Overview of the SustainEval 2025 Shared Task: identifying the topic and verifiability of sustainability report excerpts

Jakob Prange, Charlott Jakob, Patrick Göttfert, Raphael Huber, Pia Wenzel Neves, Annemarie Friedrich

Angaben zur Veröffentlichung / Publication details:

Prange, Jakob, Charlott Jakob, Patrick Göttfert, Raphael Huber, Pia Wenzel Neves, and Annemarie Friedrich. 2025. "Overview of the SustainEval 2025 Shared Task: identifying the topic and verifiability of sustainability report excerpts." In *KONVENS: 21th Conference on Natural Language Processing (KONVENS 2025); proceedings of the conference - volume 2: workshops, Hildesheim, Germany, 9.-12. September 2025*, edited by Christian Wartena and Ulrich Heid, 229–38. Hannover: HsH Applied Academics.
https://doi.org/10.25968/opus-3679.

Overview of the SustainEval 2025 Shared Task: Identifying the Topic and Verifiability of Sustainability Report Excerpts

Jakob Prange[1] **Charlott Jakob**[2,3] **Patrick Göttfert**[1]

Raphael Huber[1] **Pia Wenzel Neves**[2] **Annemarie Friedrich**[1]

[1]Universität Augsburg [2]TU Berlin [3]DFKI

Correspondence: annemarie.friedrich@uni-a.de

Abstract

The SustainEval shared task @ GermEval 2025 aims to analyze text from German sustainability reports. The shared task required solving two tasks: classifying a span's topic into one of 20 reporting criteria and estimating its verifiability on a scale from 0.0 to 1.0. The spans and their corresponding reporting criteria were retrieved from the DNK database. Furthermore, the spans were manually annotated to assess verifiability. This paper details the data collection process and provides an overview of the baselines, participating systems, and results. The submitted systems explore language-specific bidirectional and left-to-right encoders, combined with data augmentation methods. Ensembled BERT models with different sets of hyperparameters work best for content classification, while for verifiability rating, generative pretraining is competitive as well.

1 Introduction

In many ways, economic interest can be seen not only as a contributing factor to global climate change, but as one of its root causes (Leippold, 2023). Recent efforts in EU and national legislation address the corporate world directly and require companies to take responsibility by reporting on their current statuses, goals and deadlines, as well as concrete actions taken towards climate neutrality and overall environmental and social sustainability.

While corporate sustainability reporting aims to increase transparency, there is also a high incentive for companies to present themselves as environmentally friendly as possible, e.g. by selectively reporting only "good" actions or using vague optimistic language. This increasing tendency, known as *greenwashing*, makes it difficult to tell apart concrete, verifiable actions from high-level plans and plain publicity.

With this shared task as part of GermEval 2025, we aim to fuel research on the automatic analysis and detection of greenwashing by challenging teams to build systems that categorize excerpts from German sustainability reports with regard to **(A) content class** and **(B) claim verifiability**. The anonymized text excerpts in our dataset are taken from the German Sustainability Code (Deutscher Nachhaltigkeitskodex; DNK) online platform.[1]

Existing work on climate content classification (Webersinke et al., 2021; Bingler et al., 2022, 2024) and environmental claim verification (Diggelmann et al., 2020, *inter alia*) addresses exclusively English texts and often struggles to achieve good inter-annotator agreement on crucial properties like specificity and verifiability. To mitigate this issue, we rephrase the task from categorical to an ordinal scale for annotators and a real-valued continuous rating for system evaluation.

The primary goal of the shared task is the question what types of modeling approaches work best to analyze sustainability reports w.r.t. their content and clarity. Beyond that, we expect to gain insights from the participants' analyses towards the following questions: How are sustainability reports written? Are they clear and transparent as to which criteria are being addressed and how? In other words, how easy is it to train classifiers on them in general? And which parts of reports and what types of language are particularly challenging to categorize?

We find that various modeling approaches such as BERT-like bidirectional encoders and left-to-right generatively pretrained language models are feasible, when properly finetuned. The top performances (58.6% accuracy on content classification; 0.431 Kendall's τ on verifiability rating) leave room for improvement, which highlights both the complexity of the domain of German sustainability reports and the importance and future potential of the novel tasks we propose.

[1] www.deutscher-nachhaltigkeitskodex.de

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 229–238, 2025

Sec. ID	Criterion Section = Task A Label	Description	Example	Task B Label
8.	Policy Process Management **Incentive Systems**	The company discloses how target agreements and remuneration schemes for executives and employees are also geared towards the achievement of sustainability goals and how they are aligned with long-term value creation. It discloses the extent to which the achievement of these goals forms part of the evaluation of the top managerial level (board/managing directors) conducted by the monitoring body (supervisory board/advisory board).	Per our travel expenses guideline, preference is to be given to rail travel for business trips, for which **employees are regularly provided with a BahnCard**.	1.00
15.	Aspects Society **Equal Opportunities**	The company discloses in what way it has implemented national and international processes and what goals it has for the promotion of equal opportunities and diversity, occupational health and safety, participation rights, the integration of migrants and people with disabilities, fair pay as well as a work-life balance and how it will achieve these.	When it comes to promotions, **direct superiors alone** do **not make decisions regarding changes of role and salary increases.**	0.67
11.	Aspects Environment **Usage of Nat. Resources**	The company discloses the extent to which natural resources are used for the company's business activities. Possible options here are materials, the input and output of water, soil, waste, energy, land and biodiversity as well as emissions for the life cycles of products and services.	We nevertheless **measure** our use of natural resources **to the best of our ability**.	0.33
2.	Policy Strategy **Materiality**	The company discloses the aspects of its business operations that have a significant impact on sustainability issues and what material impact sustainability issues have on its operations. It analyses the positive and negative effects and provides information as to how these insights are integrated into the company's processes.	In **summary**, we **see** distinct **opportunities** for `[NAME]` to **generate new sustainability-related areas** of consultancy, **including sustainable** finance, **sustainability** risk, **sustainability** reporting, decarbonisation and digitalisation.	0.00

Table 1: Examples of reporting criteria sections, their short descriptions provided by DNK, and example text snippets taken from company reports. For brevity, we only show the target sentence here, but the snippets used in the final dataset also contain the preceding 3 sentences for context. German examples are presented in appendix A.

2 Data, Annotation, and Tasks

Here we describe what the data format looks like, how we collected, processed, and annotated the sustainability report excerpts, and we provide precise specifications of the two subtasks.

2.1 Data Format

Each sample in our dataset has a unique ID and consists of three consecutive `context` sentences, one `target` sentence which follows the context in the original document, as well as the `year` in which the report was published (2017–2021, the evaluation data also contain reports from the years 2022 and 2023). Trial, training, and development samples additionally contain a content label (`task_a_label`, indexed 1–20), a verifiability rating (`task_b_label`), which can be any real-valued number between 0.0 and 1.0 (inclusive), and the standard deviation (`task_b_stdev`) over crowdsourced verifiability annotations (see section 2.4) as a measure of uncertainty.

2.2 Data Collection and Preprocessing

The trial data (88 samples), training data (960 samples), development data (270 samples), validation data (60 samples), and evaluation data (473 samples) have been constructed from publicly available German-language company reports indexed in the German Sustainability Code DNK. DNK reports always follow the same structure, consisting of 20 sections, each corresponding to a reporting criterion (e.g. 'Incentive Systems' or 'Usage of Natural Resources'). Each criterion section not only deals with a separate topic, but also fulfills a particular communicative purpose, which is reflected in the hierarchical structure of the report outline (fig. 1). One goal of this shared task is to determine the extent to which the texts pertaining to the different sections diverge not only in content but also style and other linguistic properties. A few examples of reporting criteria sections and example text snippets are listed in table 1.

As described in section 2.1, each input to be analyzed in Tasks A and B is a text snippet of

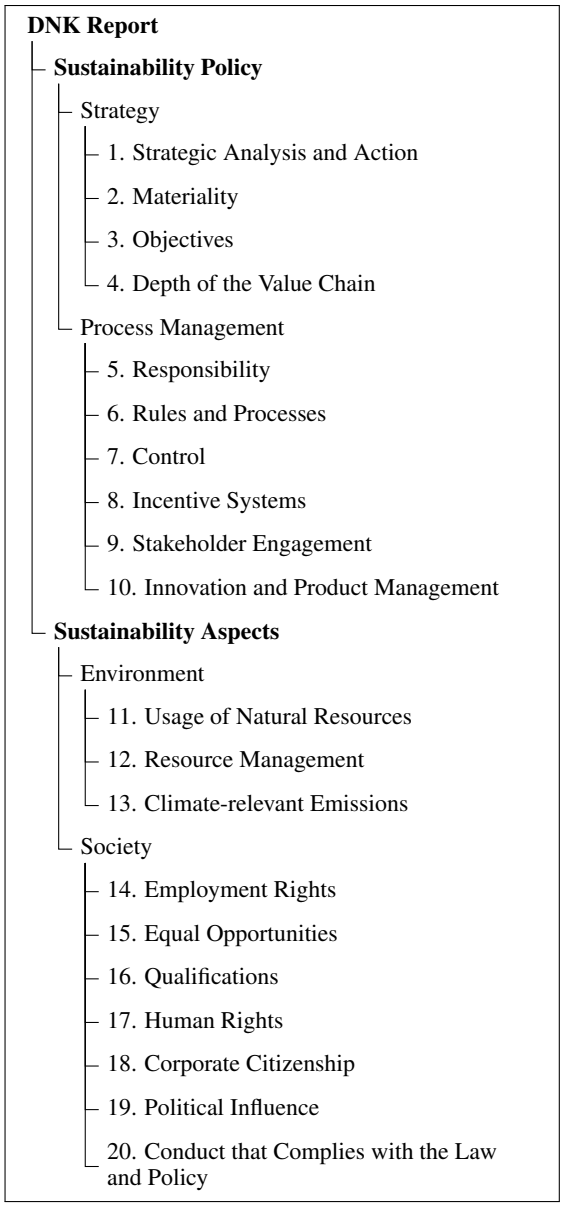

Figure 1: Structure of DNK reports.

4 consecutive sentences (3 context sentences + 1 target sentence). Text snippets were selected semi-automatically, based mostly on balanced[2] random sampling, with some filtering steps to exclude structured data such as tables, and anonymization tags replacing personally identifiable information (see appendix C). Sentence splitting and anonymization were done automatically with simple pattern-based[3] approaches at first and later manually corrected. For anonymization, we thereby prioritized high precision in the first (automatic) round and high recall in the second (manual) round.

[2]Balanced across the 20 sections and publication years.

[3]We used a punctuation-based sentence splitter and identified company names from the metadata of each report.

2.3 Task A: Content Classification (CC)

The challenge is to assign a suitable content class to each text sample. The label for each instance is the name of the DNK reporting criterion section the text snippet was sampled from (fig. 1), thus no human annotation was needed. Still, the selected text snippets were manually validated to ensure the task was neither too easy nor too difficult, and no inappropriate (e.g. noisy, personally identifiable, or offensive) data made it into the final dataset. The task is evaluated with standard accuracy.

2.4 Task B: Verifiability Rating (VR)

The challenge is to rate the verifiability of the statement, e.g., the goal or state description expressed in the last sentence of each text snippet (= the target sentence), with the previous sentences provided as context for better understanding.

We use a numerical score between 0.0 (not verifiable) and 1.0 (clearly verifiable), and predictions are evaluated by their Kendall τ-b rank correlation[4] with human ratings ($-1.0...1.0$; higher is better). We choose Kendall correlation (variant b) over Spearman for its better handling of ties, and elaborate further on the choice of metrics in appendix B.

Task B does require human annotation. The notion of verifiability is non-trivial and ambiguous. Our concrete instruction to annotators was as follows: Imagine you are an expert auditor with unlimited access to the "hard facts" of the company, e.g. technical measurements, internal communication, or legal regulations. Whenever the company claims to have implemented what they promised in the report, would it be possible for you to check whether this is true, based on how the sentences are written?

We collected answers to this question on a four-point (forced-choice) Likert scale (see table 1 for examples of each rating level):

- Not verifiable (0.0)
- Rather not verifiable (0.33)
- Somewhat verifiable (0.67)
- Clearly verifiable (1.0)

Several variants of greenwashing-detection tasks have been tried in the past: Bingler et al. (2024) applied a binary annotation scheme (specific; not specific) on English sustainability report paragraphs.

[4]https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient#Tau-b

They achieved 78% accuracy on this task with a finetuned ClimateBERT model (Webersinke et al., 2021) but only 22% raw human agreement (17% Krippendorff's α). This indicates issues with the clarity of the guidelines and potentially with the coarseness of a binary task. Burghart (2024) piloted a five-point numerical rating of "Measurability", achieving model performance of 61% macro-F1-score and human agreement of 34.4% Fleiss κ. The annotation guidelines for both of these projects mention "specific [pieces of] information on ... specific events" or that an "event is specific". This is intended to anchor annotators' decisions in event-related wording, thus increasing agreement. And indeed, human agreement on directly identifying event trigger words is much higher (κ =82.9% in Burghart (2024); 80-90% in our own pilot studies). In addition to guideline framing and textual anchoring of decision, many disagreements stem from the fact that, to some extent, this is an inherently subjective and gradient task.

The annotation for the SustainEval shared task, with our 4-point scale and detailed guidelines, was piloted by three of the organizers. During the pilot, we observed Fleiss' κ of 35.4%, similar to Burghart, but also nominal Krippendorff's α of 35.8%, twice as high as Bingler et al., as well as a ordinal and interval Krippendorff's αs of 57.7% and 58.7%, respectively. For 82.3% of instances, we found a majority vote (any 2 or 3 out of 3 annotators agree categorically). The numerical mapping of the categories has the added advantage that in cases of disagreement, we can calculate the arithmetic mean rating. Agreement, mean, and variance can also be used to distinguish clear from uncertain instances.

Once finalized, annotation was executed efficiently at scale by crowdworkers. For the development data, we worked with Crowdee GmbH, Berlin,[5] and for the training and evaluation data, we switched to Prolific[6] for annotator recruiting, while maintaining GDPR compliance by storing data and carrying out annotation on SoSciSurvey[7] servers located in Germany. No personally identifiable information was collected from annotators at any time. High quality was maintained by collecting $\approx$ 5 crowd annotations per sample. Inter-annotator agreement was measured between the crowd majority vote and the task organizers based on two control instances per annotation batch. Raw agreement between assignments to the four items on the Likert scale was 80% on the development data and 78% on the training data.

To arrive at the gold standard label, we took the majority vote, and in the case where the vote was tied, we computed the arithmetic mean between the tied values. In these cases, we also reported the standard deviation over the tied values (in cases where there was a unique majority vote, the standard deviation is 0.0). The information about the standard deviation was not strictly part of the shared task, but was provided so participants could gain insight into the uncertainty or difficulty of individual samples in the training and development data splits.

To ensure a fair and transparent competition, we established additional quality control criteria for the evaluation data split.

- 432 instances were independently annotated by 5 crowd workers and 1 trained university student (= expert) each.
  - If there was a unique crowd majority vote or two tied votes (2x2 annotators) one rating point apart, and the expert rating was at most one rating point away from the crowd vote(s), the crowd vote (or mean of two tied majority votes) was kept. This was the case for 331 instances.
  - If two tied votes were *more* than one rating point apart or the expert rating was more than one rating point away from the crowd majority vote (101 instances), a final adjudication decision was made. In this process, 76 instances were kept with the adjudicated label and 25 instances were removed because (a) the variance of ratings among crowd workers and experts was too high to warrant a useful adjudication (13), or (b) the sentence meaning was not clear to the adjudicator (12).
  - 407 out of 432 instances remained.
- 68 additional instances were annotated by trained university students, two of which were removed due to lack of clarity even to senior researchers.

In sum, this leaves a total of 407+66 = 473 evaluation instances.

[5] https://www.crowdee.com/
[6] https://www.prolific.com/
[7] https://www.soscisurvey.de/

		Task A		Task B	
Team	**Model**	**Acc** [%]	**Rank**	**Kendall's** τ	**Rank**
Baseline	Random Baseline	5.0	–	0.000	–
Baseline	bert-base-german-cased	56.0	–	0.425	–
EcoTUB	GBERT	58.6	1	–	–
SuperGLEBer	Modern-GBERT	57.3	2	0.148	(unofficial)
SuperGLEBer	LLäMmlein	52.9	(unofficial)	0.431	(unofficial)
SuperGLEBer	Llama-3	49.3	(unofficial)	0.402	1

Table 2: Official shared task rankings.

3 Baselines

For each subtask, we trained a simple baseline model in order to provide orientation to participants during the development phase and to sanity-check results after the evaluation phase. The baseline models were finetuned from the pretrained BERT-base-german-cased checkpoint[8] on the SustainEval training data. For the input, the context and target sentences were concatenated and fed to the model with left-side truncation to ensure that the target sentence is always part of the input, plus as much of the preceding context as fits.

The Task A model was trained as a 20-class classifier with cross-entropy loss, while the Task B model was trained as a single-output regressor with mean-squared error loss, both with a learning rate of 2e−5, a batch size of 4 and at most 5 epochs. In both cases, the best model checkpoint was selected based on lowest development set loss—on Task B it was reached after the first epoch, while on Task A, development set loss kept decreasing until the 5th epoch.

4 Participating Systems and Results

We used CodaBench[9] for official results submission, evaluation, and ranking. Authors were requested to upload their code together with their final CodaBench submission and to include a link to their code repository in their system description papers.

The Task A CodaBench leaderboard received 6 submissions, but only for 2 of them system description papers were submitted. As this was a formal task participation requirement, only these are thus taken into account for the official ranking. Task B had one participating team. The official rankings are shown in table 2.

Team EcoTUB. Bove et al. (2025) experimented with data augmentation through English back-translation, as well as comparing and ensembling multiple hyperparameter settings. Their main starting point is the German BERT model GBERT (Chan et al., 2020), a newer variant of bert-base-german-cased. The back-translation method generated 1,050 (near-)paraphrases of context and target sentences, doubling the training data set in size. Additionally, the team compared various hyperparameter settings during finetuning and exploited their complementary performance across classes in a final ensemble model, achieving 58.6% accuracy on Task A, beating the baseline and taking 1st place.

Team SuperGLEBer. Wunderle et al. (2025) submitted three different models: ModernGBERT (Ehrmanntraut et al., 2025), Llama-3 (Grattafiori et al., 2024), LLäMmlein (Pfister et al., 2025). All three models were used in an encoder-classifier setup during finetuning and inference, even though Llama and LLäMmlein were pretrained as generative decoders (Pfister and Hotho, 2024). In Task A, ModernGBERT performed best out of the three and is the officially ranked system, but in Task B the same system performed worst (τ=0.148). LLäMmlein and Llama fared much better, and LLäMmlein even outperformed the baseline at τ=0.431, but the officially ranked system was Llama (τ=0.402) because it was the team's final submission. One additional goal this team set for themselves was to test whether their model rankings on multiple GermEval 2025 shared tasks—not only SustainEval A and B, but also candy speech detection (Clausen et al., 2025) and harmful content detection (Felser et al., 2025)—correlate with rankings according to

[8] https://huggingface.co/google-bert/bert-base-german-cased

[9] https://www.codabench.org/

the pre-existing SuperGLEBer benchmark. This contributes a valuable perspective on cross-task transfer. Team SuperGLEBer ranks 2nd in Task A, also outperforming the official baseline, and, as the only participating team, 1st in Task B.

5 Discussion

While the methodological approaches and analyses are described in detail by Wunderle et al. (2025) and Bove et al. (2025), we observe a few overarching trends in data usage and system performance.

Encoder-classifiers vs. generative LLMs. The ongoing progress in the development of generative LLMs begs the question how they compare to encoder-based classifiers like BERT on a set of tasks that is (a) new, (b) highly domain-specific, and (c) run on German data. On Task A, while all systems perform much better than chance and lie within a 10 p.p. window of each other (49.3%–58.6%), there is a clear separation between generatively-pretrained LLMs on the lower end of that range (49.3%–52.9%) and BERT-based models taking the lead (56.0%–58.6%). On Task B, the LLMs (0.402–0.431) of Team SuperGLEBer are on par with our BERT regressor baseline (0.425), while their Modern-GBERT implementation lags far behind (0.148). This is doubly surprising: (a) Llama and LLäMmlein are ranked much better relative to other models than on Task A and (b) Modern-GBERT performs much worse than a similar model on the same task. Since all models are finetuned and applied in the same way as encoder-classifiers here (section 4; Pfister and Hotho, 2024), any "real" effects of training and inference differences would have to stem from pretraining. It remains to be tested whether some of these observations are in fact real effects or rather spurious performance drops due to suboptimal hyperparameters.

Verifiability regression as a challenging new task. SustainEval Task B sticks out among other GermEval tasks, as it is not a discrete classification task but a regression task. Next to low participation, perhaps due to unfamiliarity, this has implications for both model design and evaluation—we evaluated it with Kendall's τ rank correlation, but there are other rank correlation metrics as well as simpler closeness metrics such as mean squared error. Team SuperGLEBer observed that on this task, a diverse set of models is performance-ranked quite differently than on established tasks like POS-tagging and other classification/tagging tasks, though it is unclear whether this is due to the task being regression, due to the evaluation metric, or due to the difficulty of defining *verifiability* in the first place.

Limitations and enhancements of the training data. One obvious limitation of the data we provided is its size: 1,000 instances are not enough to train a model from scratch. Based on the current state of the field, we estimated that the dominant paradigm would be to start from pretrained model checkpoints and to either finetune it or apply prompt-based in-context learning. In fact, all submitted approaches used finetuning. Team EcoTUB tried to address the remaining data sparsity problem by generating artificial data via back-translation through English, albeit with diminishing returns.

Content classification may be multi-label. Another artifact of the data arises from the randomly sampled short excerpts. With this setup, we wanted to test in Task A how easy it is to recognize the goal of a text based only on a few sentences. But as Bove et al. (2025) point out, it is already difficult to distinguish between some similar classes, e.g. Usage of Natural Resources and Resource Management), and this difficulty only increases the shorter the text samples are. It may thus be reasonable to frame the task as multi-label classification or to back off to higher levels of the class hierarchy (fig. 1) in the future.

Expert models and ensembles. Team EcoTUB found in Task A that hyperparameter selection impacts individual classes differently, and thus "expert models" for specific sets of classes can be trained. Ensembling these expert models then improves overall performance. This also gives rise to the hypothesis that an ensemble of Task A and Task B models might mutually benefit both tasks, which should be explored in future work.

6 Related Work

Prior research has harnessed natural language processing (NLP) to understand sustainability communication better. Earlier finance-centric NLP applications predominantly employed keyword-based approaches, which lacked contextual sensitivity (Cody et al., 2015; Sautner et al., 2023). Recent advancements have embraced machine learning models such as BERT (Devlin et al., 2019), which provide contextualized representations. A variety of

BERT-based approaches have been introduced, addressing a spectrum of tasks such as climate content classification (Schimanski et al., 2023; Webersinke et al., 2021; Bingler et al., 2022, 2024), topic detection (Callaghan et al., 2021), environmental claim detection (Stammbach et al., 2023; Diggelmann et al., 2020), and environmental claim verification (Wang et al., 2021; Diggelmann et al., 2020).

Various approaches exist for categorizing sustainability texts. On the one hand, they can be classified according to their main topic(s). For this purpose, binary classification tasks can be set up, such as classifying relevance to climate change (Shiwakoti et al., 2024). Furthermore, text can be categorized within several sustainability frameworks. The simplest framework involves the categories Environmental, Social, and Governance (ESG), and is frequently used in finance-centric NLP (Schimanski et al., 2024). The Sustainable Development Goals (SDGs), with 17 categories, and the 20 criteria of DNK, provide a more granular classification of content (Jakob et al., 2024; Pukelis et al., 2020). SDGs and DNK criteria are often used for detecting topics in documents and to analyse their frequency across multiple documents, such as sustainability reports and research papers (Callaghan et al., 2021; Guisiano et al., 2022; Bedard-Vallee et al., 2023; Schimanski et al., 2024). Besides determining topics in documents, sentences can be analysed by making fine-grained semantic distinctions with regard to their factfulness (Stammbach et al., 2023; Ong et al., 2025b). This analysis is crucial for detecting greenwashing, as it distinguishes between verifiable actions and vague or misleading rhetoric (de Freitas Netto et al., 2020; Ong et al., 2025a).

Recent research uses advancements in fake news detection to verify climate claims (Thorne et al., 2018; Leippold et al., 2024). Diggelmann et al. (2020) introduced CLIMATE-FEVER, which employs evidence retrieval from Wikipedia to evaluate human-generated, verifiable climate claims and defines a claim as verifiable if it is well-formed and subjectively investigatable.

It is of utmost importance to critically assess statements made in sustainability reports. However, if we aim to detect greenwashing by verifying claims in sustainability reports, a key question arises: how can we differentiate between verifiable and non-verifiable claims? We are aware of a SemEval shared task that addresses multilingual verification classification within the ESG framework (Seki et al., 2024). However, German was not among the languages considered. With our shared task, we contribute by evaluating verifiability in German sentences and by categorizing the sentences into 20 fine-grained topics defined by the criteria of the German Sustainability Code (DNK).

7 Conclusion

The SustainEval Shared Task, with its focus on German sustainability content analysis, addresses an important gap in research on greenwashing detection. The results reveal several promising methods. However, particularly for the second task on verifiability, which relies on a fine-grained, manually annotated dataset, the full potential has yet to be realized.

Acknowledgements

We are thankful to the German Sustainability Code DNK for providing the texts for our shared task dataset and to the German Society for Computational Linguistics (GSCL) for their financial support.

References

Alexandre Bedard-Vallee, Chris James, and Guillaume Roberge. 2023. Elsevier 2023 sustainable development goals (SDGs) mapping. *Elsevier Data Repository*.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What ClimateBERT has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2024. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. *Journal of Banking & Finance*, 164:107191.

Sinan Bove, Icondy Kiba-Gassaye, Sirak Tadesse, and Lisa Raithel. 2025. EcoTUB @ SustainEval 2025: Ensembling BERT for German sustainability report classification. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. HsH Applied Academics.

Martin Burghart. 2024. Assessing corporate sustainability eorts via natural language processing. Master's thesis, Technische Universität Berlin.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina

Andrijevic, Robert J Brecha, et al. 2021. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature climate change*, 11:966–972.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yulia Clausen, Tatjana Scheffler, and Michael Wiegand. 2025. Overview of the GermEval 2025 Shared Task on Candy Speech Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. ACL.

Emily M Cody, Andrew J Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M Danforth. 2015. Climate change sentiment on twitter: An unsolicited public opinion poll. *PloS one*, 10(8):e0136092.

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):19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-FEVER: A dataset for verification of real-world climate claims. ArXiv preprint arXiv:2012.00614.

Anton Ehrmanntraut, Julia Wunderle, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. ModernGBERT: German-only 1B encoder model trained from scratch. *Preprint*, arXiv:2505.13136.

Jenny Felser, Michael Spranger, and Melanie Siegel. 2025. Overview of the GermEval 2025 Shared Task on Harmful Content Detection. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jade Eva Guisiano, Raja Chiky, and Jonathas De Mello. 2022. Sdg-meter: A deep learning based tool for automatic text classification of the sustainable development goals. In *Asian conference on intelligent information and database systems*, pages 259–271. Springer.

Charlott Jakob, Vera Schmitt, Salar Mohtaj, and Sebastian Möller. 2024. Classifying sustainability reports using companies self-assessments. In *Advances in Information and Communication*, pages 547–557, Cham. Springer Nature Switzerland.

Markus Leippold. 2023. Corporate climate disclosures: how do we weed out cheap talkers? TEDxHEC talk, Paris, last access: November 20, 2024.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, et al. 2024. Automated fact-checking of climate change claims with large language models. ArXiv preprint arXiv:2401.12566.

Keane Ong, Rui Mao, Ranjan Satapathy, Ricardo Shirota Filho, Erik Cambria, Johan Sulaeman, and Gianmarco Mengaldo. 2025a. Explainable natural language processing for corporate sustainability analysis. *Information Fusion*, 115:102726.

Keane Ong, Rui Mao, Deeksha Varshney, Erik Cambria, and Gianmarco Mengaldo. 2025b. Towards robust esg analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. *arXiv preprint arXiv:2502.15821*.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German language understanding evaluation benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Lukas Pukelis, Núria Bautista Puig, Mykola Skrynik, and Vilius Stanciauskas. 2020. Osdg–open-source approach to classify text data by un sustainable development goals (sdgs). ArXiv preprint arXiv:2005.14569.

Zacharias Sautner, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang. 2023. Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498.

Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. 2023.

ClimateBERT-netzero: Detecting and assessing net zero and reduction targets. ArXiv preprint arXiv:2310.08096.

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.

Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Hanwool Lee, Juyeon Kang, Min-Yuh Day, and Chung-Chi Chen. 2024. ML-Promise: A multilingual dataset for corporate promise verification. ArXiv preprint arXiv:2411.04473.

Shuvam Shiwakoti, Surendrabikram Thapa, Kritesh Rauniyar, Akshyat Shah, Aashish Bhandari, and Usman Naseem. 2024. Analyzing the dynamics of climate change discourse on twitter: A new annotated corpus and multi-aspect classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 984–994.

Dominik Stammbach, 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, Toronto, Canada. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. ArXiv preprint arXiv:1803.05355.

Gengyu Wang, Lawrence Chillrud, and Kathleen McKeown. 2021. Evidence based automatic fact-checking for climate change misinformation. In *Proceedings of the 15th International AAAI Conference on Web and Social Media*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. ClimateBERT: A pretrained language model for climate-related text. ArXiv preprint arXiv:2110.12010.

Julia Wunderle, Jan Pfister, and Andreas Hotho. 2025. Die SuperGLEBer at GermEval 2025 shared tasks: Growing pains - when more isn't always better. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. HsH Applied Academics.

A German Examples

See table 3.

B Metrics

As mentioned in section 2, the main ranking metrics are accuracy for Task A and Kendall's τ-b correlation for Task B. Here we elaborate on these decisions.

For Task A, other metrics, such as per-class- and macro-F1-scores and breakdowns at coarser levels of the hierarchical document structure (see fig. 1), can be very insightful. For example, hierarchical evaluation does not punish systems as much as flat accuracy evaluation for confusing closely related categories, which generally makes intuitive sense. Hierarchical evaluation may produce counter-intuitive ties, e.g. getting all instances correct at the flattest level of the hierarchy receives the same score as getting $\frac{1}{n}^{\text{th}}$ of the instances correct at the deepest (n^{th}) level of the hierarchy. And finally, we do not want to "bake" lots of hard assumptions into the task setup from the beginning, which could be favored towards certain modeling approaches.

For Task B, we choose Kendall rank correlation for its proper tie handling properties. This particular metric is well-aligned with our 4-point ordinal scale annotation method. Already when treating this scale as four categories, we are observing much better agreement in our preliminary study than in prior work, which used either binary or more fine-grained categorical schemes. Furthermore, we can resolve disagreements by averaging (after double-checking for noisy examples and annotators). This leads to a gradient ranking, for which Pearson, Spearman, or Kendall rank correlation present generally viable evaluation metrics. At the same time, the high categorical agreement also leads to many intentional ties in the gold standard data. This is where the Kendall τ metric, variant b, shows its true power, as it is the only one that handles ties well.

C Anonymization

The following categories of information in the context and target sentences were anonymized and replaced with the corresponding tags:

- Links → [LINK]
- Addresses, telephone numbers, email addresses → [CONTACT]
- Individuals → [PERSON]
- Companies and organizations → [ORG]

ID	Criterion Section = Task A Label	Description	Example	Task B Label
8.	**Anreizsysteme**	Das Unternehmen legt offen, wie sich die Zielvereinbarungen und Vergütungen für Führungskräfte und Mitarbeiter auch am Erreichen von Nachhaltigkeitszielen und an der langfristigen Wertschöpfung orientieren. Es wird offengelegt, inwiefern die Erreichung dieser Ziele Teil der Evaluation der obersten Führungsebene (Vorstand/Geschäftsführung) durch das Kontrollorgan (Aufsichtsrat/Beirat) ist.	Gemäß Reisekostenrichtlinie soll für Dienstreisen bevorzugt die Bahn genutzt werden, wofür **die Kollegen regelmäßig eine entsprechende BahnCard zur Verfügung gestellt bekommen**.	1.0
15.	**Chancen-gerechtigkeit**	Das Unternehmen legt offen, wie es national und international Prozesse implementiert und welche Ziele es hat, um Chancengerechtigkeit und Vielfalt (Diversity), Arbeitssicherheit und Gesundheitsschutz, Mitbestimmung, Integration von Migranten und Menschen mit Behinderung, angemessene Bezahlung sowie Vereinbarung von Familie und Beruf zu fördern, und wie es diese umsetzt.	Bei Beförderungen entscheiden **nicht** die **direkten Vorgesetzten allein** über Positionswechsel und Gehaltserhöhung.	0.67
11.	**Inanspruchnahme von natürlichen Ressourcen**	Das Unternehmen legt offen, in welchem Umfang natürliche Ressourcen für die Geschäftstätigkeit in Anspruch genommen werden. Infrage kommen hier Materialien sowie der Input und Output von Wasser, Boden, Abfall, Energie, Fläche, Biodiversität sowie Emissionen für den Lebenszyklus von Produkten und Dienstleistungen.	Dennoch **messen** wir, **soweit möglich**, unsere Inanspruchnahme an natürlichen Ressourcen.	0.33
2.	**Wesentlichkeit**	Das Unternehmen legt offen, welche Aspekte der eigenen Geschäftstätigkeit wesentlich auf Aspekte der Nachhaltigkeit einwirken und welchen wesentlichen Einfluss die Aspekte der Nachhaltigkeit auf die Geschäftstätigkeit haben. Es analysiert die positiven und negativen Wirkungen und gibt an, wie diese Erkenntnisse in die eigenen Prozesse einfließen.	**Zusammengefasst sehen** wir klare Chancen für [NAME], **auf Basis des Themas Nachhaltigkeit** neue Beratungsfelder zu generieren, **u.a.** zu **Sustainable** Finance, **Sustainability** Risk, Nachhaltigkeitsreporting, Dekarbonisierung und Digitalisierung.	0.0

Table 3: Examples of reporting criteria sections, their short descriptions provided by DNK, and German example text snippets taken from company reports. Note that the examples here only show the target sentence, but the snippets used in the final dataset will always be 4 sentences (3 context + 1 target) long.

- Products (e.g., names of internal software) → [PRODUCT]
- Other identifiable names or items → [NAME]

City names were generally not anonymized, unless they were part of a company name or directly associated with a company. Well-known institutions and globally recognized organizations (such as the UN) were also not anonymized. Please note that some of the anonymized elements may have changed over the course of the dataset's development.