Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
IN FINANCE

MASTER'S FINAL WORK
DISSERTATION

TALK WITHOUT WALK: A TRANSFORMER-BASED MEASUREMENT FRAMEWORK FOR CORPORATE GREENWASHING FROM VERBAL DISCLOSURES

ANWULI AUSTIN OLUH

JUNE - 2026

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
IN FINANCE

MASTER'S FINAL WORK
DISSERTATION

TALK WITHOUT WALK: A TRANSFORMER-BASED MEASUREMENT FRAMEWORK FOR CORPORATE GREENWASHING FROM VERBAL DISCLOSURES

ANWULI AUSTIN OLUH

SUPERVISION:
PROFESSOR JOÃO AFONSO BASTOS

JUNE - 2026

To Professor João Afonso
Bastos who sparked my
interest in Machine Learning
and Professor Sofia Santos
who encouraged me to write
a dissertation in the field of
Sustainable Finance

GLOSSARY

BaFin	Bundesanstalt für Finanzdienstleistungsaufsicht (German fin. regulator)
BCI	Business Conduct Issues
BERT	Bidirectional Encoder Representations from Transformers
CCQ	Claim and Commitment Quality
CCS	Claim and Commitment Specificity
EBA	European Banking Authority
EB-b	EnvironmentalBERT-Base
EB-e	EnvironmentalBERT-Environmental
ESG	Environmental, Social and Governance
EU	European Union
Fin	Financial
GCC(I)	Green Claim and Commitment (Intensity)
GHG	Greenhouse Gas
GI	Greenwashing Incident(s)
GICS	Global industry classification standard
GS	Greenwashing Score(s)
GT	Green Talk
GTI	Green Talk Intensity
GW	Green Walk
IFRS	International Financial Reporting Standards
IHPO	Improved hyperparameter optimization algorithm
LLM	Large Language Model
LoRA	Low Rank Adaptation
M&A	Merger & Acquisition
MFW	Master Final Work
ML	Machine Learning
MSCI	Morgan Stanley Capital International
NGO(s)	Non-governmental organization
NLP	Natural Language Processing
OLS	Ordinary least squares
OOF	Out-of-fold
S&P	Standard & Poor's
SBTi	Science Based Targets initiative
SEC	Securities and Exchange Commission
TCFD	Task Force on Climate-related Financial Disclosures
US EPA	United States Environmental Protection Agency
USA	United States of America

ABSTRACT, KEYWORDS AND JEL CODES

Greenwashing is the systematic gap between environmental claims and actual environmental performance and poses a growing threat to market integrity and investor decision-making. Despite increased scholarly and regulatory attention, reliable and scalable detection methods remain scarce. Existing natural language processing (NLP) approaches have focused predominantly on written corporate disclosures, leaving verbal corporate communication largely underexplored as a site of greenwashing detection. This Master Final Work addresses this gap by developing and applying an NLP-based pipeline to detect and score greenwashing in verbal corporate communication of STOXX Europe 600 companies over the period 2012-2022. A soft-voting ensemble of fine-tuned EnvironmentalBERT models classifies environmental sentences, while ClimateBERT-based classifiers identify green claims and commitments and assess their quality and specificity using dimensions grounded in TCFD and SBTi frameworks. Green Talk is operationalized as a composite score capturing the intensity, quality, and specificity of environmental claims and commitments in transcripts. Green Walk is proxied by RepRisk environmental risk incidents, weighted by severity, reach, and novelty. The Greenwashing Score is constructed as the industry-standardized gap between the two, enabling cross-sectional analysis.

Construct validity is assessed against externally identified RepRisk greenwashing incidents through bivariate correlation analysis, panel regression with company and year fixed effects and permutation placebo testing. The bivariate correlation is consistently positive and significant across all threshold specifications ($r \approx 0.26–0.33$, $p < 0.001$) and the panel regression corroborates this for the binary incident flag ($\beta = 0.060$, $p = 0.012$). Both results survive permutation placebo testing across 5,000 permutations with z-scores of 8.87 and 6.87, confirming genuine company-level signal and providing a scalable, independently validated instrument for monitoring the gap between environmental rhetoric and environmental reality.

KEYWORDS: Greenwashing Detection; Natural Language Processing; Green Talk; Green Walk; Corporate Environmental Disclosure; Earnings Call Transcripts

JEL CODES: G14; G30; M14; Q56; C55; C80.

RESUMO

O greenwashing é a discrepância sistemática entre as alegações ambientais e o desempenho ambiental real, representando uma ameaça crescente à integridade do mercado e à tomada de decisão dos investidores. Apesar da crescente atenção académica e regulatória, os métodos de deteção fiáveis e escaláveis continuam escassos. As abordagens existentes de processamento de linguagem natural (PLN) têm-se centrado sobretudo em divulgações corporativas escritas, deixando a comunicação corporativa verbal largamente pouco explorada como campo de deteção de greenwashing. Este Trabalho Final de Mestrado colmata esta lacuna através do desenvolvimento e aplicação de um pipeline baseado em PLN para detetar e pontuar o greenwashing na comunicação verbal de empresas do STOXX Europe 600, no período de 2012 a 2022. Um ensemble de modelos EnvironmentalBERT ajustados classifica as frases ambientais, enquanto classificadores baseados em ClimateBERT identificam alegações e compromissos ambientais, avaliando a sua qualidade e especificidade com base em dimensões fundamentadas nos quadros da TCFD e da SBTi. O Green Talk é operacionalizado como uma pontuação composta que capta a intensidade, a qualidade e a especificidade das alegações e compromissos ambientais nas transcrições. O Green Walk é medido por proxy através dos incidentes de risco ambiental da RepRisk, ponderados pela gravidade, alcance e novidade. O Greenwashing Score é construído como o desvio, padronizado por indústria, entre os dois, permitindo uma análise transversal.

A validade de construto é avaliada face a incidentes de greenwashing identificados externamente pela RepRisk, através de análise de correlação bivariada, regressão em painel com efeitos fixos de empresa e de ano, e testes placebo por permutação. A correlação bivariada é consistentemente positiva e significativa em todas as especificações de limiar ($r \approx 0{,}26$–$0{,}33$, $p < 0{,}001$), e a regressão em painel corrobora este resultado para o indicador binário de incidente ($\beta = 0{,}060$, $p = 0{,}012$). Ambos os resultados resistem a testes placebo por permutação em 5.000 permutações, com z-scores de 8,87 e 6,87, confirmando um sinal genuíno ao nível da empresa e fornecendo um instrumento escalável e validado de forma independente para monitorizar o desvio entre a retórica e a realidade ambiental.

TABLE OF CONTENTS

TALK WITHOUT WALK

By Anwuli Austin Oluh

Greenwashing costs lives, destroys shareholder value, and misleads the investors and regulators who depend on environmental disclosures to make consequential decisions. The Volkswagen scandal took seven years to surface. This thesis asks whether an NLP-based pipeline applied to what companies say in their verbal communication, rather than what they write in polished sustainability reports, can detect the gap between environmental rhetoric and environmental reality systematically and at scale. Validated against eleven years of incident data, the Greenwashing Score developed here offers a scalable and empirically grounded instrument for doing exactly that.

1. INTRODUCTION

An equity market loss of €27.4 billion in five days and approximately 1,200 premature deaths across the EU (Barth et al., 2022; Chossière et al., 2017; Jong & van der Linde, 2022). These were among the direct consequences of the Volkswagen emissions scandal. After installing illegal defeat devices[1] in diesel engines from 2008 onwards, Volkswagen actively marketed these vehicles under a "clean diesel" campaign (Shah et al., 2017) until the US EPA (2015) issued a notice of violation of the clean air act to Volkswagen AG in September 2015. The Volkswagen case is not an isolated incident but rather a paradigmatic example of a broader phenomenon: the systematic gap between corporate environmental claims and actual environmental performance - Greenwashing. The scale of harm that can accumulate when this gap goes undetected - seven years in the case of Volkswagen - has motivated scholars across disciplines to develop more reliable and timely methods for detecting greenwashing. Whether from a regulatory, investor, or journalistic standpoint, the demand for early detection is the same: to prevent the kind of prolonged, verifiable deception that the Volkswagen case exemplifies (Lyon & Montgomery, 2015).

The Volkswagen case represents a relatively tractable form of greenwashing as the deception was ultimately verifiable through physical measurement of tailpipe emissions. The more frequent and harder-to-detect form of greenwashing lies in the carefully chosen words of corporate disclosures, where companies can imply environmental commitment through vague targets, unverifiable claims and selective disclosure without ever crossing a legal line (Lyon & Maxwell, 2011). A prime example is the case of DWS, the asset

[1] A defeat device is any vehicle hardware, software, or design that disables or reduces emissions controls during normal driving, even when the vehicle passes official emissions tests (Barth et al., 2022).

management subsidiary of Deutsche Bank, whose former sustainability chief Desiree Fixler went public in 2021, accusing the company of misrepresenting the ESG credentials of its investment products (Kowsmann & Brown, 2021). DWS made a misleading statement in its 2020 annual report, stating that more than 50% of the group's assets were invested using ESG criteria (Sedilekova & Merchat, 2022). While DWS rejected the allegations as unfounded (Hetzner, 2022), the increased regulatory scrutiny by the United States Securities and Exchange Commission (SEC) and Germany's financial service regulator (BaFin) led them to decrease their reported ESG-invested assets by approximately 75% in their annual report of 2022 (DWS, 2023).

The DWS case illustrates that greenwashing is difficult to detect and typically requires either a whistleblower or sustained regulatory pressure to become visible. Even among investors with the resources to scrutinise disclosures, confidence remains low. EY's 2024 Institutional Investor Survey found that 85% of investors viewed greenwashing as a worsening problem and 80% considered the comparability of sustainability reporting to be in acute or substantial need of improvement (Handford & Lofts, 2024). This suggests that existing detection mechanisms are insufficient, signalling more scalable and reliable methods are demanded by market participants.

Following the 2015 Paris agreement, the EU has pursued the most ambitious corporate sustainability disclosure agenda globally (Hummel & Jobst, 2024), introducing the EU Taxonomy (2020), the Sustainable Finance Disclosure Regulation (2019), and the Corporate Sustainability Reporting Directive (2022). While these frameworks represent a significant regulatory step toward closing the gap between environmental claims and environmental reality, their effectiveness depends on the reliability of the disclosures they govern. Critically, existing EU disclosure requirements target primarily written corporate communication. These standardized reports are subject to formatting requirements, external review and increasing regulatory oversight. Verbal corporate communication, by contrast, falls largely outside these frameworks (Lerman et al., 2022), creating a disclosure channel through which firms may, whether intentionally or unintentionally, misrepresent their environmental performance without the constraints that govern written reporting. The importance of closing this gap is underscored by evidence that verbal corporate communication is a primary channel through which institutional investors

monitor companies and inform their trading decisions (Amoozegar et al., 2020), making it a consequential but underregulated site of potential environmental misrepresentation.

Traditional greenwashing detection approaches, like manual content analysis, are resource intensive, difficult to scale and inherently subjective (Haliburton et al., 2025). Recent advances in natural language processing (NLP) offer a promising alternative, enabling large-scale and systematic analysis of textual environmental disclosures (Bingler et al., 2021, 2024). However, existing NLP approaches focus almost exclusively on written corporate disclosures like sustainability reports, annual reports and corporate websites (Lublóy et al., 2025). This leaves verbal corporate communication largely underexplored as a site of greenwashing detection. Consequently, this Master Final Work (MFW) applies an NLP-based pipeline to detect and score greenwashing of environmental verbal corporate communication, drawing from 9,940 transcripts of European companies over the period of 11 years (2012-2022) answering the following research question:

"Can an NLP-based pipeline applied to verbal corporate communication measure greenwashing among STOXX Europe 600 companies (2012–2022) as the gap between multidimensional green talk and incident-based green walk, and does this measure hold up against externally documented greenwashing incidents?"

To answer this question, this thesis quantifies the Green Talk (GT) for each company-year observation in the sample. An NLP-based pipeline first classifies individual sentences from corporate transcripts as environmental or non-environmental using a fine-tuned ensemble classifier. Claims and Commitments are drawn from the environmental sentences and then scored across three dimensions: the quantity, quality, and specificity of environmental claims and commitments. These scores are aggregated into a composite green talk measure, which is then contrasted against a green walk proxy based on environmental risk incidents from RepRisk, an independent media-monitoring database. The resulting Greenwashing Score (GS) captures the company-level gap between environmental rhetoric and actual environmental performance, following the conceptual framework of greenwashing as cheap talk established by Bingler et al. (2021, 2024). This score is then validated via Pearson correlation coefficient and a panel regression against externally documented greenwashing incidents drawn from the RepRisk database.

This thesis makes four contributions to the existing literature. First, it applies these methods to verbal rather than written corporate communication, a channel that has received comparatively little academic attention despite being largely unaudited and unregulated (Calamai et al., 2026; Lublóy et al., 2025). Second, it introduces a multi-dimensional scoring framework that captures the quantity, quality, and specificity of environmental claims, offering a more granular measure than binary classification approaches. Third, it uses RepRisk environmental incident data as a validation benchmark, an approach noted as promising because it excludes corporate self-disclosure and draws instead on independently verified incidents (Lublóy et al., 2025). Fourth, it extends NLP-based greenwashing detection to the European equity market, addressing a gap in a literature dominated by US- and China-focused studies (Lublóy et al., 2025).

The remainder of this thesis is structured as follows: Section 2 reviews the relevant literature on greenwashing and NLP-based detection methods. Section 3 describes the data and pipeline methodology. Section 4 presents the results. Section 5 discusses the limitations and what future research should focus on. Section 6 concludes.

2. LITERATURE REVIEW

This section elaborates on the current literature on greenwashing and its measurement. It first sets out the theoretical foundations of greenwashing and reviews the key studies, methods, and results that have shaped the field, before turning to the research gaps that motivate this work. These gaps are organised around the three building blocks of the decoupling-based measure (Green Talk, Green Walk and the Greenwashing Score itself).

2.1. Theoretical Foundations

In their research on greenwashing Bernini and La Rosa divide the definitions of greenwashing into three perspectives (2024). The first perspective is the "omission in reporting on reality" or selective disclosure. One of the most recognized papers on selective disclosure in recent years, defines it as "a symbolic strategy whereby firms seek to gain or maintain legitimacy by disproportionately revealing beneficial […] performance indicators to obscure their less impressive overall performance" (Marquis et al., 2016). To separate from that is misleading disclosure, which forms the second perspective (Bernini & La Rosa, 2024). In contrast to selective disclosure, misleading disclosure has a clearer intention to deceive and is therefore often higher penalized as it is easier to prove. The Volkswagen emissions scandal is a prime example of misleading disclosure as Volkswagen marketed the engines of their diesel cars as "clean" while they rigged the computers of their cars with an illegal defeat device to display lower emission results when they were tested (Jong & van der Linde, 2022). The third perspective is referred to as the "gap between what companies report and what they do" (Green Talk vs. Green Walk). The majority of academic greenwashing measurement approaches focuses on measuring this gap also referred to as decoupling, because of its simplicity and data availability (Lublóy et al., 2025). Accordingly, this study adopts the same approach.

Green Talk (GT) refers to a firm's (symbolic) communication of environmental responsibility. It includes statements, claims, commitments or future-oriented goals presented in corporate communication (Walker & Wan, 2012). When computing greenwashing measurements with the decoupling approach, it is necessary to measure the Quantity (and Quality) of Green Talk.

Green Walk (GW) refers to a firm's substantive environmental actions, meaning the measures it implements to improve its environmental performance. These include

tangible initiatives such as reducing emissions, improving energy efficiency, or investing in environmental technologies, which produce measurable outcomes (Walker & Wan, 2012). When computing greenwashing measurements with the decoupling approach, there are several ways of addressing the Green Walk. Among the most popular ones are ESG scores and ratings, selected ESG data points or content analysis (Lublóy et al., 2025).

The greenwashing measure of decoupling is derived from the gap between a company's Green Talk and Green Walk, resulting in a greenwashing score. The most cited paper by Yu et al. (2020) uses measures representing a firm's relative position to its peers in the distributions[2]. In contrast, other studies do not construct peer groups and instead compare firms solely based on their relative positions within the overall distributions (He et al., 2024).

2.2. Key Studies, Methods and Results

Current practices for measuring greenwashing in corporate communications are complex due to the task's inherent challenges. Measures of decoupling are more appropriate for detecting greenwashing in contrast to measures of selective disclosure because measures of selective disclosure consider only communication and do not take environmental performance into account (Lublóy et al., 2025).

The most widely used greenwashing measure is Yu et al.'s (2020) decoupling measure which is comparing a company's peer-relative position in disclosure (Bloomberg ESG Disclosure score) to a modified ASSET4 ESG score. Its popularity is driven by the high accessibility of the required data for researchers (Lublóy et al., 2025). A positive value in the greenwashing score indicates that a firm discloses more ESG information than justified by its performance, suggesting greenwashing behavior. Using this measure within a panel regression framework, the authors examine company- and country-level determinants of greenwashing. The results indicate a significant decrease in greenwashing behavior when external scrutiny is increased, particularly through higher shares of independent directors and institutional investors. A key limitation of the decoupling measure proposed by Yu et al., as well as related approaches in the literature, is their reliance on external ESG scores, ratings, or disclosure metrics (Yu et al., 2020;

[2] Distributions means in this context the several different metrics used like ESG Disclosure scores, ESG scores, Green Talk Intensity, Environmental Incidents etc.

Zeng et al., 2025), which have been criticized for inconsistencies across providers and their strong dependence on corporate self-reporting (Berg et al., 2019).

In recent years, advancements in Machine Learning (ML) have expanded the possibilities for detecting greenwashing (Calamai et al., 2026). The authors Zeng et al. (2025) developed a predictive framework for the detection of greenwashing by combining XGBoost with an improved hyperparameter optimization algorithm (IHPO), using data from Chinese listed firms and a constructed Greenwashing Index that captures the gap between disclosure and actual performance. The IHPO enhances optimization through advanced search techniques, leading to a model that significantly outperforms standard and alternative methods in predictive accuracy and robustness.

Advances also happened in the field of NLP, a subfield of ML which focuses specifically on text and language data. “Several tasks, such as climate topic identification, climate risk classification, and the characterization of environmental claim types, now achieve near perfect performance when tackled with fine-tuned models.” (Calamai et al., 2026). Chinese researchers developed DeepGreen, a dual-stage framework for detecting greenwashing in Chinese corporate annual reports. They employ Large Language Models (LLMs) to extract potential "green" keywords and perform semantic analysis to determine if those keywords represent substantive actions or mere symbolic talk (Xu et al., 2025).

The most closely related paper and therefore key study for this MFW is the work of researchers from Australia and New Zealand (He et al., 2024). In their study they measure greenwashing intensity for public firms in the U.S. (2007-2021) by identifying the gap between what companies say (Green Talk) and what they do (Green Walk). Green Talk is measured from earnings conference call transcripts. The NLP model FinBERT (Araci, 2019), a domain-specific variant of BERT, pre-trained on financial texts (10-Ks, transcripts) is used. FinBERT was specifically fine-tuned with a manually labelled dataset of 3,500 sentences to recognize positive environmental statements with 90% accuracy, while excluding general climate discussions or explanations of past accidents. From the classification results, the Green Talk Intensity (GTI) was calculated, defined as the proportion of green sentences relative to the total number of sentences. To assess the Green Walk, the study utilized RepRisk incidents. RepRisk is a Swiss data provider specialized in ESG Data. They provide insights on reputational risk and responsible

business conduct by combining artificial intelligence with human analysis to support due diligence, risk management and compliance processes. Methodologically, RepRisk follows an outside-in approach, systematically screening large volumes of public information such as media, NGOs and regulatory sources, while explicitly excluding corporate self-disclosures to reduce bias (RepRisk, 2026). Its data is widely used by financial institutions and corporates to assess ESG-related risks (RepRisk, 2020). The study by He et al. (2024) retrieves environmental incidents from RepRisk and aggregates them at a firm-year level. In a subsequent step, the authors constructed percentile rankings for Green Talk based on the Green Talk Intensity (GTI)[3] and for Green Walk based on the frequency of environmental incidents. With the following equation they constructed an industry-level greenwashing score (GS):

$$(1) \qquad GS_{i,t} = \frac{Rank_{i,t}^{GTI} - (-1) * Rank_{i,t}^{EnvIncidents}}{100}.$$

According to their research greenwashing intensity increased significantly following the 2015 Paris agreement, with the sharpest rise concentrated in fossil fuel and stranded asset industries. The authors trace this to an agency problem: greenwashing inflates the environmental ratings companies receive from rating agencies, which in turn strengthens managerial job security and ESG-linked compensation. Managers are therefore incentivised to portray environmental performance as favourably as possible, capturing private benefits even when doing so destroys value for shareholders. This conflict matters because the gains accrue to managers while the costs fall on the company, which is a divergence that helps explain why greenwashing persists despite evidence that it does not pay at the company level. Consistent with this, the predominant finding in the literature is that greenwashing has no or negative effects on financial performance, as measured through stock returns, firm value, and accounting-based performance metrics such as return on assets and net income (Gatti et al., 2021; Ghitti et al., 2024; Wu & Shen, 2013).

In contrast to the predominant findings in the literature, a recent study conducted in China identifies a positive effect of greenwashing on corporate financial performance in developing country contexts (Li et al., 2023). Rather than contradicting the above, this points to a common moderator: detection. Part of the reason for this positive relationship

[3] The Green Talk Intensity (GTI) is measured as the number of sentences classified as Green Talk divided by the total number of sentences.

is the information asymmetry[4] between stakeholders and greenwashing companies. Information asymmetry is high where stakeholders cannot evaluate the substance behind environmental claims, making greenwashing less likely to be seen through, allowing companies to capture reputational and market benefits without incurring the penalties that follow exposure. Where disclosure is more heavily scrutinised, the same gap is more readily detected and priced, eroding any benefit. This detection mechanism is the same one underlying the agency problem above: in both cases, whether greenwashing pays depends on whether outside parties can distinguish talk from walk. This underscores the importance of developing reliable, granular and independently verifiable greenwashing measures that reduce information asymmetry between companies and their stakeholders, which is the motivation for the Greenwashing Score (GS) developed in this thesis.

2.3. Research Gaps

Several research gaps persist in the measurement of greenwashing, particularly in NLP-supported approaches, as this field is still relatively new and continuously evolving. Decoupling measures integrate quantitative indicators of both Green Talk and Green Walk to assess the discrepancy between them. Accordingly, this section is organized into three parts: Green Talk, Green Walk, and the Greenwashing Score.

2.3.1. Green Talk

(1) The vast majority of research focuses on non-verbal corporate communication such as annual reports, sustainability reports, and corporate websites (Lublóy et al., 2025). Written corporate disclosures typically follow formatting requirements and regulatory templates that constrain narrative expression. Whereas verbal corporate communication provides a dynamic and flexible platform where the less rigid and interactive nature allows executives to present environmental messages in a more conversational and unstructured manner, potentially amplifying greenwashing (Blau et al., 2015). This is particularly consequential given that corporate communication like earnings conference calls have emerged as a primary channel through which institutional investors monitor companies and inform their trading decisions, meaning that greenwashing in verbal corporate communication has direct implications for capital allocation (Amoozegar et al.,

[4] The information asymmetry refers to the inability of stakeholders to evaluate the substance of corporate communication due to the lack of insights into a firms' activities (Li et al., 2023).

2020). To the best of the author's knowledge only one study has analyzed verbal corporate communication transcripts, namely earnings conference call transcripts to measure greenwashing (He et al., 2024). Several factors explain the limited use of verbal corporate communication. Notably, recent advancements in NLP techniques were necessary to enable large-scale analysis of verbal corporate communication (Medya et al., 2022). While expert-annotated datasets are available for non-verbal corporate communication, no sufficiently large dataset exists for verbal corporate communication yet. Consequently, researchers rely on alternative approaches, such as self-supervised learning, to avoid high costs associated with manual annotation (Huang et al., 2025). Although training a new model with an expert-annotated dataset is beyond the scope of this MFW, fine-tuning an existing model trained on textual corporate disclosures is performed to achieve an improved model accuracy on the classification (environmental and non-environmental) of verbal corporate communication.

The researchers He et al. (2024) fine-tuned FinBERT, a pre-trained NLP model with a generalized understanding of the financial context, as it achieves higher accuracy in predicting ESG-related sentences compared to the base BERT model. However, Schimanski et al. (2024) demonstrated in their research that the FinBERT model performs significantly worse in classifying environmental sentences than the fine-tuned version of the EnvironmentalBERT-base (EB-b) model, EnvironmentalBERT-environmental (EB-e). This MFW advances previous research through the novel fine-tuning of EB-e, aiming to improve the classification of verbal corporate communication, bridging the domain gap. In doing so, it provides one of the first systematic approaches to quantify Green Talk (and subsequently Greenwashing) in verbal corporate communication.

(2) The multidimensional and dynamic nature of greenwashing hinders reliable, fully automated approaches, leading many methods to incorporate human input, such as expert-annotated datasets or expert interviews to quantify the Green Talk of a company (Capetz et al., 2024; Schimanski et al., 2024; Zeng et al., 2025). However, manual labelling is highly time-intensive and inherently subjective, with associated biases that robust frameworks can mitigate, but not fully eliminate (Haliburton et al., 2025). This MFW avoids extensive and complex manual labelling. Instead, it limits the manual labelling to a relatively simple classification task, namely determining whether a sentence contains environmental content. This manual labelling is performed to obtain a stratified sample

which is later used to evaluate model accuracy and fine-tune the model. By minimizing the manual labelling this MFW minimizes potential bias of the Green Talk compared to other research in the field (Capetz et al., 2024; Zeng et al., 2025) thereby enhancing the robustness and interpretability of the results.

(3) Calamai et al. (2026) found that most of the research quantifies the amount of Green Talk rather than assessing the amount of information in the Green Talk. They suggest that future research should address this and focus on the quality of Green Talk in addition to the quantity. This MFW assesses the quality of Green Talk and incorporates the results into the computation of the Greenwashing score. Furthermore, it does not rely on the classic GTI as often used (He et al., 2024). It uses a Green Claim and Commitment Intensity (GCCI) based on the work of Ruiz-Blanco et al. (2022) who defines greenwashing as "the difference between what the company says it does in terms of commitment to sustainability, and what the company actually does as evaluated by external parties". The choice directly follows Bingler et al. (2024), which established that an accurate assessment of whether a company is engaged with climate action requires distinguishing between sentences that commit to company-level climate activities and sentences that describe climate change in general terms. Restricting the measure to claims and commitments therefore removes the noise introduced by general environmental discourse, which may be more prevalent in certain industries not because of superior climate engagement but because of the structural nature of their business vocabulary.

2.3.2. Green Walk

Regarding Green Walk, Kathan et al. (2025) find for companies of the STOXX Europe 600 there is a positive correlation between ESG scores and greenwashing cases as those scores focus on apparent environmental performance rather than real environmental performance. Furthermore, the divergence of ESG ratings makes the choice of a specific rating provider highly consequential for the resulting Green Walk measure (Berg et al., 2019). Consequently, despite their widespread use and ease of application, ESG scores are inadequate for the quantification of Green Walk.

The study by He et al. (2024) measures Green Walk with RepRisk's environmental risk incidents. According to the systematic literature review by Lublóy et al. (2025) on quantifying firm-level greenwashing, the authors come to the conclusion that

"greenwashing measures based on actual incidents are promising, marking the emergence of a new branch of research.". Hence, this MFW employs RepRisk's environmental risk incidents as a metric to quantify Green Walk, an approach that is not yet widely adopted.

2.3.3. Greenwashing Measure

The greenwashing score proposed by He et al. (2024) compares percentile rankings of all the U.S. public listed companies in their sample. Lublóy et al. (2025) argue that, due to industry-specific characteristics and regulatory requirements, companies should be compared only to their respective peer groups rather than to the entire sample. Furthermore, "Communication and performance […] shall be standardized for each industry separately by the industry mean and standard deviation.". The objections are valid and have been addressed in prior research (Di & Li, 2023; Li & Zheng, 2024). Following this approach, the present study develops industry-specific measures of Green Talk and Green Walk, which are subsequently standardized within each industry.

3. METHODOLOGY

This section sets out how the Greenwashing Score (GS) is constructed and validated. It first describes the Green Talk side, namely transcript selection and sample construction, the fine-tuned NLP classification pipeline and the measurement of claim and commitment intensity, quality, and specificity. Subsequently, the methodology behind the Green Walk proxy from RepRisk environmental incidents is explained. Finally, the combination of the two components into the industry-standardized GS and its validation is described.

3.1. Green Talk: Transcript Selection and Sample Construction

The focus of this MFW is the broad market of European equities. Therefore, the STOXX Europe 600 index was selected, as it represents approximately 90% of the underlying investable market in Europe (STOXX Ltd., 2026). The sample period spans from 2012 to 2022, enabling the analysis of greenwashing behavior before, during and after the Paris agreement, as well as the emergence of advanced EU sustainability regulations. The sector classification utilized for classifying each company into distinct industry sectors (in the following only referred to as "industry") is the Global industry classification standard (GICS) by MSCI and S&P Dow Jones Indices (MSCI Inc., 2024). Out of the 11 industries, the financials industry is excluded because its environmental incidents are often structurally distinct from those of non-financial firms, exhibiting more indirect causality and different disclosure incentives (Cambridge Centre for Sustainable Finance, 2016). Furthermore, the Real Estate industry is excluded due to its relatively small representation within the index (STOXX Ltd., 2026).

TABLE I

KEYWORDS IDENTIFYING RELEVANT CORPORATE COMMUNICATION

Transition	Climate Change	Climate	Deforestation
Sustainability	Environment	ESG	Emission
Green	Clean Energy	Energy	CO2
Carbon Offset	Carbon Reduction	Greenhouse Gas	Renewable
GHG	Biodiversity	Conservation	Carbon
Ecosystem	Recycling	Circular Economy	Warming
Sustainable	Pollution	Water Quality	Global Warming
Environmental Impact	Wind	Taxonomy	Solar

Source: Bloomberg

The transcripts are collected using Bloomberg's transcript archives, which include Sales Results Calls, M&A Calls, Conference Presentation Calls, Earnings Calls and Shareholder Management Calls. The dataset is pre-filtered using relevant keywords to identify corporate communication that contains environmental content, as visible in Table I. Bloomberg's Keyword Search Logic counts the occurrence of specified keywords in each transcript and reports their frequency. Transcripts are downloaded manually, excluding those containing fewer than ten keyword occurrences to balance the computational effort required for analysis with the expected output, namely the extraction of relevant sentences.

From the transcripts each sentence is extracted and matched with its relevant information like transcript Style, Industry, date, company name and speaker. Only sentences by verified company participants are relevant, as only these should be counted when assessing the amount and quality of a company's green talk.

Prior to deployment, the classification accuracy of the EnvironmentalBERT-Environmental (EB-e) model is evaluated on a stratified transcript sample to assess the extent of the written-to-spoken domain gap. Natural language processing (NLP) models trained on written language are generally expected to yield reduced performance when applied to spoken language, as they differ systematically in syntax, vocabulary, and sentence structure (Wahde et al., 2024). For this purpose, a stratified sample of 494 sentences[5] was constructed, with observations distributed relatively evenly across the years 2012, 2014, 2016, 2018, and 2020. The sample is balanced between environmental and non-environmental observations (50/50) to ensure equal representation of both classes. Within each year, the industry distribution of sampled sentences is aligned with the corresponding distribution in the full dataset, thereby preserving the underlying industry composition. This approach prevents the model from being disproportionately fine-tuned toward specific industries or time periods, while also mitigating the risk of overfitting to year-specific language patterns. Due to the simplicity of the annotation task

[5] The stratified sample contains 494 rather than 500 sentences because the number of sentences per industry was rounded within each of the five years used for industry aggregation, so some years yielded slightly fewer than 100 sentences.

(Environmental or not environmental), the limited scope of this study and the unavailability of expert annotators, the annotation was carried out by the author.

TABLE II

AGGREGATE PERFORMANCE SUMMARY

The table below reports mean accuracy and F1 score across five cross-validation folds, together with the pooled out-of-fold accuracy and its 95% Wilson confidence interval. The ensemble row (highlighted in bold) is the final recommended classifier.

Model	Acc. mean ± std	F1 mean ± std	Prec. mean ± std	Rec mean ± std	Pool Acc.	95% CI
EB-e (zero-shot, no fine-tuning)	-	-	-	-	0.8765	[0.845, 0.903]
M1: EB-b (fine-tuned)	0.9252 ± 0.027	0.9279 ± 0.026	0.9249 ± 0.027	0.9250 ± 0.027	0.9251	[0.899, 0.945]
M2: EB-e (fine-tuned)	0.9373 ± 0.008	0.9405 ± 0.006	0.9371 ± 0.008	0.9371 ± 0.008	0.9372	[0.912, 0.955]
M3: EB-e + LoRA	0.9333 ± 0.040	0.9406 ± 0.028	0.9335 ± 0.039	0.9327 ± 0.041	0.9332	[0.908, 0.952]
Ensemble M2 + M3 (w = 0.40 / 0.60, thr. = 0.58)	**-**	**0.9615 (pooled)**	**0.9616**	**0.9615**	**0.9615**	**[0.941, 0.975]**

Applied zero-shot, without further training on verbal transcripts, EB-e achieves a pooled accuracy of 0.8765 (95% CI: [0.845, 0.903]) implying that approximately every 8th sentence gets misclassified by the EB-e model applied on verbal corporate communication as reported in Table II above. Fine-tuning both EB-b and EB-e on the stratified sample increases accuracy to 0.9251 and 0.9372 respectively, reducing the misclassification rate to approximately every 13th and 16th sentence respectively, as reported in Table II above. While this represents a meaningful improvement over the zero-shot baseline, approximately one in sixteen sentences remains incorrectly classified under the best individual model, which motivates further improvement.

A low rank adaptation (LoRA) (Hu et al., 2021) is applied to the EB-e (M2) model to preserve the domain knowledge and reduce the risk of overfitting on the limited corpus of the stratified sample. Rather than updating every model parameter, LoRA inserts small trainable adapter matrices into the attention layers while keeping the pre-trained backbone frozen. Every of the three models is fine-tuned on the stratified sample and evaluated under five-fold stratified cross-validation, yielding predictions on all the observations without any sentence appearing both in the training and testing set. This produces a stable performance estimate despite the small sample size (Hastie et al., 2009). The result

indicates that the new model (M3: EB-e + LoRA) is within the tight cluster of the prior fine-tuned models, as visible in Table II above. The range (between 92.5% and 93.7% pooled accuracy) is narrow, implying statistical indistinguishability. The per-fold results in Appendix Table V draw an interesting picture. The LoRA model shows the highest variance (std. = 0.040) in F1 accuracy across folds, dropping to 88.9% in fold 1 and reaching its best result in fold 2 with 97.0%. Examining the per-class performance in Appendix Table VI reveals the asymmetry in the LoRA model. It is the most cautious non-environmental predictor (Precision = 0.978) but misses 11.3% of actual non-environmental sentences (Recall = 0.887), flagging them as environmental instead. Conversely, the LoRA model almost never misses an environmental sentence (recall = 0.980). This asymmetry motivates a soft voting ensemble (Hastie et al., 2009) in which the predicted class probabilities of M2 and M3 are blended as a weighted average. The blending weight and classification threshold are jointly optimized over the pooled out-of-fold predictions to maximize the macro F1 score:

(2) $\hat{p}_{ens} = w * \hat{p}_{EB-e} + (1 - w) * \hat{p}_{LoRA}$.

Appendix Table VII shows that the ensemble achieves its highest performance at $w = 0.4$ (threshold of 0.58) with a pooled accuracy and F1 score of 0.9615. The ensemble cuts total misclassifications from 31 sentences in M2 to just 19 sentences (39% reduction in errors) as reported in Appendix Table VIII. Notably, this is achieved without sacrificing either class. It makes fewer false positives than any model and similar false negatives as the best model (8 vs. 5). This confirms that the two constituent models are genuinely complementary as reported in the Appendix Table VI.

To verify that the performance gain reflects a true difference rather than a numerical fluctuation, a McNemar's test is conducted on the pooled out-of-fold predictions. The results indicate that the ensemble produces a statistically significant improvement over EB-e ($p = 0.010$) and EB-e + LoRA ($p = 0.011$) as reported in the Appendix Table IX. Table II above shows that it thereby surpasses each individual model and represents the only performance gain statistically confirmed by McNemar's test. However, as this value represents a point estimate, it may not fully reflect generalization performance. The true performance on unseen data is more appropriately captured by the 95% confidence interval (0.941, 0.975), which therefore provides a more reliable scope for evaluating the

model. The ensemble model is applied to classify all sentences in the full transcript corpus as environmental or non-environmental.

3.2. Green Talk: Measurement of Intensity and Annotation of Sentences

Green Talk Intensity (GTI) is defined as the share of environmental sentences in a company's total transcript sentences in a given year (He et al., 2024).[6] Within this universe of environmental sentences, a second classification step identifies sentences that constitute environmental (green) claims or commitments. From this, the Green Claim and Commitment Intensity (GCCI) is defined as the share of claim and commitment sentences within the environmental sentence universe. Looking forward, only sentences that are labelled as Green Claims and/or Commitments are considered for Green Talk Computation.

To identify environmental claims within the sample of environmental sentences, the environmental-claims model proposed by Stammbach et al. (2023) is employed. This model is a fine-tuned version of the distilroberta-base-climate-f model, a ClimateBERT-based model (Webersinke et al., 2022). It demonstrates high accuracy in detecting environmental claims at the sentence level. In addition to the claims identified by the environmental-claims model, commitments are identified using the distilroberta-base-climate-commitment, a further ClimateBERT-based model, fine-tuned on the climate commitments actions dataset by Bingler et al. (2024). Since this model was trained on a paragraph level, inference is performed using a three-sentence centred window constructed around each target sentence, bounded by speaker turn boundaries to ensure that context never crosses into adjacent speaker turns while preserving consistency with the model's training data. Fine-tuning the models on verbal corporate communication (as performed on the environmental classification task) is outside of the scope of the study, so spoken language characteristics may affect classification accuracy. The result is a final sample of 109,319 environmental (green) Claims and/or Commitments (GCC) for our Green Talk Computation.

[6] This measure is overstated, as the transcripts were pre-filtered to include only those containing environmental content. Consequently, the reported value for GTI does not reflect the full sample and is expected to be significantly lower in an unfiltered dataset. Same applies to the GCCI

The quality of a claim/commitment is determined using a rule-based scoring approach of theoretically grounded dimensions that capture the substantive content of corporate Green Talk. The idea to include the claim/commitment quality and specificity into the Green Talk score is based on the approach followed by Capetz et al. (2024). The researchers aim "to capture the essence of greenwashing: overly positive language without specific commitments or details, […].". In this MFW a modified version is used which incorporates the quality and specificity of a claim/commitment in addition to its quantity. On claim quality a sentence receives credit for containing climate-domain vocabulary (+0.30), numeric figures (+0.40 in climate context, +0.20 otherwise), strong climate terms (+0.10), and a tiered ambition bonus for the magnitude of any percentage reduction present (+0.25 / +0.15 / +0.08 for high, medium, and low ambition respectively). Science Based Targets initiative (SBTi) validation adds a flat +0.20, reflecting that external verification against a science-based methodology is the strongest credibility signal available in corporate climate disclosure (SBTi, 2023). Commitment quality follows the same logic but adds an extra time-horizon component, consistent with the Task Force on Climate-Related Financial Disclosures (TCFD)[7] (2017). For commitments, SBTi validation is weighted more heavily and in a tiered fashion: +0.30 when the sentence also contains a numeric target, a time horizon, and climate vocabulary which is the combination that constitutes the gold standard of a credible corporate climate commitment in the cheap talk literature (Bingler et al., 2021, 2024). Since the specific weights assigned to each quality dimension are theoretically motivated but not empirically calibrated, their sensitivity is tested through two alternative specifications in the Section: 3.5 Validation and Robustness of the Greenwashing Score.

Specificity is assessed using the distilroberta-base-climate-specificity model (Bingler et al., 2024) a fine-tuned ClimateBERT classifier model, trained to distinguish specific from vague climate-related sentences. The model produces a continuous specificity score for each GCC, which is treated as a standalone output rather than merged into the quality scores. This separation reflects a conceptual distinction between the substantiveness of a

[7] The TCFD fulfilled its remit and disbanded in October 2023 following the publication of its final status report. The IFRS Foundation (via the International Sustainability Standards Board) has since assumed responsibility for monitoring climate-related disclosure progress, and the TCFD recommendations have been substantially incorporated into IFRS S2 (2025). The quality dimensions operationalised here, reflect criteria that persist across both frameworks.

disclosure (Claim and Commitment Quality) and its concreteness (Claim and Commitment Specificity), which concerns whether the language is precise and actionable rather than aspirational and diffuse. The distinction matters because a sentence can score highly on quality dimensions (containing a numeric target, a time horizon, and climate vocabulary) while still being expressed in vague or hedged language, and conversely a highly specific sentence may lack the commitment framing that quality scoring rewards. Retaining both dimensions separately allows downstream steps to distinguish between disclosures that are substantive but vague and those that are both substantive and concrete.

3.3. Green Walk: RepRisk Environmental Incidents

The Swiss ESG data-science company RepRisk identifies 28 Business Conduct Issues (BCI), and each risk incident is systematically assigned to one of the BCI. These are further classified into environmental, social, governance, and cross-cutting issues (RepRisk, 2025). For the purposes of this MFW, only environmental issues are relevant, as the focus is on greenwashing. Environmental issues are further subdivided into categories such as impacts on ecosystems and biodiversity or local pollution. However, no additional granularity is required for this analysis. Therefore, only environmental risk incidents in general are considered.

RepRisk's methodology provides a robust academic foundation for detecting greenwashing because it employs an outside-in approach that intentionally excludes company self-disclosures, which are often considered unreliable for identifying actual risks (Chopra et al., 2024). This assessment is performed through a hybrid model combining machine learning and human intelligence. Automated systems screen sources daily across 23 languages to identify and pre-tag relevant entities and issues using proprietary algorithms. Following that, a specialized analyst team reviews these outputs to approve tagging and scoring according to a rules-based system. These analysts curate the data into risk incident briefs, and a senior analyst performs a final quality assurance check to ensure adherence to methodological standards before the information is published (RepRisk, 2021).

Each identified incident is analyzed based on three critical parameters: (1) Severity, which assesses the harshness of the incident based on its consequences (e.g., health and safety impacts), the extent of its impact, and the level of intent as visible in the Appendix

Table X. (2) Reach rates the influence and readership of the source from limited to global as visible in the Appendix Table XI. (3) Novelty determines the "newness" of an incident, meaning if the incident marks the first time a company is exposed to a specific issue in a certain location as visible in the APPENDIX TABLE XII.

In total, 16,845 environmental incidents were recorded between January 2012 and December 2022 for STOXX Europe 600 companies (ex. Financials & Real Estate). From this universe 1,126 Greenwashing incidents were excluded to mitigate circularity.

3.4. Construction of the Greenwashing Score

The construction of the Green Talk score extends Bingler et al.'s (2024) cheap talk index by a quality dimension derived from the TCFD (2017) and SBTi (2023) criteria. Green Talk is quantified along three equally weighted pillars: the Green Claim and Commitment Intensity (GCCI), the Claim and Commitment Quality (CCQ), and the Claim and Commitment Specificity (CCS). The raw pillars are defined as follows:

(3) $$GCCI_{c,t} = \frac{Green\ Claims\ and\ Commitments_{c,t}}{Environmental\ Sentences_{c,t}},$$

(4) $$CCQ_{c,t} = \frac{\overline{Claim\ Quality}_{c,t} + \overline{Commitment\ Quality}_{c,t}}{2},$$

(5) $$CCS_{c,t} = \overline{Specificity\ score}_{c,t}.$$

Following the common standardization approach (Di & Li, 2023; Yu et al., 2020), confirmed by Lublóy et al. (2025) each pillar is normalized relative to the cross-sectional distribution of companies within the same industry i at time t. The standardized Green Talk score for company c in year t is defined as:

(6) $$GT_{c,t}^{std} = \frac{1}{3}\sum_k \frac{p_{c,t}^k - \mu_{i,t}^k}{\sigma_{i,t}^k},$$

where $k \in \{GCCI, CCQ, CCS\}$ indexes the three pillars, $p_{c,t}^k$ denotes the raw pillar value of company c at time t, and $\mu_{i,t}^k$ and $\sigma_{i,t}^k$ are the industry-time-specific mean and sample standard deviation across all companies in industry i at time t. Demeaning by $\mu_{i,t}^k$ removes structural industry effects and common temporal shifts in disclosure norms, while scaling by $\sigma_{i,t}^k$ places all three pillars on a common unit-variance scale prior to aggregation. This procedure ensures that each contributes equally to the composite regardless of its empirical distribution. The Green Talk score $GT_{c,t}^{std}$ captures a firm's

relative environmental communication quality within its industry each year, with positive values indicating above-average levels across amount, quality and specificity dimensions.

On the side of the Green Walk, the measure relies on environmental risk incidents sourced from the RepRisk database, following Bingler et al. (2024). In constructing this measure, not only the frequency of environmental risk incidents is considered, but also their severity, reach, and novelty as assessed by RepRisk's proprietary scoring methodology. This approach reflects the rationale that incident quantity alone does not adequately capture a companies' actual environmental risk profile. A company causing a high number of incidents of low severity and limited reach should be evaluated more favourably than a company with equal incidents of higher severity, widely impactful and reoccurring. The Green Walk score for company c in year t is defined as:

(7) $$GW_{c,t} = (N_{c,t}^{env} - N_{c,t}^{Greenwashing}) * \frac{1}{3}(\overline{Severity}_{c,t} + \overline{Reach}_{c,t} + \overline{Novelty}_{c,t}),$$

where $N_{c,t}^{env}$is the total number of environmental risk incidents and $N_{c,t}^{Greenwashing}$ is the total number of incidents flagged as greenwashing. Excluding $N_{c,t}^{Greenwashing}$ from the Green Walk score avoids circularity in the validation exercise as discussed in the section 3.5 Validation and Robustness of the Greenwashing Score. $\overline{Severity}_{c,t}$, $\overline{Reach}_{c,t}$, and $\overline{Novelty}_{c,t}$ are the equally weighted company-year averages of RepRisk's incident-level scores respectively. Higher values of $GW_{c,t}$ indicate worse environmental performance, reflecting a greater number of incidents and/or incidents of higher impact. To construct the main specification of the Greenwashing Score, the Green Walk score is standardized within industry i and year t analogously to the Green Talk score:

(8) $$GW_{c,t}^{std} = \frac{GW_{c,t} - \mu_{i,t}}{\sigma_{i,t}},$$

ensuring that GT & GW are on a comparable scale and that industry- and time-specific effects are removed prior to aggregation. A positive $GT_{c,t}^{std}$ indicates that a company talks "greener" than its industry average, while a positive $GW_{c,t}^{std}$ indicates more frequent environmental incidents (and/or incidents of greater severity, reach, and/or novelty) than its industry average.

The Greenwashing Score is therefore defined as:

(9) $$GS_{c,t}^{std} = GT_{c,t}^{std} + GW_{c,t}^{std}.$$

A positive $GS_{c,t}^{std}$ indicates that a company's Green Talk exceeds its actual Green Walk relative to industry peers, indicating greenwashing (Bernini & La Rosa, 2024; Pizzetti et al., 2021; Ruiz-Blanco et al., 2022). A negative $GS_{c,t}^{std}$ indicates the opposite, meaning a company that delivers more (fewer incidents) than it communicates relative to peers, called greenhushing (Font et al., 2017). Consequently, $GS_{c,t}^{std}$ close to zero indicate a company "walks their talk" neither indicating greenwashing nor greenhushing.

To assess the sensitivity of the results to the industry normalization and to enable cross-industry comparisons, a robust specification is additionally reported. Both sides of the GS enter as cross-sectional percentile ranking across all companies each year:

(10) $$GS_{c,t}^{pct} = Rank_{c,t}^{GT} + Rank_{c,t}^{GW},$$

where $Rank_{c,t}^{GT}$ is the cross-sectional percentile rank of the Green Talk (GCCI, CCQ and CCS), and $Rank_{c,t}^{GW}$ is the cross-industry percentile rank of Green Walk (Equation (7)), both computed across all companies c in year t. Unlike the main GS (Equation (9)), this measure preserves between-industry variation, enabling claims about which industries exhibit systematically higher greenwashing gaps and how these gaps evolve over time.

3.5. Validation and Robustness of the Greenwashing Score

To assess the validity of the Greenwashing Score, a multi-step validation framework is employed using third-party greenwashing incidents that RepRisk records from external media sources. The contemporaneous and lagged[8] count of RepRisk greenwashing incidents per company-year ($GI_{c,t}$) is a rather informative validation outcome, while the primary validation outcome is the binary indicator of incident occurrence:

(11) $$GW\_Flag_{c,t} = \mathbb{1}[GI_{c,t} > 0],$$

where $\mathbb{1}$ is an indicator function that returns 1 when its specified condition is true and 0 when it is false. If the Greenwashing Score captures genuine greenwashing behavior, companies with higher GS in year t should be more likely to receive independently identified greenwashing incident flags from RepRisk in the same or subsequent years.

[8] In the validation literature, the terms lead and lag are perspective-dependent: from the incident side, incidents in year t+1 are lagged one period relative to the score in year t, while from the score side, the score leads the incidents by one year. Throughout this study the lag notation follows the incident perspective, such that lag k denotes incidents observed k years after the score, which is equivalent to the score leading incidents by k years.

A bivariate correlation analysis and a panel regression with company and year fixed effects are conducted to assess the validity of the GS. Both are methodologically complementary validation tests operating on different sources of variation: The correlation examines whether companies with systematically higher GS ranks attract disproportionately more greenwashing incidents in the cross-section. The panel regression absorbs between-company variation through company fixed effects and examines whether within-company increases in GS relative to industry peers predict a higher probability of greenwashing incident detection in the same year. Before the validation tests are performed both get tested for their robustness against different specifications. For each specification the entire Green Talk pipeline is rerun, with the Green Walk scores held fixed.

To verify that the validation does not hinge on the chosen data-quality thresholds, both validation steps are re-estimated under a Lenient (ENV $\geq$ 5, CC $\geq$ 2, industry-year N $\geq$ 3) and Strict (ENV $\geq$ 15, CC $\geq$ 5, industry-year N $\geq$ 5) threshold specification, bracketing the main medium specification (ENV $\geq$ 10, CC $\geq$ 3, industry-year N $\geq$ 4). Those specifications set minimum thresholds on environmental sentences (ENV), claims and/or commitments (CC) and industries in each year. Stability is assessed as the maximum absolute difference in the coefficient between any two specifications, with values below 0.05 taken as evidence of threshold insensitivity. The results confirm robustness on both validation steps as reported in the Appendix Table XIII and Appendix Table XIV. In the bivariate correlation, Pearson's r is virtually unchanged and highly significant throughout: the incident-count correlation is flat (r = 0.262 / 0.257 / 0.262; maximum difference 0.005) and the flag correlation varies only marginally (r = 0.327 / 0.332 / 0.334; maximum difference 0.007), both at $p < 0.001$. The panel regression is similarly stable. The incident-count coefficient stays within a narrow band (β = 0.114 / 0.128 / 0.136; maximum difference 0.022) but remains insignificant under every specification. The flag coefficient is likewise stable (β = 0.055 / 0.060 / 0.043; maximum difference 0.017) and remains positive and significant across all three specifications. Two features are worth commenting on. First, the flag significance eases from the 5% level under the Lenient and Medium thresholds to the 10% level under the Strict threshold in the panel regression. This reflects the contraction of the estimation sample as the thresholds tighten (N falls from 1,212 to 997 company-years). Second, both the flag

correlation and the flag regression coefficient peak at the medium specification rather than moving monotonically. It is the balanced choice as lenient thresholds include company-years with too little environmental content to yield a reliable Green Talk signal, whereas strict thresholds shrink both the sample and the industry-year peer groups used for standardization. Taken together, the validations are both insensitive to the threshold definitions. Going forward, the medium specification is used throughout.

Since the rule-based scoring approach within the Claim and Commitment Quality is only theoretically motivated but not empirically calibrated, the sensitivity of the results to the weighting scheme is tested against two alternative specifications. The equal-weighted specification assigns a uniform weight of 0.2 to each of the binary signals (climate context, numeric figure, time horizon, strong climate term and SBTi validation). In the binary specification each sentence that contains at least one of the binary signals gets a quality score of 1.0 and 0 otherwise. The results confirm that the validation results are insensitive to chosen weighting schemes as visible in the Appendix Table XV. The bivariate correlation barely moves ($r = 0.255 / 0.257 / 0.261$ for Binary / Tiered / Equal-weighted; all $p < 0.001$), with no pairwise difference exceeding 0.0053, and the regression flag coefficient is similarly stable across the full period ($\beta = 0.0649 / 0.0595 / 0.0628$) and post-Paris ($\beta = 0.0835 / 0.0735 / 0.0747$), remaining significant under every scheme.[9] The theoretically grounded tiered specification is used going forward.

One concern that arises is the fact that both the Green Walk score and the validation benchmark come from RepRisk, which might lead to some mechanical correlation. However, there are three main reasons why this concern is less significant. First, the Green Walk score is based only on environmental incidents (ex. Greenwashing incidents). Second, the Green Talk pillar is completely based on transcripts from Bloomberg, making it an independent source of data for the Greenwashing Score. Third, a placebo test is conducted to see if the relationship observed could come from shared characteristics of the data source. If that were the case, randomly shuffled Greenwashing Scores would be expected to yield similar statistics, but they don't, as detailed in 4.3 Placebo Testing.

[9] While the stability of results across quality specifications is reassuring, it should be interpreted cautiously. Since CCQ represents only one third of Green Talk and one sixth of the total Greenwashing Score, its marginal influence on the final score may be limited after industry-year standardization, meaning the observed robustness may partly reflect the dominance of the quantity and specificity pillars rather than genuine insensitivity to quality measurement choices.

4. RESULTS

This section reports the multi-step validation of the Greenwashing Score against the third-party greenwashing benchmark. First, the methodology of the bivariate correlation analysis, the panel regression and the permutation placebo testing is clarified. Then their results are presented and interpreted. Together, these tests establish whether the score captures a statistically significant company-level signal rather than an artifact of the shared data structure.

4.1. Bivariate Correlation Validation Results

In the first step, a bivariate correlation analysis examines whether a positive relationship exists between the company-year Greenwashing Score $GS_{c,t}^{pct}$ and greenwashing incidents $GI_{c,t}$ at the contemporaneous and lagged horizons, as well as the binary greenwashing flag $GW\_Flag_{c,t}$:

(12) $\rho\left(GS_{c,t}^{pct}, GI_{c,t+k}\right) \quad k \in \{0, 1, 2\}$,

(13) $\rho\left(GS_{c,t}^{pct}, GW\_Flag_{c,t}\right)$.

The percentile-rank specification $GS_{c,t}^{pct}$ is used here rather than the standardized specification $GS_{c,t}^{std}$. Because $GS_{c,t}^{std}$ is normalized within industry-year cells, values are only comparable within the same industry and year, making cross-sectional correlation with RepRisk greenwashing incidents (non-normalized) methodologically unsound. $GS_{c,t}^{pct}$, computed as the sum of cross-sectional percentile ranks of raw Green Talk and Green Walk scores within each year, preserves between-industry variation and is directly comparable across the full cross-section, aligning with the industry-agnostic RepRisk validation benchmark. A significantly positive Pearson correlation coefficient provides initial evidence that companies with higher Greenwashing Scores are independently identified as more prone to greenwashing behavior, validating the directional accuracy of the measure. The Fisher z-transformation is applied to construct 95% confidence intervals. Both the raw incident count $GI_{c,t+k}$ and the binary flag $GW\_Flag_{c,t}$ are used as outcome variables. The analysis is repeated separately for the full sample period (2012-2022), the pre- and post-Paris agreement sub-periods (2012-2016 and 2017-2022).

TABLE III

BIVARIATE CORRELATION OF THE MAIN SPECIFICATION (MEDIUM) OVER TIME

Pearson correlations between GS^{pct} (cross-sectional percentile-rank Greenwashing Score) and RepRisk GI counts and binary flags at contemporaneous (lag 0) and lagged (lag 1, lag 2) horizons. Pearson's r (*** p < 0.01, ** p < 0.05, * p < 0.10). In this context GW refers to Greenwashing not Green Walk.

Period	Outcome labels	Lag 0	Lag 1	Lag 2
Full period	GW incident count	0.257***	0.189***	0.204***
(2012-2022)	GW incident flag	0.332***	-	-
Pre-Paris	GW incident count	0.286***	0.155	0.270**
(2012-2016)	GW incident flag	0.295***	-	-
Post-Paris	GW incident count	0.259***	0.209***	0.215***
(2017-2022)	GW incident flag	0.349***	-	-

At lag 0, the binary greenwashing flag correlation is about 8 percentage points higher than raw counts, while at lag 1 and lag 2 the flag correlation cannot be computed due to insufficient incident variation in the smaller lagged subsamples. Therefore, count correlations are used exclusively at those horizons. Incident counts tend to be noisier and right skewed while the binary flag is more robust as visible in Table III above. This suggests the Score is better suited to identifying companies prone to greenwashing than to predicting the intensity of incident exposure. The bivariate correlation illustrates that the post-Paris agreement period yields a stronger correlation than the pre-Paris period (flag lag 0: 0.295 pre vs. 0.349 post), suggesting that the Greenwashing score becomes more predictive of independently identified greenwashing incidents with an intensifying external scrutiny as visible in Table III above. This is consistent with the broader post-Paris institutional context: the European Banking Authority (EBA) (2023, p. 5) documents a clear increase in potential greenwashing cases across all sectors since 2012, attributing this to heightened public attention and growing corporate accountability for environmental claims. Table III above shows that a further structural shift in the results is that the lag 1 incident count correlation is not statistically significant in the pre-Paris period ($r = 0.155$, $p = 0.15$) but is significant post-Paris ($r = 0.209$, $p < 0.01$), suggesting that the detection horizon may have shortened after 2015 as scrutiny intensified and detection became faster.[10] Notably, the lag 2 correlation ($r = 0.204$) marginally exceeds lag 1 ($r = 0.189$) in the full period, which may reflect that some greenwashing incidents

[10] The non-significance of the lag 1 correlation in the pre-Paris period ($r = 0.155$, $p = 0.15$) should be interpreted with caution given the small sample size ($N = 88$). At this sample size, a minimum r of approximately 0.21 is required to achieve statistical significance at the 5% level, meaning the result may reflect insufficient power rather than a genuine absence of predictive validity.

take longer to formally surface, though given the small and compositionally different subsamples at each lag horizon the difference should not be over-interpreted. Nevertheless, some predictive power is evident: the year-t Score remains significantly correlated with greenwashing incidents one and two years ahead (full-period r = 0.189 and 0.204, both $p < 0.01$), indicating that a high GS anticipates incidents that surface later. Even though the lag 1 and lag 2 correlations are weaker than those at lag 0, they show that the GS has some predictive power.

4.2. Robustness of the Panel Regression Results

In the second step, a panel regression is estimated to test whether this relationship holds after controlling for potential confounders. A Panel regression with company and year fixed effects is preferred over pooled OLS because it absorbs all stable firm-level characteristics that might jointly drive both the Greenwashing Score and greenwashing incident occurrence, ensuring that the coefficient is identified purely from within-firm year-to-year variation rather than from between-firm differences. The panel regression formulas are:

(14) $$GI_{c,t} = \alpha + \beta * GS_{c,t}^{std}, + \gamma X_{c,t} + \delta_c + \nu_t + \epsilon_{c,t},$$

(15) $$GW\_Flag_{c,t} = \alpha + \beta * GS_{c,t}^{std}, + \gamma X_{c,t} + \delta_c + \nu_t + \epsilon_{c,t},$$

where β is the coefficient of interest, expected to be positive and statistically significant if the GS captures genuine greenwashing behavior. γ is a vector of coefficients and $X_{c,t}$ is a vector of control variables including firm size (log market capitalisation), total environmental incident exposure (log (environmental incidents +1)) and industry-year peer group size (log number of companies), all measured at the company-year level. The log (+1) transformation is applied to environmental incidents to retain company-years with zero incidents, avoiding the selection bias that would arise from dropping clean firms. δ_c and ν_t denote company and year fixed effects respectively, absorbing time-invariant firm characteristics and common time trends. Standard errors are clustered at the company level to account for serial correlation within firms over time. The standardized specification $GS_{c,t}^{std}$ is employed as the regressor in this step. Unlike $GS_{c,t}^{pct}$, which is constructed as a cross-sectional percentile rank within each year, $GS_{c,t}^{std}$ is normalised relative to industry-year peers and therefore captures variation along two dimensions: between companies within the same industry-year cell, and within the same

company across years as its relative communication and environmental performance evolve. Company fixed effects absorb the former by subtracting each company's time-average from every observation, leaving only the latter to identify β. This within-firm temporal variation is precisely the variation that $GS_{c,t}^{std}$ is designed to capture and that the fixed effects estimator exploits. $GS_{c,t}^{pct}$ is poorly suited to this identification strategy, because cross-sectional ranks are recomputed annually across all companies. A company's rank in any given year reflects both genuine changes in behavior and mechanical reranking due to sample composition shifts. Its residual year-to-year variation after company fixed effects demeaning is both noisy and strongly collinear with the environmental incident control.

Fixed effects are removed via the within-transformation, which subtracts each variable's group mean from itself so that unit-specific intercepts vanish algebraically before OLS is applied to the demeaned variables (Arellano, 1987; Mundlak, 1978). With two fixed-effect dimensions (Company and Year) simultaneous demeaning is not feasible in a single pass, as demeaning by one dimension alters the means of the other. The alternating projections algorithm (Guimarães & Portugal, 2010) resolves this iteratively: the outcome and each regressor are demeaned by company, then by year, repeating until convergence and producing the same result as including a full set of unit and time dummies (Frisch & Waugh, 1933; Lovell, 1963).

Because each company appears in multiple years, the errors within a company are likely serially correlated (greenwashing tends to persist across years) and ignoring this understates the standard errors. They are therefore clustered by company using the Liang-Zeger sandwich estimator (1986), which allows arbitrary error correlation within clusters while assuming independence across them[11].

Three models are estimated for Equation (14) and (15), differing only in their fixed-effect structure. Model A (company + year fixed effects, SE clustered by company) is the primary specification: by including a separate intercept for each company, it absorbs all stable firm-level characteristics that might jointly drive the Greenwashing Score and

[11] A small-sample correction of G/(G-1)·(n-1)/(n-k) is applied to guard against downward bias in the variance estimate when the number of clusters G is small, as recommended by Cameron and Miller (2015). This correction is particularly relevant for Model B, where the nine industry clusters provide limited degrees of freedom for cluster-robust inference.

incident occurrence, so that β is identified purely from within-company year-to-year variation. Model B replaces company intercepts with industry intercepts (industry + year FE, SE clustered by industry), identifying β from variation across companies within the same industry; its clustered standard errors should however be interpreted with caution, as the nine industry clusters fall well below the minimum of 20-50 recommended for reliable sandwich estimation (Cameron & Miller, 2015) and it is therefore reported for completeness rather than as a primary result.

An industry-level breakdown additionally runs Model A for each industry meeting minimum thresholds (≥ 30 company-year observations and ≥ 5 unique companies), testing whether the aggregate result is driven by specific sectors and whether the effect varies across industries with differing environmental profiles.

TABLE IV

PANEL REGRESSION RESULTS MODEL A

Model A at lag 0: Company and year fixed effects, standard errors clustered by company. Controls: log (market capitalization), log (environmental incidents + 1), log (industry-year peer group size). Additionally, an industry sub-regression for Utilities, the only industry yielding a significant result, stable across all three threshold specifications. n = company-year observations; G = number of clusters.

Period	Outcome	β	SE	t	p	Sig	n	G
Full Period	GW incident count	0.124	0.086	1.449	0.149		1,090	240
(2012-2022)	GW incident flag	0.060	0.024	2.521	0.012	**	1,090	240
Pre-Paris	GW incident count	0.030	0.162	0.187	0.852		345	132
(2012-2016)	GW incident flag	-0.007	0.048	-0.134	0.894		345	132
Post-Paris	GW incident count	0.117	0.100	1.164	0.246		745	226
(2017-2022)	GW incident flag	0.074	0.032	2.335	0.020	**	745	226
Full Period	GW incident count	0.478	0.186	2.579	0.016	**	229	27
Utilities	GW incident flag	0.091	0.052	1.743	0.093	*	229	27

Table IV above shows that the results of the panel regression are directionally consistent with the findings of the bivariate correlation analysis. Estimating Equation (15) with company and year fixed effects, the Greenwashing score yields a statistically significant positive coefficient on the binary greenwashing incident flag at lag 0 across the full period (β = 0.060, p = 0.012**) and in the post-Paris sub-period (β = 0.074, p = 0.020**). The pre-Paris period yields no significant results. The noisiness of the raw incident count yields insignificant outcomes. Lag 1 and lag 2 results are not reported. At those horizons, lagged incident counts are only available for the subset of firms with a RepRisk record in the respective lagged year, producing a selected sample of incident-

prone firms that is not representative of the full panel. The binary flag is undefined at lag 1 and lag 2 due to insufficient outcome variation after correct missing value propagation.[12]. As anticipated, Model B yields no significant results across any outcome or period, as visible in the Appendix Table XVI.

The weaker aggregate result relative to the bivariate correlation is expected and methodologically explicable. Company fixed effects remove all stable between-company variation, leaving only within-company year-to-year changes in GS to identify the coefficient. This demanding identification strategy discards much of the cross-sectional signal the bivariate analysis exploits. The within-company standard deviation ratios range from 0.33 to 0.51 across most industries, meaning fixed effects remove between 49% and 67% of total GS variation before estimation begins. A further constraint is that the Greenwashing Score is substantially correlated with the log environmental incident control across most industries. Pearson's r exceeds 0.75 in Communications, Materials, Consumer Discretionary and Utilities, consistent with a partial data overlap between Green Walk and the log environmental incident control. Nevertheless, substantive co-movement between green communication and environmental activity cannot be ruled out. This collinearity causes the control to absorb much of the remaining GS signal, pressing the coefficient toward zero in most industries even where a true effect may exist.

The industry-level breakdown reveals that Utilities is the only industry yielding consistently significant results across both outcome variables: As Table IV above illustrates, a one standard deviation increase in GS^{std} is associated with 0.478 additional greenwashing incidents in the same year ($\beta = 0.478, p = 0.016^{**}$) and a 9.1 percentage point higher probability of receiving at least one greenwashing incident flag ($\beta = 0.091, p = 0.093^{*}$). These results are stable across all three threshold specifications ($|\beta_{Lenient} - \beta_{Strict}| = 0.022$), representing the strongest industry-level validation finding in the analysis and consistency with the work of He et al. (2024) which similarly identify Utilities as the industry with the highest greenwashing intensity among all industries in their US sample. Whether the Utilities finding reflects a genuinely stronger

[12] In an earlier implementation, missing lagged incident values were treated as zero rather than as missing observations, artificially inflating the no-incident class at lag 1 and lag 2 and retaining the full panel at those horizons. The corrected implementation propagates missing values correctly, ensuring only firm-years with a confirmed RepRisk record in the lagged year are included in the lagged regressions.

greenwashing dynamic or simply more favourable estimation conditions, including a sufficient incident rate, adequate within-firm variation and a sample size large enough to identify the coefficient, cannot be determined from the available data. The absence of significant results in other industries should be interpreted as inconclusive rather than as evidence against the validity of the Greenwashing Score in those sectors, given the identifiable data and estimation constraints present in each case.

4.3. Placebo Testing

For the two validation steps, a permutation placebo test is performed to confirm that the validation results represent actual company-level signal rather than an artifact of the data structure or shared RepRisk data source. In both cases, Greenwashing Score values are randomly reassigned across companies while preserving temporal structure. The observed test statistic is compared against the distribution generated across 5,000 permutations. The percentage of permutations where the absolute permuted statistic is equal or greater than the absolute observed statistic is known as the permutation p-value:

$$(16) \quad \mathrm{p_{perm}} = \left(\frac{1}{\mathrm{B}}\right) * \sum_{b=1}^{B} 1[|\mathrm{stat}_b| \geq |stat_{obs}|],$$

where B = 5,000 and stat_b denotes the statistic computed on the b-th permuted dataset. For the correlation (Test 1), GS^{pct} values are shuffled across companies within each year. For the regression (Test 2), GS^{std} values are shuffled within each industry-year cell. For the correlation the Appendix Table XVII shows that across 5,000 permutations, the observed r = 0.257 lies 8.87 standard deviations above the permuted mean of 0.000, with a permuted 95% confidence interval of [-0.057, 0.057], placing the observed result entirely outside the null distribution. For the panel regression, the observed β = 0.060 lies 6.87 standard deviations above the permuted mean of -0.001, with a permuted 95% confidence interval of [-0.019, 0.016]. In both cases the permutation p-value is below 0.001. The validation findings cannot be attributed to mechanical correlation arising from the shared data source or the panel structure. Therefore, the Greenwashing Score captures a genuine and statistically distinguishable company-level signal.

5. LIMITATIONS AND FUTURE RESEARCH

Several limitations of this study should be noted. First, the framework only detects greenwashing in companies that actively make environmental claims meaning firms engaging in strategic non-disclosure remain invisible to the score, as the Green Talk pillar requires sufficient environmental language to produce a meaningful signal. Future research could complement the talk-based score with a disclosure-expectation model. This model could predict how much environmental language a company should produce given its size, sector and incident exposure. Therefore, companies communicating far less than expected (greenhushing or strategic silence) become visible rather than dropping out of the score.

Second, the underlying NLP models struggle with contextual ambiguity. Common words frequently produce false positives (e.g. "The demand in Germany is sustainable") and context related words produce false negatives (e.g. missing "Our commitment to the Paris agreement"). This is not unique to this study. Tasks requiring ambiguity, subjectivity or contextual reasoning remain fundamentally unsolved, meaning companies using such language more frequently may receive a systematically distorted Green Talk score. Future research could analyse industry-specific false-positive- and false-negative-rates to systematically correct distorted Green Talk scores ex post.

Third, the underlying models were trained on written ESG disclosures rather than verbal corporate communication. This introduces an additional domain shift that may affect classification precision. An expert-annotated corpus of verbal corporate communication would be a natural next step to develop a verbal transcript-native model.

Fourth, even though there is no data overlap between the Greenwashing Score and the validation benchmark, both the Green Walk component of GS and the log environmental incident control variable used in the panel regression, draw from the same RepRisk environmental incident data. Sourcing the environmental incident control variable from an independent provider would be a promising alternative approach for future research.

Fifth, the company fixed effects estimator is highly demanding by design. It removes a substantial share of GS variation before estimation begins, which likely explains the limited number of significant regression results beyond the flag outcome. Future research

should extend the panel across more years, regions and company sizes to restore statistical power of industry-specific regression results.

Sixth, the lagged validation is constrained by RepRisk coverage structure. At lag 1 and lag 2 only firms with a confirmed incident record in the lagged year are retained, producing a selected subsample that is not representative of the full panel.

Seventh, although this MFW minimizes manual labelling by limiting it to a binary classification of environmental content, this step remains inherently subjective and may introduce bias into the GTI, despite the task's relative simplicity.

Eighth, the claim and commitment quality (CCQ) is theoretically motivated. Even though the robustness to different quality specifications is validated, future research or maybe even regulatory authorities should develop a robust framework on the quality-assessment of environmental claims and commitments.

Finally, the sample is limited to large-cap European companies over 2012-2022 and findings may not generalise to smaller firms or markets outside this coverage universe.

6. CONCLUSION

This thesis began with a simple but consequential question: when company executives and investor relations speak about the environment, can a machine tell whether their words are matched by the companies' actions? More precisely, it asks whether an NLP-based pipeline can measure greenwashing based on the gap between verbal corporate communication and real-world environmental risk incidents of large-cap European listed companies over the period 2012-2022. The findings suggest a meaningful but conditional answer. The gap between Green Talk and Green Walk is measurable and quantifiable at scale, yet only for companies that choose to speak.

The constructed Greenwashing Score earns its claim to validity on three fronts. First, its bivariate correlation with third-party greenwashing incidents is positive, highly significant and remarkably stable. The Pearson correlations of approximately 0.26 to 0.33 (full period) across every threshold specification tested, strengthen in the post-Paris sub-period when external scrutiny intensified. The relationship is also forward-looking: the year-t Score remains significantly correlated with greenwashing incidents one and two years later ($r = 0.189$ and 0.204, both $p < 0.01$), so a high GS today anticipates incidents that surface later. This predictive signal is weaker than the contemporaneous one, but it points to a measure that can flag companies before the incidents appear rather than only alongside them. The panel regression provides partially confirming evidence: The binary incident flag is significant across both the full sample and the post-Paris years, with the Utilities sector providing the clearest and only statistically significant, industry-level confirmation. The placebo test settles the most obvious objection by confirming that across 5,000 permutations not one produced a statistic approaching the observed values ($z = 8.87$ and $z = 6.87$). Thereby a genuine company-level relationship is confirmed rather than an artifact of a shared data source.

The Greenwashing Score can therefore detect greenwashing-prone firms robustly, while quantification of intensity is supported weaker. Its reach is bound by the communicative behavior of the companies it covers. Companies engaging in strategic non-disclosure remain invisible to the score by construction. Current limitations of NLP-models in parsing contextually ambiguous spoken language introduce measurement uncertainty that cannot be fully resolved with existing tools.

Within those bounds, this MFW makes four contributions to literature. First, it applies NLP-based greenwashing detection to verbal rather than written corporate communication. The verbal communication channel that is largely unaudited and unregulated yet consequential for capital allocation, as earnings calls are a primary source through which institutional investors monitor companies. Second, it introduces a multi-dimensional Green Talk scoring framework that captures quantity, quality and specificity of environmental claims and commitments, offering a more granular measure than binary classification approaches and responding to the call by Calamai et al. (2026) for research that quantifies the information content of green communication rather than its volume alone. Third, it employs RepRisk environmental incident data as a validation benchmark, an approach identified by Lublóy et al. (2025) as a promising but underutilized direction in the greenwashing measurement literature. Fourth, it extends NLP-based detection to the European equity market, addressing a geographic gap in a literature dominated by US- and China-focused studies.

That the thesis stops at measurement and validation, rather than reaching for financial implications, is a deliberate choice: a score must be shown to mean something before it can be used to explain anything. With the foundation of the score future research can now ask the financial questions this thesis intentionally leaves open: whether high-GS companies face a higher cost of capital, whether the score predicts abnormal returns around greenwashing incident disclosures or whether institutional investors systematically price greenwashing risk in portfolio allocation decisions. The strengthening of the validation signal in the post-Paris period suggests that these financial implications may be particularly pronounced in a regulatory environment of increasing scrutiny - precisely the environment European markets now operate in.

The thesis opened with Volkswagen, whose "clean diesel" campaign took seven years to unravel. The pipeline built here cannot prevent the next such greenwashing incident. But it offers regulators, investors and researchers something they currently lack: a scalable, externally grounded and statistically validated instrument for timely measurement and early warning of greenwashing in the channel where companies speak most freely.

Words: 11,953

REFERENCES

Amoozegar, A., Berger, D., Cao, X., & Pukthuanthong, K. (2020). Earnings Conference Calls and Institutional Monitoring: Evidence from Textual Analysis. *Journal of Financial Research*, *43*(1), 5–36. https://doi.org/10.1111/jfir.12199

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models* (arXiv:1908.10063). arXiv. https://doi.org/10.48550/arXiv.1908.10063

Arellano, M. (1987). PRACTITIONERS' CORNER: Computing Robust Standard Errors for Within-groups Estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434. https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x

Barth, F., Eckert, C., Gatzert, N., & Scholz, H. (2022). Spillover Effects from the Volkswagen Emissions Scandal: An Analysis of Stock and Corporate Bond Markets. *Schmalenbach Journal of Business Research*, *74*(1), 37–76. https://doi.org/10.1007/s41471-021-00121-9

Berg, F., Kölbel, J. F., & Rigobon, R. (2019). *Aggregate Confusion: The Divergence of ESG Ratings* (SSRN Scholarly Paper No. 3438533). Social Science Research Network. https://doi.org/10.2139/ssrn.3438533

Bernini, F., & La Rosa, F. (2024). Research in the greenwashing field: Concepts, theories, and potential impacts on economic and social value. *Journal of Management and Governance*, *28*(2), 405–444. https://doi.org/10.1007/s10997-023-09686-5

Bingler, J., Kraus, M., & Leippold, M. (2021). *Cheap Talk and Cherry-Picking: What ClimateBert has to say on Corporate Climate Risk Disclosures* (SSRN Scholarly Paper No. 3796152). Social Science Research Network. https://doi.org/10.2139/ssrn.3796152

Bingler, J., Kraus, M., Leippold, M., & Webersinke, N. (2024). How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, *164*, 107191. https://doi.org/10.1016/j.jbankfin.2024.107191

Blau, B. M., DeLisle, J. R., & Price, S. M. (2015). Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from

short sales. *Journal of Corporate Finance*, *31*, 203–219. https://doi.org/10.1016/j.jcorpfin.2015.02.003

Calamai, T., Balalau, O., Guenedal, T. L., & Suchanek, F. M. (2026). *Detecting Greenwashing: A Natural Language Processing Literature Survey* (arXiv:2502.07541). arXiv. https://doi.org/10.48550/arXiv.2502.07541

Cambridge Centre for Sustainable Finance. (2016). *Environmental Risk Analysis by Financial Institutions*. Cambridge Institute for Sustainability Leadership.

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *The Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Capetz, M., Vinella, A., Pattichis, R., Chance, C., Ghosh, R., & Chang, K.-W. (2024). *Leveraging Language Models to Detect Greenwashing* (arXiv:2311.01469). arXiv. https://doi.org/10.48550/arXiv.2311.01469

Chopra, S. S., Senadheera, S. S., Dissanayake, P. D., Withana, P. A., Chib, R., Rhee, J. H., & Ok, Y. S. (2024). Navigating the Challenges of Environmental, Social, and Governance (ESG) Reporting: The Path to Broader Sustainable Development. *Sustainability*, *16*(2), 606. https://doi.org/10.3390/su16020606

Chossière, G. P., Malina, R., Ashok, A., Dedoussi, I. C., Eastham, S. D., Speth, R. L., & Barrett, S. R. H. (2017). Public health impacts of excess NOx emissions from Volkswagen diesel passenger vehicles in Germany. *Environmental Research Letters*, *12*(3), 034014. https://doi.org/10.1088/1748-9326/aa5987

Di, R., & Li, C. (2023). The cost of hypocrisy: Does corporate ESG decoupling reduce labor investment efficiency? *Economics Letters*, *232*, 111355. https://doi.org/10.1016/j.econlet.2023.111355

Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 Amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as Regards Corporate Sustainability Reporting (Text with EEA Relevance), EP, CONSIL, 322 OJ L (2022). http://data.europa.eu/eli/dir/2022/2464/oj

DWS. (2023, March 17). *DWS publishes 2022 Annual Report | DWS*. https://www.dws.com/en-gb/our-profile/media/media-releases/dws-publishes-2022-annual-report/

EBA. (2023, June 1). *ESAs present common understanding of greenwashing and warn on related risks | European Banking Authority*. https://www.eba.europa.eu/publications-and-media/press-releases/esas-present-common-understanding-greenwashing-and-warn

Font, X., Elgammal, I., & Lamond, I. (2017). Greenhushing: The deliberate under communicating of sustainability practices by tourism businesses. *Journal of Sustainable Tourism*, *25*(7), 1007–1023. https://doi.org/10.1080/09669582.2016.1158829

Frisch, R., & Waugh, F. V. (1933). Partial Time Regressions as Compared with Individual Trends. *Econometrica*, *1*(4), 387–401. https://doi.org/10.2307/1907330

Gatti, L., Pizzetti, M., & Seele, P. (2021). Green lies and their effect on intention to invest. *Journal of Business Research*, *127*, 228–240. https://doi.org/10.1016/j.jbusres.2021.01.028

Ghitti, M., Gianfrate, G., & Palma, L. (2024). The agency of greenwashing. *Journal of Management and Governance*, *28*(3), 905–941. https://doi.org/10.1007/s10997-023-09683-8

Guimarães, P., & Portugal, P. (2010). A Simple Feasible Procedure to fit Models with High-dimensional Fixed Effects. *The Stata Journal*, *10*(4), 628–649. https://doi.org/10.1177/1536867X1101000406

Haliburton, L., Leusmann, J., Welsch, R., Ghebremedhin, S., Isaakidis, P., Schmidt, A., & Mayer, S. (2025). Uncovering labeler bias in machine learning annotation tasks. *AI and Ethics*, *5*(3), 2515–2528. https://doi.org/10.1007/s43681-024-00572-w

Handford, M., & Lofts, G. (2024, December 10). *EY Global Institutional Investor Survey 2024*. EY. https://www.ey.com/en_gl/insights/climate-change-sustainability-services/institutional-investor-survey

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

He, Q., Marshall, B. R., Nguyen, J. H., Nguyen, N. H., Qiu, B., & Visaltanachoti, N. (2024). *Greenwashing: Measurement and Implications* (SSRN Scholarly Paper No. 4981856). Social Science Research Network. https://doi.org/10.2139/ssrn.4981856

Hetzner, C. (2022, May 31). *Deutsche Bank raided by authorities over ESG "greenwashing" claims*. Fortune. https://fortune.com/2022/05/31/deutsche-bank-dws-esg-greenwashing-raid-evidence-seized-whistleblower-fixler/

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Huang, Y., Tai, W., Zhou, F., Gao, Q., & Zhong, T. (2025). Extracting key insights from earnings call transcript via information-theoretic contrastive learning. *Information Processing & Management*, *62*(3), 103998. https://doi.org/10.1016/j.ipm.2024.103998

Hummel, K., & Jobst, D. (2024). An Overview of Corporate Sustainability Reporting Legislation in the European Union. *Accounting in Europe*, *21*(3), 320–355. https://doi.org/10.1080/17449480.2024.2312145

IFRS. (2025, December 1). *IFRS S2 Climate-related Disclosures*. https://www.ifrs.org/issued-standards/ifrs-sustainability-standards-navigator/ifrs-s2-climate-related-disclosures/

Jong, W., & van der Linde, V. (2022). Clean diesel and dirty scandal: The echo of Volkswagen's dieselgate in an intra-industry setting. *Public Relations Review*, *48*(1), 102146. https://doi.org/10.1016/j.pubrev.2022.102146

Kathan, M. C., Utz, S., Dorfleitner, G., Eckberg, J., & Chmel, L. (2025). What you see is not what you get: ESG scores and greenwashing risk. *Finance Research Letters*, *74*, 106710. https://doi.org/10.1016/j.frl.2024.106710

Kowsmann, P., & Brown, K. (2021, August 1). WSJ News Exclusive | Fired Executive Says Deutsche Bank's DWS Overstated Sustainable-Investing Efforts. *Wall Street Journal*. https://www.wsj.com/finance/investing/fired-executive-says-deutsche-banks-dws-overstated-sustainable-investing-efforts-11627810380

Lerman, A., Steffen, T. D., & Zhang, K. (2022). *Earnings Conference Calls and the SEC Comment Letter Process* (SSRN Scholarly Paper No. 4501693). Social Science Research Network. https://doi.org/10.2139/ssrn.4501693

Li, W., Li, W., Seppänen, V., & Koivumäki, T. (2023). Effects of greenwashing on financial performance: Moderation through local environmental regulation and media coverage. *Business Strategy and the Environment*, *32*(1), 820–841. https://doi.org/10.1002/bse.3177

Li, W., & Zheng, X. Y. (2024). *Government Environmental Attention and Enterprise Greenwashing Behavior* (SSRN Scholarly Paper No. 4712124). Social Science Research Network. https://doi.org/10.2139/ssrn.4712124

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lovell, M. (1963). Seasonal Adjustment of Economic Time Series and Multiple Regression. *Cowles Foundation Discussion Papers*. https://elischolar.library.yale.edu/cowles-discussion-paper-series/380

Lublóy, Á., Keresztúri, J. L., & Berlinger, E. (2025). Quantifying firm-level greenwashing: A systematic literature review. *Journal of Environmental Management*, *373*, 123399. https://doi.org/10.1016/j.jenvman.2024.123399

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

Lyon, T. P., & Montgomery, A. W. (2015). The Means and End of Greenwash. *Organization & Environment*, *28*(2), 223–249. https://doi.org/10.1177/1086026615575332

Marquis, C., Toffel, M. W., & Zhou, Y. (2016). Scrutiny, Norms, and Selective Disclosure: A Global Study of Greenwashing. *Organization Science*, *27*(2), 483–504. https://doi.org/10.1287/orsc.2015.1039

Medya, S., Rasoolinejad, M., Yang, Y., & Uzzi, B. (2022). *An Exploratory Study of Stock Price Movements from Earnings Calls* (arXiv:2203.12460). arXiv. https://doi.org/10.48550/arXiv.2203.12460

MSCI Inc. (2024). *The Global Industry Classification Standard (GICS®)*. https://www.msci.com/indexes/index-resources/gics

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, *46*(1), 69. https://doi.org/10.2307/1913646

Oluh, A.-A. (2026). *Anwuli-Austin/environmental-bert-transcript-classification: Initial release - Master thesis submission* (Version v1.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.19551656

Pizzetti, M., Gatti, L., & Seele, P. (2021). Firms Talk, Suppliers Walk: Analyzing the Locus of Greenwashing in the Blame Game and Introducing 'Vicarious Greenwashing.' *Journal of Business Ethics*, *170*(1), 21–38. https://doi.org/10.1007/s10551-019-04406-2

Regulation (EU) 2019/2088 of the European Parliament and of the Council of 27 November 2019 on Sustainability-related Disclosures in the Financial Services Sector (Text with EEA Relevance), 317 OJ L (2019). http://data.europa.eu/eli/reg/2019/2088/oj

Regulation (EU) 2020/852 of the European Parliament and of the Council of 18 June 2020 on the Establishment of a Framework to Facilitate Sustainable Investment, and Amending Regulation (EU) 2019/2088 (Text with EEA Relevance), 198 OJ L (2020). http://data.europa.eu/eli/reg/2020/852/oj

RepRisk. (2020, July 21). *RepRisk reaches new milestone of evaluating more than 150,000 companies on ESG risks*. News Powered by Cision. https://news.cision.com/reprisk/r/reprisk-reaches-new-milestone-of-evaluating-more-than-150-000-companies-on-esg-risks,c3157441

RepRisk. (2021). *RepRisk | RepRisk methodology overview*. https://www.reprisk.com/insights/resources/methodology

RepRisk. (2025, December 1). *RepRisk Research Scope: Business Conduct Issues*. RepRisk. https://www.reprisk.com/content/static/reprisk-esg-issues-definitions.pdf

RepRisk. (2026). *RepRisk | Technology*. https://www.reprisk.com/technology

Ruiz-Blanco, S., Romero, S., & Fernandez-Feijoo, B. (2022). Green, blue or black, but washing–What company characteristics determine greenwashing? *Environment, Development and Sustainability*, *24*(3), 4024–4045. https://doi.org/10.1007/s10668-021-01602-x

SBTi. (2023, April 1). *SBTi Corporate Manual*. https://files.sciencebasedtargets.org/production/files/SBTi-Corporate-Manual-v2.1.pdf

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, *61*, 104979. https://doi.org/10.1016/j.frl.2024.104979

Sedilekova, Z., & Merchat, I. (2022, June 6). *DWS and Deutsche Bank raided over greenwashing allegations*. Lexology. https://www.lexology.com/library/detail.aspx?g=80ebdc10-5617-4d8c-b78b-05ce4ae2495d

Shah, R., Singh, G., & Puri, S. (2017, April 25). *Volkswagen Emission Scandal: Reputation Recovery and Recall Strategy | Harvard Business Impact Education*. Ivey Publishing. https://hbsp.harvard.edu/product/W17228-PDF-ENG

Stammbach, D., Webersinke, N., Bingler, J. A., Kraus, M., & Leippold, M. (2023). *Environmental Claim Detection* (arXiv:2209.00507). arXiv. https://doi.org/10.48550/arXiv.2209.00507

STOXX Ltd. (2026). 1-STOXX® Europe 600 (SXXP) (EU0009658202). *STOXX*. https://stoxx.com/index/sxxp/

TCFD. (2017). *Recommendations of the Task Force on Climate-Related Financial Disclosures*. https://www.fsb-tcfd.org/recommendations/

US EPA. (2015, September 18). *VW Notice of Violation, Clean Air Act (September 18, 2015)* [Letter]. https://www.epa.gov/enforcement/notices-violation

Wahde, M., Della Vedova, M. L., Virgolin, M., & Suvanto, M. (2024). An interpretable method for automated classification of spoken transcripts and written text.

Evolutionary Intelligence, *17*(1), 609–621. https://doi.org/10.1007/s12065-023-00851-1

Walker, K., & Wan, F. (2012). The Harm of Symbolic Actions and Green-Washing: Corporate Actions and Communications on Environmental Performance and Their Financial Implications. *Journal of Business Ethics*, *109*(2), 227–242. https://doi.org/10.1007/s10551-011-1122-4

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). *ClimateBert: A Pretrained Language Model for Climate-Related Text* (arXiv:2110.12010). arXiv. https://doi.org/10.48550/arXiv.2110.12010

Wu, M.-W., & Shen, C.-H. (2013). Corporate social responsibility in the banking industry: Motives and financial performance. *Journal of Banking & Finance*, *37*(9), 3529–3547. https://doi.org/10.1016/j.jbankfin.2013.04.023

Xu, C., Liu, J., Li, Z., & Lin, C. (2025). *DeepGreen: Effective LLM-Driven Greenwashing Monitoring System Designed for Empirical Testing -- Evidence from China*. https://doi.org/10.48550/ARXIV.2504.07733

Yu, E. P., Luu, B. V., & Chen, C. H. (2020). Greenwashing in environmental, social and governance disclosures. *Research in International Business and Finance*, *52*, 101192. https://doi.org/10.1016/j.ribaf.2020.101192

Zeng, F., Wang, J., & Zeng, C. (2025). An optimized machine learning framework for predicting and interpreting corporate ESG greenwashing behavior. *PLOS ONE*, *20*(3), e0316287. https://doi.org/10.1371/journal.pone.0316287

APPENDICES

APPENDIX TABLE V

PER-FOLD RESULTS

The fold-by-fold performance reveals how stable each model is across different subsets of the data. A high standard deviation signals sensitivity to which sentences end up in the test set.

Fold	M1 Acc.	M1 F1	M2 Acc.	M2 F1	M3 Acc.	M3 F1
1	0.9394	0.9426	0.9495	0.9509	0.8586	0.8889
2	0.9293	0.9294	0.9394	0.9400	0.9697	0.9698
3	0.9192	0.9259	0.9293	0.9386	0.9495	0.9509
4	0.8788	0.8815	0.9293	0.9311	0.9293	0.9341
5	0.9592	0.9599	0.9388	0.9417	0.9592	0.9592

APPENDIX TABLE VI

PER-CLASS PERFORMANCE (POOLED OOF)

The fold-by-fold performance reveals how stable each model is across different subsets of the data. A high standard deviation signals sensitivity to which sentences end up in the test set.

Model	Class	Precision	Recall	F1
M1: EB-b	Non-environmental (0)	0.9203	0.9315	0.9259
M1: EB-b	Environmental (1)	0.9300	0.9187	0.9243
M2: EB-e	Non-environmental (0)	0.9465	0.9274	0.9369
M2: EB-e	Environmental (1)	0.9283	0.9472	0.9376
M3: EB-e + LoRa	Non-environmental (0)	0.9778	0.8871	0.9302
M3: EB-e + LoRa	Environmental (1)	0.8959	0.9797	0.9359
Ensemble M2+M3	Non-environmental (0)	0.9673	0.9556	0.9615
Ensemble M2+M3	Environmental (1)	0.9558	0.9675	0.9616

APPENDIX TABLE VII

ENSEMBLE WEIGHT SENSITIVITY

The optimal blend weights M2 at 40% and M3 at 60%, with a classification threshold of 0.58. The table below shows how performance changes as the blend shifts from pure M3 (top) to pure M2 (bottom), at the fixed optimal threshold.

M2 weight	M3 weight	Macro F1	Accuracy	Note
0.00	1.00	0.9331	0.9332	Pure M3
0.10	0.90	0.9372	0.9372	
0.20	0.80	0.9351	0.9352	
0.30	0.70	0.9433	0.9433	
0.40	0.60	0.9615	0.9615	Optimal
0.50	0.50	0.9555	0.9555	
0.60	0.40	0.9494	0.9494	
0.70	0.30	0.9433	0.9433	
0.80	0.20	0.9372	0.9372	

0.90	0.10	0.9372	0.9372	
1.00	0.00	0.9372	0.9372	Pure M2

APPENDIX TABLE VIII

CONFUSION MATRICES (POOLED OOF, N = 494)

Each cell shows: actual class (rows) vs predicted class (columns).

Model	Pred: Non-env (0)	Pred: Env (1)	Total errors
M1: EB-b | Actual: Non-env	231	17	37 total
M1: EB-b | Actual: Env	20	226	
M2: EB-e | Actual: Non-env	230	18	31 total
M2: EB-e | Actual: Env	13	233	
M3: LoRA | Actual: Non-env	220	28	33 total
M3: LoRA | Actual: Env	5	241	
Ensemble M2+M3 | Actual: Non-env	237	11	19 total
Ensemble M2+M3 | Actual: Env	8	238	

APPENDIX TABLE IX

STATISTICAL SIGNIFICANCE - MCNEMAR'S TEST

McNemar's test is the corresponding chi-square test for paired data, and it tests whether two models make errors on the same sentence or systematically different errors. A significant result ($p < 0.05$) is the statistical requirement for claiming one classifier improves on another.

Comparison	chi^2	p-value	Significance	Interpretation
M3 (LoRA) vs. M2	0.023	0.880	ns[a]	Indistinguishable individually
M3 (LoRA) vs. M1	0.173	0.677	ns[a]	Indistinguishable individually
M2 vs. M1	0.962	0.327	ns[a]	Indistinguishable individually
Ensemble vs. M2	6.722	0.010	ss[b]	Ensemble makes different errors
Ensemble vs. M3	6.500	0.011	ss[b]	Ensemble makes different errors

(a) ns = not significant; (b) ss = statistically significant

APPENDIX TABLE X

LEVEL OF SEVERITY

	Consequences of the risk incident	Extent of the impact	Cause of the Risk Incident
Low (1)	No further	One person	By Accident
Medium (2)	Injury	Group of people	By negligence
High (3)	Death	Large number of people	In a systematic way

Source: (RepRisk, 2021)

APPENDIX TABLE XI

LEVEL OF REACH

	Classification by Reach
Low (1)	Local media, smaller NGOs, local governmental bodies and social media
Medium (2)	Most national and regional media, international NGOs, and state, national, and international governmental bodies
High (3)	Few truly global media outlets

Source: (RepRisk, 2021)

APPENDIX TABLE XII

LEVEL OF NOVELTY

	Classification by Novelty
Low (1)	Story development: a new development appears related to the same issue.
Medium (2)	Source escalation: the issue appears again in a more influential source.
High (3)	Six-week rule: the issue appears again for the same company in the same country after a six-week period, which is a potential signal that the issues are unresolved.

Source: (RepRisk, 2021)

APPENDIX TABLE XIII

ROBUSTNESS COMPARISON BIVARIATE CORRELATION – THRESHOLD SPECIFICATIONS

Stage 1 excludes company-years below the ENV & CC sentence threshold; Stage 2 excludes those whose industry-year cell is below the industry-year threshold. GT retained is the surviving Green Talk sample, and N is the subset also matched to a Green Walk score (only those observations enter the correlation). For robustness the maximum difference is computed. All results are full-period results.

Metric	Lenient	Medium (main)	Strict
ENV sentence threshold	5	10	15
CC sentence threshold	2	3	5
Industry-year N threshold	3	4	5
Stage 1 excluded	640	962	1,223
Stage 2 excluded	5	18	37
GT retained (post-exclusion)	2,488	2,153	1,873
N (observations)	1,339	1,203	1,096
Pearson r (GW incidents, lag 0)	0.262	0.257	0.262
p-value	< 0.001	< 0.001	< 0.001
Significance	***	***	***
$\lvert r_{Medium} - r_{Strict} \rvert$	**0.005 (< 0.05 → robust)**		
Pearson r (GW flag, lag 0)	0.327	0.332	0.334
p-value	< 0.001	< 0.001	< 0.001
Significance	***	***	***
$\lvert r_{Strict} - r_{Lenient} \rvert$	**0.007 (< 0.05 → robust)**		

APPENDIX TABLE XIV

ROBUSTNESS COMPARISON PANEL REGRESSION – THRESHOLD SPECIFICATIONS

Stage 1 excludes company-years below the ENV & CC sentence threshold; Stage 2 excludes those whose industry-year cell is below the industry-year threshold. GT retained is the surviving Green Talk sample, and N is the subset also matched to a Green Walk Score, less the observations removed by listwise deletion of the three controls (predominantly those missing log market capitalization). For robustness the maximum difference is computed. All results are full-period results. Stability metrics below 0.05 indicate insensitivity to the weighting choice (*** p < 0.01, ** p < 0.05, * p < 0.10).

Metric	Lenient	Medium (main)	Strict
ENV sentence threshold	5	10	15
CC sentence threshold	2	3	5
Industry-year N threshold	3	4	5
Stage 1 excluded	640	962	1,223
Stage 2 excluded	5	18	37
GT retained (post-exclusion)	2,488	2,153	1,873
N (observations)	1,212	1,090	997
β(GW incidents, lag 0)	0.114	0.128	0.136
p-value	0.135	0.134	0.167
Significance			
$\lvert\beta_{Lenient} - \beta_{Strict}\rvert$	**0.022 (< 0.05 → robust)**		
β (GW flag, lag 0)	0.055	0.060	0.043
p-value	0.015	0.012	0.090
Significance	**	**	*
$\lvert\beta_{Medium} - \beta_{Strict}\rvert$	**0.016 (< 0.05 → robust)**		

APPENDIX TABLE XV

ROBUSTNESS COMPARISON – QUALITY SPECIFICATIONS

Because weighting does not affect which company-years are retained, the sample is identical across all three schemes (correlation N = 1,203; regression N = 1,090 after control deletion). Each scheme is run through the full pipeline and validated against the RepRisk greenwashing benchmark. Pairwise stability metrics below 0.05 indicate insensitivity to the weighting choice (*** p < 0.01, ** p < 0.05, * p < 0.10).

Metric	Binary	Tiered (main)	Equal-weighted
Pearson r (full period, lag 0)	0.255	0.257	0.261
p-value	< 0.001	< 0.001	< 0.001
Significance	***	***	***
Regression β (full, flag lag 0)	0.0649	0.0595	0.0628
p-value (regression full)	0.0086	0.0124	0.0091
Significance (regression full)	***	**	***
Regression β (post-Paris flag)	0.0835	0.0735	0.0747
p-value (regression post)	0.0050	0.0204	0.0216
Significance (post-Paris)	***	**	**

$\lvert r_{Tiered} - r_{Equal} \rvert$ (stability metric)	0.0032 (< 0.05 threshold → robust)
$\lvert r_{Tiered} - r_{Binary} \rvert$ (stability metric)	0.0021 (< 0.05 threshold → robust)
$\lvert r_{Equal} - r_{Binary} \rvert$ (stability metric)	0.0053 (< 0.05 threshold → robust)
$\lvert \beta_{Tiered} - \beta_{Equal} \rvert$ (stability metric)	0.0033 (< 0.05 threshold → robust)
$\lvert \beta_{Tiered} - \beta_{Binary} \rvert$ (stability metric)	0.0054 (< 0.05 threshold → robust)
$\lvert \beta_{Equal} - \beta_{Binary} \rvert$ (stability metric)	0.0021 (< 0.05 threshold → robust)

APPENDIX TABLE XVI

PANEL REGRESSION RESULTS MODEL B

Model A at lag 0: Company and year fixed effects, standard errors clustered by company. Controls: log (market capitalization), log (environmental incidents + 1), log (industry-year peer group size). Additionally, an industry sub-regression for Utilities, the only industry yielding a significant result, stable across all three threshold specifications. n = company-year observations; G = number of clusters.

Period	Outcome	β	SE	t	p	Sig	n	G
Full Period (2012-2022)	GW incident count	-0.063	0.110	-0.570	0.584		1,090	9
	GW incident flag	0.023	0.022	1.047	0.326		1,090	9
Pre-Paris (2012-2016)	GW incident count	0.033	0.096	0.347	0.739		345	8
	GW incident flag	-0.012	0.036	-0.342	0.743		345	8
Post-Paris (2017-2022)	GW incident count	-0.127	0.166	-0.763	0.468		745	9
	GW incident flag	0.041	0.027	1.513	0.169		745	9

APPENDIX TABLE XVII

PLACEBO TEST RESULTS

Test 1 shuffles GS^{pct} values across companies within each year, preserving the annual cross-sectional distribution while breaking the company-specific link to greenwashing behavior. Test 2 shuffles GS^{sti} values within each industry-year cell, preserving the industry-year normalization structure while breaking the within-firm temporal association with incidents. The permutation p-value is the share of permutations where the absolute permuted statistic equals or exceeds the absolute observed statistic. Random seed = 42.

Metric	Test 1	Test 2
Observed statistic	r = 0.257	β = 0.060
Permuted mean	0.000	-0.001
Permuted SD	0.029	0.009
Permuted 95% CI	[-0.057, 0.057]	[-0.019, 0.016]
Z-score	8.87	6.87
Permutation p-value	< 0.001	< 0.001
Permutations yielding |stat| ≥ |observed|	0 / 5,000	0 / 5,000

DISCLOSURE AND DISCLAIMERS

AI Disclosure

This project was developed with strict adherence to the academic integrity policies and guidelines set forth by ISEG, Universidade de Lisboa. The work presented herein is the result of my own research, analysis, and writing, unless otherwise cited. In the interest of transparency, I provide the following disclosure regarding the use of artificial intelligence (AI) tools in the creation of this thesis/internship report/project:

2. I disclose that AI tools were employed during the development of this thesis as follows:

- AI-based research tools were used to assist in literature review and data collection.
- Claude was used for formulating/rephrasing
- Claude Code was used for the construction of the pipeline and finetuning code.
- Generative AI tools were consulted for brainstorming and outlining purposes. However, all final synthesis, and critical analysis are my own work. Instances where AI contributions were significant are clearly cited and acknowledged.

Nonetheless, I have ensured that the use of AI tools did not compromise the originality and integrity of my work. All sources of information, whether traditional or AI-assisted, have been appropriately cited in accordance with academic standards. The ethical use of AI in research and writing has been a guiding principle throughout the preparation of this thesis.

I understand the importance of maintaining academic integrity and take full responsibility for the content and originality of this work.

Anwuli Austin Oluh 22.07.2026