Scouting out the Border: Leveraging Explainable AI to Generate Synthetic Training Data for SDG Classification

Norman Süsstrunk and **Albert Weichselbraun** and **Andreas Murk**
Roger Waldvogel and **André Glatzl**
University of Applied Sciences of the Grisons
7000 Chur, Switzerland
{firstname.lastname}@fhgr.ch

Abstract

This paper discusses the use of synthetic training data towards training and optimizing a DistilBERT-based classifier for the SwissText 2024 Shared Task which focused on the classification of the United Nation's Sustainable Development Goals (SDGs) in scientific abstracts. The proposed approach uses Large Language Models (LLMs) to generate synthetic training data based on the test data provided by the shared task organizers. We then train a classifier on the synthetic dataset, evaluate the system on gold standard data, and use explainable AI to extract problematic features that caused incorrect classifications. Generating synthetic data that demonstrates the use of the problematic features within the correct class, aids the system in learning based on its past mistakes. An evaluates demonstrates that the suggested approach significantly improves classification performance, yielding the best result for Shared Task 1 according to the accuracy performance metric.

1 Introduction

The United Nation's Sustainable Development Goals (SDG) cover 17 interlinked global objectives that aim at achieving a better and more sustainable future. The SDGs address a wide range of issues, including poverty, inequality, climate change, environmental degradation, peace, and justice, emphasizing that development must balance social, economic, and environmental sustainability. The SwissText 2024 Shared Task 1 requested researchers to design systems that assign scientific abstracts to the most appropriate SDG, or to a non-relevant category, if no SDG applies. The shared task organizers provided a dataset of over 400 labeled abstracts which has been highly unbalanced in regard to the class distribution (Figure 1).

The challenge within this shared task has been developing a classifier based on a highly unbalanced dataset of 18 classes (17 SDGs + the non-relevant category) which can lead to significant model biases towards the majority classes and poor performance on the minority classes.

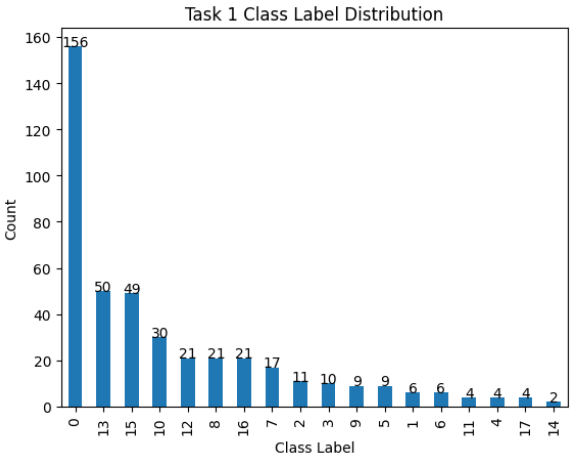

Figure 1: Label distribution within the training dataset for task 1

The rest of this paper is organized as follows: Section 2 outlines the method introduced in this work. Afterward, Section 3 presents and discusses evaluation results. The paper concludes with Section 4 which is followed by a short discussion of limitations.

2 Method

Figure 2 outlines the process used for training and optimizing the SDG classifier. At first, we draw upon GPT-3.5 (chatgpt.com) and Llama 3 (llama.meta.com/llama3/) to generate synthetic training data for all minority classes with the aim to better balance the dataset (Section 2.1).

We then train a transformer-based sequence classifier on both the training and synthetic dataset, and use it to classify the publicly available test dataset (Section 2.2). Finally, we apply explainable AI techniques to identify terminology within the incorrectly classified documents that has contributed significantly to misclassifications. Using an LLM

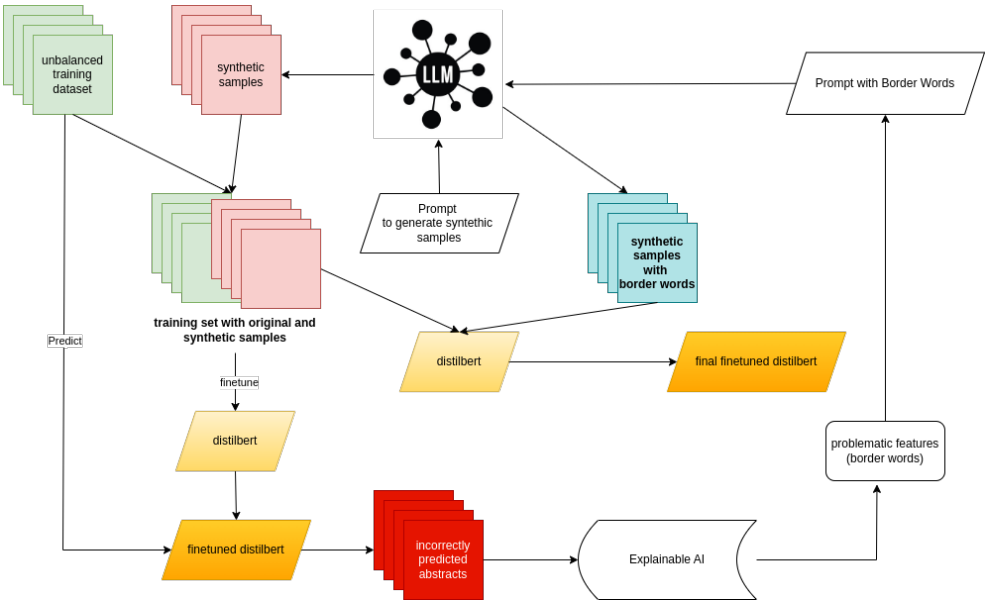

Figure 2: Process used for training the SDG classifier

allows to generate additional synthetic examples that contain this problematic terminology together with the correct class label. This additional synthetic training data aim at enabling the classifier to better learn the distinction between the affected classes, since it provides samples that have been inspired by prior mistakes and are aligned along the class boundaries.

2.1 Creating synthetic training data with GPT-3.5 and Llama 3

The first step utilizes GPT-3.5 and Llama 3 to generate synthetic training data for the minority classes, thus mitigating class imbalances and improving the overall performance of the text classification model. The following prompt was used to create the synthetic data:

```
You are a helpful assistant designed to
    generate synthetic data.

Create a JSONL file with 10 rows of data
    .
The data comes from the United Nations'
    Sustainable Development Goals.

This is an example row from my current
    data

{"ID":"oai:www.zora.uzh.ch:126666",
 "TITLE":"Identifying phrasemes...",
 "ABSTRACT":"In corpus linguistics...",
  ... , "SDG":0}

These are the SDGs that are available
    for the data:

0:"Non-Relevant",
1:"No Poverty",
2:"Zero Hunger",
3:"Good Health and Well-being",
4:"Quality Education",
5:"Gender Equality",
6:"Clean Water and Sanitation",
7:"Affordable and Clean Energy",
8:"Decent Work and Economic Growth",
9:"Industry, Innovation, and
    Infrastructure",
10:"Reduced Inequality",
11:"Sustainable Cities and Communities",
12:"Responsible Consumption and
    Production",
13:"Climate Action",
14:"Life Below Water",
15:"Life on Land",
16:"Peace, Justice, and Strong
    Institutions",
17:"Partnerships for the Goals"

Make sure that the text makes sense (i.e
    ., the title and abstract are
    coherent) and that the SDG is one of
     the 18 options listed above.
Also only respond with the resulting
    JSONL file.
```

Figure 3 summarizes the label distribution with the added synthetic samples, therefore, outlining the impact of the additional data on class imbalances.

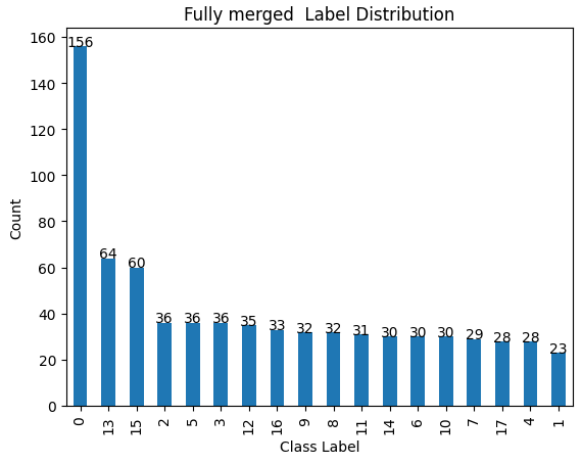

Figure 3: Label distribution after extending the gold standard with synthetic data generated by ChatGPT 3.5 and Llama 3

2.2 Transformer classifier

The proposed classification approach draws upon the Hugging Face library's *AutoModelForSequenceClassification* class in conjunction with the *distilbert-base-multilingual-cased* model (Wolf et al., 2019). This approach leverages the pre-trained DistilBERT transformer model (Sanh et al., 2020), i.e., a distilled version of BERT, which is well-suited for rapid prototyping, since it provides quick training times in conjunction with a good performance for various natural language processing tasks.

The model was fine-tuned on the training and synthetic data using the AdamW optimizer and a learning rate scheduler. Cross-entropy loss function was utilized, as it is standard for multi-label classification tasks.

In addition, we draw upon Optuna (Akiba et al., 2019), a hyperparameter optimization framework, to identify the best hyperparameters for our model. The search space was defined using the `optuna_hp_space` function, specifying ranges for key hyperparameters such as learning rate, batch size, and the number of epochs. Optuna's efficient search algorithms, such as Tree-structured Parzen Estimator (TPE), were utilized to explore this space and determine the optimal set of hyperparameters (Table 1).

Table 1: Hyperparameter configuration

Hyperparameter	Value
learning_rate	4.53e-05
per_device_train_batch_size	16
per_device_eval_batch_size	3
num_train_epochs	30
weight_decay	0.01

2.3 Scouting out the border

We developed an approach for identifying incorrectly classified abstracts, and extracting terminology that significantly contributed to the misclassification (i.e., problematic terms or border words), to create additional synthetic training data which are tailored towards addressing the classifier's weaknesses. The hypothesis is that incorporating these synthetic samples into the training set will further enhance classification accuracy, by providing samples that are well-suited towards learning class boundaries.

2.3.1 Boundary scouting process

The boundary scouting process involves the following steps:

1. *Identify incorrectly classified abstracts*: Applying the developed DistilBERT classifier (Section 2.2) to the test data yields a set of incorrectly classified abstracts.

2. *Extraction of problematic terms*: The *SequenceClassificationExplainer* tool which is part of the Transformer Interpret package (pypi.org/project/transformers-interpret/) is used for analyzing the incorrectly classified abstract. The package draws upon research by Janizek et al. (2020) and Sundararajan et al. (2017) which leverages attribution methods to assign importance scores to individual tokens in the input sequence. These scores indicate the contribution of each token to the model's prediction, and help in understanding the model's decision-making process by highlighting text that contributed most to the predicted class (i.e., the terminology responsible for misclassifications).

3. *Synthetic Sample Generation*: We use the identified problematic terminology in conjunction with Llama3-8B-8192 to generate synthetic abstracts that demonstrate the use of the

problematic terminology (i.e., border words) in the correct class. For example, if an abstract has been misclassified as "Climate Action" (SDG 13) rather than "Affordable and Clean Energy" (SDG 7) due to the use of the phrase "solar energy", we would ask the model to generate synthetic examples that use the phrase "solar energy" in the context of SDG 7. Automatizing this process yields additional training data that specifically address the classifier's current weaknesses.

4. *Model Retraining*: Retrain the classifier with the original and synthetic samples.

5. *Evaluation*: Evaluate the performance of the retrained classifier to assess improvements.

2.3.2 Example

The following example demonstrates the use of the proposed approach based on an abstract that has been misclassified by the initial classifier model (Section 3):

- *Gold standard label*: 4 ("Quality Education")
- *Predicted label*: 8 ("Decent Work and Economic Growth")

Figure 4 in the paper's appendix shows the classified example text together with the interpretation obtained from the Transformer Interpret package with tokens that contributed significantly to the incorrect classification (`labour, ter, market, differenti, academic, the`) marked in green.

The model likely considered the text's focus on the labor market and economic implications of education as more relevant to SDG 8. Words related to economic growth and employment outcomes provided strong signals that outweighed the educational content, despite the text's clear relevance to the quality of education.

We, therefore, use the following prompt to instruct the LLM to generate a synthetic abstract that belongs to the correct class:

```
Invent a title and an abstract of a
    research paper about Sustainable
    Development Goals (SDG) that has
- the subject = Quality Education
- the abstract should contain and focus
    the content around following words
    extensively: ['labour', 'ter', '
    market', 'differenti', 'academic', '
    the']
Do not include the words Sustainable
    Development Goals (SDG) in the
    abstract or the title.
```

The model then returned the following output (shortened):

```
Title:  Bridging Academic Pathways and
    Labour Market Needs: Analyzing the
    Impact of Quality Education on
    Economic Differentiation
Abstract: This paper examines the
    intricate relationship between
    quality education and its impact on
    labour market differentiation. By
    analyzing academic programs and
    their alignment with the evolving
    needs of the labour market, this
    study highlights the critical role
    of education in fostering economic
    growth and social stability...
```

The generated synthetic abstracts have been added to the training set and used to retrain the transformer classifier. Section 3 outlines the performance gains obtained through this process.

3 Evaluation

We submitted two evaluation runs. One in which the DistilBERT classifier has been trained on the test and synthetic dataset (Syn; submission name: *NLPChur_TASK1__merged_synthetic_data_task1_report_goldlabel.txt*), and a second one which used in addition the synthetic abstracts generated based on the problematic words (Syn+; submission name: *NLPChur_TASK1__merged_synthetic_data_bad_words_task1_report_goldlabel.txt*).

Table 2 presents the overall classification performance of both approaches. As outlined in the table, the classifier trained on the improved synthetic dataset that has been extended based on the method introduced in Section 2.3 (Syn+) outperforms the classifier trained on the initial synthetic dataset (Syn) in every single evaluation metric.

Table 3 presents the per class classification performance for the classifier trained on the Syn+ dataset. The presented results indicate that although the overall performance improved significantly with the boundary scouting process, there are still classes where the classifier clearly failed to produce viable results. Investigating and mitigating

Table 2: Overall classification performance on Task 1 (correct prediction of the primary SDG) for the classifier trained on (i) the training and synthetic dataset (Syn), and (ii) the training, synthetic dataset and the synthetic data created based on the border words (Syn+).

Metric	Syn	Syn+
Accuracy	0.46	0.52
Macro Precision	0.49	0.53
Macro Recall	0.51	0.60
Macro F1 Score	0.44	0.51
Weighted Precision	0.59	0.65
Weighted Recall	0.46	0.52
Weighted F1 Score	0.49	0.55

these shortcomings will be an interesting direction for future work.

Table 3: Per label classification performance on Task 1 (correct prediction of the primary SDG) of the classifier trained on the Syn+ dataset.

SDG	f1	precision	recall
0	0.59	0.48	0.77
1	0.67	0.67	0.67
2	0.89	1.00	0.80
3	0.18	0.33	0.12
4	0.25	0.17	0.50
5	0.73	1.00	0.57
6	0.86	0.75	1.00
7	0.67	1.00	0.50
8	0.06	0.20	0.04
9	0.50	0.60	0.43
10	0.40	0.75	0.27
11	0.67	0.50	1.00
12	0.60	0.50	0.75
13	0.36	1.00	0.22
14	0.89	0.80	1.00
15	0.91	1.00	0.83
16	0.00	0.00	0.00
17	0.00	0.00	0.00

4 Outlook and Conclusions

This paper introduced an approach for creating and optimizing a transformer-based Sustainable Development Goals (SDG) classifier that was used in SwissText Shared Task 1 that focuses on identifying the majority SDG class for scientific abstracts. We leverage LLMs and explainable AI to generate synthetic training data that aims at (i) mitigating class imbalances, and (ii) aiding the classifier in learning class boundaries. The system obtained the top accuracy for the Shared Task 1, demonstrating the method's potential.

Future work will focus on further improving the system's performance by adding a binary classifier to distinguish between abstracts that contain references to SDGs and those that do not. Additionally, efforts will be directed towards enhancing the methodology for generating synthetic training data. This includes improving the handling of subtokens to ensure more accurate and representative synthetic samples. In addition, we plan to investigate cases where the respective classes did not benefit from the improved synthetic dataset and research strategies to address this issue.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Joseph D. Janizek, Pascal Sturmfels, and Su-In Lee. 2020. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. ArXiv:2002.04138 [cs, stat].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*. ArXiv: 1910.01108.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328. PMLR. ISSN: 2640-3498.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

A Transformer Interpret

Figure 4 illustrates the interpretations obtained from the Transformer Interpret package, which highlights tokens that contributed significantly to the classification label marked in green, and tokens that provided contrary information in red.

Word Importance

[CLS] This paper analyses whether ter ##tiary education of different types , i . e . , academic or vocation ##al ter ##tiary education , leads to more or less favour ##able labour market outcomes . We study the problem for Switzerland , where more than two third ##s of the work ##force gain vocation ##al secondary degrees and a substantial number go on to a vocation ##al ter ##tiary degree but only a small share gain an academic ter ##tiary degree . As outcome variables , we ex ##amine the risk of being une ##mp ##lo ##yed , monthly earning ##s , and variation in earning ##s (reflect ##ing financial risk) . We study these outcomes at career entry and later stages . Our em ##pir ##ical results reveal that the type of ter ##tiary education has various effects on these outcomes . At career entry , we observe equal une ##mployment risk but higher average wa ##ges and lower financial risk for vocation ##al graduates . At later career stages , we find that these higher average wa ##ges disa ##ppe ##ar and risk of une ##mployment becomes lower for vocation ##al graduates . Thus , by differenti ##ating the ter ##tiary system into vocation ##al and academic institutions graduates face a variety of valuable options allowing them to self - select into an educational type that best matches their individual pre ##ference ##s . [SEP]

Figure 4: Example explanation provided the Transformer Interpret package