

Topic-Aware Environmental Claim Extraction from Corporate News

Tanay Kumar Gupta, Vaishali Aggarwal, Ishan Verma

TCS Research

New Delhi, India

{gupta.tanay, aggarwal.vaishali, ishan.verma}@tcs.com

Abstract

Corporate environmental claims frequently appear in news articles, providing an external view of sustainability-related communications. However, extracting such claims remains challenging as they are often implicit, distributed across multiple sentences, and embedded within narrative contexts. In this work, we propose an end-to-end system for automated environmental claim extraction enriched with contextual metadata. Our pipeline combines multi-label topic classification, topic-conditioned candidate claim-evidence extraction using large language models (LLMs), and candidate filtering with ASE-BERT, a lightweight aspect- and sentiment-aware classifier for claim detection. To support evaluation, we construct a curated dataset of corporate environmental news, annotated with claims, supporting evidence, and contextual attributes. Experimental results demonstrate that for claim classification, ASE-BERT attains the highest F1 (0.87) across the four combined environmental claim datasets, exceeding the best baseline without task-specific fine-tuning by 4 points. For end-to-end extraction, topic conditioning alone improves LLM F1 by an average of 15 points, and our full pipeline adds a further 5 points over the best topic-conditioned setting. These findings demonstrate that incorporating topic awareness and claim filtering substantially improves environmental claim extraction from corporate news, supporting downstream sustainability intelligence and decision-making tasks.

1 Introduction

The increasing emphasis on environmental sustainability has intensified scrutiny of how companies communicate their environmental impact. Corporate environmental claims that suggest or imply environmentally beneficial actions or outcomes play a central role in shaping public perception. News media amplifies these claims, influencing how investors, regulators, and other stakeholders inter-

pret corporate sustainability performance and ESG (Environmental, Social and Governance) related risks (Carroll and McCombs, 2003; Vogler and Eisenegger, 2021).

Despite the growing volume of environment-related news, automatically identifying and analyzing environmental claims remains challenging. Unlike structured disclosures, claims in news narratives are often context-dependent, may span multiple sentences, and involve implicit attribution to entities such as companies, regulators, or third parties (Reddy et al., 2022). Moreover, claims frequently combine evaluative language with factual descriptions, making them difficult to distinguish from general environmental statements. As illustrated in Figure 1, interpreting such claims often requires integrating cross-sentence context, attribution cues, and topic-specific information.

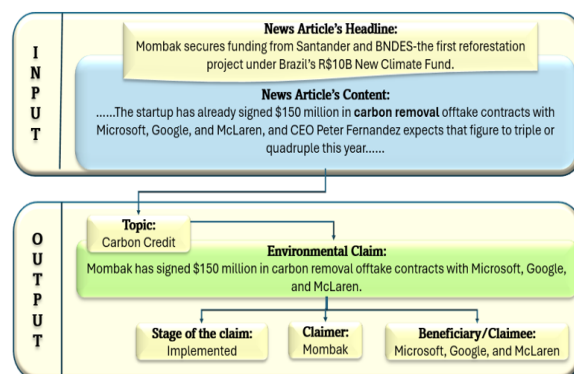


Figure 1: Example illustrating cross-sentence dependencies and contextual attributes in environmental claims.

In line with regulatory guidelines and prior literature (Stammbach et al., 2023), we define an environmental claim as *a statement that suggests or creates the impression that a product, service, process, or corporate action is environmentally friendly, has a positive environmental impact, or is less harmful than competing alternatives*. Although some claims reflect genuine sustainability

initiatives, the broader regulatory and academic discourse highlights increasing concerns around greenwashing (de Freitas Netto et al., 2020). This motivates the need for scalable methods that can identify, contextualize, and organize environmental claims for downstream ESG analysis, risk assessment, disclosure verification, and accountability.

Prior work has explored claim extraction in sustainability reports (Mahdavi et al., 2024; Ong et al., 2025) and analyzed ESG signals in news (Tseng et al., 2023; Gupta et al., 2024). However, existing datasets in climate communication, such as Climate Fever (Diggelmann et al., 2020), ClimateCheck (Ahmad et al., 2025), and ClimateMiSt (Choi et al., 2024) primarily focus on misinformation or social media and do not model corporate environmental claims in news or their topic-specific structure. The closest news-oriented benchmark, NewsClaims (Reddy et al., 2022), provides claimer and claim object attribution but is domain-general (largely COVID-19) rather than environmentally grounded. This creates a gap between general purpose claim verification and the requirements of analyzing market-facing corporate environmental narratives for ESG intelligence and decision support.

To address this gap, we propose a topic-conditioned framework for extracting and classifying environmental claims from news articles. Our design rests on two observations: (i) environmental claims are inherently topic-dependent, requiring grounding in environmental domains, and (ii) they exhibit an evaluative linguistic signal that can be captured through sentiment-aware representations. Building on these observations, our system first assigns document-level environmental topics using a multi-label classification module. It then performs topic-conditioned candidate claim-evidence pair extraction using an LLM. These candidate claims are subsequently classified using ASE-BERT, a lightweight aspect and sentiment aware transformer designed to distinguish environmental claims from non-environmental claims. In addition, the system enriches extracted claims with contextual metadata, including the identifying entity (claimer), affected party (beneficiary), and claim stage, enabling structured analysis. To support this task, we also construct a curated, expert-annotated dataset¹ from corporate environmental news articles. The dataset includes labeled claims, supporting evidence, topic

annotations, and relational metadata, providing a structured corpus for evaluating environmental claim extraction systems. The resulting framework converts unstructured environmental narratives into structured, attributable claims, supporting downstream tasks such as ESG intelligence, risk assessment, and disclosure verification. Our main contributions are:

- We propose a topic-conditioned system for environmental claim extraction that models claims as context-dependent and evaluative linguistic constructs.
- We introduce ASE-BERT, a lightweight classifier that integrates topic information and sentiment signals for improved claim detection.
- We construct an expert-annotated dataset of corporate environmental news with claims, evidence spans, topics, and contextual metadata.
- We demonstrate that the proposed approach improves performance over strong transformer and LLM-based baselines for both classification and end-to-end extraction, highlighting the importance of topic conditioning and sentiment-aware modeling.

2 System Architecture

Figure 2 illustrates the end-to-end system for extracting and analyzing environmental claims from ESG-related news. The design is motivated by the observation that environmental claims are both *topic-dependent* and *evaluative*, and therefore require modeling beyond sentence-level classification. To this end, the pipeline decomposes the task into topic grounding, structured extraction, and claim filtering stages.

Given a news article A_i , the system first assigns document-level topics $\mathcal{T} = \{T_1, \dots, T_K\}$. Conditioned on $\mathcal{T}$, it then extracts a set of candidate claim-evidence pairs $C_i = \{(x, e)\}$, where x is a candidate claim and e its supporting evidence span in A_i . Extracted candidate claims x are then filtered using an aspect and sentiment-aware classifier (ASE-BERT). Filtered claims are later enriched with contextual metadata. This decomposition reduces extraction ambiguity and improves alignment with environmental context.

¹<https://github.com/datasharerepo/EnvClaim-TA>

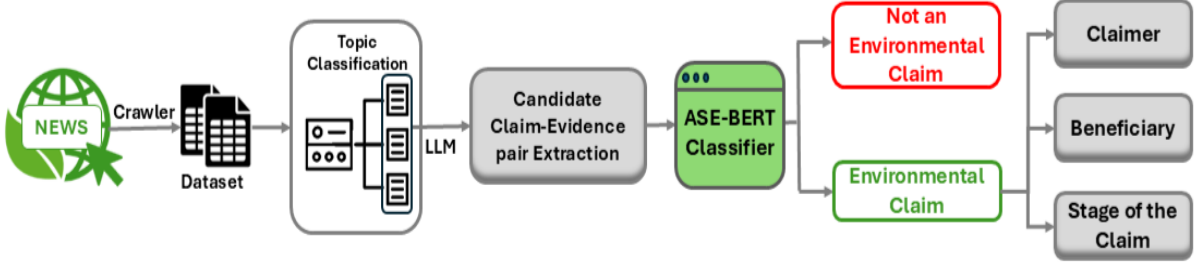


Figure 2: System Architecture

2.1 Multi-label Topic Classification of Articles

Long-form news articles often span multiple environmental themes. We assign each article to one or more topics from a predefined taxonomy, *e.g.* *Energy, Waste, Water, Sustainable Investment, Biodiversity, etc.*, using a prompt-based LLM classifier with deterministic decoding to minimize variance and ensure reproducibility. To improve consistency, the prompt incorporates curated keyword cues and concise semantic definitions for each topic, enabling the model to better distinguish between closely related environmental categories. The predicted topics serve two purposes: (i) constraining downstream candidate claim-evidence extraction, and (ii) providing aspect input to the ASE-BERT classifier.

Validation: To assess reliability, we evaluated the topic classifier (Llama 70B (Grattafiori et al., 2024)) on 20 randomly sampled articles, with a human reviewer and GPT 5.5 model (OpenAI, 2026) serving as judges. Table 1 reports the corresponding Precision, Recall, and F1 under micro, macro, and weighted averaging.

Evaluation Type	Micro			Macro			Weighted		
	P	R	F1	P	R	F1	P	R	F1
Human	.82	.84	.83	.86	.78	.82	.88	.81	.84
GPT 5.5	.86	.87	.86	.87	.81	.84	.92	.84	.88

Table 1: Multi-label topic classification performance on 20 human-annotated articles (micro/macro/weighted P/R/F1)

2.2 Topic-conditioned Candidate Claim-Evidence Pair Extraction

For each predicted topic, we extract structured candidate claim-evidence pairs using an LLM with schema-constrained generation. The model outputs normalized claims along with supporting evidence spans in JSON format. Unlike sentence-level approaches, this design captures claims distributed across multiple sentences or cross-sentence, en-

abling recovery of contextual information, such as the claimer. Evidence spans provide explicit provenance, while normalized claims are used for classification. To ensure reliability, we enforce a predefined schema and verify extracted spans against the source text.

Validation: We evaluated Llama 70B for topic-conditioned candidate claim-evidence pair extraction in three different settings (i) candidate claims x extraction only, (ii) evidence e extraction only, and (iii) candidate claim-evidence pair (x, e) extraction simultaneously, using a human reviewer and GPT 5.5. Table 2 reports the corresponding accuracies.

	Claim	Evidence	Claim-Evidence
Human	89	94	82
GPT 5.5	91	97	84

Table 2: Component-level validation of topic-conditioned extraction (% agreement with a human reviewer) for claim-only, evidence-only, and joint extraction.

2.3 Aspect Sentiment Aware ASE-BERT Classifier

Given a candidate claim sentence x and its associated document-level topic vector $\mathcal{T} \in \{0, 1\}^K$, ASE-BERT performs two tasks: (i) aspect-based sentiment prediction, and (ii) environmental claim classification. The classification task determines whether x constitutes an environmental claim ($y^{claim} \in \{0, 1\}$). Environmental claims typically express an evaluative stance towards environmental impact. To capture this signal, ASE-BERT incorporates sentiment conditioned on the assigned topics, enabling the model to distinguish claim-like statements from descriptive or neutral content.

As shown in Figure 3, the input candidate claim sentence is encoded using a transformer model to obtain contextual embedding $h = \text{Enc}(x)$, which

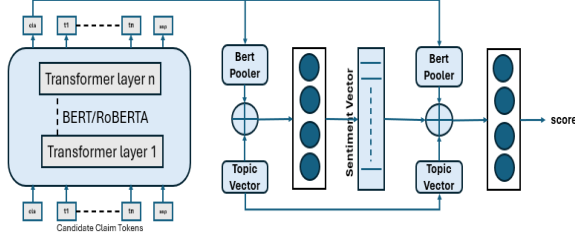


Figure 3: ASE-BERT architecture depicting sentiment & topic encoding with transformer for classification

is concatenated with a one-hot topic vector $\mathcal{T}$. This joint representation is passed to a sentiment regression head to estimate topic-conditioned sentiment $\hat{s} = f_{\text{sent}}([h; \mathcal{T}])$. The predicted sentiment, together with the sentence and topic representations, is then fed into a final classification layer $\hat{y}^{\text{claim}} = f_{\text{cls}}([h; \mathcal{T}; \hat{s}])$. Training proceeds in two stages: the sentiment head is first trained on an aspect-based sentiment analysis task, after which the full model is fine-tuned for claim classification. This lets ASE-BERT exploit sentiment signals while optimizing directly for the downstream objective.

2.4 Contextual Metadata Enrichment

For each detected claim, we extract additional contextual attributes using LLMs with deterministic decoding. Specifically, we identify: (i) the **claimer**, i.e., the entity making the statement (e.g., company, regulator), inferred from attribution cues such as reporting verbs, quoted text, and nearby named entities; (ii) the **beneficiary**, i.e., the entity affected by or benefiting from the claim (e.g., environment, community, or organization); and (iii) the **stage**, a discrete lifecycle category indicating whether the claim reflects planning, has been implemented or is indeterminate. Because beneficiaries are often implicit, the model sometimes returns generic environmental recipients (e.g., *forest*, *environment*, *nature*) that carry no specific accountability signal. To keep the beneficiary field informative, we apply a small set of manually defined filtering rules that map such generic mentions to *no beneficiary*, retaining only explicitly named affected parties.

These annotations provide structured context without altering the claim text, enabling downstream analyses such as attribution tracking, impact analysis, and cross-referencing with external sustainability disclosures. Full prompts, number of topics and other implementation details are provided in Appendix G for reproducibility.

3 News Dataset Construction

We construct a curated dataset using corporate environmental news articles to support environmental claim extraction. The news articles are collected from the Environmental section of the ESG News website². In total, the dataset includes 219 articles spanning diverse environmental topics and approx. $2k$ labeled claim instances.

Annotation Guidelines: The dataset was annotated by four domain experts with backgrounds in sustainability reporting/assurance, ESG risk analysis, and environmental policy/compliance, in a two-phase process.

Phase 1 - Curation (single expert): One annotator was given the complete pool of candidate claims extracted by all models for each article. This annotator (i) removed redundant or duplicate claims and rewrote fragmented ones into complete, self-contained statements, and (ii) verified that every retained claim was explicitly grounded in the source article. This step produced a clean, de-duplicated set of candidate claim-evidence pairs.

Phase 2 - Independent labeling (three experts): The curated claims, together with their source articles, were then independently labeled by three separate experts. For each candidate claim sentence, every annotator assigned a binary label (1 = environmental claim, 0 = otherwise) and confirmed that the claim was grounded in the article. Each annotator worked through a blinded interface, without access to the model identity or to the other annotators’ labels. Final labels were assigned by majority vote. Because claim interpretation is often topic-dependent, all annotators were provided with the definition of an environmental claim (Section 1) and the environmental topics associated with each claim-evidence pair. Inter-annotator agreement among the three Phase-2 experts was substantial, with Fleiss’ $\kappa = 0.81$ (Fleiss, 1971).

Dataset Statistics: The annotated dataset contains 1463 environmental claims and 284 non-environment claim instances. To improve coverage of the negative class, we augment the dataset with an additional 253 sentences sampled from the news corpus that are environmentally relevant but do not express explicit claims (e.g., descriptive or contextual statements). These sentences are manually annotated by domain experts following the same protocol. The final dataset consists of 1463 claims and 537 non-environmental claim instances.

²<https://esgnews.com/category/environmental/>

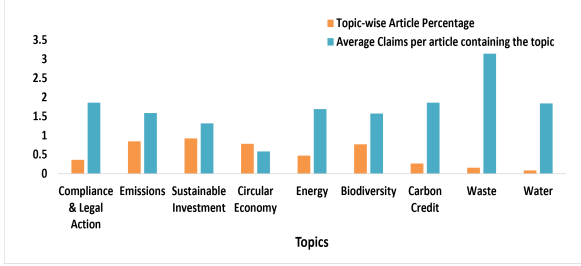


Figure 4: Topic-wise distribution of article volume and extracted candidate claims across the dataset

Topic Distribution: Figure 4 presents the distribution of topic-wise articles and average number of claims per article where the topic appears. While *Emissions* and *Sustainable Investment* dominate in terms of article coverage, high claim density is concentrated in relatively underrepresented topics, such as *Waste*, *Water*, and *Carbon Credit*. In contrast, widely covered themes like *Circular Economy* exhibit notably lower claim intensity (≈ 0.58). This variation highlights the uneven distribution of claim-bearing content across environmental domains, motivating the need for topic-aware extraction and classification.

4 Experiments & Results

We evaluate the proposed framework along two dimensions: (i) end-to-end claim extraction performance and the impact of topic conditioning, and (ii) claim classification performance of ASE-BERT.

4.1 End-to-End System Evaluation

We evaluate the proposed pipeline for joint extraction of environmental claims, comparing it against LLM-based baselines under two settings: (i) *Direct LLM extraction*, where models are prompted to extract environmental claims in a single inference step, and (ii) *Topic-conditioned extraction*, where topic information is explicitly provided to guide generation. The dataset is partitioned into 70/30 train-test split at the article level, ensuring that all claims from a given article appear exclusively in either the training or test partition. All results below are reported on the held-out test set consisting of 65 articles.

We experiment with GPT-OSS (120B) (OpenAI, 2025), Qwen (32B) (Ahmed et al., 2025), and Llama (8B and 70B) (Grattafiori et al., 2024) using identical prompts and structured output formats. Extractions are validated against the annotated ground truth, where an extracted claim is

counted as correct if a human judge confirms it is semantically equivalent to a ground-truth claim and grounded in the article. Performance is measured using Precision (P), Recall (R), and F1, along with coverage (Cvr@k%), defined as the percentage of articles with $\geq k\%$ of ground-truth claims retrieved. The *direct* GPT-OSS run failed to produce usable output and is therefore reported only in its topic-conditioned form.

Models	P	R	F1	Cvr@50%
Qwen 32B	.76	.49	.60	.51
Topics + Qwen 32B	.83	.67	.75	.65
Topics + GPT-OSS-120B	.59	.46	.52	.72
Llama 8B	.54	.47	.51	.43
Topics + Llama 8B	.66	.63	.64	.68
Llama 70B	.74	.49	.59	.33
Topics + Llama 70B	.78	.72	.75	.87
Ours	.83	.76	.80	.94

Table 3: End-to-end claim extraction on the held-out test set. “+ Topics” denotes topic-conditioned extraction; “Ours” adds ASE-BERT filtering. P/R/F1 and Cvr@50% are defined in the text.

Impact of Topic Conditioning: Table 3 shows incorporating topic information consistently improves performance across all evaluated LLMs, yielding an average gain of $\approx 15\%$ F1 and 19% in recall. This increase indicates that explicit topic grounding reduces ambiguity and focuses generation on relevant content. The significant 30% increase in Cvr@50% also supports our choice of stepwise pipeline, as jointly inferring topics and claims introduces noise, whereas prior topic grounding enables more precise extraction. Quantitatively, on a set of 65 news articles, the best unified LLM (Llama 70B) yields 213 extracted claims, whereas topic conditioned LLM produces 311 claims.

Our System vs LLM-Based Approaches: The proposed pipeline achieves the best overall performance, with Recall (.76) and F1 (.80), while maintaining competitive precision. Although direct LLM extraction via Llama 8B generates a larger number of claims, it often introduces noise and redundancy. GPT-OSS exhibits both redundancy and occasional hallucinations, while Qwen 32B and Llama 70B demonstrate relatively lower redundancy ($\approx 8\%$ redundancy for Llama 70B). Leveraging this, Llama 70B is used within our system. The pipeline further improves reliability through structured extraction and ASE-BERT-based filtering, resulting in cleaner and more accurate claim detection. (Refer Appendix D & E for detailed analysis and examples.)

Models	ECD			CF			A3CG			News (Ours)			ALL		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
ClimateBERT	.93	.75	.83	.54	.56	.55	.92	.37	.52	.99	.46	.63	.94	.35	.51
Llama 3.1 8B	.62	.82	.71	.76	.59	.67	.95	.60	.74	.97	.64	.77	.81	.65	.72
Llama 3.1 8B (FS)	.62	.99	.77	.73	.73	.73	.98	.78	.87	.96	.92	.94	.81	.85	.83
Llama 3.1 8B (Topic)	.63	.97	.76	.74	.67	.70	.98	.72	.83	.97	.92	.94	.82	.82	.82
Qwen 2.5 7B	.68	.97	.80	.71	.65	.68	1	.65	.79	.97	.84	.90	.82	.77	.79
Qwen 2.5 7B (FS)	.60	.97	.74	.73	.74	.74	1	.75	.86	.97	.90	.93	.81	.84	.82
Qwen 2.5 7B (Topic)	.59	.98	.74	.72	.78	.75	.98	.73	.84	.96	.91	.93	.80	.85	.82
EnvGPT	.88	.34	.49	.86	.14	.25	1	.19	.32	.97	.42	.58	.94	.28	.43
EnvGPT (FS)	.82	.48	.60	.77	.19	.31	1	.25	.40	.96	.42	.58	.89	.33	.48
EnvGPT (Topic)	.90	.40	.56	.77	.17	.28	1	.16	.30	.99	.18	.30	.93	.31	.46
Mistralv0.3 7B	.50	.97	.66	.70	.71	.70	.97	.78	.86	.96	.91	.94	.76	.83	.79
Mistralv0.3 7B (FS)	.54	.97	.70	.71	.71	.71	.98	.77	.86	.97	.91	.94	.78	.83	.81
Mistralv0.3 7B (Topic)	.47	.97	.63	.71	.73	.72	.91	.78	.84	.96	.92	.94	.74	.85	.79
BERT-Base FT	.81	.90	.85	.71	.88	.79	.86	.95	.90	.91	.93	.92	.76	.88	.82
BERT-Base + Topic	.78	.94	.85	.69	.98	.81	.95	.90	.93	.96	.97	.96	.78	.93	.85
BERT-Base + Sentiment	.80	.92	.86	.68	1	.81	.92	.93	.92	.95	.97	.96	.79	.91	.84
ASE-BERT 110M	.74	.96	.83	.70	.98	.82	.95	.93	.94	.96	.98	.97	.80	.94	.87

Table 4: Environmental claim classification: Precision (P), Recall (R), and F1 on each dataset and on the combined set (ALL). FS = few-shot; Topic = topic-conditioned prompting. **Bold** marks the best value per column.

4.2 Environmental Claim Classification

Performance of ASE-BERT

We evaluate ASE-BERT for environmental claim classification across multiple datasets to assess its generalization, robustness, and effectiveness relative to transformer and LLM-based baselines.

Training Setup: ASE-BERT is trained in two stages. First, the sentiment regression head is pre-trained on the MEMD-ABSA dataset (Cai et al., 2025) to capture aspect-specific sentiment signals. The full model is then fine-tuned for claim classification using a weighted binary cross-entropy loss to address class imbalance.

Datasets: We evaluate our pipeline on three public benchmarks and our curated news dataset: (i) *ECD* (Stammach et al., 2023), consisting of corporate disclosure sentences from multiple reports; (ii) *Climate Fever* (Diggelmann et al., 2020), derived from social media; (iii) *A3CG* (Ong et al., 2025), containing structured sustainability statements; and (iv) *News(Ours)* dataset. (For more dataset details, refer Appendix C)

Baselines and Evaluation Protocol: To establish a strong performance baseline, we evaluate ASE-BERT against two categories of baselines: a) Vanilla models, including BERT (Koroteev,

2021), fine-tuned with a classification head using Binary Cross-Entropy (BCE) with logits loss, trained with early stopping, and optimized using a learning rate of $2 * 10^{-5}$, and ClimateBERT (Webersinke et al., 2021) representing standard transformer architectures commonly used for text classification. b) Large Language Models, evaluated under zero-shot, few-shot (FS) and Topic conditioned settings, including Llama3.1 8B (Grattafiori et al., 2024), Qwen2.5 (Ahmed et al., 2025), EnvGPT (Zhang et al., 2025), and Mistralv0.3 (Jiang et al., 2024), each representing leading advances in reasoning-centric and general-purpose language modeling. Experiments are conducted both on individual datasets and on a combined(ALL) dataset, where each dataset is partitioned using a claim-wise 70/30 train-test split (70% training, 30% test) at the level of individual claim instances. Performance is measured using Precision, Recall, and F1 score. Table 4 summarizes classification performance across all datasets and models.

Overall Performance: ASE-BERT achieves consistently high recall across most datasets, indicating strong sensitivity to implicit and context-dependent claims. For the combined (ALL) dataset, it achieves an F1 score of 0.87 with a recall of 0.94, outperforming both transformer baselines

and LLM-based approaches. Unlike LLMs, which exhibit unstable precision-recall trade-offs, ASE-BERT maintains a reliable balance, making it well-suited for recall-critical applications.

Dataset-Specific Results: i) On ECD, ASE-BERT achieves high recall (0.96), close to Llama3.1 8b (Topic), and a competitive F1 (0.83), slightly below BERT-Base + Sentiment. The lower precision is due to the larger number of more implicit claim cases in data. ii) On Climate Fever, ASE-BERT achieves highest F1 (0.82) with very high recall (0.98), outperforming both transformer and LLM baselines. LLMs exhibit comparatively lower recall in this setting, highlighting challenges in identifying claim structures in noisier text. iii) On A3CG, ASE-BERT achieves an F1 of 0.94, outperforming BERT-Base and demonstrating robust performance on structured inputs. While some LLMs achieve high precision, they consistently underperform in recall, missing diverse claim expressions. iv) On News(Ours) dataset, ASE-BERT achieves the best overall performance ($F1 = 0.97$, $\text{recall} = 0.98$), demonstrating the effectiveness of topic-aware and sentiment-informed modeling in the target domain.

LLMs vs Fine-tuned Models: LLM performance varies significantly across datasets. Zero-shot configurations often yield high precision but low recall, leading to weak F1 scores. Few-shot and Topic-conditioned prompting improve performance for some models but do not consistently resolve recall deficiencies. In contrast, fine-tuned models provide more stable and balanced performance, particularly for domain-specific claim detection.

Ablation Study: We analyze the contribution of key components by comparing variants of ASE-BERT: (i) **BERT-Base FT** (ii) **BERT-Base + Topic** (iii) **BERT-Base + Sentiment** (iv) **ASE-BERT**: full model combining both. Incorporating topic features consistently improves recall across datasets, with the exception of A3CG, by strengthening contextual alignment. Incorporating sentiment signals further enhances the detection of evaluative language, improving F1 scores. The full ASE-BERT model, combining both topic conditioning and sentiment information, achieves the strongest performance on the combined dataset and most individual settings, indicating that these components provide complementary benefits.

Figure 5 represents the cross dataset generalization of ASE-BERT when tuned using combined dataset and tested on all four datasets separately. Our approach achieved higher or comparable recall

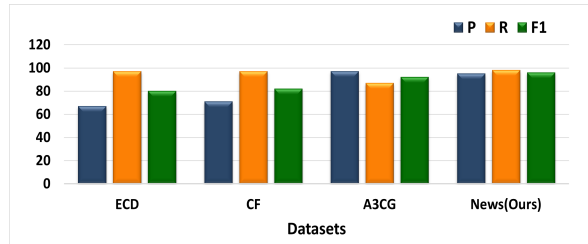


Figure 5: Results on other dataset while ASE-BERT is trained on combined data

for all the datasets. ASE-BERT achieves lower precision for ECD (67%) and CF (71%) while achieving higher precision on A3CG (97%) and News(Ours) data (95%).

5 Entity Attribution and Stage Analysis

The analysis below is reported only for 1463 instances labeled as environmental claims from the expert-labeled corpus. Figure 1 shows a representative example of how a claim in news text is enriched with its claimer, beneficiary, and lifecycle stage.

5.1 Entity Attribution

For each claim, we identify the *claimer* (who makes the claim) and the *beneficiary* (who is affected by or benefits from it), enabling accountability and downstream cross-checks. We manually validate both fields on the annotated corpus; Table 5 reports the results, where *accuracy* denotes exact-match correctness of the extracted entity and P/R/F1 are computed against the presence of a correct entity.

Entity Type	Accuracy	Precision	Recall	F1
Claimer	91.67	95.2	93.4	94.3
Beneficiary	76.70	75.2	98.3	85.2

Table 5: Manual validation of claimer and beneficiary extraction

Claimer identification is strong and well-balanced ($F1 = 94.3\%$), reflecting that claimers are usually explicit named entities like companies, executives, or regulators stated in the surrounding attribution context. Beneficiary extraction is harder: the model rarely misses a genuine beneficiary when one is present ($R = 98.3\%$) but tends to over-generate, proposing a beneficiary even when none is explicitly stated. Manual inspection shows that 31.5% of predicted beneficiaries are broad, generic recipients (e.g., “environment,” “nature,”

“ecosystem”) rather than a specific named stakeholder. While these outputs may be factually correct, they provide limited analytical value, since accountability in corporate news depends on identifying the specific stakeholder behind a claim. Applying the filtering rules specified in section 2.4, to map these generic mentions to *no beneficiary* improves beneficiary precision by 23 points over direct extraction.

Entity Distribution: Attribution is widely available: at least one party is mentioned in 96.6% of claims, both roles are present in 81%, and neither is present in only 3.4%. This allows most claims to be linked to both a speaker and a target, supporting accountability and downstream cross-referencing.

5.2 Stage Enrichment

To capture the maturity of each claim, we assign one of three stages: *Planning*, *Implemented*, and *Indeterminate*. Approximately 50% of the claims fall into the *Indeterminate* category, reflecting vague or non-action-specific statements, while *Implemented* and *Planning* each account for about 25%, indicating a balance between forward-looking commitments and reported actions. The large share of *Indeterminate* claims is expected, as statements often describe an environmental impact or outcome without explicitly anchoring it in time, lacking clear future-oriented or completed-action language. As a result, the boundary between *Implemented* and *Indeterminate* is often subtle, as a claim describing a positive outcome without an explicit time marker can reasonably be interpreted either way.

6 Conclusion

We presented a topic-aware framework for environmental claim extraction from corporate news articles. The proposed approach combines document-level topic grounding, topic-conditioned claim-evidence extraction, ASE-BERT based claim filtering, and contextual metadata enrichment to transform unstructured news narratives into structured environmental claims. Using an expert-annotated corpus of corporate environmental news, we show that for end-to-end extraction, topic conditioning alone improves LLM extraction F1 by an average of 15 points, and that our full pipeline adds a further 5 points over the best topic-conditioned setting. For claim classification, ASE-BERT attains the highest F1 (0.87) across the four combined environmental claim datasets, exceeding the best baseline without

task-specific fine-tuning by 4 points and demonstrating strong generalization beyond the target domain. In addition, the framework enables extraction of contextual information such as claimers, beneficiaries, and claim stages, supporting attribution and downstream analysis. Overall, the results demonstrate the value of topic-aware extraction and claim filtering for structuring environmental claims from corporate news, providing a foundation for downstream sustainability intelligence, ESG analytics, and environmental information discovery tasks.

7 Limitations

While the proposed framework demonstrates strong performance for environmental claim extraction and classification, several limitations remain.

Dependence on LLM-based Components: Topic classification, candidate claim-evidence extraction, and contextual metadata enrichment rely on LLMs. Although deterministic prompting is used to improve consistency, LLM outputs may still exhibit variability and occasional inaccuracies, particularly in complex or ambiguous cases.

Dataset Scope and Coverage: The dataset is constructed from a single news source, which may introduce biases in writing style, topic distribution, and reporting patterns. As a result, model performance may not fully generalize to other news domains, such as regional media, financial reporting, or non-English corpora.

Dataset Size: Despite expert annotation, the dataset remains relatively small compared to large-scale NLP benchmarks. Although sufficient for evaluating environmental claim extraction and classification, this may constrain the ability of supervised models to generalize to diverse linguistic patterns and rare claim formulations.

Contextual Metadata Extraction: While claimer extraction achieves high reliability, beneficiary identification remains more challenging. Manual analysis shows that many generated beneficiaries correspond to broad environmental recipients such as environment or nature rather than specific stakeholders, requiring post-processing rules to improve precision. Similarly, stage assignment can be difficult for claims that describe outcomes without clear temporal cues, resulting in ambiguity between implemented and indeterminate categories.

Claim verification: We note that our pipeline

structures and organizes environmental claims for analysis; it does not assess whether extracted claims are factually correct, materially significant, or consistent with reported sustainability performance, as such assessment requires verification against external evidence and is beyond the scope of the present work.

References

- Raia Abu Ahmad, Aida Usmanova, and Georg Rehm. 2025. The climatecheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 42–56.
- Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Naseef Ur Rahman Chowdhury, and Ahshanul Haque. 2025. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potential to surpass all competitors. *Authorea Preprints*.
- Ahmed Al-Rawi, Derrick O’Keefe, Oumar Kane, and Aimé-Jules Bizimana. 2021. Twitter’s fake news discourses around climate change and global warming. *Frontiers in Communication*, 6:729818.
- Simone Angioni, Sergio Consoli, Danilo Dessì, Francesco Osborne, Diego Reforgiato Recupero, and Angelo Salatino. 2024. Exploring environmental, social, and governance (esg) discourse in news: An ai-powered investigation through knowledge graph analysis. *IEEE Access*, 12:77269–77283.
- Darpan Aswal and Manjira Sinha. 2025. Efficient environmental claim detection with hyperbolic graph neural networks. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 24–35.
- Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting companies’ esg ratings from news articles using multivariate timeseries analysis. *arXiv preprint arXiv:2212.11765*.
- Md Badsha Biswas and Ozlem Uzuner. 2026. Contradiction to consensus: Dual perspective, multi source retrieval based claim verification with source level disagreement using llm. *arXiv preprint arXiv:2602.18693*.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Heqing Ma, Jianfei Yu, and 1 others. 2025. Memd-absa: a multi-element multi-domain dataset for aspect-based sentiment analysis. *Language Resources and Evaluation*, 59(3):2501–2529.
- Craig E Carroll and Maxwell McCombs. 2003. Agenda-setting effects of business news on the public’s images and opinions about major corporations. *Corporate reputation review*, 6(1):36–46.
- Yan Chen, Yijia Zheng, Gaotian Lv, Yahan Gu, and Wenjie Zhang. 2024. Quantifying the effect of esg-related news on chinese stock movements. *Journal of Environmental Management*, 354:120301.
- YeonJung Choi, Lanyu Shang, and Dong Wang. 2024. Climatedist: Climate change misinformation and stance detection dataset. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 321–330. Springer.
- Sebastião Vieira de Freitas Netto, Marcos Felipe Falcão Sobral, Ana Regina Bezerra Ribeiro, and Gleibson Robert da Luz Soares. 2020. Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32:1–12.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leipold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.
- Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.
- Gregor Dorfleitner and Rongxin Zhang. 2024. Esg news sentiment and stock price reactions: a comprehensive investigation via bert. *Schmalenbach Journal of Business Research*, 76(2):197–244.
- Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. 2022. Saving earth one tweet at a time through the lens of artificial intelligence. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Tanay Gupta, Tushar Goel, and Ishan Verma. 2025. Leveraging large language models for sustainability disclosures extraction and verification.
- Tanay Kumar Gupta, Tushar Goel, Ishan Verma, Lipika Dey, and Sachit Bhardwaj. 2024. Knowledge graph aided llm based esg question-answering from news. In *KG4S@ ESWC*, pages 65–78.
- Omar Mohammed Hassan Nassar, Fahimeh Jafari, and Chanchal Jain. 2025. From news to knowledge: Leveraging ai and knowledge graphs for real-time esg insights. *Sustainability*, 17(24):11128.

- Mohammad Izzat Raihan Imron. 2025. Sustainability in the corporate sector: A news textual analysis approach to measuring esg performance. *Junior Management Science (JUMS)*, 10(2):369–401.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.
- Julian F Kölbel, Markus Leippold, Jordy Rillaerts, and Qian Wang. 2022. Ask bert: How regulatory disclosure of transition and physical climate risks affects the cds term structure*. *Journal of Financial Economics*, 22(1):30–69.
- Haein Lee, Seon Lee, Kyeo Lee, and Kim Jang. 2023. Esg discourse analysis through bertopic: comparing news articles and academic papers. *Computers, Materials, & Continua*, 75(3):6023.
- Gladié Lui. 2024. Natural language processing analysis of sustainability disclosures and corporate operation efficiency. In *International Conference on Operations Research*, pages 66–71. Springer.
- Mohammad Mahdavi, Ramin Baghaei Mehr, and Tom Debus. 2024. Combat greenwashing with goalspotter: Automatic sustainability objective detection in heterogeneous reports. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4752–4759.
- Mahsa Mohammadrezaei, José Carlos Marques, and Asif Huq. 2024. Use of text mining and natural language processing techniques in analyzing sustainability reports: A systematic literature review and assessment. *Available at SSRN 5071032*.
- Gaku Morio, Soh Young In, Jungah Yoon, Harri Rowlands, and Christopher D Manning. 2024. Report-parse: A unified nlp tool for extracting document structure and semantics of corporate sustainability reporting. In *IJCAI*, pages 8749–8753.
- Keane Ong, Rui Mao, Deeksha Varshney, Erik Cambria, and Gianmarco Mengaldo. 2025. Towards robust esg analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. *arXiv preprint arXiv:2502.15821*.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). Preprint, arXiv:2508.10925.
- OpenAI. 2026. Gpt-5.5 models documentation. <https://developers.openai.com/api/docs/models/gpt-5.5>. Accessed: 2026-07-16.
- Mariya Pavlova, Bernard Casey, and Miaosen Wang. 2024. Esg-ftse: A corpus of news articles with esg relevance labels and use cases. *arXiv preprint arXiv:2405.20218*.
- Naiara Pikatza-Gorrotxategi, Jon Borregan-Alvarado, Aitor Ruiz-de-la Torre-Acha, and Izaskun Alvarez-Meaza. 2024. News and esg investment criteria: What’s behind it? *Social Network Analysis and Mining*, 14(1):47.
- Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed Elsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and 1 others. 2022. Newsclaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018.
- Diya Saha, Manjira Sinha, and Tirthankar Dasgupta. 2024. Enclaim: A style augmented transformer architecture for environmental claim detection. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 123–132.
- Diya Saha, Manjira Sinha, and Tirthankar Dasgupta. 2025. Green by design: Detecting environmental claims in corporate web content. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5176–5180.
- Tesfaye Salarin, Jacopo Staiano, Flavio Bazzana, and Sandra Paterlini. 2024. Examining intraday stock market response to esg news classified by ai: An event study approach. *Available at SSRN 5040946*.
- Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.
- George Serafeim and Aaron Yoon. 2022. Which corporate esg news does the market react to? *Financial Analysts Journal*, 78(1):59–78.
- George Serafeim and Aaron Yoon. 2023. Stock price reactions to esg news: the role of esg ratings and disagreement: G. serafeim, a. yoon. *Review of accounting studies*, 28(3):1500–1530.
- Dominik Stambach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066.
- Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicsesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

- Daniel Vogler and Mark Eisenegger. 2021. Csr communication, corporate reputation, and the role of the news media as an agenda-setter in the digital age. *Business & society*, 60(8):1957–1986.
- Thanh Nam Vu, Juha-Pekka Junttila, and Heikki Lehkonen. 2024. Esg news and long-run stock returns. *Finance Research Letters*, 60:104915.
- Mansi Wang, Xinxin Yu, Shuran Feng, and Chaohui Xu. 2025. The application of artificial intelligence in corporate climate risk: Information disclosure, information advantage, and information prediction. *Journal of Cleaner Production*, 534:147040.
- Nicolas Webersinke, Mathias Kraus, Julia Anna Binger, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.
- Vinicius Woloszyn, Joseph Kobti, and Vera Schmitt. 2021. Towards automatic green claim detection. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 28–34.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in neural information processing systems*, 37:95716–95743.
- Yuanxin Zhang, Sijie Lin, Yaxin Xiong, Nan Li, Lijin Zhong, Longzhen Ding, and Qing Hu. 2025. Fine-tuning large language models for interdisciplinary environmental challenges. *Environmental Science and Ecotechnology*, page 100608.
- Shi Zhi, Yicheng Sun, Jiayi Liu, Chao Zhang, and Jiawei Han. 2017. Claimverif: a real-time claim verification system using the web and fact databases. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2555–2558.

A Related Work

NLP for Sustainability Report & ESG News Analysis:

Research on NLP for sustainability disclosures has evolved from broad text-mining surveys to structure-aware extraction, automated scoring, and finance-oriented ESG measurement. Reviews such as (Mohammadrezaei et al., 2024) highlight the potential of NLP while emphasizing transparency and reproducibility. ReportParse (Morio et al., 2024) addresses long-report layout and semantic parsing, while subsequent studies connect NLP-derived disclosure characteristics to firm-level outcomes (Lui, 2024). Moving beyond extraction, Gupta et al. (2025) combine layout-aware PDF parsing with LLM-based modules that both extract GRI disclosures and verify them against reporting-standard requirements, classifying non-substantive

disclosures as potential greenwashing. Domain-specific models quantify E/S/G communication and explain variation in ESG ratings (Schimanski et al., 2024), and financial NLP benchmarks relate ESG analysis to information extraction, forecasting, risk management, and decision-making (Xie et al., 2024). A3CG (Ong et al., 2025) further introduces a claim-centric, topic-aware benchmark for sustainability statements.

In ESG news analysis, ESG-FTSE (Pavlova et al., 2024) offers ESG relevance, category, and target-company labels for news articles. DynamicESG (Tseng et al., 2023) offers multilingual annotated news for dynamically inferring ESG ratings. Prior work examines the relationship between ESG news and investment criteria (Pikatz-Gorrotxategi et al., 2024), compares ESG themes across news and academic writing using BERTopic (Lee et al., 2023), estimates firm-level ESG performance from large-scale news corpora (Imron, 2025), predicts firm-level ESG ratings from news streams (Aue et al., 2022). Studies show that stock prices react primarily to financially material ESG news (Serafeim and Yoon, 2022), that rating disagreement weakens the incorporation of ESG news into prices (Serafeim and Yoon, 2023), and links ESG sentiment to stock-price reactions (Dorfleitner and Zhang, 2024). Cross-market and recent studies extend this: ESG news sentiment and source characteristics have a significant impact on Chinese Stock market (Chen et al., 2024), ESG news effects on the US market is heterogenous across E,S,G themes (Vu et al., 2024), and GPT-4-classified intraday ESG news produces measurable, pillar-specific short-term effects (Salarin et al., 2024). Complementing this, Knowledge-based methods further organize ESG news into real-time knowledge graphs (Hassan Nassar et al., 2025), analyze longitudinal ESG discourse across E/S/G dimensions (Angioni et al., 2024) and support KG-RAG question answering over sustainability news articles (Gupta et al., 2024).

NLP for Environmental Claim Detection: Environmental claim detection has emerged as an important step for verification, ESG monitoring, and greenwashing analysis. Environmental Claim Detection (Stammbach et al., 2023) introduced an expert-annotated dataset and baselines, establishing claim detection as a prerequisite for scalable verification. Later, (Saha et al., 2024) improves claim identification across web and social media content using stylistic cues, while a recent study (Aswal

and Sinha, 2025) showed that parameter-efficient hyperbolic graph neural networks can surpass transformers for environmental claim detection. Complementary research (Woloszyn et al., 2021) detects and compares green claims automatically, and extends claim identification to heterogeneous corporate reports via GoalSpotter (Mahdavi et al., 2024) to capture environmental and social claims. Further research spans climate discussions on social media (Al-Rawi et al., 2021; Duong et al., 2022), claim verification methods (Zhi et al., 2017; Diggelmann et al., 2020; Dmonte et al., 2024; Biswas and Uzuner, 2026), and regulatory environmental disclosure analysis (Kölbel et al., 2022; Wang et al., 2025). Extending detection to corporate web properties, Green by Design (Saha et al., 2025) identifies environmental claims in short-form web content. NEWSCLAIMS (Reddy et al., 2022) provides an attribute-aware benchmark for news claim detection with claimer and claim-object extraction, however, it focuses on general news claims, primarily in the COVID-19 domain, rather than environmentally specific corporate claims.

Although sustainability-report NLP, ESG news analytics, ESG rating and market-impact prediction, and environmental-claim detection have each advanced substantially, these strands are often studied separately. Our work fills this gap by bringing them together within a unified, news-specific formulation, positioning it as a complementary but clearly distinct contribution to the ESG NLP and Claim Detection landscape. This claim-level view supports ESG intelligence, investment research, market monitoring, risk assessment, disclosure verification, and greenwashing analysis.

B Number of Topic Class

Selecting the optimal number of topic classes in the environmental domain is challenging. To guide this choice, we applied Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to the collected news corpus and evaluated topic quality using coherence scores. The resulting coherence maximizing configuration produced eight interpretable topic categories. To further refine these topics and ensure that each category reflected a coherent and environmentally meaningful vocabulary, we conducted manual word-level clustering on the LDA-derived terms. This manual consolidation allowed us to reorganize and enrich the eight LDA outputs into **nine final topic classes**, as defined above. To qual-

itatively inspect the coherence of these manually validated clusters, Figure 6 presents three representative word clouds selected from these nine classes. Topics such as Sustainable Investment and Biodiversity exhibit clear, domain-aligned vocabularies, whereas the Non-Environmental cluster highlights broader corporate terminology that helps distinguish true environmental content from general news discourse. These visual examples reinforce the interpretability and thematic clarity of the final topic classes used in downstream extraction.

C Other datasets detailed description

1. **Environmental Claim Detection(ECD) (Stammbach et al., 2023):** This dataset contains 2,647 sentences, sourced from corporate sustainability reports, earnings calls, and annual filings. After filtering tied annotations, we had 665 environmental claims and 1982 items belonging to not an environmental claim category.
2. **Climate Fever(CF) (Diggelmann et al., 2020):** This dataset contains 2669 sentences, with 1535 environmental claims and 1134 not an environmental claim, derived from social-media content.
3. **Aspect-Action Analysis with Cross-Category Generalization:** A3CG Dataset (Ong et al., 2025) consists 1790 sentences drawn from sustainability statements from Sustainability Reports of Singapore Exchange Companies. Under this scheme, the dataset contains 800 environmental claims and 990 instances belonging to not an environmental claim category.

D Examples of claims

This section presents different types of claim examples. Table 6 presents representative examples highlighting key challenges in the claim extraction task.

- The first three examples illustrate redundant claims generated via one of the experimented models. In first case, both claims pertain to water usage, where water intensity reduction is directly proportional to reduced water consumption. Additionally, the cited collaboration and the Nuton bioleaching process refer to the same underlying evidence, resulting in redundancy. In the second and third

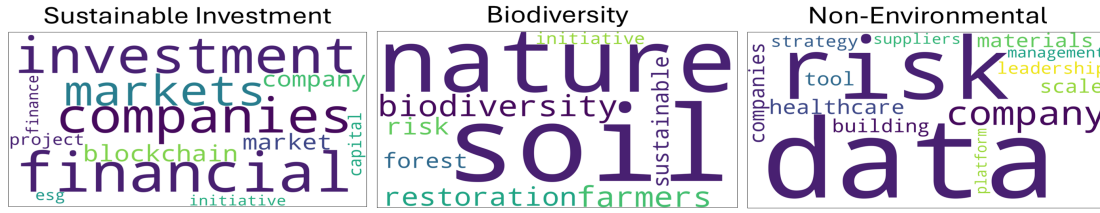


Figure 6: Example word-clouds illustrating three of the nine topic classes.

Example	Type
The Nuton bioleaching process reduces water intensity relative to conventional concentrator routes in copper production.	Redundant Claim
The collaboration aims to reduce water use in copper production.	
Apple’s direct display suppliers have committed to abate 90% of F-GHG emissions by 2030.	Redundant Claim
Apple’s suppliers have committed to abate 90% of F-GHG emissions by 2030.	
Apple achieved a 60% reduction in global greenhouse gas emissions compared to its 2015 baseline	Redundant Claim
Apple has cut global greenhouse gas emissions by over 60% since 2015.	
ANEO will more than double its wind power production in Sweden.	Hallucination
Microsoft reused over 6.2 million components in 2024.	Hallucination
Channeling \$10 billion target for 2030, into restoration and bioeconomy could shift Brazil’s economic trajectory and global climate outcomes.	Correct (Cross-sentence)
Microsoft diverted over 2,500 metric tonnes of waste from landfills through redesigned datacenter packaging.	Correct (Claimer from other sentence)
Voyager Ventures invests in carbon management technologies that enable capture, reuse, and removal, which can help achieve carbon neutrality targets.	Correct (Cross-sentence)
Singapore commits to channel 5% of proceeds to climate adaptation in Chile and cancel 2% of credits to ensure net global emission reductions. (Claimer: Singapore, Beneficiary: Chile)	Correct (Explicit Claimer & Beneficiary)
Nonprofits using Salesforce’s AgentForce AI platform are projected to reduce emissions by 20% for disaster relief organization Good360. (Claimer: Salesforce, Beneficiary: Good360)	Correct (Explicit Claimer & Beneficiary)
The project will support Microsoft’s broader portfolio of high-quality carbon removal solutions. (Claimer: Brian Marrs, Senior Director for Energy & Carbon Removal at Microsoft)	Correct (Explicit Claimer)

Table 6: Example Table

cases, redundancy arises due to alternative phrasings of semantically equivalent claims. Moreover, a manual evaluation over 20 articles shows that our system generates 7 redundant claims, compared to 34 for Llama 8B, 18 for Llama 70B, 9 for Qwen 32B, and 17 for GPT OSS-120B under topic-conditioned generation. This demonstrates that our system achieves lower redundancy than other baseline models.

- The fourth and fifth examples demonstrate hallucinated claims generated. In the fourth ex-

ample, the generated claim refers to doubling wind power, whereas the supporting evidence only indicates an increase in wind capacity (124 MW onshore and 48 MW nearshore) with no mention of baseline capacity. Similarly, in the fifth example, the generated claim is inconsistent with the reported value of ‘3.2 million’ in the evidence.

- The sixth, seventh, and eighth examples highlight cross-sentence dependencies. In the sixth and eighth cases, claims are constructed by integrating information across multiple sen-

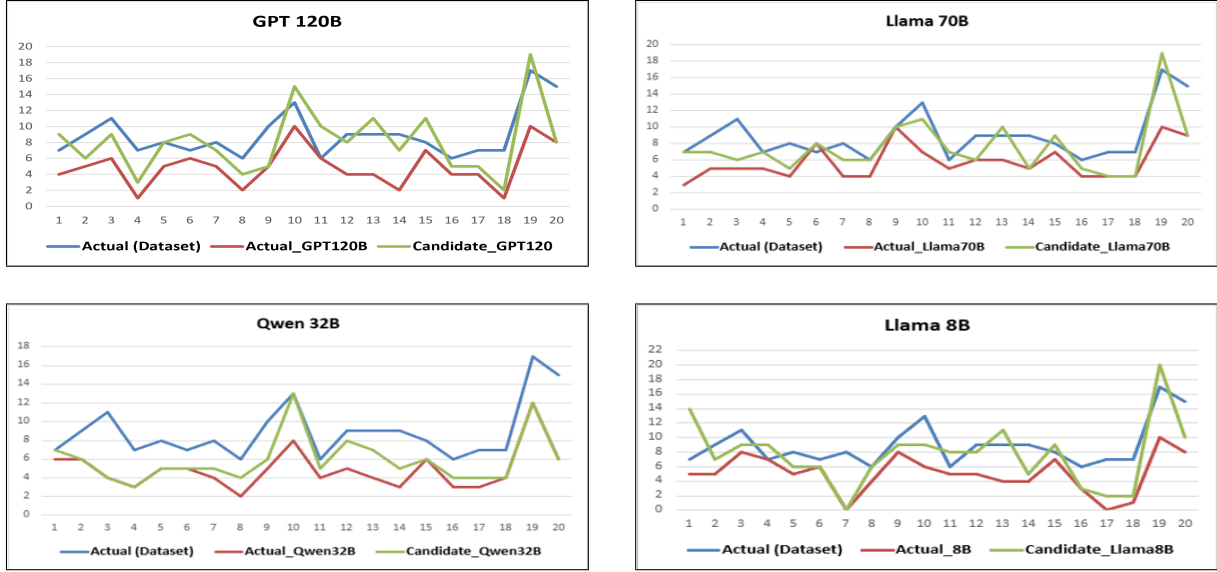


Figure 7: Comparison of ground-truth and model-generated claims across four models

tences, whereas in the seventh case, only the claimer is derived from a preceding sentence.

- The last three examples emphasize the importance of metadata-driven contextual enrichment. In the ninth and tenth examples, both the claimer and beneficiary are explicitly stated in the generated claim. However, in the last example, initially it may seem like some third party is claiming about Microsoft. However, closer inspection shows that the claimer is Brian Marrs, Senior Director for Energy & Carbon Removal at Microsoft, thereby indicating that the claim is made on behalf of Microsoft.

E Analysis of Model-wise Claim Trends

Figure 7 compares ground-truth claim counts (Actual Dataset) with model-generated actual and candidate claims across a representative subset of 20 articles, as visualizing all 219 articles is not practical. The trends shown correspond to topic-conditioned LLM outputs, which demonstrated better performance than direct LLM generation and are therefore used for analysis.

All models capture the overall trend of the ground-truth distribution, but fail to consistently match exact claim counts, indicating limited extraction precision. Candidate-generated claims (green curves) often overestimate counts, leading to redundancy due to over-generation. In contrast, validated outputs (red curves) are typically lower than the

ground truth, reflecting conservative filtering that improves precision but reduces coverage. Notable variability is also observed across models. Overall, all models struggle to capture variations in claim counts, highlighting challenges with heterogeneous article structures and varying claim densities.

F Hardware Details

The system was deployed for internal use. It utilises 2 NVIDIA Tesla V100 GPUs, equipped with 9 virtual CPUs, 60 GB of RAM, and 32 GB of dedicated GPU memory. All components of the system are implemented in Python. Based on our experimental results and infrastructure constraints, Llama 70B was deployed in INT8 precision for the internal system. For experimentation and comparative evaluation, larger models were accessed through Groq API calls. To assess model consistency, we repeated all experiments three times on a smaller evaluation dataset.

G Prompt & Cue Details

This section presents the complete set of prompts used throughout the system and evaluations, including those for multi-label topic classification in Table 7, candidate claim-evidence pair generation in Table 8, contextual meta-data extraction and LLM only baselines in Table 9. We also include prompt variants designed for in-context learning to enhance model performance.

Component	Content
System Prompt	You are an environmental content analyst. Given a news article text, determine if it is environment-related. If yes, extract the most relevant Topics from: Water, Waste, Emissions, Energy, Biodiversity, Carbon Credit, Compliance and Legal Action, Sustainable Investment, Circular Economy. Instructions: • Multi-label classification task • Output only relevant Topics • Use only predefined Topics • Return NA if not relevant • Output format: comma-separated Topics Topic Cues: Water (stress, recycling), Waste (recycling, landfill), Energy (renewable, efficiency), Emissions (GHG, pollutants), Biodiversity (habitat, conservation), Carbon Credit (trading, off-sets), Compliance (ISO, audits), Sustainable Investment (ESG, green bonds), Circular Economy (reuse, lifecycle)
User Prompt	Here is the article text: {final_text[i]}. Select the most relevant Topics from the predefined list. Follow the output format strictly and do not include explanations.

Table 7: Prompt structure for Environmental Topic Identification

Role	Content
System	Role: Expert information extraction system for environmental and sustainability claims. Task: Extract ONLY explicit and verifiable environmental claims grounded in the provided content. Constraints: • Do not infer, generalize, or hallucinate claims • Extract claims strictly from given text • If no claim exists for the given Topic, return an empty list
User	Instruction: Extract environmental claims corresponding to each Topic. Definition of Environmental Claim: A statement that asserts, predicts, or interprets environmental impact and can be agreed or disagreed with. A claim qualifies if it: • Asserts change (reduction, improvement, harm, benefit, targets) • Predicts or commits to future environmental performance • Describes environmental properties of processes/products • Represents a disputable assertion Not a claim if: • Neutral description without impact • Operational detail without environmental assertion Guidelines: • Evaluate each Topic independently • Extract only explicit claims • Do not infer claims • Provide supporting text span for each claim Input: Headline: {final_headlines[i]} Article Content: {final_text[i]} Topics: {final_topics[i]} Output Format (JSON): <pre>{ "claims": [{ "Topic": "<topic>", "claims": [{ "claim": "<concise rewritten claim>", "evidence": "<exact supporting text>" }] }] }</pre>

Table 8: Prompt structure for Environmental Claim Extraction by Topic

Role	Content
System	You are an expert in extracting environmental claims from news articles with respect to given topics. Task: Extract ONLY explicit and verifiable environmental claims grounded in the input text. Requirements: • Do not infer, generalize, or hallucinate claims • Extract claims per provided topic • Identify: – Claimer (entity making the claim) – Claimee (beneficiary, if any) – Stage of claim: Planning, Indeterminate, Implemented • If no claims exist, return an empty list
User	Instruction: Extract explicit environmental claims from the article for each given topic. Input: Article Content: {final_text[i]} Topics: {final_categories[i]} Output Format (JSON): <pre>{ "claims": [{ "Topic": "<associated topic>", "claim": "<concise rewritten claim>", "evidence": "<exact supporting sentence(s)>", "claimer": "<identified claimer>", "claimee": "<identified beneficiary>", "stage of the claim": "<Planning Indeterminate Implemented>" }] }</pre>

Table 9: Prompt structure for Environmental Claim Extraction

Role	Content
System	You are an expert in analyzing sustainability and environmental claim. Task: Classify the stage of action in a company's claim for the given aspect, given an environmental claim and the corresponding aspect. Use the definitions for classification Definitions: • PLANNING: Future intent, commitment, or target; no implementation evidence. • IMPLEMENTED: Actions taken, ongoing work, or measurable progress. • INDETERMINATE: Mentions the aspect but unclear if action is taken or planned.
User	Instructions: • Classify strictly for the given aspect. • Output ONLY one label: PLANNING, IMPLEMENTED, or INDETERMINATE. • No explanations. Input: Claim: {final_text[i]} Topics: {final_categories[i]} Output: One of PLANNING / IMPLEMENTED / INDETERMINATE

Table 10: Prompt structure for Stage of the Claim Extraction

Role	Content
System	You are an expert in extracting entities from sustainability claims. Task: Given a claim and its evidence, identify the entity making the claim ('claimer'), and if explicitly stated, the entity benefiting from or affected by the claim ('beneficiary' or 'claimee') considering both the claim and the evidence.
User	Instructions: Return the required entities ONLY in the specified format, no explanations. Input: Environmental Claim: {final_claim[i]} Evidence: {final_evidence[i]} Output: Claimer: <claimer>; Beneficiary: <beneficiary>

Table 11: Prompt structure for Entity Extraction

Role	Content
System	Role: Text classification assistant. Task: Classify a given statement into one of the following: • Environmental Claim • Non-Environmental Claim Definition: An Environmental Claim is any statement referring to environmental impact, sustainability, climate change, carbon emissions, energy efficiency, pollution, conservation, recycling, biodegradability, renewable resources, eco-friendliness, or protection of nature. Instructions: • Read the statement carefully • Decide if it qualifies as an Environmental Claim • Output ONLY one label: – Environmental Claim – Non-Environmental Claim • Do not provide explanations or additional text Examples: {e}
User	Instruction: Classify the following statement: {q}

Table 12: Prompt structure for Environmental Claim Classification