Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Say-Do Gap Index: A Reproducible NLP Pipeline for Detecting ESG Rhetoric–Reality Divergence in Corporate Media Coverage

Alfredo Merlet

alfredo.merlet@gmail.com

University of Seville https://orcid.org/0009-0007-8558-164X

Method Article

Keywords: NLP pipeline, BETO, BERT, ESG classification, greenwashing detection, GDELT, reproducible research, text-as-data, say-do gap, natural language processing

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10757614/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The author is affiliated with two institutions — Universidad del Alba and Instituto Profesional IACC — that provide publication incentives conditioned on research output. This constitutes a potential conflict of interest, disclosed here in accordance with the author's institutional policies.

Abstract

This paper documents a reproducible natural language processing (NLP) pipeline for detecting divergence between corporate ESG communication and observable outcomes — a "say-do gap" — using large-scale media coverage as an external verification signal. Rather than reporting the substantive findings of an empirical application, which are presented separately (Merlet Echavarría, under review), this paper treats the pipeline itself as the object of contribution: a fully documented, reusable architecture spanning corpus extraction (GDELT/BigQuery), a six-category ESG behavior taxonomy, a targeted annotation enrichment protocol, and a fine-tuned BETO (Spanish BERT) classifier achieving Macro F1 = 0.76. The pipeline was developed and validated on 131,151 news articles covering 34 constituent firms of the IBEX 35 index (2019–2025), yielding a firm-year Say-Do Gap Index. We discuss which components of the architecture are likely market- and language-agnostic versus which require recalibration for other institutional or linguistic contexts, and we flag construct-validity concerns in interpreting textual discrepancies as behavioral accountability gaps rather than legitimate strategic adjustment. All code, the annotation protocol, and classifier training scripts are deposited at Zenodo under a CC-BY 4.0 license.

[VERIFICAR ANTES DE ENVIAR: *Borrador de abstract generado a partir de tus respuestas sobre Introducción/Discusión — revisa que el framing (herramienta reutilizable + foco en construct validity) refleje lo que quieres priorizar antes de enviar.***]**

1. Introduction

The paper "Say-Do Gap in the Age of AI" (Merlet Echavarría, under review at Meditari Accountancy Research) uses the pipeline described here to answer a substantive empirical question: whether externally verified ESG behavior — as detected through media coverage — is associated with firm value in a sample of IBEX 35 constituents. That paper's contribution is the empirical answer. This paper's contribution is different: it treats the pipeline itself as a reusable research infrastructure, and documents in full the methodological decisions that a journal-length empirical paper necessarily compresses — corpus extraction and entity resolution, the annotation protocol and its iterative correction, the classifier's development trajectory, and the validation procedures behind the resulting index.

This distinction matters because the pipeline is not specific to the empirical question it was first built to answer. Its architecture — extracting firm-level media coverage, classifying it against a disclosure-behavior taxonomy, and constructing a firm-year discrepancy index — is directly transferable to other markets, time periods, firms, languages, and related research questions involving disclosure, earnings-call rhetoric, or other forms of divergence between what firms say and what they subsequently do. Publishing it as a standalone methodological contribution allows other researchers to adopt, adapt, or critique the pipeline independently of the specific empirical claim it was used to test.

The paper is written for an interdisciplinary audience at the intersection of computational accounting and NLP/data science. It is deliberately not framed as a pure NLP contribution, since the value of the

classification task depends on the accounting and financial disclosure problem it is designed to measure; nor is it framed purely as an accounting methods paper, since a substantial part of the contribution lies in how the textual analysis is operationalized and automated. The intended readership includes accounting, finance, and corporate disclosure researchers seeking to incorporate computational methods into their own work, as well as NLP and data science researchers interested in reproducible applications in financial and disclosure domains.

2. Related Work

This section positions the pipeline within three literatures: automated greenwashing and say-do gap detection using NLP, BERT-family classifiers applied to Spanish-language text, and open, reproducible NLP pipelines in accounting and sustainability research.

Textual-analysis approaches to greenwashing detection have used NLP to compare disclosure content and tone against verifiable outcomes; Gorovaia and Makrominas (2025) show that firms with environmental violations systematically alter the length, positivity, and readability of their CSR reporting relative to firms without violations, evidencing a detectable rhetoric-reality gap in disclosure text itself. More automated, model-based approaches have since extended this line of work: Zeng, Wang, and Zeng (2025) combine gradient-boosted classification with explainable AI methods to predict and interpret ESG greenwashing behavior directly from corporate text and disclosure features, and Calamai, Balalau, Le Guenedal, and Suchanek (2025) provide a comprehensive methodological survey of the NLP tasks used across this literature — climate-topic detection, green-claim extraction, and deceptive-communication classification — situating the present pipeline's taxonomy-based classification approach within that broader task decomposition.

The classifier underlying this pipeline is a fine-tuned version of BETO, a BERT-family language model pre-trained exclusively on Spanish-language corpora (Cañete, Chaperon, Fuentes, Ho, Kang, & Pérez, 2020), selected over multilingual alternatives given the pipeline's exclusive focus on Spanish-language media coverage of IBEX 35 firms.

3. Data: Corpus Construction

3.1 Source and Extraction

The corpus is built from GDELT's Global Knowledge Graph, queried via Google BigQuery, restricted to press coverage of firms constituting the IBEX 35 index (Spain). Firm identification uses URL-slug matching rather than GDELT's native V2Organizations field, which was found to corrupt short acronyms (e.g., "BBVA") during entity resolution.

Extraction was performed across two billing periods to remain within BigQuery Sandbox free-tier quotas: July 2022–2025 (~ 80,000 articles) and August 2019–2021 (~ 51,000 articles, extracted month-by-

month). The merged, deduplicated corpus totals 131,151 articles across 34 of the 35 IBEX 35 constituent firms.

3.2 Taxonomy

A six-category ESG behavior taxonomy was operationalized following Walker and Wan (2012)'s substantive/symbolic distinction, adapted for media-coverage classification: (1) substantive ESG action, (2) unverified/aspirational commitment, (3) ambiguous disclosure/greenwashing signal, (4) negative ESG controversy, (5) informative/neutral ESG mention, and (6) non-ESG noise (approx. 26% of raw corpus).

4. Annotation Protocol

A stratified seed sample of 700 articles was manually annotated. Because category 2/3 (unverified commitment) was underrepresented in the initial stratified draw, a targeted keyword-search enrichment procedure was applied to increase representation before training. A critical train–inference format inconsistency was identified during development: the classifier had learned input length as a spurious proxy for categories 2/3, producing a 15.6% over-prediction rate against an expected ~ 2.3%. Standardizing text format (URL-slug only) across training and inference corrected this to 2.4%.

5. Classifier Development and Validation

The taxonomy classifier is a fine-tuned BETO model (Spanish BERT), trained iteratively across three rounds. The first round (v1, six separate categories) yielded Macro F1 = 0.49, with categories 2 (aspirational) and 3 (ambiguous/greenwashing) effectively uncaptured (F1 = 0.00) due to severe under-representation (2–3 validation cases). Categories 2 and 3 were merged into a single "unverified commitment" class — consistent with the Say-Do Gap Index formula, which already treats them identically — yielding v2 (Macro F1 = 0.59), still limited by only 16 seed examples for the merged class. A targeted keyword-search enrichment procedure increased training examples for this class, yielding the final model (v3): Macro F1 = 0.76, with the merged category improving from F1 = 0.00 to F1 = 0.73, and all six categories reaching usable performance (range 0.63–0.91).

Manual precision-checking on a validation sample of predictions for the critical "unverified commitment" category yielded approximately 56% precision, reported transparently as a known limitation of this category specifically rather than of the classifier overall.

6. Index Construction

The firm-year Say-Do Gap Index is computed as:

$$\textit{Say-Do Gap Index} = (N_{2/3} - N_1 - 2 \times N_4) / N_ESG$$

where N_1, $N_2/_3$, and N_4 are firm-year article counts in the substantive, unverified/ambiguous, and controversy categories respectively, and N_ESG is the total count of ESG-relevant articles for that firm-year. The resulting panel covers 238 firm-year observations (34 firms × 7 years, 2019–2025).

7. Reproducibility and Availability

The full pipeline — extraction scripts, annotation tooling, and classifier training code — is deposited at Zenodo under a CC-BY 4.0 license, as a dedicated deposit distinct from the replication package accompanying the empirical application of this pipeline (Merlet Echavarría, under review at Meditari Accountancy Research). This paper's deposit excludes the firm-year financial panel and econometric analysis scripts, which are specific to that empirical study. Raw article text is not redistributed; all source URLs are publicly accessible press URLs, and the corpus is fully reproducible from GDELT's public dataset via the included extraction scripts. Generative AI (Claude, Anthropic) assisted with organizing prose and verifying citations for this manuscript; see Declarations for the full disclosure.

[VERIFICAR ANTES DE ENVIAR: *Falta crear el Zenodo deposit nuevo (solo pipeline: extracción + anotación + clasificador, sin panel financiero ni script econométrico) e insertar aquí su DOI real cuando lo tengamos — dime cuándo quieres que lo armemos.***]**

8. Discussion: Generalizability and Limitations

8.1 Generalizability

The pipeline's architecture — corpus extraction, taxonomy-based classification, and index construction — is expected to generalize reasonably well to other stock market contexts, provided that comparable textual sources (news coverage, disclosure documents, or transcripts) and structured financial data are available. The overall architecture should transfer with limited modification, though specific extraction rules and data sources would need recalibration for each new market.

Generalization across languages is a separate and more delicate question. While the methodological structure itself is transferable, the linguistic components — embeddings, claim extraction, semantic classification, and the identification of commitments or promises — are more likely to lose precision outside Spanish-language contexts and would likely require language-specific models rather than a direct transfer of the current BETO classifier.

We therefore distinguish between architectural generalizability (the pipeline's structure and logic, which travels well) and empirical generalizability (the specific parameters, trained models, and classification thresholds, which do not necessarily travel with it). A natural extension of this work would be to identify explicitly which components of the architecture are market- and language-agnostic and which require context-specific adaptation — a question we consider open rather than resolved by the present pipeline.

8.2 Limitations

Beyond the technical limitations already documented (Sections 4–5), the most consequential limitation concerns the construct validity of the Say-Do Gap Index itself: whether an observed divergence between a corporate statement and a subsequent outcome genuinely reflects a gap between what a firm said it would do and what it actually did. A firm may fail to fulfill a stated commitment for reasons unrelated to misrepresentation — external shocks, regulatory change, macroeconomic conditions, or legitimate strategic revision — without this reflecting deceptive or inconsistent conduct.

The pipeline, as designed, measures an observable textual and outcome-based discrepancy with reasonable reliability. Interpreting that discrepancy as a behavioral or accountability gap is a separate inferential step that requires caution and is not resolved by improving the underlying NLP components. This is fundamentally a question of construct validity and causal inference rather than classification accuracy, and we consider it one of the more consequential open problems for future work building on this pipeline.

9. Conclusion

The contribution of this paper is not simply that the Say-Do Gap can be measured, but that it can be operationalized through a fully reproducible, reusable pipeline. What is documented here is a piece of open methodological infrastructure: the corpus extraction logic, the annotation protocol, and the classifier training procedure can be adapted by other researchers to other markets, firms, time periods, and — with appropriate recalibration — other languages, extending the study of say-do gaps well beyond the specific empirical application this pipeline was first built to serve.

At the same time, measuring an observable discrepancy is not the same as explaining why it occurred, and it is not, on its own, evidence of misconduct. The pipeline's strength lies in identifying and systematizing the phenomenon at scale; its substantive interpretation still requires theoretical and contextual judgment, applied case by case. In short, this pipeline offers a way to scale and reproduce the study of the Say-Do Gap, but it does not eliminate the need for careful, context-sensitive interpretation of what any given gap actually means.

Declarations

Conflict of Interest

The author is affiliated with two institutions — Universidad del Alba and Instituto Profesional IACC — that provide publication incentives conditioned on research output. This constitutes a potential conflict of interest, disclosed here in accordance with the author's institutional policies.

[VERIFICAR ANTES DE ENVIAR: *Esta es la versión genérica no identificante (correcta para el manuscrito anónimo si el destino usa doble-ciego). Si Research Square u otro portal pide además una versión con nombres de institución explícitos en un campo separado de comentarios al editor, dime y te la preparo aparte — nunca debe decir "no conflict of interest" mientras declares estas afiliaciones.***]**

AI Use Disclosure

Generative AI (Claude, Anthropic) was used to assist with organizing and translating the author's own arguments into academic prose for the Introduction and Discussion sections, drafting the technical Methods/Results sections based on verified project records, and conducting literature verification for citations. All substantive methodological decisions, the underlying pipeline, and its empirical validation are the author's own work.

Data Availability

Replication materials are available at Zenodo (DOI to be confirmed per note above).

References

1. Calamai T, Balalau O, Le Guenedal T, Suchanek FM (2025) Corporate greenwashing detection in text — a survey. https://doi.org/10.48550/arXiv.2502.07541. arXiv:2502.07541
2. Cañete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J (2020) Spanish pre-trained BERT model and evaluation data. PML4DC at ICLR 2020
3. Gorovaia N, Makrominas M (2025) Identifying greenwashing in corporate-social responsibility reports using natural-language processing. Eur Financ Manag. https://doi.org/10.1111/eufm.12509
4. Merlet Echavarría AA (under review). Say-Do Gap in the Age of AI: External Media Verification of ESG Disclosure Quality and Firm Value in IBEX 35 Firms. Meditari Accountancy Research
5. Walker K, Wan F (2012) The harm of symbolic actions and green-washing: Corporate actions and communications on environmental performance and their financial implications. J Bus Ethics 109(2):227–242
6. Zeng F, Wang J, Zeng C (2025) An optimized machine learning framework for predicting and interpreting corporate ESG greenwashing behavior. PLoS ONE. https://doi.org/10.1371/journal.pone.0316287

[VERIFICAR ANTES DE ENVIAR Lista de referencias completa según lo citado en el texto (Related Work + Taxonomía). Verificar formato exigido por el portal de destino (Research Square no exige un estilo único, pero si luego este paper migra a una revista específica, puede requerir Harvard/APA/Vancouver según la revista).]