Detecting Greenwashing Hints in ESG Reports: Linguistic and Claim Analysis in Traffic-Related Emissions Disclosures

Johannes Florstedt and **Jonas Fahlbusch** and **Moritz Sontheimer**
Technische Universität Berlin
florstedt@campus.tu-berlin.de

Abstract

Detecting greenwashing in corporate Environmental, Social, and Governance (ESG) reports presents challenges due to data scarcity and ambiguity, particularly concerning complex topics like traffic emissions. This paper introduces a machine learning framework to identify potential greenwashing indicators by analyzing linguistic patterns and claim substantiation in 150 ESG reports from German DAX companies, 2020-2023. We evaluate sentiment polarity (VADER), linguistic specificity (ClimateBERT), and internal claim verification (Sentence-BERT). Key findings reveal two complementary signals suggesting areas for scrutiny: first, an external discrepancy where high reporting positivity coincides with lower relative external performance proxy scores (Refinitiv Emission Score), identifying specific firms potentially overstating performance; second, an internal inconsistency where low linguistic specificity correlates with weak internal claim substantiation, indicating potential *cheap talk*. While automated external claim verification proves difficult, the framework provides quantitative indicators to help stakeholders prioritize the assessment of ESG reports in the critical traffic sector.

1 Introduction

Heightened stakeholder demand for corporate Environmental, Social, and Governance (ESG) transparency has driven a significant increase in sustainability reporting. This trend, however, is accompanied by growing concerns regarding *greenwashing* – the practice where organizations portray their environmental performance more favorably than warranted by their actions (European Securities and Markets Authority, 2023). Evidence suggests this is a considerable issue, with studies finding misleading claims prevalent across various markets (UK Competition and Markets Authority, 2021; Australian Competition and Consumer Commission, 2023). The systematic detection of greenwashing is complicated by the lack of a universally standardized definition and the scarcity of reliably labeled datasets, which limits the applicability of conventional supervised machine learning methods (Calamai et al., 2025). As a result, research increasingly utilizes Natural Language Processing (NLP) techniques to analyze the extensive textual content of ESG reports for linguistic and semantic patterns that might indicate misrepresentation (Bingler et al., 2022; Vinella et al., 2024).

These detection challenges are particularly pronounced in the context of disclosures related to **traffic and transportation emissions**. This area holds critical importance due to the transport sector's substantial contribution to global greenhouse gas emissions (Shukla et al., 2022). Disclosures often involve complex Scope 3 emissions data (e.g., logistics, business travel, employee commuting), which are notoriously difficult to measure, report consistently, and verify externally (Berthe et al., 2025). This inherent complexity and potential data opacity may provide avenues for companies to engage in greenwashing within their traffic-related narratives (Robinson, 2022).

This paper presents and evaluates an ML framework specifically developed to identify potential greenwashing indicators within these traffic-related disclosures. We investigate the interplay between sentiment polarity and linguistic specificity, and their relationship to external performance proxies. Furthermore, we assess the degree to which specific claims are substantiated by internal textual evidence using semantic similarity techniques, while also exploring the practical challenges associated with attempts at external verification. A key objective is to understand how linguistic analysis and claim verification compare and potentially complement each other in highlighting potential greenwashing risks. Our aim is not to achieve definitive greenwashing classification, but rather to provide quantitative, data-driven indicators that enable stakeholders to prioritize and focus their scrutiny efforts more effectively.

2 Background and Related Work

Greenwashing often involves the strategic deployment of language, such as using excessively positive framing or ambiguous terminology, potentially to divert attention from unfavorable environmental performance (Delmas and Burbano, 2011). Indicators commonly associated with greenwashing include communication that appears overly positive relative to actual performance, the use of vague or non-specific language lacking concrete details, and the presentation of unsubstantiated claims regarding

environmental benefits (European Securities and Markets Authority, 2023; European Parliament, 2023). The traffic sector, characterized by its complex and often difficult-to-verify Scope 3 emissions footprint (Berthe et al., 2025), represents an area susceptible to such practices, as highlighted by public controversies involving the automotive and aviation industries (Robinson, 2022; Plucinska, 2023).

Existing ML approaches for detecting greenwashing signals are diverse. Supervised learning methods frequently grapple with the scarcity of labeled data, sometimes employing synthetically generated labels (Vinella et al., 2024), weak supervision based on aggregated firm-level scores (Sharma et al., 2024), or requiring substantial manual annotation efforts (Bingler et al., 2024). Unsupervised techniques often focus on identifying discrepancies between corporate narratives and external benchmarks. Common strategies found in the literature involve comparing report content and tone against quantitative ESG performance scores (Chen and Ma, 2024; Lagasio, 2024), while others analyze alignment with public discourse, such as media sentiment and topic coverage (Lipenkova et al., 2023; Zhao et al., 2023).

Common linguistic features analyzed include sentiment polarity (Chen and Ma, 2024; Zhao et al., 2023) and the degree of linguistic specificity versus vagueness (Bingler et al., 2024; Vinella et al., 2024). Domain-specific language models, particularly ClimateBERT (Webersinke et al., 2022), have shown improved effectiveness in analyzing the specialized vocabulary and context of climate-related text compared to general models (Bingler et al., 2024; Trajanov et al., 2023). While stylistic analysis is relatively common, the systematic evaluation of internal claim substantiation within reports seems less explored in the context of greenwashing detection. Automated fact-checking tools like LOKI (Li et al., 2024) offer potential pathways for external verification but face considerable hurdles when applied to the complex and nuanced nature of ESG claims (Leippold et al., 2024).

Our work integrates insights from these varied approaches. We employ a primarily unsupervised framework focusing on quantifiable indicators (positivity, specificity, internal consistency) tailored specifically to the traffic domain. We utilize accessible tools, including ClimateBERT variants and Sentence-BERT, and importantly, compare derived communication patterns against an external performance proxy.

3 Methodology

Our analysis is based on a corpus of 150 English-language ESG reports collected from German DAX companies for the years 2020 through 2023. Text was extracted from PDF documents using the `Kreuzberg` tool, chosen for its ability to produce cleaner textual output suitable for NLP tasks compared to some standard libraries. A multi-pipeline framework was implemented to analyze disclosures related to traffic emissions.

1. Filtering Traffic-Related Content: The core analysis focused on relevant text segments identified through a sequential filtering process applied to 500-character chunks (with a 20-character overlap, intended to preserve context across boundaries). First, the ClimateBERT Detector model (Bingler et al., 2024) classified chunks based on climate relevance, retaining those exceeding a confidence score threshold of 0.5. Second, these climate-relevant chunks were further filtered using a custom-developed traffic lexicon (keywords including 'fleet', 'electric vehicle', 'transport', 'fuel', 'logistics', 'business travel', 'commuting', 'aviation', 'shipping') to isolate segments specifically discussing traffic-related issues. This filtering cascade aimed to focus the analysis efficiently on the most pertinent text passages.

2. Language Analysis Module: This module evaluated the stylistic properties of the filtered chunks. Linguistic specificity was assessed using the ClimateBERT Specificity model (Bingler et al., 2024), classifying each chunk as either 'specific' (containing concrete data, metrics, or detailed actions) or 'non-specific' (general, vague statements). The proportion of 'specific' chunks per document was calculated to derive a document-level Specificity Score (0-100). Sentiment polarity was determined using VADER (Hutto and Gilbert, 2014), selected for its capability to handle contextual nuances like negation and intensifiers found in narrative text. The average VADER compound score across a document's filtered chunks was linearly transformed into a Positivity Score (0-100 scale, where 50 indicates neutrality).

3. Claim Verification Module: This module examined the substantiation of claims. For *internal verification*, potential claim sentences (identified heuristically via modal/assertive keywords + traffic terms) and potential proof sentences (identified via evidence-related keywords) were extracted. Sentence-BERT (Reimers and Gurevych, 2019), specifically the efficient `all-MiniLM-L6-v2` model, generated embeddings for claims and proofs. Cosine similarity was computed between each claim and all potential proof sentences from the same report. The highest similarity score to a non-identical proof sentence was considered the measure of internal support. An average Internal Claim Score (0-100) per document summarized this semantic coherence. While pragmatic, these heuristic extraction steps influence the inputs to the similarity assessment and represent a known limitation. For *external verification*, a limited, exploratory analysis was performed on a small set of claims (prioritizing those with low internal scores) using the public LOKI web interface (Li et al., 2024) to investigate the feasibility and challenges of automated web-based verification.

4. Performance Proxy: We utilized the Refinitiv Emission Score (0-100), accessed via the Refinitiv Eikon database, as an external proxy for corporate environmental performance. This score was chosen due to its focus on emissions within the broader ESG context, its consideration of Scope 1-3 emissions data (though not specifically isolating traffic), and its methodology

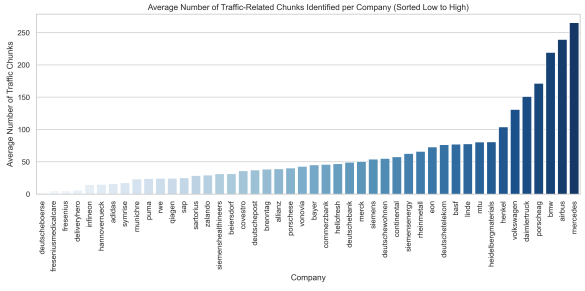

Figure 1: Average Number of Traffic-Related Chunks Identified per Company (Sorted Low to High).

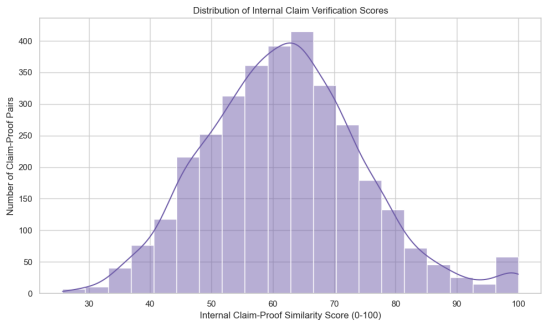

Figure 2: Distribution of Internal Claim Verification Scores (Individual Claim-Proof Pairs).

which integrates company disclosures with external controversy screening, offering a relatively comprehensive benchmark available for this study (LSEG Data & Analytics, 2024).

5. Analysis: The core analysis involved calculating correlations (Pearson, Spearman) between the company-level average indicators (Positivity, Specificity, Claim Score, Refinitiv Score). Visual discrepancy analysis using scatterplots was employed to identify specific companies exhibiting patterns potentially indicative of greenwashing risk relative to the observed trends.

4 Results

4.1 Reporting Intensity and Linguistic Style

The analysis revealed substantial variation in the extent to which companies elaborated on traffic-related climate issues. Figure 1 illustrates the wide range in the average number of filtered, relevant text chunks per company, with firms in transport-intensive sectors generally providing more content, though significant intra-sector variation exists. This variability in reporting intensity affects the statistical robustness of metrics for companies with minimal relevant text.

On average, the linguistic style within these disclosures tended towards positive sentiment (mean Positivity Score 70.5) and moderate specificity (mean Specificity Score 65.4%). Importantly, no statistically significant correlation was found between a company's average positivity score and its average specificity score, suggesting these represent largely independent dimensions

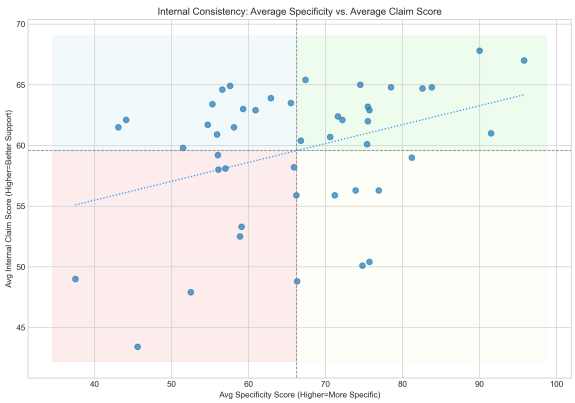

Figure 3: Internal Consistency: Avg Specificity vs. Avg Claim Score. Bottom-left (red) suggests potential inconsistency.

of communication style in this context.

4.2 Internal Claim Verification

The internal claim verification assessed semantic similarity between identified claims and potential supporting sentences within the same document. The distribution of individual claim-proof similarity scores (Figure 2) was centered around a mean of 62.0 (0-100 scale). This suggests that, typically, claims found moderately related textual evidence within the report. However, the broad distribution of scores indicates varying degrees of internal substantiation, with some claims finding strong semantic links while others lacked clear support.

4.3 Correlation Analysis and Discrepancies

Analysis of company-level average indicators over the 2020-2023 period revealed significant relationships:

Internal Consistency Signal: A statistically significant, moderate positive correlation was observed between Average Specificity and Average Internal Claim Score (Spearman rho=0.310, p=0.036; Pearson r=0.363, p=0.013). This key finding indicates that companies employing more specific and detailed language in their traffic disclosures also tend to exhibit stronger internal semantic coherence, meaning their claims are better supported by other statements within the report (visualized in Figure 3). This linkage between linguistic style and internal evidence provides a measurable indicator of reporting consistency.

External Alignment Signal: Average Positivity showed a significant positive correlation with the Average Refinitiv Emission Score proxy (Spearman rho=0.332, p=0.024; Pearson r=0.472, p=0.001). On average, companies assessed as having better emissions performance (via the proxy) tended to use more positive language in their traffic-related sections. No significant correlations were found between Specificity or Claim Score and the Refinitiv score. Analyzing discrepancies from the main Positivity-Refinitiv trend is crucial here. Figure 4 identifies companies (marked with red

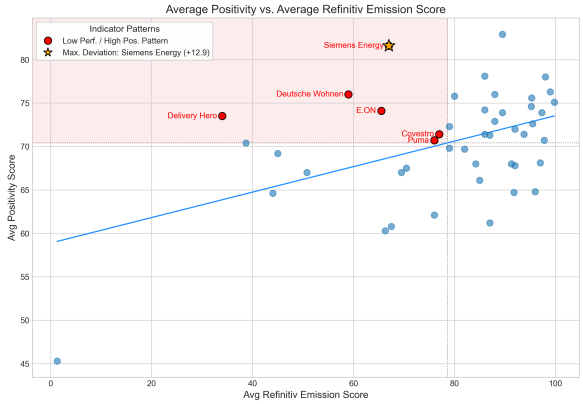

Figure 4: External Discrepancy: Avg Positivity vs. Avg Refinitiv Score. Upper-left (red circles) flags potential risk.

circles: Delivery Hero, Dt. Wohnen, E.ON, Covestro, Puma, Siemens Energy) situated in the upper-left area, characterized by high reporting positivity despite lower relative performance proxy scores. This pattern aligns with theoretical greenwashing risk profiles (Delmas and Burbano, 2011). Siemens Energy exhibited the largest positive deviation from the overall trend line.

5 Discussion

The framework applied in this study provides quantitative indicators and reveals communication patterns that can aid in the systematic identification of potential greenwashing risks within the challenging domain of traffic-related ESG disclosures.

A key finding is the significant positive correlation between linguistic specificity and internal claim substantiation, serving as an informative **internal consistency check**. Reports characterized by both vague language (low specificity) and weak internal support for claims (low internal claim score) – corresponding to the bottom-left quadrant in Figure 3 – represent a pattern suggesting potential concern. This combination might indicate instances of *cheap talk*, where commitments are stated vaguely or lack concrete detail and verifiable grounding within the report itself. Identifying such internal inconsistencies allows analysts to focus attention on disclosures that appear potentially insubstantial or poorly documented.

The analysis of **external alignment** revealed that higher reporting positivity, on average, correlated with better assessed performance according to the Refinitiv Emission score proxy. This suggests that positive communication is not solely the domain of poorer performers. However, the true value lies in identifying deviations from this general trend. Companies exhibiting high positivity relative to their performance proxy score (upper-left area in Figure 4) display a pattern consistent with established definitions of greenwashing risk (Delmas and Burbano, 2011) – potentially creating an impression of sustainability leadership not fully matched by the external benchmark. While acknowledging the proxy's limitations, this discrepancy analysis provides a data-driven basis for flagging specific companies (e.g., those marked red, and particularly the largest deviator, Siemens Energy) whose optimistic framing merits closer qualitative investigation.

Importantly, these two signals – internal inconsistency and external discrepancy – offer **complementary diagnostic perspectives**. A report might be flagged by one signal but not the other. Using both allows for a more comprehensive risk assessment. For instance, a report could be internally coherent but externally misaligned, or vice versa. This multi-signal approach enhances the ability of stakeholders (investors, regulators, researchers) to prioritize their limited resources, directing in-depth qualitative analysis and verification efforts towards the reports and companies exhibiting the most salient risk indicators. Effective use of these signals can streamline the otherwise daunting task of evaluating large volumes of ESG reporting.

The research also highlights persistent **methodological challenges**. NLP models for specificity or sentiment analysis are not infallible and can misinterpret context, particularly in formal reporting language. Heuristic methods for filtering content or extracting claims, while computationally efficient, inherently limit precision and recall. The exploratory external verification using LOKI confirmed substantial difficulties in reliably automating fact-checking for nuanced ESG claims via standard web search; the tool struggled with context, comparative language, and source reliability, limiting the utility of its outputs without careful manual validation. These limitations underscore that automated tools are best viewed as aids to, rather than replacements for, critical human analysis.

Despite these limitations, the framework provides a valuable advancement by offering structured, data-driven indicators. It moves the assessment of reporting credibility beyond subjective interpretation towards identifying specific, quantifiable patterns associated with potential greenwashing risk in the critical domain of corporate traffic emissions reporting.

6 Conclusion

This research developed and evaluated a machine learning framework to identify potential greenwashing indicators in traffic-related ESG disclosures. By analyzing linguistic style (positivity, specificity) and internal claim substantiation, and correlating these with an external performance proxy, we identified two complementary signals meriting further scrutiny: 1) *Internal inconsistency* (low specificity combined with weak internal claim support), potentially indicating *cheap talk*, and 2) *External discrepancy* (high reporting positivity relative to assessed performance). These quantitative indicators provide stakeholders with a data-driven methodology to prioritize the assessment of reporting credibility, contributing to efforts towards greater transparency and

accountability in this vital sustainability domain.

Limitations

The findings should be interpreted considering several limitations. **Scope and Data:** The analysis focused on English-language reports from German DAX companies (2020-2023) and specifically on traffic-related disclosures, limiting broader generalizability. The lack of a standardized, labeled greenwashing dataset necessitated using proxy indicators. A key limitation is the reliance on the Refinitiv Emission Score as an external performance proxy. This score reflects overall corporate emissions performance and is not specific to traffic-related activities. Comparing communication patterns within the traffic domain to this aggregate score assumes a degree of correlation between general performance and specific reporting, an assumption which requires caution as traffic-specific trends might diverge. Furthermore, any ESG score represents a specific assessment methodology with its own potential biases. **Methodology and Tools:** Standard PDF-to-text conversion potentially introduced noise and missed non-textual information. Resource constraints led to heuristic methods for filtering and claim/proof identification, impacting precision/recall. The accuracy of employed NLP models (e.g., ClimateBERT Specificity, VADER) affects result reliability. Sentiment analysis tools may misinterpret neutral technical language. **Verification Challenges:** Internal claim scores reflect semantic similarity based on heuristically extracted sentences, not guaranteed factual accuracy. Exploratory external verification using the public LOKI interface revealed significant limitations in reliably assessing specific, complex ESG claims against web data due to issues with context, source evaluation, and reasoning capabilities. **Conceptual Ambiguity:** Defining and operationalizing greenwashing remains inherently challenging, limiting objective measurement. The identified indicators signal risk, not definitive proof of intent.

Future Work

The implementation of the EU's Corporate Sustainability Reporting Directive (CSRD), mandating standardized, machine-readable formats (XHTML/iXBRL) and detailed Scope 3 data (European Parliament and Council of the European Union, 2022; European Commission, 2023), offers significant opportunities. Future research should leverage these formats to potentially overcome current text extraction issues and enable more robust analysis of granular data. Applying this framework to CSRD reports will allow investigation into whether reporting patterns evolve under this stricter regulation.

Methodological advancements could involve replacing heuristic steps with more sophisticated NLP techniques for claim extraction (cf. Stammbach et al., 2022) and contextual filtering, possibly using semantic topic modeling. Refining specificity analysis (e.g., distinguishing numerical vs. qualitative detail) could yield richer insights. Addressing the challenge of reliable external verification remains crucial, likely requiring integration of curated authoritative databases or domain-specific knowledge graphs, moving beyond generic web search tools. Expanding this analytical approach to other sectors, regions, and ESG topics will further contribute to understanding and enhancing corporate sustainability reporting credibility.

References

Australian Competition and Consumer Commission. 2023. Greenwashing by businesses in australia. findings of the accc's internet sweep of environmental claims. Last accessed on April 11, 2025.

Tegwen Berthe, Sandrine Nguiakam, and Mathieu Journeau. 2025. Measuring scope 3 emissions: implications challenges for investors. Technical report, Amundi Research Center.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

Tom Calamai, Oana Balalau, Théo Le Guenedal, and Fabian M. Suchanek. 2025. Corporate greenwashing detection in text – a survey. arXiv preprint arXiv:2502.07541.

Yan Chen and Ding Ma. 2024. Detection of greenwashing in esg reports of chinese listed companies based on word2vec and tf-idf. In *Proceedings of the 2024 International Conference on Data Mining*, New York, NY, USA. Association for Computing Machinery.

Magali A Delmas and Vanessa Cuerel Burbano. 2011. The drivers of greenwashing. *California management review*, 54(1):64–87.

European Commission. 2023. Commission delegated regulation (eu) 2023/2772 adopting european sustainability reporting standards (esrs) - set 1. Official Journal of the European Union, L, 2023/2772 (Published 22 December 2023).

European Parliament. 2023. Green claims directive: Protecting consumers from greenwashing. Last accessed on April 11, 2025.

European Parliament and Council of the European Union. 2022. Directive (eu) 2022/2464 as regards corporate sustainability reporting (csrd). Official Journal of the European Union, L 322, p. 15–80.

European Securities and Markets Authority. 2023. Guidelines on greenwashing in sustainability reporting. Last accessed on April 11, 2025.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Valentina Lagasio. 2024. Esg-washing detection in corporate sustainability reports. *International Review of Financial Analysis*, 96:103742.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *Preprint*, arXiv:2401.12566.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2024. Loki: An open-source tool for fact verification. *Preprint*, arXiv:2410.01794.

Janna Lipenkova, Guang Lu, and Susie Rao. 2023. Detecting greenwashing signals through a comparison of esg reports and public media.

LSEG Data & Analytics. 2024. Esg scores | lseg. LSEG webpage. Last accessed on April 11, 2025.

J. Plucinska. 2023. Greenwashing cases against airlines in europe and the us. Reuters. Last accessed on April 11, 2025.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

D. Robinson. 2022. 10 companies called out for greenwashing. Published on Earth.org. Last accessed on April 13, 2025.

Ujjwal Sharma, Stevan Rudinac, Joris Demmers, Willemijn van Dolen, and Marcel Worring. 2024. Greenscreen: A multimodal dataset for detecting corporate greenwashing in the wild. In *International Conference on Multimedia Modeling*, pages 96–109. Springer.

P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley. 2022. *Chapter 10: Transport*. Cambridge University Press.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims. *arXiv preprint*.

Dimitar Trajanov, Georgi Lazarev, Lubomir Chitkushev, and Irena Vodenska. 2023. Comparing the performance of chatgpt and state-of-the-art climate nlp models on climate-related text classification tasks. In *Proceedings of the 4th International Conference on Environmental Design (ICED2023)*.

UK Competition and Markets Authority. 2021. Global sweep finds 40% of firms' green claims could be misleading. Last accessed on April 11, 2025.

Avalon Vinella, Margaret Capetz, Rebecca Pattichis, Christina Chance, Reshmi Ghosh, and Kai-Wei Chang. 2024. Leveraging language models to detect greenwashing. *Preprint*, arXiv:2311.01469.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. *Preprint*, arXiv:2110.12010.

Yue Zhao, Leon Kroher, Maximilian Engler, and Klemens Schnattinger. 2023. Detecting greenwashing in the environmental, social, and governance domains using natural language processing. In *KDIR*, pages 175–181.