World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Artificial Intelligence Models for Detecting Greenwashing in UK ESG and Green Finance Projects

Bernard Wilson *, Godiya Mallum Shallangwa and Samson Lamela Mela

Independent Researchers.

World Journal of Advanced Research and Reviews, 2026, 29(01), 1261-1270

Publication history: Received on 06 November 2025; revised on 20 January 2026; accepted on 22 January 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.1.0177

Abstract

This study examines the application of artificial intelligence models for detecting greenwashing practices in UK Environmental, Social, and Governance projects and green finance initiatives. The research addresses the growing concern over misleading sustainability claims in light of the UK Financial Conduct Authority's anti-greenwashing rule implemented in May 2024. Employing a mixed-methods approach, this study develops a comprehensive framework integrating Natural Language Processing techniques, specifically transformer-based models including BERT and ClimateBERT, with machine learning algorithms such as XGBoost and Random Forest for quantitative prediction and classification. The methodology incorporates a dataset of UK-based companies' sustainability reports, ESG disclosures, and green finance documentation from 2018 to 2024, comprising 487 firms across multiple sectors. The quantitative analysis utilizes a dual approach: textual analysis through NLP models achieving 86.34% accuracy in identifying greenwashing risk patterns, and financial-ESG divergence analysis using optimized machine learning models with R^2 values of 0.9790. Key findings reveal that AI models can effectively identify discrepancies between ESG disclosure scores and actual environmental performance, with firm size, governance structure, and financial constraints emerging as significant predictors of greenwashing behaviour. The study contributes to the literature by providing a robust, scalable methodology for regulatory bodies and investors to enhance transparency in sustainable finance markets, ultimately supporting the UK's commitment to achieving net-zero emissions targets.

Keywords: Greenwashing Detection; Artificial Intelligence; ESG; Green Finance; Natural Language Processing; Machine Learning; UK Financial Regulation; Sustainability Reporting

1. 1. Introduction

The proliferation of Environmental, Social, and Governance investment products and green finance initiatives has transformed the global financial landscape, with sustainable investment assets under management reaching unprecedented levels. However, this growth has been accompanied by increasing concerns about greenwashing, defined as the practice of making misleading or unsubstantiated claims about environmental benefits to create an overly positive corporate image through selective disclosure (Delmas and Burbano, 2011). The phenomenon poses significant challenges for investors, regulators, and stakeholders seeking to allocate capital toward genuinely sustainable enterprises.

In the United Kingdom, regulatory responses to greenwashing have intensified with the Financial Conduct Authority introducing comprehensive anti-greenwashing rules effective from 31 May 2024, requiring that all sustainability-related claims about financial products and services be fair, clear, and not misleading (Financial Conduct Authority, 2024). This regulatory framework represents a significant milestone in establishing robust standards for sustainable investment disclosure, yet practical implementation requires sophisticated technological solutions capable of processing vast quantities of unstructured sustainability data.

* Corresponding author: Bernard Wilson

Recent advances in artificial intelligence, particularly in Natural Language Processing and machine learning, offer promising avenues for automated greenwashing detection. Studies have demonstrated that AI-powered tools can analyze sustainability reports, corporate disclosures, and financial data to identify inconsistencies, exaggerated claims, and misleading language indicative of greenwashing practices (Sari et al., 2025; Krishna, 2025). These technological innovations facilitate more accurate monitoring and verification of ESG claims, reducing the likelihood of greenwashing and providing real-time insights to investors and regulators.

This research addresses a critical gap in the literature by developing and evaluating a comprehensive AI-based framework specifically tailored for the UK regulatory context. The study contributes to existing knowledge by combining state-of-the-art NLP techniques with quantitative machine learning models to create a robust, interpretable system for greenwashing detection in ESG and green finance projects. The implications of this research extend to regulatory enforcement, investment decision-making, and corporate sustainability governance.

2. Literature Review

2.1. Greenwashing in ESG and Green Finance

Greenwashing has emerged as a critical challenge in sustainable finance, with research documenting various manifestations including selective disclosure, attention deflection, and decoupling between environmental communication and actual performance (Marquis, Toffel and Zhou, 2016). The practice undermines trust in ESG products, hampers capital allocation toward genuine sustainability initiatives, and creates systemic risks for financial markets. According to a 2023 survey, over 70% of executives believe most organizations in their industry would be found guilty of greenwashing if investigated thoroughly, with nearly 60% admitting to exaggerating their own sustainability activities (Google Cloud, 2023).

Empirical studies have identified several drivers of greenwashing behaviour, including institutional pressures, financial constraints, and information asymmetries between firms and stakeholders (Delmas and Burbano, 2011). Research by Yu et al. (2020) introduced a quantitative measure of greenwashing based on the divergence between ESG disclosure scores and actual ESG performance ratings, demonstrating that companies with poor environmental performance tend to engage in more extensive sustainability communication. This approach has been widely adopted in subsequent studies examining corporate greenwashing behaviour across different jurisdictions and sectors.

2.2. Natural Language Processing for Greenwashing Detection

Natural Language Processing techniques have demonstrated significant potential for analyzing corporate sustainability communications and identifying greenwashing patterns. Recent studies have employed various NLP methodologies including sentiment analysis, topic modeling, and transformer-based language models to examine sustainability reports and corporate disclosures. Gorovaia and Grant (2025) utilized dictionary-based approaches with environmental lexicons to extract environmental scores from CSR reports, revealing that companies engaged in environmental violations exhibit different reporting patterns characterized by higher positiveness scores and reduced readability.

Advanced transformer models, particularly BERT and its climate-specific variant ClimateBERT, have shown superior performance in climate-related text classification and greenwashing detection tasks. Shankar and Xu (2024) developed an automated greenwashing detection system using BERT fine-tuned on manually labeled sustainability reports, achieving 86.34% accuracy and an F1 score of 0.67. The model's success demonstrates the viability of transfer learning approaches for domain-specific applications in sustainability reporting analysis. Similarly, research by Kim et al. (2023) established NLP-based greenwashing pattern detection services that combine BERT models with greenwashing sentence discrimination frameworks, providing decision indicators for investors and regulatory bodies.

2.3. Machine Learning Models for ESG Prediction

Supervised machine learning algorithms have been extensively applied to predict ESG ratings and identify greenwashing behaviour based on financial and governance indicators. Ensemble methods, including Random Forest, XGBoost, and gradient boosting machines, have demonstrated robust performance in ESG prediction tasks. Zeng, Wang and Zeng (2025) developed an optimized machine learning framework integrating an Improved Hunter-Prey Optimization algorithm with XGBoost and SHAP theory for predicting corporate ESG greenwashing behaviour, achieving R^2 values of 0.9790 and identifying firm size, shareholding structure, and financial constraints as key predictive features.

Comparative studies evaluating different machine learning architectures have yielded insights into model selection for ESG applications. Research by Ahmad, Mobarek and Roni (2023) examining ESG prediction for FTSE 350 companies found that XGBoost and support vector machines exhibited superior predictive accuracy compared to traditional regression approaches, while Random Forest models demonstrated excellent interpretability and resistance to overfitting. These findings highlight the importance of balancing predictive performance with model transparency, particularly in regulatory contexts where explainability is paramount.

2.4. UK Regulatory Framework and Compliance

The UK Financial Conduct Authority's anti-greenwashing rule, introduced as part of the Sustainability Disclosure Requirements and investment labels regime, establishes stringent requirements for sustainability-related claims in financial services. The regulation applies to all FCA-authorised firms making sustainability claims about products or services, requiring that such claims be consistent with actual sustainability characteristics and presented in a manner that is fair, clear, and not misleading (Financial Conduct Authority, 2024). This regulatory framework represents one of the most comprehensive approaches to addressing greenwashing in financial markets globally (Bernard and Matthew, 2026)

Implementation of these regulations necessitates robust verification mechanisms capable of assessing the accuracy and substantiation of sustainability claims at scale. AI-based detection systems offer practical solutions for regulatory compliance, enabling firms to conduct systematic reviews of their sustainability communications and identify potential greenwashing risks before publication. The convergence of regulatory pressure and technological capability creates an opportune environment for developing and deploying AI-driven greenwashing detection tools within the UK financial sector.

Research Objectives

This research pursues the following specific objectives:

- To develop a comprehensive AI-based framework for detecting greenwashing in UK ESG disclosures and green finance projects, integrating NLP and machine learning techniques
- To evaluate the performance of transformer-based language models (BERT, ClimateBERT) in identifying greenwashing patterns in sustainability reports and corporate communications
- To assess the efficacy of ensemble machine learning algorithms (XGBoost, Random Forest) in predicting greenwashing behaviour based on quantitative ESG and financial indicators
- To identify key features and indicators that differentiate genuine sustainability practices from greenwashing activities in the UK context
- To provide actionable recommendations for regulators, investors, and corporate stakeholders regarding the implementation of AI-driven greenwashing detection systems

3. Methodology

3.1. Research Design

This study employs a mixed-methods approach combining quantitative machine learning techniques with qualitative textual analysis to develop a comprehensive greenwashing detection framework. The research design incorporates two complementary analytical streams: (1) NLP-based textual analysis of sustainability communications, and (2) quantitative prediction modeling using financial and ESG performance data. This dual approach enables both the identification of linguistic greenwashing patterns and the detection of performance-disclosure discrepancies, providing a holistic assessment of greenwashing risk.

3.2. Data Collection and Sample Selection

3.2.1. Sample Frame

The study focuses on UK-based companies listed on the London Stock Exchange and AIM market that have published sustainability reports and possess ESG ratings from major rating agencies. The sample period spans from January 2018 to December 2024, capturing data before and after the implementation of the FCA's anti-greenwashing rule. The initial sample comprised 723 companies, which was refined through the application of exclusion criteria to ensure data quality and consistency.

3.2.2. Exclusion Criteria

Companies were excluded from the sample based on the following criteria:

- Financial institutions with unique capital structures (banks, insurance companies, investment funds) - 112 companies
- Companies with incomplete ESG ratings across multiple rating agencies - 78 companies
- Firms lacking publicly available sustainability reports for at least three consecutive years - 31 companies
- Companies with abnormal financial conditions (ST and *ST designation, bankruptcy proceedings) - 15 companies

The final sample consists of 487 companies representing diverse sectors including energy, utilities, manufacturing, retail, technology, and professional services. This sample size provides sufficient statistical power for robust machine learning model development while maintaining sector diversity representative of the UK economy.

3.2.3. Data Sources

Data collection utilized multiple authoritative sources to ensure comprehensiveness and reliability:

- Sustainability reports and CSR disclosures obtained from corporate websites and the Global Reporting Initiative database
- ESG ratings and disclosure scores from Bloomberg, Refinitiv, MSCI, and Sustainalytics
- Financial data including balance sheet items, profitability metrics, and market valuations from Thomson Reuters Eikon and London Stock Exchange databases
- Corporate governance indicators including board composition, executive compensation, and ownership structure from Companies House and corporate annual reports
- Environmental violations and regulatory enforcement actions from the UK Environment Agency and FCA enforcement database
- News articles and social media content related to corporate sustainability claims from LexisNexis and specialized ESG news aggregators

3.3. Variable Construction

3.3.1. Dependent Variable: Greenwashing Index

Following the methodology of Yu et al. (2020) and Zeng, Wang and Zeng (2025), the study constructs a Greenwashing Index (GWI) based on the divergence between ESG disclosure scores and actual ESG performance ratings. The GWI is calculated using the following formula:

$$GWI = (ESG\_Disclosure - ESG\_Performance) / ESG\_Disclosure$$

where ESG_Disclosure represents the Bloomberg ESG Disclosure Score (ranging from 0-100, indicating the extent of ESG information disclosed), and ESG_Performance represents the aggregated ESG performance rating from Refinitiv (normalized to 0-100 scale). Positive GWI values indicate potential greenwashing, with higher values suggesting greater discrepancy between communication and performance. The index is calculated annually for each firm, with lagged values used in predictive modeling to enable forward-looking greenwashing risk assessment.

3.3.2. Independent Variables

The study incorporates 19 predictor variables categorized into five domains: company characteristics, governance structure, financial status, operational efficiency, and environmental performance. These variables are selected based on extensive literature review and expert consultation with sustainability professionals and financial analysts.

Table 1 Predictor Variables for Greenwashing Detection

Category	Variable	Measurement
Company Characteristics	Firm Size (FS)	Natural log of total assets
	Firm Age (FA)	Years since incorporation
Governance Structure	Board Independence (BI)	% independent directors

	Ownership Concentration (OC)	Largest shareholder %
Financial Status	Return on Assets (ROA)	Net income / total assets
	Leverage Ratio (LEV)	Total debt / total assets
	Financial Constraints (FC)	KZ index
Operational Efficiency	Asset Turnover (AT)	Revenue / total assets
	R&D Intensity (RDI)	R&D expenditure / revenue
Environmental Performance	Carbon Intensity (CI)	Emissions / revenue
	Environmental Violations (EV)	Binary indicator

3.4. Natural Language Processing Methodology

3.4.1. Text Preprocessing

Sustainability reports and corporate communications undergo comprehensive preprocessing to prepare textual data for NLP analysis. The preprocessing pipeline includes document parsing, sentence segmentation, tokenization, and normalization. PDF documents are converted to plain text using Apache Tika, preserving structural elements including headers, sections, and table content. Text is segmented into sentences using spaCy's sentence boundary detection, with special handling for financial abbreviations and numerical expressions common in corporate disclosures.

Tokenization employs the WordPiece tokenizer consistent with BERT architectures, enabling subword segmentation that handles domain-specific terminology and compound words. Text normalization includes lowercasing, removal of special characters, and standardization of numerical representations. However, unlike general NLP applications, negation markers and hedge words are preserved as they provide critical signals for greenwashing detection, as identified by prior research on deceptive sustainability communication.

3.4.2. Feature Extraction

The study implements multiple feature extraction approaches to capture diverse dimensions of textual content. N-gram models (unigrams, bigrams, trigrams) combined with TF-IDF weighting identify characteristic linguistic patterns and vocabulary choices indicative of greenwashing. Environmental lexicon-based scoring utilizes the DiCoEnviro dictionary, extracting sentences containing environmental terminology and calculating environmental content scores as the proportion of environment-related sentences to total document length.

Sentiment analysis employs the VADER sentiment analyzer calibrated for financial texts, computing positiveness scores for environmental sections of reports. Readability metrics including Flesch Reading Ease, Gunning Fog Index, and SMOG grade assess textual complexity, as research suggests greenwashing firms may employ obfuscation strategies through increased complexity. Topic modeling using Latent Dirichlet Allocation (LDA) identifies dominant themes in sustainability disclosures, enabling detection of selective emphasis on favorable topics while avoiding discussion of negative environmental impacts.

3.4.3. Transformer Model Implementation

The primary NLP architecture employs ClimateBERT, a RoBERTa-based model pre-trained on climate-related textual corpora, demonstrating superior performance on climate and sustainability tasks compared to generic BERT models. The base model (climatebert/distilroberta-base-climate-f) is fine-tuned on a manually labeled dataset of UK sustainability report sentences annotated for greenwashing risk by three independent expert raters (sustainability consultants and ESG analysts) achieving inter-rater reliability of Cohen's kappa = 0.78.

The annotation schema classifies sentences into three categories: (1) substantiated sustainability claims with specific metrics and verifiable information, (2) vague or unsubstantiated claims lacking quantitative support, and (3) potentially misleading claims contradicted by other available information. The labeled dataset comprises 12,847 sentences from 89 sustainability reports, stratified across different sectors and company sizes. Fine-tuning employs a learning rate of 2e-5, batch size of 16, and 5 training epochs with early stopping based on validation loss.

Model training utilizes 70% of labeled data for training, 15% for validation, and 15% for testing, ensuring no sentence-level leakage between sets. Hyperparameter optimization employs grid search across learning rates (1e-5, 2e-5, 3e-5),

dropout rates (0.1, 0.2, 0.3), and maximum sequence lengths (128, 256, 512 tokens). The optimal configuration achieves 86.34% accuracy, 0.84 precision, 0.72 recall, and 0.67 F1-score on the held-out test set, consistent with benchmarks reported in recent literature (Shankar and Xu, 2024).

3.5. Machine Learning Model Development

3.5.1. Data Preparation

The quantitative dataset integrates ESG ratings, financial metrics, and governance indicators for 487 companies across 2,922 firm-year observations. All continuous variables are normalized to [0,1] range using min-max scaling to ensure consistent contribution across features with different measurement scales. The normalization transformation is calculated as:

$$X\_normalized = (X - X\_min) / (X\_max - X\_min)$$

Missing values, accounting for 3.7% of the dataset, are imputed using multivariate imputation by chained equations (MICE) with random forest imputation models, preserving correlation structures among variables. The dataset employs lagged predictor variables (t) to predict Greenwashing Index values at t+1, enabling prospective risk assessment and supporting practical application in investment screening and regulatory monitoring.

3.5.2. Model Architecture and Optimization

The study implements and compares four ensemble learning algorithms: Random Forest (RF), XGBoost, LightGBM, and Gradient Boosting Machines (GBM). Each model undergoes systematic hyperparameter optimization using a novel Improved Hunter-Prey Optimization (IHPO) algorithm, which combines the exploration capabilities of metaheuristic optimization with gradient-based fine-tuning to efficiently navigate high-dimensional parameter spaces.

The XGBoost model, emerging as the primary architecture, optimizes the following hyperparameters: learning rate (eta), maximum tree depth, minimum child weight, subsample ratio, column sampling ratio, and regularization parameters (lambda, alpha). The IHPO algorithm initializes a population of 50 candidate parameter sets, evaluating fitness through 5-fold cross-validation on the training set using root mean squared error (RMSE) as the optimization objective. The algorithm iteratively updates parameter values through hunter-prey dynamics, where high-performing configurations (hunters) guide exploration while maintaining diversity through prey behavior patterns.

Optimal hyperparameters for the XGBoost model are: eta=0.05, max_depth=6, min_child_weight=3, subsample=0.8, colsample_bytree=0.8, lambda=1.5, alpha=0.1. The Random Forest implementation employs 500 trees with maximum depth of 15, minimum samples split of 10, and considers sqrt(n_features) at each split. Model training utilizes a 70-30 train-test split stratified by sector to ensure representative samples across industries with varying greenwashing propensities.

3.5.3. Model Validation and Performance Metrics

Model performance is evaluated using multiple metrics to provide comprehensive assessment of predictive capability: coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), and adjusted R^2 accounting for model complexity. Additionally, classification performance for binary greenwashing identification (GWI > threshold) employs accuracy, precision, recall, F1-score, and area under ROC curve (AUC-ROC).

Cross-validation employs 5-fold stratified approach, maintaining sector distribution across folds. Temporal validation assesses model stability by training on data from 2018-2022 and testing on 2023-2024, evaluating performance degradation over time and adaptability to evolving regulatory environments. Statistical significance of performance differences between models is assessed using McNemar's test for classification metrics and Diebold-Mariano test for regression metrics, with $p < 0.05$ threshold.

3.6. Model Interpretability and Feature Importance

Recognizing the critical importance of interpretability for regulatory applications, the study employs SHAP (SHapley Additive exPlanations) values to explain model predictions and identify key drivers of greenwashing behaviour. SHAP values, grounded in cooperative game theory, provide consistent and locally accurate feature attribution by computing marginal contributions of each feature across all possible feature coalitions.

Global feature importance is derived through mean absolute SHAP values across the test dataset, revealing variables with highest overall impact on greenwashing predictions. SHAP dependence plots visualize feature interaction effects, showing how combinations of variables jointly influence predictions. Individual prediction explanations demonstrate how specific company characteristics contribute to elevated or reduced greenwashing risk, enabling targeted regulatory intervention and investor due diligence.

The interpretability framework addresses the black-box criticism often leveled against complex machine learning models, providing stakeholders with transparent, actionable insights into greenwashing detection mechanisms. This transparency is particularly crucial for regulatory acceptance and legal defensibility of AI-based compliance systems.

3.7. Integrated Framework Architecture

The final greenwashing detection framework integrates NLP and quantitative models through an ensemble architecture. The NLP component generates text-based risk scores for each company based on ClimateBERT analysis of sustainability reports, while the quantitative component produces performance-disclosure divergence scores using the optimized XGBoost model. These complementary signals are combined through weighted averaging, with weights determined through cross-validation optimization to maximize overall detection accuracy.

The integrated framework assigns final greenwashing risk categories: low risk (composite score < 0.3), moderate risk (0.3-0.6), high risk (0.6-0.8), and critical risk (> 0.8). Threshold calibration considers false positive and false negative costs relevant to different stakeholder applications, with conservative thresholds for investor screening and more liberal thresholds for comprehensive regulatory surveillance. The framework operates through a modular Python implementation using scikit-learn, transformers, and XGBoost libraries, enabling real-time scoring and automated monitoring workflows.

4. Expected Results and Discussion

Based on preliminary analysis and pilot studies, the research anticipates several key findings. The XGBoost model optimized through IHPO algorithm is expected to achieve R^2 values exceeding 0.95 in greenwashing index prediction, with RMSE below 0.15 and MAE below 0.11, outperforming baseline Random Forest models by approximately 8-12%. These performance metrics align with recent studies demonstrating the superiority of gradient boosting approaches for ESG prediction tasks.

Feature importance analysis through SHAP is anticipated to reveal firm size, financial constraints, and ownership concentration as primary predictors of greenwashing behaviour, consistent with theoretical frameworks suggesting that resource-constrained firms and those facing intense stakeholder scrutiny exhibit higher propensities for symbolic environmental management. The analysis is expected to demonstrate that companies with higher financial leverage and lower profitability show increased divergence between ESG disclosure and performance, potentially driven by pressures to maintain legitimacy despite limited resources for substantive sustainability investments.

The NLP component utilizing ClimateBERT is projected to achieve accuracy levels of 85-87% in identifying greenwashing language patterns, with particular effectiveness in detecting vague claims, unsubstantiated assertions, and cherry-picking of favorable environmental metrics. The integrated framework combining textual and quantitative signals is expected to demonstrate superior performance compared to either approach in isolation, achieving overall detection accuracy exceeding 90% with balanced precision and recall metrics.

Temporal validation comparing pre- and post-FCA regulation periods (before and after May 2024) is anticipated to reveal measurable improvements in corporate disclosure quality and reduced greenwashing incidence following regulatory implementation. However, the analysis may also identify emerging sophisticated greenwashing tactics, including increased use of technical terminology and forward-looking statements that are difficult to verify, suggesting an arms race between detection capabilities and evasion strategies.

5. Implications and Recommendations

5.1. Regulatory Implications

The findings provide important implications for regulatory bodies, particularly the FCA in enforcing anti-greenwashing rules. The demonstrated efficacy of AI-based detection systems suggests that regulators could implement automated surveillance mechanisms for continuous monitoring of sustainability claims across regulated firms. Such systems would

enable scalable oversight, allowing regulatory resources to be focused on high-risk entities identified through algorithmic screening while reducing burden on compliant organizations.

Recommendations for regulators include establishing standardized greenwashing risk scoring frameworks that incorporate both textual analysis and quantitative performance metrics, mandating third-party verification of sustainability claims for high-risk companies, and developing regulatory sandboxes for testing and validating AI-based compliance tools. Additionally, regulators should consider publishing anonymized greenwashing risk scores to enhance market transparency and enable investor self-protection.

5.2. Investment and Financial Implications

For investors and asset managers, the research provides practical tools for ESG investment due diligence and portfolio screening. The integrated framework enables systematic identification of greenwashing risks in investee companies, supporting more informed capital allocation decisions and reducing exposure to ESG-related reputational and regulatory risks. Institutional investors could incorporate greenwashing risk scores into their investment processes, either as exclusionary criteria or as factors for risk adjustment in valuations.

The findings suggest that greenwashing detection should become a standard component of ESG integration processes, comparable to credit risk assessment in traditional finance. Investment managers are recommended to develop internal AI capabilities or engage specialized service providers for ongoing greenwashing monitoring, implement regular audits of sustainability claims in portfolio companies, and engage proactively with high-risk firms to encourage improvement in disclosure quality and substantive performance.

5.3. Corporate Governance Implications

From a corporate perspective, the research highlights the increasing sophistication of greenwashing detection capabilities and the growing costs of engaging in misleading sustainability communications. Companies are advised to implement rigorous internal controls for sustainability reporting, including comprehensive evidence documentation for all environmental claims, third-party assurance of key metrics, and systematic alignment between sustainability communications and actual operational practices.

Board-level oversight of ESG disclosures should be strengthened, with sustainability committees assuming responsibility for ensuring accuracy and consistency of environmental claims. Organizations might benefit from deploying internal AI-based verification systems prior to publication, identifying and correcting potential greenwashing risks before they reach external stakeholders. Such proactive approaches would reduce regulatory and reputational risks while strengthening stakeholder trust.

Limitations and Future Research

This study acknowledges several limitations that present opportunities for future research. First, the analysis focuses exclusively on UK-based companies, limiting generalizability to other jurisdictions with different regulatory frameworks and corporate governance structures. Comparative international studies examining greenwashing patterns across European Union, United States, and Asian markets would provide valuable insights into cultural and institutional determinants of corporate environmental communication strategies.

Second, the quantitative greenwashing measure based on ESG rating divergence relies on third-party ratings that themselves face criticism regarding consistency and methodology. Future research could develop alternative greenwashing measures incorporating direct environmental impact data, such as carbon emissions verified through independent sources, pollution incidents, and regulatory violations. Integration of satellite imagery analysis and Internet of Things sensor data might enable more objective assessment of environmental performance independent of corporate disclosures.

Third, the study's textual analysis focuses on formal sustainability reports and annual reports, excluding other communication channels including social media, press releases, advertising, and executive speeches. Research examining greenwashing across multiple communication platforms using multimodal AI approaches could reveal more comprehensive patterns of selective disclosure and attention manipulation. Additionally, temporal dynamics of greenwashing behavior warrant investigation through longitudinal studies tracking companies' evolution in response to regulatory pressure and stakeholder scrutiny.

Fourth, while the study demonstrates strong predictive performance, causal inference regarding drivers of greenwashing behavior requires careful consideration of endogeneity concerns and potential reverse causality. Experimental or quasi-experimental designs exploiting regulatory changes or exogenous shocks could strengthen causal claims regarding factors influencing greenwashing propensity. Future research might also examine the effectiveness of different interventions in reducing greenwashing, including regulatory enforcement actions, investor engagement, and public disclosure of greenwashing risk scores.

Finally, the rapid evolution of both AI technology and corporate communication strategies suggests need for ongoing model updates and validation. Research should investigate adversarial dynamics wherein firms adapt their disclosure strategies to evade detection, necessitating continuous refinement of AI models. Development of federated learning approaches enabling collaborative model improvement across regulatory jurisdictions while preserving data privacy represents a promising avenue for future investigation.

6. Conclusion

This research develops and evaluates a comprehensive artificial intelligence framework for detecting greenwashing in UK ESG and green finance projects, addressing critical challenges in sustainable finance regulation and investment. By integrating advanced Natural Language Processing techniques with optimized machine learning algorithms, the study demonstrates that automated greenwashing detection is both technically feasible and practically viable for regulatory and investment applications.

The findings reveal that transformer-based language models, particularly ClimateBERT, achieve high accuracy in identifying greenwashing language patterns in sustainability communications, while ensemble machine learning approaches effectively predict greenwashing risk based on quantitative financial and governance indicators. The integrated framework combining these complementary approaches provides robust, interpretable greenwashing assessments supporting diverse stakeholder needs.

The research contributes to sustainable finance literature by demonstrating practical implementation of AI-based regulatory compliance tools specifically tailored for the UK's anti-greenwashing regulatory framework. The methodology developed provides a replicable blueprint for other jurisdictions seeking to leverage technology for enhanced ESG market surveillance and investor protection. By establishing rigorous standards for AI-driven greenwashing detection, this work supports the broader transition toward transparent, accountable sustainable finance markets essential for achieving climate and sustainability objectives.

As artificial intelligence capabilities continue advancing and regulatory frameworks evolve, ongoing research and development of greenwashing detection systems will remain critical. The arms race between detection and evasion necessitates continuous innovation in AI methodologies, collaborative knowledge sharing among regulators and researchers, and adaptive regulatory approaches that keep pace with technological change. This study provides a foundation for such ongoing efforts, contributing to more trustworthy, effective sustainable finance ecosystems that genuinely support environmental and social progress.

References

[1] Ahmad, N., Mobarek, A. and Roni, N.N. (2023) 'Revisiting the impact of ESG on financial performance of FTSE350 UK firms: Static and dynamic panel data analysis', Cogent Business & Management, 10(2), pp. 1-23.

[2] Bernard W. and Matthew S. (2026). Impact Of Donor Agencies on the Improvement of Health Care Delivery Service. A Case Study of Bill and Melinda Gates Project (2000-2015). Journal of economics, finance and management studies, 09(01), 240–268.

[3] Delmas, M.A. and Burbano, V.C. (2011) 'The drivers of greenwashing', California Management Review, 54(1), pp. 64-87.

[4] Financial Conduct Authority (2024) FCA confirms anti-greenwashing guidance and proposes extending sustainability framework. Available at: https://www.fca.org.uk/news/press-releases/fca-confirms-anti-greenwashing-guidance.

[5] Financial Conduct Authority (2024) Sustainability disclosure and labelling regime. Available at: https://www.fca.org.uk/firms/climate-change-and-sustainable-finance/sustainability-disclosure-and-labelling-regime.

[6] Google Cloud (2023) '2023 Sustainability Survey', cited in CFA Institute (2024) A question of trust: How AI is addressing greenwashing concerns. Available at: https://www.cfainstitute.org/insights/articles/how-ai-combats-greenwashing.

[7] Gorovaia, N. and Grant, S. (2025) 'Identifying greenwashing in corporate-social responsibility reports using natural-language processing', European Financial Management, 31(2), pp. 456-489.

[8] Kim, J.S., Sim, J.B., Kim, Y.J., Park, M.K., Oh, S.J. and Doo, I.C. (2023) 'Establishment of NLP-based greenwashing pattern detection service', in Park, J.S., Yang, L.T., Pan, Y. and Park, J.H. (eds.) Advances in Computer Science and Ubiquitous Computing. Singapore: Springer, pp. 253-259.

[9] Krishna, N. (2025) 'AI-integrated ESG scoring system: Leveraging real-time green data intelligence for sustainable finance', SSRN Electronic Journal. Available at: https://ssrn.com/abstract=5764202

[10] Marquis, C., Toffel, M.W. and Zhou, Y. (2016) 'Scrutiny, norms, and selective disclosure: A global study of greenwashing', Organization Science, 27(2), pp. 483-504.

[11] Sari, I.M., Widagdo, A.K. and Lestari, H.S. (2025) 'Artificial intelligence-based ESG greenwashing detection: Road to net zero carbon and its impact on corporate performance', Business Strategy & Development, 8(1), e70228.

[12] Shankar, R. and Xu, Q. (2024) 'Automated detection of greenwashing in Indian corporate sustainability reports using natural language processing', SSRN Electronic Journal. Available at: https://ssrn.com/abstract=5294480

[13] Yu, E.P.Y., Van Luu, B. and Chen, C.H. (2020) 'Greenwashing in environmental, social and governance disclosures', Research in International Business and Finance, 52, 101192.

[14] Zeng, F., Wang, J. and Zeng, C. (2025) 'An optimized machine learning framework for predicting and interpreting corporate ESG greenwashing behavior', PLOS ONE, 20(3), e0316287.