MeHuBe at SwissText 2024 Shared Task 1: Ensembling and QLoRA with Retrieved Citations for Fine-Grained Classification of Sustainable Development Goals

Fernando de Meer Pardo
University of Zurich
fernando.demeerpardo@uzh.ch

Hanna Hubarava
University of Zurich
hanna.hubarava@uzh.ch

Vera Bernhard
University of Zurich
veralara.bernhard@uzh.ch

Abstract

This study is a contribution to the SwissText Shared Task 2024, aimed at an improvement of the automatic classification of scientific abstracts related to the United Nations' Sustainable Development Goals (SDGs). Using Semantic Scholar API, we augment the small and highly imbalanced training set by retrieving additional abstracts from citations of the original dataset. The enriched dataset is then used to fine-tune a number BERT-based models as well as the Mistral-7B model, which is fine-tuned in a parameter-efficient way with QLoRA. Experimentation with various ensembling strategies reveals a positive influence of prediction confidence, with the best ensembling strategy yielding the accuracy of 77% and macro F1 of 77%. The submission occupied place 10 when evaluated on the Shared Task's test set.

Keywords: Sustainable Development Goals, SDG, NLP, QLoRA, BERT, ensembling, text classification, data augmentation

1 Introduction

This study is a team submission to Task 1 of the SwissText Shared Task 2024[1]. The Shared Task is a part of the Swiss Text Analytics Conference. The goal of the shared task is to improve upon the existing ways of automatic classification of scientific abstracts related to the United Nations' Sustainable Development Goals (SDGs)[2].

Natural Language Processing (NLP) has demonstrated significant potential in addressing various socio-economic challenges, from healthcare and education to climate change: For example, BERT models have been successfully used for identifying geographic entities in climate literature which enables monitoring evolution of climate issues (Mallick et al., 2024). Additionally, larger LLMs like Mistral-7B has been effectively applied to tasks such as medical dialogue systems (Zhao et al., 2024).

By leveraging the tools offered by Natural Language Processing, the Shared Task aims to build a better approach towards identification and analysis of academic research pertinent to sustainable development. This, in its turn, should help to both quantify and leverage the impact of academia on the progress in achieving these global goals.

The key contributions and findings of this study are the following:

- We study the effect of augmenting the SDG training data with weakly labeled citations fetched via the Semantic Scholar API.
- We fine-tune and evaluate 4 different BERT variants and an adapter for Mistral-7B with QLoRa.
- We experiment with various ensembling techniques for all the models we fine-tuned.

2 Dataset

2.1 Dataset Description

Task 1 focusses on the classification at the level of the SDGs, where each abstract is to be mapped to a single SDG number (class label). There are 18 classes in total: 17 SDGs and an additional class zero ("non-relevant"). The latter is reserved for the articles that should not be classified as relevant to any of the SDGs, posing a challenge and demanding a creative solution.

The final version of the train set includes 430 entries in jsonl format from the UZH publication repository Zora. It contains Zora ID, title, abstract, URL and a target SDG number. The dataset is extremely unbalanced, with nearly half of the train set represented by the four largest classes (life on land, good health and well-being, climate action, and non-relevant).

[1] see Appendix A for our code implementation and the software specifications used for this submission.

[2] see https://sdgs.un.org/goals

In addition to the train set provided by the Task organizers, there exists the OSDG Community Dataset (OSDG et al., 2021), which appears to be a popular resource in the research on automatic SDG classification (Pukelis et al., 2022).

2.2 Dataset Modifications

The train set history on GitHub[3] reveals two rounds of modifications between the task release and the submission deadline. The modifications of the train set during the Shared Task had a twofold impact: Firstly, they caused the participants to reconsider gold standard labels for the modified records and the features which could be most informative. Secondly, they pointed towards a potential lack of consensus among annotators regarding some problematic classes.

Crucially, the ad hoc modifications raised a concern whether the final split into the train and test was done truly randomly, i.e. preserving the original class distribution and in accordance with the same annotation guidelines. A chi-square test confirmed our suspicion that the train and test set are extremely unlikely to come from the same distribution (Chi2: 63.6, p-value: 2.63×10^{-7}). See Figure 2 and Appendix D for a more detailed analysis.

3 Methods

3.1 Data Augmentation

In order to acquire more data to fine-tune our models with, we resorted to the Semantic Scholar API (Kinney et al., 2023). For each original record, we gathered all of its citations and generated new records by concatenating their titles and abstracts. Each new record was then assigned the same label as the work it referenced, without any manual supervision.

This weak labeling approach carries some risks. Although scientific papers usually cite thematically related manuscripts as part of their related work, there remains a possibility that such overlaps may be unrelated to the SDG labels (see Section 6).

We hypothesize that following along "citation paths", we can better capture the trends within SDGs. For example, out of the 17 records labeled with SDG 7 (renewable energy) in the provided train set, 15 concern solar energy and water splitting, only one publication is on the topic of adaptive energy consumption, while the remaining record is related to cloud networks. Crucially, the majority of those records were published in 2018-2019, ignoring the developments and themes of other years and decades. By including the publications that build upon previous research, we strive to harvest the evolution of topics over time and thus mitigate the limitations of the train set.

3.2 Models

3.2.1 BERT Variants

The following BERT variants were selected for fine-tuning:

- **Multilingual BERT**: BERT pretrained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective (Devlin et al., 2018).
- **SciBERT**: BERT pretrained on a corpus of 1.14M papers, 3.1B tokens built through the Semantic Scholar API. Additionally, it has its own vocabulary (scivocab) fit to match the training corpus (Beltagy et al., 2019).
- **Aspect-Based SciBERT**: BERT pretrained to perform pairwise document classification on a dataset of citing and cited papers originating from the ACL Anthology (Ostendorff et al., 2020).
- **BioBERT**: BERT pre-trained on large-scale biomedical corpora consisting on PubMed abstracts and PMC full-text articles (Lee et al., 2019).

The selection of these models took into account the type of data used in their pretraining, which should make the models efficient at processing scientific and technical texts in our task.

3.2.2 Mistral-7B with QLoRA

Due to promising results from manual zero-shot experiment using Mistral-7B on the SDG classification tasks, we decided to leverage the semantic knowledge of this large language model and fine-tuned it on the classification task (Jiang et al., 2023). Since our available resources were limited to a single T4 GPU (see Appendix C), we opted for the parameter efficient fine-tuning approach QLoRA (Dettmers et al., 2024): It injects trainable low rank adapters into a frozen, quantized large language model and enables memory efficient fine-tuning by introducing a 4-bit float data type, using double quantization and preventing memory spikes.

[3] https://github.com/ZurichNLP/sdg_swisstext_2024_sharedtask

3.2.3 Ensembling

We experimented with the following ensembling methods:

- **Majority Voting:** The class voted for by the majority of the models in the ensemble is selected. In the case of a tie, the first one of the tied classes is chosen as a default.

- **Majority Voting with Tie Breakup via Logits:** Same as above, except ties are broken via adding the logits of the tied classes and choosing the one with the largest sum.

- **Soft Voting:** Adds all the logits of all the models and chooses the class with the largest sum.

4 Experiments

To be able to validate the performance of fine-tuned models, we performed a stratified partition across the SDG labels of the SwissText dataset into a train/test split with 1/3 and 2/3 of the records, respectively. The stratification ensures all SDG classes are present across the splits and their proportions are preserved.

In the following subsections, we describe the setup of each of the different experiments we carried out and eventually led us to our final methodology. We include the results of each experiment and the respective discussion in Section 5.

4.1 Data Augmentation Experiment

In order to estimate the effect on performance of the additional records described in Section 3.1 we fine-tuned the **Multilingual BERT** model with the following sets of records:

- **SwissText records:** Train split of the SwissText original records as described in Section 4.

- **Enlarged SwissText:** All the citations of SwissText records obtained via the Data Augmentation procedure as described in Section 3.1.

- **Original and Enlarged SwissText and OSDG:** Combination of the two datasets described above along with all of the OSDG records with an agreement score bigger than 0.5 and their citations.

Both enlarged datasets are heavily unbalanced, as the total number of citations is not equal across SDGs. In order to have balanced training datasets, we randomly sampled at most 1000 records for each SDG label, leading to datasets of 17k records in both instances.

4.2 Experiments with BERT Variants

We fine-tuned all BERT variants described in Section 3.2.1 for 5 epochs. In all cases we observed an increase in accuracy for each epoch. We report the results of the 5th epoch for all models. See Appendix B for the full set of fine-tuning hyperparameters employed.

4.3 Experiments with QLoRA

We fine-tuned an adapter for Mistral-7B with QLoRA for 4 epochs on the full combination of records. The hyperparameters and infrastructure specification can be found in Appendix C.

4.4 Experiments with Ensembling

As specified in Section 3.2.3, we ensemble all the fine-tuned models listed in 3.2.1 and 3.2.2. Having observed a positive correlation between prediction confidence and F1-score (see Figure 1), we opted for the inclusion of confidence (logits) into some of the ensembling experiments.

5 Results & Discussion

In this section, we discuss the results of each of the experiments previously discussed. All of the scores correspond to our test split (containing 2/3 of the train records) of the SwissText records described in Section 4[4].

5.1 Main Results

5.1.1 Data Augmentation Experiment

Our hypothesis, as described in 3.1, posited that additional labeled data could be retrieved via citations. This hypothesis finds support in the performance statistics, as measured on 2/3 of the original SwissText data, held out as our test set: As shown in Table 1, the overall accuracy increased by 23 percentage points and the macro F1 score even

[4] The SwissText dataset was modified with multiple records being relabeled during the course of development. As a consequence all scores should only be considered as orientative, some scores were obtained in the original and others on the relabeled version of the dataset, see Appendix D for details. We did not re-run all experiments due to time and hardware constraints.

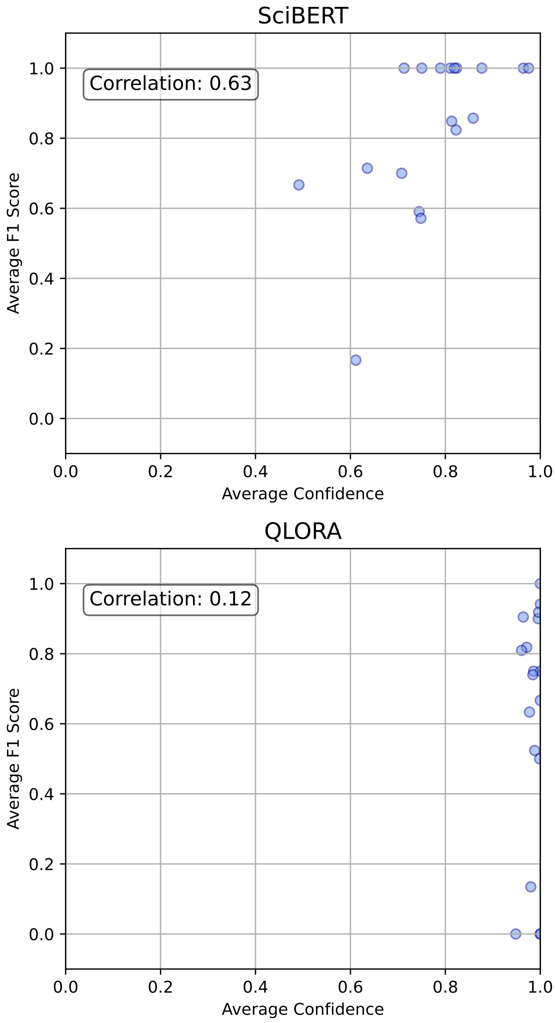

Figure 1: A positive correlation between average per-class F1 score and confidence of SciBERT suggest it could be beneficial to take logits into account during ensembling. QLoRA's extreme confidence in all predictions indicates it could outvote other models in the ensemble, if its logits are taken into account.

by 53 percentage points when the training set was augmented. Adding the OSDG records[5] and their citations further improved the overall accuracy by additional 8 percentage points.

5.1.2 Experiments with BERT variants

The four different BERT variants described in 3.2.1 were trained on the augmented dataset (original, enlarged SwissText, OSDG). Table 1 demonstrates that SciBERT performed the best, achieving an accuracy of 73% and macro F1 of 72%. The other BERT models followed closely, with Aspect-based SciBERT even surpassing SciBERT in terms of average precision.

5.1.3 Experiments with QLoRA

Despite the promising results of QLoRA reported by Dettmers et al., 2024, fine-tuning Mistral-7B on the augmented dataset did not achieve the performance level of BERT variants, lagging 21 percentage points behind the accuracy of SciBERT (see Table 1). This performance gap may be due to our limited resources for more extensive experimentation with QLoRA (see Section 6).

5.1.4 Experiments with Ensembling

Since the fine-tuned BERT variants outperformed Mistral-7B fine-tuned with QLoRA, they were ensembled via majority voting, resulting in slight increase of F1 by 1 percentage point compared to SciBERT, the best performing individual model (see Table 1). The experimental addition of QLoRA to the ensemble increased the accuracy to 75% and the macro F1 to 74%.

Leveraging the correlation between prediction probability and F1 score (as shown in Section 4.4) by breaking up ties via probabilities, the F1 score and accuracy rose to 76%-77%, depending on the inclusion of QLoRA.

Since QLoRA tends to predict with overly high confidence (see Figure 1), we applied soft voting only to an ensemble of the BERT variants, achieving performance comparable to majority voting with logits.

In sum, the experiments involving ensembling (see Table 1) demonstrate that ensembling and incorporating logits into the voting strategy is effective.

5.2 Error Analysis

As pointed out in Section 2.1, the provided train dataset is highly unbalanced. Some classes, either due to topical overlaps or a lack of detailed annotation guidelines (see Section 2.2), are particularly challenging to classify both for human annotators and the models (see confusion matrices in Section F of the Appendix).

Classes 0 (non-relevant), 8 (decent work and economic growth), 9 (industry, innovation and infrastructure), 16 (peace, justice and strong institutions) and 17 (global partnership for sustainable development goals) yield the lowest results across all models and ensembling configurations. SDG 9 (industry, innovation and infrastructure), for example, most commonly gets confused with SDG 12 (responsible consumption and production), reflecting

[5] With an agreement score > 0.5.

Model	Acc.	Pre.	Rec.	F1
Data Augmentation Experiment				
SwissText	0.37	0.08	0.06	0.03
SwissText big	0.60	0.51	**0.70**	0.56
SwissText + SwissText big + OSDG	**0.68**	**0.59**	**0.70**	**0.62**
Experiments with BERT variants				
mBert	0.68	0.59	0.70	0.62
SciBERT	**0.73**	0.68	**0.83**	**0.72**
Aspect SciBERT	0.72	**0.71**	0.76	0.70
BioBert	0.69	0.62	0.71	0.63
Experiment with QLoRA				
Mistral-7B	0.52	0.46	0.61	0.48
Experiments with Ensembling				
Majority w/o QLoRA	0.73	0.72	0.79	0.73
Majority with QLoRA	0.75	0.75	0.81	0.74
*Majority + Logit w/o QLoRA	0.76	**0.76**	**0.83**	**0.77**
*Majority + Logit with QLoRA	**0.77**	0.74	**0.83**	0.76
*Soft voting w/o QLoRA	0.76	0.75	0.82	0.76

Table 1: Results of all experiments. *Acc.* stands for overall accuracy, *Pre.*, for the precision averaged over all labels *Rec.*, for the recall averaged over all labels, *F1* for macro F1, i.e. F1 averaged over all labels. *SwissText big* is used synonymously for SwissText enlarged. **Bold** signifies the highest values within one experiment. The * marks the configurations chosen for the final submission. All experiments were evaluated on the stratified 2/3 of the original SwissText dataset, held out as our test set.

the semantic overlaps between industry, production and consumption. The confusion between the non-relevant class and SDG 3 (good health and well-being), analysed in more detail in the context of dataset modifications (see Section D of the Appendix), is also visible in model classification results, suggesting a more transparent distinction between medical science and general health topics would benefit the annotation.

5.3 Final Submission

Our final submission comprised three prediction runs with the following components:

- *MeHuBe_RUN1*: Ensemble with QLoRA using majority voting with logits.
- *MeHuBe_RUN2*: Ensemble without QLoRA using majority voting with logits.
- *MeHuBe_RUN3*: Ensemble without QLoRA using soft voting.

The above ensembles showed the most promising results when fine-tuned on the enriched dataset + 1/3 of the train data and evaluated on 2/3 of the train data, see Table 1.

The results of our runs on the test data of the shared task are shown in Table 2. With this results, we achieved place 10 in the Shared Task. The performance of our models on the shared task dataset was markedly lower than on the training dataset, reaching accuracies around 40% and F1 score of maximally 47%.

One reason for this performance gap may be the notable difference in non-relevant data proportions between the two sets. While non-relevant abstract built around a third of the original train set, they represented nearly a half of the test set (see Figure 2). As noted in Section 3.1 , classifying class 0 was particularly susceptible to annotation changes, with examples in Appendix D demonstrating incomprehensible decisions, suggesting broader issues with class 0. Furthermore, our models exhibited difficulties with class 0 already during the training phase, as documented in Section 5.2.

Our approach showed a better performance in

Run	Acc.	F1
Evaluation 1: main SDG		
MeHuBe_RUN1	0.39	0.42
MeHuBe_RUN2	0.38	0.41
MeHuBe_RUN3	0.38	0.41
Evaluation 2: secondary SDG		
MeHuBe_RUN1	0.42	0.47
MeHuBe_RUN2	0.42	0.45
MeHuBe_RUN3	0.41	0.45

Table 2: Results on the shared task test data. *Acc.* stands for overall accuracy, *F1* for macro F1, i.e. F1 averaged over all labels.

the evaluation 2 where secondary SDGs were also taken into consideration (see 2. This indicates that our approach occasionally selected secondary SDGs, a phenomenon observed in our error analysis (see Section 5.2), which identified topical overlaps in the dataset. Across the three runs, there was no notable performance disparity; they exhibited comparable results, mirroring our observations during training data evaluation.

6 Limitations

The scope of technical experimentation was restricted by limited computational resources. This pushed us towards fine-tuning compact BERT-based models, as well as parameter-efficient fine-tuning of Mistral-7B with QLoRA. With each QLoRA epoch taking around 18 hours, we were forced to refrain from exploring the effect of various dataset types and hyperparameters on the model performance.

As a consequence of the extremely unbalanced dataset, no cross-validation could be employed, since the smallest classes contained very few records and could not be split into k-folds. No hyperparameter optimization took place when fine-tuning BERT-based models, which should be taken into account if conducting further experiments.

The study would have also benefited from a thorough qualitative analysis of the records obtained via data augmentation. We hypothesize that the results could have been further improved through better preprocessing as well as pruning of the enriched dataset from irrelevant entries. Importantly, dataset enrichment with studies citing the given record only yields papers published later on. To gather trends and topics which preceded the record, one would also need to collect the publications referenced by that given record.

Due to a late publication of the gold standard labels shortly before the paper submission deadline, no qualitative analysis of the test set errors could be carried out. Lastly, the lack of access to the annotator guidelines limited our understanding of the definitions of classes. An analysis of errors and modifications of the train set points towards inconsistencies in labeling of some abstracts, e.g. those related to medicine and healthcare.

7 Conclusions

Our submission aims at implementing automatic classification of scientific abstracts related to the UN Sustainable Development Goals (SDGs). Leveraging an experimental data augmentation technique, specifically by integrating citations from the original dataset, we enriched the training data and subsequently improved classification performance. Ensembling various fine-tuned BERT-based models, including Mistral-7B fine-tuned with QLoRA, notably improved accuracy and F1 scores on the training data. Using QLoRA provided an interesting proof of concept of fine-tuning larger language models, yet its utility requires further experimentation in a setting with greater computational power. The modest performance of our approach highlights the complexity of SDG classification, particularly identifying abstracts as relevant and non-relevant and modeling it as a single-label task.

Acknowledgements

This paper was written not only as a submission for the Shared Task at SwissText 2024 but also as part of the course Machine Learning for Natural Language Processing 2 at the University of Zurich. We therefore thank the the University of Zurich for providing both the computational infrastructure and valuable feedback throughout the project, and the University of Applied Sciences (ZHAW) for providing supplementary GPU access.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv e-prints*, page arXiv:1903.10676.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform. *ArXiv*, abs/2301.10140.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746.

Tanwi Mallick, John Murphy, Joshua David Bergerson, Duane R Verner, John K Hutchison, and Leslie-Anne Levy. 2024. Analyzing regional impacts of climate change using natural language processing techniques. *arXiv preprint arXiv:2401.06817*.

OSDG, UNDP IICPSD SDG AI Lab, and PPMI. 2021. Osdg community dataset (osdg-cd).

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020. Aspect-based Document Similarity for Research Papers. *arXiv e-prints*, page arXiv:2010.06395.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Lukas Pukelis, Nuria Bautista-Puig, Gustė Statulevičiūtė, Vilius Stančiauskas, Gokhan Dikmener, and Dina Akylbekova. 2022. OSDG 2.0: a multilingual tool for classifying text data by UN Sustainable Development Goals (SDGs). ArXiv:2211.11252 [cs] version: 1.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. 2024. Chatcad+: Towards a universal and reliable interactive cad using llms. *IEEE Transactions on Medical Imaging*, pages 1–1.

A Software and Code

The code to train and evaluate our results can be found on GitHub: https://github.com/vera-bernhard/SwissTextSDG

In addition, the following Python packages have been used:

- PyTorch, pytorch-transformers (Paszke et al., 2017)
- Hugging Face's transformer, accelerate, peft, bitsandbytes (Wolf et al., 2019)
- NLTK (Bird et al., 2009)

B Hyperparemeters of BERT Variants Training

We fine-tuned the following pretrained Bert variants from Hugging Face (Wolf et al., 2019) with the hyperparameters as specified in Table 3:

bert-base-multilingual-uncased,
dmis-lab/biobert-base-cased-v1.2,
allenai/scibert_scivocab_uncased,
malteos/aspect-acl-scibert-scivocab-uncased

Fine-tuning each model for 5 epochs on the enlarged datasets required 1 hour of GPU time, utilizing a Tesla T4 GPU (16GB).

C QLoRA Training: Hyperparameters and Infrastructure

We had access to a single Tesla T4 GPU (16GB) and trained QLoRA with the hyperparameters in Table 4. The first training experiment took ~74h for 4 epochs, while training on the full dataset for the same number of epochs required ~95h.

Hyperparameter	Value
Learning Rate	2e-5
Learning Rate Scheduler	Warmup-LinearScheduler
Batch Size	32
Adam Epsilon	1e-8
Number of Epochs	5
Max Gradient Norm	1.0
Max Sequence Length	265

Table 3: Hyperparameters for BERT Variants Training

Hyperparameter	Value
General Training Parameters	
Learning Rate	2e-5
Learning Rate Scheduler	WarmupLinearScheduler
Batch Size	1
Adam Epsilon	1e-8
Number of Epochs	4
Max Gradient Norm	1.0
Max Sequence Length	265
PEFT Parameters	
Pretrained Model	mistralai/Mistral-7B-v0.1
Lora alpha	16
Lora dropout	0.1
Rank of low-rank	2
Total # Parameters	7,111,659,520
Trainable # Parameters	925,696

Table 4: Hyperparameters for QLoRA Training

D Dataset Modifications

The modifications in early March encompassed nine label changes, while major modifications of over 300 lines took place in April. The problematic classes which caused the largest number of modifications are the following: SDG 0 (non-relevant), 3 (good health and well-being), 5 (gender equality), and 16 (peace, justice and strong institutions). A notably large share of modifications concerned the non-relevant class (SDG 0), with seven out of nine first-round modifications related to this class.

Record 130 of the last train set edition serves as an example of the problematic classes and modifications: “Newspaper coverage of female candidates during election campaigns: evidence from a structural topic model”). The article is clearly concerned with gender inequality and bias, yet it was modified to be labeled as SDG 0 (non-relevant). Record 23 exemplifies another case of the modification towards non-relevant class. The article titled "Internal auditory canal volume in normal and malformed inner ears" is a medical publication on the topic of hearing abnormalities, yet it was modified as belonging to SDG 0.

E Correlation between Confidence and F1-score

A positive correlation between average per-class prediction confidence (logits) and macro F1-score could be observed in all BERT-based models (see Figures 1 and 3). As specified in Section 4.4, this observation led us to include confidence into some ensembling configurations, resulting in an improved performance.

F Confusion Matrices

The confusion matrices presented below are those of the best-performing model (SciBERT) and ensembling strategy (majority voting with QLoRA and logits). The models were trained on the enriched dataset and evaluated on the remaining 2/3 of the train set.

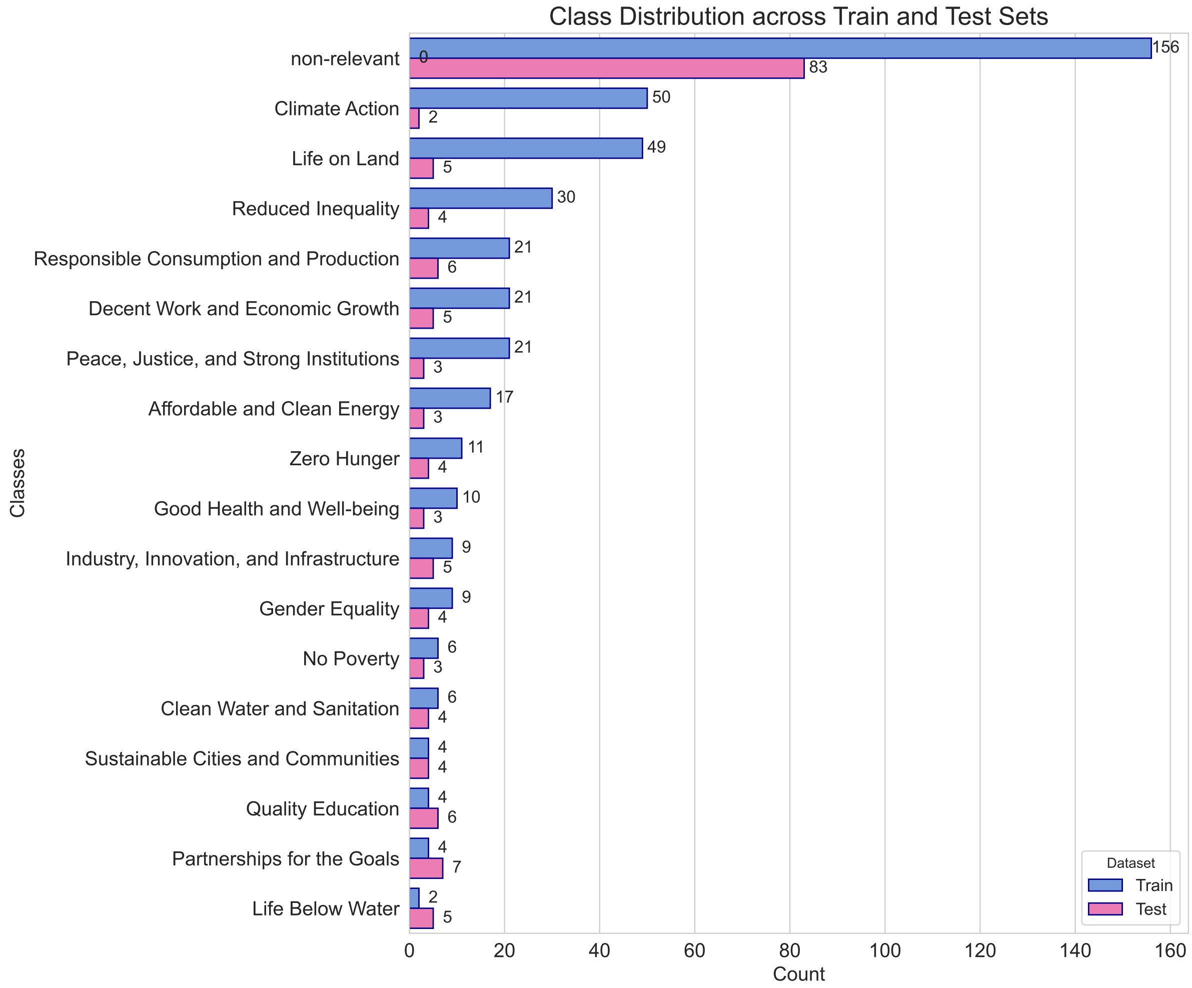

Figure 2: Class sizes in the train and test sets provided by the Shared Task, in decreasing order of class size in the train set.

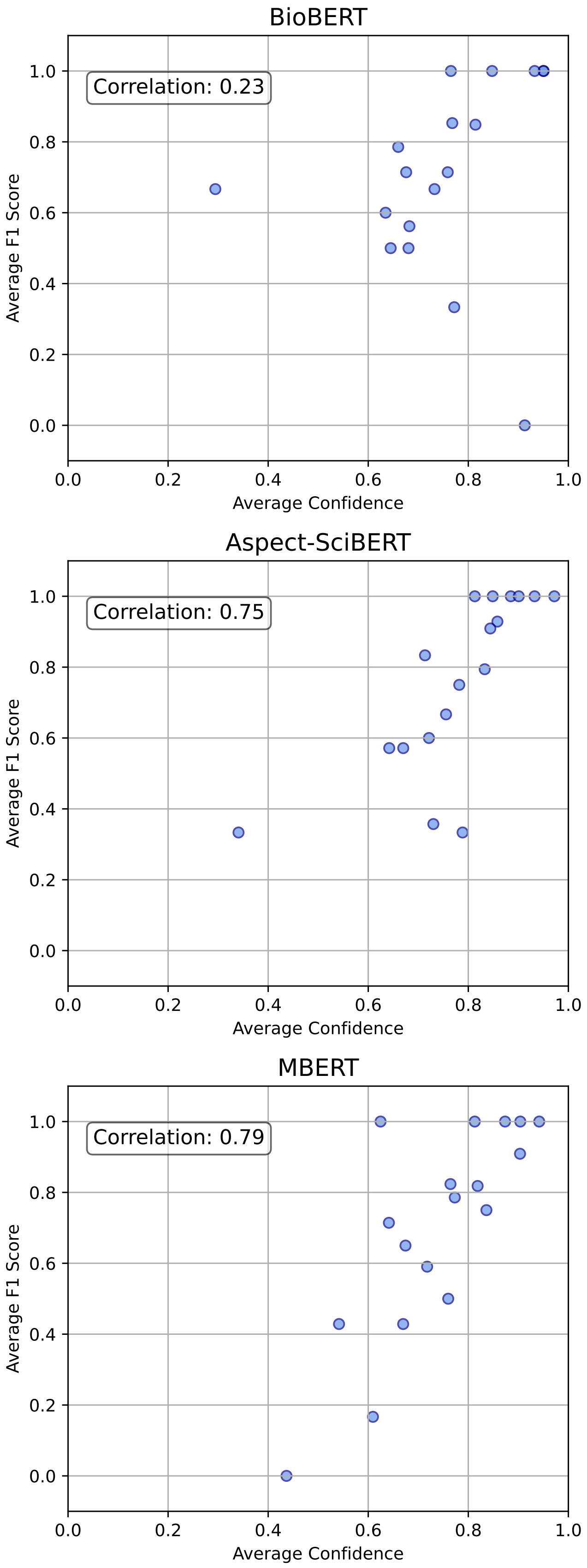

Figure 3: A positive correlation between model's prediction average per-class confidence and F1-score can be observed in all BERT-based models.

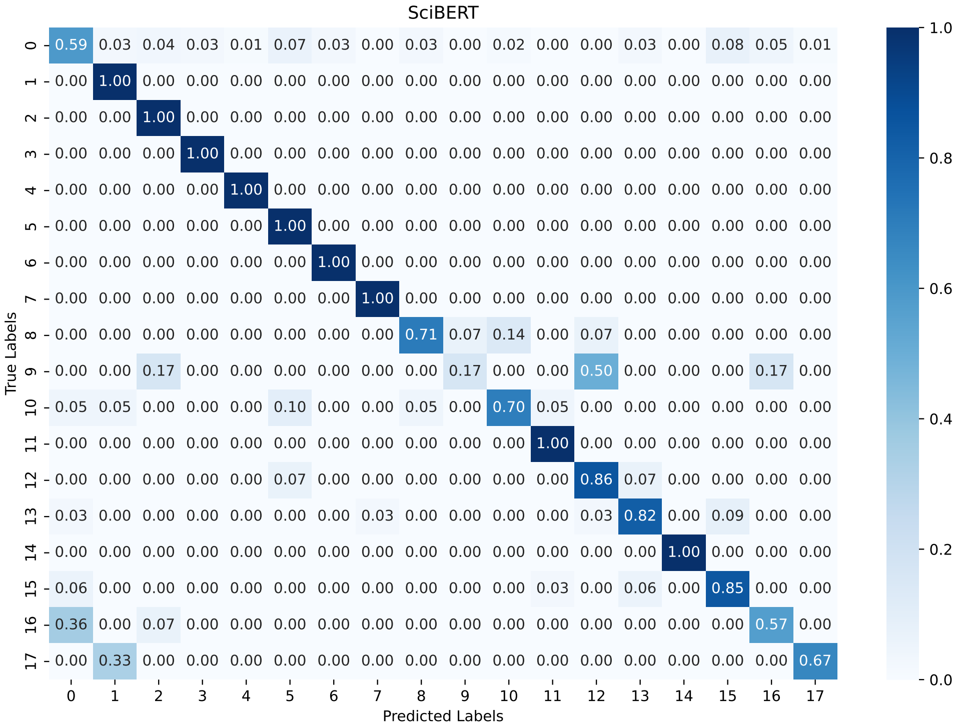

Figure 4: Normalized confusion matrix of SciBERT, the best performing individual model.

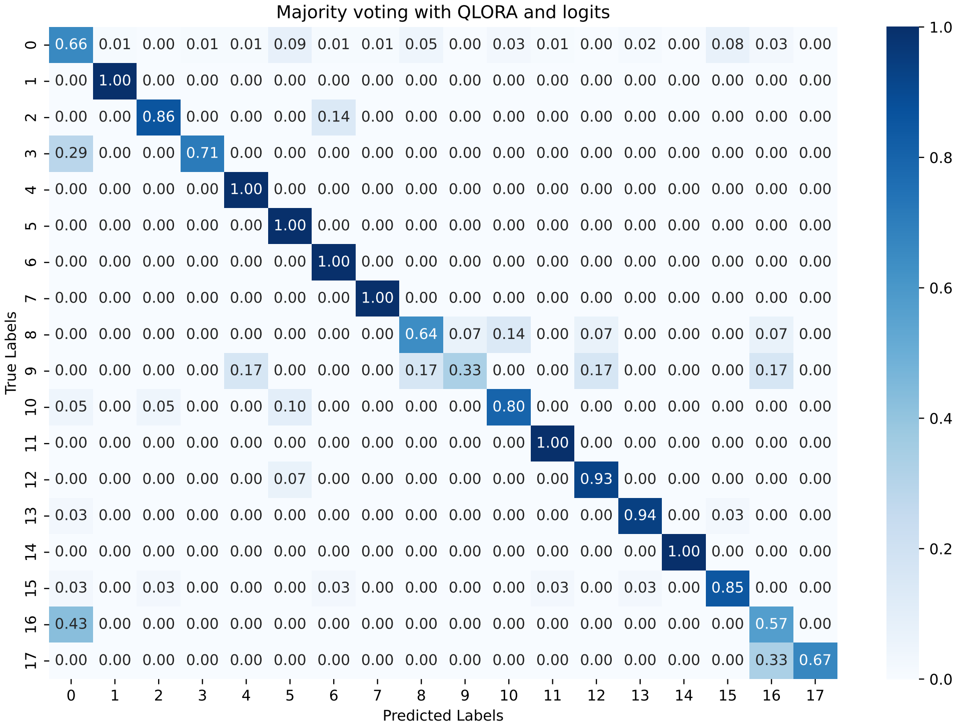

Figure 5: Normalized confusion matrix of the best ensembling strategy, including all BERT-based models and QLoRA and taking into account logits.