Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish

Fred Philippy[1,2], Shohreh Haddadan[1], Siwen Guo[1]

[1]Zortify S.A., Luxembourg [2]University of Luxembourg, Luxembourg

{fred, siwen}@zortify.com, shohreh.haddadan@gmail.com

Abstract

In NLP, zero-shot classification (ZSC) is the task of assigning labels to textual data without any labeled examples for the target classes. A common method for ZSC is to fine-tune a language model on a Natural Language Inference (NLI) dataset and then use it to infer the entailment between the input document and the target labels. However, this approach faces certain challenges, particularly for languages with limited resources. In this paper, we propose an alternative solution that leverages dictionaries as a source of data for ZSC. We focus on Luxembourgish, a low-resource language spoken in Luxembourg, and construct two new topic relevance classification datasets based on a dictionary that provides various synonyms, word translations and example sentences. We evaluate the usability of our dataset and compare it with the NLI-based approach on two topic classification tasks in a zero-shot manner. Our results show that by using the dictionary-based dataset, the trained models outperform the ones following the NLI-based approach for ZSC. While we focus on a single low-resource language in this study, we believe that the efficacy of our approach can also transfer to other languages where such a dictionary is available.

Keywords: Less-Resourced/Endangered Languages, Document Classification, Corpus

1. Introduction

Zero-shot classification (ZSC) allows to classify a text document into a category for which no labeled examples are available. A common technique for ZSC is to leverage pre-trained language models that have learned general semantic representations from large corpora. These models can be fine-tuned on a natural language inference (NLI) dataset and then be used to infer the entailment between the document and the labels (Yin et al., 2019). In this approach, each potential target label is considered as a hypothesis in natural language, and the NLI model is used to evaluate the level of entailment between the input document and potential labels. For example, given a document "I always eat my soup with a spoon" and the labels "food" and "animals", the model can predict a score of how likely the document entails each label. The label with the highest entailment score can be selected as the predicted class.

Directly adopting NLI datasets for ZSC poses several challenges and limitations in real-world scenarios. We identify and highlight three main limitations of such an approach. First, there is a mismatch between the NLI and ZSC tasks. Second, the performance of this approach depends on the availability and quality of NLI datasets, which are challenging and costly to obtain. Third, for many low-resource languages, the lack of pre-training data hinders the model's ability to solve complex reasoning tasks such as NLI. In this work, we discuss the case of Luxembourgish, a West Germanic language spoken by around 400,000 people in Luxembourg. There is no large NLI dataset for the language, and only a small amount of unlabeled pre-training data is available. Therefore, using NLI datasets for ZSC in Luxembourgish results in poor performance.

In this work, we propose an alternative solution that provides sufficient data for low-resource languages in the context of ZSC. The proposed approach exploits dictionaries as a source of data for ZSC. More specifically, this dictionary-based approach offers two main advantages: 1) it provides data that is more relevant to the task of ZSC, and 2) it leverages resources that are more readily available in many low-resource languages. We demonstrate our approach on the Luxembourgish language, for which we construct two new topic relevance classification datasets based on a dictionary.[1] In short, our main contributions are as follows:

1. We introduce a new approach for creating datasets that allow to adapt models to ZSC for low-resource languages where a dictionary is available.
2. Using this approach, we construct and release two new datasets for Luxembourgish that are more suitable for ZSC tasks than existing NLI datasets.
3. We evaluate our datasets on the task of zero-shot topic classification by comparing the performance of models trained on our datasets and NLI datasets

[1] Our code and datasets are accessible via https://github.com/fredxlpy/LETZ/

2. Motivation

Our work aims to address the following limitations and challenges that hinder the effectiveness of zero-shot classification for low-resource languages such as Luxembourgish:

1. The mismatch between the fine-tuning task, NLI, and the inference task, topic classification, as the former requires reasoning about logical relations between sentences (entailment, contradiction, neutral), while the latter evaluates the relevance of labels to a sentence (relevant, irrelevant) (Ma et al., 2021).
2. The difficulty and the expense of creating NLI data, especially for low-resource languages. NLI data requires high-quality annotations that capture the subtle nuances of entailment and contradiction between sentence pairs. Moreover, such annotations are often prone to inter-annotator disagreement, which undermines the validity and reliability of NLI datasets (Pavlick and Kwiatkowski, 2019; Kalouli et al., 2023).
3. The poor performance of language models on high-level tasks such as NLI for low-resource languages (Ebrahimi et al., 2022). Low-resource language models suffer from insufficient training data and vocabulary coverage, which affects their ability to encode rich semantic representations and handle complex reasoning tasks such as NLI.

3. Related Work

A common method for ZSC is the *entailment approach* (Yin et al., 2019), which uses NLI datasets to fine-tune pre-trained language models and then apply them to ZSC tasks. However, this approach has several drawbacks, as discussed by Ma et al. (2021). They identify issues such as label mismatch, data imbalance, and semantic ambiguity that affect the performance and generalization of the entailment approach. Moreover, Ebrahimi et al. (2022) show that NLI models perform cross-lingual transfer poorly for low-resource languages, which in turn affects their ZSC capability. Therefore, they argue for the need of creating annotated datasets for semantic tasks in low-resource languages.

Luxembourgish Language
Luxembourgish is one of the three national languages of Luxembourg and is spoken by roughly 400,000 people (≈ 70% of the population). According to UNESCO *World Atlas of Languages*[2], Luxembourgish belongs to the world's *potentially vulnerable* languages.

However, Luxembourgish has seen significant transformations over the past century, including its development into a national language, expansion into written and digital media, and its role as a symbol of national identity.

The sociolinguistic landscape of Luxembourg, with its unique multilingual setup (Purschke and Gilles, 2023) and the dynamic evolution of Luxembourgish from a dialect to a national language with increasing digital presence, provides a fertile ground for NLP research. Researching Luxembourgish through the lens of NLP contributes to the field of lesser-studied languages by developing methodologies that can be applied to other multilingual and language variation contexts.

4. Our Dataset

Based on a publicly available online dictionary, we create two new topic relevance classification datasets that allow to adapt pre-trained language models to zero-shot topic classification in Luxembourgish.

4.1. Data Collection

Luxembourg Online Dictionary[3] (LOD) is a publicly available platform hosting a multilingual dictionary with the aim of promoting Luxembourgish as the language of communication, integration and literature. In the following, we present some statistics relevant to our work about the data provided by the Center for the Luxembourgish Language (ZLS[4]) in a report[5] in 2022.

The dictionary contains around **10,000 synonyms** and **48,000 example sentences** on approximately **31,000 entries**. Words with multiple meanings are treated separately for each of their distinct meanings, with corresponding synonyms and example sentences. For most entries, the dictionary provides translations from/to 5 languages: German, French, English, Portuguese and Sign Language. In addition, it features 20,000 phonetic transcriptions, 30,000 audio recordings, 9,300 conjugation and declension tables as well as 5,000 proverbs and idiom explanations.

ZLS released all of their data on the Luxembourgish Open Data platform[6] under a *Creative Commons Zero* (CC0) license. In this work, we use the dataset version released on June 5, 2023.

[2] https://en.wal.unesco.org

[3] https://lod.lu

[4] *Zenter fir d'Lëtzebuerger Sproochh*

[5] https://gouvernement.lu/fr/actualites/toutes_actualites/communiques/2022/06-juin/21-lod-neie-look.html

[6] https://data.public.lu/en/organizations/zenter-fir-dletzebuerger-sprooch/

4.2. From Dictionary to Dataset

We first extract the part-of-speech tag, synonyms, and example sentences for each meaning of every word in the raw LOD data, and filter out the non-nouns.

Next, we assign all the synonyms of a word meaning as labels to its example sentences. To prevent the model from exploiting the shortcut of matching the label with the word occurrence in the sentence, we exclude the word itself from the label set .

Moreover, since many Luxembourgish words are orthographic variants of French or German words[7], we discard noun-synonym pairs that have a low Levenshtein distance.

Finally, we generate “non-entailment” samples by randomly selecting a word from the entire noun vocabulary as a label for each example sentence. However, we exclude any words that are similar to any of the words in the sentence based on the Levenshtein distance.

Following the exact same approach, we additionally create a separate dataset based on the word translations available in the dictionary instead of synonyms.

This new type of dataset is termed ***L**uxembourgish **E**ntailment-based **T**opic classification via **Z**ero-shot learning* (LETZ), with the synonym-based dataset being referred to as **`LETZ-SYN`** and the one derived from word translations as **`LETZ-WoT`**.

The number of "entailment"/"relevant" ("1") and "non-entailment"/"irrelevant" ("0") samples is balanced for all sets. The dataset split sizes are provided in Table 1. We provide examples and more details of our data sets in Appendix A.

Dataset	|Train|	|Dev|	|Test|
`LETZ-SYN`	11,822	1,478	1,478
`LETZ-WoT`	39,132	4,892	4,892

Table 1: Dataset statistics

5. Implementation

5.1. Training

We conduct experiments using two different models that have been pre-trained on Luxembourgish data: **LuxemBERT** (Lothritz et al., 2022), a monolingual Luxembourgish model, and **mBERT** (Devlin et al., 2019), a multilingual BERT model that has been pre-trained on 102 languages, including Luxembourgish.

In order to perform the classification task, we append an additional layer to the pre-trained model that consists of a linear layer and a tanh activation function. The classification layer has two output nodes which are used to determine whether a given document contains a topic or not (Figure 2a). Considering the limited amount of fine-tuning data, which could lead to variability in performance outcomes, we conduct each experiment four times using distinct random seeds. We then report the average results to account for any inconsistencies.

Besides fine-tuning both models on our new datasets, we use additional training datasets for comparison:

- **NLI-lb** (Lothritz et al., 2022), a Luxembourgish NLI dataset consisting of 568 train and 63 validation samples. The dataset only contains entailment ("1") and contradiction samples ("0").
- **XNLI-de**, **XNLI-en** & **XNLI-fr**, German, English and French subsets of the XNLI (Conneau et al., 2018) dataset respectively.

In addition, we perform experiments in "high-resource" (11,822 train and 1,478 validation samples)[8] and "low-resource" (568 train and 63 validation samples)[9] settings.

5.2. Evaluation

Due to the inherent limitations associated with Luxembourgish being a low-resource language, there is a conspicuous lack of labeled datasets available. Within the context of topic classification, we could only identify two evaluation datasets that were suitable for our study:

- The Luxembourgish subset of ***SIB-200*** (Adelani et al., 2024), a multilingual topic classification dataset, containing seven categories, namely: `science/technology`, `travel`, `politics`, `sports`, `health`, `entertainment`, and `geography`.
- A Luxembourgish News Classification dataset introduced by Lothritz et al. (2022), consisting of news articles from a Luxembourg-based news platform. For our experiments we restrict it to the following 5 (out of 8) categories: `Sports`, `Culture`, `Gaming`, `Technology`, `Cooking recipes`. We exclude `National news`, `International news` and `European news` to avoid overlap with other categories. In what follows we will refer to this dataset as ***LuxNews***.

[7] Examples: “alerte” → “Alert”, “Million” → “Millioun”.

[8] Number of samples in `LETZ-SYN`.

[9] Number of samples in the Luxembourgish NLI dataset (Lothritz et al., 2022).

		n = 568		n = 11.822	
Model	Train data	SIB-200	LuxNews	SIB-200	LuxNews
mBERT	NLI-lb	17.52 (16.56)	15.87 (12.51)	\	\
	NLI-de	25.61 (24.69)	30.22 (25.88)	48.04 (43.76)	43.06 (35.18)
	NLI-en	22.67 (22.38)	28.55 (23.20)	49.51 (44.34)	50.73 (38.18)
	NLI-fr	22.30 (21.30)	25.02 (20.01)	49.75 (45.77)	46.30 (37.65)
	LETZ-WoT	49.39 (49.50)	59.81 (43.18)	53.55 (52.46)	59.96 (52.13)
	LETZ-SYN	**52.08 (51.45)**	**65.08 (49.20)**	**53.80 (54.13)**	**66.07 (47.73)**
LuxemBERT	NLI-lb	14.58 (12.91)	24.69 (16.53)	\	\
	LETZ-SYN	**18.50 (15.86)**	**30.63 (19.48)**	**65.07 (64.07)**	**51.81 (38.27)**

Table 2: Results of our experiments on two topic classification datasets. Experiments are conducted for different number of training samples **n** from the different training sets. The performance metrics are reported as "**Accuracy (F1 score)**" for each task.

Following Yin et al. (2019), we use an entailment approach (Figure 2b in Appendix B) to evaluate the models on these datasets, instead of a traditional supervised classification approach, where the number of output nodes corresponds to the number of categories. To be more exact, for a given sample **x** and potential topics/categories $T = \{T_1, \ldots, T_n\}$, we compute the entailment probability for each pair $(\mathbf{x}, T_i)_{i\in\{1,\ldots,n\}}$ denoted by $\mathbf{P}_{i,1}$ and select T_{i^*} where

$$i^* = \underset{i\in\{1,\ldots,n\}}{\operatorname{argmax}} \mathbf{P}_{i,1}$$

The details of the training and evaluation methodology and the datasets employed are presented in Appendix B.

6. Results

Table 2 shows that models fine-tuned on our datasets exceed the performance of those trained on NLI data, especially in the "low-resource" setting. More exactly, mBERT, with only 568 samples from our dictionary-based datasets, exceeds the results achieved with 20x more NLI samples in French, German, or English.

However, fine-tuning on German, French, or English NLI datasets markedly improves results over Luxembourgish data for which the performance is comparable to that of the random baseline. This suggests that the limited size of the Luxembourgish pre-training corpus may hinder the model's ability to acquire a sufficient level of semantic and pragmatic understanding to solve complex reasoning tasks such as NLI.

In the "low-resource" setting, LuxemBERT underperforms mBERT, suggesting it needs more data for task-specific knowledge compared to mBERT's general cross-lingual knowledge acquired during pre-training from high-resource languages. Nonetheless, in the "high-resource" setting, LuxemBERT outperforms mBERT on *SIB-200* but underperforms on *LuxNews*, possibly due to its inability to interpret multilingual speech excerpts or quotes.

7. Discussion

While we focus on Luxembourgish as an example of low-resource languages in this paper, we believe that this approach can be generalized to other languages where such dictionaries are available as well.

While we acknowledge that our method depends on the availability of dictionaries for low-resource languages, it is crucial to note that dictionaries often receive priority due to their fundamental role in educational and cultural preservation efforts. They are typically more prevalent because they form the bedrock for literacy and basic education, which are more fundamental needs than specialized datasets like those required for NLI. The creation of NLI datasets demands advanced linguistic knowledge and resources, making it a less immediate concern compared to building basic language tools. Initiatives, such as the *Dictionaria*[10] journal, the *Living Dictionaries*[11] or the *Webonary*[12] platform, support the development of dictionaries for low-resource and even indigenous languages. So, while both dictionaries and NLI datasets may not be universally available, there is a stronger, more widespread

[10] https://dictionaria.clld.org

[11] https://livingdictionaries.app

[12] https://www.webonary.org

motivation behind the creation of dictionaries, rendering them relatively more accessible and likely to exist for low-resource languages.

Additionally, our experiments suggest that these dictionaries would not require tens of thousand of entries to be effective, as it appears that a multilingual language model can attain satisfactory performance with just a few hundred sentence-synonym or sentence-word translation pairs.

8. Conclusion

This paper presents a new but simple approach to construct datasets that enable a language model to perform zero-shot topic classification in a low-resource language, such as Luxembourgish. We argue that the conventional approach of transferring from NLI to ZSC is ineffective for such languages, due to the semantic complexity of NLI and the scarcity of linguistic resources. We propose an alternative approach that leverages a dictionary to create a dataset that is more aligned with the ZSC task. We demonstrate that our dataset enables the model to outperform the ones that employ cross-lingual NLI transfer or in-language NLI fine-tuning on Luxembourgish ZSC, using over 20 times fewer training samples. In future work, we intend to explore the effectiveness of our approach when applied to other low-resource languages, as well as to high-resource ones.

Limitations

One of the limitations of our study is that we only focus on a single low-resource language, Luxembourgish, and we do not test our approach on other languages. Therefore, the generalizability of our method may be limited by the availability and quality of dictionaries for different languages. Another limitation is that we rely on a single source of data, namely a dictionary, which may not capture all the nuances and variations of natural language.

Ethics Statement

Our study aims to provide a novel solution for zero-shot classification in low-resource languages, which can potentially benefit various applications and users who need to classify textual data without labeled examples. While our method could potentially benefit any language, we specifically emphasize its usefulness for low-resource languages that suffer from data scarcity and lack of adequate tools. We believe that our method can contribute to the promotion of linguistic diversity, as well as to the empowerment and inclusion of speakers of low-resource languages.

However, we also acknowledge that some dictionaries may contain outdated, inaccurate, or offensive information that could harm certain groups or individuals. Therefore, we urge future researchers and practitioners to carefully select and evaluate the dictionaries they use and to adhere to the ethical principles and guidelines of their respective fields and communities.

9. Bibliographical References

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.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Hai Hu, Alexander F. Webb, Lawrence S. Moss, and Valeria De Paiva. 2023. Curing the SICK and Other NLI Maladies. *Computational Linguistics*, 49(1):199–243.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with Entailment-based Zero-shot Text Classification. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 786–796, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Christoph Purschke and Peter Gilles. 2023. Sociolinguistics in Luxembourg. In *The Routledge Handbook of Sociolinguistics Around the World*, 2 edition. Routledge.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

10. Language Resource References

Adelani, David and Liu, Hannah and Shen, Xiaoyu and Vassilyev, Nikita and Alabi, Jesujoba and Mao, Yanke and Gao, Haonan and Lee, En-Shiun. 2024. *SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects*. Association for Computational Linguistics.

Conneau, Alexis and Rinott, Ruty and Lample, Guillaume and Williams, Adina and Bowman, Samuel and Schwenk, Holger and Stoyanov, Veselin. 2018. *XNLI: Evaluating Cross-lingual Sentence Representations*. Association for Computational Linguistics.

Lothritz, Cedric and Lebichot, Bertrand and Allix, Kevin and Veiber, Lisa and Bissyande, Tegawende and Klein, Jacques and Boytsov, Andrey and Lefebvre, Clément and Goujon, Anne. 2022. *LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish*. European Language Resources Association.

A. Our Dataset

Figure 1 shows the distribution of the sample length of LETZ-SYN, expressed as word count, and Table 3 shows a small example subset of LETZ-SYN. Both datasets, LETZ-SYN and LETZ-WoT, are publicly available under a *Creative Commons Attribution 4.0 International* (CC BY 4.0) license.

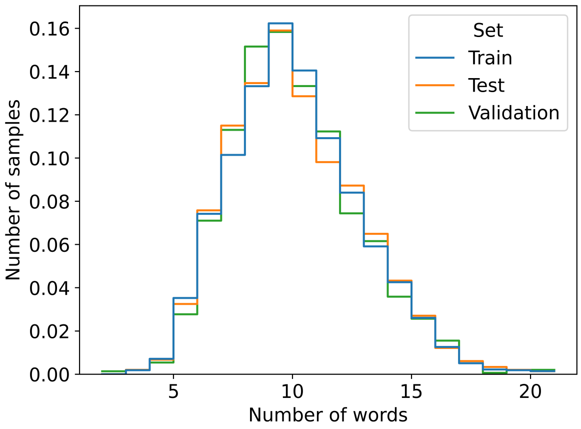
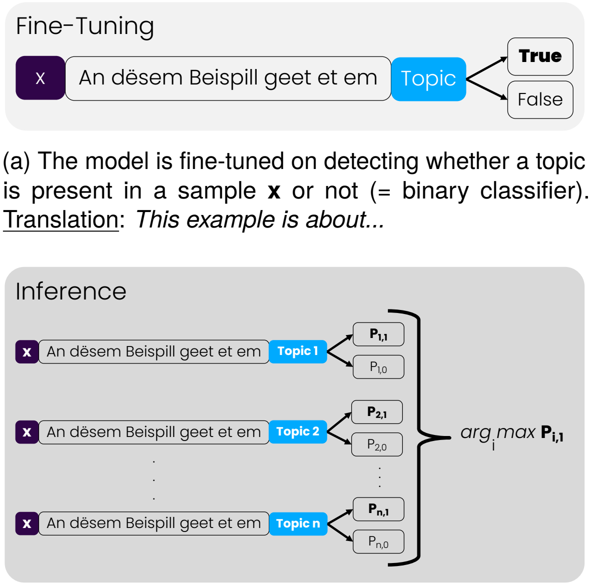


Figure 1: Distribution of text sample length, expressed in terms of word count, for the training, validation and test sets of LETZ-SYN

B. Implementation Details

B.1. Methodology

We provide a visual illustration of the *entailment approach* (Yin et al., 2019) that we use in our experiments in Figure 2. The natural language label description words and number of samples per class during evaluation are provided in Table 4.



(a) The model is fine-tuned on detecting whether a topic is present in a sample **x** or not (= binary classifier). Translation: *This example is about...*

(b) The model estimates the likelihood of each candidate topic independently at the inference stage and then the topic with the maximum probability is chosen.

Figure 2: Illustration of the *entailment approach* (Yin et al., 2019) for ZSC

B.2. Models

We conduct our experiments on the base multilingual BERT (cased) (Devlin et al., 2019) and LuxemBERT (Lothritz et al., 2022) models. Both models are based on the same architecture and have 12 attention heads and 12 transformer blocks with a hidden size of 768. mBERT and LuxemBERT have a vocabulary size of 30,000 and 119,547 respectively. Both models have 110 million parameters.

B.3. Reproducibility

To reduce the computational expenses, we refrain from conducting hyper-parameter tuning and employ the configurations that yielded satisfactory results in our initial experiments. We conduct all the experiments using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 2e-5 with 10% warm-sup steps and linear decay and a batch size of 32. We fine-tune, with 10 warm-up steps, over 5 epochs. We perform validation after each epoch and select the optimal checkpoint based on the lowest validation loss. The maximum sequence length, during training, is set to 128 tokens. During evaluation, we set the maximum length to 128 tokens for SIB-200, and to 512 for the LuxNews dataset. For each evaluation dataset, we output the accuracy and macro-averaged F1 score.

B.4. Computational Resources

All experiments were run within a few hours on 4 A100 40GB GPUs in parallel, using 4 different random seeds (one per GPU).

Text	Label	Class
Gedëlleg dech a waart op de richtegen **Abléck**! (*Be patient and wait for the right* ***point in time****!*)	Moment (*moment*)	1
Däin Auto huet hannen um Parechoc eng Téitsch. (*Your car has a dent on the rear bumper.*)	Libell (*dragon-fly*)	0
Bei esou vill Kandidate muss eng **Auswiel** gemaach ginn. (*With so many candidates, a* ***choice*** *must be made.*)	Selektioun (*selection*)	1
Ech schécken der d'Adress vun engem lëschtege Site. (*I am sending you the link to a funny website.*)	Schrauwenzéier (*screwdriver*)	0

Table 3: Examples from our dataset (*with English translations*).

Dataset	**Class**	**Class Label**	**n**
LuxNews	**Sports**	`Sport`	567
	Culture	`Konscht`	266
	Technology	`Technologie`	199
	Gaming	`Videospiller`	82
	Cooking recipes	`Rezept`	20
	National news	/	
	International news	/	
	European news	/	
SIB-200	**Science/Technology**	`Technologie`	51
	Travel	`Rees`	40
	Politics	`Politik`	30
	Sports	`Sport`	25
	Health	`Gesondheet`	22
	Entertainment	`Entertainment`	19
	Geography	`Geografie`	17

Table 4: The original classes and their corresponding translated Luxembourgish class labels that were used our experimental setup. We used the classes marked in **bold** for evaluation, and discarded the rest from the evaluation set. **n** is the number of samples used for evaluation.