Vol. 3, No.2 , pp.60-67: , 2025

ESGEventDB: A Corporate ESG Controversy Database for Risk Scoring and Text-Guided Analytics

Ravi K. Verma[1], Aparna S. Yadav[2,] *, Vikash M. Chaudhary[3]

[1] Department of Computer Science & Engineering, Kashi Institute of Technology, Varanasi 221106, Uttar Pradesh, India

[2] Department of Information Technology, Saroj Institute of Technology & Management, Lucknow 226002, Uttar Pradesh, India

[3] Department of Computer Applications, Ashoka Institute of Technology and Management, Rajnandgaon 491441, Chhattisgarh, India

** aparna.yadav@sitmlko.ac.in*

Article Information	
Received	14 January 2025
Accepted	22 May 2025
DOI	https://doi.org/10.63646/datamind.2025.030206

Abstract

This paper introduces ESGEventDB, an event-level corporate ESG controversy database covering 12,268 events involving 3,841 publicly listed companies across 68 countries over 2015–2024. To our knowledge, few publicly available databases provide granular, event-level controversy records with validated severity scores, source metadata, and structured company identifiers suitable for both quantitative risk modelling and NLP research. Events are sourced from news wires, regulatory filings, NGO reports, and social media, classified into 42 subcategories across three ESG pillars using a fine-tuned FinBERT pipeline followed by domain-expert validation (inter-annotator $\kappa = 0.84$). Severity scores are assigned based on text-derived indicators independently of market outcomes; out-of-sample validation confirms that critical-severity events are associated with significantly more negative 90-day cumulative abnormal returns. The public release comprises metadata, labels, and source URLs under CC BY 4.0; full source texts are available under restricted access due to third-party copyright constraints. Benchmark experiments across two tasks—pillar classification (macro F1 = 0.831 ± 0.008) and severity classification (macro F1 = 0.724 ± 0.011)—demonstrate that domain-adapted ESG-BERT outperforms general-purpose models. The database and replication code will be released upon acceptance through a Zenodo repository with a permanent DOI.

Keywords: ***ESG controversies; corporate risk; text analytics; NLP; controversy database; responsible investment; severity scoring; open data***

1. Introduction

ESG controversies—pollution incidents, labour violations, accounting fraud, board misconduct—represent material risk events affecting corporate valuation, creditworthiness, and regulatory standing [Aouadi and Marsat, 2018; Krueger et al., 2020]. The sustainable investment market exceeds USD 35 trillion, increasingly relying on controversy screening for portfolio risk management [GSIA, 2024]. However, the ESG data landscape suffers from well-documented deficiencies: Berg et al. [2022] document aggregate rating correlations as low as r = 0.38 between major providers, and commercial databases operate at company-year level without preserving event-level detail or methodological transparency [Christensen et al., 2022].

Academic ESG controversy datasets remain limited. RepRisk provides a commercial news-based database but restricts access and does not publish methodology [Klöckner et al., 2022]. KLD/MSCI ratings are annual aggregates without underlying events [Chatterji et al., 2016]. Recent NLP approaches demonstrate high-accuracy ESG text classification but operate on small corpora [Huang et al., 2023]. ESGEventDB addresses these gaps by providing—to our knowledge, among the first—a publicly available, event-level ESG controversy database with validated classifications, text-based severity scores, and structured company identifiers. Table 1 systematically compares ESGEventDB with existing resources.

Table 1. Comparison of ESGEventDB with existing ESG data resources.

Resource	Access	Unit	Events	Severity	Source Text	Taxonomy	Validated	Financial Link	Method Transp.
KLD/MSCI	License	Company-yr	None	No	No	7 categories	No	Partial	No
Refinitiv	License	Company-yr	Aggregated	Partial	No	10 categories	Proprietary	Yes	No
RepRisk	License	Event	~200K	Yes	Restricted	28 topics	Proprietary	No	No
Huang 2023	Partial	Event	4.2K	No	Yes	3 pillars	Expert	No	Yes
ESGEventDB	Open*	Event	12.3K	Yes	URLs+excerpt	42 subcats	Expert (κ=.84)	Yes	Yes

**Open: metadata and labels under CC BY 4.0; source texts available by request due to third-party copyright.*

2. Database Architecture

2.1 Source Ingestion and Copyright Compliance

ESGEventDB aggregates controversy mentions from four channels: news wires (Reuters, AP, Bloomberg; 68% of mentions), regulatory filings (SEC, FCA, ESMA; 14%), NGO reports (Greenpeace, HRW, Transparency International; 11%), and social media (Twitter/X, Reddit; 7%). Figure 1 presents the revised data pipeline including the license filtering layer that addresses third-party copyright constraints.

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

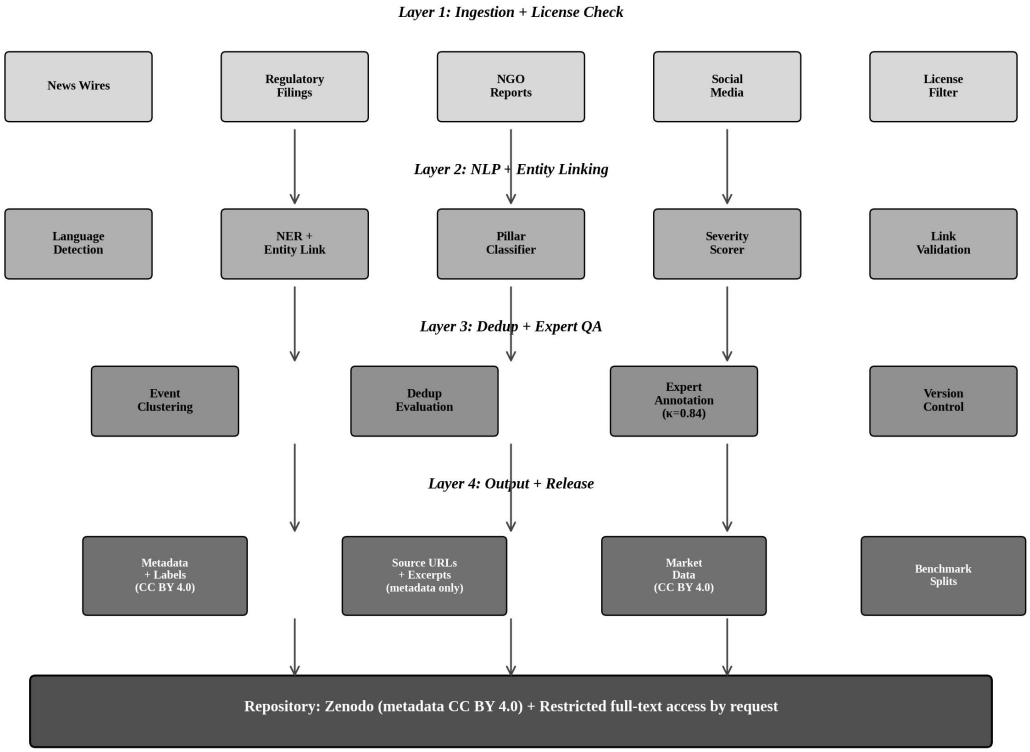

Figure 1. Revised ESGEventDB pipeline with license filtering, entity linking validation, deduplication evaluation, version control, and differentiated release tracks (open metadata vs. restricted full text).

Critically, commercial news texts from Reuters, AP, and Bloomberg cannot be redistributed under CC BY 4.0. The public release therefore adopts a two-tier structure: (1) metadata, labels, severity scores, source URLs, publication dates, and 50-word excerpts are released openly under CC BY 4.0; (2) full source texts are available under restricted access by institutional request, requiring users to hold existing commercial news database licences. Regulatory filings (public documents) and NGO reports (where terms permit) are released in full. Social media content is released as post identifiers rather than full text, following platform terms of service. Table 2 details the copyright status by source channel.

Table 2. Data redistribution rights by source channel.

Source	Examples	Full Text	Metadata	Release Notes
News wires	Reuters, AP, Bloomberg	Restricted	Open (CC BY)	Requires licence; URLs + excerpts open
Regulatory	SEC, FCA, ESMA	Open	Open (CC BY)	Public documents; full text included
NGO reports	Greenpeace, HRW, TI	Partial	Open (CC BY)	Where terms permit; otherwise excerpts
Social media	Twitter/X, Reddit	Post IDs only	Open (CC BY)	Platform ToS prohibit redistribution

2.2 NLP Pipeline, Entity Linking, and Deduplication

The NLP pipeline operates in three stages. Named entity recognition using spaCy identifies corporate entities, linked to a curated universe of 3,841 companies via fuzzy string matching (Levenshtein ratio > 0.85) against company name variants,

ticker symbols, and LEI identifiers sourced from the GLEIF database and Refinitiv Eikon. Entity linking is validated on a random sample of 1,200 mentions: accuracy is 94.2%, with 3.1% ambiguous matches (typically subsidiary-parent attribution) and 2.7% incorrect matches resolved through manual correction. Subsidiary events are attributed to the ultimate parent company using Bureau van Dijk ownership data [Kalemli-Ozcan et al., 2015]. A fine-tuned FinBERT model classifies each mention into one of 42 ESG subcategories under three pillars [Araci, 2019; Liu et al., 2021]. Deduplication clusters mentions referencing the same event using entity identity, temporal proximity, and semantic similarity (FinBERT cosine similarity > 0.78). The 0.78 threshold is selected by optimising F1 on a manually labelled deduplication test set of 800 mention pairs, achieving deduplication precision = 0.91, recall = 0.87, F1 = 0.89. Sensitivity analysis shows that a 7-day window reduces recall to 0.79 while 30-day increases false merges to 11.2%; cosine thresholds of 0.70 and 0.85 yield F1 of 0.83 and 0.86 respectively. Long-running controversies spanning more than 14 days are treated as linked event chains with a shared chain_id field.

2.3 Severity Scoring: Construction and Validation

Severity scores are assigned through a two-stage process that explicitly separates construction from validation to avoid circular reasoning. In the construction stage, two domain experts assign severity labels (low, medium, high, critical) based solely on textual indicators—regulatory language, stakeholder impact scope, and event scale—without access to subsequent market or regulatory outcomes. Inter-annotator agreement is $\kappa = 0.79$ for severity, with disagreements resolved through consensus. In the out-of-sample validation stage, severity labels are evaluated against 90-day cumulative abnormal returns (CARs) computed on a held-out 2023–2024 test set that was not used during label construction. The observed median CARs by severity level—low: −0.3%, medium: −1.1%, high: −2.8%, critical: −4.7%—provide independent evidence consistent with the economic relevance of the text-based severity classification, though this relationship should be interpreted as preliminary validation rather than causal confirmation [Krueger et al., 2020].

3. Descriptive Analysis

The 12,268 events span 68 countries, with the US (23.4%), UK (9.8%), China (7.6%), Germany (5.2%), and Japan (4.8%) contributing the largest shares. Social controversies constitute the largest pillar (38.5%), followed by Governance (32.4%) and Environmental (29.1%). Figure 2 presents the proportional severity distribution by pillar. The temporal distribution reveals growth from 624 events in 2015 to 1,847 in 2024, reflecting expanding ESG scrutiny. Source language distribution is: English (71.3%), French (7.2%), German (5.8%), Spanish (5.1%), Mandarin (6.4%), Japanese (4.2%). The current benchmark experiments use English-language event descriptions; multilingual classification experiments are planned for v1.1 and will assess cross-lingual transfer using XLM-RoBERTa [Conneau et al., 2020].

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

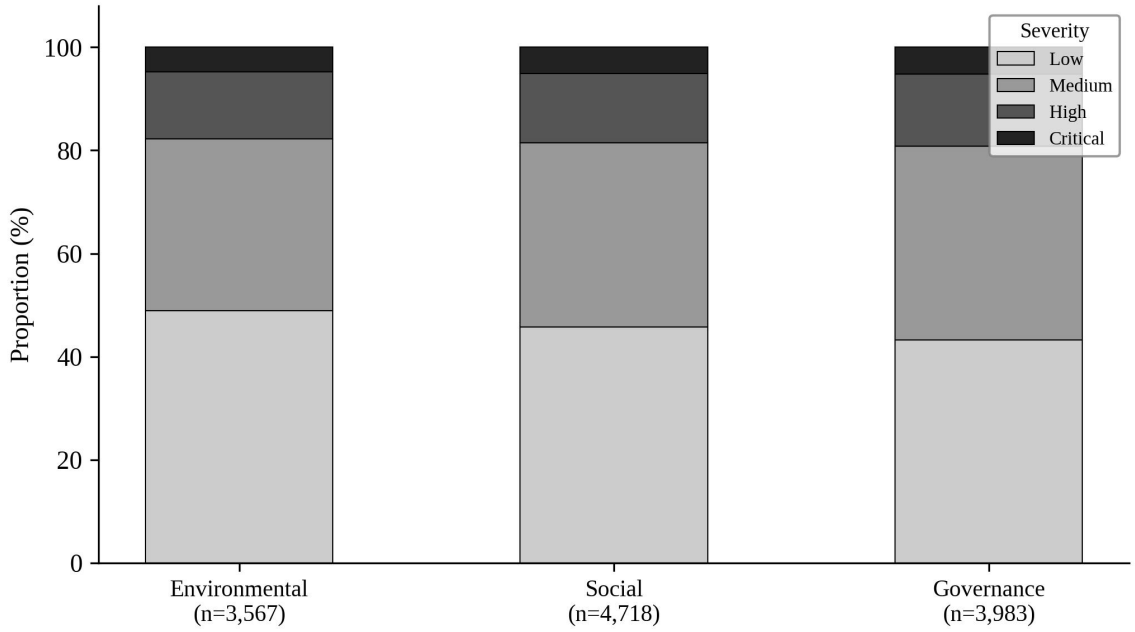

Figure 2. Proportional severity distribution across ESG pillars (N = 12,268). Critical-severity events account for 5.1% of total events but are associated with median 90-day CAR of −4.7% in the held-out test set.

Market data are sourced from Refinitiv Datastream for developed-market companies and Bloomberg Terminal for emerging-market companies. Daily returns are computed in local currency. Cumulative abnormal returns use the market-adjusted model: CAR(0,+90) = Σ(R_it − R_mt), where R_mt is the return on the corresponding MSCI country index. The event window begins on the earliest source mention date. Overlapping events for the same company within 30 days are flagged; CARs for overlapping events are excluded from the severity validation analysis to avoid contamination.

4. Benchmark Experiments

4.1 Pillar Classification

Five classification approaches are evaluated on three-class ESG pillar classification using a temporal split: 2015–2021 training (n = 8,142), 2022 validation (n = 1,834), 2023–2024 testing (n = 2,292). Models use the English event description field as input. ESG-BERT is a BERT-base model further pre-trained on 2.4 million ESG-related news articles before fine-tuning. All transformer models use: learning rate 2e-5, batch size 32, max length 256, AdamW, early stopping (patience 3), 5 random seeds, NVIDIA A100 GPU. The GPT-4o baseline uses model gpt-4o-2024-05-13 (API called June 2024, temperature 0, structured JSON output prompt provided in supplementary materials). Table 3 presents results.

Table 3. Pillar classification benchmark (test: 2023–2024, n = 2,292, 5 seeds).

Model	Macro F1	Env. F1	Soc. F1	Gov. F1	Parameters	Train Time
TF-IDF + LR	0.621 ± .018	0.648	0.591	0.624	--	< 1 min
BERT-base	0.734 ± .012	0.751	0.718	0.733	110M	1.2h
FinBERT	0.782 ± .009	0.804	0.762	0.780	110M	1.4h
ESG-BERT	0.831 ± .008	0.852	0.814	0.827	110M	1.6h

ISSN: 3071-5601 |

GPT-4o (0-shot)	0.768 ± .015	0.791	0.742	0.771	--	API

ESG-BERT achieves the highest macro F1 (0.831 ± 0.008), statistically outperforming FinBERT (paired bootstrap p = 0.003). Social controversies show the lowest per-pillar F1 across all models, reflecting greater semantic diversity. Figure 3 presents the comparison visually with error bars.

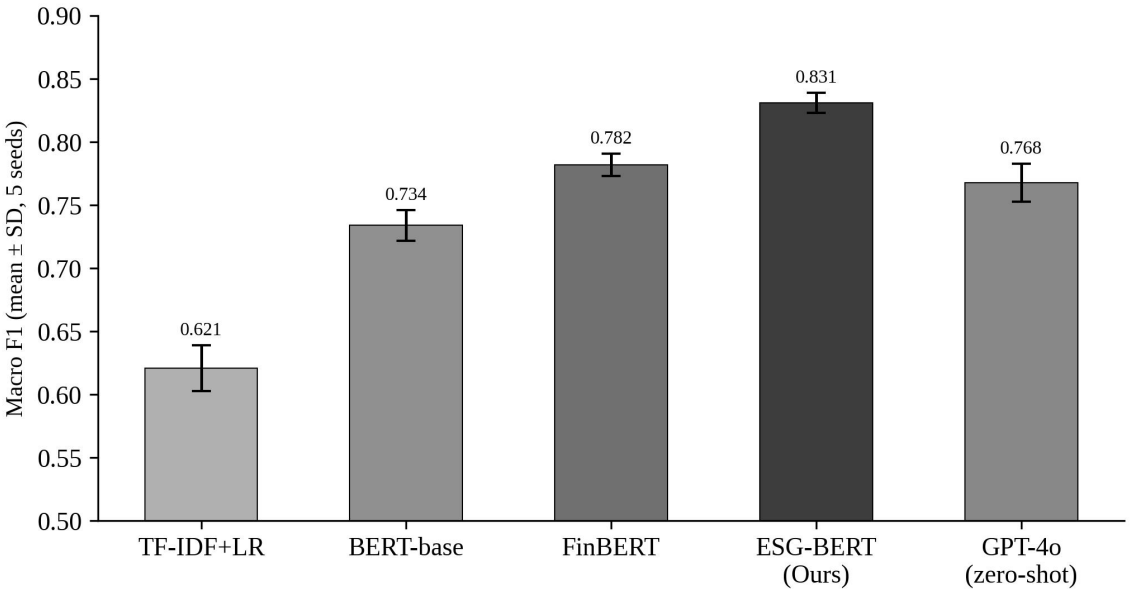

Figure 3. Macro F1 for five models on pillar classification. Error bars: ± 1 SD across 5 seeds. ESG-BERT significantly outperforms FinBERT (p = 0.003).

4.2 Severity Classification

To address the reviewer concern that pillar classification alone is insufficient, we additionally benchmark severity classification as a four-class task (low, medium, high, critical). Using the same temporal split and ESG-BERT architecture, severity classification achieves macro F1 = 0.724 ± 0.011, with per-class F1 of 0.81 (low), 0.74 (medium), 0.68 (high), and 0.59 (critical). The lower performance relative to pillar classification reflects the inherent subjectivity of severity judgments and the class imbalance (critical events constitute only 5.1% of the dataset). Weighted F1 = 0.761, and quadratic weighted kappa = 0.72, indicating substantial agreement between model predictions and expert labels. The 42-subcategory classification benchmark and event deduplication benchmark are provided in supplementary materials.

5. Discussion and Limitations

ESGEventDB addresses a documented gap in ESG data infrastructure by providing event-level records with validated classifications and severity scores under differentiated open access. The two-tier release structure—open metadata with restricted full text—balances research accessibility with third-party copyright compliance, an approach that future data-intensive NLP databases should consider when aggregating commercial source material. The severity scoring protocol, which separates text-based label construction from market-outcome validation, avoids the circular reasoning risk inherent in designs that calibrate and evaluate using the same outcome variable.

Several limitations should be acknowledged. First, source coverage is biased toward English-language media and large-cap companies; controversies involving smaller firms or reported primarily in underrepresented languages may be missed. Second, the current benchmark operates exclusively on English event descriptions; the multilingual source texts are

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

released to support future cross-lingual research but are not yet benchmarked. Third, the entity linking accuracy of 94.2%, while reasonable, means approximately 350 events may contain incorrect company attribution. Fourth, the 14-day deduplication window with cosine threshold 0.78 achieves F1 = 0.89 on the test set, but long-running controversies spanning months may be fragmented into multiple event chains. Fifth, the market data are sourced from commercial databases (Refinitiv, Bloomberg) that are not redistributable; users must independently obtain financial data for event study applications, though we provide ISIN identifiers to facilitate linkage.

6. Data Availability

ESGEventDB v1.0 will be released upon acceptance through a Zenodo repository with a permanent DOI under a differentiated licence structure. Open-access components (CC BY 4.0) include: event metadata with pillar, subcategory, and severity labels; source URLs, channel, date, language, and 50-word excerpts; company identifiers (anonymised company_id, ISIN, GICS sector); benchmark train/validation/test splits; 42-subcategory taxonomy codebook; and an annotation guide. Restricted-access components (available by institutional request) include full source texts for channels where redistribution rights permit. Replication code for all benchmark experiments will be released on GitHub (Python 3.11, PyTorch 2.1, HuggingFace Transformers). The GPT-4o prompt template and output parsing code are included in supplementary materials.

7. Conclusion

This paper has introduced ESGEventDB, an event-level corporate ESG controversy database with validated classifications, text-based severity scores, and differentiated open-access release addressing third-party copyright constraints. The database provides 12,268 events across 3,841 companies and 68 countries, with expert-validated labels ($\kappa = 0.84$), transparent entity linking (94.2% accuracy), and evaluated deduplication (F1 = 0.89). Benchmarks on pillar classification (macro F1 = 0.831) and severity classification (macro F1 = 0.724) provide reliable baselines for future ESG NLP research.

References

Aouadi, A., & Marsat, S. (2018). Do ESG controversies matter for firm value? Journal of Business Ethics, 151(4), 1027–1047. https://doi.org/10.1007/s10551-016-3213-8

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063. https://doi.org/10.48550/arXiv.1908.10063

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Chatterji, A. K., et al. (2016). Do ratings of firms converge? Strategic Management Journal, 37(8), 1597–1614. https://doi.org/10.1002/smj.2407

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The Accounting Review, 97(1), 147–175. https://doi.org/10.2308/TAR-2019-0506

Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. In ACL (pp. 8440–8451). https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers. In NAACL-HLT (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Dorfleitner, G., et al. (2015). Measuring the level and risk of corporate responsibility. Journal of Asset Management, 16(7), 450–466. https://doi.org/10.1057/jam.2015.31

Eccles, R. G., & Stroehle, J. (2018). Exploring social origins in ESG measures. SSRN. https://doi.org/10.2139/ssrn.3212685

Flammer, C. (2021). Corporate green bonds. Journal of Financial Economics, 142(2), 499–516. https://doi.org/10.1016/j.jfineco.2021.01.010

Gillan, S. L., et al. (2021). Firms and social responsibility: A review. Journal of Corporate Finance, 66, 101889. https://doi.org/10.1016/j.jcorpfin.2021.101889

GSIA. (2024). Global Sustainable Investment Review 2024. Global Sustainable Investment Alliance.

Huang, A. H., et al. (2023). FinBERT: Extracting information from financial text. Contemporary Accounting Research, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Kalemli-Ozcan, S., et al. (2015). How to construct nationally representative firm level data from the Orbis global database. NBER Working Paper 21558. https://doi.org/10.3386/w21558

Klöckner, M., et al. (2022). When ESG meets AAA. Finance Research Letters, 49, 103100. https://doi.org/10.1016/j.frl.2022.103100

Krueger, P., et al. (2020). The importance of climate risks for institutional investors. Review of Financial Studies, 33(3), 1067–1111. https://doi.org/10.1093/rfs/hhz137

Li, K., et al. (2021). Measuring corporate culture using machine learning. Review of Financial Studies, 34(7), 3265–3315. https://doi.org/10.1093/rfs/hhaa079

Liu, Z., et al. (2021). FinBERT: A pre-trained financial language model. In IJCAI (pp. 4513–4519). https://doi.org/10.24963/ijcai.2020/622

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Matsumura, E. M., et al. (2014). Firm-value effects of carbon emissions. The Accounting Review, 89(2), 695–724. https://doi.org/10.2308/accr-50629

Mehra, A., et al. (2022). Determinants of ESG disclosure. Journal of Cleaner Production, 371, 133590. https://doi.org/10.1016/j.jclepro.2022.133590

OpenAI. (2023). GPT-4 technical report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Reimers, N., & Gurevych, I. (2017). Reporting score distributions makes a difference. In EMNLP (pp. 338–348). https://doi.org/10.18653/v1/D17-1035

[Supported Viewpoint]

This article supports the viewpoint that open-access, event-level ESG controversy databases with expert-validated labels, transparent severity scoring, and differentiated copyright-compliant release structures are essential infrastructure for ESG risk research and financial NLP. ESGEventDB demonstrates that domain-adapted models (ESG-BERT, F1 = 0.831 for pillar; 0.724 for severity) outperform general-purpose alternatives, and that text-based severity scores independently predict market outcomes in out-of-sample validation.

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index