White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/212308/

Version: Published Version

Article:
Razuvayevskaya, O., Wu, B., Leite, J.A. et al. (2024) Comparison between parameter-efficient techniques and full fine-tuning: a case study on multilingual news article classification. PLoS ONE, 19 (5). e0301738. ISSN: 1932-6203

https://doi.org/10.1371/journal.pone.0301738

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

PLOS ONE

RESEARCH ARTICLE

Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification

Olesya Razuvayevskaya[1☯*], **Ben Wu**[1☯], **João A. Leite**[1☯], **Freddy Heppell**[1☯], **Ivan Srba**[2], **Carolina Scarton**[1], **Kalina Bontcheva**[1], **Xingyi Song**[1]

1 Department of Computer Science, The University of Sheffield, Sheffield, United Kingdom, **2** Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia

☯ These authors contributed equally to this work.
* o.razuvayevskaya@sheffield.ac.uk

Check for updates

OPEN ACCESS

Citation: Razuvayevskaya O, Wu B, Leite JA, Heppell F, Srba I, Scarton C, et al. (2024) Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. PLoS ONE 19(5): e0301738. https://doi.org/10.1371/journal.pone.0301738

Editor: Mario Graff-Guerrero, Infotec: Centro de Investigacion e Innovacion en Tecnologias de la Informacion y Comunicacion, MEXICO

Received: August 7, 2023

Accepted: March 20, 2024

Published: May 3, 2024

Copyright: © 2024 Razuvayevskaya et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying the results presented in the study are available from the organisers of SemEval 2023 Shared Task 3: https://propaganda.math.unipd.it/semeval2023task3/ The access to the data is currently only granted through a token-based authorisation to the individuals registered for the shared task. The official data access page: https://propaganda.math.unipd.it/semeval2023task3/. The use of the data, for those granted access, is

Abstract

Adapters and Low-Rank Adaptation (LoRA) are parameter-efficient fine-tuning techniques designed to make the training of language models more efficient. Previous results demonstrated that these methods can even improve performance on some classification tasks. This paper complements existing research by investigating how these techniques influence classification performance and computation costs compared to full fine-tuning. We focus specifically on multilingual text classification tasks (genre, framing, and persuasion techniques detection; with different input lengths, number of predicted classes and classification difficulty), some of which have limited training data. In addition, we conduct in-depth analyses of their efficacy across different training scenarios (training on the original multilingual data; on the translations into English; and on a subset of English-only data) and different languages. Our findings provide valuable insights into the applicability of parameter-efficient fine-tuning techniques, particularly for multilabel classification and non-parallel multilingual tasks which are aimed at analysing input texts of varying length.

Introduction

The development of language models has led to a significant increase in the number of trainable parameters needed to fine-tune such models, with state-of-the-art models comprising of millions or even billions of parameters [1, 2]. This poses a serious constraint on the process of fine-tuning such models, often relying on significant computational resources. Many of the recent research efforts are therefore focused on the development of more efficient training techniques [3–5]. Methods that can decrease the computational costs make language models more accessible to researchers and practitioners with limited computational resources, and reduces the carbon footprint of their training.

In this study, we investigate the effectiveness of parameter-efficient fine-tuning techniques (PEFTs) in multilingual, monolingual and cross-lingual text classification scenarios. We have included PEFT techniques that have been evaluated on large volumes of data (>20B) in prior

restricted to research purposes only within the context of the shared task. Our usage of the data complied with these conditions. It is expected that the data will be made publicly available by the task organisers after a certain period of time.

Funding: This research is partially supported by the UK's innovation agency (InnovateUK) grants 10039055 (vera.ai, https://cordis.europa.eu/project/id/101070093) and 10039039 (Vigilant, https://cordis.europa.eu/project/id/101073921), a European Union grant INEA/CEF/ICT/A2020/2381686 (European Digital Media Observatory (EDMO) Ireland, https://edmohub.ie/), and Transnational Access funds from "SoBigData++: European Integrated Infrastructure for Social Mining and BigData Analytics" (EU H2020, Grant Agreement n.871042, http://www.sobigdata.eu). Freddy Heppell is supported by a University of Sheffield Faculty of Engineering PGR Prize Scholarship. João A. Leite and Ben Wu are supported by a University of Sheffield Engineering & Physical Sciences Research Council (EPSRC) Doctoral Training Partnership Scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

research [6]—namely, LoRA, adapter, BitFit, prefix tuning and prompt tuning. Since the majority of these techniques were previously evaluated on transformer-based models (BERT and RoBERTa) for text classification tasks, we select the subset of those that can be applied to such models—namely LoRA, adapter and BitFit. We exclude prefix tuning [7] and prompt tuning [8] as these techniques are applied to generative Large Language Models (LLMs) models for text-to-text generation tasks and cannot be directly compared to the above-mentioned three methods.

In more detail, adapter-based fine-tuning represents a family of efficiency techniques that work by freezing a pre-trained language model and adding a small number of trainable parameters in the layers of the language model [9–11]. This significantly reduces the training time at the cost of a small or no performance penalty. Another method of reducing the number of trainable parameters is based on performing Low-Rank Adaptation (LoRA) [12]. The main idea behind the LoRA approach is to freeze the weights of pre-trained language models and insert low-rank decomposition matrices into the transformer layers. BIas-Term FIne-Tuning (BitFit) is a PEFT technique that uses only the bias term and the task-specific linear classification layer during training, omitting most of the parameters in the encoder-decode layers [3]. This results in a significant reduction in computational costs as bias terms represent only a small fraction (up to 0.1%) of all the parameters of the models.

Prior studies report that, in addition to reducing computational costs, adapter-based methods can outperform full fine-tuning (FFT) in zero-shot cross-lingual settings [11, 13, 14]. However, the tasks addressed in these studies focus on parallel multilingual texts [13], meaning that inference is not performed in a realistic zero-shot manner [14], on short inputs [11] or in highly specific domains [14]. Given that adapter methods are additionally known to have limited capabilities in processing long text inputs due to reserving a part of the sequence length for adaptation [12], there is a motivation for evaluating this method in non-parallel cross-lingual and multilingual settings and on different input lengths. LoRA, on the other hand, does not reduce the input sequence length, and prior comparisons between FFT and LoRA suggest that, in addition to being parameter-efficient, LoRA can, for certain models, outperform FFT [12]. However, there is currently a research gap in terms of evaluating LoRA in similar multilingual training scenarios, which creates a motivation for including this PEFT in our study. There is no prior evidence regarding the cross-lingual capabilities of BitFit, however, this PEFT technique was shown to have comparable performance to adapters and to also be comparable to FFT on the GLUE benchmark. This motivated the inclusion of the BitFit method in this research.

This study is therefore motivated by the lack of consistent evaluation of PEFT techniques in multilingual and cross-lingual zero-shot classification tasks performed on long non-parallel texts. This gap in current research is addressed through a systematic comparative investigation of how adapter, LoRA and FitBit techniques perform on multilingual multilabel classification tasks, both in terms of classification performance and computation costs. We first perform an ablation study on one classification task to identify the PEFT techniques that demonstrate the best performance in multilingual and cross-lingual scenarios. Those best performing PEFT methods are then investigated further in three training scenarios (multilingual and cross-lingual) on three diverse classification tasks.

In particular, we study the behaviour of these PEFTs on three multilingual multilabel news article classification tasks introduced as separate sub-tasks of the recent SemEval 2023 Shared Task 3 [15]: news genre, framing and persuasion technique detection. All three tasks are based on multilingual data and contain 'unseen' languages, i.e. languages not available in the training set, but present in the test set. Further motivation for this research came from the success of our three original best performing approaches in each of these three different sub-tasks [16, 17]. Our best method for sub-task 1 was based on an ensemble of FFT and adapter-based

models, as well as language-specific checkpoint selection. Our best sub-task 2 approach was based on mono- and multilingual ensembles, one of which combined FFT and adapter methods with task adaptive pre-training [18]. Finally, our models for sub-task 3 included language-specific classification threshold selection and the incorporation of unlabelled data into the training corpus. Overall, we found that adapters improved performance in certain monolingual scenarios in sub-task 1 and for multilingual ones in sub-task 2.

The main contributions of this paper are the following:

- **Provide an evaluation of PEFTs on diverse text classification tasks** as compared to fully fine-tuned models.
- **Compare the computational costs for training FFT and PEFT models on the three sub-tasks as such an evaluation has not been performed before on these datasets and tasks**. We show that PEFTs significantly reduce the number of trainable parameters (between 140 and 280 times less parameters) and achieve shorter training times (between 32% and 44%). Unlike previous analyses between bottleneck adaptors, LoRA and FFT, our comparison is novel is providing more fine-grained statistics, such as the peak VRAM usage and the relative train time duration. The latter statistic is particularly important because the training process typically more time-consuming than inference.
- **Carry out an in-depth comparison of the performance of PEFTs in different training scenarios**, investigating both joint multilingual training and two types of single source language training scenarios. We evaluate how each method performs on seen languages and generalises to unseen ones.
- **Improve on the original highest SemEval 2023 results**. For sub-task 3, we achieved a better performance on eight out of the nine languages compared to the top results in the official leaderboard. For sub-tasks 1 and 2 the results reported here are mostly comparable to our original SemEval 2023 submissions despite the fact that in this study we use significantly less complex models (the original solutions utilised multiple sub-task tailored steps and complex ensembles).

Related work

Parameter-efficient fine-tuning techniques

As already mentioned above, PEFTs are computationally efficient due to limiting the number of trainable parameters. All such techniques freeze the pre-trained model, but differ in the location of the inserted trainable parameters. We focus specifically on Bottleneck adapters and on Low-Rank Adaptation (LoRA).

Bottleneck adapters [10] have a structure similar to autoencoders. The transformer hidden state h is first down-projected to some smaller dimensionality $d_{\text{bottleneck}}$ with matrix W_{down}, passes through non-linearity function f, and is then up-projected to the original dimensionality with matrix W_{up}, and a residual connection r. This is defined formally in Eq 1.

$$h \leftarrow W_{\text{up}} \cdot f(W_{\text{down}} \cdot h) + r \tag{1}$$

The location of the adapter layer depends on the adapter type. Houlsby adapters [10] place adapter layers after the multi-head attention and feed-forward block, whereas Pfeiffer adapters [19] place the adapter layer only after the feed-forward block. Although adding the extra layer reduces the number of trainable parameters and thus increases the speed of fine-tuning, it also increases the number of overall parameters permanently, slowing down inference.

LoRA [12] adds low-rank decomposition matrices to the query (Q), key(K), value (V) and pre-trained W_0 matrices of the self-attention sub-layers of the transformer.

Given a layer expressed as the matrix multiplication $h \leftarrow W_0 x$ (where $W_0 \in \mathbb{R}^{d \times k}$), during the fine-tuning process, the value of W_0 is modified by some ΔW. LoRA represents this delta as the low-rank decomposition $\Delta W = BA$ (where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$) of rank $r \ll \min(d, k)$. Here, W_0 is frozen, while B and A are initialised randomly and updated during fine-tuning. The decomposition is scaled by hyperparameter α and rank r, thus giving the new expression in Eq 2.

$$h \leftarrow W_0 x + \frac{\alpha}{r} BAx \tag{2}$$

Once the fine-tuning stage has been completed, the additional matrices can be removed by simplifying W_0, A and B to a single matrix W_0', thus giving the same number of parameters as the original pre-trained model. This solves the problem of increased inference time. When performing hyperparameter search, the α hyperparameter can be fixed since it is proportional to the learning rate [20].

Transformer models consist of six weight matrices, W_0, W_K, W_V, W_Q, and two matrices in multilayer perceptron (MLP) layer. In principle, it is possible to adapt any number of weight matrices, however, the authors claim that adapting only self-attention matrices (W_0, W_K, W_V, W_Q) produces results comparable to adapting all layers [12].

Evaluating using LoRA adaptation in other parts of the model (all attention layers, all feed-forward layers, all layers, and attention and feed-forward output layers) revealed that inserting LoRA adaptation in all layers results in the highest performance, and that in this configuration the hyperparameter r has no effect [20].

BitFit [3] uses only a small percentage of model parameters called *bias terms*. This allows a significant reduction in the number of trainable parameters.

Each self-attention head m at layer l consists of the key (K), query(Q) and value (V) linear layers. Unlike in the FFT fine-tuning method, only the bias terms b^l are considered during fine-tuning:

$$\begin{aligned} Q^{m,l}(x) &= W_q^{m,l} x + b_q^{m,l} \\ K^{m,l}(x) &= W_k^{m,l} x + b_k^{m,l} \\ V^{m,l}(x) &= W_v^{m,l} x + b_v^{m,l} \end{aligned} \tag{3}$$

Multiple attention heads are then combined using an attention mechanism and fed into the MLP with layer-norm (LN). The biases from the query, key and value layers, attention and normalisation layers are the only fine-tuned parameters of the network. The $W^{(l),\cdot}$ matrices and other MLP parameters are kept frozen. This reduces the number of trainable parameters to 0.08%–0.09%.

The authors additionally show how the number of trainable parameters can be further reduced by focusing only on the bias terms from the query layer and the second MLP layer.

Related work on the comparison of adapters, LoRA and BitFit to FFT

Despite significantly reducing the number of trainable parameters, bottleneck adapters were previously found to have minimal negative impact on the performance of fine-tuned models for simple sentence classification tasks. In particular, the evaluation on the GLUE benchmark dataset [21] (a collection of sentence and sentence-pair classification tasks) showed the

bottleneck adapter performance to be within 0.8% of the performance of FFT, whilst only training 3.6% of parameters [10]. Additionally, in the field of machine translation, Bapna and Firat [9] found that adapters produce equivalent or even better results compared to FFT.

Adapters received special attention in the context of multilingual tasks, with various studies reporting a consistent advantage of adapter models in comparison to FFT. In particular, Pfeiffer et al. [19] proposed a modular adapter-based architecture, which combines task-specific adapters in the source language with 'language' adapters trained on unlabelled data in the target language using the masked language modelling (MLM) objective. The authors report that this framework is able to outperform the traditional fully fine-tuned language models in cross-lingual transfer inference for the majority of language pairs. Different from Pfeiffer et al. [19], He et al. [11] investigated the zero-shot cross-lingual capabilities of adapters without additionally training them on unlabelled target language data. The authors found that adapters still outperform FFT on named entity recognition (NER), part of speech tagging (POS) tagging and cross-lingual Natural Language Inference (XNLI) tasks. They report that the former method is particularly beneficial in low-resource and cross-lingual tasks, since it mitigates forgetting effects by minimising the differences between the representations of the fine-tuned and the pretrained model.

Unlike previous studies that mainly focused on the analysis of short texts, Chalkidis et al. [13] investigated the performance of adapter models on long legal documents. Consistently with previous works, the authors found that bottleneck adapters outperform FFT, and provide better zero-shot cross-lingual capability. Their findings are based on the MultiEURLEX dataset, which consists of 65,000 EU law texts in 23 languages, categorised at multiple levels of detail (between 21 and 567 categories). In some aspects, this dataset is comparable to the dataset for sub-task 3 that we explore in this work, as both datasets are multilabel and multilingual, and have a comparable number of labels at MultiEURLEX's lowest level. However the style of EU law is naturally much more rigid and very different to news articles. Additionally, due to the specific data collection methodology used for MultiEURLEX, this dataset is not likely to contain irrelevant text. Finally, the MultiEURLEX dataset contains parallel multilingual data, meaning that the model making a cross-lingual zero-shot prediction on the target language has already 'seen' this text at training time in another language.

The latter limitation of the MultiEURLEX dataset became the focus of the study by Xenouleas et al. [14] who questioned whether the findings of Chalkidis et al. would generalise to non-parallel datasets. When the dataset is modified to include only non-parallel documents, the authors found that translation-based methods (translate-test and translate-train) outperform multilingual models. Consistently with previous research, however, the authors observed that adapters still outperform the FFT in each of the settings, cross-lingual and translation-based. We note that this is likely dependent on domain and on whether the relevant properties are significantly affected by translation: legal documents are more likely to be properly represented in the target language than, for example, the language-specific linguistic properties signalling certain persuasion techniques. Additionally, the joint multilingual experiments conducted by Xenouleas et al. [14] do not leave out any languages to perform the zero-shot cross-lingual inference, which makes this approach not comparable to the monolingual ones due to the difference in the size of the training set.

Taking into account the limitations of prior research in terms of the input length, high specificity of the domains, and parallelism of multilingual data, we aim to perform a wider comparison of adapter models with FFT on a variety of tasks and on non-parallel input texts that are not limited to a specific domain and vary in length. Another important limitation of prior work is that all the studies mentioned above evaluate cross-lingual transfer capabilities in a one-to-many manner, while the joint multilingual training setting does not perform zero-shot

cross-lingual inference, making it not comparable with the monolingual zero-shot cross-lingual scenarios. To fill this gap and bring understanding into how the joint multilingual training data can affect cross-lingual capabilities of the methods, we aim to introduce a many-to-many joint inference into our comparison scenarios while keeping certain languages as 'unseen' for all the training scenarios.

To our knowledge, no comparison of LoRA and BitFit to the FFT method in a similar multilingual scenario exists. However, given the prior evidence that LoRA and BitFit can, for certain tasks, outperform FFT [3, 12], there exists a clear motivation of performing a pioneering evaluation of these PEFT techniques in multilingual scenarios. LoRA technique can be particularly promising for our objective due to its ability to perform better on longer texts than adapter methods due to the fact that LoRA does not reduce the input sequence length [12].

Dataset and task description

To select the tasks and datasets suitable for our objective, we analysed the corpora from the most recent survey on multilingual datasets [22] that keeps real-time track of the multilingual data. We selected the datasets for the similar tasks of classification and sentiment analysis, and filtered out the datasets that contain multilingual data for short inputs—lemmas, word-pairs, sentences, tweets and short statements. We then manually analysed each dataset based on how the multilingual data was collected, and filtered out parallel multilingual datasets. Finally, we filtered out the datasets that focus on specific narrow domains. This method narrowed down the scope of our interest to the dataset that was created recently as part of *SemEval-2023 Task 3: "Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup"* [15].

Prior to SemEval 2023, a number of other related challenging multilingual misinformation and propaganda detection tasks were addressed in SemEval (https://semeval.github.io) shared tasks, including detection of hyperpartisan content [23], sarcasm [24], and a smaller set of persuasion techniques in textual [25] and multimodal [26] data. The Shared Task 3 within SemEval 2023 challenge extended this prior work on persuasion techniques by introducing new kinds of persuasion techniques, as well as addressing two other related sub-tasks, namely news genre categorisation and framing detection.

Sub-task 1: News genre categorisation

Given a news article, determine whether it is objective news reporting, an opinion piece, or satire.

Sub-task 2: Framing detection

Given a news article, identify one or more of fourteen framing dimensions used: *Economic*, *Capacity and resources*, *Morality*, *Fairness and equality*, *Legality, constitutionality and jurisprudence*, *Policy prescription and evaluation*, *Crime and punishment*, *Security and defense*, *Health and safety*, *Quality of life*, *Cultural identity*, *Public opinion*, *Political*, *External regulation and reputation*. The set of framing techniques used in this shared task was defined following a pre-existing taxonomy [27].

Sub-task 3: Persuasion techniques detection

Given a paragraph of a news article, identify zero or more out of 23 persuasion techniques used (see S1 Appendix for a detailed list of the techniques). The set of techniques represents an

Table 1. Properties of genre, framing and persuasion technique detection tasks.

Comparison criteria	Sub-task 1	Sub-task 2	Sub-task 3
Task type	Genre Detection	Framing Detection	Persuasion Techniques
Input Type	Whole document	Whole document	Paragraphs
Granularity	Multiclass	Multilabel	Multilabel
Official scoring metric	$F1_{macro}$	$F1_{micro}$	$F1_{micro}$
Number of classes	3	14	23
Avg number of tokens	1,157	1,157	74
Train set size	1,234	1,238	10,927
Number of source languages	6	6	6
Number of target languages	9	9	9
Task subjectivity	High	Low	Medium
Data imbalance	12.7	4.6	44

https://doi.org/10.1371/journal.pone.0301738.t001

extension of the taxonomy used in previous SemEval datasets [26, 28]. The task additionally provides 6 high-level categories that subsume similar persuasion techniques. Although the task is paragraph level, each of the articles has at least one labelled paragraph.

It should be noted that three of the systems that participated in the original SemEval-2023 Task 3 [15] evaluation exercise used adapters. Teams HHU [29] and NAP [30] entered only sub-task 3, in which they used adapters, whereas SheffieldVeraAI [16] applied adapters to sub-tasks 1 and 2. Initial performance analysis in these sub-tasks showed the effect of adapters to be inconsistent across the different sub-tasks. Namely, adapters achieved higher average performance for monolingual models in sub-task 1, while hindering performance of monolingual models in sub-task 2 but achieving better results there for multilingual models. This evidence provides an even stronger motivation gaining better understanding of the effectiveness of adapter methods across a range of classification tasks ranging in difficulty.

These three sub-tasks use broadly overlapping data, i.e. same input articles, however, differ in properties and the summary statistics for the datasets, which are summarised in Table 1.

We estimate the data imbalance as a ratio of the class samples for the most frequent class in the training set to that for the most rare one. Below, we discuss some of the individual properties in more detail.

Class imbalance

In addition to posing multilingual and multiclass classification challenges, the data for these three sub-tasks is highly imbalanced, which adds further difficulty. In particular, the class distribution for sub-task 1 is highly skewed, with 76% falling into the *opinion* class and *satire* accounting for only less than 6% of the data. For sub-task 2, the distribution of classes is also uneven, but is less skewed compared to sub-task 1. The most common frame is *Political* which appears in 49.4% of the training articles. The least common frame is *Cultural Identity* which appears in just 10.8% of the articles. Finally, in sub-task 3 *loaded language*, *doubt* and *name calling* are the most common persuasion techniques, accounting for 22%, 15.6% and 12.8% of the training paragraphs, respectively. The remaining 20 classes, on average, account for 2.5% of the training set, totalling 49.6% together. Particularly, *appeal to time*, *whataboutism* and *red herring* are the least frequent persuasion techniques, representing 0.5%, 0.5% and 0.7% of the training paragraphs, respectively.

Dataset statistics: Multilinguality, size and input length

Three sets of data are provided for each language and task: labelled *training* and *development* (except for unseen languages), and unlabelled *testing*.

The task organisers provided test data in nine languages: English (EN), French (FR), German (DE), Georgian (KA), Greek (EL), Italian (IT), Polish (PL), Russian (RU), and Spanish (ES). Three of the languages (Georgian, Greek and Spanish) are 'surprise' languages, meaning that no corresponding labelled training data exists in the dataset. Therefore, in order to make predictions for these languages, their test set must either be translated to a 'seen' language, or a multi-lingual approach capable of supporting zero-shot evaluation must be applied. For the remaining 6 languages, labelled training and validation data is included in the dataset.

It must be noted that the task organisers have not yet released the gold labels for the test set in order to prevent researchers from overfitting their systems. This means that detailed error analysis can only be carried out on the six languages for which the development sets are available.

Tables 2–4 show the detailed analysis of the training, development and test data used in sub-tasks 1, 2 and 3 respectively. The average length, calculated in the number of tokens, was estimated using the tokenizer for RoBERTa-large model [31], since this is the model we use in our experiments. For the training and development sets in sub-task 3, the average length was calculated for the paragraphs that have at least one persuasion technique assigned. For the test sets in sub-task 3, the average length includes every paragraph due to the lack of gold-standard

Table 2. Data statistics per language for sub-task 1: Genre detection.

Language	Number of examples			Average number of tokens		
	Training	Development	Test	Training	Development	Test
EN	433	83	54	1,307	1,066	978
FR	158	54	50	1,241	1,025	927
DE	132	45	50	995	913	1,203
IT	226	77	61	975	856	958
PL	144	50	47	1,354	1,438	1,935
RU	142	49	72	1020	797	547
ES	0	0	30	N/A	N/A	838
EL	0	0	64	N/A	N/A	1,071
KA	0	0	29	N/A	N/A	429

https://doi.org/10.1371/journal.pone.0301738.t002

Table 3. Data statistics per language for sub-task 2: Framing detection.

Language	Number of examples			Average number of tokens		
	Training	Development	Test	Training	Development	Test
EN	433	83	54	1,307	1,066	978
FR	158	53	50	1,196	1,059	927
DE	132	45	50	1,008	875	1,203
IT	227	76	61	965	885	958
PL	145	49	47	1,369	1,397	1,935
RU	143	48	72	1,009	827	547
ES	0	0	30	N/A	N/A	838
EL	0	0	64	N/A	N/A	1,071
KA	0	0	29	N/A	N/A	429

https://doi.org/10.1371/journal.pone.0301738.t003

Table 4. Data statistics per language for sub-task 3: Persuasion techniques.

Language	Number of examples			Average number of tokens		
	Training	Development	Test	Training	Development	Test
EN	3,610	1,103	11,466	88	37	65
FR	1,693	437	7,140	97	108	91
DE	1,251	405	11,060	95	87	76
IT	1,742	594	8,302	99	91	99
PL	1,228	415	14,084	109	122	90
RU	1,232	310	8,414	85	80	66
ES	0	0	1,320	N/A	N/A	76
EL	0	0	3,792	N/A	N/A	72
KA	0	0	640	N/A	N/A	78

https://doi.org/10.1371/journal.pone.0301738.t004

labels for the test data. This is why the number of examples in the test-sets for sub-task 3 is significantly higher than that for the training and development sets. However, not all of these examples are expected to contain at least one persuasion technique.

Sub-tasks 1 and 2 use the same set of articles in the test set, while the accumulative set of articles used in the development and training sets is also identical for these two sub-tasks, their assignment to a certain set varies slightly. This is why, as we can see from Tables 2 and 3, the data statistics for these sub-tasks are quite similar. As can be seen, the distribution of the training examples across the languages is not even, with EN accounting for almost 4 times as many articles as DE, RU and PL. We can also observe that the average length of the articles is highly dependent on the language, with articles in the test set for Georgian (KA) being more than 4.5 times shorter than that for the articles in the test set for Polish (PL). This aspect is important for our experiments since it suggests that the models are more likely to omit important information for certain languages compared to others, due the the limitation of transformer models in terms of the input length. Another observation is that the training set is not always representative of the test set. For example, the articles in the test set for Russian (RU) are twice shorter, on average, than the articles in the training set for this language. The difference in terms of the input length between training and test data is less significant for sub-task 3, which uses paragraphs as an input, and for which all the inputs are within the limit of transformer models.

We analysed the languages in the training and test sets in terms of the amount of resources. For each language, we considered the number of training examples in that language or in a language from the same language family and the amount of pre-training data in that language for the model that we use in our experiments. Our analysis places Georgian as a clear outlier and a low-resource language, having significantly less pre-training data than the other 8 languages and no training data in Georgian or a related language. English, on the other hand, is the highest-resource language, benefiting from most training and pre-training data and additional training data in a related language. The details of this analysis are provided in S4 Appendix.

Task subjectivity

As specified in the annotation instructions [32], the subjectivity differs across the sub-tasks. Sub-task 1 relies on a very nuanced analysis of the whole article and on commonsense knowledge, since, as mentioned by the organisers, the satirical articles often "tend to mimic true articles" and to mention "real-world individuals, organisations, and events", while the distinction between opinionated and objective reporting can lie in the certain ways the reporters tend to balance out the reported opinions. The authors also highlight that "the borders between

opinion and reporting might be sometimes blurred" and "a news article which contains some small text fragment, e.g., a sentence that appears satirical" does not normally trigger a satire genre. The task of framing detection tends to be more reliant on a certain linguistic information as the annotation instructions ask the annotators to specify exact text spans corresponding to a certain frame. The authors additionally provide examples of the discussion topics that trigger certain frames. For example, "costs, benefits, or other financial implications" usually are indicative of the "economic" frame. Finally, sub-task 3 is the most fine-grained task as it provides span-level annotations of the propaganda techniques. The detection of persuasion techniques usually relies on a certain argumentative structure and linguistic triggers, such as the mention of an entity that is considered an authority ("appeal to authority" technique), associating an opponent with a group, event or concept that has negative connotations ("guilt by association" technique), "the noun phrase, the adjective that constitutes the label and/or the name" ("name calling or labeling" technique) or "text fragments that repeat the same message or information that was introduced earlier" ("repetition" technique). More objective assessment of the subjectivity is not possible due to the lack of the inter-annotator agreement scores [15].

Dataset collection

The data is extracted from both mainstream and alternative media sources, collected through news aggregators (e.g. Google News, Europe Media Monitor) and fact-checking organisations (e.g. MediaBiasFactCheck, NewsGuard), respectively. All of the news articles were published between 2020 to mid 2022. The text of each article was extracted automatically from the HTML source of each web page by using either the text-gathering tool Trafilatura [33] or a site-specific procedure. Notably, this process is error prone as it sometimes includes textual content which is not a part of the news article itself, such as web polls, newsletter sign-up forms, and author information. For English, a pre-existing dataset was also utilised [28], but the organisers of the shared task did not make it sufficiently clear as to what other English data was included in the new dataset.

Methods

Training scenarios

While our primary focus is on the multilingual fine-tuning scenario, we introduce two additional settings where models are trained on English-only data in order to investigate whether the effectiveness of each training method differs depending on the composition and the size of the training set.

These three different training scenarios are summarised below:

- **Multilingual Joint (many-to-many)**: models are fine-tuned using all training data in the original 6 languages.
- **English + Translations (one-to-many)**: models are fine-tuned on all the original English training data and English translations of the training data in the other 5 languages. It is important to mention that the test data in this scenario was kept in its original languages, meaning that the predictions on all the languages except for English were made in a zero-shot cross-lingual way.
- **English Only (one-to-many)**: models are fine-tuned on only the original English data in the training set. Similarly to the 'English + Translations' scenario, the test set was not translated into English.

The choice of the three training scenarios above is motivated by the fact that we want to evaluate the effect of two different factors on each of the training techniques:

1. The 'English + Translations' training scenario enables the evaluation of the effect of multilinguality. In particular, we want to compare the effect of having multilingual training data against the scenario where training data is available in only one language. By translating other languages into English, we compose a dataset consisting of the same number of training examples but without having the language diversity. This eliminates the possibility that differences in performance across the three methods could be due to the size of the training data for each language. At the same time, machine translation, as a specific transfer paradigm for cross-lingual learning [34], may introduce some level of noise and thus break the required correspondence between the original and translated sample.
2. The 'English Only' training scenario enables the analysis of the effect of training data size on each method. This can be achieved by comparing performance on 'English Only' training data against performance on 'English + Translations' data, where the only difference between the two is in the number of training examples. This training scenario, however, is not directly comparable against the multilingual training scenario, since it does not eliminate the possibility that differences in the method's performance could be due to the different linguistic characteristics of the multilingual training data.

Training techniques

In selecting the PEFT methods for our further analysis, we performed a detailed ablation study on sub-task 1, where we compare the adapter, LoRA and BitFit methods in multilingual and cross-lingual classification scenarios (see S2 Appendix). As can be seen, the BitFit method does not reach a performance comparable to that of LoRA and adaptor methods for any of languages in any of the classification scenarios. We therefore exclude this PEFT technique from our further experiments.

We experiment with XLM-RoBERTa Large [35], using the following training techniques:

- **Full fine-tuning (FFT)**: All parameters of the model are updated during fine-tuning.
- **Low-Rank Adaptation (LoRA)**: The model's parameters are frozen and LoRA matrices (key, query, value) are added to both the MLP and attention layers.
- **Bottleneck Adapter (Adapter)**: The model's parameters are frozen and bottleneck adapters in the Pfeiffer configuration [19] are added to all the layers. As we discuss further, the choice of the adapter configuration was suggested by our ablation experiments that demonstrated the Pfeiffer configuration to outperform the Houlsby version on average (S3 Appendix).

The next section describes the methodology behind each training technique and training scenario.

Experimental setup

Model and hyperparameter selection. When selecting the model size, we took into account prior comparison of the LoRA and adapter-based PEFT methods for different sizes of RoBERTa model across a wide range of tasks and found a clear preference for a larger model size [12, 19]. For both FFT and PEFT methods, we additionally conducted the comparison of XLM-RoBERTa Base with XLM-RoBERTa Large on our sub-tasks in the default training scenario which includes all the data for each sub-task in the original form (the 'Multilingual Joint'

scenario). The results of this comparison are presented in S3 Appendix. As can be observed, a large model demonstrates consistently better average performance across all three sub-tasks. Therefore, based on both prior evidence for PEFT techniques and our ablation studies for FFT, we resolved to using XLM-RoBERTa Large in our main experimental setup.

Choosing between the Pfeiffer [36] and Houlsby [10] configuration of the adapter method, we performed the comparison of both adapters for the 'Multilingual Joint' training scenario. The results of this comparison are presented in S5 Appendix. As can be seen, Pfeiffer adapter shows slight average advantage over Houlsby one in all three sub-tasks. We therefore use the Pfeiffer configuration in the rest of our experiments.

For the best hyperparameters, we first perform a search for each training scenario and training technique within each sub-task. The search is performed on the original development set as provided by the organisers of SemEval 2023 Task 3. The best configuration obtained for each method can be found in Table 5. One needs to bear in mind that our objective is to maximise model performance per training scenario for each sub-task rather than to minimise the computational costs. In other words, greater parameter efficiency could be possible, but at the cost of model performance.

For the bottleneck adapter, we use the default Pfeiffer configuration [36] by adding the adapter after each 'ffn' sub-layer. Despite prior ablation studies showing that the lowest 4 layers have little contribution to the performance [20], our objective is to maximise our performance and to make the setting comparable with LoRA where we use all the layers.

Text preprocessing. As shown in Table 1, sub-tasks 1 and 2 have an average number of tokens per article of 1,157. Thus in those sub-tasks, 80.0% of articles are truncated to a maximum of 512 tokens. In contrast, sub-task 3 presents an average number of tokens per sequence of 74, thus all training sentences are fully encoded without loss of information. For sub-task 1, the articles that are longer than 512 tokens are separated into sentences, which are then sampled sequentially from the beginning and the end of the article, preserving the original order, until the maximum of 512 tokens is reached. Such a truncation approach is motivated by our experiments on sub-task 1 data during the competition stage of SemEval 2023 Task 3 [16]. This approach yielded a significant improvement in the $F1_{macro}$ score over the setting that

Table 5. Hyperparameters.

Full fine-tuning (FFT)			
Hyperparameter	Sub-task 1	Sub-task 2	Sub-task 3
Learning Rate	1.00E-05	3.00E-05	3.40E-04
Batch size	16	8	32
Low-Rank Adaptation (LoRA)			
Hyperparameter	Sub-task 1	Sub-task 2	Sub-task 3
Learning rate	7.00E-06	3.00E-4	1.59E-03
Batch size	16	8	32
Rank	8	8	2
Layers	all	all	all
Dropout	5.00E-01	5.00E-01	5.00E-01
Attention matrix	k,q,v	k,q,v	k,q,v
Bottleneck Adapter			
Hyperparameter	Sub-task 1	Sub-task 2	Sub-task 3
Learning rate	3.16E-05	2.00E-4	4.30E-04
Batch size	16	8	32
Reduction factor	4	8	8

https://doi.org/10.1371/journal.pone.0301738.t005

simply truncates texts to the first 512 tokens. This improvement can potentially be explained by the fact that the instructions for human annotators in sub-task 1 highlighted the importance of opinionated sentences which tend to be found towards the end of the articles.

We perform text preprocessing for sub-tasks 1 and 2 by applying the following steps for all languages:

- a full stop is added at the end of each title;
- duplicate sentences directly following each other are removed;
- the @ symbol is removed from any Twitter handles;
- hyperlinks to websites and images are also removed.

English articles were further preprocessed as follows:

- text promoting sharing on different social media platforms was removed from the bottom of the articles;
- sentences encouraging user participation in online polls, comments, or advertisements were also removed;
- sentences stipulating the site's terms of use were removed;
- removal of sentences indicating licensing and containing phrases such as 'reprinted with permission', 'posted with permission' and 'all rights reserved';
- sentences detailing author biographies were also removed.

For sub-task 3, preliminary experiments found no performance gains when text preprocessing was applied, thus our experiments for this sub-task use directly the original text. Importantly, for sub-task 3 experiments, we include sentences that do not have assigned labels into the training data by assigning them a vector of zeros to indicate that they do not belong to any class. This approach was shown to significantly improve classification performance on this sub-task in our initial experiments [16, 17]. The size of the training set displayed in Table 1 (10,927 examples) is based on the number of labelled examples. When unlabelled sentences are added, training data for sub-task 3 grows to 20,704 instances.

The multilabel sub-tasks 2 and 3 use confidence thresholds of 50% and 30%, respectively, after applying a sigmoid activation function to the logits. The confidence threshold for sub-task 3 is purposefully lower and was selected according to our previous experiments [17], which revealed that its careful calibration can significantly influence the performance of the model.

Training scenarios. We experiment with the three training scenarios described previously. All models are trained on the original training split provided by the task organisers, using either data in all 6 seen languages; all EN data and the translations into English of the data in the other 5 languages; or only using the EN part of the training data. Each model is then evaluated on the task organiser's test split (6 seen and 3 surprise languages), without translation.

Aligned with previous research [11, 14, 37], we define the 'unseen' language is a 'target' language on which a cross-lingual zero-shot prediction is done and which is different from the 'source' language on which the task-specific fine-tuning of the transformer model was performed. The three unseen test set languages—Greek, Georgian, and Spanish—allow us to evaluate the zero-shot cross-lingual transfer learning capabilities of the training techniques trained in the first, fully multilingual, setting. In the other two training scenarios ('English + Translations'

and 'English Only'), the remaining 8 languages (FR, DE, IT, PL, RU, ES, EL and KA) provide an insight into the models' performance in the cross-lingual zero-shot setting.

Evaluation metrics. The performance of the different training techniques is then compared using two sets of criteria: (1) computational resource efficiency; (2) classification performance. For the latter, both $F1_{micro}$ and $F1_{macro}$ are reported as performance metrics for all three sub-tasks. However, it must be noted that the SemEval 2023 Task 3 organisers used only $F1_{macro}$ as the official scoring metric for sub-task 1, whereas sub-tasks 2 and 3 used only $F1_{micro}$. Therefore, where a more detailed language-specific analysis is carried out in this paper, only the respective official metric for each sub-task is provided.

Mean and standard deviation are computed over three different random seed initialisations.

Resource efficiency is measured through four metrics: (i) the peak amount of VRAM used during training; (ii) speedup relative to the fully fine-tuned method, which is the number of training steps per second N_m/t_m of the respective method (LoRA or Adapter) divided by the number of training steps per second of the fully fine-tuned method N_{FFT}/t_{FFT} (Eq 4); (iii) the number of trainable parameters; and (iv) the number of non-trainable parameters.

$$S^* = \frac{\frac{N_m}{t_m}}{\frac{N_{FFT}}{t_{FFT}}} \tag{4}$$

Implementation details. All experiments were performed with the AdapterHub framework [36].

In order to obtain the 'English + Translations' data, we translate all available training and development data into English using Google Cloud Translation API. The choice of the Machine Translation (MT) system was motivated by the recent extensive report on the evaluation of MT systems [38]. The results across all 11 language pairs and 9 domains analysed by the authors show that Google Translate is the state-of-the-art system based on the COMET score. We believe that this choice helps to reduce the potential noise caused by the translation, however, it is difficult to quantify the effect of the noise for the general case of news articles, since, as the report shows, the performance of the systems varies highly depending on the domain. Moreover, while we have the performance estimates for EN-DE, EN-FR, EN-ES and EN-IT pairs (which are comparable across all domains), we could not find the relevant results for the remaining 4 language pairs (EN-PL, EN-KA, EN-RU, EN-EL).

Our code is available on GitHub (https://github.com/GateNLP/PEFT_FFT_multilingual) and the Zenodo repository (https://doi.org/10.5281/zenodo.10066649).

Results

The analysis of our results is structured around the following three main research questions:

RQ1: How does the classification performance and computational costs of each training technique differ for each sub-task?

RQ2: How do the training scenarios (determining the diversity of the languages in the training set and its size) affect the performance of each training technique?

RQ3: How do the training techniques compare with each other for each training scenario and language?

The importance of the research question posed in this study is motivated by the limitations of prior research on the performance of PEFT techniques in the multilingual article-level classification tasks.

RQ1 provides a high-level analysis of each training technique under the best training scenario. This comparison is important because it provides the first to our knowledge examination of the PEFT and FFT techniques on non-parallel texts that are not restricted to a narrow domain and range in length from paragraphs to long articles. This RQ is particularly novel for the LoRA method since, as we highlighted above, no comparison of LoRA with adapter methods or FFT was previously performed for the multilingual multilabel classification task. We additionally provide the comparison of the computational costs as such a comparison was not previously performed for any of the sub-tasks in this study. It is not possible to assume that previous results would hold for our tasks given that the computational efficiency of PEFT methods can be sensitive to the input length [12]. Unlike previous analyses between bottleneck adaptors, LoRA and FFT that either perform a task-agnostic inference latency analysis or provide the number of trainable parameters for specific tasks [12], our comparison focuses on more fine-grained statistics, such as the peak VRAM usage and the relative train time duration. The latter statistic is particularly important because the training process is typically more time-consuming than inference, and the relation between the number of trainable parameters and the training time is not linearly proportional.

RQ2 provides better insights into how the performance of each training technique changes in different training scenarios. As mentioned in the 'Related work' section, previous experiments [14] conducted on non-parallel multilingual tests conclude that the zero-shot translation-based (translate-train and translate-test) approaches outperform cross-lingual ones for both FFT and adapter training techniques. However, their cross-lingual zero-shot approaches are limited to one-to-many scenarios, making it difficult to fairly compare the translation-based training settings to the joint multilingual scenarios due to the differences in the sizes of the training sets. We aim to address this gap by sequentially changing one aspect of the training process, by first eliminating the multilinguality of the training set whilst keeping the same size, and then reducing the size and eliminating potentially noisy translated texts from the training set. Additionally, previous joint multilingual experiments [14] do not leave any languages out to perform the zero-shot cross-lingual inference, which also makes it impossible to compare this approach to the monolingual ones. Finally, we add LoRA to our set of PEFT methods and provide novel insights on the performance of this technique under different scenarios in multilingual tasks.

RQ3 focuses on a different dimension of the problem and tries to answer which method is preferable depending on the amount and type of the training data. This analysis is motivated by the prior work by He et al. [11] on short texts that concludes that adapter methods are particularly beneficial for low-resource and cross-lingual tasks. We therefore investigate how PEFT methods and FFT compare to each other as we reduce the number of training resources in certain languages and reduce the overall amount of training data.

Comparison of the computational and performance properties of training techniques

To answer the first research question (RQ1), we select the best training scenario for each method and sub-task and compare the performance of the FFT model against the performance of the LoRA and adapter methods for each sub-task. The results of this comparison are reported in Table 6. We report mean scores after three runs with different random seeds along

Table 6. Performance and computational costs for each sub-task and training technique.

Sub-task 1: Genre Detection			
	FFT	LoRA	Adapter
$F1_{macro}$ *	**59.9 ± 3.1**	57.9 ± 6.3	58.0 ± 2.0
$F1_{micro}$	61.7 ± 7.5	60.2 ± 3.9	**62.8 ± 5.7**
Best training scenario	Multilingual Joint	Multilingual Joint	Multilingual Joint
Peak VRAM usage	∼39GB	∼**24GB**	∼28GB
Train Time relative to FFT	1	**0.59**	0.67
Trainable parameters	∼560M	∼3.2M	∼26M
Non-trainable parameters	0	∼560M	∼560M
Sub-task 2: Framing Detection			
	FFT	LoRA	Adapter
$F1_{macro}$	**49.2 ± 7.4**	45.3 ± 7.1	47.1 ± 7.9
$F1_{micro}$ *	**56.7 ± 6.1**	53.4 ± 6.0	54.8 ± 7.1
Best training scenario	Multilingual Joint	Multilingual Joint	Multilingual Joint
Peak VRAM usage	∼23GB	∼18GB	∼**14GB**
Train Time relative to FFT	1	0.68	**0.56**
Trainable parameters	∼560M	∼4M	∼7M
Non-trainable parameters	0	∼560M	∼560M
Sub-task 3: Persuasion Techniques			
	FFT	LoRA	Adapter
$F1_{macro}$	**23.7 ± 5.0**	**23.7 ± 6.9**	20.6 ± 6.3
$F1_{micro}$ *	41.8 ± 8.6	**42.9 ± 9.5**	42.2 ± 9.5
Best training scenario	Multilingual Joint	Multilingual Joint	Multilingual Joint
Peak VRAM usage	∼20GB	∼**13GB**	∼16GB
Train Time relative to FFT	1	**0.56**	0.71
Trainable parameters	∼560M	∼2M	∼7M
Non-trainable parameters	0	∼560M	∼560M

The best scores and performance metrics appear in **bold**. The main metric for a certain sub-task is marked with an asterisk (*).

https://doi.org/10.1371/journal.pone.0301738.t006

with standard deviations. The standard deviation is the square root of the average of the squared deviations from the mean.

The results demonstrate that:

(1) **FFT and adapters perform better in sub-tasks 1 and 2, while LoRA performs better for sub-task 3**. We observe that for longer texts, such as the articles analysed in sub-tasks 1 and 2, FFT and adapter-based classification demonstrates better results on average than LoRA. At the same time, LoRA on average outperforms FFT and adapters for sub-task 3, which is trained on shorter texts.

(2) **'Multilingual Joint' training scenario performs best, regardless of the sub-task and training technique**. We observe a pattern of 'Multilingual Joint' training scenario achieving the best results for all three sub-tasks as well as all three training techniques. This implies that, in general, training models on larger datasets with a variety of languages, can be beneficial for both FFT and PEFT methods applied to the tasks with various properties. This effect has not been, however, consistently observed for all combinations of training scenarios and training techniques (the more detailed analysis per individual training scenarios is provided in the following section).

(3) **LoRA and adapters can save computational costs significantly**. By design, the PEFTs reduce the number of trainable parameters significantly (between 140 and 280 times less parameters). As a result, for sub-task 1 and 3, the utilisation of LoRA led to a significant decrease in the memory consumption—from 39GB to 24GB (38%), and from 20GB to 13GB (35%) respectively. For sub-task 2, the adapter achieved the best memory efficiency, while decreasing the peak VRAM usage from 23GB to 14GB (39%). The similar pattern can be observed for a total training time, which decreased to 56–71% of the FFT training time.

(4) **Saving computational costs results in lower performance, some exceptions, however, exist**. Saving the VRAM usage and shortening training time is naturally reflected in the lower performance compared to that of FFT. Adapters are consistently outperformed by FFT in all three sub-tasks. However, in the case of sub-task 3, LoRA not only achieved highly comparable results, but also outperformed FFT for the most of languages (the difference is, however, not statistically significant—the more detailed analysis is provided in the next sections). For sub-tasks 2 and 3, we also observe a higher standard deviation of the results, implying a higher instability of fine-tuning when PEFTs are applied.

Comparison of the effect of a training scenario on each training technique

To answer the second question (RQ2), we compare FFT, LoRA and adapter training techniques across the three training scenarios introduced above, namely 'Multilingual Joint', 'English + Translations' and 'English Only'. It should be noted that in the latter two scenarios all languages in the test set are unseen (except English) as the model did not have access to training data in those languages. The results are measured in the official sub-task metrics ($F1_{macro}$ for sub-task 1 and $F1_{micro}$ for sub-tasks 2 and 3) and are shown in Tables 7–9.

(1) **The diversity of languages in the training set improves the average performance of the FFT technique**. For all three sub-tasks, we observe a significantly better average classification performance when the training data is provided in the original 6 different languages as opposed to providing the same amount of data in English only. When looking at the individual languages, this effect holds for all languages in sub-task 1 except for English (the only seen

Table 7. Sub-task 1: Genre detection—Mean ± 1 STD $F1_{macro}$ scores.

Language	Multilingual Joint			English + Translations			English Only		
	FFT	LoRA	Adapter	FFT	LoRA	Adapter	FFT	LoRA	Adapter
EN	52.7±0.5	49.4±0.4	**52.8±0.2**	53.1±1.1	52.4 ±0.6	*__53.5±0.9__	40.9±1.3	39.1±1.2	**42.2±0.6**
FR	*__69.7±1.2__	67.4±2.3	67.5±0.9	68.0±1.9	**69.2±1.5**	68.4±0.7	65.0±2.2	66.3±0.5	**66.7±3.7**
DE	66.3±0.5	64.8±1.2	*__67.2±0.8__	65.4±2.5	63.9±3.0	**65.8±1.2**	62.1±4.2	**64.2±2.8**	63.6±0.7
IT	52.2±1.4	*__53.4±1.8__	52.0±3.1	52.1±1.7	52.0±1.5	**52.9±0.8**	45.1±3.6	**47.3±1.8**	44.2±1.1
PL	*__69.2±1.1__	66.8±0.4	65.2±1.5	61.3±0.7	**64.0±0.7**	61.8±2.4	58.9±2.2	**60.6±1.4**	59.6±3.0
RU	*__57.4±0.6__	55.7±1.7	52.8±0.9	52.4±1.3	54.2±0.8	**54.9±2.5**	**49.7±0.9**	49.4±2.4	48.7±0.6
Average	*__61.3±2.7__	59.6±1.9	59.6±3.1	58.7±1.8	59.3±5.2	**59.6±2.7**	53.6±3.5	**54.5±6.1**	54.2±2.9
ES	*__47.1±1.4__	41.8±0.5	44.2±0.7	44.5±2.7	46.0±1.3	**46.2±5**	**42.3±2.9**	41.7±0.8	40.9±1.1
EL	40.8±2.4	**41.4±2.7**	40.9±1.7	*__43.5±2.1__	42.9±1.7	42.2±3.6	**38.9±2.7**	38.6±1.7	37.5±0.8
KA	*__83.3±2.1__	80.8±5.0	79.2±1.8	**81.0±3.2**	79.6±4.1	77.5±2.2	**76.7±3.3**	72.9±1.5	74.8±2.4
Average	*__57.1±2.3__	54.7±2.4	54.8±1.8	**56.3±2.4**	56.2±3.3	55.3±1.9	**52.6±2.9**	51.1±4.0	51.1±3.3
All	*__59.9±3.1__	57.9±6.3	58.0±2.0	57.9±3.4	**58.2±3.8**	58.1±4.1	**53.3±5.2**	53.3±2.6	53.1±3.6

Best scores by language are marked with an asterisk (*). Best scores by training method for each training scenario are in **bold**.

https://doi.org/10.1371/journal.pone.0301738.t007

Table 8. Sub-task 2: Framing detection—Mean ± 1 STD $F1_{micro}$ scores.

Language	Multilingual Joint			English + Translations			English Only		
	FFT	LoRA	Adapter	FFT	LoRA	Adapter	FFT	LoRA	Adapter
EN	*55.8±0.2	52.2±1.7	55.7±2.0	54.9±1.9	54.3±1.7	**55.1±1.5**	46.5±1.3	46.2±1.3	**47.8±0.7**
FR	**53.3±3.3**	47.3±1.5	50.8±3.6	*54.5±2.5	52.9±2.5	50.3±2.5	**44.2±1.6**	41.0±0.8	41.7±0.4
DE	63.1±1.9	62.3±2.2	*64.2±1.0	58.9±1.6	55.8±1.8	**59.1±3.3**	46.6±3.3	45.7±1.1	**49.1±2.9**
IT	*59.9±1.9	56.8±1.6	58.2±1.0	57.0±1.9	55.7±0.4	**59.5±1.0**	54.9±1.9	**55.3±2.0**	**55.3±0.6**
PL	*65.2±0.8	61.0±0.8	64.1±1.7	55.3±3.8	54.1±1.5	**56.3±1.9**	45.5±0.5	**46.3±0.1**	46.1±2.7
RU	**45.3±3.0**	43.6±0.6	41.7±2.0	40.3±2.0	38.2±1.2	**41.0±2.7**	**35.1±1.1**	31.1±1.6	31.7±1.3
Average	*57.1±7.3	53.9±7.5	55.8±8.6	53.5±6.7	51.8±6.8	**53.6±7.0**	**45.5±6.3**	44.3±7.9	45.3±8.0
ES	*52.7±2.1	51.7±3.0	49.1±2.0	49.7±0.7	**49.9±1.0**	46.6±2.5	**41.8±1.9**	35.9±3.8	36.4±2.1
EL	*54.9±1.7	51.4±1.2	54.1±2.9	53.2±0.1	50.7±0.7	**55.5±1.1**	**47.2±1.2**	43.6±0.2	47.1±1.5
KA	*60.1±4.2	53.9±4.8	55.3±1.6	55.3±4.9	52.8±4.0	**57.0±4.6**	**51.4±0.9**	45.3±1.7	50.1±2.1
Average	*55.9±3.8	52.3±1.4	52.8±3.3	52.7±2.8	51.1±1.5	**53.1±5.6**	**46.8±4.8**	41.6±5.0	44.5±7.2
All	*56.7±6.1	53.4±6.0	54.8±7.1	53.2±5.5	51.6±5.4	**53.4±6.2**	**45.9±5.6**	43.4±6.9	45.0±7.3

Best scores by language are marked with an asterisk (*). Best scores by training method for each training scenario are in **bold**.

https://doi.org/10.1371/journal.pone.0301738.t008

language in the 'English + Translations' training scenario) and Spanish (one of the unseen languages in the joint multilingual setting). The decreased performance on the English test set in the 'Multilingual Joint' scenario for this task can potentially be explained with the 'negative inference' effect highlighted in previous studies [39]. However, we do not observe this effect consistently across the sub-tasks. In particular, for sub-task 2, the only exception is French, which benefits from a monolingual training setting. Notably, the performance on the English test set decreases in the 'English + Translations' scenario for sub-task 2, despite being trained on much more data in this language. Finally, for sub-task 3, French is the only language in the test set that benefits from being trained on monolingual English data, which is consistent with sub-task 2.

Table 9. Sub-task 3: Persuasion techniques—Mean ± 1 STD $F1_{micro}$ scores.

Language	Multilingual Joint			English + Translations			English Only		
	FFT	LoRA	Adapter	FFT	LoRA	Adapter	FFT	LoRA	Adapter
EN	34.9±1.7	**37.7±0.9**	37.5±2.9	42.2±0.8	42.9±0.4	*44.0±1.0	34.0±1.6	**34.2±5.5**	33.3±3.4
FR	45.9±1.2	*48.6±0.8	45.7±1.9	42.5±0.5	41.9±1.8	**42.9±1.1**	**28.9±2.5**	28.3±6.0	24.2±4.3
DE	52.1±1.9	52.3±0.9	*53.0±0.7	43.2±1.5	**45.6±2.0**	44.0±1.9	**29.1±2.5**	27.4±4.2	26.7±6.9
IT	55.1±2.5	*58.7±0.5	58.1±1.6	53.4±2.5	52.9±0.9	**54.0±1.5**	**33.6±3.0**	33.0±7.5	32.6±3.0
PL	40.4±3.2	*42.1±0.6	41.9±2.7	**37.8±1.7**	36.9±1.6	37.5±1.2	**23.4±2.2**	22.3±1.9	19.6±5.2
RU	40.9±1.4	*42.3±0.3	39.2±2.6	36.2±1.3	35.4±0.8	**37.1±0.4**	22.9±2.5	**25.5±4.5**	17.9±3.7
Average	44.9±7.7	*46.9±7.7	**45.9±8.1**	42.5±6.0	42.6±6.3	**43.5±6.4**	**28.6±4.8**	28.4±4.5	25.7±6.4
ES	37.7±2.3	*39.0±1.6	36.7±1.3	**38.3±0.9**	35.3±1.5	37.8±0.8	**26.1±1.2**	26.0±4.7	23.0±5.1
EL	26.7±0.9	*25.5±0.6	25.4±1.5	**22.6±0.1**	21.8±0.9	21.9±0.9	16.2±1.9	**16.7±0.9**	12.5±1.2
KA	*42.6±1.0	40.4±3.1	42.2±2.4	**36.8±3.1**	35.4±1.6	36.7±1.8	**29.4±4.2**	24.2±6.1	22.4±8.7
Average	*35.7±8.1	34.9±8.2	34.7±8.5	**32.6±8.7**	30.8±7.8	32.1±8.9	**23.9±6.9**	22.3±4.9	19.3±5.9
All	41.8±8.6	*42.9±9.5	42.2±9.5	39.2±8.1	38.7±8.7	**39.7±8.8**	**27.1±5.6**	26.4±5.3	23.6±6.7

Best scores by language are marked with an asterisk (*). Best scores by training method for each training scenario are in **bold**.

https://doi.org/10.1371/journal.pone.0301738.t009

(2) **The diversity of languages in the training set has an inconsistent effect on the performance of the LoRA training technique across the three sub-tasks**. While we observe an average decrease in classification performance in the 'English + Translations' training scenario for sub-tasks 2 and 3, this setting improves the average results for sub-task 1. For sub-task 3, this effect holds for every language in the test set, while for sub-task 2, LoRA improves performance on EN and FR when trained using monolingual English data. For sub-task 1, LoRA benefits from the multilingual training data when making predictions on 5 out of the 9 languages (DE, IT, PL, RU, and KA) and favours the monolingual English training setting for the 4 remaining languages (EN, FR, EL and ES). The difference is particularly high for Spanish, resulting in a slightly better average performance in a monolingual training scenario for sub-task 1.

(3) **The diversity of languages in the training set has an inconsistent effect on adapter classification performance across all tasks**. While the 'English + Translations' training scenario insignificantly improves average performance in sub-task 1, it decreases the average performance in sub-tasks 2 and 3. Adapters benefit from a monolingual training scenario when making predictions on 6 out of the 9 languages (EN, FR, IT, RU, ES and EL) in sub-task 1. For sub-tasks 2 and 3, this is the case only for 3 languages (IT, EL and KA) and 2 languages (EN and ES) respectively.

(4) **Decrease in the size of the training set decreases the performance of the FFT, LoRA as well as adapter method on all seen and zero-shot languages**. We observe a significant decrease in FFT's performance and both PEFTs in the 'English Only' training scenario compared to the 'English + Translation' setting across all the sub-tasks and for every language within each sub-task. This result is particularly noteworthy for the English test set as it indicates that even potentially noisy translated data is able to improve the performance on a given language as compared to using a smaller but better quality dataset.

In summary, the effect of removing language diversity from the training set is not consistent across SemEval 2023 sub-tasks and training techniques. Sub-task 1, in particular, demonstrates slight improvement in performance for LoRA and adapter methods when trained on 'English + Translations' data. All training techniques show a significantly decreased performance when trained on the original English-only data, demonstrating the importance of the size of the training data.

In the next section, we compare the results of the three training techniques within each of the three training scenarios.

Comparison of training techniques for each training scenario and language

To answer the third research question (RQ3), we analyse whether the performance of FFT, LoRA and adapter methods is consistent across training scenarios or are certain scenarios more prone to the lack of training data and or language diversity within the training data. This analysis, on the top of results presented in Tables 7–9 for sub-tasks 1, 2 and 3 respectively, complement the previous RQ2.

(1) **FFT outperforms LoRA and adapter methods for sub-tasks 1 and 2 in a 'Multilingual Joint' training scenario, while for sub-task 3, only zero-shot predictions on unseen languages benefit from the FFT method**. We observe that FFT yields best performance in a multilingual setting for most of the seen and unseen languages in sub-tasks 1 and 2. While the differences between FFT and LoRA are often less than 1% for sub-task 1, sub-task 2 demonstrates a clear preference for the FFT classification approach.

(2) **In an 'English + Translations' training scenario, adapters outperform FFT across all sub-tasks and show better or on-par performance compared to LoRA**. A particularly clear

preference for adapters can be observed for sub-task 2, where the majority of seen and unseen languages benefit from this training technique. For sub-tasks 1 and 2, most of the zero-shot predictions on unseen languages favour FFT. The performance of the adapter method is particularly consistent on English, the only seen language in this training scenario, with all sub-tasks favouring the adapter classifier.

(3) **In an 'English Only' training scenario with less data, differences between FFT, LoRA and adapters become less obvious across all sub-tasks**. In this setting, the differences in the average classification performances between FFT, adapters and LoRA across all sub-tasks is often insignificant, with less than 1% of one method over the other. For example, the difference between the adapter method and LoRA for sub-task 1 is 0.2%, and the average performance of FFT and LoRA is the same and differs only in the confidence intervals. Similarly, for sub-task 2, FFT is comparable to the adapter method, and for sub-task 3, FFT is very close in the performance to LoRA. We observe that for sub-tasks 1 and 2, adapters perform better for EN (the only seen language), while for sub-task 3, LoRA yields better performance for the seen language (EN).

(4) **The overall best performance across all training scenarios and training techniques is achieved in a 'Multilingual Joint' training scenario**. While the FFT method works better for sub-tasks 1 and 2 in this setting, sub-task 3 shows a clear improvement when trained using the LoRA method. While sub-tasks 1 and 2 are consistent in favouring FFT for both seen and unseen (ES, EL, KA) languages, sub-task 3 favours LoRA when making the predictions on seen languages only. For unseen languages, sub-task 3 agrees with sub-tasks 1 and 2 in favouring the FFT classification approach.

Some of the languages demonstrate strong preferences towards certain training techniques and training scenarios:

(1) **English consistently favours adapter classification approach across all the training scenarios in sub-task 1**.

(2) **FFT method yields best performance on Georgian zero-shot predictions across all settings in sub-tasks 1 and 3**.

(3) **In a 'Multilingual Joint' training scenario, German demonstrates a consistent preference for the adapter training technique across all three sub-tasks**. In particular, German is the only language in sub-task 2 where FFT is not producing the best performance for the 'Multilingual Joint' training scenario. It is also the only case where adapter models demonstrate better performance than LoRA for seen languages for sub-task 3 in the joint training scenario.

We also observe that LoRA, in a 'Multilingual Joint' training scenario, shows the best overall performance for sub-task 3. Since this method was not used by any of the teams participating in the shared task, we want to look into how this approach compares to the official leaderboard results after the competition. Table 10 demonstrates the scores of the winning system for sub-task 3 along with the scores achieved in sub-task 3 in these experiments. As can be seen, the LoRA method for sub-task 3 outperforms most of the results of the winning systems. We achieve an increase in the performance of up to 19.63% for all the languages except for Georgian (KA). 6 out of 9 languages (FR, IT, PL, RU, ES and EL) achieve the best result in the 'Multilingual Joint' training scenario when applying LoRA. Not surprisingly, the best score for English is achieved in one-to-one 'English only' scenario. This increase also results in the first placings in 8 out of 9 languages.

To summarise this analysis, we observe that adapter models and LoRA produce comparable or even better results for sub-task 3 or in the scenarios when the training data is scarce or is only available in one language. This is a promising result given that these approaches require less memory and training time, as discussed in the previous section.

Table 10. Sub-task 3 official test set leaderboard comparison.

Language	1st Place Team	$F1_{micro}$ (1st Place)	$F1_{micro}$ (Ours)	$F1_{micro}$ Increase	Final Placing (Ours)
EN	APatt	0.37562	0.44937	19.63%	1
FR	NAP	0.46869	0.49238	5.05%	1
DE	KInITVeraAI	0.51304	0.54174	5.30%	1
IT	KInITVeraAI	0.55019	0.59919	8.91%	1
PL	KInITVeraAI	0.43037	0.44964	4.48%	1
RU	KInITVeraAI	0.38682	0.42635	10.22%	1
ES	TeamAmpa	0.38106	0.40674	6.74%	1
EL	KInITVeraAI	0.26733	0.27668	3.38%	1
KA	KInITVeraAI	0.45714	0.448	-2.00%	2

https://doi.org/10.1371/journal.pone.0301738.t010

Limitations and discussion

The results reported in this study are limited to using one dataset annotated with different classification tasks. Further analysis would be beneficial to test these findings on a diverse set of corpora and tasks. Additionally, we examined the effect of one adapter technique, which prevents us from generalising our findings to all types of adapter methods. Finally, since sub-tasks 1 and 2 are performed at article-level where the average length of the article is more than 512 tokens (Table 1), many of the signals present in the rest of the articles could potentially be ignored by all of the methods compared in this study.

Our findings provide novel insights into the effectiveness of PEFT techniques in multilingual classification tasks. In particular, we found that the multilingual training scenario on average improves the performance on all sub-tasks. However, the best performing methods within this setting differed depending on the sub-task.

Our results on the LoRA method are novel as this PEFT technique was not previously investigated in multilingual tasks. While we observe that sub-tasks 1 and 2 benefit from FFT method in a 'Multilingual Joint' training scenario, our results demonstrate a particularly interesting behaviour of the LoRA method in this training scenario for sub-task 3, where it consistently outperforms the FFT approach. While a thorough analysis is needed to conclude why this is the case, the differences in the properties of these tasks that we presented in Table 1 could potentially bring light to this question. Firstly, sub-task 3 is trained on and applied to much shorter texts than sub-tasks 1 and 2, while having substantially more training examples. Additionally, sub-task 3 is characterised by a high number of classes (23 compared to 14 in sub-task 2 and 3 in sub-task 1) and a much more severe data imbalance, as shown in Table 1. The nature of the tasks is also quite different. The task of framing detection relies less on commonsense knowledge and pragmatics, while sub-tasks 1 and 3 are more subjective for human experts. All of these factors can potentially influence the performance of each method in the respective sub-tasks, and further research is needed to make conclusions regarding the types of classification tasks that benefit from from LoRA method.

We observe the adapter method to outperform FFT only in the zero-shot cross-lingual scenario when a sufficient amount of data is provided in the source language ('English + Translations' scenario) while being much less efficient when the training data contains a variety of languages or is limited and monolingual. This is an important insight that complements previous studies which report the adapter approach to always outperform FFT in a cross-lingual zero-shot inference [11, 13]. Our results also challenge a prior conclusion by Xenouleas et al. [14] which suggests that adapter methods are universally beneficial for low-resource and cross-lingual tasks. Additionally, our results disagree with the prior reports that the FFT

approach is always outperformed by the adapter method in a zero-shot cross-lingual inference [11, 13]. Contrary to this, we found that for the 'unseen' languages consistent across all training scenarios ('ES', 'EL' and 'KA'), FFT outperforms adapter method in all training scenarios for sub-task 1 and sub-task 3, and for two our three scenarios in sub-task 2, 'Multilingual Joint' and 'English Only'. Finally, our study provides the first to our knowledge comparison of the 'Multilingual Joint' scenario with the translation-based monolingual one in terms of the zero-shot cross-lingual inference.

While sub-task 3 performs better for the LoRA approach in a multilingual setting on average, this is not consistent across seen and zero-shot languages, with FFT performing better in a zero-shot cross-lingual setting. This preference is particularly strong in the cases when the amount of training data is limited. This can indicate the fact that the adapter method is potentially not good at cross-lingual generalisation, however, more analysis is needed on a wider variety of tasks.

In the tasks with highly skewed data, adapter method can potentially be more prone than FFT to favoring the most frequent class, since as can be observed from Table 6, $F1_{micro}$ score for sub-task 1 is significantly higher for adapter method compared to FFT, while FFT results in a higher $F1_{macro}$. However, a detailed error analysis to confirm or refute this assumption is currently not possible since the gold-standard labels for the test set are not released.

Low-resource languages may have a preference for FFT in all training scenarios when predictions are made in a cross-lingual zero-shot way. This assumption is suggested by the fact that Georgian is the only language that consistently prefers the FFT approach across all training scenarios and for all sub-tasks.

Interestingly, the performance on Georgian in sub-tasks 1 and 2 is higher than that on seen languages, despite being low-resource and zero-shot. One potential reason for this observation could be the fact that, as was previously shown in Tables 2 and 3, the texts in the test set for Georgian are, on average, within the limit of XLM-R and are much shorter than the input lengths for other 8 languages. This could explain why the same effect is not observed for sub-task 3, where all the inputs are within the transformer token limit and are relatively short. However, with the lack of the gold standard labels for surprise languages, it is not possible to eliminate other reasons for this phenomenon, as it could also be explained with the lack of particularly difficult to predict classes in the test set for Georgian for both sub-tasks.

Conclusion

In this work, we performed the first (to our knowledge) analysis of the performance of Low-Rank Adaptation (LoRA) technique and its comparison with the adapter and full fine-tuning (FFT) methods in a multilingual multiclass scenario.

We found that parameter-efficient fine-tuning techniques (PEFTs), LoRA and bottleneck adapter, provide significant computation efficiency compared to FFT in terms of the training time, the number of trainable parameters and the amount of VRAM memory required. In particular, they reduce the number of trainable parameters between 140 and 280 times and achieve between 32% and 44% shorter training time.

The comparison between LoRA and adapter methods in terms of the parameter efficiency suggests that their performance depends on a certain sub-task and hyperparameters used. This observation is aligned with the results of the previous study by He et al. [11], who found the benefit of the adapter approach to be task-dependent. While we observe LoRA to be more efficient than the adapter method for tasks of news articles' genre and framing detection, the adapter method takes less average time and uses less training parameters for the latter one.

Moreover, we found the performance of the methods to be highly dependent on the training scenario. Adapter method performs better than LoRA and FFT in the scenario where there is a lack of language diversity in the training set across the sub-tasks.

The differences between all three methods become insignificant, often less than 1% on average, as the size of the training data decreases. This indicates that it is possible to achieve high computational efficiency by using PEFT methods without losing much in terms of the classification performance in this setting. More experiments on this result involving a gradual decrease in the size of the training set would be beneficial in future to find the threshold when the performances across the methods match or when PEFT methods become more classification-efficient.

The performance on the unseen languages is often highly dependent on the training scenario. We found that FFT performs better than PEFT methods in zero-shot cross-lingual predictions when trained on a joint multilingual dataset, which is different from the results reported by Chalkidis et al. [13]. However, we observe the effect reported by the authors in a monolingual training scenario, where adapter method performs better on zero-shot languages.

Finally, the multilingual joint LoRA setting allowed us to significantly improve our official results on the SemEval 2023 sub-task 3 (persuasion techniques detection) and to outperform most of the official leaderboard-best results, placing first in all languages except Georgian, where we are in the second place compared to the official leaderboard results.

Supporting information

S1 Appendix. Complete list of categories and their hierarchy for sub-task 3.
(PDF)

S2 Appendix. Ablation study for PEFT techniques.
(PDF)

S3 Appendix. Model size selection for XLM-RoBERTa.
(PDF)

S4 Appendix. Amount of resources per-language.
(PDF)

S5 Appendix. Comparison of Houlsby and Pfeiffer adapters across three sub-tasks.
(PDF)

Acknowledgments

We thank Joanna Wright for providing her comments and suggestions on the draft of this work.

Author Contributions

Conceptualization: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell, Ivan Srba.

Formal analysis: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell.

Funding acquisition: Carolina Scarton, Kalina Bontcheva.

Investigation: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell.

Methodology: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell, Ivan Srba.

Project administration: Carolina Scarton, Kalina Bontcheva, Xingyi Song.

Supervision: Carolina Scarton, Kalina Bontcheva, Xingyi Song.

Validation: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell.

Writing – original draft: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell, Ivan Srba.

Writing – review & editing: Olesya Razuvayevskaya, Ben Wu, João A. Leite, Freddy Heppell, Ivan Srba, Carolina Scarton, Kalina Bontcheva.

References

1. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: North American Chapter of the Association for Computational Linguistics; 2019. Available from: https://api.semanticscholar.org/CorpusID:52967399.
2. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research. 2020; 21 (1):5485–5551.
3. Ben Zaken E, Goldberg Y, Ravfogel S. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics; 2022. p. 1–9. Available from: https://aclanthology.org/2022.acl-short.1.
4. Jiang H, He P, Chen W, Liu X, Gao J, Zhao T. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 2177–2190. Available from: https://aclanthology.org/2020.acl-main.197.
5. Xu R, Luo F, Zhang Z, Tan C, Chang B, Huang S, et al. Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning. In: Moens MF, Huang X, Specia L, Yih SWt, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 9514–9528. Available from: https://aclanthology.org/2021.emnlp-main.749.
6. Lialin V, Deshpande V, Rumshisky A. Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv:230315647. 2023;.
7. Li XL, Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 4582–4597. Available from: https://aclanthology.org/2021.acl-long.353.
8. Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning. In: Moens MF, Huang X, Specia L, Yih SWt, editors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 3045–3059. Available from: https://aclanthology.org/2021.emnlp-main.243.
9. Bapna A, Firat O. Simple, Scalable Adaptation for Neural Machine Translation. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 1538–1548. Available from: https://aclanthology.org/D19-1165.
10. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, et al. Parameter-Efficient Transfer Learning for NLP. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. PMLR; 2019. p. 2790–2799. Available from: https://proceedings.mlr.press/v97/houlsby19a.html.
11. He R, Liu L, Ye H, Tan Q, Ding B, Cheng L, et al. On the Effectiveness of Adapter-based Tuning for Pre-trained Language Model Adaptation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 2208–2222. Available from: https://aclanthology.org/2021.acl-long.172.

12. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations; 2022. Available from: https://openreview.net/forum?id=nZeVKeeFYf9.

13. Chalkidis I, Fergadiotis M, Androutsopoulos I. MultiEURLEX—A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 6974–6996. Available from: https://aclanthology.org/2021.emnlp-main.559.

14. Xenouleas S, Tsoukara A, Panagiotakis G, Chalkidis I, Androutsopoulos I. Realistic Zero-Shot Cross-Lingual Transfer in Legal Topic Classification. In: Proceedings of the 12th Hellenic Conference on Artificial Intelligence. SETN'22. New York, NY, USA: Association for Computing Machinery; 2022. Available from: https://doi.org/10.1145/3549737.3549760.

15. Piskorski J, Stefanovitch N, Da San Martino G, Nakov P. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). Toronto, Canada: Association for Computational Linguistics; 2023. p. 2343–2361. Available from: https://aclanthology.org/2023.semeval-1.317.

16. Wu B, Razuvayevskaya O, Heppell F, Leite JA, Scarton C, Bontcheva K, et al. SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). Toronto, Canada: Association for Computational Linguistics; 2023. p. 1995–2008. Available from: https://aclanthology.org/2023.semeval-1.275.

17. Hromadka T, Smolen T, Remis T, Pecher B, Srba I. KInITVeraAI at SemEval-2023 Task 3: Simple yet Powerful Multilingual Fine-Tuning for Persuasion Techniques Detection. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). Toronto, Canada: Association for Computational Linguistics; 2023. p. 629–637. Available from: https://aclanthology.org/2023.semeval-1.86.

18. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 8342–8360. Available from: https://aclanthology.org/2020.acl-main.740.

19. Pfeiffer J, Vulić I, Gurevych I, Ruder S. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 7654–7673. Available from: https://aclanthology.org/2020.emnlp-main.617.

20. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized llms. arXiv preprint arXiv:230514314. 2023;.

21. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 353–355. Available from: https://aclanthology.org/W18-5446.

22. Yu X, Chatterjee T, Asai A, Hu J, Choi E. Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 3725–3743. Available from: https://aclanthology.org/2022.findings-emnlp.273.

23. Kiesel J, Mestre M, Shukla R, Vincent E, Adineh P, Corney D, et al. SemEval-2019 Task 4: Hyperpartisan News Detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 829–839. Available from: https://aclanthology.org/S19-2145.

24. Abu Farha I, Oprea SV, Wilson S, Magdy W. SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic. In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Seattle, United States: Association for Computational Linguistics; 2022. p. 802–814. Available from: https://aclanthology.org/2022.semeval-1.111.

25. Da San Martino G, Barrón-Cedeño A, Wachsmuth H, Petrov R, Nakov P. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. Barcelona (online): International Committee for Computational Linguistics; 2020. p. 1377–1414. Available from: https://aclanthology.org/2020.semeval-1.186.

26. Dimitrov D, Bin Ali B, Shaar S, Alam F, Silvestri F, Firooz H, et al. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In: Proceedings of the 15th International Workshop on

Semantic Evaluation (SemEval-2021). Online: Association for Computational Linguistics; 2021. p. 70–98. Available from: https://aclanthology.org/2021.semeval-1.7.

27. Card D, Boydstun AE, Gross JH, Resnik P, Smith NA. The Media Frames Corpus: Annotations of Frames Across Issues. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China: Association for Computational Linguistics; 2015. p. 438–444. Available from: https://aclanthology.org/P15-2072.
28. Da San Martino G, Yu S, Barrón-Cedeño A, Petrov R, Nakov P. Fine-Grained Analysis of Propaganda in News Article. 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). Hong Kong, China: Association for Computational Linguistics; 2019. p. 5636–5646. Available from: https://aclanthology.org/D19-1565.
29. Billert F, Conrad S. HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). Toronto, Canada: Association for Computational Linguistics; 2023. p. 1166–1171. Available from: https://aclanthology.org/2023.semeval-1.162.
30. Falk N, Eichel A, Piccirilli P. NAP at SemEval-2023 Task 3: Is Less Really More? (Back-)Translation as Data Augmentation Strategies for Detecting Persuasion Techniques. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). Toronto, Canada: Association for Computational Linguistics; 2023. p. 1433–1446. Available from: https://aclanthology.org/2023.semeval-1.198.
31. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Computing Research Repository. 2019;arXiv:1907.11692.
32. Piskorski J, Stefanovitch N, Bausier VA, Faggiani N, Linge J, Kharazi S, et al. News Categorization, Framing and Persuasion Techniques: Annotation Guidelines. European Commission, Ispra, JRC132862; 2023.
33. Barbaresi A. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Association for Computational Linguistics; 2021. p. 122–131. Available from: https://aclanthology.org/2021.acl-demo.15.
34. Pikuliak M, Šimko M, Bieliková M. Cross-lingual learning for text processing: A survey. Expert Systems with Applications. 2021; 165:113765. https://doi.org/10.1016/j.eswa.2020.113765
35. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 8440–8451. Available from: https://aclanthology.org/2020.acl-main.747.
36. Pfeiffer J, Rücklé A, Poth C, Kamath A, Vulić I, Ruder S, et al. AdapterHub: A Framework for Adapting Transformers. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 46–54. Available from: https://aclanthology.org/2020.emnlp-demos.7.
37. Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4996–5001. Available from: https://aclanthology.org/P19-1493.
38. Savenkov K, Lopez M. The State of the Machine Translation 2022. In: Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track). Orlando, USA: Association for Machine Translation in the Americas; 2022. p. 32–49. Available from: https://aclanthology.org/2022.amta-upg.4.
39. Wang Z, Lipton ZC, Tsvetkov Y. On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 4438–4450. Available from: https://aclanthology.org/2020.emnlp-main.359.