EcoTUB @ SustainEval 2025: Ensembling BERT for German Sustainability Report Classification

Sinan Bove*
Technische Universität Berlin
bove@campus.tu-berlin.de

Icondy Kiba-Gassaye*
Technische Universität Berlin
kiba-gassaye@campus.tu-berlin.de

Sirak Tadesse*
Technische Universität Berlin
sirak.tadesse@campus.tu-berlin.de

Lisa Raithel
BIFOLD
QU Lab, Technische Universität Berlin
raithel@tu-berlin.de

Abstract

This paper outlines our contribution to the SustainEval 2025 shared task (Prange et al., 2025), which focuses on the classification of German sustainability report snippets into one of 20 predefined content categories defined by the German Sustainability Code (DNK). We fine-tuned a transformer-based model, deepset/gbert-base, and explored multiple methods to improve classification performance, including hyperparameter tuning, data augmentation through back-translation, and model ensembling. While our ensemble model achieved a accuracy of 0.74 on our internal validation set, its performance dropped to 0.58 on the final test set evaluated by the organizers, highlighting challenges in adaptability to new data. We compare our results to several baselines and conduct error analysis to identify common misclassifications patterns, such as overlapping categories and ambiguous language. Our findings demonstrate both the potential and the limitations of NLP approaches for structured content analysis in German sustainability reports.

1 Introduction

Sustainability reporting is an essential practice for organizations to show their commitment to environmental, social, and economic standards. They serve as a tool for transparency and accountability for employees, investors, the public, and especially regulators. European law requires large companies and corporations to publish regular reports on the social and environmental risks they face, and on how their activities impact people and the environment (European Union, 2022). With the enactment of the new directive in 2022, mandatory sustainability reporting requirements have been extended to a significantly larger number of companies across the EU, leading to a rapid increase in published reports (European Union, 2022). These reports come in many different languages, formats and structures and their content needs to be processed, reviewed and regulated.

This work was conducted as part of the SustainEval 2025 Shared Task (Prange et al., 2025), which comprised two subtasks: Content Classification (Task A) and Verifiability Rating (Task B). Our submission addresses Task A, which involves classifying short text snippets extracted from sustainability reports according to one of 20 predefined criteria, as shown in Table 1 and defined by the German Sustainability Code (Deutscher Nachaltigkeitskodex, DNK) (Rat für Nachhaltige Entwicklung (RNE), 2020).

> *Neubauten statten wir mit Solarthermie aus. In weiteren Objekten haben wir Blockheizkraftwerke installiert. Die CO_2-Emissionen für das Jahr 2017 wurden separat ermittelt. Beim Vergleich der Jahre 2017 zu 2015 ergibt sich bereits eine spürbare Reduzierung.*

Figure 1: Example snippet with label 13: Climate-Relevant Emissions

In this project, we address the task of automatically classifying German-language text snippets from corporate sustainability reports into predefined content categories. Each text snippet corresponds to a specific reporting criteria defined by the German Sustainability Code, and the goal is to accurately assign the appropriate label to each sample. An example of one of these snippets together with its assigned label is shown in Figure 1. Our work attempts to answer the following research questions: How effectively can NLP and machine learning models perform this classification task? And which approaches are best to improve a baseline model?

Sustainability reports play a crucial role in documenting a company's compliance with environmental, social, and governance standards, and they significantly shape public perception and corpo-

*These authors contributed equally. This work was done in an Advanced Study Project of the Quality & Usability Lab at TU Berlin.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 239–249, 2025

Strategy	Process Management	Environment	Society
1. Strategic Analysis and Action	5. Responsibility	11. Usage of Natural Resources	14. Employment Rights
2. Materiality	6. Rules and Processes	12. Resource Management	15. Equal Opportunities
3. Objectives	7. Control	13. Climate-Relevant Emissions	16. Qualifications
4. Depth of the Value Chain	8. Incentive Systems		17. Human Rights
	9. Stakeholder Engagement		18. Corporate Citizenship
	10. Innovation and Product Management		19. Political Influence
			20. Conduct that Complies with the Law and Policy

Table 1: Predefined reporting criteria in the German Sustainability Code (DNK) (Rat für Nachhaltige Entwicklung (RNE), 2020)

rate image. Automating the classification of such reports offers several benefits: It saves time and resources, reduces human error, and enables large-scale analysis and benchmarking of sustainability activities across companies. NLP-based models can scale effectively to process vast volumes of textual data, making them ideal tools for this domain. However, challenges remain, particularly due to the complexity of sustainability-related language and the limited amount of research focused on German-language corporate reporting. Our project addresses this gap by contributing to a field where automated, language-specific solutions are still underexplored.

Our key contribution in this paper is the development of a model that classifies German text snippets from sustainability reports into one of twenty predefined classes from the German Sustainability Code. We further implement a baseline and an improved ensemble model to compare performances. To evaluate the models we the standard metrics accuracy, precision recall and F_1 score. Finally, we conducted an error analysis to identify misclassifications and gain an understanding of the faults in the model.

Our results show that the improved model outperforms our own baseline model across the key evaluation metrics and the official SustainEval baseline on accuracy (Prange et al., 2025). However, we also observed that certain classes were consistently easier to classify than others, most likely due to overlapping terminology or similar language usage across categories. Through detailed error analysis, we identified key challenges and limitations in distinguishing between semantically related classes. Overall, our findings demonstrate the strong potential of NLP-based models for the automated classification of sustainability report content. In this paper, we

1. propose a classification system tailored to German-language sustainability reports
2. evaluate its performance using multiple metrics and synthetic data augmentation, and
3. show that fine-tuned NLP models can significantly improve classification quality.

2 Related Work

Previous research has explored different approaches to automate the classification and analysis of sustainability reports, particularly with a focus on the United Nations Sustainable Development Goals (SDGs). Jakob et al. (2024) propose a method for classifying content in sustainability reports by leveraging SDG icons embedded by companies within their own reports. Their approach employs page-level annotations derived from these self-assessments to fine-tune multi-label transformer-based models (Vaswani et al., 2017), including Longformer (Beltagy et al., 2020) and Transformer-XL (Dai et al., 2019). In addition, they also experimented with OpenAI's GPT models (OpenAI et al., 2024) as part of their approach. Another study by Angin et al. (2022) introduced a fine-tuned RoBERTa (Liu et al., 2019) model trained on the OSDG Community Dataset (OSDG et al., 2024), comparing classical and deep learning models for both binary and multi-class classification of report segments, and found that deep learning approaches outperformed traditional ones. Shahi et al. (2011) looked at supervised text classification of corporate sustainability reports based on the Global Reporting Initiative (GRI) framework, investigating how machine learning can assess report completeness against official performance indicators. Additionally, Maibaum et al. (2024) compared various NLP techniques, such as dictionary-based methods, topic modeling, word embeddings, and large language models (LLMs) on a large labeled dataset, concluding that LLMs like ChatGPT and fine-tuned BERT (Devlin et al., 2019) models outperform earlier approaches, with fine-tuning being essential for optimal performance. Lastly, another paper developed a novel RoBERTa-based classifier to trace thematic transitions in European banks'

sustainability reporting over time, concluding with strong performance in SDG-related classification. (Li and Rockinger, 2024)

Our work shares several similarities with prior studies that use natural language processing techniques in the context of sustainability reporting. We focus on the classification of sustainability-related content and employ transformer-based models such as BERT for this task. However, while most related studies focus on English-language data and SDG alignment, we specifically address German-language sustainability reporting. Moreover, our classification task is based on the German Sustainability Code (Deutscher Nachhaltigkeitskodex, DNK, Rat für Nachhaltige Entwicklung (RNE) (2020)), using the 20 reporting criteria defined by DNK.

3 Methodology

Experimental Setting Our approach followed several steps to develop a strong classification system for German sustainability report text snippets. We first built a baseline model by fine-tuning a pretrained German BERT model variant[1] (Chan et al., 2020) specifically for our classification task. We then evaluated the baseline on a development set using accuracy, precision, recall, and F_1-score, and examined the confusion matrix to see where the model made mistakes. Based on these results, we improved the model by tuning hyperparameters and adding early stopping to prevent overfitting. After confirming better performance on the development set, we created an ensemble that combined several models, to boost overall accuracy and robustness.

Models and Fine-Tuning Specifically, for the classification task, we chose 'gbert-base' (Chan et al., 2020), a German language-specific variant of BERT that is well-suited for processing complex German text. In preliminary experiments, we tested several models, with the deepset model and the bert-base-german-cased variants emerging as the best-performing ones. While bert-base-german-cased was used as our own baseline, the final model was built using the deepset variant due to its superior results (Figure 2). This model serves as the backbone for our fine-tuning approach, allowing us to adapt its deep contextualized representations to the domain of sustainability report classification. We used the HuggingFace Transformers library (Wolf et al., 2020) to facilitate integration of the model, tokenizer, and training pipeline.

To optimize model performance, we implemented hyperparameter tuning using the Optuna framework (Akiba et al., 2019), which efficiently explores the parameter spaces transformer model, epochs, learning rate, weight decay, batch size, and warm-up ratio using Bayesian optimization. This search helped us identify configurations that improved accuracy while reducing overfitting risks. For training management and experiment tracking, we integrated the Weights & Biases framework (Biewald, 2020), providing transparent logging and visualization of training progress and metrics, and at the same time ensuring reproducibility.

Baselines To assess the performance of our models, we compared them against two baselines. First, we used the official baseline result provided by the shared task organizers, which achieved an accuracy of approximately 0.63 on the validation dataset (Prange et al., 2025). Additionally, we tested our own baseline model using random hyperparameters (Table 5), which achieved a similar accuracy of about 0.63 on the validation dataset, but only 0.61 on the development dataset.

Data The dataset used is publicly available and was curated as part of the Shared Task SustainEval 2025 (Prange et al., 2025). It consists of approximately 1,000 samples for training, 300 for development, around 500 for validation, and an additional trial set of about 80 samples. Each text snippet is between 3 to 5 sentences long and was derived from publicly available German-language company reports indexed in the German Sustainability Code (Rat für Nachhaltige Entwicklung (RNE), 2020). The structure of each sample included several variables: an ID, the year, a context string, a target string, and one assigned label out of 20, as shown in Table 1. An exemplary text snippet is shown in Figure 1.

Data Augmentation To improve our models performance, we augmented the original data with synthetic data generated through back translation using a model[2] provided by Tiedemann and Thottingal (2020). We translated the German context and target strings into English and then back into German in order to create paraphrased versions of the original text snippets. In total, we added 1,050

[1] deepset/gbert-base

[2] 'Helsinki-NLP/opus-mt-de-en'

mores text snipped using this method. Each generated data sample follows a similar structure as the original data with variables including context, target and task A label.

Ensemble We noticed that many models had close to the same accuracy, however performed differently on the different classes, depending on their hyperparameters. Based on that information, we decided to take an ensemble of models for our final submission, where the collective model can cancel out weaknesses of the individual models. The final ensemble is a combination of seven models where each has its own hyperparameters. The predictions per model are weighted equally. One of those models was fine-tuned on the synthetic data described above. For the remaining models, the synthetic data did not improve performance.

4 Results and Discussion

The main goal of our evaluation was to measure how well our models correctly classify each text snippet. We used the shared task's development dataset and the official validation phase, applying standard metrics: accuracy, precision, recall, and F_1-score. We also analyzed confusion matrices to identify common misclassifications and better understand which labels posed the most challenges.

After hyperparameter tuning, our best single model achieved an accuracy of 0.68 on both the validation and development sets, ranking first in the validation phase of Task A. Notably, although many models reached similar overall accuracy, their performance varied widely across labels. To address this, we combined seven diverse models into an ensemble, which improved accuracy to 0.74 on the development set (Table 2). In the final test phase, however, the ensemble's accuracy dropped to 0.58, placing second overall.

Per-class results show that the ensemble performed well on clear-cut categories like 'Incentive Systems' and 'Climate-Relevant Emissions', but struggled with more ambiguous ones like 'Strategic Analysis and Action' or 'Employment Rights'. The recall distribution (Figure 2) confirms that the ensemble reduced variation across labels compared to individual models.

The confusion matrices (Figure 3, Figure 4, Figure 5) illustrate this trend: while the ensemble reduced overall errors, some categories with overlapping semantics, e.g., 'Rules and Processes' vs. 'Responsibility', remained difficult to separate.

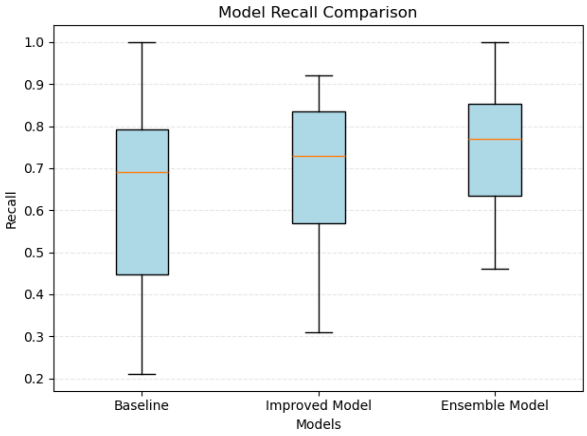

Figure 2: This illustrates the distribution of recall scores across individual class labels for the Baseline, Improved, and Ensemble models. Recall was computed separately for each label to assess the models' ability to correctly identify positive instances within each class.

The roughly 16-point drop from development to test accuracy suggests that our models may have overfit on the small in-domain datasets and did not generalize well to unseen data. This highlights a key limitation: larger or more diverse training data might benefit this task more than generic data augmentation. Moreover, the conceptual overlap among classes underlines an annotation challenge, real-world text snippets often fulfill multiple criteria, complicating single-label predictions.

5 Conclusion

Through our participation in the SustainEval 2025 shared task (Prange et al., 2025), we explored the classification of German sustainability report snippets into 20 predefined DNK criteria using NLP methods. We reconfirmed that BERT-based models like deepset's German BERT are well-suited for domain-specific classification tasks in German, particularly when paired with careful hyperparameter tuning and ensemble techniques. However, we also observed that back-translation and data augmentation did not lead to performance improvements in this context, and that evaluation results can vary significantly between internal validation and external benchmarks. Automating the analysis of sustainability reports offers a scalable solution to processing large volumes of complex, multilingual data. It contributes to greater transparency, comparability, and regulatory compliance in corporate sustainability practices, ultimately supporting better accountability in environmental, social,

Class	Precision	Recall	F1-score	Support
Strategic Analysis and Action	0.58	0.50	0.54	14
Materiality	0.67	0.46	0.55	13
Objectives	0.73	0.53	0.62	15
Depth of the Value Chain	0.75	0.64	0.69	14
Responsibility	0.67	0.77	0.71	13
Rules and Processes	0.78	0.54	0.64	13
Control	0.65	0.85	0.73	13
Incentive Systems	0.92	1.00	0.96	12
Stakeholder Engagement	0.86	0.86	0.86	14
Innovation and Product Management	0.56	0.64	0.60	14
Usage of Natural Resources	0.85	0.73	0.79	15
Resource Management	0.67	0.67	0.67	12
Climate-Relevant Emissions	0.88	0.93	0.90	15
Employment Rights	0.48	0.79	0.59	14
Equal Opportunities	0.89	0.62	0.73	13
Qualifications	0.79	0.85	0.81	13
Human Rights	0.91	0.83	0.87	12
Corporate Citizenship	0.91	0.77	0.83	13
Political Influence	0.80	1.00	0.89	12
Conduct that Complies with the Law and Policy	0.80	0.92	0.86	13
Accuracy			0.74	267
Macro avg F_1	0.76	0.74	0.74	267
Weighted avg F_1	0.75	0.74	0.74	267

Table 2: Score metrics for the final ensemble on the development dataset.

and governance reporting. For future work, several promising directions can be explored. One avenue is the implementation of hierarchical classification approaches, where the model first predicts a superclass and then classifies the snippet into a subclass. Additionally, the use of large language models could be investigated for zero-shot or few-shot classification, especially in scenarios with limited labeled data. Ultimately, this work lays a foundation for the broader use of NLP in sustainability and corporate accountability efforts.

6 Limitations

While our approach delivered competitive results, it has some limitations. First, semantic overlap between labels made it challenging to distinguish similar classes, which affected overall precision. Standard data augmentation, such as back translation, did not improve performance, suggesting that domain-specific strategies may be needed for specialized texts. Also, the drop in accuracy from internal validation to the final test set indicates potential overfitting and limited generalization. Finally, due to time constraints, we did not explore Task B (Verifiability Rating) or more advanced approaches such as multi-label or zero-shot classification with large language models, which could address some of these challenges.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BIFOLD25B).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Merih Angin, Beyza Taşdemir, Cenk Arda Yılmaz, Gökcan Demiralp, Mert Atay, Pelin Angin, and Gökhan Dikmener. 2022. A RoBERTa approach for automated processing of sustainability reports. 14(23):16139. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arxiv:2004.05150 [cs].

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

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.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. *Preprint*, arxiv:1901.02860 [cs].

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, Minneapolis, Minnesota. Association for Computational Linguistics.

European Union. 2022. Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting (Text with EEA relevance). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464. Accessed: 2025-07-10.

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. Springer Nature Switzerland.

Yao Li and Michael Rockinger. 2024. Unfolding the transitions in sustainability reporting. 16(2):809. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Frederik Maibaum, Johannes Kriebel, and Johann Nils Foege. 2024. Selecting textual analysis tools to classify sustainability information in corporate reporting. 183:114269.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OSDG, UNDP IICPSD SDG AI Lab, and PPMI. 2024. Osdg community dataset (osdg-cd).

Jakob Prange, 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 *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, Hildesheim, Germany. HsH Applied Academics.

Rat für Nachhaltige Entwicklung (RNE). 2020. Der Deutsche Nachhaltigkeitskodex: Maßstab für nachhaltiges Wirtschaften. https://www.nachhaltigkeitsrat.de/wp-content/uploads/2020/03/RNE_DNK_BroschuereA5_2019_DE.pdf. Accessed: 2025-07-10.

Amir Mohammad Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2011. Analysis of supervised text classification algorithms on corporate sustainability reports. In *Proceedings of 2011 International Conference on Computer Science and Network Technology*, volume 1, pages 96–100.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017*, page 11. Tex.ids= vaswani_attention_nodate.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Appendix

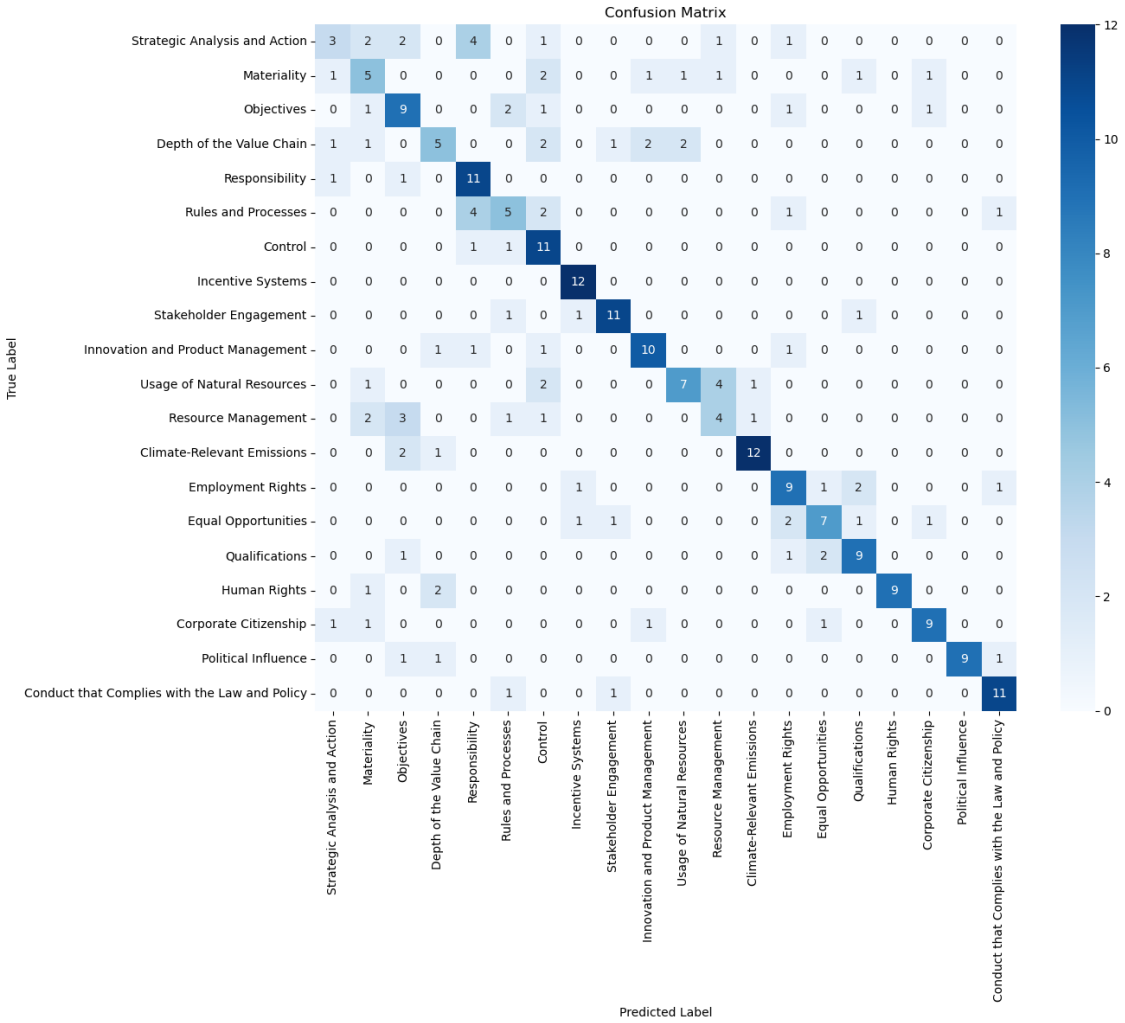

Figure 3: Confusion matrix of the baseline model or the development dataset.

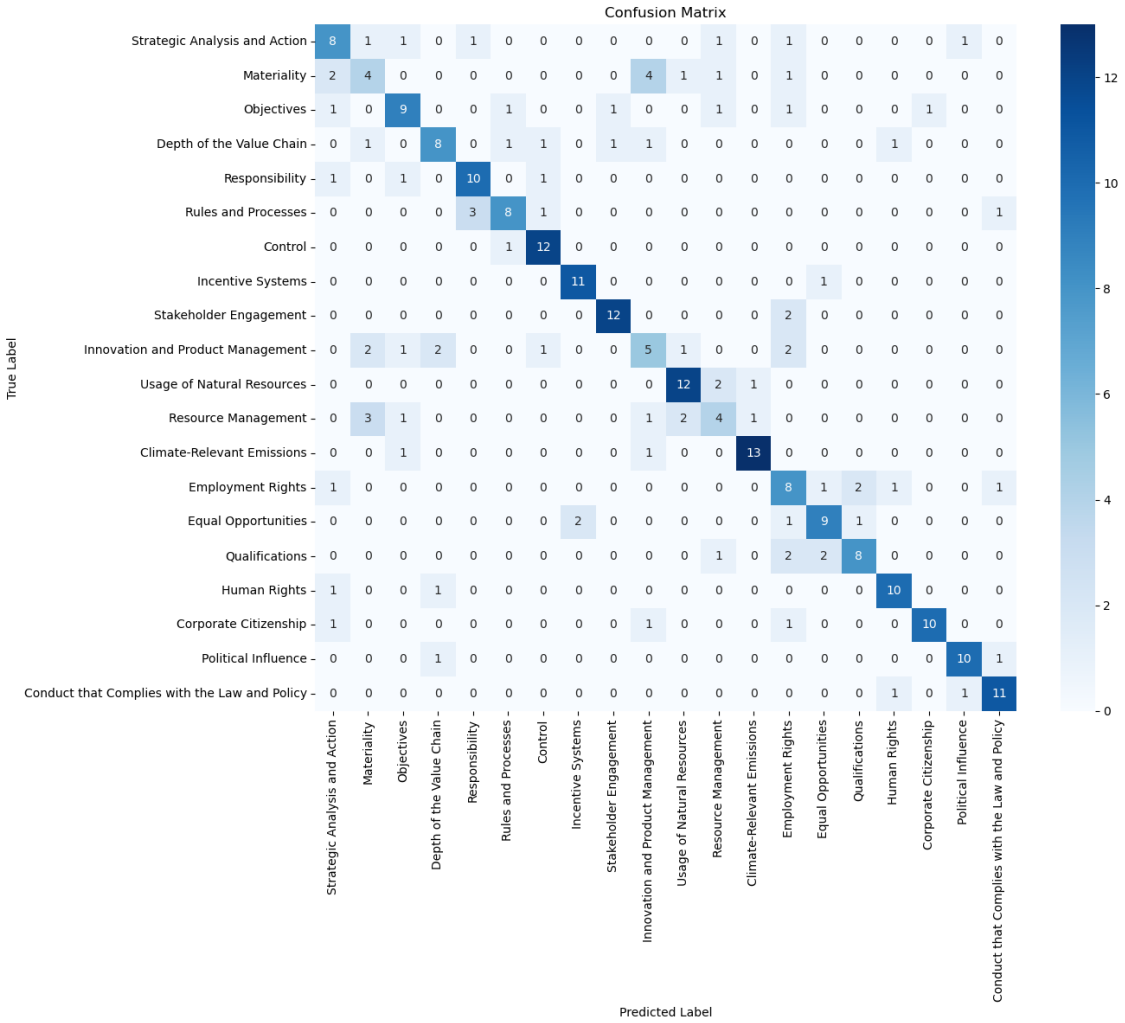

Figure 4: Confusion matrix of the improved model or the development dataset.

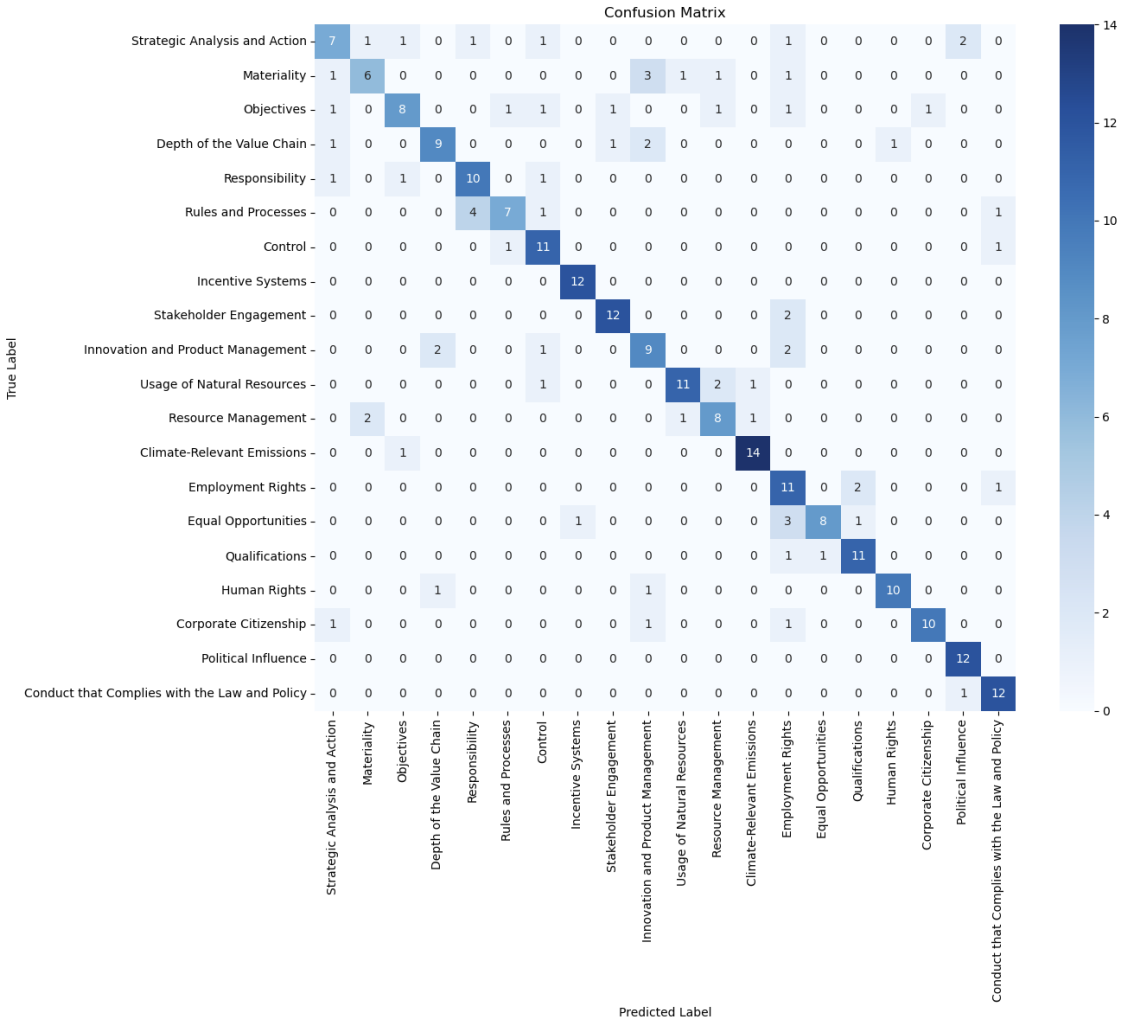

Figure 5: Confusion matrix of the final ensemble for the development dataset.

Class	Precision	Recall	F1-score	Support
Strategic Analysis and Action	0.43	0.21	0.29	14
Materiality	0.36	0.38	0.37	13
Objectives	0.47	0.60	0.53	15
Depth of the Value Chain	0.50	0.36	0.42	14
Responsibility	0.52	0.85	0.65	13
Rules and Processes	0.45	0.38	0.42	13
Control	0.48	0.85	0.61	13
Incentive Systems	0.80	1.00	0.89	12
Stakeholder Engagement	0.79	0.79	0.79	14
Innovation and Product Management	0.71	0.71	0.71	14
Usage of Natural Resources	0.70	0.47	0.56	15
Resource Management	0.40	0.33	0.36	12
Climate-Relevant Emissions	0.86	0.80	0.83	15
Employment Rights	0.56	0.64	0.60	14
Equal Opportunities	0.64	0.54	0.58	13
Qualifications	0.64	0.69	0.67	13
Human Rights	1.00	0.75	0.86	12
Corporate Citizenship	0.75	0.69	0.72	13
Political Influence	1.00	0.75	0.86	12
Conduct that Complies with the Law and Policy	0.79	0.85	0.81	13
Accuracy			0.63	267
Macro avg F_1	0.64	0.63	0.63	267
Weighted avg F_1	0.64	0.63	0.62	267

Table 3: Score metrics for the baseline on the development dataset.

Class	Precision	Recall	F1-score	Support
Strategic Analysis and Action	0.53	0.57	0.55	14
Materiality	0.36	0.31	0.33	13
Objectives	0.64	0.60	0.62	15
Depth of the Value Chain	0.67	0.57	0.62	14
Responsibility	0.71	0.77	0.74	13
Rules and Processes	0.73	0.62	0.67	13
Control	0.75	0.92	0.83	13
Incentive Systems	0.85	0.92	0.88	12
Stakeholder Engagement	0.86	0.86	0.86	14
Innovation and Product Management	0.38	0.36	0.37	14
Usage of Natural Resources	0.75	0.80	0.77	15
Resource Management	0.40	0.33	0.36	12
Climate-Relevant Emissions	0.87	0.87	0.87	15
Employment Rights	0.42	0.57	0.48	14
Equal Opportunities	0.69	0.69	0.69	13
Qualifications	0.73	0.62	0.67	13
Human Rights	0.77	0.83	0.80	12
Corporate Citizenship	0.91	0.77	0.83	13
Political Influence	0.83	0.83	0.83	12
Conduct that Complies with the Law and Policy	0.79	0.85	0.81	13
Accuracy			0.68	267
Macro avg F_1	0.68	0.68	0.68	267
Weighted avg F_1	0.68	0.68	0.68	267

Table 4: Score metrics for the improved model on the development dataset.

Model	Pretrained Model	Epochs	Learning Rate	Weight Decay	Batch Size	Warm-up Ratio
Baseline	bert-base-german-cased	8	0.00002	0.01	8	0.1
Improved Model	deepset/gbert-base	10	0.00007	0.25209	16	0.26270
Ensemble 1	deepset/gbert-base	10	0.00010	0.27410	16	0.26045
Ensemble 2	deepset/gbert-base	8	0.00004	0.27222	16	0.32183
Ensemble 3	deepset/gbert-base	9	0.00006	0.13132	16	0.25980
Ensemble 4	deepset/gbert-base	9	0.00004	0.20540	16	0.31962
Ensemble 5	deepset/gbert-base	9	0.00005	0.25659	16	0.28312
Ensemble 6	deepset/gbert-base	10	0.00004	0.21600	16	0.27289

Table 5: Hyperparameters of each model (The improved model is part of the ensemble).