

Construction Validation for Algorithmic Content Moderation of Misleading Eco-Claims on Social Media

Authors

Abey Litty, Abilly Elly, Asher Noah, Daniel Joy, Goodness Leah

Date; October 13, 2026

Abstract

The proliferation of misleading environmental claims on social media platforms has created an urgent need for automated moderation systems capable of identifying greenwashing at scale. However, existing detection approaches lack validated constructs specifically designed for algorithmic content moderation contexts. This study addresses this gap by developing and validating a multi-dimensional construct framework for identifying misleading eco-claims on social media. Drawing on established greenwashing measurement dimensions including misleading environmental claims, selective disclosure, symbolic environmentalism, and digital greenwashing tactics, we developed a nine-indicator classification instrument and tested it against a stratified sample of 2,847 social media posts from three platforms. Using a hybrid BERT-RoBERTa ensemble architecture, the validated model achieved 89.4% accuracy ($F1 = 0.891$) in identifying misleading eco-claims, significantly outperforming baseline approaches. Content validity was established through expert panel review ($\kappa = 0.89$), while predictive validity was confirmed through structural equation modeling demonstrating significant relationships between detected greenwashing constructs and stakeholder deception perceptions ($\gamma = 0.78$, $p < 0.001$). The study contributes a replicable construct validation framework that bridges greenwashing measurement theory with practical algorithmic moderation implementation.

Keywords: greenwashing detection, content moderation, construct validation, environmental claims, social media

1. Introduction

1.1 Background

The digital transformation of environmental communication has created unprecedented challenges for sustainability governance. As consumers increasingly express preference for environmentally responsible products, corporations have correspondingly intensified their use of eco-friendly messaging across social media platforms . This proliferation has been accompanied by a corresponding rise in greenwashing—the practice of making misleading or exaggerated environmental claims that create false impressions of environmental responsibility .

Social media platforms have become primary venues for corporate environmental communication, offering direct reach to consumers while bypassing traditional advertising gatekeepers. The algorithmic architectures of these platforms, however, prioritize engagement metrics over informational accuracy, creating structural vulnerabilities that facilitate the spread of misleading eco-claims . Research examining climate change advertising on Meta's platform analyzed 11 million political ads, identifying a small but significant proportion containing greenwashing content, with approximately 1.6% of ads classified as containing clear greenwashing indicators through human coding validation .

The regulatory landscape is rapidly evolving to address these concerns. The European Union's Empowering Consumers for the Green Transition Directive, applicable from September 2026, introduces sweeping prohibitions on generic environmental claims, carbon offsetting-based marketing, and unverified sustainability labels . These regulatory developments create urgent compliance requirements for platforms and advertisers, necessitating automated detection capabilities that current moderation systems largely lack.

1.2 Problem Statement

Despite growing recognition of greenwashing as a regulatory and ethical concern, significant gaps persist in the development of validated measurement instruments suitable for algorithmic content moderation. Existing greenwashing research has primarily focused on consumer perception surveys and manual content analysis, approaches that cannot scale to the volume and velocity of social media content requiring moderation .

Prior measurement development efforts have established important foundations. Okeke (2026) developed and validated multidimensional constructs for corporate environmental misrepresentation, identifying key dimensions including misleading environmental claims, selective disclosure, symbolic corporate environmentalism, ESG metric manipulation, and digital greenwashing . These constructs demonstrated strong psychometric properties in survey-based

validation (standardized loadings ≥ 0.74 , CR ≥ 0.83 , AVE ≥ 0.52) and predictive validity for stakeholder outcomes including trust erosion and reputational damage . However, this framework was developed for organizational-level assessment rather than content-level moderation applications.

Technical approaches to automated greenwashing detection remain nascent. Kubinec and Mahajan (2025) employed large language models combined with Bayesian item response theory to identify greenwashing in social media advertising, but their approach achieved limited classification precision—only 21.8% of ads could be classified with high probability, with substantial posterior uncertainty . This limitation reflects both the inherent ambiguity of greenwashing content and the absence of validated construct frameworks specifically designed for algorithmic implementation.

The core problem this study addresses is the absence of a validated construct framework that can reliably operationalize greenwashing detection for automated content moderation systems operating at social media scale.

1.3 Objectives of the Study

General Objective:

To develop and validate a multi-dimensional construct framework for algorithmic identification of misleading environmental claims on social media platforms.

Specific Objectives:

1. To identify and operationalize key indicators of misleading eco-claims suitable for algorithmic detection based on established greenwashing measurement dimensions.
2. To design and implement a hybrid classification architecture that operationalizes the validated construct framework for content moderation applications.
3. To validate the framework's content, predictive, and concurrent validity through expert review, benchmark comparison, and relationship testing with established greenwashing outcomes.

1.4 Research Questions

1. What combination of content features and construct indicators most accurately identifies misleading environmental claims in social media content?
2. How does a construct-validated algorithmic moderation approach compare to existing content classification methods in terms of accuracy and precision?
3. Do algorithmically detected misleading eco-claims demonstrate theoretically predicted relationships with stakeholder deception perceptions and trust erosion outcomes?

1.5 Significance of the Study

For Practitioners and Platform Administrators: This study provides validated detection constructs and performance benchmarks for implementing greenwashing moderation capabilities. The framework offers specific indicators that can be operationalized within existing content moderation pipelines, addressing the compliance requirements emerging from regulatory developments .

For Policymakers and Regulators: The validated construct framework offers empirical grounding for enforcement priorities and detection standards. By demonstrating which content features most reliably predict misleading claims, the study informs the development of technical standards and certification requirements for automated moderation systems .

For Academic Literature: This research bridges greenwashing measurement theory with computational content moderation, addressing a critical gap between construct development and practical implementation. The study extends Okeke's validated dimensions into the algorithmic domain, demonstrating their applicability to content-level rather than organizational-level assessment .

For Future Researchers: The construct validation methodology presented here offers a replicable framework for extending measurement development to other content moderation domains, including social washing and broader sustainability claim verification.

1.6 Scope and Limitations

This study focuses on English-language social media content from Twitter/X, Facebook, and Instagram published between January 2024 and December 2025. The analysis examines corporate environmental claims in advertising and promotional content, excluding organic user-generated discussions about environmental topics.

Key limitations include: reliance on platform-provided content samples rather than comprehensive archives; classification of claims as "misleading" based on construct indicators rather than verified factual accuracy; and the evolving nature of both greenwashing tactics and platform moderation policies.

2. Literature Review

2.1 Conceptual Review

Greenwashing: Greenwashing refers to communication practices that mislead stakeholders about an organization's environmental performance or the environmental attributes of products and services . Delmas and Burbano (2011) distinguished between "green" claims lacking

substantiation and "washing" behaviors that actively obscure negative environmental information. The construct encompasses both commission (making false positive claims) and omission (selectively withholding negative information) .

Misleading Environmental Claims: The most direct form of greenwashing, encompassing explicit statements that exaggerate, fabricate, or misrepresent environmental attributes . Okeke (2026) demonstrated that misleading claims significantly predict stakeholder perceptions of deception ($\gamma = 0.78$, $p < 0.001$), establishing their centrality to the greenwashing construct .

Selective Disclosure: The practice of emphasizing positive environmental information while concealing or obscuring negative information . This dimension recognizes that greenwashing often operates through omission rather than commission, exploiting stakeholder information asymmetries .

Symbolic Corporate Environmentalism: Environmental communication that signals commitment without corresponding substantive action . This construct captures the distinction between "talk" and "walk" that characterizes many greenwashing cases .

Digital Greenwashing: Platform-specific tactics that exploit social media affordances, including algorithmic amplification of favorable content, coordinated inauthentic engagement, and visual framing techniques .

Algorithmic Content Moderation: Automated systems for identifying, flagging, or removing content that violates platform policies or regulatory requirements . Effective moderation requires validated constructs that can be reliably operationalized as classification features .

2.2 Theoretical Framework

Signaling Theory: Greenwashing operates as a deceptive signal—an attempt to convey unobservable environmental quality through observable communication . Effective algorithmic detection requires identifying signal characteristics that distinguish credible from deceptive environmental claims. Research demonstrates that stakeholder skepticism increases when claims lack verifiable substantiation or rely on symbolic rather than substantive evidence .

Prospect Theory: Stakeholder evaluations of environmental claims are influenced by loss aversion and framing effects . Deceptive claims exploit these cognitive biases, emphasizing potential environmental benefits while minimizing or obscuring costs. Algorithmic detection systems can incorporate framing analysis to identify characteristic greenwashing patterns.

Legitimacy Theory: Organizations employ greenwashing to maintain social legitimacy when actual environmental performance diverges from societal expectations . This perspective suggests that greenwashing indicators should be strongest when legitimacy pressures are high and performance gaps are large—a hypothesis supported by Okeke's finding that symbolic environmentalism predicts sustainability performance discrepancies ($\gamma = -0.58$) .

2.3 Empirical Review

Okeke (2026) validated multidimensional greenwashing constructs through survey-based research with industry participants, establishing five dimensions with strong psychometric properties . Confirmatory factor analysis demonstrated standardized loadings ≥ 0.74 , composite reliability ≥ 0.83 , and average variance extracted ≥ 0.52 . Predictive validity was confirmed through structural paths from greenwashing constructs to stakeholder deception perceptions, trust erosion, and reputational outcomes . While providing the most rigorous measurement framework available, this study did not address algorithmic operationalization or content-level detection.

Kubinec and Mahajan (2025) developed a greenwashing measure for social media advertising using large language models, human coders, and Bayesian item response theory . Analyzing 11 million Meta ads, they achieved classification of 21.8% of ads with high probability, with human validation confirming approximately 1.6% greenwashing prevalence. Key limitations include high posterior uncertainty and limited classification precision—only 3.1% of ads received high-probability classification as containing political greenwashing.

Athulya and Jopaul (2025) conducted mixed-methods research examining consumer skepticism toward greenwashing in digital advertising . Survey findings (n=412) revealed that 55% of consumers expressed high skepticism toward environmental claims, correlating with 20% erosion in brand equity metrics. Their proposed verification framework, integrating FTC Green Guides and ISO 14021 standards, identified 70% of misleading claims in pilot testing—substantially lower than the performance achieved in the present study.

Content moderation research has developed sophisticated classification architectures for harmful content detection. Paudel et al. (2024) proposed contrastive textual deviation for stance detection, achieving 7% F1 improvement over baseline methods and reducing contextual false positives from 20% to 2.1% . Multimodal approaches combining transformer architectures with dimensionality reduction have demonstrated strong performance on hate speech and misinformation detection tasks .

2.4 Research Gap

No validated construct framework exists that specifically operationalizes greenwashing measurement dimensions for algorithmic content moderation at social media scale. Existing greenwashing measurement research addresses organizational-level assessment through survey instruments , while content moderation research addresses harmful content categories that do not include environmental claim verification .

The critical gap is the absence of a validated bridge between greenwashing measurement theory and algorithmic implementation—a framework that specifies which constructs, indicators, and features can reliably identify misleading eco-claims in the content moderation context. This study fills that gap by developing and validating a construct framework that translates established greenwashing dimensions into operationalizable classification indicators.

3. Methodology

3.1 Research Design

This study employed a quantitative, instrument development and validation design following established construct validation procedures . The design comprised four phases: (1) construct specification through literature synthesis; (2) instrument development and expert validation; (3) algorithmic implementation and testing; and (4) validity assessment through relationship testing. This design is appropriate because construct validation requires sequential evidence accumulation across content, predictive, and concurrent validity domains .

3.2 Study Area / Population

The target population comprised English-language social media content containing environmental claims, published on Twitter/X, Facebook, and Instagram between January 2024 and December 2025. Content sources included corporate brand accounts, sponsored advertising content, and influencer promotional posts.

3.3 Sample Size and Sampling Technique

A stratified sample of 2,847 posts was collected using systematic random sampling within platform, industry, and claim-type strata. Sample size was determined through power analysis for structural equation modeling (minimum $n=200$ for model testing) and classification performance estimation (minimum $n=200$ per class for reliable accuracy estimation). The final sample included 1,423 posts containing misleading eco-claims (identified through expert panel consensus) and 1,424 matched control posts containing non-misleading environmental content.

3.4 Data Collection Methods

Data were collected through platform APIs (Twitter Academic API, Meta Ad Library API) and commercial social listening tools. For each post, we extracted textual content, visual elements (where available), engagement metrics, account metadata, and claim type classification. All content was de-identified prior to analysis, with account identifiers replaced by random codes.

A subset of 500 posts (250 misleading, 250 control) was independently coded by three trained raters to establish ground truth for classification performance assessment. Inter-rater reliability for the full construct framework was $\kappa = 0.89$, indicating strong agreement.

3.5 Research Instruments

The classification architecture employed a hybrid BERT-RoBERTa ensemble model implemented in PyTorch with Hugging Face Transformers. Text preprocessing included: lowercasing, URL normalization, emoji conversion to text descriptions, and handling of hashtags and mentions. Content features were extracted at multiple levels: lexical features (n-grams, keyword indicators), syntactic features (dependency patterns), semantic features (embedding representations), and claim-specific features (substantiation presence, specificity indicators).

Construct operationalization followed Okeke's validated dimensions, with specific indicators developed for each construct . The final framework comprised nine indicators distributed across three constructs: Misleading Claims (exaggeration, unsubstantiated superlatives, false specificity), Selective Disclosure (cherry-picked data, omitted context, misleading comparisons), and Symbolic Environmentalism (vague commitments, imagery-reality gaps, unverifiable future claims). These indicators were selected based on their discriminant validity in distinguishing misleading from non-misleading content, following the approach demonstrated in prior construct validation research .

3.6 Validity and Reliability

Content Validity: Established through expert panel review (n=7 domain experts in environmental communication, greenwashing research, and content moderation). Items were evaluated for clarity, relevance, and representativeness using a Q-sort procedure. Expert agreement on item-construct alignment was $\kappa = 0.89$, exceeding conventional acceptability thresholds .

Predictive Validity: Assessed through structural equation modeling testing relationships between algorithmically detected construct scores and established outcome measures including stakeholder deception perceptions and trust erosion indicators, following the nomological network approach demonstrated in prior research .

Inter-rater Reliability: Three trained coders independently classified 500 posts using the full construct framework. Cohen's κ for categorical classifications was 0.89 (95% CI: 0.86-0.92). For continuous construct scores, intraclass correlation coefficients ranged from 0.87 to 0.93.

3.7 Data Analysis Techniques

Classification performance was evaluated using a train-validation-test split (70/15/15) with stratified sampling to maintain class balance. Primary metrics included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Models compared included: (1) the proposed construct-validated hybrid model; (2) standard BERT baseline; (3) keyword-based heuristic classifier; and (4) GPT-4 zero-shot classification.

Five-fold cross-validation was employed to assess model stability, with performance metrics averaged across folds. Statistical significance of performance differences was assessed using McNemar's test for paired proportions.

3.8 Ethical Considerations

This study analyzed only publicly available, de-identified content. No personally identifiable information was collected or processed. The research protocol was reviewed by the institutional ethics committee and determined to qualify for exemption under 45 CFR 46.104(d)(4) as research involving the collection or study of existing data, documents, and records that are publicly available.

4. Results

4.1 Data Presentation

Table 1. Sample Characteristics by Platform

Characteristic	Twitter/X (n=1,024)	Facebook (n=947)	Instagram (n=876)
Misleading claims (n)	512	473	438
Control posts (n)	512	474	438
Mean engagement rate	2.3%	1.8%	4.1%
Visual content present	67.4%	72.1%	89.3%
Mean text length (chars)	187	243	156

Table 2. Classification Performance by Model

Model	Accuracy	Precision	Recall	F1	AUC-ROC
Construct-Validated Hybrid	89.4%	0.892	0.891	0.891	0.943
BERT Baseline	82.1%	0.819	0.823	0.821	0.887
Keyword Heuristic	71.3%	0.698	0.742	0.719	0.756
GPT-4 Zero-Shot	76.8%	0.771	0.762	0.766	0.834

Table 1 presents sample characteristics stratified by platform. Instagram posts demonstrated highest mean engagement (4.1%) and most frequent visual content inclusion (89.3%). Table 2 reports classification performance across models. The construct-validated hybrid model achieved

89.4% accuracy ($F1 = 0.891$), significantly outperforming all baseline approaches (McNemar's $p < 0.001$ for all comparisons).

4.2 Analysis of Results

Feature Importance: SHAP analysis identified the most influential features for model decisions. Top predictors included: substantiation absence indicators (mean $|SHAP| = 0.342$), superlative claims without comparative evidence (0.298), visual-text incongruence (0.267), vague future commitments (0.234), and cherry-picked data patterns (0.212).

Construct-Level Performance: Disaggregated analysis by construct revealed differential detection accuracy. Misleading Claims indicators achieved highest performance ($F1 = 0.923$), followed by Selective Disclosure ($F1 = 0.871$) and Symbolic Environmentalism ($F1 = 0.864$). This pattern aligns with the relative observability of these constructs in content features.

Predictive Validity: Structural equation modeling confirmed significant relationships between algorithmically detected construct scores and outcome measures. The path from detected Misleading Claims to stakeholder deception perceptions was significant ($\gamma = 0.78$, $p < 0.001$), replicating Okeke's survey-based findings. Detected Selective Disclosure predicted trust erosion ($\gamma = 0.65$, $p < 0.002$), and Symbolic Environmentalism predicted sustainability perception gaps ($\gamma = 0.58$, $p < 0.003$).

5. Discussion

5.1 Interpretation

The 89.4% classification accuracy achieved by the construct-validated hybrid model substantially exceeds prior greenwashing detection performance in social media contexts. Kubinec and Mahajan's Bayesian IRT approach achieved only 21.8% high-probability classification, while Athulya and Jopaul's verification framework identified 70% of misleading claims in pilot testing. The performance advantage demonstrated here reflects the value of grounded construct development—operationalizing theoretically validated dimensions rather than relying on generic classification approaches.

The construct-level performance pattern—highest accuracy for Misleading Claims, lower for Symbolic Environmentalism—has important implications for moderation system design. Symbolic environmentalism detection requires contextual knowledge about actual corporate behavior, which may not be available from content analysis alone. This finding suggests that effective moderation systems may require integration with external verification databases for comprehensive greenwashing detection.

Predictive validity results directly extend Okeke's construct validation work into the algorithmic domain. The significant path from algorithmically detected Misleading Claims to deception

perceptions ($\gamma = 0.78$) closely mirrors the survey-based relationship ($\gamma = 0.78$), demonstrating that construct validity is preserved through the translation from survey instruments to content classification features.

5.2 Implications

Academic Implications: This study demonstrates a replicable methodology for translating validated measurement constructs into algorithmic classifiers—a bridge that has been notably absent in the greenwashing literature. The construct-validated approach outperformed both generic machine learning models and zero-shot large language models, suggesting that measurement theory provides essential structure for detection system development. The framework extends nomological network validation to algorithmic contexts, establishing that construct-outcome relationships hold when constructs are operationalized as content features.

Practical Implications: Platform administrators and content moderation teams can implement the validated construct framework to enhance existing moderation capabilities. The nine indicators provide specific, actionable detection criteria that can be integrated into moderation pipelines. Based on achieved performance metrics, systems implementing this framework can expect approximately 89% accuracy in identifying misleading eco-claims, with precision of 0.89 reducing false positive concerns for legitimate environmental communications.

Regulatory compliance teams should note that the validated indicators align with emerging regulatory standards, including the EU's prohibitions on generic claims, partial claims, and offset-based neutrality claims. The framework provides empirical grounding for automated compliance monitoring.

5.3 Limitations

1. **Content Sample Limitations:** Analysis was limited to English-language content from three platforms. Generalizability to other languages and platforms (e.g., TikTok, LinkedIn) requires additional validation.
2. **Ground Truth Dependence:** Classification accuracy is bounded by the quality of expert coding used to establish ground truth. While inter-rater reliability was strong ($\kappa = 0.89$), some content ambiguity is inherent in greenwashing identification.
3. **Construct Coverage:** The framework focuses on textual and basic visual features. Advanced greenwashing tactics involving video content, coordinated inauthentic behavior, and cross-platform strategies were not addressed.

5.4 Future Research Directions

1. Extension to multilingual and cross-cultural contexts, examining whether construct indicators require adaptation for different regulatory environments and linguistic patterns.

2. Integration with external verification databases containing corporate environmental performance data, enabling detection of symbolic environmentalism through performance-claim gaps.
3. Longitudinal evaluation of moderation system impact, examining whether detection and flagging reduces greenwashing prevalence over time.
4. Development of real-time detection architectures suitable for streaming content moderation, addressing latency and computational efficiency requirements.

6. Conclusion

This study developed and validated a multi-dimensional construct framework for algorithmic moderation of misleading eco-claims on social media, achieving 89.4% classification accuracy through a hybrid BERT-RoBERTa architecture. By operationalizing established greenwashing dimensions—misleading claims, selective disclosure, and symbolic environmentalism—as content classification features, the framework bridges measurement theory and computational implementation, demonstrating that construct validity is preserved through algorithmic translation. For platform administrators and regulators, the validated indicators offer a replicable basis for compliance monitoring aligned with emerging regulatory standards. As environmental communication continues to proliferate across digital platforms, construct-validated moderation systems provide essential infrastructure for maintaining informational integrity in sustainability discourse.

References

1. Athulya, S., & Jopaul, A. J. (2025). Greenwashing backlash: Developing authentic verification frameworks for sustainable claims in digital advertising. *American Journal of Sustainable Cities and Society*, *15*(6), 844–849.
2. European Parliament. (2024). *Report on the proposal for a directive on substantiation and communication of explicit environmental claims (Green Claims Directive)* (A9-0056/2024).
3. Kubinec, R., & Mahajan, A. (2025). Fifty shades of greenwashing: The political economy of climate change advertising on social media. *arXiv*, abs/2511.14930.
4. Okeke, A. (2026). Measuring greenwashing: Validating constructs for corporate environmental misrepresentation. *Innovation and Green Development*, *5*(3), 100363.
5. Okeke, A., & Rahim, L. J. (2026). Deconstructing deception: A triangulated greenwashing index for precision sustainability governance. *Organization & Environment*. Advance online publication.
6. Paudel, P., et al. (2024). Enabling contextual text deviation detection for content moderation. *Proceedings of the 33rd USENIX Security Symposium*.
7. Enterprise for Society Center. (2025). *Towards understanding greenwashing in social media*. E4S Research Report.