An Efficient Workflow Towards Improving Classifiers in Low-Resource Settings with Synthetic Data

Adrian M.P. Braşoveanu[1,2], Albert Weichselbraun[3,4], Lyndon J.B. Nixon[1,2] & Arno Scharl[1,3]
[1] Modul University Vienna, Am Kahlenberg 1, 1190, Vienna, Austria
[2] Modul Technology GmbH, Am Kahlenberg 1, 1190, Vienna, Austria
[3] webLyzard technology gmbh, Liechtensteinstraße 41/26, 1090 Vienna, Austria
[4] University of Applied Sciences of the Grisons, Pulvermühlestrasse 57, 7000 Chur, Switzerland
{adrian.brasoveanu,lyndon.nixon,arno.scharl}@modul.ac.at
albert.weichselbraun@fhgr.ch

Abstract

The correct classification of the 17 Sustainable Development Goals (SDG) proposed by the United Nations (UN) is still a challenging and compelling prospect due to the Shared Task's imbalanced dataset. This paper presents a good method to create a baseline using RoBERTa and data augmentation that offers a good overall performance on this imbalanced dataset. What is interesting to notice is that even though the alignment between synthetic gold and real gold was only marginally better than what would be expected by chance alone, the final scores were still okay.

1 Introduction

Automated SDG classification, the main topic of the SwissText 2024 Shared Task 1, is one of the most interesting research topics in light of the United Nations Agenda 2030. The goals offer a holistic approach to global challenges covering various issues from hunger and lack of healthcare to energy security and well-being. Alignment with these goals helps create targeted policies and investments in critical sectors while simultaneously enhancing accountability and measurements. The goals are the core of promoting global partnerships and cooperation that drive innovation and guide educational and awareness efforts. Additionally, goals are broken down into specific targets (typically between 8 and 12) to make them actionable for public and private organizations around the globe.

The correct classification of the 17 Sustainable Development Goals (SDG) proposed by the United Nations (UN) is still a challenging and compelling prospect, especially when coupled with an imbalanced dataset, as it happens in the SwissText 2024 Shared Task 1. The main objective was to classify scientific abstracts based on their relevance to one of the 17 SDGs or tag them as non-relevant if they cannot be aligned with any of the SDGs. In a certain sense, due to the addition of non-relevant cases, the task incorporated both classification and alignment, making it more difficult than a pure classification task.

The primary goal of the Shared Task was to evaluate the accuracy and overall performance of automated systems for classifying scientific abstracts with the appropriate SDG classes. An additional class was added to cover situations for which it was difficult to select a proper SDG class. This split between relevant and non-relevant classes was a novel element for this type of classification. A second element that made the task worthy of pursuing was the fact that the training data was quite imbalanced. The task asked for methods to help classify documents in low-resource settings.

The classes with the highest number of examples were the non-relevant class (SDG 0 with 156 examples, approximately one-third of all examples), followed by SDGs 13 (40 examples) and 15 (49 examples). On the lower end of the spectrum, four classes included less than five examples: SDG 4, 11 and 17 with four examples each, and SDG 14 with as few as three. These imbalances between the classes and the lack of training data led to the idea of using data augmentation techniques to improve the results. The selected approach helps mitigate the impact of class imbalance and can work across different datasets.

The rest of the paper is organized as follows: method and related work are presented in Section 2, Section 3 presents the evaluation results, which are then discussed in Section 4. The paper concludes with future work and insights on enhancing the process of automatically classifying imbalanced datasets.

2 Method

The classification required for this Shared Task is built a little bit differently than the usual SDG classification, as it used 18 classes instead of the expected 17, the additional class being used for non-relevant examples. It was not explicitly defined if these non-relevant examples should be examples that could fit into multiple categories or simply examples that do not belong to any SDG categories. Obviously the second category was much larger than the first one, therefore the non-relevant examples we included in our synthetic datasets mostly belonged to this category. As already mentioned, the high number of non-relevant examples contributed significantly to the severe imbalances found in the dataset.

Based on literature, BERTopic (Grootendorst, 2022) coupled with RoBERTa (Liu et al., 2019) or SetFit (Tunstall et al., 2022) seem to be provide quick and good solutions for any type of classification, especially when MPNet embeddings are used (Sayed et al., 2023).

Due to the fact that the dataset is imbalanced, only several solutions are deemed to perform well, zero-shot or few-shot learning (Brown et al., 2020) or data augmentation (Feng et al., 2021) being the first two that come to mind. Our paper focuses on data augmentation, and given the limited time allocated to the Shared Task, it seemed a rather good option.

We started by looking for a set of topics that could trigger the various SDGs. A recent study by Amel et al. (Amel-Zadeh et al., 2021) suggested a list, but while we examined it, we realized that there was some overlap between the SDGs, as it became quite clear that topics like "innovation", "energy", or even "economic aspects" tend to span across multiple SDGs. While their approach indicates higher scores are possible when applying this method, we wanted a fully automated approach that required no additional lists. This was the main reason we decided to focus on data augmentation.

Since the launch of ChatGPT, the number of articles on data augmentation techniques has increased exponentially. Reviewing all of them in such a short timespan would have been impossible. Therefore we limited ourselves to a study about pre-ChatGPT data augmentation techniques (Feng et al., 2021) and surveying modern articles about the impact of reasoning strategies on NLP task (Wadhwa et al., 2023). We surveyed two ideas related to data augmentation: i) using LLMs to generate synthetic data and ii) using existing datasets for augmentation. Each approach came with its own set of issues. For the LLM approach, it was clear that the data quality would be an issue, as it could increase or decrease over time. For the second approach, we felt strongly that open datasets like OSDG (Pukelis et al., 2020, 2022) could provide a good solution. While the quality of the datasets is certainly much higher than the current generation of LLMs, and their construction is well covered through their open papers and a series of notebooks, we have quickly uncovered some issues. The annotation process for these datasets followed strict rules, but they did not include out-of-domain cases or non-relevant cases.

After carefully examining both approaches, we decided to test the LLM data augmentation with GPT 4.0 for this shared task, as we wanted to see whether it could help us achieve good results quickly. Drawing upon data augmentation also made the results more interesting, since the chosen approach can be easily adapted to other datasets.

We tested several models before submitting the three requested runs. The examined models included: BERT (Guisiano et al., 2022), SVMs (Morales-Hernández et al., 2022), and RoBERTa (Liu et al., 2019).

The best performing approach extended a RoBERTa base model (125M parameters, similar to BERT) with dropout and fully connected layers. Adding the dropout layer set to 0.5 helped us prevent overfitting. The fully connected layer mapped the model output to the 18 classes. The model was evaluated using a 5-fold stratified cross-validation strategy to ensure a similar number of samples of each target class in each fold. For each fold, the model was trained for five epochs using the AdamW optimizer with a learning rate of 1E-5. The experiments were run in Google Colab Pro using L4 GPUs.

3 Evaluation Results

The synthetic dataset was generated using GPT 4.0 and classic prompts, which included 12 examples after the prompt (e.g., "Please annotate the following documents with their corresponding SDG class") (Wadhwa et al., 2023).

Since the labelled test results were not available upfront, we created a labelled test set using the same procedure we used for the augmentation. While this test set differed significantly from the dataset published after the Shared Task's conclusion, we considered it necessary to help select the best runs for the submission. A discussion related

Table 1: Results for the primary metric — correct prediction for primary SDG ordered by accuracy. P, R, F1 represent precision, recall and F1 metrics.

Run	Accuracy	Macro P	Macro R	Macro F1	Weighted P	Weighted R	Weighted F1
run3	0.49	0.65	0.66	0.56	0.74	0.49	0.53
run1	0.46	0.61	0.58	0.51	0.69	0.46	0.49
run2	0.40	0.58	0.61	0.51	0.64	0.40	0.42

Table 2: Results for the secondary metric - average F1 score per SDG. P, R, F1 represent precision, recall and F1 metrics.

Run	Accuracy	Macro P	Macro R	Macro F1	Weighted P	Weighted R	Weighted F1
run3	0.52	0.68	0.67	0.59	0.75	0.52	0.55
run1	0.50	0.66	0.63	0.58	0.69	0.50	0.53
run2	0.43	0.62	0.63	0.56	0.65	0.43	0.45

to these differences is included in Section 4.

One of the three submitted test runs was found to offer the best average performance on this imbalanced dataset. Although all submissions scored above average in both evaluation settings, the variability between runs suggests that improvements can be made.

Table 1 showcases the performance obtained by the submitted runs (named run31 to run33) for the primary metric, which optimizes for accuracy. Table 2 presents the results obtained for the secondary metric, i.e., the average F1 score per class. One of our runs (run33) obtained the best performance for the secondary metric from all submitted runs. The runs were submitted using the cover name: test_roberta_base_synth_TASK1_RUN31 to RUN33, which included the name of the dataset (test), model (Roberta base), method (synth), task (TASK1), and run (RUN31, RUN32 and RUN33). The evaluation reports can be found on the Shared Task's GitHub folder [1].

4 Discussion

We see at least several avenues for improving the data augmentation strategies. Perhaps the most obvious one is using a modern reasoning strategy like Chain-of-Thought (CoT) (Wei et al., 2022). Adding a justification for each example generated by the LLM would have further improved the quality of the synthetic dataset and, therefore, led to even better classification results (Wadhwa et al., 2023).

Such techniques are known to work better for LLMs or larger Transformer models, which tend to generalize better. Consequently, we didn't use this approach in our experiments, as we relied upon the smaller roberta-base model (125M parameters) due to time restrictions.

Given the counts for the relevant (SDGs 1 to 17) and non-relevant (SDGs marked as 0) classes, the classification results would have been considerably better if we had started with a binary classifier to separate relevant from non-relevant classes.

As outlined in Table 3, more than half of the test examples were non-relevant. This severely skewed the results since LLMs tend to overfit. In fact, an LLM will not be able to reproduce this setting a priori unless it is made more transparent through a detailed prompt (e.g., by adding a line like: "Please be aware that half of the examples I will ask you to annotate will be non-relevant").

An evaluation of the augmented data revealed that the synthetic dataset was missing the non-relevant class. The LLMs failed to produce examples for the non-relevant class despite being instructed to provide examples for all 18 classes (which includes class 0 for non-relevant cases). This error suggests that data augmentation is still the way to go, as even with all the errors that followed, the results were still the most balanced.

We assume that in a real-world setting, non-relevant entries will likely be even more prominent than in the provided dataset, which amounted to approximately one-third of the provided training examples and one-half of the provided test examples. In addition, overlaps between various classes

[1] https://github.com/ZurichNLP/sdg_swisstext_2024_sharedtask/tree/main/evaluation

are probably also more likely in a production setting.

While the length of the abstract was not considered a key parameter for our prompt, it is important to notice that the generated abstracts from the synthetic dataset were, on average, shorter than the ones from the real repository used for collecting the abstracts for this task. This suggests that the prompts need to be further refined to consider this aspect.

Table 3: Alignment between Gold and Synth Gold SDG Counts.

SDG	Gold	Synth Gold	Difference
0	83	12	71
1	3	2	1
2	4	5	1
3	3	25	22
4	6	10	4
5	4	7	3
6	4	5	1
7	3	6	3
8	5	24	19
9	5	12	7
10	4	8	4
11	4	4	0
12	6	4	2
13	2	8	6
14	5	3	2
15	5	7	2
16	3	1	2
17	7	2	5

5 Future Work

Future work will focus on improving the data augmentation strategies. The top priority will be creating synthetic datasets that are closer to the train and test distributions. Some other datasets should be based on classic distributions (e.g., multinomial, Poisson, log-normal, etc.). Pairing existing SDG datasets with non-relevant examples generated by LLMs could be another viable strategy to improve the training data. We also plan to test on multiple SDG datasets using the same approach.

Limitations

A main limitation of the presented approach was its failure to generate synthetic examples for the zero (non-relevant) class. Likely, even a few non-relevant examples in the synthetic dataset would have further improved the results. Another major shortcoming is that we have not considered various data distributions for this particular set of runs. This limitation will be addressed in future work, as already mentioned. Processing speed was another significant problem since a typical 5-fold stratified cross-validation with 5 epochs per fold took over 10 minutes to run, which could be considered a bit too much given the size of the training and test datasets (430 and 156 examples).

References

Amir Amel-Zadeh, Mike Chen, George Mussalli, and Michael Weinberg. 2021. Nlp for sdgs: Measuring corporate alignment with the sustainable development goals. *Columbia Business School Research Paper*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 968–988. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Jade Eva Guisiano, Raja Chiky, and Jonathas De Mello. 2022. Sdg-meter: A deep learning based tool for automatic text classification of the sustainable development goals. In *Intelligent Information and Database Systems - 14th Asian Conference, ACIIDS 2022, Ho Chi Minh City, Vietnam, November 28-30, 2022, Proceedings, Part I*, volume 13757 of *Lecture Notes in Computer Science*, pages 259–271. Springer.

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. *CoRR*, abs/1907.11692.

Roberto Carlos Morales-Hernández, David Becerra-Alonso, Eduardo Romero Vivas, and Joaquín Gutiérrez. 2022. Comparison between SVM and distilbert for multi-label text classification of scientific papers aligned with sustainable development goals. In *Advances in Computational Intelligence - 21st Mexican International Conference on Artificial Intelligence, MICAI 2022, Monterrey, Mexico, October 24-29, 2022, Proceedings, Part II*, volume 13613 of *Lecture Notes in Computer Science*, pages 57–67. Springer.

Lukas Pukelis, Núria Bautista-Puig, Mykola Skrynik, and Vilius Stanciauskas. 2020. OSDG - open-source approach to classify text data by UN sustainable development goals (sdgs). *CoRR*, abs/2005.14569.

Lukas Pukelis, Núria Bautista-Puig, Guste Statuleviciute, Vilius Stanciauskas, Gokhan Dikmener, and Dina Akylbekova. 2022. OSDG 2.0: a multilingual tool for classifying text data by UN sustainable development goals (sdgs). *CoRR*, abs/2211.11252.

Mohamad Al Sayed, Adrian M. P. Brasoveanu, Lyndon J. B. Nixon, and Arno Scharl. 2023. Unsupervised topic modeling with bertopic for coarse and fine-grained news classification. In *Advances in Computational Intelligence - 17th International Work-Conference on Artificial Neural Networks, IWANN 2023, Ponta Delgada, Portugal, June 19-21, 2023, Proceedings, Part I*, volume 14134 of *Lecture Notes in Computer Science*, pages 162–174. Springer.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *CoRR*, abs/2209.11055.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.