SheffieldVeraAI at SemEval-2023 Task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification

Ben Wu[#], **Olesya Razuvayevskaya**[#], **Freddy Heppell**[#], **João A. Leite**[#],
Carolina Scarton, **Kalina Bontcheva**, and **Xingyi Song**
Department of Computer Science, The University of Sheffield, Sheffield, UK
`{bpwu1, o.razuvayevskaya, frheppell1, jaleite1}@sheffield.ac.uk`

Abstract

This paper describes our approach for SemEval-2023 Task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. For Subtask 1 (News Genre), we propose an ensemble of fully trained and adapter mBERT models which was ranked joint-first for German, and had the highest mean rank of multi-language teams. For Subtask 2 (Framing), we achieved first place in 3 languages, and the best average rank across all the languages, by using two separate ensembles: a monolingual RoBERTa-MUPPET$_{\text{LARGE}}$ and an ensemble of XLM-RoBERTa$_{\text{LARGE}}$ with adapters and task adaptive pretraining. For Subtask 3 (Persuasion Techniques), we trained a monolingual RoBERTa-Base model for English and a multilingual mBERT model for the remaining languages, which achieved top 10 for all languages, including 2nd for English. For each subtask, we compared monolingual and multilingual approaches, and considered class imbalance techniques. [1]

1 Introduction

With the rise of opinion-manipulating news and misinformation surrounding COVID-19, elections and wars, the task of propaganda and hyperpartisan detection has received much attention over the last five years. Since 2019, various SemEval tasks have addressed detecting hyperpartisan (Kiesel et al., 2019), sarcasm (Abu Farha et al., 2022), and persuasion techniques in textual and multimodal data (Da San Martino et al., 2020; Dimitrov et al., 2021). This task (Piskorski et al., 2023) can be seen as an extension of the latter two tasks, suggesting an expanded ontology of persuasion techniques and addressing other related aspects of persuasion, such as satire, opinionated news, and framing detection.

The three subtasks presented in this shared task are the detection of: 1) genre: opinion, objective reporting or satire; 2) framing techniques: 14 multilabel frames; 3) persuasion techniques: 23 multilabel techniques, which can be grouped into 6 high-level classes.

The data consists of labelled training and development sets in English, French, German, Italian, Polish and Russian, and unlabelled test sets in the same languages plus three zero-shot languages: Spanish, Greek and Georgian.

The main contributions of this paper are twofold: 1) evaluation of the viability of monolingual versus multilingual models for each of the subtasks; and 2) presentation of the models which ranked first in four subtask-language pairs, top three in 16 subtask-language pairs, and were within the top 10 for all.

Our approaches for the three subtasks differ, therefore we present each subtask separately in sections 3-5 respectively. An overview of the techniques used in each subtask is shown in Table 1.

2 Background

Fine-grained propaganda technique classification was first introduced by Da San Martino et al. (2019), who suggested a multi-granularity network, where the lower and higher granularity tasks refer to the fragment and sentence-level classification respectively. Other state-of-the-art approaches to this task used an ensemble of RoBERTa models with class weighting, where some models perform a semi-supervised task of span detection (Jurkiewicz et al., 2020) and an ensemble of 5 different transformer models (Tian et al., 2021), namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), XLNet (Yang et al., 2019), DeBERTa (He et al., 2020) and ALBERT (Lan et al., 2019).

Framing detection specifically has been addressed primarily for political news, with the models exploring unsupervised probabilistic topic models combined with autoregressive distributed-lag

[#]Equal contribution, listed randomly.
[1]Our code is available at `https://github.com/GateNLP/semeval2023-multilingual-news-detection`

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1995–2008
July 13-14, 2023 ©2023 Association for Computational Linguistics

Tasks	Text Clean	External Data	Oversampling	Class Weights	Adapters	TAPT	Unseen languages	Ensemble
Subtask1	✔⋆	✔	✔⋆	-	✔†	-	Zero-shot	✔
Subtask2	✔	-	-	✔	✔	✔	⇒ EN	✔
Subtask3	-	-	-	✔	-	-	⇒ EN	-

Table 1: An overview of approaches used. TAPT: Task-adaptive Pre-training, ⋆ not used in adapter model for submission. † as part of ensemble.

models (Tsur et al., 2015), finetuning BERT (Liu et al., 2019a) and multilingual BERT (mBERT) (Akyürek et al., 2020). The latter system is the closest to our task since it explores the multilabel multilingual setting and the effect of translating texts for use in monolingual models. However, it uses article headlines instead of the full texts as the classification data. The authors found that English $\text{BERT}_{\text{BASE}}$ uncased trained on translated data and tested on the data in the target language often outperforms the multilingual model. We perform similar comparison experiments for all three subtasks of this shared task.

Wang and Banko (2021) performed a series of experiments comparing monolingual and multilingual approaches for hate speech detection and sentiment analysis and found that different task-language combinations favour either monolingual and multilingual settings. The authors also concluded that data augmentation in the form of translation and task-adaptive pretraining (TAPT) (Gururangan et al., 2020) helps to further improve the results.

Another important task addressed in fake news detection is satire detection, with the methods ranging from convolutional neural networks (CNNs) (Guibon et al., 2019) to adversarial training (McHardy et al., 2019) and BERT-based architectures with long-short-term memory (LSTM) (Pandey and Singh, 2022; Liu and Xie, 2021) and CNN (Kaliyar et al., 2021) layers on top.

Bottleneck Adapters Adapters (Houlsby et al., 2019; Bapna and Firat, 2019) represent a family of techniques aimed at improving parameter efficiency in finetuning by freezing a pretrained model and inserting low-dimension adapter modules within each layer. Houlsby et al. found that, despite training only 3.6% of the parameters compared to a full model, performance only decreased by 0.4%, while Bapna and Firat found that adapters produced comparable, or in some cases better, results. Particularly relevant for our task is He et al. (2021)'s finding that adapter-based tuning of LLMs is particularly effective for low-resource and cross-lingual tasks. For our system, we used two different configurations of the bottleneck adapter modules: 1) the original Houlsby et al. bottleneck configuration, which places adapter modules after the multi-head attention block and feed-forward block of each transformer layer; 2) the Pfeiffer et al. configuration, which places adapter modules only after the feed-forward block of each layer.

Chalkidis et al. (2021) also found that for XLM-R on the MultiEURLEX dataset, training bottleneck adapters outperforms traditional full finetuning and improves zero-shot cross lingual capability. Similarly to this task, the MultiEURLEX dataset was used for multilingual multilabel classification, though it is significantly larger than the data in this task (covering 23 languages and classifying hierarchically from 21 to 567 labels).

3 System Description for Subtask 1

3.1 System Overview

The system consisted of an ensemble of four models, comprising 1) three mBERT models each finetuned using the organiser training set and 2/3 of the development set; 2) one frozen mBERT model with a finetuned Houlsby adapter[2]. The ensemble predictions were decided by majority vote, with rare tie cases handled by selecting the model with the best validation performance.

Full Finetuning $\text{mBERT}_{\text{BASE}}$ (Devlin et al., 2019) was finetuned on a shuffled combination of all languages. Different to previous approaches (Wu and Dredze, 2020; Adebara et al., 2020), we chose the epoch with the best validation performance per-language, instead of overall, since the best overall epoch is not necessarily the best for a given language[3]. By using three identically-configured models in the ensemble, the data sacrificed for model selection can be rotated between them, so overall no data is truly unseen.

Adapter Model A Houlsby bottleneck adapter was applied to a pretrained $\text{mBERT}_{\text{BASE}}$ model, with a reduction factor of 8 (i.e. $d = 96$), using the AdapterHub (Pfeiffer et al., 2020a) framework.

[2]Using a data split described in section 3.2

[3]See Appendix C.1

The mBERT model parameters were frozen, so only the adapter and classification head parameters were trained.

Data Preprocessing Since task data is obtained from webpages, it often contains unwanted content, such as hyperlinks, account handles, dates and author biographies. We applied the preprocessing described in Appendix B to remove this content.

Long Article Truncation The organiser annotation instructions indicate that even human annotators find it difficult to distinguish *opinion* from *reporting* and *satire*. This is due to subtle differences in how opinionated direct speech could be balanced or reported on. The instructions also mention that single opinionated sentences, which would trigger the *opinion* genre, often appear at the end of an article. Given the limit on input length for BERT models, for the articles that were longer than 512 tokens, we sequentially selected sentences from the beginning and the end of the article, preserving the original sentence order, until the length of 512 tokens was reached.

External Satire Due to the lack of satire data, our training set was supplemented with 203 English-language satire articles from Golbeck et al. (2018).

Data Oversampling The training data is severely imbalanced, with less than 6% of articles annotated as *satire* and 18% of articles annotated as *reporting*. Although the external satire data improved the balance for English, the performance of the satire class in other languages still remained inadequate. To address this, we performed oversampling for both *satire* and *opinion* classes by repeating random oversampling without replacement on the original data for a given language and class until the classes were balanced. For English *satire* class, we applied the same approach but oversampled the external satire data mentioned earlier rather than the original training set. We also compared the effectiveness of the oversampling approach with the class weighting approach in our experiment, and the results showed a slight advantage for the oversampling approach on average.

3.2 Experimental Setup

For the final submission, mBERT transformer models were finetuned on the organiser training set and a part of the development set for 30 epochs with the learning rate of 1e-5, AdamW optimiser (ε=1e-8) and ReLU activation function. The organiser development set was split into three parts, stratified by label and language. We then finetuned three models, using $\frac{1}{3}$ of the development set as a test and merging $\frac{2}{3}$ of the development set with the training data and shuffling the dataset. The held-out part of the development set was used to identify the language-specific best checkpoints for each model. We utilised the checkpoint with the best overall $F1_{macro}$ on the held-out set to make predictions on the surprise languages. As described in section 3.1, all articles were preprocessed and the training data was oversampled for *satire* and *reporting* classes of each language.

Language	$F1_{macro}$	Place	Language	$F1_{macro}$	Place
English	61.282	3	Italian	72.040	3
French	68.157	5	Polish	76.455	3
German	81.951	*1	Russian	72.871	2
Spanish	44.293	4	Greek	68.681	6
Georgian	96.268	2			

Table 2: Subtask 1 final leaderboard results. * joint.

Adapter models were trained on the combined organiser and development set (resplit 80% train, 10% validation, 10% held-out test, stratified by label), for 20 epochs with the learning rate of 1e-4, AdamW optimiser (ε=1e-8) and Tanh activation function in the classifier, and selected the checkpoint with the highest overall validation $F1_{macro}$ score. The above preprocessing and oversampling were not used, and articles were truncated at 512 tokens.

After submission, we conducted additional experiments using the organiser training and development sets for consistency. For monolingual models, all articles in the training set and the external satire were translated with Google Translate into the language of each monolingual model in question. Due to the character length limitation, particularly long articles were translated sentence-by-sentence.

3.2.1 Results and Reflections

The final submission results of the ensemble are listed in Table 2.

The final ensemble results achieved a higher $F1_{macro}$ score than in the supplementary multilingual results in all languages, except Polish. In English, the ensemble achieved $F1_{macro}$ score 25% higher than the single mBERT transformer or adapter model. It should be noted that the final models were trained on both training and development data. However, since the development sets are only $\approx \frac{1}{3}$ the size of the training sets, the difference in the amount of the training data was not dramatic.

In the absence of gold standard labels for the test set, it is difficult to analyse why the model achieved a high score in Georgian, despite being zero-shot. However, our ensemble predictions suggest that there is likely to be no satire articles in the Georgian test set, which was consistently the most difficult class to detect.

Table 3 shows the differences between monolingual and multilingual versions of adapter bottleneck and transformer models, evaluated against the organiser development set. The multilingual transformer models always perform better than monolingual ones, while for 4 out of 6 languages, adapter bottleneck models benefit from the monolingual setup. This may be due to using a fixed reduction factor across all languages. Interestingly, the mBERT model demonstrates the best average result in English for both transformer and bottleneck adapter models. For Italian, XLM-R yields the best results for both transformer and adapter bottleneck models. It is also notable that the results for English are by far the worst across all the models, possibly because the models are overly focused on capturing semantic meaning and are not as effective in genre classification.

Even though transformer XLM-R demonstrated significantly better results than transformer mBERT for Italian and German, these differences were only marginal in our main setting where the validation set was smaller, while the marginally better results for Russian were not observed at all. Given the above observations and the fact that XLM-R yielded higher $F1_{\text{macro}}$ fluctuations, sometimes reaching 10%, we opted for the mBERT model as our main submission.

3.3 Post-competition Findings

Since, in our final submission, all languages were evaluated without translation (including the three surprise languages), a natural question we wanted to explore after the competition was whether translating texts into a different language for evaluation (the 'translate-test' approach) would have yielded better results.

We selected the checkpoints that, during training, achieved the highest validation $F1_{\text{macro}}$ for each individual language. These 'language-optimal' checkpoints were then used to evaluate translations of the other test sets. For example, using the 'French'-optimal checkpoint, we translated all tests sets into French and made predictions.

Surprisingly, we found that the translate-test English $\rightarrow$ Russian and Italian $\rightarrow$ French each improved the $F1_{\text{macro}}$ performance by 1% on English and Italian respectively, while French $\rightarrow$ Russian improved by over 6%.

Two out of the three surprise languages, Spanish and Greek, also benefited from being translated into other languages and tested using the corresponding best checkpoints. The Spanish $\rightarrow$ English setting showed particularly striking increase in $F1_{\text{macro}}$, from 68.7 to 81.7, which is also 21% higher than the score of the winning team for Spanish and is 20% above the other post-competition results. Except for German, translating the Greek test set into the other 5 main languages and testing using the corresponding checkpoints also provided significant improvements in the range of 5%-13%, which is over 1% above the result of the winning system and is the current leaderboard-best.

4 System Description for Subtask 2

4.1 System Overview

Two systems were used for submission, depending on the language. For English and the three surprise languages, we used **a monolingual English ensemble** of 3 RoBERTa-MUPPET$_{\text{LARGE}}$ models. For the remaining languages (French, German, Italian, Russian, Polish), we used **a multilingual ensemble** of 3 XLM-R$_{\text{LARGE}}$ models (with adapters and task-adaptive pre-training (TAPT)).

An overview of the two models is shown in Table 4. A key difference between these two systems is that the monolingual MUPPET models were trained using traditional finetuning of all parameters, whereas the XLM-R models 1) underwent task-adaptive pre-training; and 2) were finetuned using Pfeiffer bottleneck adapters.

For both systems, we trained our models jointly on articles in all languages (using English translations for our monolingual model). This meant that we produced a single monolingual or multilingual system that was able to make predictions for all languages. We chose this approach of joint training across all languages in order to maximise the number of examples seen for each class, since the dataset for Subtask 2 is quite small, particularly when split by language. Our early experiments showed that this approach was superior to training

Language	Transformer			Adapter		
	Monolingual	mBERT	XLM-R	Monolingual	mBERT	XLM-R
English	30.0 ± 5.6	*$\mathbf{36.2 \pm 2.5}$	36.1 ± 2.1	20.5 ± 3.3	*21.4 ± 6.0	20.0 ± 3.1
French	51.2 ± 3.3	62.5 ± 4.6	*65.5 ± 4.3	*$\mathbf{68.3 \pm 0.6}$	64.2 ± 0.8	61.3 ± 2.7
German	59.9 ± 4.1	59.9 ± 5.0	*$\mathbf{66.9 \pm 1.0}$	*65.7 ± 3.6	57.8 ± 2.8	62.0 ± 3.2
Italian	56.7 ± 6.5	55.1 ± 4.3	*$\mathbf{72.6 \pm 6.4}$	51.9 ± 4.7	47.8 ± 2.3	*60.3 ± 3.1
Polish	71.7 ± 6.6	*$\mathbf{81.9 \pm 3.7}$	79.4 ± 1.1	*77.6 ± 2.9	72.9 ± 5.1	76.7 ± 2.2
Russian	52.8 ± 8.8	52.9 ± 1.6	*54.7 ± 9.8	*$\mathbf{56.9 \pm 9.5}$	48.3 ± 2.1	48.0 ± 0.8

Table 3: Mean $F1_{macro} \pm 1$ std (over 3 runs) on subtask 1 organiser development set for multilingual and monolingual models for transformer and adapter-only architecture. * denotes the best per model per language and **bold** denotes the overall best per language.

MUPPET ensemble	XLM-R ensemble
MUPPET$_{\text{LARGE}}$	XLM-R$_{\text{LARGE}}$
Monolingual (English)	Multilingual
Trained on all articles (in translation)	Trained on all articles (original)
No TAPT	TAPT
Full finetuning	Adapter finetuning
Ensemble size 3	Ensemble size 3
Submitted for EN, EL, KA, ES	All other languages

Table 4: Summary of the monolingual vs multilingual systems submitted for subtask 2

models on the individual articles in each language.

4.1.1 The English Monolingual System

Full Finetuning Our monolingual system used a finetuned RoBERTa-MUPPET$_{\text{LARGE}}$ (Aghajanyan et al., 2021) ensemble. RoBERTa-MUPPET improves on its baseline RoBERTa counterpart by adding an additional 'pre-finetuning' stage of multitask learning. We opted not to use adapters, because our cross validation experiments showed this worsened performance (see Appendix Table 13).

English Translations Because RoBERTa-MUPPET is a monolingual model, we translated all articles into English for training, and used them for finetuning alongside the original English articles. We performed inference on non-English languages by translating the articles into English: a 'translate-test' approach.

4.1.2 Multilingual System

We used XLM-RoBERTa$_{\text{LARGE}}$ (Conneau et al., 2020) for our multilingual model. Our system uses two techniques to improve performance: TAPT and adapter layers.

Task-adaptive Pre-training We performed task-adaptive pre-training on the entire XLM-R model, following the approach suggested by Gururangan et al. (2020). Masked-language modelling was performed, using all available articles (including the organisers' development and test sets). We trained for 60 epochs with a learning rate of 1e-4 and batch size of 128 (for full hyperparameters, see Appendix Table 11).

TAPT could alternatively be performed by freezing the base model and training the adapters with an MLM objective. Despite being a faster approach, this has been found to sometimes decrease performance (Kim et al., 2021).

Adapters Our multilingual system used the Pfeiffer bottleneck adapter configuration, with a reduction factor of 8, which for XLM-R$_{\text{LARGE}}$ corresponds to a bottleneck hidden size of 128.

Although using adapters did result in slightly improved performance, we found that their main advantage came from their low parameter number, which allowed for faster training and more experimentation.

4.2 Ensemble

Predictions made by our ensembles were decided by a majority vote. Each ensemble consisted of 3 individual models (MUPPET models for monolingual; adapter-finetuned XLM-R + TAPT for multilingual). Within each ensemble, two models were trained with class-weighting, and one-without.

Class Weighting Class weighting helped to account for class imbalance by balancing the impact of under- and over-represented classes. When calculating the loss, the logit for each class was multiplied by a class weight that was inversely proportional to the frequency of that class in the dataset.

Overall $F1_{micro}$ scores were similar for models with or without class weights. Class-weighting did help to improve performance on less frequent frames (such as *Cultural Identity* and *Public Opinion*), but at the expense of more frequent classes (such as *Political*). Additionally, class weights were problematic in the joint language setting, causing varying performance across languages while maintaining similar overall $F1_{micro}$. (A comparison for XLM-R is provided in the Appendix Table 13.) For this reason, we chose to use a mix of class-weighted and non-class-weighted models for our ensembles in order to reduce the variance of our final systems.

4.3 Experimental setup

4.3.1 Data Preprocessing

For both monolingual and multilingual models, we cleaned and preprocessed the article text using a set of steps described in appendix B and truncated it to the first 512 tokens. For monolingual English models, we used Google Translate to produce English translations.

4.3.2 Data Split

For subtask 2, we merged the organiser training and development sets, and used 3-fold cross-validation (stratified by language) to identify the best model configurations. We then produced final models by training on the entire training and organizer development set (rather than a 2/3 fold).

Although this meant we did not have a validation set to judge the final models that went into our ensemble, it enabled training on all available data, which was important due to the small size of the dataset.

4.3.3 Hyperparameters

We finetuned our monolingual models for a fixed 20 epochs using a learning rate of 3e-5 (warm-up ratio 0.1; linear decay), a batch size of 8, and the AdamW optimiser.

We used the same hyperparameters for adapter finetuning, except we raised the learning rate to 1e-4. In general, adapters require a higher learning rate than traditional finetuning, and this is reflected by the findings of Chalkidis et al. (2021) for a similar task.

4.3.4 Cross-validation Findings and Language Selection

Table 5 displays a condensed summary of our cross-validation results (for full version, see Appendix Table 13).[4] [5]

For monolingual models, $\text{MUPPET}_{\text{LARGE}}$ achieves an $F1_{micro}$ of 70.4 on English, outperforming the RoBERTa baseline by 2 points. Similarly, our XLM-R + TAPT + Adapters demonstrates small but consistent improvements over the multilingual XLM-R baseline across most languages.

When comparing across monolingual and multilingual models, we see that for English, XLM-R models are unable to compete with the performance of monolingual MUPPET. (They are, however, able to match the performance of their monolingual counterpart RoBERTa). In contrast to this, the multilingual models generally demonstrated better performance on non-English languages. This is reflected by the overall $F1_{micro}$ scores: MUPPET's 63.5 vs XLMR+TAPT+Adapter's 64.2. Based on these results, we decided to use MUPPET for English and XLM-R for other languages.

For the unseen languages, we decided to use the monolingual 'translate-test' approach based on additional holdout experiments that indicated better MUPPET performance. Although this decision enabled us to achieve 1st place on the evaluation test set for Greek and Georgian, our post-competition findings (section 4.5) discovered that submitting our multilingual model may have achieved even better results.

4.4 Results and Reflections

The scores and positions of our model are shown in Table 6.

The strong performance of our monolingual model, which achieved first place in 3 out of 4 languages submitted, suggests that the 'translate-test' approach is competitive for performing multilingual classification, especially for a zero-shot cross-lingual scenario.

Although our multilingual model also performed well, the competition results suggest that even better performance may be achieved by applying our monolingual approach to other languages i.e. for each language, training a native monolingual model with translations.

Article truncation poses a limitation on our findings, since news frames can potentially be located in parts of the article that are unseen by the model. One of important future extension of these experiments would be to apply long-document processing techniques to this task.

[4]This table shows the average performance of individual models trained during cross-validation, and not the performance of any ensembles.

[5]The Overall $F1_{micro}$ column refers to the $F1_{micro}$ of the entire validation fold and not the mean $F1_{micro}$ score across languages.

Monolingual English	EN	DE	FR	IT	PO	RU	Overall $F1_{micro}$
RoBERTa-Large	68.4 ± 2.0	63.5 ± 2.0	57.9 ± 2.9	60.9 ± 0.2	65.8 ± 3.4	54.5 ± 2.7	63.6 ± 0.1
MUPPET-Large	**70.4 ± 2.0**	62.1 ± 3.7	**59.0 ± 0.9**	58.3 ± 1.5	65.7 ± 0.9	52.9 ± 1.7	63.5 ± 0.7
Multilingual							
XLM-R	68.3 ± 1.4	64.4 ± 1.4	58.5 ± 0.7	60.6 ± 0.5	66.5 ± 3.3	54.9 ± 2.0	64.0 ± 1.2
XLM-R + TAPT + Adapters	68.2 ± 0.9	**65.0 ± 1.8**	58.5 ± 2.8	**61.0 ± 0.6**	**66.7 ± 3.0**	**55.7 ± 3.1**	**64.2 ± 0.3**

Table 5: Mean $F1_{micro}$ scores of class-weighted models (over 3-fold cross-validation). For complete version with ablations and other configurations, see Table 13

Language	Test $F1_{micro}$	Place
Monolingual MUPPET		
English	57.895	1
Spanish*	50.829	3
Greek*	54.630	1
Georgian*	65.421	1
Multilingual XLM-R (+ TAPT + Adpt)		
French	53.425	3
German	65.251	3
Italian	57.079	7
Polish	64.516	2
Russian	44.144	2

Table 6: Subtask 2 final leaderboard results for monolingual and multilingual systems. *Translated to English

4.5 Post-competition Findings

Is a 'translate-test' monolingual model really better than a multilingual model?

After the competition ended, we wanted to compare the performance of our two ensembles across all test set languages, as each system had only been submitted for a subset of languages.

Surprisingly, in contrast to our cross-validation experiments, we found that for English, our multilingual system outperformed our monolingual submission: 58.475 (multi) vs 57.895 (mono). The multilingual system also performed better on two of the surprise languages: Greek (58.0 vs 54.63) and Spanish (52.023 vs 50.829). In contrast to the findings of Xenouleas et al. (2022), who found translation-based approaches "vastly outperform cross-lingual finetuning with adapters", this suggests that the two approaches are competitive with each other. It is difficult, however, to draw firm conclusions from this finding due to the small size of the unseen test set as well as the impact of TAPT. For more details, see Appendix D.

5 System Description for Subtask 3

5.1 System Overview

In subtask 3, our focus was not only to maximize the overall $F1_{micro}$, but also to ensure a balanced model performance across all classes[6]. Due to the highly imbalanced nature of the 23 classes in subtask 3, achieving a balanced model performance across all classes is challenging. To address this issue, similarly to subtask 2, we explored cross-lingual training and implemented class weighting. We also explored an oversampling technique, but it did not provide any additional benefit compared to class weighting and increased the training time.

In contrast to the organiser baseline, which discards paragraphs without a label, we assigned them a class vector of zeros and included them in the training set. This method lead to a significant improvement in performance across all languages[7].

We evaluated the performance of the models on unseen languages by training multilingual models while holding out each language. However, the performance significantly decreased for every language when held out. Therefore, we translated the articles of the three unseen languages into English for the final predictions.

Our final models were a **RoBERTa$_{\mathbf{BASE}}$** for the submission of English and translated unseen languages, and an **mBERT$_{\mathbf{BASE}}$** for the remaining languages.

5.2 Experimental Setup

RoBERTa$_{\mathbf{BASE}}$ was finetuned for 20 epochs only on English data[8], with a batch size of 32, truncated to 256 tokens, and AdamW optimizer with a learning rate of 0.00005, 20% of the training steps as linear warm-up and weight decay of 0.1. A classification threshold of 0.4 was used.

mBERT$_{\mathbf{BASE}}$ was finetuned on all languages combined and shuffled, with the same hyperparameters as the RoBERTa$_{BASE}$ model, except a batch size of 16.

We conducted an experiment to explore cross-

[6]We further discuss this in appendix E.3

[7]A comparison of these approaches are in appendix E.1.

[8]Unlike subtask 2, we finetuned the model only with the original English dataset without adding other translated languages.

Language	$F1_{micro} \pm 1$ std	
	Monolingual	Multilingual
English	**36.2 ± 0.3**	31.8 ± 0.6
French	40.5 ± 0.4	**43.4 ± 0.4**
German	36.9 ± 0.5	**40.9 ± 0.7**
Italian	43.4 ± 1.3	**47.5 ± 0.3**
Polish	28.9 ± 0.8	**30.2 ± 0.8**
Russian	31.5 ± 1.3	**37.5 ± 1.7**

Table 7: Monolingual vs. multilingual $F1_{micro}$ scores on the development set for each language on subtask 3. Best $F1_{micro}$ per language are marked as **bold**.

Language	$F1_{micro} \pm 1$ std
	Zero-Shot
English	21.6 ± 0.4
French	35.8 ± 0.7
German	28.0 ± 0.6
Italian	34.3 ± 0.4
Polish	21.7 ± 0.6
Russian	18.6 ± 1.1

Table 8: Subtask 3 $F1_{micro}$ for the zero-shot experiments.

Language	Test $F1_{micro}$	Place
Monolingual RoBERTa$_{BASE}$		
English	36.802	2
Spanish*	27.497	9
Greek*	17.426	7
Georgian*	24.911	10
Multilingual mBERT		
French	41.436	4
German	44.726	6
Italian	52.494	3
Polish	34.700	7
Russian	31.841	5

Table 9: Subtask 3 final leaderboard results for monolingual and multilingual systems. *Translated to English.

lingual training by comparing monolingual vs. multilingual models for each language with fixed hyperparameters, across three runs. The monolingual models were finetuned and validated on only one language, using either RoBERTa-Base for English or mBERT for the other languages, while the multilingual version was an mBERT finetuned on all the languages and validated on each language separately.

For the final submissions, we merged the training and development sets and finetuned the models without validation data. The model with the lowest training loss across all epochs was selected as the best model. We used the random seed that produced the best $F1_{micro}$ in the previous development set experiments while training the final model.

5.2.1 Results

Table 9 presents the final leaderboard results for subtask 3. Although the main metric for this subtask is $F1_{micro}$, we highlight that our placings for subtask 3 improve considerably when measuring through $F1_{macro}$, placing top 5 in all languages except the three surprise languages and first place for Italian and French, as shown in Table 16 in the Appendix.

Results for the monolingual and multilingual experiments on the official development set are displayed on Table 7, where we report the $F1_{micro}$ for each language. Scores are reported over three runs on different random seed initialisations. Notice that English is the only language in which the monolingual model outperforms the multilingual version.

Table 8 shows the $F1_{micro}$ for the zero-shot experiments, in which a multilingual model was finetuned on all languages except the one being evaluated. By comparison with Table 7, we see that zero-shot drastically hinders the performance on all languages. For this reason, we decided to translate the test sets for the three surprise languages into English and perform inference using the English monolingual model. We were unable to experiment with translating into different languages other than English due to time constraints.

6 Conclusion

We presented three systems aimed at solving three subtasks within SemEval-2023 Task 3. Our systems applied a variety of state-of-the-art techniques including adapters and TAPT, and consistently achieved a high rank across all available languages, including zero-shot low-resource languages. We additionally presented an analysis of the viability of monolingual vs multilingual approaches for each subtask, and found that the results of the comparison vary depending on the subtask. For subtask 1, multilingual transformer models demonstrate better average performance than monolingual models with translations. A similar effect was observed for subtask 2 and subtask 3 where multilingual settings achieved better performance than monolingual ones in all languages except English. We found the impact on bottleneck adapters to be unpredictable across tasks – despite performing on average better for monolingual models in subtask 1, they were more beneficial for multilingual models in subtask 2 (and hindered monolingual performance). Finally, we presented post-competition findings, which suggest that subtask 2 would have benefited from a zero-shot prediction using multilingual models, while subtask 1 could have achieved much better results with the 'translate-test' approach. Further analysis of this will be possible when test labels are released.

7 Acknowledgments

This work has been co-funded by the European Union under the Horizon Europe vera.ai (grant 101070093) and Vigilant (grant 101073921) projects and the UK's innovation agency (Innovate UK) grants 10039055 and 10039039.

References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. 2020. Translating similar languages: Role of mutual intelligibility in multilingual transformers. In *Proceedings of the Fifth Conference on Machine Translation*, pages 381–386, Online. Association for Computational Linguistics.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8614–8624, Online. Association for Computational Linguistics.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. 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*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. 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)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA. Association for Computing Machinery.

Gaël Guibon, Liana Ermakova, Hosni Seffih, Anton Firsov, and Guillaume Le Noé-Bienvenu. 2019. Multilingual fake news detection with satire. In *CICLing: International Conference on Computational Linguistics and Intelligent Text Processing*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *Computing Research Repository*, arXiv:2006.03654. Version 6.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained 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)*, pages 2208–2222, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Seungwon Kim, Alex Shum, Nathan Susanj, and Jonathan Hilgart. 2021. Revisiting pretraining with adapters. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 90–99, Online. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *Computing Research Repository*, arXiv:1905.07213. Version 1.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *Computing Research Repository*, arXiv:1909:11942. Version 6.

Hongying Liu and Ling Xie. 2021. Research on sarcasm detection of news headlines based on Bert-LSTM. In *2021 IEEE international conference on emergency science and information technology (ICESIT)*, pages 89–92. IEEE.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. *Computing Research Repository*, arXiv:1907.11692. Version 1.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 660–665, Minneapolis, Minnesota. Association for Computational Linguistics.

Rajnish Pandey and Jyoti Prakash Singh. 2022. BERT-LSTM model for sarcasm detection in code-mixed social media post. *Journal of Intelligent Information Systems*, pages 1–20.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. 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)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. 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 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Junfeng Tian, Min Gui, Chenliang Li, Ming Yan, and Wenming Xiao. 2021. MinD at SemEval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1082–1087, Online. Association for Computational Linguistics.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China. Association for Computational Linguistics.

Cindy Wang and Michele Banko. 2021. Practical transformer-based multilingual text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Stratos Xenouleas, Alexia Tsoukara, Giannis Panagiotakis, Ilias Chalkidis, and Ion Androutsopoulos. 2022. 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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

A Language Models Used

The models used in each subtask are shown in Table 10.

B Article Preprocessing

Articles were preprocessed with the following steps, for all languages:

- add a full stop at the end of the title.
- remove duplicate sentences directly following each other;
- remove the @ symbol from twitter handles;
- remove hyperlinks to websites and images;

English articles were further preprocessed:

- remove lines indicating the possibility to share the article on different social media platforms;
- remove sentences suggesting the user to take part in online polls, comments, or advertisements;
- remove sentences indicating the terms of use;
- remove sentences indicating the licenses and containing phrases such as 'reprinted with permission', 'posted with permission' and 'all rights reserved';
- remove sentences relating to the article author biographies

C Subtask 1

C.1 Language-specific performance after each epoch on development set

Figure 1 shows $F1_{macro}$ scores on the held-out development set for one of the finetuned transformer models. As can be seen, Polish reaches its best performance quite early on, while German and Russian need more than 17 epochs to achieve the best score.

D Subtask 2

Hyperparameters for Subtask 2 TAPT are shown in Table 11. The detailed cross-validation results are shown in Table 13.

Language	Huggingface Model Name	Publication	Subtasks		
English	`bert-base-cased`	Devlin et al. (2019)	1		
English	`RoBERTa-base`	Liu et al. (2019b)			3*
English	`RoBERTa-large`	Liu et al. (2019b)		2	
English	`MUPPET-large`	Aghajanyan et al. (2021)		2*	
French	`camembert-base`	Martin et al. (2020)	1		
German	`deepset/gbert-base`	Chan et al. (2020)	1		
Italian	`dbmdz/bert-base-italian-cased`	-	1		
Polish	`dkleczek/bert-base-polish-cased-v1`	-	1		
Russian	`DeepPavlov/rubert-base-cased`	Kuratov and Arkhipov (2019)	1		
Multilingual	`bert-base-multilingual-cased`	Devlin et al. (2019)	1*		3*
Multilingual	`xlm-RoBERTa-base`	Conneau et al. (2020)	1		3
Multilingual	`xlm-RoBERTa-large`	Conneau et al. (2020)		2*	

Table 10: Models used in each subtask. * denotes the models used in the final submission. Based on model selection of Chalkidis et al. (2021).

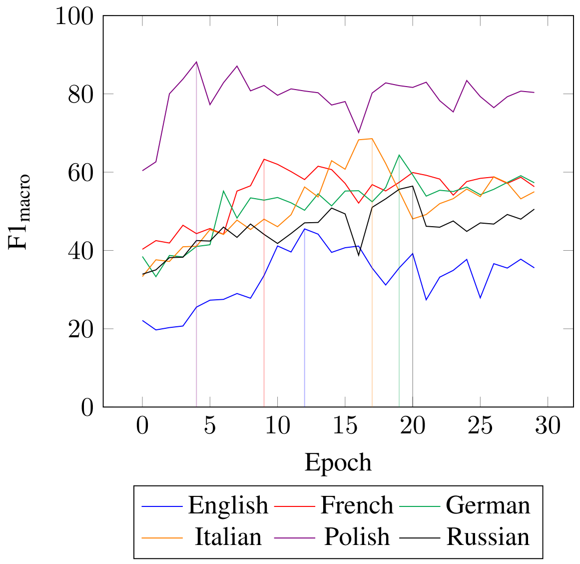

Figure 1: Validation $\mathrm{F1}_{\mathrm{macro}}$ of each language over time, with maximal epoch indicated.

Epochs	60
Effective Batch Size	128
Max learning rate	1e-4
Warmup ratio	0.06, linear
Learning rate decay	linear
Optimiser	AdamW
Adam epsilon	1e-6
Adam beta weights	0.9, 0.98
Weight decay	0.01

Table 11: Subtask 2: TAPT Hyperparameters

Language	Multilingual	Monolingual
English	**58.475**	57.895 (1)
French	53.425 (3)	**54.181**
German	**65.251** (3)	62.069
Italian	57.079 (7)	**60.577**
Polish	**64.516** (2)	63.581
Russian	**44.144** (2)	40.800
Spanish	**52.023**	50.829 (3)
Greek	**58.000**	54.630 (1)
Georgian	60.870	**65.421** (1)

Table 12: Subtask 2: $\mathrm{F1}_{\mathrm{micro}}$ Perfomance on test set - post-competition comparison. () indicates ranking for official submissions.

D.1 Post-competition Findings

Table 12 shows the full set of post-competition results. The multilingual model outperforms our official monolingual submission for test set English, Greek, and Spanish, but is much worse for Georgian. Conversely, our monolingual model does better than our official multilingual submission for French and Italian. This suggests that neither monolingual nor multilingual models are consistently better than the other across all languages.

Traditional zero-shot cross-lingual experiments focus on finetuning a multilingual model on a single source language and testing in a target language. In our situation, we jointly finetune on multiple languages, which encourages the model to retain multilingual representations, even for languages not in its training, thus improving its zero-shot cross-lingual capabilities. This may help to explain the improved performance of our multilingual model. However, it is important to note that these results are not representative of true zero-shot classification, since our multilingual model did perform task-adaptive pre-training on articles from the surprise languages. Unfortunately, because the organisers have not released labels for the test set, we are unable to perform error analysis. As mentioned in the main section, the small size of the test set also makes it difficult to draw firm conclusions on whether translate-test is better than multilingual zero-shot classification.

Monolingual English	EN	DE	FR	IT	PO	RU	Overall $F1_{micro}$
RoBERTa-Large	68.4 ± 2.0	63.5 ± 2.0	57.9 ± 2.9	60.9 ± 0.2	65.8 ± 3.4	54.5 ± 2.7	63.6 ± 0.1
MUPPET-Large	70.4 ± 2.0	62.1 ± 3.7	59.0 ± 0.9	58.3 ± 1.5	65.7 ± 0.9	52.9 ± 1.7	63.5 ± 0.7
MUPPET-Large + Adapters	68.0 ± 1.0	59.5 ± 1.6	54.5 ± 1.9	58.0 ± 0.7	61.9 ± 2.1	51.0 ± 3.7	61.1 ± 0.9
Multilingual Models							
XLM-R	68.3 ± 1.4	64.4 ± 1.4	58.5 ± 0.7	60.6 ± 0.5	66.5 ± 3.3	54.9 ± 2.0	64.0 ± 1.2
XLM-R + Adapters	69.0 ± 1.2	64.0 ± 1.4	58.4 ± 3.0	61.3 ± 0.9	67.4 ± 1.0	53.1 ± 2.7	64.3 ± 0.4
XLM-R + TAPT + Adapters	68.2 ± 0.9	65.0 ± 1.8	58.5 ± 2.8	61.0 ± 0.6	66.7 ± 3.0	55.7 ± 3.1	64.2 ± 0.3
XLM-R (no class weights)	68.8 ± 1.7	57.1 ± 2.2	61.6 ± 4.0	61.7 ± 1.2	67.4 ± 2.1	57.0 ± 2.2	65.1 ± 0.2

Table 13: Full version of Subtask 2 cross-validation results. Comparison of averaged $F1_{micro}$ scores on 3-fold cross-validation (merged training and organiser-dev set). All models have class-weighting, except where indicated otherwise.

Language	Training Set Size	
	Without Non-Labelled	With Non-Labelled
English	3760	9498 (+152%)
French	1693	2259 (+33%)
German	1252	1555 (+24%)
Italian	1745	2623 (+50%)
Polish	1232	2310 (+32%)
Russian	1245	1962 (+57%)

Table 14: Subtask 3 train set sizes for each language without and with the addition of examples that weren't assigned a class during labelling.

Language	$F1_{micro} \pm 1$ std	
	Without Classless	With Classless
English	27.1 ± 1.0	**36.2 ± 0.3**
French	41.3 ± 0.1	**43.4 ± 0.4**
German	40.8 ± 0.1	**40.9 ± 0.8**
Italian	44.1 ± 0.6	**47.5 ± 0.4**
Polish	27.8 ± 0.9	**30.2 ± 0.1**
Russian	35.7 ± 0.9	**37.5 ± 2.0**

Table 15: Subtask 3 $F1_{micro}$ for best model configurations for each language with and without the addition of classless examples. Best $F1_{micro}$ per language are marked as **bold**.

E Subtask 3

E.1 Training With vs. Without Non-Labelled Examples

In the combined training set across all the languages, there are 9,280 paragraphs that do not have a label. Although it is expected that this also occurs on the test set, the organizer's baseline approach discards these train samples, so it never explicitly trains on unlabelled examples. Table 14 displays the sizes of the train set for each language without adding non-labelled examples vs. adding them. Table 15 shows the $F1_{micro}$ results of both approaches, with means and stds computed over three random seed initialisations. Note that adding the non-labelled examples contributes to a considerable increase in performance for all languages, particularly English, which is also the language that had the biggest increase in train set size.

E.2 Development Set Fine-grained Results

Table 17 shows the fine-grained results for the English official development set. Results are obtained from the best random seed over three runs. Although *Appeal_to_Time*, *Appeal_to_Values*, *Consequential_Oversimplification* and *Questioning_the_Reputation* classes do not have a single example in development set, there are six other classes in which we also obtain 0.0 F1-Score, namely *Appeal_to_Hypocrisy*, *Appeal_to_Popularity*, *Obfuscation-Vagueness-Confusion*, *Red_Herring*, *Straw_Man* and *Whataboutism*, although together they account for only 5% of the development set. The three biggest classes, *Loaded_Language*, *Name_Calling-Labeling* and *Doubt* account for 29%, 15% and 11% of the development set, respectively, thus having a large impact on $F1_{micro}$.

	Final Submission			
	Test $F1_{micro}$	Place	Test $F1_{macro}$	Place
English	36.802	2	17.194	2
French	41.436	4	32.424	1
German	44.726	6	23.679	3
Italian	52.494	3	28.22	1
Polish	34.7	7	19.102	4
Russian	31.841	5	20.522	2
Greek	17.426	7	11.028	8
Spanish	27.497	9	13.042	8
Georgian	24.911	10	29.553	4

Table 16: Subtask 3 final submission $F1_{micro}$ and $F1_{macro}$ and our placement according to both of them.

Class	Precision	Recall	F1-Score	Samples
Appeal_to_Authority	0.11	0.07	0.09	28
Appeal_to_Fear-Prejudice	0.39	0.23	0.29	137
Appeal_to_Hypocrisy	0	0	0	8
Appeal_to_Popularity	0	0	0	34
Appeal_to_Time	0	0	0	0
Appeal_to_Values	0	0	0	0
Causal_Oversimplification	0.03	0.04	0.04	24
Consequential_Oversimplification	0	0	0	0
Conversation_Killer	0.11	0.28	0.16	25
Doubt	0.26	0.36	0.3	187
Exaggeration-Minimisation	0.21	0.34	0.26	115
False_Dilemma-No_Choice	0.26	0.16	0.2	63
Flag_Waving	0.34	0.49	0.4	96
Guilt_by_Association	0.33	0.25	0.29	4
Loaded_Language	0.39	0.64	0.48	483
Name_Calling-Labeling	0.42	0.69	0.52	250
Obfuscation-Vagueness-Confusion	0	0	0	13
Questioning_the_Reputation	0	0	0	0
Red_Herring	0	0	0	19
Repetition	0.12	0.24	0.16	141
Slogans	0.21	0.43	0.29	28
Straw_Man	0	0	0	9
Whataboutism	0	0	0	2
micro avg	0.31	0.44	0.36	1666
macro avg	0.14	0.18	0.15	1666

Table 17: Subtask 3 fine-grained results for the English development set.

E.3 Full Leaderboard Results

Table 16 shows our full final submission scores and placements according to both $F1_{micro}$ and $F1_{macro}$. As we previously point out in section 5.1, we aimed towards a model capable of identifying all the 23 classes, thus having high $F1_{macro}$, even though the main metric for the subtask is $F1_{micro}$. We believe that a realistic application of a model for this particular label scheme should not disregard under-represented classes, otherwise they should simply be removed from the label scheme. Although our placings according to $F1_{macro}$ are considerably higher, we acknowledge that because the main metric for the subtask is not $F1_{macro}$, other teams' submissions are likely not focusing on maximizing it, thus making their scores lower on average.