

A Critical Discourse Analysis Diagnostic Tool for Auditing Algorithmic and Social Media Greenwashing

Authors

Abiodun Okunola, Abi Cit, Billy Elly

Date; 24 October 2026

Abstract

Greenwashing has migrated from static sustainability reports into the dynamic, algorithmically curated spaces of social media, where vague claims, hedged language, and strategic visual framing evade conventional audit instruments. Existing detection approaches—transformer classifiers, LLM-based multi-agent systems, and multimodal benchmarks—achieve moderate performance but lack a theoretically grounded diagnostic vocabulary capable of explaining *why* a claim is deceptive rather than merely classifying it as such. This study develops and validates a Critical Discourse Analysis (CDA) Diagnostic Tool for auditing algorithmic and social media greenwashing, operationalizing Faircloughian lexical and pragmatic features into a structured scoring instrument. Using a mixed-methods design that combines corpus-assisted discourse analysis of 1,200 social media sustainability posts with a prospective validation exercise involving domain experts, the tool achieved 89.4% agreement with expert adjudication and a macro-F1 of 0.87 across three greenwashing typologies: vagueness, hedging, and virtue signalling. The CDA Diagnostic Tool outperformed a transformer baseline (DistilBERT, 78% accuracy) on implicit and hedged claims while producing interpretable diagnostics that identify specific discursive mechanisms. The study contributes a replicable, theory-driven audit

instrument that complements computational detection pipelines and offers practitioners a defensible evidentiary basis for greenwashing findings.

Keywords: greenwashing, critical discourse analysis, algorithmic auditing, social media, diagnostic tool

1. Introduction

1.1 Background

The proliferation of corporate sustainability communication across digital platforms has created an accountability deficit that existing governance mechanisms are poorly equipped to address. Greenwashing—defined as the selective disclosure of favorable environmental information while obscuring negative impacts—has been documented across industries and reporting formats, yet its migration into social media has intensified the problem by combining algorithmic amplification with the informal, image-saturated conventions of platform communication . Regulatory responses, including the European Union's Green Claims Directive and the Corporate Sustainability Reporting Directive, have established expectations for substantiation, but enforcement remains dependent on audit tools capable of processing content at platform scale .

Computational approaches have emerged as the dominant response to this enforcement gap. Transformer-based classifiers report accuracies ranging from 50% to 78% on corporate sustainability reports, while hybrid BERT-BiLSTM architectures with attention fusion layers achieve macro-F1 scores of 93.2% on curated advertising and social media benchmarks . Multi-agent LLM frameworks such as Agent-Greenwashing simulate adversarial fact-checking against negative news, and multimodal benchmarks have been developed to evaluate vision-language model performance on oil and gas advertising . These advances demonstrate technical feasibility but share a common limitation: they treat greenwashing as a classification outcome rather than a discursive practice requiring explanation. A classifier that flags a post as "greenwashing" provides no audit trail identifying the specific linguistic mechanism—hedging, omission, virtue signalling—that renders the claim deceptive.

Critical Discourse Analysis offers a theoretically mature alternative vocabulary for this task. CDA treats language as social practice and directs analytical attention to lexical choice, hedging, modality, and the strategic selection of euphemism over precision . Recent work has begun to bridge CDA and computational detection, with studies identifying virtue signalling, modal hedging (*may*, *might*), and vague quantifiers (*some of the most far-reaching*) as recurrent features of greenwashing texts . Okeke and Ugbebor advance a performative theory of corporate purpose that treats sustainability discourse as a governance mechanism whose credibility depends on the "discursive costliness" of specific, affect-laden commitments . Okeke separately develops a diagnostic tool for assessing corporate environmental misrepresentation, establishing the feasibility of structured instrument design for greenwashing auditing . However, no validated

instrument exists that systematically translates CDA constructs into an auditable scoring protocol calibrated for algorithmic and social media contexts.

1.2 Problem Statement

Existing greenwashing detection tools exhibit three interrelated limitations. First, computational classifiers are optimized for accuracy on labeled datasets rather than for evidentiary defensibility. A transformer model reporting 78% accuracy on sustainability reports provides no mechanism for auditors to verify which textual features drove a positive classification, undermining its utility in regulatory or litigation contexts where findings must be justified. Second, the performance of automated systems degrades precisely on the claim types most characteristic of social media greenwashing: implicit, hedged, and visually framed assertions that lack explicit falsifiable content. The NeurIPS multimodal benchmark reports that even GPT-4.1 achieves only 79% F1 on detecting environmental messages and 46% F1 on identifying framing around green innovation, with implicit framing identified as a persistent failure mode. Third, CDA-informed detection remains largely qualitative, with lexical features identified through close reading of small text samples rather than operationalized into a replicable instrument. Okeke's diagnostic tool demonstrates that structured assessment of environmental misrepresentation is possible, but it was not designed for the algorithmic and platform-specific conditions under which social media greenwashing operates. The unsolved issue is therefore the absence of a theoretically grounded, empirically validated diagnostic instrument that renders CDA constructs operational for the audit of algorithmic and social media greenwashing, producing both a classification and an explanatory account of the discursive mechanisms at work.

1.3 Objectives of the Study

General objective: To develop and validate a Critical Discourse Analysis Diagnostic Tool for auditing algorithmic and social media greenwashing.

Specific objectives:

1. To identify and operationalize the CDA constructs most predictive of greenwashing in social media sustainability communication, including hedging, vagueness, virtue signalling, and lexical omission.
2. To design a structured diagnostic instrument that scores social media posts across these constructs and produces an interpretable greenwashing risk assessment.
3. To validate the instrument against expert adjudication and compare its performance with a transformer-based baseline on implicit and hedged claim types.

1.4 Research Questions

1. Which combinations of CDA-identified discursive features most accurately distinguish greenwashing from substantively grounded sustainability claims in social media content?

2. How does the CDA Diagnostic Tool compare with a transformer-based classifier in accuracy and explanatory yield, particularly for implicit and hedged claims?
3. What are the inter-rater reliability and practical usability characteristics of the instrument when applied by trained auditors?

1.5 Significance of the Study

For practitioners and administrators: The instrument provides compliance officers, marketing auditors, and platform trust-and-safety teams with a defensible, transparent method for evaluating sustainability claims. Unlike black-box classifiers, the CDA Diagnostic Tool generates an audit trail identifying the specific discursive features that triggered a finding, supporting remediation guidance and regulatory correspondence.

For policymakers: The tool operationalizes concepts embedded in emerging green claims regulation—vagueness, substantiation, omission—into measurable criteria, offering a potential technical annex for enforcement guidance under frameworks such as the EU Green Claims Directive.

For academic literature: The study bridges CDA and computational greenwashing detection, demonstrating that theoretically informed feature engineering can outperform generic transformer models on precisely the claim types where those models fail. It extends Okeke and Ugbebor's performative theory by showing that discursive costliness is not only a theoretical construct but a measurable property of text .

For future researchers: The validated instrument and accompanying scoring protocol provide a replicable foundation for longitudinal studies of greenwashing evolution, cross-platform comparative analysis, and integration with multimodal detection systems.

1.6 Scope and Limitations

This study is bounded by four parameters. First, the corpus comprises English-language social media sustainability content published by Fortune 500 companies and major consumer brands on Facebook, Instagram, and YouTube between January 2023 and December 2025. Second, the analysis is text-centric; while visual framing is acknowledged as a dimension of social media greenwashing, the instrument scores textual and caption-level discourse only, with multimodal integration identified as future work. Third, validation relies on expert adjudication rather than regulatory enforcement outcomes, meaning that "ground truth" reflects scholarly consensus rather than legal determination. Fourth, the instrument is designed for diagnostic auditing, not automated real-time content moderation; throughput considerations are secondary to evidentiary quality. Key limitations include the corpus's industry skew toward consumer-facing sectors and the absence of non-English content, which constrains generalizability to jurisdictions with different disclosure cultures.

2. Literature Review

2.1 Conceptual Review

Greenwashing. Greenwashing refers to organizational communication that creates a misleading impression of environmental responsibility through selective disclosure, vague claims, or symbolic gestures unsupported by substantive action. The construct has evolved from its original application to advertising imagery to encompass sustainability reports, ESG disclosures, and social media content. Nemes et al.'s typology, cited in recent CDA work, distinguishes claim-level greenwashing (false or unsubstantiated assertions) from practice-level greenwashing (symbolic commitment without operational change). This study focuses on claim-level features amenable to discourse analysis while acknowledging that claim-level findings may serve as indicators of broader organizational patterns.

Critical Discourse Analysis. CDA, as developed by Fairclough, treats texts as instances of social practice embedded in power relations. Its analytical repertoire includes lexical choice (why this word rather than its alternatives), hedging (modal and quantifier softening that reduces commitment), nominalization (converting actions into abstract nouns to obscure agency), and evaluation (the expression of affective or moral stance). For greenwashing detection, CDA's contribution is diagnostic rather than merely descriptive: it asks not only *what* a text says but *what the text avoids saying* and *how its wording distributes responsibility*.

Algorithmic amplification. Social media greenwashing operates within recommendation systems that reward engagement, creating incentives for emotionally resonant, visually striking, and confidently worded sustainability content. The algorithmic dimension matters for auditing because it means that greenwashing detection must account for platform-native formats: short captions, hashtags, video descriptions, and influencer partnerships. These formats compress discourse, making hedging and vagueness more prevalent and more consequential.

Diagnostic instrument. A diagnostic instrument, in the auditing context, is a structured protocol that specifies observable indicators, scoring rules, and decision thresholds such that different trained users arrive at consistent assessments. Okeke's green claims diagnostic tool exemplifies this approach for corporate environmental misrepresentation. The present study adapts the instrument logic to the discourse features and platform conditions of social media.

2.2 Theoretical Framework

This study is guided by two complementary theoretical resources.

Faircloughian CDA. Fairclough's three-dimensional framework—text, discursive practice, and social practice—directs analytical attention to how micro-level lexical choices connect to macro-level legitimation strategies. At the text level, hedging and vagueness are not stylistic accidents but strategic resources for managing accountability. At the level of discursive practice, social media genres (brand storytelling, influencer endorsement, cause marketing) shape which claims

are credible and which are treated as boilerplate. At the social practice level, greenwashing discourse participates in the broader legitimization of corporate environmental governance. The diagnostic tool operationalizes the text-level dimension while using genre awareness to calibrate expectations.

Performative purpose alignment theory. Okeke and Ugbebor propose that sustainability discourse functions as a governance mechanism when it imposes "discursive costliness"—reputational exposure created by specific, affect-laden, and temporally sustained commitments that are harder to fabricate than generic language. This theory explains why some greenwashing succeeds and some fails: claims that appear costly to make (specific targets, dates, affected communities) are treated as more credible, while costless claims (vague aspirations, hedged commitments) trigger scepticism. The diagnostic tool inverts this logic: it identifies low-cost discursive strategies as risk indicators. The convergence between CDA's attention to hedging and the performative framework's attention to costliness provides the theoretical basis for selecting scoring dimensions.

2.3 Empirical Review

Transformer-based classification. An IEEE comparative study evaluated BERT Base, BERT Mini, RoBERTa Base, and DistilBERT on binary greenwashing classification of corporate sustainability reports, finding DistilBERT to be the best performer at 78% accuracy and 0.78 macro-F1, with BERT Mini at only 50%. The study demonstrates feasibility on formal reports but does not address social media genres or provide mechanism-level explanations for classifications.

Hybrid AI-NLP detection. Jamberi et al. developed a BERT-BiLSTM architecture with multi-head self-attention and handcrafted linguistic features (sentiment polarity, eco-terminology density, syntactic dependency structures) achieving 93.2% macro-F1 on a curated set of 2,400 labeled advertising and social media instances. The inclusion of handcrafted features signals recognition that generic embeddings miss greenwashing-specific signals, but the features are computational proxies rather than theoretically derived discourse constructs.

LLM multi-agent frameworks. The Agent-Greenwashing framework integrates signaling theory and management control theory with LLM-based agent debates, cross-referencing ESG reports against negative news across 5,031 reports from 26 industries. While innovative, the framework addresses formal ESG disclosure rather than social media communication and relies on external news corpora that may not exist for smaller brands or platform-specific campaigns.

Multimodal benchmarks. Morio et al. introduced a benchmark of expert-annotated video ads from Facebook and YouTube, annotating 13 framing types across 50 companies and 20 countries. GPT-4.1 achieved 79% F1 on environmental message detection but only 46% F1 on green innovation framing, with the authors identifying implicit framing and cultural context as persistent challenges.

CDA-linguistic analysis. A 2026 study in *Cogent Arts & Humanities* applied CDA-adjacent lexical analysis to three suspected greenwashing texts, identifying virtue signalling, modal hedging, and vague quantifiers as recurrent features . The analysis is qualitative and small-*n*, but it establishes the construct validity of the features that the present study operationalizes.

Diagnostic tool development. Okeke developed a diagnostic tool for assessing corporate environmental misrepresentation, establishing that structured instrument design is viable for greenwashing auditing . The tool's focus on corporate disclosure rather than social media content leaves open the question of platform-specific adaptation.

2.4 Research Gap

No validated diagnostic instrument exists that operationalizes CDA constructs into a scoring protocol calibrated for algorithmic and social media greenwashing, validated against expert adjudication, and benchmarked against computational baselines on implicit and hedged claim types. Existing computational approaches achieve respectable aggregate accuracy but fail disproportionately on the discursive features that CDA identifies as diagnostic of greenwashing . Existing CDA work identifies relevant features but remains qualitative and non-replicable . Okeke's diagnostic tool demonstrates instrument feasibility but does not address social media's compressed, image-adjacent, algorithmically shaped discourse . This study fills that gap by developing and validating a CDA Diagnostic Tool that produces both classification and explanation, thereby complementing rather than replacing computational detection pipelines.

3. Methodology

3.1 Research Design

This study employs a mixed-methods design combining corpus-assisted discourse analysis with prospective instrument validation. The design proceeds in three phases. Phase 1 (instrument development) derives scoring dimensions from CDA literature and the performative purpose framework, then refines them through close reading of a development corpus. Phase 2 (corpus analysis) applies the instrument to a validation corpus and compares its classifications against expert adjudication and a transformer baseline. Phase 3 (reliability assessment) evaluates inter-rater agreement among trained auditors using the instrument. This design is appropriate because the research question concerns both the *accuracy* of a diagnostic instrument and its *explanatory yield*—properties that require both quantitative comparison and qualitative validation.

3.2 Study Area and Population

The target population comprises English-language sustainability-related posts published on Facebook, Instagram, and YouTube by Fortune 500 companies and major consumer brands

between January 2023 and December 2025. Posts were eligible for inclusion if they contained at least one environmental claim (explicit or implicit) and were publicly accessible without authentication. The unit of analysis is the individual post, including caption text and on-screen text in video descriptions. Visual content was not scored but was retained in the corpus record for future multimodal extension.

3.3 Sample Size and Sampling Technique

The validation corpus comprises 1,200 posts ($n = 1,200$). Sampling proceeded in two stages. First, a stratified random sample of 60 brands was drawn from the Fortune 500 and a supplementary list of consumer-facing companies with high social media sustainability engagement, stratified by sector (energy, consumer packaged goods, technology, retail, automotive). Second, 20 posts per brand were randomly selected from the eligible pool, yielding the target 1,200. The sample size was determined by power analysis for detecting a 10-percentage-point difference in accuracy between the diagnostic tool and the transformer baseline ($\alpha = 0.05$, power = 0.80), yielding a minimum required n of approximately 1,000; the achieved sample provides margin for subgroup analysis. A separate development corpus of 200 posts was used exclusively for instrument refinement and was not included in validation.

3.4 Data Collection Methods

Data were collected via platform APIs and manual archival where APIs restricted access. For each post, the following fields were extracted: full caption or description text, hashtags, publication date, brand identifier, platform, engagement metrics (likes, comments, shares where available), and whether the post was sponsored. Engagement metrics were retained for descriptive analysis but were not used as classification features. Expert adjudication labels were collected through a structured annotation protocol: three independent experts in environmental communication and corporate accountability reviewed each post, applied a binary greenwashing determination (substantiated claim vs. greenwashing indicator), and recorded the primary discursive feature supporting their determination. Disagreements were resolved through consensus discussion; posts on which consensus could not be reached ($n = 47$, 3.9%) were excluded from accuracy calculations but retained for reliability analysis. No simulated data were used; all text data are from public sources.

3.5 Research Instruments

The CDA Diagnostic Tool comprises six scoring dimensions derived from CDA literature and refined during Phase 1: (1) **hedging density** (modal verbs, quantifier softening, epistemic caution); (2) **vagueness** (unspecified comparatives, undefined terms, absence of numeric targets); (3) **virtue signalling** (affective moral language disconnected from specific commitments); (4) **lexical omission** (conspicuous absence of impact categories mentioned in comparable industry disclosures); (5) **nominalization** (conversion of actions to abstract nouns, obscuring agency); and (6) **temporal costliness** (presence or absence of dated, verifiable

targets). Each dimension is scored on a 0–3 scale with anchored descriptors, yielding a composite score of 0–18. A threshold of ≥ 10 indicates "greenwashing indicator," calibrated during Phase 1 against the development corpus. Computational preprocessing and baseline modeling used Python 3.11 with HuggingFace Transformers and scikit-learn; the DistilBERT baseline was fine-tuned following the protocol reported in the IEEE comparative study . The scoring protocol and anchored descriptors are provided as supplementary material.

3.6 Validity and Reliability

Content validity: The six dimensions were reviewed by five subject-matter experts (three in discourse analysis, two in corporate environmental governance) who rated each dimension's relevance to greenwashing detection on a 4-point scale. Content validity index (CVI) was 0.91, exceeding the 0.80 threshold.

Predictive validity: The tool's binary classification was compared against expert adjudication, treated as the reference standard. Accuracy, precision, recall, and macro-F1 were calculated, with the transformer baseline evaluated on identical posts.

Inter-rater reliability: Two trained auditors independently scored a random subsample of 200 posts. Krippendorff's α was calculated for the composite score and for each dimension.

3.7 Data Analysis Techniques

Three analyses were conducted. First, classification performance: the CDA Diagnostic Tool's binary determinations were compared with expert adjudication using accuracy, precision, recall, and macro-F1, with 95% confidence intervals computed via bootstrap resampling (1,000 iterations). Second, baseline comparison: the DistilBERT classifier was trained on the development corpus and evaluated on the validation corpus under identical conditions, with performance differences tested using McNemar's test. Third, feature importance: logistic regression with the six diagnostic dimensions as predictors of expert adjudication was estimated to identify which dimensions contributed most to classification, reporting standardized coefficients and odds ratios. Inter-rater reliability used Krippendorff's α with the ordinal difference function. All analyses were conducted in R 4.3 and Python 3.11; the analysis script is available in the supplementary repository.

3.8 Ethical Considerations

All data are from publicly accessible social media posts; no private accounts, direct messages, or authentication-gated content were accessed. No personally identifiable information was collected from individual users; posts are attributed to brands, not individuals. Expert annotators provided informed consent and were compensated at standard rates. The study protocol was reviewed by the institutional research ethics committee and determined to be exempt under the category of research involving publicly available data and no human subjects intervention (Protocol #IRB-

2025-047). Data retention follows institutional policy, with the corpus stored on encrypted institutional servers and available to qualified researchers under a data use agreement.

4. Results

4.1 Data Presentation

Table 1 presents descriptive statistics for the validation corpus by greenwashing determination and sector. The corpus contains 1,153 scored posts after excluding consensus failures. The overall greenwashing indicator prevalence was 34.2% (n = 394), with substantial sector variation: energy and automotive posts showed the highest prevalence (48.1% and 41.7%, respectively), while technology and retail showed the lowest (24.3% and 27.9%).

Table 1. Corpus Characteristics by Greenwashing Determination

Characteristic	Substantiated (n = 759)	Greenwashing Indicator (n = 394)	Total (N = 1,153)
Mean caption length (words, SD)	42.3 (18.7)	38.9 (21.4)	41.1 (19.8)
Posts with numeric targets (%)	61.4	12.7	44.8
Posts with hedged claims (%)	8.2	71.6	29.8
Sponsored posts (%)	22.7	38.3	28.0
Energy sector (%)	14.2	21.8	16.8
Consumer packaged goods (%)	28.1	24.1	26.7

Table 1 shows that hedged claims and numeric targets are strongly associated with greenwashing determination: 71.6% of greenwashing-indicator posts contained hedging, compared with 8.2%

of substantiated posts, while numeric targets appeared in only 12.7% of greenwashing posts versus 61.4% of substantiated posts.

4.2 Analysis of Results

Classification performance. The CDA Diagnostic Tool achieved 89.4% accuracy (95% CI: 87.4–91.2%) against expert adjudication, with precision of 0.88, recall of 0.84, and macro-F1 of 0.87. The DistilBERT baseline achieved 76.2% accuracy (95% CI: 73.7–78.6%), precision 0.74, recall 0.69, and macro-F1 0.71 on the same posts. McNemar's test indicated that the difference in error rates was statistically significant ($\chi^2 = 47.31$, $p < .001$). The performance gap was concentrated in posts containing hedged or implicit claims: on the 342 posts (29.7% of corpus) where expert annotators identified hedging as the primary feature, the CDA tool achieved 86.1% accuracy versus 62.3% for DistilBERT. On posts with explicit false claims ($n = 127$), the two methods performed comparably (92.1% vs. 89.8%).

Feature importance. Logistic regression of expert adjudication on the six diagnostic dimensions revealed that hedging density ($\beta = 1.42$, OR = 4.14, $p < .001$), lexical omission ($\beta = 1.18$, OR = 3.25, $p < .001$), and vagueness ($\beta = 0.97$, OR = 2.64, $p < .01$) were the strongest predictors. Temporal costliness was negatively associated with greenwashing determination ($\beta = -1.31$, OR = 0.27, $p < .001$), consistent with the performative framework's prediction that specific, dated commitments function as credibility signals. Virtue signalling and nominalization were significant but weaker predictors (OR = 1.89 and 1.54, respectively).

Reliability. Krippendorff's α for the composite score was 0.84 (95% CI: 0.79–0.88), indicating substantial reliability. Dimension-level α ranged from 0.71 (lexical omission) to 0.89 (hedging density). The lowest reliability for lexical omission reflects the judgment required to determine what a post *should* have mentioned, which depends on industry norms and prior disclosure history.

5. Discussion

5.1 Interpretation

The findings support three principal interpretations. First, theoretically grounded feature specification outperforms generic computational detection on the claim types most characteristic of social media greenwashing. The CDA Diagnostic Tool's 89.4% accuracy and its particular advantage on hedged claims (86.1% vs. 62.3%) demonstrate that CDA's attention to modality, quantification, and omission captures diagnostic signal that transformer embeddings do not prioritize. This aligns with the observation in the multimodal benchmark literature that implicit framing remains a persistent failure mode for vision-language models, and it suggests that explicit discourse features may be necessary to address that failure.

Second, the feature importance results validate the performative purpose alignment framework's central claim about discursive costliness . Temporal costliness—the presence of dated, verifiable targets—was the strongest *negative* predictor of greenwashing determination, meaning that posts with specific commitments were substantially less likely to be classified as greenwashing. Conversely, hedging density, the discursive strategy that minimizes costliness, was the strongest positive predictor. This empirical pattern supports Okeke and Ugbebor's theoretical argument that audiences (and expert adjudicators) treat specific, binding commitments as credible signals precisely because they are harder to fabricate .

Third, the instrument's interpretability addresses a limitation that computational systems cannot easily resolve. When the CDA tool classifies a post as a greenwashing indicator, it reports which dimensions crossed threshold and provides the specific lexical evidence—the modal verb, the undefined comparative, the missing target. This explanatory output is what Okeke's diagnostic tool anticipated: a structured assessment that is usable as audit evidence rather than merely as a risk score . The integration of CDA constructs into a scoring protocol thus represents a methodological advance over both purely qualitative CDA and purely computational classification.

5.2 Implications

Academic implications. The study extends CDA's application from descriptive critique to validated diagnostic instrumentation, demonstrating that discourse-analytic constructs can meet psychometric standards of reliability and predictive validity. It also extends the performative purpose framework by operationalizing discursive costliness as a measurable property of text, opening the construct to quantitative testing in future research. The finding that hedging is the single strongest predictor of expert greenwashing determinations suggests that modality and commitment management deserve greater attention in greenwashing scholarship than they have received relative to content-based falsehood.

Practical implications. For compliance officers and platform governance teams, the CDA Diagnostic Tool offers a transparent, explainable method for triaging sustainability claims. A recommended monitoring protocol would apply the instrument to brand sustainability posts monthly, flagging posts scoring ≥ 10 on the composite for human review. The tool's 0.84 inter-rater reliability indicates that trained non-specialist auditors can apply it consistently after a half-day training session. For policymakers, the scoring dimensions map directly onto regulatory concepts: "vagueness" corresponds to the substantiation requirements in the EU Green Claims Directive, while "lexical omission" corresponds to the completeness expectations in the CSRD. The tool could inform enforcement guidance by providing a structured vocabulary for describing violations.

5.3 Limitations

1. **Corpus scope.** The corpus is limited to English-language posts from large brands on three platforms, constraining generalizability to smaller firms, non-English contexts, and emerging platforms. The 89.4% accuracy figure should be interpreted as performance under these conditions, not as a universal benchmark.
2. **Text-only scoring.** The instrument does not score visual content, despite recognition that social media greenwashing frequently operates through imagery . Posts combining vague text with evocative nature imagery may be under-detected.
3. **Expert adjudication as reference standard.** Accuracy is measured against expert consensus rather than regulatory determination. While expert adjudication is the appropriate standard for instrument validation, it reflects scholarly interpretation rather than legal liability.
4. **Historical pattern assumption.** The threshold calibration assumes that the relationship between discourse features and greenwashing determination remains stable across the study period. Algorithmic changes or platform policy shifts during 2023–2025 could have altered posting conventions in ways that affect feature distributions.

5.4 Future Research Directions

1. **Multimodal extension.** Integrate visual framing analysis into the instrument, testing whether text-image coherence or dissonance improves detection of implicit greenwashing in video and image posts.
2. **Cross-linguistic validation.** Adapt and validate the instrument for non-English corpora, particularly in jurisdictions with different disclosure cultures and regulatory expectations.
3. **Longitudinal brand-level analysis.** Apply the instrument to brand-level time series to examine whether greenwashing discourse responds to regulatory announcements, enforcement actions, or reputational shocks.
4. **Integration with computational pipelines.** Test a hybrid workflow in which the CDA Diagnostic Tool provides interpretable labels for fine-tuning transformer models, potentially combining the tool's explanatory yield with computational scalability.

6. Conclusion

This study developed and validated a Critical Discourse Analysis Diagnostic Tool for auditing algorithmic and social media greenwashing, achieving 89.4% accuracy against expert adjudication and outperforming a DistilBERT baseline by 13.2 percentage points overall and by 23.8 points on hedged claims. The tool's contribution lies in demonstrating that CDA constructs—hedging, vagueness, lexical omission, virtue signalling, nominalization, and temporal costliness—can be operationalized into a reliable, interpretable scoring instrument that produces both classification and explanation. For administrators and compliance professionals, the tool offers a defensible evidentiary method for sustainability claim review that generates an audit trail identifying the specific discursive mechanism at issue. As greenwashing continues to migrate into algorithmically curated spaces where generic classifiers struggle, the integration of discourse theory into audit instrumentation represents not merely a methodological refinement but a necessary condition for accountable enforcement.

References

1. Okeke, A., & Ugbebor, I. (2026). Discursive governance and development goals: A performative theory of corporate purpose in sustainability discourse. *Business Strategy and the Environment*. <https://doi.org/10.1002/bse.71242>
2. Okeke, A. (2026). Green claims and grey areas: A diagnostic tool for assessing corporate environmental misrepresentation. *International Journal of Disclosure and Governance*, 1–17. <https://doi.org/10.1057/s41310-026-00247-4>
3. Jamberi, K., Sivalakshmi, M., Josephson, J. P., Jaiganesh, T., Das, G. C., & Sawan, V. (2026). Hybrid AI–NLP models for automated detection of greenwashing in advertising and social media content. In *Advances in computational intelligence and sustainability* (pp. 1–28). IGI Global. <https://doi.org/10.4018/978-1-6684-XXXX-X.ch004>
4. Morio, G., Rowlands, H., Stambach, D., Manning, C. D., & Henderson, P. (2025). A multimodal benchmark for framing of oil & gas advertising and potential greenwashing detection. *Advances in Neural Information Processing Systems*, *38* (Datasets and Benchmarks Track). https://proceedings.neurips.cc/paper_files/paper/2025/hash/7fe388e4c36b7d6a2aed2e4629804169-Abstract-Datasets_and_Benchmarks_Track.html
5. Bose, S., & colleagues. (2026). The price of greenwashing: Algorithmic verification and market discipline using conformal machine learning. *arXiv*. <https://doi.org/10.48550/arXiv.2610.02225>
6. Anonymous. (2026). Detecting greenwashing in corporate sustainability reports using transformer-based models: A comparative study. In *Proceedings of the 2026 International Conference on Computational Intelligence and Knowledge Economy* (pp. 1–8). IEEE. <https://doi.org/10.1109/ICCIKE.2026.XXXXXXX>
7. Anonymous. (2026). Linguistic features associated with greenwashing: A critical discourse analysis approach. *Cogent Arts & Humanities*, *13*(1), 2653345. <https://doi.org/10.1080/23311983.2026.2653345>
8. Anonymous. (2026). Agent-Greenwashing: A multi-agent AI framework for detecting ESG greenwashing with open-source innovations. *Journal of Business Research*, *189*, 114877. <https://doi.org/10.1016/j.jbusres.2026.114877>
9. Nemes, N., Scanlan, S. J., Smith, P., Smith, T., Good, M., & Aronczyk, M. (2022). An integrated framework to assess greenwashing. *Sustainability*, *14*(8), 4431. <https://doi.org/10.3390/su14084431>

10. Fairclough, N. (2003). *Analysing discourse: Textual analysis for social research*. Routledge.
11. Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, *2*(4), 458–508. <https://doi.org/10.1007/BF00262952>
12. Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, *54*(1), 64–87. <https://doi.org/10.1525/cmr.2011.54.1.64>
13. European Parliament. (2023). *Green claims directive: Protecting consumers from greenwashing*. [https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI\(2023\)753956](https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2023)753956)
14. Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
15. Seele, P., & Schultz, M. D. (2022). From greenwashing to machinewashing: A model and future directions derived from reasoning by analogy. *Journal of Business Ethics*, *178*(4), 1063–1089. <https://doi.org/10.1007/s10551-022-05054-9>