Zhang, 2022) (Gururangan et al., 2020) (Huang et al., 2023; Webersinke et al., 2022).

3.2. Model architecture selection

A concurrent first step in the conceptual framework is the selection of an appropriate base model architecture to be used for the creating on the extended model. This conceptual framework builds on the previous demonstrated successes of pretrained transformer based LLMS. This step therefore involves the selection of an appropriate pre-trained transformer based LLM that has previously been pre-trained on the sustainability domain (e.g. ClimateBert, ESGBert etc.). The selected model (main domain model or MDM) would then be further pre-trained on a subdomain-specific text corpora specifically created to extend the selected model.

3.3. Model enhancement

The chosen model (MDM) is further enhanced by increasing the model vocabulary. This consists of the inclusion of words specific to the STC not present in the original MDM vocabulary. This is consistent with the work of Webersinke et al. (2022) who find model performance improvements as a result of using this technique. Specifically, tokens for words present in the STC and not in the specific domain are added to the model's vocabulary.

3.4. Model pre-training

The next step is the pre-training of the model. The selected enhanced vocabulary MDM is trained on the large, unlabeled STC, thereby extending the approach first proposed by Devlin et al (2019) and modified by Gururangan et al. (2020). In essence, this involves subdomain-specific pretraining of the selected model – which in turn was

© 2023 The Authors. Journal Compilation © 2023 European Center of Sustainable Development.

previously trained on a main domain text corpus. The model is also pre-trained on a downstream task, namely, text classification. The results and performance of the extended model is assessed, based on the loss on a validation set from the STC. The results and performance are assessed and compared in relation to the results and performance of the MDM on the same task, as well as the BERT model underlying the MDM.

3.5. Downstream task training

Lastly, the model is trained (fine-tuned) on a downstream task namely binary text classification. We select binary text classification as our goal is for the model to classify between text that relates to the subdomain, and text which does not form part of or relate to the subdomain.

This step requires a labelled subdomain-specific dataset, with each class labelled. Such a dataset may be constructed or sourced from existing public repositories. For example, the open-source LLM repository, Huggingface.co, contains a number of LLMs with sustainability related labelled datasets that could be used. Again, the performance of the model in executing the downstream task may be compared to that of the MDM and the BERT model underlying the MDM.

As noted by Devlin et al (2019), the hyperparameters for fine-tuning would depend on the task, though the authors suggest epochs of 2, 4, or 6 with batch sizes of 16 or 32 for underlying original BERT model. However, in practice, we find a wide range of hyperparameters across pre-trained models in the literature (Friederich et al., 2021; Stammbach et al., 2022) , and therefore suggest hyperparameters that are tailored to both the model architecture, text corpus, processing power availability and requirements, and planned environmental impact.

3.6. Model outputs

The results of the MDM pretrained on the subdomain specific text corpora are then evaluated and compared from an F1 score perspective and a cross entropy loss perspective for the binary text classification task across the three models evaluated:

- BERT Model (general domain)
- MDM (main domain)
- SPM (subdomain-specific and fine-tuned for text classification downstream task)

Other models may also be chosen for comparison or for a baseline.

4. Discussion

The aim of the conceptual framework presented is the creation of the SPM i.e. the use of a model architecture previously trained on the main domain, further pre-trained on the subdomain, and fine-tuned for the downstream task of text classification. This results in a model, SPM, trained more specifically to identify items within a specific sustainability disclosure subdomain. This is the first step toward identifying green claims, as once sustainability disclosures are classified as being within the subdomain by the SPM, this narrows the search for green claims within that subdomain. As the SPM model architecture is an LLM, this paves the way for large datasets to be efficiently parsed to allow for such classification and to reduce the time required to identify green claims within subdomain-specific sustainability disclosures. The result of the text classification task is, in essence, a labelled dataset showing text which is classified as in-subdomain or outside the subdomain. That labelled dataset may be used for further studies relating to the use of further pre-trained LLMS for green claim detection.

5. Conclusion

5.1. Contribution

We propose a conceptual framework for an extended domain-specific pre-trained model, further trained on a specific subdomain. We do so in order to tackle the issue of greenwashing within sustainability disclosures. Within relevant literature, we find that greenwashing is often difficult to detect as a result of the vast volume and complexity of sustainability disclosures, and because there exists a wide-range of greenwashing or green claim typologies. However, we also find within the literature that AI tools such as LLMs provide the potential to automate the identification of green claims. The accuracy of the outputs of such models relies on the correct choice of model, as well as the process by which such models are trained. Crucially, the general domain or domain-specific text corpora used for training LLMs significantly affect the output of these models. We find within the literature that domains for text corpora used for pre-training such LLMs are amorphous and often not clearly delineated. We therefore propose, as part of the conceptual framework, the creation of subdomain specific text corpora, taking into account the heterogeneity of general domain, or domain-specific, text corpora.

The conceptual framework provides the following contributions:

- the extension of existing BERT-based model architectures previously pre-trained on sustainability-related disclosures,
- a conceptual understanding of the criteria required for selection of subdomain specific text,
- the creation of subdomain-specific text corpora for model pre-training,
- the creation of a subdomain specific BERT-based model pre-trained on subdomain specific text (the SPM), and
- an SPM fine-tuned for the downstream task of text classification, capable of identifying subdomain-specific text.

These contributions extend those found in current literature. Finally, we envisage that the output of the text classification task of the SPM is a labelled dataset, providing binary class labels indicating whether specific text is in-subdomain or not.

5.2. Limitations

An inherent limitation of the model, which results from applying the developed conceptual framework, is that the model created as a result of further pre-training, is based upon the training dataset, which may initially be limited. However, we expect the size and availability of such datasets to expand over time as the framework is applied in further research. A second limitation is that BERT models are based on single word or subword associations. That being said, an inherent benefit is that BERT models generate bi-directional contextualised word embeddings. Lastly, we note that any model created by applying the conceptual framework would not be suitable for detecting green claims

© 2023 The Authors. Journal Compilation © 2023 European Center of Sustainable Development.

relating to all forms of greenwashing, but would, given the nature of the subdomain and conceptual framework, relate to green claims arising from specific subdomains.

6. Acknowledgments

The authors would like to acknowledge the professional support of the University of Johannesburg's Johannesburg Business School.

References

Aharoni, R., & Goldberg, Y. (2020). *Unsupervised Domain Clusters in Pretrained Language Models*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.692

Bingler, J. A., Kraus, M., & Leippold, M. (2021). *Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures*. Finance Research Letters, 47(B), 102776. https://doi.org/10.1016/j.frl.2022.102776

Cojoianu, T., Hoepner, A., Ifrim, G., & Lin, Y. (2020, June 15). Greenwatch-shing: Using AI to Detect Greenwashing. AccountancyPlus - CPA Ireland. https://ssrn.com/abstract=3627157

de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. da L. (2020). *Concepts and forms of* greenwashing: a systematic review. Environmental Sciences Europe, 32(1), 1–12. https://doi.org/10.1186/s12302-020-0300-3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

European Commission. (2020). *Environmental performance of products & businesses – substantiating claims*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12511-Environmental-performance-of-products-businesses-substantiating-claims/public-consultation_en

European Commission. (2023, March 22). *Proposal for a Directive on green claims*. https://environment.ec.europa.eu/publications/proposal-directive-green-claims_en

Friederich, D., Kaack, L. H., Luccioni, A., & Steffen, B. (2021). *Automated Identification of Climate Risk Disclosures in Annual Corporate Reports*. https://doi.org/10.48550/arXiv.2108.01415

Gu, Y. U., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H., & Gu, Y. (2021). *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing*. ACM Transactions on Computing for Healthcare, 3(1), 1–23. https://doi.org/10.1145/3458754

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020, July). *Don't Stop Pretraining: Adapt Language Models to Domains and Tasks*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2004.10964

Huang, A. H., Wang, H., & Yang, Y. (2023*). FinBERT: A Large Language Model for Extracting Information from Financial Text*. Contemporary Accounting Research, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

In, S. Y., & Schumacher, K. (2021). *Carbonwashing: A New Type of Carbon Data-Related ESG Greenwashing*. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.3901278

Jestratijevic, I., Uanhoro, J. O., & Creighton, R. (2022). *To disclose or not to disclose? Fashion brands' strategies for transparency in sustainability reporting*. Journal of Fashion Marketing and Management, 26(1), 36–50. https://doi.org/https://doi.org/10.1108/JFMM-09-2020-0182

JSE Limited. (2021). *Leading the way for a better tomorrow JSE Sustainability Disclosure Guidance*. https://www.jse.co.za/sites/default/files/media/documents//JSE%20Sustainability%20Disclosure%20Guidance%202021_SPS.pdf

Kobti, J., Schmitt, V., & Woloszyn, V. (2021). *Towards Automatic Green Claim Detection; Towards Automatic Green Claim Detection*. Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation. https://doi.org/10.1145/3503162.3503163

Li, M., Trencher, G., & Asuka, J. (2022). *The clean energy claims of BP, Chevron, ExxonMobil and Shell: A mismatch between discourse, actions and investments*. PLoS ONE, 17(2), e0263596. https://doi.org/10.1371/journal.pone.0263596

Luccioni, A., Baylor, E., & Duchene, N. (2020). *Analyzing Sustainability Reports Using Natural Language Processing*. https://doi.org/10.48550/arxiv.2011.08073

Luccioni, A., & Palacios, H. (2019). *Using Natural Language Processing to Analyze Financial Climate Disclosures*. https://s3.us-east-1.amazonaws.com/climate-change-ai/papers/icml2019/34/paper.pdf

Lyon, T. P., & Maxwell, J. W. (2011*). Greenwash: Corporate Environmental Disclosure under Threat of Audit*. Journal of Economics & Management Strategy, 20(1), 3–41. https://doi.org/10.1111/J.1530-9134.2010.00282.X

Macpherson, M., Gasperini, A., & Bosco, M. (2021). *Implications for Artificial Intelligence and ESG Data*. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.3863599

Nemes, N., Scanlan, S. J., Smith, P., Smith, T., Aronczyk, M., Hill, S., Lewis, S. L., Montgomery, A. W., Tubiello, F. N., & Stabinsky, D. (2022). *An Integrated Framework to Assess Greenwashing*. Sustainability, 14(8), 4431. https://doi.org/10.3390/su14084431

Ning, X., Yim, D., & Khuntia, J. (2021). *Online Sustainability Reporting and Firm Performance: Lessons Learned from Text Mining*. Sustainability, 13(3), 1069. https://doi.org/10.3390/SU13031069

Pimonenko, T., Bilan, Y., Horák, J., Starchenko, L., & Gajda, W. (2020). *Green Brand of Companies and Greenwashing under Sustainable Development Goals*. Sustainability, 12(4), 1679. https://doi.org/10.3390/SU12041679

Sanchez, C., & Zhang, Z. (2022). *The Effects of In-domain Corpus Size on pre-training BERT*. https://doi.org/10.48550/arXiv.2212.07914

Siano, A., Vollero, A., Conte, F., & Amabile, S. (2017). *"More than words": Expanding the taxonomy of greenwashing after the Volkswagen scandal*. Journal of Business Research, 71, 27–37. https://doi.org/10.1016/j.jbusres.2016.11.002

Smeuninx, N., De Clerck, B., & Aerts, W. (2016). *Measuring the Readability of Sustainability Reports: A Corpus-Based Analysis Through Standard Formulae and NLP*. International Journal of Business Communication, 57(1), 52–85. https://doi.org/10.1177/2329488416675456

Stammbach, D., Zurich, E., Webersinke, N., Bingler, J. A., Kraus, M., & Leippold, M. (2022). *A dataset for detecting real-world environmental claims*. https://doi.org/10.3929/ethz-b-000568978

Steiner, G., Geissler, B., Schreder, G., & Zenk, L. (2018). *Living sustainability, or merely pretending? From explicit self-report measures to implicit cognition*. Sustainability Science, 13(4), 1001–1015. https://doi.org/10.1007/s11625-018-0561-6

Testa, F., Boiral, O., & Iraldo, F. (2015*). Internalization of Environmental Practices and Institutional Complexity: Can Stakeholders Pressures Encourage Greenwashing?*. Journal of Business Ethics, 147(2), 287–307. https://doi.org/10.1007/S10551-015-2960-2

Trewartha, A., Walker, N., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K. A., Ceder, G., & Jain, A. (2022*). Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science*. Patterns, 3(4), 100488. https://doi.org/10.1016/j.patter.2022.100488

Webersinke, N., Kraus, M., Bingler, J., & Leippold, M. (2022). *CLIMATEBERT: A Pretrained Language Model for Climate-Related Text*. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4229146

Zheng, Z., Lu, X. Z., Chen, K. Y., Zhou, Y. C., & Lin, J. R. (2022). *Pretrained domain-specific language model for natural language processing tasks in the AEC domain*. Computers in Industry, 142, 103733. https://doi.org/10.1016/j.compind.2022.103733

© 2023 The Authors. Journal Compilation © 2023 European Center of Sustainable Development.