Classify and Label the Content from the German Text Snippets of the Sustainable Development Report

Kongqiang Wang[1*], Qingli Tan[2], and Peng Zhang[1]
[1]School of Information Science and Engineering, Yunnan University, 650500, Kunming, Yunnan, China
[2]College of Ecology and Environment, Yunnan University, 650500, Kunming, Yunnan, China
*Corresponding author: wangkongqiang60@gmail.com

Abstract

The ability to understand text snippets content is an essential component of human-like artificial intelligence, as text snippets content greatly influence human cognition, decision making, and social interactions. In addition to intention recognition in sustainable development report, the task of identifying the potential categories behind an individual's text state in german text snippets is of great importance in many application scenarios. The main content of our research is content classification and labeling from German text snippets of sustainable development reports, which aims at assign a content classification label to German text snippets taken from sustainability reports. Each snippet contains 3–5 sentences and corresponds to one of the predefined categories based on the German Sustainability Code (DNK). We used a context-based text prediction method and combine with a pre-trained model. During this process, we repeatedly tested different pre-trained models in an effort to achieve the best results. The best result on the test set was an Accuracy of 0.635285412. We have reached the most advanced level in this field compared with other models. The project code is available from `https://github.com/WangKongQiang/Sustaineval-2025`.

Index Terms— Text Content Analysis, Text Classification, Multi-category labeling, Pre-trained Model, Bert.

1 Introduction

Understanding text snippets content is crucial to achieve human-like artificial intelligence, as text snippets are intrinsic to humans life and significantly influence our cognition, decision-making, and social interactions. Sustainable development report is an important form of human communication and contains a large amount of information. Furthermore, given that sustainable development report in its natural form is textual modality, many studies have explored textual modality intention recognition in sustainable development report using language modalities.

Sustainability reports, as a specialized form of corporate disclosure, possess distinctive linguistic and structural characteristics that differentiate them from general-purpose texts. These reports frequently employ conditional expressions, incorporate a large number of quantitative indicators, and make extensive references to external standards and regulatory frameworks. Such features introduce significant challenges for traditional language models, which are typically trained on general-domain corpora and often struggle to capture the deeper semantic and contextual nuances embedded in sustainability-related discourse.

From the perspective of computational linguistics [1], analyzing sustainability reports requires the development of a domain-specific theoretical framework. This framework should not only account for surface-level linguistic features—such as syntax, terminology, and discourse patterns—but also integrate the underlying business logic, reporting intentions, and verification mechanisms that govern the construction of these documents. In other words, effective analysis must bridge the gap between textual representation and the real-world corporate practices that the text aims to describe and justify.

To address these challenges, this study draws upon principles from information theory and causal reasoning to construct a novel computational model tailored to sustainability texts. By incorporating both statistical properties of language and causal relationships among reported indicators, the proposed model aims to enhance the interpretability and analytical accuracy of automated systems. Ultimately, this work seeks to lay a solid theoretical and methodological foundation for the automated analysis of sustainability reports, contributing to more reliable and scalable evaluation of corporate sustainability performance.

We proposing numerous well-designed pipeline systems. Moreover, we applied advanced pre-trained models for sustainable development report analysis and achieved promising results. These pre-trained models include *dbmdz/bert-base-german-cased*, *deepset/bert-base-german-cased-hatespeech-GermEval18Coarse*, *albert/albert-base-v2*, *google-bert/bert-base-german-cased*, *FacebookAI/roberta-large-mnli*, and *heripov/deid-roberta-i2b2- fine-tuned-german*.

2 Related Works

With the in-depth advancement of the global sustainable development goals, corporate sustainability reports have become

an important basis for regulatory compliance and investment decisions [2]. However, current automated analysis methods face three major challenges: Firstly, traditional text classification neglects the multimodal characteristics and temporal evolution patterns of sustainability reports; Secondly, the current verifiability assessment lacks a causal reasoning basis, making it difficult to distinguish between relevance and genuine verification capabilities. Finally, the single-task optimization strategy fails to capture the intrinsic correlation between content classification and verifiability assessment [3].

2.1 Intention Recognition in Sustainable Development Report

Research on automated sustainability-report analysis has gradually emerged at the intersection of natural language processing (NLP) and corporate disclosure studies. Early work in ESG text mining mainly focused on general-purpose text classification and sentiment analysis, where machine learning models were used to identify environmental or social themes in corporate disclosures [4]. However, these approaches were largely developed for English financial documents and often overlooked the unique linguistic patterns of sustainability reports, such as technical terminology, vague commitments, and externally referenced reporting standards.

Recent studies have emphasized the importance of domain-specific modeling for sustainability language. Prior work on climate-related text classification and environmental claim verification has shown that transformer-based language models can capture thematic information in corporate reports, but they still face challenges in identifying whether a statement is concrete, measurable, and externally verifiable. This limitation is particularly relevant for detecting greenwashing, where companies may use positive but unverifiable language to portray stronger sustainability performance than can be objectively supported.

To address this gap, the SustainEval 2025[1] shared task was introduced as an official GermEval challenge co-located with the KONVENS 2025 conference. The task focuses on the automatic analysis of German sustainability reports and defines two complementary subtasks: **content classification**, which assigns report excerpts to one of twenty sustainability reporting categories, and **verifiability rating**, which estimates how objectively a statement can be verified on a continuous scale. Participants have the opportunity to explore various machine learning (ML) and natural language processing (NLP) methods to address these challenges. By framing sustainability understanding as both a semantic and factual assessment problem, SustainEval provides one of the first benchmark datasets specifically designed for studying transparency and accountability in corporate sustainability communication. The shared task has also demonstrated that conventional pre-trained language models remain insufficient for this domain. Baseline and participating systems reported moderate performance, suggesting that sustainability-report analysis requires models capable of integrating linguistic cues with domain knowledge and causal reasoning. This motivates the development of specialized architectures that can jointly model textual meaning, numerical evidence, and reporting logic. Our work builds upon these findings by proposing a computational framework tailored to the deeper semantic structure of sustainability disclosures, aiming to improve both interpretability and automated evaluation in this emerging research area.

[1] https://sustaineval.github.io/

2.2 The Evolution of Deep Learning in Text Classification

Text classification, as a core task in natural language processing (NLP), has undergone a significant transformation from traditional machine learning to deep learning. Early studies mainly relied on Bag-of-Words models and TF-IDF feature engineering methods [5], which performed well in short text classification but had difficulty capturing complex semantic relations. The introduction of convolutional neural networks (CNNs) has brought new breakthroughs to text classification. Yoon Kim et al. [6] proved that multi-scale convolutional kernels can effectively extract local semantic features. Subsequently, recurrent neural networks, especially long short-term memory networks (LSTM), have demonstrated advantages in processing long sequential texts [7]. The emergence of the Transformer architecture has completely transformed the technical landscape of text classification. BERT [8] has achieved significant performance improvements on multiple text classification tasks through bidirectional encoder and mask language model pre-training. Subsequent improved models such as RoBERTa [9] and DeBERTa [10] have further promoted the development in this field. However, these general pre-trained models still have domain adaptability issues in professional domain text processing, especially when dealing with enterprise reports with special language features.

2.3 Research on Domain-specific Text Classification

Text classification in professional fields is confronted with challenges such as terminology specialization, context complexity, and scarcity of labeled data. In the field of financial text analysis, Dogu Araci et al. [11] proposed the FinBERT model, which is specifically pre-trained for financial documents and performs well in financial sentiment analysis and document classification tasks. In the field of medical text classification, BioBERT [12] and ClinicalBERT [13] have effectively improved the classification performance of medical literature and clinical records through continuous pre-training on biomedical corpora. In the research of legal text classification, Ilias Chalkidis et al. [14] developed LegalBERT, which specifically handles the language features of legal documents. The success of these domain-specific models validates the importance of domain adaptation for professional texts. However, as an emerging professional field, the language features and classification requirements of sustainability reports have not been fully studied.

Named Entity	Content
[ORG]	Names of companies and organizations
[PERSON]	Names of persons
[PRODUCT]	Products of a company or organization, e.g. name of a software

Named Entity	Content
[KONTAKT]	Addresses, phone numbers, emails
[LINK]	Links
[NAME]	Everything else that might identify a company or person but does not fit one of the other categories

Table 1: Example content for personally identifiable words or phrases are replaced by one of the tags.

2.4 The Application of Multi-task Learning in Text Processing

Multi-task learning enhances the performance of related tasks by sharing underlying representations and has been widely applied in the field of text processing. The MT-DNN framework proposed by Xiaodong Liu et al. [15] demonstrates the effectiveness of multi-task pre-training in natural language understanding tasks. The pioneering work of Ronan Collobert and Jason Weston [16] demonstrated that the joint training of part-of-speech tagging, named entity recognition, and semantic role tagging can enhance the performance of each sub-task. In the field of document analysis, Zichao Yang et al. [17] proposed a hierarchical attention network to handle both document classification and sentence importance assessment tasks simultaneously. However, the existing multi-task learning methods mainly focus on the combination of tasks at the grammatical and semantic levels, and rarely involve the joint modeling of content understanding and credibility evaluation, which provides an innovative space for this study.

2.5 Text Analysis of Sustainability and Corporate Social Responsibility

The text snippets are preprocessed with a named entity recognition (NER) tool, and then checked manually for further personally identifiable information [18]. Personally identifiable words or phrases are replaced by one of the tags below: See Table 1.

Sustainability report analysis, as an emerging research direction, currently focuses its work mainly on information extraction and sentiment analysis. Travis Dyer et al. [19] used machine learning methods to automatically identify environmental, social and governance (ESG) related information from corporate social responsibility reports. Tim Loughran and Bill McDonald [20] constructed a sentiment dictionary of financial texts, providing a fundamental tool for sentiment analysis of enterprise reports. Recent research has begun to focus on the quality assessment of sustainability reports. Katrin Hummel et al. [21] proposed an evaluation index for report quality based on text complexity and information density. Aminul Islam Chy and Md Mahbubur Rahman [22] used deep learning methods to analyze the completeness and accuracy of corporate sustainability disclosure. However, most of these studies are confined to English texts and lack in-depth analysis of the verifiability of the reported content.

2.6 German Natural Language Processing

German, as a language with rich morphological variations, faces unique challenges in natural language processing (NLP). The compound word formation, case change system and word order flexibility of German add complexity to text analysis. Pre-trained models such as German BERT [23] and GBERT [24] have been specifically optimized for German and have achieved good results in multiple German text processing tasks. In terms of German text classification, Julian Risch and Ralf Krestel [25] compared the performance of different pre-trained models in the German news classification task. Gregor Wiedemann et al. [26] studied the domain adaptation problem in German sentiment analysis. However, research on text classification in the specific field of German sustainability reports remains scarce, and the SustainEval 2025 shared task fills this significant gap.

3 Task and Dataset

3.1 Task Description

This Task has two sub-tasks, namely Task A: Content Classification and Task B: Verifiability Rating. The following describes these two sub-tasks respectively [27].

Task A: Content Classification. Participants are tasked with assigning a content class to text snippets from sustainability reports. The text snippets are sampled from different sections of German-language company reports indexed in the German Sustainability Code (DNK). Each snippet corresponds to one of the predefined reporting criteria in DNK, and the goal is to classify the snippet according to its corresponding criterion section. Evaluation of this task: Accuracy. There are 20 classes for this task [28]. Their specific situations are shown in Table 2.

Task B: Verifiability Rating. This task focuses on rating the verifiability of the last sentence in each text snippet, with the prior sentences serving as context. The verifiability score is assigned on a scale from 0.0 (not verifiable) to 1.0 (clearly verifiable). The task is evaluated by the Kendall τ rank correlation with human ratings. Evaluation of this task: Kendall τ rank correlation.

Our team mainly participated in Task A: Content Classification and achieved significant development results by using pre-trained model technology.

3.2 Task Dataset Instance

The training data will consist of text snippets from sustainability reports, which contains target content and context content, an illustrative example of the task dataset is shown in Figure 1.

4 Methodology

In this task, the main model training framework we use is a supervised training framework based on text modality. In this

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 2: The specific details of the 20 categories in the content classification of Task A: Content Classification.

Figure 1: This is an illustrative example for our task data, it does not actually appear in the task dataset, and its content class is Resource Management. The verifiability rating is 0.8.

framework, the context is part of the training text, concatenated at the end of the target text, and prompt words are inserted in the middle to inform the pre-trained model that the following text is the context content. Finally, it is fed to the pre-trained model together with the label of Task A to obtain the final model result. The overall text concatenation flow and text pre-training process. As shown in the Figure 2.

4.1 Data Preprocessing

Firstly, the official dataset is given in the JSONL format. JSONL (JSON Lines) is a special JSON file format. Unlike traditional JSON files, each line in a JSONL file is an independent JSON object without commas or other delimiters. This format makes the data clearer and more readable, especially when dealing with large amounts of data.

Based on this, we concatenate the target fragment and context fragment in training_data.jsonl, trial_data.jsonl and development_data.jsonl to form a new target fragment, so as to enable the model to perform content classification refer to the context [29]. The task_a_label fragment therein serves as labels_list [30]. Table 2 shows the classes supported by the text snippets.

In validation_data.jsonl file and evaluation_data.jsonl file, perform the same operation, except that they do not have the task_a_label fragment, and thus there is no way to extract labels_list, which is also the main step of the content and tasks we need to predict.

4.2 Dataset Division

We collect training dataset, the trials dataset and development dataset three datasets for a dataset, named TrainTrialDevelopment CSV file. The merged dataset is then divided into the training set and the validation set in a ratio of train: valid= 8:2. During the allocation process, stratified sampling and sample shuffling are carried out based on labels. You can obtain the complete code and datasets file from `https://github.com/WangKongQiang/Sustaineval-2025/tree/main/sustaineval2025_data/`.

4.3 Metrics Equations

The calculation formula for the model accuracy rate is shown below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Among them, TP represents the true example, TN represents the true counter example, FP represents the false positive example, and FN indicates a false counter example.

4.4 Fine-tuning Pre-trained Models

First, we choose the pre-trained model as dbmdz/bert-base-german-cased[2]. It is the MDZ Digital Library team (dbmdz) at the Bavarian State Library open sources another German BERT models [31]. The hyperparameters we provided to the model are shown in Table 3. The accuracy rate of this model on our validation set is 0.6853.

Table 3: Model hyperparameter Settings for $dbmdz/bert-base-german-cased$

hyperparameter	value
require_improvement	1000
num_epochs	30
batch_size	8
pad_size	128
learning_rate	2×10^{-5}
hidden_size	768
dropout	0.2

Second, we choose the pre-trained model as albert/albert-base-v2[3]. It is pre-trained model on English language using a masked language modeling (MLM) objective. This model as

[2]https://huggingface.co/dbmdz/bert-base-german-cased
[3]https://huggingface.co/albert/albert-base-v2

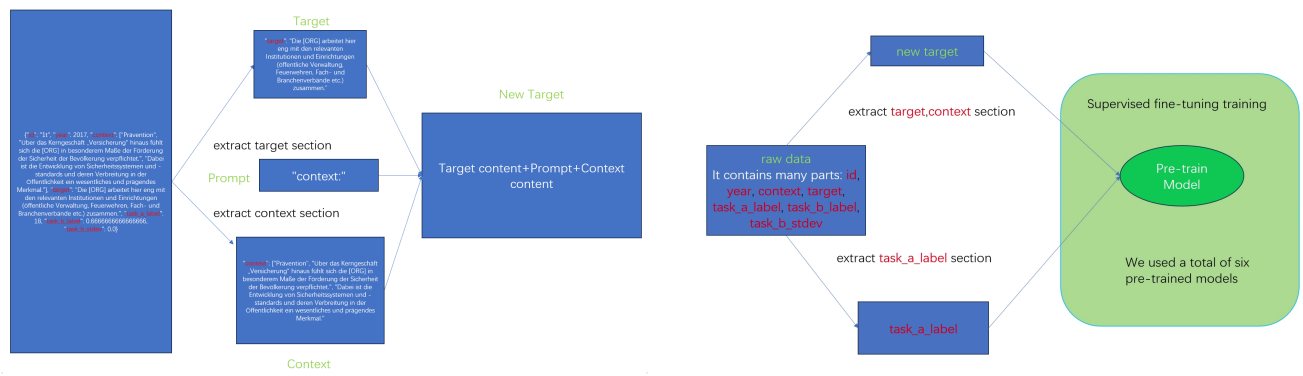

Figure 2: The $left$ figure is the text concatenation flow, and the $right$ figure is the text fine-tuning training process of six pre-train models in total.

all ALBERT models is uncased: it does not make a difference between english and English. The hyperparameters we provided to the model are shown in Table 4. The accuracy rate of this model on our validation set is 0.2849.

Table 4: Model hyperparameter Settings for $albert/albert-base-v2$

hyperparameter	**value**
require_improvement	10000
num_epochs	20
batch_size	8
pad_size	128
learning_rate	2×10^{-6}
hidden_size	768
dropout	0.2

Third, we choose the pre-trained model as google-bert/bert-base-german-cased[4]. It is pre-trained model as training data we used the latest German Wikipedia dump (6GB of raw txt files), the OpenLegalData dump (2.4 GB) and news articles (3.6 GB). The hyperparameters we provided to the model are shown in Table 5. The accuracy rate of this model on our validation set is 0.5341.

Table 5: Model hyperparameter Settings for $google-bert/bert-base-german-cased$

hyperparameter	**value**
require_improvement	1000
num_epochs	30
batch_size	8
pad_size	128
learning_rate	2×10^{-5}
hidden_size	768
dropout	0.2

Fourth, we choose the pre-trained model name is deepset/bert-base-german-cased-hatespeech-GermEval18Coarse[5]. It is latest pretrained model as a German BERT v1 (`https://deepset.ai/german-bert`) trained to do hate speech detection on the GermEval18Coarse dataset. The hyperparameters we provided to the model are shown in Table 6. The accuracy rate of this model on our validation set is 0.5530 .

Table 6: Model hyperparameter Settings for $asdeepset/bert-base-german-cased-hatespeech-GermEval18Coarse$

hyperparameter	**value**
require_improvement	1000
num_epochs	30
batch_size	8
pad_size	128
learning_rate	2×10^{-5}
hidden_size	768
dropout	0.2

Fifth, we choose the pre-trained model name is FacebookAI/roberta-large-mnli[6]. It is latest pretrained model as roberta-large-mnli is the RoBERTa large model fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus. The model is a pre-trained model on English language text using a masked language modeling (MLM) objective. The hyperparameters we provided to the model are shown in Table 7. The accuracy rate of this model on our validation set is 0.3939.

Sixth, we choose Sheripov/deid-roberta-i2b2-fine-tuned-german[7]. It is latest pretrained model as the RoBERTa large model fine-tuned on the problem type of Token Classification. The hyperparameters we provided to the model are shown in Table 8. The accuracy rate of this model on our validation set is 0.5076.

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

[5]https://huggingface.co/deepset/bert-base-german-cased-hatespeech-GermEval18Coarse

[6]https://huggingface.co/FacebookAI/roberta-large-mnli

[7]https://huggingface.co/Sheripov/deid-roberta-i2b2-fine-tuned-german

Table 7: Model hyperparameter Settings for $FacebookAI/roberta-large-mnli$

hyperparameter	value
require_improvement	10000
num_epochs	30
batch_size	8
pad_size	128
learning_rate	2×10^{-6}
hidden_size	1024
dropout	0.2

Table 8: Model hyperparameter Settings for $Sheripov/deid-roberta-i2b2-fine-tuned-german$

hyperparameter	value
require_improvement	10000
num_epochs	50
batch_size	8
pad_size	128
learning_rate	2×10^{-6}
hidden_size	1024
dropout	0.2

5 Result

The accuracy rates of these different pre-trained models on our official test set are shown in Table 9. the accuracy indicators were measured through the evaluation script and the gold standard data used for scoring.

As shown in Figure 3, we can compare the confusion matrix of the suboptimal prediction results submitted by our team with that of the prediction results of the best pre-trained model. It can be found that the prediction results of our best pre-trained model can classify the text more delicately in these 20 categories. This is reflected in the accuracy index rising from 0.5158 to 0.6353.

Table 9: The accuracy rates of these different pre-trained models on our official test set.

pre-train model	accuracy
dbmdz/bert-base-german-cased	**0.6353**
deepset/bert-base-german-cased-hatespeech-GermEval18Coarse	0.5158
albert/albert-base-v2	0.2649
google-bert/bert-base-german-cased	0.4587
acebookAI/roberta-large-mnli	0.3572
heripov/deid-roberta-i2b2- fine-tuned-german	0.3742

Below are the final rankings for the two sub-tasks, see Table 10 and Table 11. In the official ranking. Our team ranked 1st in the Task A-Content Classification with an Accuracy index of 0.635285412.

Table 10: Task A - Content Classification

Rank	Team Name	Accuracy
1	**wangkongqiang**	**0.635285412**
2	22520474	0.625792812
3	1234566	0.585623679
4	s1nbo	0.579281184
5	janpf	0.572938689
6	supachoke	0.486257928

Table 11: Task B - Verifiability Rating

Rank	Team Name	Kendall's Tau
1	janpf	0.40233707

6 Conclusion

The accuracy of these different pre-trained models on our official test set is not particularly outstanding. A major reason for this is that the pre-trained models cannot handle text features well. Since we provide the context of the text to the pre-trained models by concatenating the context content, it may result in the data content not being well utilized. Another point is that there are fewer pre-trained models in German than in English, and the models used are relatively limited. If large language models (LLMs) [32] such as GPT 4 and Llama 3 are used for fine-tuning, and then prompt word prediction is used, it is believed that the prediction can be more accurate. Better text feature extraction is also a future direction for improving the performance and accuracy indicators of the model.

References

[1] Jie Tao, Xing Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," Journal of Big Data (7), 1–26, 2020.

[2] Uladzimir Sidarenka, "Potts: the potsdam twitter sentiment corpus," In Proceedings of the Tenth In- ternational Conference on Language Resources and Evaluation (LREC'16), 1133–1141, 2016.

[3] Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, "A unified generative framework for aspect-based sentiment analysis," arXiv preprint arXiv:2106.04300, 2021.

[4] Rie Kubota Ando, Tong Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," Journal of Machine Learning Research (6), 1817–1853, 2005.

[5] Gerard Salton, Michael J McGill, "Introduction to modern information retrieval," McGraw-Hill, 1986.

[6] Yoon Kim, "Convolutional neural networks for sentence classification," In Proceedings of the 2014 Conference

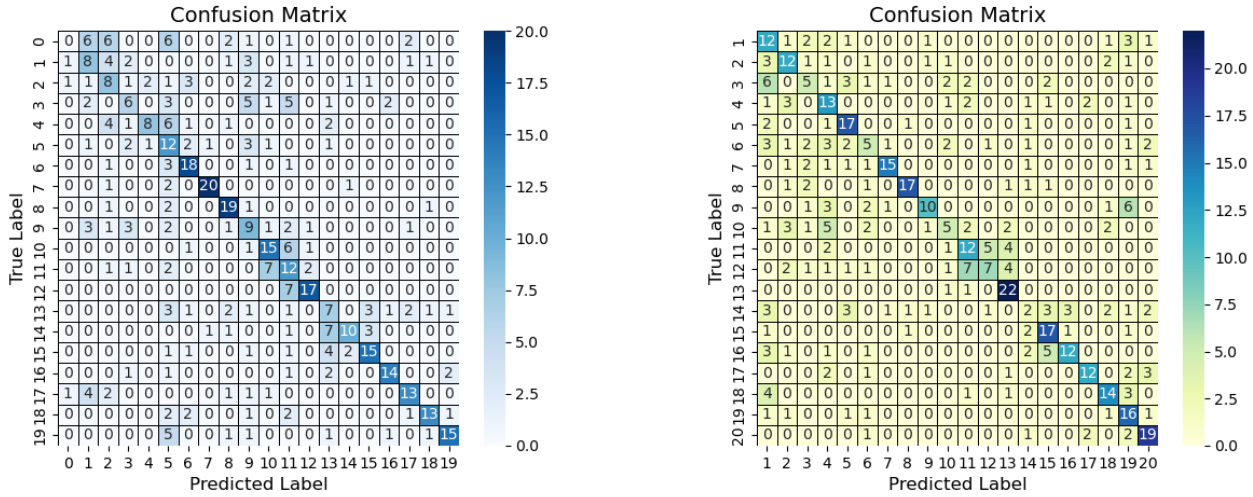

Figure 3: The *left* figure is the confusion matrix of the suboptimal prediction results submitted by our team, and the *right* figure is the confusion matrix of the prediction results of the best pre-trained model.

on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751, 2014.

[7] Sepp Hochreiter, Jürgen Schmidhube, "Long short-term memory," Neural computation, 9(8):1735–1780, 1997.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

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

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, "Deberta: Decoding-enhanced bert with disentangled attention," In International Conference on Learning Representations, 2021.

[11] Dogu Araci, "Finbert: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908.10063, 2019.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics,36(4):1234–1240 2020.

[13] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, Matthew McDermott, "Publicly available clinical bert embeddings," In Proceedings of the 2nd Clinical Natural Language Processing Workshop, 72–78, 2019.

[14] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, Ion Androutsopoulos, "Legal-bert: The muppets straight out of law school ," In Findings of the Association for Computational Linguistics: EMNLP 2020,2898–2904, 2020.

[15] Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, "Multi-task deep neural networks for natural language understanding," In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4487–4496, 2019.

[16] Ronan Collobert, Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," In Proceedings of the 25th international conference on Machine learning, 160–167, 2008.

[17] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification," In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 1480–1489. 2016.

[18] Sebastian Haunss, Jonas Kuhn, Sebastian Pado, Andre Blessing, Nico Blokker, Erenay Dayanik, Gabriella Lapesa, "Integrating manual and automatic annotation for the creation of discourse network data sets," Politics and Governance, volume 8(2), 2020.

[19] Travis Dyer, Mark Lang, Lorien Stice-Lawrence, "Language as a window into corporate culture," California Management Review, 59(3):56–69, 2017.

[20] Tim Loughran, Bill McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," The Journal of finance, 66(1):35–65, 2011.

[21] Katrin Hummel, Caroline Schlick, Matthias Fifka, "The role of sustainability performance and accounting assurors in sustainability assurance engagements," Journal of Accounting and Public Policy, 36(1):71–83, 2017.

[22] Aminul Islam Chy, Md Mahbubur Rahman, "Automated analysis of corporate sustainability reports," In Proceedings of the International Conference on Data Science and Applications, 145–158, 2021.

[23] Branden Chan, Stefan Schweter, Timo Möller, "German bert," arXiv preprint arXiv:2010.10906, 2020.

[24] Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, Martin Boeker, "Gottbert: a pure german language model," arXiv preprint arXiv:2012.02110, 2020.

[25] Julian Risch, Ralf Krestel, "Toxic comment detection in german," In Proceedings of the Conference on Natural Language Processing (KONVENS), 91–101, 2020.

[26] Gregor Wiedemann, Steffen Remus, Arpan Chawla, Chris Biemann, "Transfer learning for affective computing: A case study on valence-arousal prediction," In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 9–15, 2018.

[27] Jakob Prange, Charlott Jakob, Patrick Göttfert, Raphael Huber, Pia Wenzel, Annemarie Friedrich, "Overview of the SustainEval 2025 Shared Task," In Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops, Hildesheim, Germany. HsH Applied Academics, 2025.

[28] Grigorios Tsoumakas, Ioannis Katakis, "Multilabel classification: An overview," International Journal of Data Warehousing and Mining (IJDWM), 3(3):1–13, 2007.

[29] Yichun Yin, Yangqiu Song, Ming Zhang, "Document-level multi-aspect sentiment classification as machine comprehension," In Proceedings of the 2017 conference on empirical methods in natural language processing, 2044–2054, 2017

[30] Min-Ling Zhang, Zhi-Hua Zhou, "A review on multi-label learning algorithms," IEEE transactions on knowledge and data engineering, 26(8):1819-1837,2013.

[31] Telmo Pires, Eva Schlinger, Dan Garrette, "How multilingual is multilingual bert?" In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. Association for Computational Linguistics, 4996–5001, 2019.

[32] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, Lidong Bing, "Sentiment analysis in the era of large language models: A reality check," In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 1–15, 2023.