AttrPrompt in Action: Evaluating Synthetic Data Generation for SDG Classification

Manuel Bolz
University of Zurich
manuel.bolz@uzh.ch

Andreas Loizidis
University of Zurich
andreas.loizidis@uzh.ch

Kevin Bründler
University of Zurich
kevin.bruendler@uzh.ch

Abstract

This workshop paper presents the methodology and results of our participation in the SwissText Shared Task 2024, focusing on classifying scientific abstracts into Sustainable Development Goals (SDGs). To address data sparsity and class imbalance, we employed synthetic data generation using large language models, including GPT-4, Mixtral-8x22B, and Llama-3-70b. We utilized a domain-adjusted version of AttrPrompt (Yu et al., 2024) to generate 16,600 synthetic abstracts, leveraging models such as GPT-4, Mixtral-8x22B, and Llama-3-70b to fine-tune pre-trained SciBERT (Beltagy et al., 2019) and Muppet-RoBERTa (Aghajanyan et al., 2021) models. Our findings indicate that synthetic data significantly enhances model performance, though the optimal data generation model varies with the classifier. Notably, SciBERT (Beltagy et al., 2019) consistently outperformed Muppet-RoBERTa (Aghajanyan et al., 2021) across various metrics. The most human-like synthetic texts, generated by GPT-4, yielded the best performance. Our approach achieved third place in the shared task, demonstrating the potential of synthetic data in improving classification accuracy for complex, multiclass settings.

1 Introduction

To address critical global issues such as climate change, poverty, and inequality, all United Nations (UN) Member States have adopted the 2030 Agenda for Sustainable Development, encompassing 17 Sustainable Development Goals (SDGs) with diverse humanitarian, environmental, and developmental objectives. Each SDG includes several sub-targets representing the different facets of the 17 main SDGs. To facilitate the classification of research abstracts into SDGs, the University of Zurich's Sustainability Team has curated the ZORA dataset consisting of labeled abstracts, where each abstract is given one of the 17 SDG labels or 0 (null class), if the abstract does not relate to any of the SDGs. Such classification aids in understanding research trends, identifying knowledge gaps, and ultimately informing policy decisions aimed at addressing these pressing global issues.

Previous attempts at SDG document classification utilized the labeled data from the OSDG Community Dataset (OSDG et al., 2023). For example, Sadick (2023) has fine-tuned a BERT-based text classification model trained on OSDG data, available on Huggingface. However, the model currently only supports the first 16 goals and does not contain a null class. Extending this, Roady (2023) explored various data configurations and language models to classify SDG labels in scientific abstracts with variable success, primarily caused by data sparsity, class imbalance, and vague class definitions, while also omitting a null class.

This paper investigates whether synthetic data generated by large language models (LLMs) can enhance model performance in multiclass classification tasks characterized by sparse and imbalanced data with poorly separated classes. While transformer-based models have shown promising results in text classification, they frequently struggle with generalization, particularly when confronted with limited data for certain classes and label noise.

Our approach builds on previous findings that synthetic data can improve classification accuracy on multiclass settings, particularly when certain classes are rare (Kochanek et al., 2023; Møller et al., 2024). We aim to employ LLM-generated synthetic data to expand the training dataset, thus improving the model's capacity to learn from varied and representative examples across all SDG classes. By systematically evaluating the effectiveness of this approach when human-labeled data is not only sparse and imbalanced, but also suffers from label noise, we seek to contribute insights into improving the robustness and generalization capabilities of SDG document classification models.

We apply a domain-adjusted version of AttrPrompt (Yu et al., 2024) to increase representation of underrepresented SDG classes. AttrPrompt enriches a prompt with a range of domain-specific attributes to generate synthetic data points and has demonstrated superior performance to simple class-conditional prompts. We test three models— GPT-4, Mixtral-8x22B, and Llama-3-70b—to generate 16,600 synthetic abstracts each to fine-tune pretrained SciBERT (Beltagy et al., 2019) and Muppet-Roberta (Aghajanyan et al., 2021) models. We evaluate the performance of each model against a baseline model trained on the OSDG and ZORA dataset.

We find that the SciBERT model trained on synthetic data generated by GPT-4 performs best reaching an accuracy of 0.47. In both models, including synthetic data moderately increases the accuracy and generally, SciBERT outperforms Muppet-RoBERTa. However, accuracy stays overall low therefore leaving room for alternative approaches.

2 Methodology

2.1 Synthetic Data

To generate the synthetic data, we applied a three-step prompt to generate the attributes configuration for each SDG:

1. *Which 20 web-of-science research domains will most likely be related to the UN SDG goal number {sdg_id}: {description}?*

2. *Fill in the following structure for studies on the UN SDG goal number {sdg_id} with 10 diverse sub-topics per research domain: {json_structure}*

3. *Analyze the following research domains likely to contain studies on the UN SDG goal number {sdg_id} in terms of completeness. If there is a web-of-science research domain missing that could contain such studies, please generate these domains including 10 diverse sub-topics. Return the generated content in a json structure as shown in the following input: {json_structure}*

This process ensures a comprehensive set of potential research areas, each with multiple sub-topics, resulting in approximately 23 research areas and 230 sub-topics per SDG.

Further, we specify attributes such as *length*, *style*, and *abstract start*, which are described in the appendix. We utilize GPT-4, Llama-3-70b, and Mixtral-8x22B to generate synthetic abstracts by randomly combining these attributes from 34,500 potential combinations using the following prompt:

Write an abstract of a scholarly article from the Web of Science database concerning {main_topic}. Ensure the abstract:

1. *Aligns subtly with the themes of the UN SDG goal {sdg_goal}, though without explicit mention of the goal itself;*
2. *Focuses on '{subtopic}';*
3. *Starts by {abstract_start}*
4. *Is between {length} and {int(length) + 60} words in length*
5. *Reflects a study that {style}*

We excluded SDG 16 from synthetic data generation due to its over-representation in the OSDG and ZORA datasets. Post-generation, all synthetic abstracts were cleansed of any LLM-specific artifacts such as *"Here is a potential abstract:"* or *"(narrative hook)"*, commonly found in outputs from Llama-3-70b and Mixtral-8x22B.

2.2 Null Class

To ensure robustness and validate the specificity of classification, we generated a null class consisting of abstracts unrelated to any SDG topics. This process involved several steps. Initially, we utilized GPT-4 to generate topics that are as unrelated as possible to any SDG by querying it with SDG labels and definitions. We then used these unrelated topics to scrape paper abstracts from Semantic Scholar.

Next, we conducted topic modeling on the collected abstracts. The text data was preprocessed, and we applied TF-IDF vectorization to extract relevant features. Using Non-negative Matrix Factorization (NMF), we identified prominent topics for each SDG and determined the most significant words associated with these topics.

To create the null class, we identified and excluded any abstracts containing specific keywords revealed by the topic modeling.

2.3 Data and Splits

The data used for training comes from three primary sources: the OSDG dataset, the synthetic dataset detailed in the previous subsection, and the given shared task training set. The synthetic dataset included 1,000 samples for every class but 16 and 17. Class 16 was excluded as mentioned before, and for class 17 we generated 1,600 samples to compensate for it not appearing in the OSDG dataset. Synthetic data for the classes for each model amounts to 16,600 total samples[1]. We generated 2185 samples for the null class.

To internally evaluate the models and choose the best hyperparameters, we used a 80/20 stratified train/test split, ensuring that at least two abstracts per class from the given dataset were included in the test set, maintaining class balance and representation. For the final submission as well as the experiments shown here, we used a 95/5 split for training, and the released test set for evaluation.

3 Experiments

3.1 Models

We considered two transformer models: SciBERT and Muppet (Massive Multi-task Representations with Pre-Finetuning) RoBERTa. SciBERT (Beltagy et al., 2019) is a variant of BERT pre-trained on a large corpus of scientific text, making it particularly suitable for academic and research-based tasks. Muppet (Aghajanyan et al., 2021) is a pre-finetuned variant of RoBERTa, trained using multi-task learning to enhance its performance across various natural language processing tasks.

3.2 Finetuning

In all settings, we apply a dropout of 0.1 and optimize cross-entropy loss using Adam (Kingma and Ba, 2017). We experiment with different hyperparameter settings and fine-tune the models for 2 to 5 epochs using batch sizes of 16 and 32, and a learning rate of 5e-6, 1e-5, 2e-5, or 5e-5 with a slanted triangular schedule (Howard and Ruder, 2018), which is equivalent to the linear warmup followed by linear decay (Devlin et al., 2019). For each dataset and BERT variant, we pick the best learning rate and number of epochs on the development set and report the corresponding test results. We found that the setting that works best across most datasets and models is 2 epochs, a batch size of 16, and a learning rate of 2e-5. While task-dependent, optimal hyperparameters for each task are often the same across BERT variants.

3.3 Results

Model	Accuracy	F1 Score
muppet-no-synth	0.38	0.27
muppet-llama	0.38	0.34
muppet-gpt-4	**0.42**	**0.41**
muppet-ensemble	0.40	0.41
muppet-mixtral	0.39	0.40
scibert-no-synth	0.38	0.33
scibert-llama	0.43	**0.47**
scibert-gpt-4	**0.47**	0.44
scibert-ensemble	0.45	0.43
scibert-mixtral	0.45	0.45

Table 1: Accuracy and Avg. F1 Score per Model

Table 1 showcases the performance of each model variant in terms of accuracy and F1 score. The Muppet variants showed relatively similar performances in terms of accuracy, with the Muppet-GPT-4 achieving the highest accuracy and F1 score at 0.42 and 0.41, respectively. Compared to the baseline model Muppet-no-synth, most variants show a slight improvement in accuracy.

On the other hand, SciBERT generally performed better, particularly the SciBERT-GPT-4 for the highest accuracy at 0.47 and the SciBERT-Llama for the highest F1 Score. The consistently higher performance across different datasets suggests a the SciBERT model is better suited for tasks concerning scientific text, likely benefiting from its training on a scientific corpus. Overall, SciBERT models generally outperformed Muppet models on the same data, indicating a possible advantage in handling task-specific nuances.

4 Conclusion

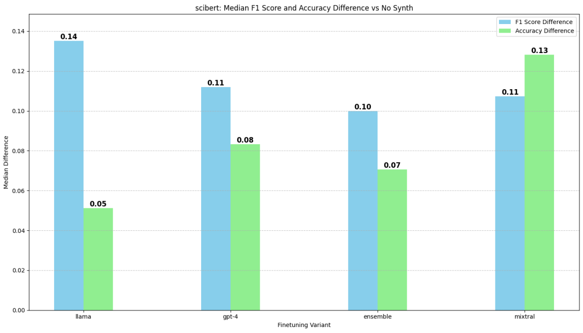

Figure 1: Median Improvement of Accuracy and F1 Score of the SciBERT model by Adding Synthetic Data

The results demonstrate that synthetic data significantly enhances performance in the multi-class

[1] We were only able to generate 16,518 samples using Mixtral.

classification task. However, determining the most effective model for data generation remains inconclusive. Notably, the Muppet model exhibited the highest boost in F1 score when trained with synthetic data generated by the GPT-4 model. In contrast, SciBERT achieved better performance with synthetic data from llama. This variation suggests that the optimal choice of synthetic data generation model might be contingent upon specific model architectures and their inherent characteristics.

The most coherent and human-like synthetic texts, as assessed by the authors, were generated by GPT-4. In line with preliminary expectations, the GPT-4 generated data did yield the best classification performance for both Muppet and SciBERT in terms of accuracy. This indicates that the quality of synthetic data, in terms of human-likeness and coherence, could correlate with improved model performance. However, the effectiveness of synthetic data appears to be influenced by how well the generated data aligns with the specific characteristics and requirements of the target model.

Our findings contradict the general wisdom that "there is no data like more data." The ensembling of synthetic data from different models did not result in the largest F1 boost, suggesting that simply increasing the volume of synthetic data does not automatically enhance performance. It underscores the importance of the quality and compatibility of the synthetic data with the specific model being used.

Furthermore, our experiments underline the potential of leveraging large language models to mitigate issues of data sparsity and class imbalance in multiclass classification. The generated synthetic data contributed to noticeable improvements across several evaluation metrics, indicating its viability as a supplementary resource in training robust classification models.

While our study highlights the benefits of synthetic data, it also opens avenues for further research. Future work could explore a broader range of language models for synthetic data generation and investigate the underlying factors contributing to the varying performance boosts across different models. Additionally, a deeper examination of the attributes and configurations used in synthetic data generation could offer insights into optimizing these processes for enhanced classification outcomes.

Overall, our findings advocate for the integration of synthetic data into training pipelines, especially in scenarios with limited labeled data. This approach not only augments model performance but also aligns with the growing trend of using advanced language models to address complex challenges in natural language processing tasks.

Acknowledgements

The authors would like to thank the organizers of the SwissText Shared Task 2024 and the UZH Sustainability Team for providing the data.

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *CoRR*, abs/2101.11038. ArXiv preprint arXiv:2101.11038.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3615–3620. Association for Computational Linguistics.

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. ArXiv preprint arXiv:1801.06146.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. ArXiv preprint arXiv:1412.6980.

Mateusz Kochanek, Przemysław Kazienko, Jan Kocon, Igor Cichecki, Oliwier Kaszyca, and Dominika Szydło. 2023. Can Innovative Prompt Engineering with ChatGPT Address Imbalances in Machine Learning Datasets? *Authorea Preprints*. Authorea preprint.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192.

OSDG, UNDP IICPSD SDG AI Lab, and PPMI. 2023. OSDG community dataset (OSDG-CD). `https://zenodo.org/records/7540165`. Zenodo.

Jessica Saemi Roady. 2023. Automatic classification of academic papers according to the UN sustainable development goals – an interdisciplinary perspective. Master's thesis, University of Zurich, Zurich, 12. Supervisor: Dr. Simon Clematide, Department of Computational Linguistics.

A. M. Sadick. 2023. SDG classification with BERT. https://huggingface.co/sadickam/sdg-classification-bert.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. *Advances in Neural Information Processing Systems*, 36.

A Appendix

A.1 Attributes: Abstract Start

1. Posing a question to frame the abstract in an engaging manner.
2. Mentioning the methodology used in the study.
3. Highlighting the significance or novelty about the research.
4. Using a narrative hook to grab attention.
5. Outlining the purpose or objective of the study.

A.2 Attributes: Length

1. 40
2. 100
3. 160

A.3 Attributes: Style

1. Tests hypotheses by manipulating variables to establish cause-and-effect relationships, using controlled experiments.
2. Constructs and articulates abstract concepts to develop theoretical frameworks for real-world application.
3. Compiles and evaluates existing research to summarize findings and highlight research gaps and patterns.
4. Provides an in-depth analysis of a specific event or individual to understand underlying principles.
5. Observes and describes phenomena as they occur naturally, detailing the observed features without manipulation.
6. Investigates relationships between variables to assess the strength and direction of associations.
7. Observes the same subjects over time to document changes and trends.
8. Gathers data from a population at a single time point to provide a snapshot of various characteristics.
9. Collaboratively addresses real-world problems, combining research with practical action for iterative improvements.
10. Integrates qualitative and quantitative methods to leverage their strengths for comprehensive insights.

A.4 Example Attributes: Main Topic SDG 1

1. Development Studies
2. Economics
3. Social Sciences - Interdisciplinary
4. Sociology
5. Environmental Science
6. Public, Environmental & Occupational Health
7. Anthropology
8. Political Science
9. Geography
10. Urban Studies
11. Education & Educational Research
12. Business & Economics
13. Agricultural Economics & Policy
14. Psychology - Applied
15. Law
16. Social Work
17. Demography
18. Health Care Sciences & Services

19. International Relations
20. Energy & Fuels
21. Human Geography
22. Behavioral Economics
23. Public Health

A.5 Example Attributes: Sub-topics SDG 1, Development Studies

1. Impact of microfinance programs on rural poverty
2. Effectiveness of conditional cash transfers
3. Role of technology in poverty alleviation
4. Sustainable livelihood frameworks
5. Community-based development projects
6. Urban vs. rural poverty dynamics
7. International development aid effectiveness
8. Gender and poverty
9. Education’s impact on poverty reduction
10. Poverty and climate change resilience