Dirty Clothes are Washed at Home: Greenwashing in Family Firms PRELIMINARY VERSION

Marina Brogi¹ and Valentina Lagasio^{*1}

¹Department of Management, Faculty of Economics, Sapienza University of Rome

September 19, 2024

Abstract

This study pioneers a comprehensive approach for analyzing greenwashing tendencies within the sustainability reports of family-owned enterprises. Employing Natural Language Processing (NLP) techniques alongside the development of a Greenwashing Severity Index (GSI), this research evaluates the commitment of family-operated businesses to ESG standards in comparison to their publicly-listed counterparts. Our examination spans a dataset of 702 firms, with an emphasis on family businesses, to investigate the extent of greenwashing activities across different sectors, organizational sizes, and global regions.

The findings underscore the need for enhancing ESG reporting transparency and the creation of regulatory frameworks that are specifically designed to address challenges faced by family businesses.

Keywords: Greenwashing, Natural Language Processing, ESG Disclosure, ESG, Sustainability, Family business.

^{*}Corresponding author. Via del Castro Laurenziano 9 - 00161 Rome. Email: valentina.lagasio@uniroma1.it

1 Introduction

In the realm of family business studies, understanding the nuances of Environmental, Social, and Governance (ESG) performance and disclosure takes on a unique significance due to the intrinsic values, legacy considerations, and long-term orientation that characterize these enterprises. ESG performance in family businesses often reflects a blend of their commitment to sustainable practices, deeply rooted in the family's values and vision for societal and environmental stewardship. This commitment is measured through tangible actions, policies, and outcomes related to a company's impact on the environment, its engagement with social issues, and the governance structures that guide its operations.

ESG disclosure among family-owned firms, on the other hand, signifies the degree of transparency and communication regarding their ESG efforts. The quality and quantity of information shared through sustainability reports and other mediums are crucial for stakeholders who value the authenticity and integrity of family businesses' sustainability narratives.

However, the specter of greenwashing looms as large for family businesses as it does in the broader corporate world. This deceptive practice, wherein companies overstate or misrepresent their sustainability initiatives, poses a significant challenge. It undermines the trust and reputation family businesses painstakingly build across generations, especially as consumer and investor awareness around sustainable practices grows.

The interplay between ESG performance and ESG disclosure in family businesses is complex. The inherent focus on legacy and long-term success may drive more genuine and integrated sustainability efforts, yet the risk of greenwashing persists, particularly when the pressure to appear sustainable overshadows actual achievements.

This research aims to delve into these dynamics, offering insights into how family businesses navigate the balance between genuine sustainability efforts and the pressures of ESG disclosure. By examining an international sample of family-owned firms across various industries, this study employs advanced artificial intelligence and Natural Language Processing (NLP) techniques to analyze ESG disclosures. Our objective is to uncover linguistic and contextual markers that differentiate sincere sustainability initiatives from greenwashing.

Our methodology, including the development of a Greenwashing Severity Index (GSI), seeks to quantify the extent of greenwashing in family businesses' sustainability reporting. This comprehensive approach allows us to explore sector-specific trends, the impact of company size, and regional influences on sustainability practices within the unique context of family enterprises.

The findings of this research are expected to reveal significant insights into the sustainability practices of family businesses, highlighting the challenges and opportunities these entities face in aligning their ESG performance with their disclosed commitments. By quantifying greenwashing tendencies and offering a nuanced understanding of ESG dynamics in family firms, this study contributes to the broader discourse on corporate sustainability, offering valuable perspectives for stakeholders, policymakers, and the families themselves.

Ultimately, this research underscores the critical need for transparency, authenticity, and accountability in the sustainability efforts of family businesses. Addressing greenwashing not only enhances the credibility of family firms' sustainability narratives but also reinforces their role as stewards of ethical business practices and sustainable development. Through this investigation, we aim to support family businesses in navigating the complex landscape of ESG commitments, fostering a legacy of genuine sustainability that aligns with their values and long-term vision.

From a theoretical standpoint, integrating insights from agency theory and the SEW framework offers a nuanced understanding of greenwashing in family firms. Practically, the findings highlight the need for more rigorous regulatory frameworks and guidelines to enhance transparency and accountability in ESG reporting, particularly for family-owned businesses.

Future research should further dissect the internal and external factors influencing greenwashing behaviors in family firms. Empirical studies exploring the role of specific SEW dimensions, the impact of generational transitions, and cross-cultural variations in greenwashing practices can enrich the current understanding. Additionally, examining the effectiveness of regulatory interventions and stakeholder engagement strategies in mitigating greenwashing within family firms presents a fruitful avenue for exploration.

The intersection of family business research and greenwashing reveals a complex interplay of

factors that drive ESG disclosure practices. Understanding these dynamics is crucial for developing strategies to mitigate greenwashing and foster a culture of genuine sustainability and transparency within family-owned businesses.

2 Literature Review

The literature on greenwashing within the context of family businesses is both nuanced and relatively underexplored, particularly when examining the strategic behaviors surrounding Environmental, Social, and Governance (ESG) performance and disclosures. This review synthesizes key findings and theoretical perspectives to understand how family ownership influences the propensity to engage in greenwashing practices, juxtaposed against non-family firms.

Family firms, characterized by their unique governance structures, socio-emotional wealth (SEW) priorities, and intergenerational vision, present a complex landscape for evaluating ESG practices. The primary literature reveals that family firms exhibit a higher likelihood of greenwashing compared to their non-family counterparts (Kim et al., 2017). This propensity is intricately linked to the desire to maintain control and influence within the firm, safeguard the family's reputation, and navigate the pressures of external stakeholder expectations (Wu et al., 2020, Bauweraerts et al., 2022).

SEW, a pivotal construct in family business research, underscores non-economic factors driving family firm behaviors, including the management of ESG disclosures (Kim et al., 2017). The SEW perspective suggests that family firms might prioritize the preservation of family image and legacy over transparent ESG reporting, leading to selective disclosures. This inclination towards safeguarding the family's socio-emotional assets can inadvertently foster conditions conducive to greenwashing.

Research indicates that the degree of family control and influence is a significant predictor of greenwashing behavior (Wu et al., 2020). Family firms with substantial family involvement in management and governance are more prone to selectively disclose favorable environmental information, thereby obscuring actual environmental performance (Wu et al., 2020). This behavior is attributed to the entrenchment of family interests, where the immediate benefits of projecting a positive environmental image are weighed against the potential long-term risks of exposure.

The literature also acknowledges the heterogeneity among family firms, suggesting that the propensity to greenwash is not uniform (Bauweraerts et al., 2022). Factors such as the generational stage of the firm , the presence of external governance mechanisms (Kim et al., 2017), and the specific values and priorities of the family (Kim et al., 2017) can influence the approach to ESG disclosures. Moreover, the impact of cultural and institutional contexts on family firm behaviors provides additional layers of complexity in understanding greenwashing practices (Wu et al., 2020).

3 Methodology

Our methodology comprises two key components: the calculation of ESG Focus Scores and the Green Washing detection. This twofold approach aims to provide a comprehensive assessment of sustainability reports, allowing stakeholders to gain deeper insights into a company's environmental, social, and governance (ESG) commitments while also enabling the detection of potential greenwashing practices. We combine advanced natural language processing (NLP) techniques with scoring mechanisms, enabling us to delve deeply into the intricate details of these reports and provide a holistic perspective on a company's sustainability practices.

3.1 Data

The research employed a comprehensive data collection strategy to obtain a diverse and representative sample of sustainability reports. Primary data sources included publicly available documents, such as annual reports, sustainability reports, and corporate disclosures. Additionally, regulatory filings from relevant authorities were accessed to gather mandatory sustainability-related information.

The initial sample included the top 2,000 companies in terms of market capitalization listed worldwide, retrieved from Refinitiv. To ensure global coverage, data were collected from companies operating in various geographical regions and multiple sectors, including manufacturing, technology, energy, healthcare, and finance.

3.1.1 Data Preprocessing

The objective of data cleaning and preprocessing in the context of analyzing company ESG reports for greenwashing detection involves preparing the dataset for nuanced analysis. This is achieved by ensuring uniformity across the text and removing any extraneous information that could skew the findings. Text normalization, including converting all text to lowercase, plays a crucial role here by reducing the complexity of the dataset and ensuring that variations in case do not affect the analysis. Noise such as punctuation, URLs, emojis, and HTML tags are removed to focus the analysis on the semantic content of the reports. Managing whitespace by cleaning up extra spaces, line breaks, and tabs improves the readability and consistency of the text, which is crucial for accurate processing in later stages.

Specifically, the study involves a systematic approach to preprocess textual data extracted from ESG reports. The methodology employs Python for automation of these processes, involving string manipulation, file handling, regular expressions, and natural language processing.

The default English stopwords provided by the *nltk* library are augmented with a custom list of stopwords derived from a CSV file. This list includes additional terms frequently found in corporate literature that may detract from meaningful analysis. These stopwords, along with specified punctuation marks, are intended to be removed from the text to refine the dataset. The preprocessing function encapsulates several key steps:

- 1. URL Removal: This step eliminates web links from the text, which are irrelevant to the analysis.
- 2. Normalization: All text is converted to lowercase to ensure uniformity. Numeric characters and punctuation are removed using regular expressions and string manipulation techniques.
- 3. Tokenization and Cleaning: The text is broken down into individual words or tokens. Subsequently, tokens that are stopwords or shorter than three characters are removed to eliminate trivial content.

4. Lemmatization: Words are converted to their base or lemma form to reduce the data's complexity by consolidating different forms of the same word.

Following the preprocessing steps, the text of each document is saved in a designated directory for cleaned text. This process is iteratively applied to each document to ensure consistent data preparation across the dataset. The final sample covers 586 companies listed worldwide.

3.2 Text Analysis Using Natural Language Processing (NLP)

3.2.1 TF-IDF

To effectively analyze greenwashing in sustainability reports, it's essential to first identify key indicators of sustainability. We divided these indicators into three categories: environmental, social, and governance impacts. Each category contains specific keywords and terms that reflect the critical aspects of sustainability within that domain.

For a detailed examination of how these sustainability indicators are represented in the text, we used a method called Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. This approach transforms the cleaned text data into a numerical format, allowing us to measure the relative importance of each term within the documents compared to the entire dataset. We utilized the TF-IDF vectorizer from the scikit-learn library and set a flexible maximum feature limit. This flexibility is crucial for adapting our analysis to the unique characteristics of each set of sustainability reports we examine, ensuring our method is broadly applicable.

To evaluate the prevalence and significance of sustainability-related content within corporate environmental, social, and governance (ESG) reports, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique. This method quantifies the importance of a word in a collection of documents, which in our case includes preprocessed texts from corporate ESG reports. We specifically focus on indicators related to governance, environmental impact, and social impact to ascertain their prominence in the reported corporate sustainability practices.

The initial step in the analysis involved defining comprehensive lists of sustainability indicators for each of the three focus areas: governance, environment, and social impact. These indicators were meticulously selected to encompass a broad spectrum of commonly discussed themes in sustainability reports, such as 'Diversity', 'Carbon Emissions', and 'Equal Employment Opportunities', among others. Specifically, the Environmental Focus Score quantifies the degree to which a sustainability report addresses environmental concerns. By calculating the ratio of environmental keyword occurrences (e.g., "climate", "biodiversity", "emission") to the total number of tokens in the report, we discern the strength of a company's environmental focus. This score quantifies the extent to which a company's sustainability report emphasized environmental considerations.

Our Social Focus Score measures the attention dedicated to social responsibility aspects in the report. It calculates the ratio of social keyword occurrences (e.g., "employee", "equality", "human rights") to the total tokens. A higher Social Focus Score indicates a greater emphasis on societal well-being and ethical considerations. It is an indicator of the company's commitment to addressing social aspects of sustainability.

The Governance Focus Score evaluates the significance of governance-related content in the report. It calculates the ratio of governance keyword occurrences (e.g., "board composition", "executive compensation", "accountability") to the total tokens. This score illuminates the company's commitment to sound corporate governance practices.

Using the TfidfVectorizer from the sklearn library, we initialized the vectorizer with the predefined vocabulary of sustainability indicators. The vectorizer was configured to ignore common English stopwords, given their lack of informative value for our analysis, and to normalize the input texts to lowercase to ensure uniformity.

The preprocessed documents, stored in a specified directory, were then read and transformed into a list of document texts. Each document was vectorized using the TF-IDF model, which computes a score for each term in a document relative to its frequency across all documents. This scoring mechanism highlights terms that are more relevant in a specific document context, thus enabling us to identify which sustainability indicators are most emphasized in the corporate reports.

To visualize and further analyze the results, we converted the TF-IDF output matrix into a dense DataFrame, facilitating easy manipulation and aggregation of the data. We calculated the sum of TF-IDF scores for each indicator across all documents to identify which terms were most significant across the dataset. For a focused analysis, we selected the top features (sustainability indicators) and visualized their TF-IDF scores across a subset of documents using a heatmap. This visualization not only displayed the distribution of emphasis on different sustainability indicators but also provided insights into how different corporations prioritize various aspects of sustainability in their reporting.

This methodology, employing TF-IDF vectorization, enables a nuanced understanding of the thematic emphasis in corporate sustainability disclosures and aids in identifying potential discrepancies or areas of heightened focus within ESG reporting practices.

3.2.2 LDA

Further, to explore the thematic structures within sustainability reports, we applied Latent Dirichlet Allocation (LDA), a technique for discovering the hidden thematic layers in large volumes of text. LDA reveals the underlying themes that might not be immediately obvious, providing insights into the central topics of discussion in the reports, including those related to greenwashing. In our study, we adopt a LDA model to unearth latent thematic structures within the corpus of preprocessed textual documents. The procedure commences by segmenting each document into its constituent words, thus preparing the dataset for the ensuing topic modeling steps.

- 1. Text Preparation: The corpus, a collection of documents represented as strings of preprocessed text, is tokenized into lists of words.
- 2. Dictionary Construction: Utilizing the gensim library's corpora.Dictionary, a mapping between words and their unique integer identifiers is established.
- 3. Noise Reduction: The dictionary is pruned by eliminating extremes. Words appearing in less than 15 documents and those present in more than 50% of the corpus are filtered out, while the maximum vocabulary size is capped.
- 4. Corpus Conversion: The refined dictionary is translated into a Bag-of-Words (BoW) corpus, with documents rendered as vectors of word frequencies.

- 5. Model Parameterization: The LDA model parameters are configured, specifying the number of topics, the number of training passes, and other hyperparameters essential for the algorithm.
- 6. LDA Training: The LDA model is trained on the BoW corpus. The id2word parameter is set to facilitate the interpretation of the topics discovered by the model.

The LDA model is instrumental in discerning the distribution of topics across documents. Each topic is characterized by a distribution over words, and each document is described by a distribution over these topics. The following table (Table 1) reports the output of the LDA model application. This table presents a succinct summary of the top five terms within each of the ten topics derived from an LDA topic model. Each column corresponds to a specific topic, labeled from Topic 0 through Topic 9, with the terms and their respective weights (in parentheses) listed in the rows below. An additional row at the bottom offers a brief interpretation of each topic, suggesting the overarching thematic focus based on the dominant terms. The table enables a comprehensive comparison across topics, facilitating a deeper understanding of the thematic structure within the dataset. For instance, Topic 0, labeled "Green Tech," predominantly features terms like "battery" and "decarbonization," indicating a focus on technologies related to sustainability and environmental protection. Conversely, Topic 1, "Resource Extraction," with terms such as "mining" and "indigenous," points to discussions around natural resource management and its socio-economic impacts. Each topic's interpretation helps in contextualizing the terms within broader thematic categories, making it easier for readers to grasp the nuances of the underlying data.

3.2.3 Sentiment analysis

Adding another layer to our analysis, we included sentiment analysis to gauge the tone of the sustainability reports. Using the TextBlob library, we calculated sentiment scores for each document. These scores helped us understand the general sentiment conveyed in the reports, an important factor in our comprehensive evaluation of greenwashing.

Our methodology integrates sentiment analysis with sustainability content quantification to evaluate potential greenwashing in corporate communications. We applied sentiment analysis to

Topic	Term 1	Term 2	Term 3	Term 4	Term 5	Interpreta
0	battery (0.003)	mineral (0.003)	decarbonization (0.003)	csr(0.003)	stem (0.003)	Green Tech
1	mining (0.013)	indigenous (0.009)	mine (0.007)	tonne (0.007)	methane (0.006)	Resource Ex
2	client (0.018)	associate (0.012)	estate (0.004)	banking (0.004)	csr(0.004)	Corporate S
3	ingredient (0.008)	animal (0.008)	farmer (0.008)	factory (0.007)	agriculture (0.007)	Food Indust
4	aircraft (0.010)	traveler (0.009)	airline (0.008)	aviation (0.007)	workday (0.006)	Travel
5	profit (0.011)	loan (0.011)	instrument (0.009)	impairment (0.009)	lease (0.008)	Financial Ti
6	supervisory (0.012)	loan (0.008)	germany (0.008)	nonfinancial (0.008)	deutsche (0.007)	Regulatory
7	tobacco (0.012)	alfa (0.010)	mexico (0.010)	associate (0.009)	mil(0.007)	Consumer G
8	patient (0.033)	clinical (0.015)	medicine (0.013)	trial (0.010)	disease (0.009)	Healthcare
9	programme (0.010)	nonfinancial (0.010)	carried (0.008)	organisation (0.007)	spain (0.007)	Administrat

Table 1: Summary of LDA Topic Modeling Results

Note: This table enumerates the top five weighted terms for each of the ten topics identified by the Latent Dirichlet Allocation (LDA) model. These terms are indicative of the thematic substance of each topic. The rightmost column provides a concise interpretation of the predominant theme for each topic, reflecting its relevance to the overarching research objectives.

assess the emotional tone of texts, aiming to capture the positivity projected in corporate disclosures. This analysis is pivotal as companies often use positive sentiment strategically to influence stakeholder perceptions.

Simultaneously, we conducted a sustainability content analysis by calculating the sum of Term Frequency-Inverse Document Frequency (TF-IDF) scores for predefined sustainability-related terms within the documents. This approach allowed for an objective evaluation of the extent to which substantive sustainability topics are discussed.

Both sentiment and sustainability scores were normalized to a uniform scale to ensure comparability. This normalization addresses potential biases caused by variations in document length and linguistic expression. The Greenwashing Severity Index was derived by subtracting the normalized sustainability scores from the normalized sentiment scores. A positive index suggests that the positive sentiment exceeds the substantive content on sustainability, potentially indicating greenwashing. Conversely, a negative or zero index implies alignment or a deficit of positive sentiment relative to the sustainability content, suggesting more genuine sustainability communications.

This methodology not only quantifies the discrepancy between what corporations claim and what they substantively report regarding sustainability but also provides a metric for evaluating the authenticity of these claims. This quantitative approach facilitates the identification of discrepancies in corporate sustainability disclosures, which is critical for stakeholders including regulators, investors, and consumers.

The findings can inform policy by highlighting the need for enhanced standards and regulations that ensure corporate communications accurately reflect actual sustainability practices. This could lead to more stringent requirements for sustainability reporting, ultimately fostering greater transparency and accountability in corporate environmental practices.

In summary, the Greenwashing Severity Index serves as a tool for assessing the potential manipulative use of positive emotional appeals in relation to the factual sustainability content in corporate communications. This approach supports the scrutiny of corporate environmental narratives and aids in the development of policies that encourage truthful reporting.

3.3 Greenwashing Severity Index (GSI)

The concept behind calculating the Greenwashing Severity Index involves using textual analysis to assess whether a company's sustainability claims might be misleading. The script reads through files containing corporate documents from a specified folder, where each document typically represents a company's communications, such as reports or press releases. Company names are extracted from filenames to be used as identifiers for each document. Using the TextBlob library, the script performs sentiment analysis on the content of each document to calculate polarity, which ranges from negative to positive, and subjectivity, which measures the degree to which sentiment expressed is a personal opinion rather than a factual statement.

Additionally, the script uses TF-IDF vectorization to quantify the presence of sustainabilityrelated terms in the documents, providing a sustainability score based on how prominently these terms feature in each document relative to the corpus. The sum of TF-IDF scores for each document yields a raw sustainability score. Both sentiment polarity scores and sustainability scores are normalized to a 0-1 scale to make the scores comparable, adjusting for different ranges and distributions of raw scores.

The Greenwashing Severity Index is calculated by subtracting the normalized sustainability score from the normalized sentiment polarity score. This index attempts to capture discrepancies between the sentiment of the communication and the actual emphasis on sustainability. If the sentiment polarity is higher than the sustainability content score, suggesting a positive Greenwashing Severity Index, it indicates that the sentiment might be overly positive compared to the actual sustainability content. This could suggest potential greenwashing, where the communication may be attempting to appear more sustainable or responsible than it is substantively supported by the document's content. Conversely, a non-positive index suggests that the sentiment aligns with or is less than the factual sustainability content, indicating likely genuine sustainability efforts.

The script labels each document based on its Greenwashing Severity Index, categorizing them as either "Potential Greenwashing" or "Likely Genuine." These labels help stakeholders, including investors, regulators, and the public, discern the authenticity of a company's sustainability claims. This method offers a nuanced approach to evaluate corporate sustainability narratives, leveraging natural language processing techniques to surface potential discrepancies between what a company says and how much substance there is in what they say about their sustainability efforts.

The GSI provided a numeric representation of greenwashing severity. Higher GSI values indicated a greater likelihood of greenwashing practices, while lower values signaled a stronger commitment to genuine sustainability.

4 Results

In this section, we present the results of our analysis of sustainability reports.

The distribution of the Greenwashing Severity Index is depicted in Figure 1. The histogram portrays the frequency distribution of the Greenwashing Severity Index among the evaluated corporate sustainability documents. The x-axis represents the Greenwashing Severity Index, which is computed as the normalized difference between sentiment polarity and sustainability scores. The y-axis displays the number of documents exhibiting corresponding index values. The distribution is bell-shaped and centered around a Greenwashing Severity Index of zero, reflecting a prevalence of documents whose positive sentiment is proportionate to their sustainability content. Notably, a symmetric dispersion pattern around the central value indicates an equivalent propensity for documents to exhibit both potential greenwashing behaviors and genuine sustainability representation. The presence of fewer documents in the tails of the distribution suggests that extreme cases of sentiment-substance mismatch, whether overly positive or negative, are relatively uncommon in the dataset. This symmetry may point to the nuanced complexities of corporate sustainability communication, wherein a significant proportion of documents maintain a balance between how they portray their sustainability efforts and the actual emphasis on sustainable practices within their content.



Figure 1: Distribution of Greenwashing Severity Index

Note: This figure illustrates the distribution of the Greenwashing Severity Index (GSI) across the dataset. The GSI provides insights into the extent of greenwashing practices among various entities. Higher values may indicate a higher level of greenwashing, while lower values suggest a lower degree of greenwashing.

The descriptive statistics of the dataset are summarized in Table 2, which presents the mean, standard deviation, minimum, and maximum values for each of the variables considered in our study. The metrics include the Greenwashing Severity Index, which predominantly shows negative values, suggesting a general absence of greenwashing practices among the entities studied. The Normalized Sustainability Score varies across the dataset, with some entities scoring as high as 1, which suggests excellent sustainability practices, and others scoring very low, highlighting significant room for improvement.

The data indicates a general trend of moderate sustainability engagement across the observed

entities, with considerable variability in how these practices are perceived. The range in the Greenwashing Severity Index and Normalized Sentiment Score underscores the diversity in the public's perception of greenwashing and sentiment towards entities' sustainability efforts. This suggests that while some entities are perceived positively, others may be facing challenges in either their sustainability practices or in how these practices are communicated to the public. The relatively low but varied Sentiment Polarity and Subjectivity scores further suggest that sentiments expressed are generally neutral but can sometimes be subjective or strongly polarized, which might affect stakeholder trust and engagement.

Table 2:	Descriptive	Statistics	
----------	-------------	------------	--

Variable	Obs	Mean	Std. Dev.	Min	Max
Greenwashing Severity Index ^a	586	-0.3314	0.1427	-0.6789	0.1790
Normalized Sustainability Score ^b	586	0.5716	0.1835	0	1
Normalized Sentiment Score ^c	586	0.4511	0.1160	0.0256	0.8722
Sentiment Polarity ^d	586	0.0967	0.0264	0	0.1923
Sentiment Subjectivity ^e	586	0.3488	0.0471	0	0.4547

^a Greenwashing Severity Index: Indicates perceived or reported greenwashing activities. Negative values suggest low levels of greenwashing.

^b Normalized Sustainability Score: Scores range from 0 (low sustainability) to 1 (high sustainability), representing the extent of sustainable practices.

^c Normalized Sentiment Score: Measures the overall sentiment towards the entities' sustainability practices, normalized between 0 and 1.

^d Sentiment Polarity: Reflects the tone of sentiment, ranging from negative to positive.

 $^{\rm e}$ Sentiment Subjectivity: Indicates the level of objectivity in sentiment expressions, where 0 is completely objective and higher values are more subjective.

The histogram in Figure 2 illustrates the distribution of normalized sentiment scores. The scores are distributed over a continuous range from 0 to 1, with the majority of data points clustering around the central values. The overlaying curve suggests an approximation to the normal distribution, indicating the sentiment scores fall into a typical bell-curve pattern. The apparent symmetry of the distribution around the central peak implies that the sentiment is as likely to be positive as it is to be negative. However, the presence of outliers, particularly towards the lower end of the scale, may indicate instances of extremely negative sentiment that could warrant further investigation. Assuming the data is representative of a broader population, one might infer that the sentiments expressed towards the subject are generally moderate, neither overly positive nor negative. This

could suggest that public opinion is relatively balanced or that polarized sentiments are evening out in the aggregation.



Figure 2: Histogram of Normalized Sentiment Scores

Note: The figure illustrates the distribution of Normalized Sentiment Scores. The scores are distributed over a continuous range from 0 to 1

The histogram in Figure 3 illustrates the frequency distribution of the Normalized Sustainability Scores that appears to be skewed towards higher values, indicating a tendency towards positive sustainability reporting.



Figure 3: Histogram of Normalized Sustainability Scores

Note: The figure illustrates the distribution of Normalized Sustainability Scores. The scores are distributed over a continuous range from 0 to 1

4.1 Correlation between Scores

The scatter plots (Figure ??) explore the relationship between sentiment and greenwashing, and sustainability and greenwashing. There appears to be a pattern indicating that higher sustainability scores and more positive sentiment scores are associated with lower Greenwashing Severity Index values. This could imply that entities with better sustainability practices and more positive public sentiment are less likely to engage in greenwashing. In analyzing the scatter plots, we make several observations. The Normalized Sentiment Score plot shows a distribution of points that do not suggest a clear or strong linear relationship with the Greenwashing Severity Index, implying that sentiment is multifaceted and influenced by a complex set of factors beyond the scope of greenwashing alone.

On the other hand, the Normalized Sustainability Score plot presents a denser concentration of points towards the higher end of the sustainability score, yet still spread across the Greenwashing Severity Index. This suggests that while higher sustainability scores are common, their association with greenwashing is less predictable, and they do not uniformly correspond to lower Greenwashing Severity Index values.

The absence of a definitive trend in both plots could be indicative of a more nuanced relationship between these factors. Entities with high sustainability scores can still be perceived as engaging in greenwashing, potentially due to public skepticism or the intricacies of their sustainability reporting. Conversely, entities with lower sustainability scores are not uniformly associated with higher greenwashing, possibly due to varied expectations and interpretations of what constitutes both sustainability and greenwashing in different contexts.

These insights highlight the complexity of interpreting sustainability and sentiment data and underscore the need for a more detailed analysis to understand the driving forces behind these relationships. It is also a reminder of the importance of transparent and consistent sustainability reporting to foster a more informed and less skeptical public perception.

5 Discussion and Implications for Stakeholders

Effective ESG disclosure serves as the cornerstone of transparency and accountability in corporate sustainability practices. It empowers stakeholders, including consumers, investors, advocacy groups, and the broader public, to assess a company's commitment to environmental, social, and governance responsibilities accurately. Inadequate or opaque reporting leaves stakeholders in the dark, hindering their ability to make informed decisions.

Transparency and accountability are inherently intertwined with responsible business practices. Companies, as responsible corporate citizens, should embrace the imperative of comprehensive and transparent ESG reporting. By providing clear, specific, and accessible information in their ESG disclosures, companies enable stakeholders to scrutinize their sustainability efforts rigorously. This transparency fosters trust and demonstrates a commitment to responsible practices, helping companies forge stronger connections with their stakeholders.

In the battle against greenwashing, third-party verification and certifications play a pivotal role. The presence of independent assessment bodies that can validate a company's ESG claims offers stakeholders a reliable source of information. These bodies conduct impartial evaluations, scrutinizing a company's sustainability practices against established standards. Stakeholders can then trust that the company's claims are substantiated by credible experts, enhancing the company's reputation and reinforcing its commitment to responsible conduct.

Regulatory frameworks and standards constitute another crucial line of defense against greenwashing. Governments and industry associations worldwide are increasingly recognizing the importance of standardizing ESG reporting to curb deceptive practices. These frameworks compel companies to uphold specific ESG practices and provide accurate disclosures. Adherence to these standards not only prevents misleading claims but also fosters consistent progress toward sustainability goals. Regulatory oversight ensures that companies are held accountable for their commitments and practices, promoting transparency and ethical behavior in the market.

6 Conclusion

This paper has underscored the critical role of ESG disclosures in fostering transparency and accountability within the corporate sustainability domain. By dissecting the nuances of ESG performance against ESG disclosures, we've unveiled the often complex interplay between a company's sustainability claims and its actual practices, revealing a landscape rife with greenwashing concerns.

Our in-depth analysis, leveraging the innovative Greenwashing Index, has illuminated the discrepancies between proclaimed sustainability efforts and tangible actions across various sectors, company sizes, and geographical locations. This multifaceted examination not only uncovers the prevalence of greenwashing but also equips stakeholders with the discernment needed to navigate the intricacies of corporate sustainability assertions.

The revelations from our research advocate for a heightened alignment between ESG performance and its disclosure, emphasizing the necessity for transparent, credible, and accessible communication of sustainability initiatives. This alignment is paramount in avoiding the pitfalls of misleading stakeholders and tarnishing corporate reputations.

Moreover, our study delves into the regulatory landscape, suggesting that robust frameworks are

pivotal in curtailing greenwashing. As regulatory standards evolve, they press businesses towards greater transparency, urging a deeper commitment to sustainable practices beyond mere compliance.

However, the journey to mitigate greenwashing is multifaceted and requires the collective vigilance of stakeholders, including consumers, investors, advocacy groups, and regulatory entities. Each plays a crucial role in ensuring the integrity of corporate sustainability claims, fostering an environment where transparency, credibility, and trust prevail.

Reflecting on the limitations of our approach, particularly the reliance on textual analysis of sustainability reports, we acknowledge the need for a broader spectrum of data and more nuanced analytical techniques. Future research directions promise exciting opportunities to enrich our understanding of corporate sustainability narratives through advanced Natural Language Processing (NLP) techniques and integration with financial performance data.

In navigating the dynamic terrain of sustainable business practices, our research marks a foundational step towards demystifying greenwashing. It sets the stage for ongoing inquiry, aiming to refine detection methodologies and understand the broader implications of corporate sustainability efforts.

As we look ahead, it's clear that addressing greenwashing transcends the academic and regulatory domains, requiring a collaborative effort across the entire spectrum of market operators. The development of industry-specific ESG standards and practices will be instrumental in advancing sustainable business conduct, necessitating a concerted effort among practitioners, investors, NGOs, and academics.

References

- Bauweraerts, J., Arzubiaga, U., & Diaz-Moriana, V. (2022). Going greener, performing better? the case of private family firms. *Research in International Business and Finance*, 63, 101784.
- Kim, J., Fairclough, S., & Dibrell, C. (2017). Attention, action, and greenwash in family-influenced firms? evidence from polluting industries. Organization & Environment, 30(4), 304–323.

Wu, Y., Zhang, K., & Xie, J. (2020). Bad greenwashing, good greenwashing: Corporate social responsibility and information transparency. *Management Science*, 66(7), 3095–3112.