Team Hitachi at SemEval-2023 Task 3: Exploring Cross-lingual Multi-task Strategies for Genre and Framing Detection in Online News

Yuta Koreeda[*], **Ken-ichi Yokote**[*], **Hiroaki Ozaki**,
Atsuki Yamaguchi, **Masaya Tsunokake** and **Yasuhiro Sogawa**
Research and Development Group, Hitachi, Ltd.
Kokubunji, Tokyo, Japan
{yuta.koreeda.pb, kenichi.yokote.fb, hiroaki.ozaki.yu,
atsuki.yamaguchi.xn, masaya.tsunokake.qu, yasuhiro.sogawa.tp}@hitachi.com

arXiv:2303.01794v1 [cs.CL] 3 Mar 2023

Abstract

This paper explains the participation of team *Hitachi* to SemEval-2023 Task 3 *"Detecting the genre, the framing, and the persuasion techniques in online news in a multilingual setup"*. Based on the multilingual, multi-task nature of the task and the setting that training data is limited, we investigated different strategies for training the pre-trained language models under low resource settings. Through extensive experiments, we found that (a) cross-lingual/multi-task training, and (b) collecting an external balanced dataset, can benefit the genre and framing detection. We constructed ensemble models from the results and achieved the highest macro-averaged F1 scores in Italian and Russian genre categorization subtasks.

1 Introduction

As we pay more and more attention to the socially influencing problems like COVID-19 and the Russo-Ukrainian war, there has been an increasing concern about *infodemic* of false and misleading information (Piskorski et al., 2023). In particular, cross-lingual understanding of such information is becoming more important due to polarization of political stances, economical decoupling and echo chamber effect in social media. To that end, Piskorski et al. (2023) put together SemEval-2023 Task 3 *"Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup"*. The shared task aims to analyze several aspects of what makes a text persuasive and to foster development of building blocks for multilingual media analysis.

Creating annotated data for media analysis is time consuming thus we cannot assume that we can obtain training data of enough quality and quantity. To tackle the problem, we investigated and compared strategies for multilingual media analysis under low resource settings. Through extensive experiments, we found that (a) cross-lingual/multi-task training, and (b) collecting an external balanced dataset, can benefit the genre and framing detection. We constructed ensemble models from the results and participated in genre categorization (subtask 1) and framing detection (subtask 2) in six languages, where we achieved the highest macro-averaged F1 scores in Italian and Russian subtask 1.

2 Task Definition and our Strategy

SemEval-2023 Task 3 aims to analyze several aspects of what makes a text persuasive. It offers three subtasks on news articles in six languages: German (de), English (en), French (fr), Italian (it), Polish (pl) and Russian (ru).

Subtask 1: News genre categorization Given a news article, a system has to determine whether it is an opinion piece, it aims at objective news reporting, or it is a satire piece. This is multi-class document classification and the official evaluation measure is macro average F1 score (macro-F1) over the three classes.

Subtask 2: Framing detection Given a news article, a system has to identify what key aspects (frames) are highlighted the rhetoric from 14 frames (see (Card et al., 2015) for the taxonomy and definitions). This is multi-label document classification and the official evaluation measure is micro average F1 score (micro-F1) over the 14 frames.

Subtask 3: Persuasion techniques detection Given a news article, a system has to identify the persuasion techniques in each paragraph from 23 persuasion techniques. This is multi-label paragraph classification.

The target articles are those identified to be potentially spreading mis-/disinformation and are collected from 2020 to mid-2022. They revolve around widely discussed topics such as COVID-

[*] Equal contribution

19, migration, the build-up leading to the Russo-Ukrainian war, and some country-specific local events such as elections.

We observed that the numbers of articles are limited for the relative large label space and there exist considerable overlaps of articles between subtask 1 and 2 (see Appendix A.1). Hence, we decided to investigate if models trained on multiple languages or another subtask can benefit the target task in this low resource setting (Section 5). Since subtask 1 and 2 conveniently share the task format, we opted to participate in subtask 1 and 2 in all the six languages.

We also noticed that the English training dataset exhibits significant difference in label distribution to other languages and it is unbalanced. Hence, we decided to collect additional external dataset for English subtask 1 in a wish to improve task performance in English and to help with other languages through cross-lingual training (Section 3).

3 External Data for English Genre Categorization

In a preliminary analysis of the English subtask 1 dataset, we found that label distribution is quite unbalanced and it is different in the training and the development data. Therefore, we did not make any assumption about the distribution of the test data and decided to increase the number of rare labels in order to create a new, balanced dataset for English genre categorization. First, we undersampled articles from the training dataset for subtask 1 such that the numbers of articles for each label are equal, i.e., ten articles for each label.

We referred to a survey on fake news detection datasets (D'Ulizia et al., 2021) and checked a total of 27 datasets to see if they can be converted to subtask 1 dataset format using the following criteria:

Label similarity We checked whether the labels defined by an external dataset are close to subtask 1. For example, we focused on whether they used identical label names, such as "satire".

Text similarity We checked if the text type of a dataset is similar to subtask 1, such as whether they use news articles.

Task similarity We checked whether the task setting employed by a dataset is a method of classifying them into different classes rather than, for example, scoring them with a scale of 1 to 5.

After these checks, we adopted the Random Political News Data (Horne and Adali, 2017) which holds 75 articles for each of three labels and added the total of 225 articles to the sampled 30 original articles. This resulted in the *Augmented (small)* dataset which contains 255 articles in total. Since Horne and Adali (2017) describe the news media from which the data was collected, we independently collected 31 additional articles for each label and constructed *Augmented (large)* with 348 articles altogether.

4 Cross-lingual Multi-task Transformers

We utilized pretrained language models (PLMs) in a simple sequence classification setup (Vaswani et al., 2017; Devlin et al., 2019). We employed XLM-RoBERTa large[1] (Conneau et al., 2020) and RemBERT[1] (Chung et al., 2021). For English, we also utilized RoBERTa large[1] (Liu et al., 2019) for single-language multi-task training. In order to allow multi-task training, we added a classifier head for each subtask[2] on top of each model followed by softmax for subtask 1 and sigmoid for subtask 2. Hence, each model shares most of the parameters for the two subtasks. We used Transformers library (Wolf et al., 2020) for the implementations.

In multi-task training we simply took sum of losses from two tasks. Since there exist articles that only have either of subtask 1 or 2 labels, we ignore predictions for missing labels from loss calculation. For cross-lingual training, we simply concatenate all articles. All parameters are shared for the same subtask in different languages. For preprocessing, we simply concatenated all sentences from each article and tokenized them using the default tokenizer for each PLM.

5 Exploring Cross-lingual Multi-task Strategies

It is empirically known that further fine-tuning a model trained in a multi-task or cross-lingual setting on the target downstream task/language improves its performance. Also, models tend to require different hyperparameters for different training paradigms or languages. Hence, we decided to explore different multi-task and cross-lingual strategies through a series of random hyperparameter searches (Figure 1).

First, we ran a random hyperparameter search

[1] https://huggingface.co/{xlm-roberta-large, rembert, roberta-large}

[2] dropout→linear→tanh→dropout→linear for (XLM-)RoBERTa, and dropout→linear for RemBERT

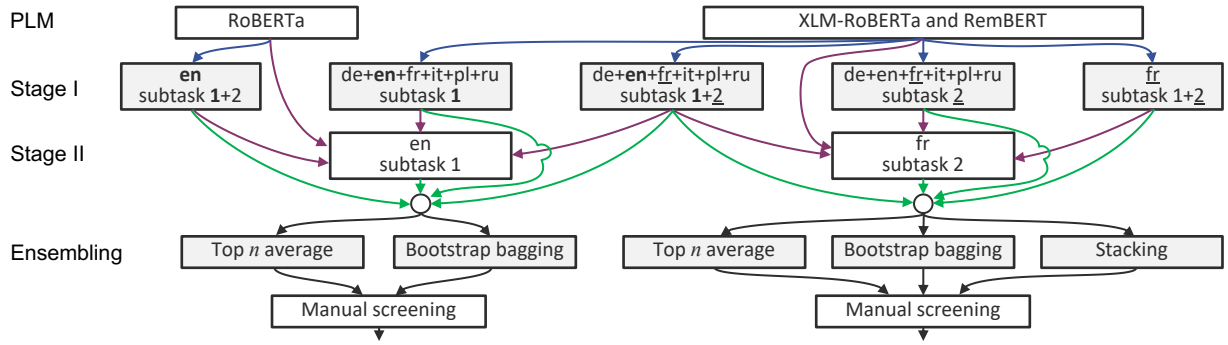

Figure 1: Series of random hyperparameter searches for exploring cross-lingual multi-task strategies. It only shows English subtask 1 and French subtask 2 but we did the same for all other languages in subtask 1 and 2.

in cross-lingual and/or multi-task settings (Stage I). Regarding the resulting Stage I models as an additional hyperparameter, we ran another random hyperparameter search to optimize the choice of the pretraining paradigm along with other hyperparameters (Stage II). Finally, we construct an ensemble for each language-subtask pair from all models in Stage I and II using their performance in the development dataset.

Unlike more sophisticated hyperparameter search methods, this approach has an advantage that we can compare and evaluate different training paradigms post hoc.

The choice of the subtask 1 English datasets (Section 3) is also incorporated as an additional hyperparameter. The hyperparameter search spaces are shown in Appendix A.2.

5.1 Stage I Training

In Stage I, we fine-tuned PLM in three settings.

(1) Multi-task (30 hyperparameter sets for each language = 180 models)
(2) Cross-lingual (50 hyperparameter sets for each subtask = 100 models)
(3) Cross-lingual multi-task (50 hyperparameter sets = 50 models)

Hence, we trained 330 models in Stage I.

5.2 Stage II Training

Stage I results in three groups of models that have been trained on each language-subtask pair. For example, "en subtask 1" in Figure 1 has incoming arrows from (1) multi-task ("en subtask 1+2"), (2) cross-lingual ("de+en+fr+it+pl+ru subtask 1"), and (3) cross-lingual multi-task ("de+en+fr+it+pl+ru subtask 1+2"). We also utilize vanilla PLMs for Stage II training (see the arrow from RoBERTa).

For each language-subtask pair, we picked four models from each group, resulting in 12 models for each language-subtask pair. The four models were chosen whose macro-F1, micro-F1, ROC-AUC or mAP was the best in the development dataset for the target language-subtask pair. This means that the same model can be chosen multiple times (e.g., a model which was the best in macro-F1 and micro-F1). We did not remove the duplicates in that case — such model will be sampled twice as much as a model which was the best only in a single metric.

Regarding these Stage I models and vanilla PLMs as an additional hyperparameter, we carried out Stage II random hyperparameter search on each language. We sampled Stage I models three times more than PLMs, so that all groups (i.e., the four arrows entering "en subtask 1" in Figure 1) are sampled equally. We trained 50 models for each language-subtask pair (50 models $\times$ 6 languages $\times$ 2 subtasks = 600 models).

5.3 Ensembling

Finally, we created an ensemble for each language-subtask pair from the results of hyperparameter search. In a rare case, fine-tuning the model on the downstream task can degrade the performance. Hence, we also considered the Stage I models for the ensemble.

We implemented multiple ensemble methods and manually chose the best one for each language-subtask pair while monitoring multiple leave-one-out metrics on the development dataset. The details of ensembling are described in Appendix A.3.

6 Results

6.1 Subtask 1: News Genre Categorization

Excerpts from the official leaderboards (Piskorski et al., 2023) for subtask 1 are shown in Table 1.

	Team	macro	micro
1	UMUTeam	81.95	82.00
	vera	81.95	82.00
	⋮	⋮	⋮
5	MELODI	77.89	78.00
6	**Hitachi**	**77.66**	**76.00**
7	SharoffAndLepekhin	71.27	72.00

(a) German (15 teams)

	Team	macro	micro
1	MELODI	78.43	81.48
2	MLModeler5	61.63	62.96
	⋮	⋮	⋮
6	Unisa	58.62	61.11
7	**Hitachi**	**55.29**	**59.26**
8	UnedMediaBiasTeam	52.36	57.41

(b) English (22 teams)

	Team	macro	micro
1	UMUTeam	83.55	88.00
2	QCRITeam	76.74	80.00
3	**Hitachi**	**74.36**	**78.00**
4	DSHacker	71.05	72.00
5	vera	68.16	74.00
6	SharoffAndLepekhin	67.14	78.00

(c) French (16 teams)

	Team	macro	micro
1	**Hitachi**	**76.83**	**85.25**
2	QUST	76.68	83.61
3	DSHacker	72.04	83.61
	vera	72.04	83.61
5	MELODI	58.67	75.41
6	UnedMediaBiasTeam	58.41	62.30

(d) Italian (16 teams)

	Team	macro	micro
1	SharoffAndLepekhin	78.55	93.62
2	**Hitachi**	**77.92**	**87.23**
3	vera	76.45	85.11
4	MELODI	70.86	85.11
5	UMUTeam	66.43	80.85
6	SinaaAI	66.35	80.85

(e) Polish (16 teams)

	Team	macro	micro
1	**Hitachi**	**75.49**	**75.00**
2	vera	72.87	72.22
3	SharoffAndLepekhin	66.80	69.44
4	UMUTeam	64.54	68.06
5	MELODI	58.64	62.50
6	QCRITeam	56.66	65.28

(f) Russian (16 teams)

Table 1: The official leaderboard for subtask 1 showing the rank, macro F1 and micro F1 score

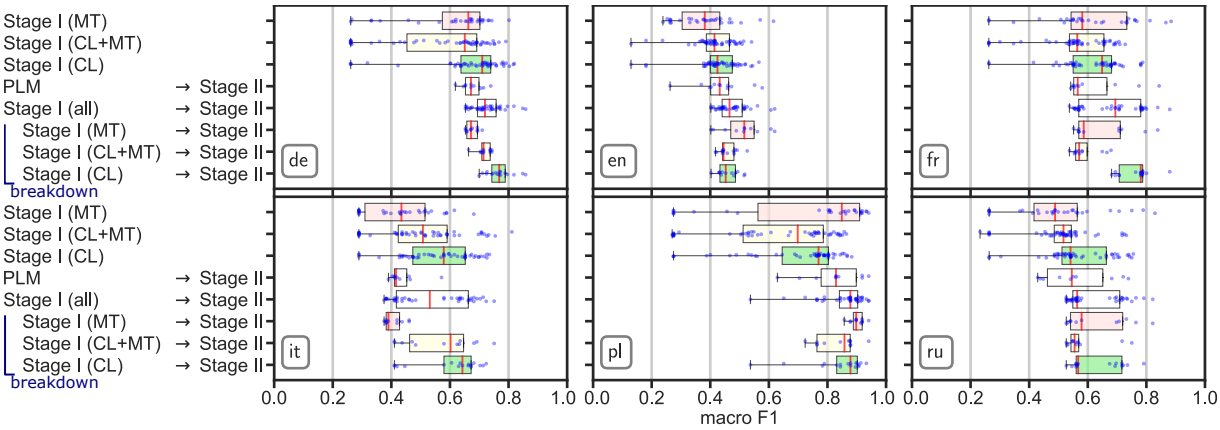

Figure 2: Comparison of subtask 1 macro-F1 under different training paradigms (CL: cross-lingual/MT: multi-task)

	Team	macro F1	micro F1
1	**Hitachi**	**72.93**	**76.79**
2	vera	72.13	77.88
3	MELODI	68.35	76.08
4	DSHacker	67.58	73.52
5	UMUTeam	65.52	75.60
6	MLModeler5	61.63	62.96

Table 2: An unofficial subtask 1 leaderboard by the mean macro-F1 in six languages

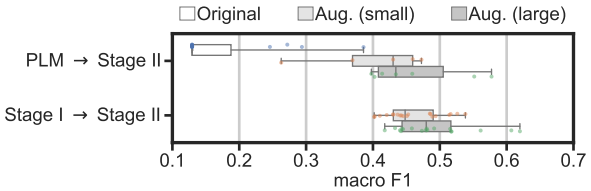

Figure 3: The effect of different training dataset

We were the first place in Italian and Russian and within top threes in French and Polish. In an unofficial ranking of mean macro-F1 of six languages, we were the first place (Table 2).

In Figure 2, we show macro-F1 for the development dataset of all the models considered for the ensemble construction. In all six languages, the models fine-tuned from cross-lingual and/or multi-lingual pretraining tend to perform better (i.e., have better median macro-F1) than the single language/task models trained from PLM ("PLM → Stage II"). This shows that cross-lingual multi-task training was overall useful for genre categorization. In most cases, fine-tuning Stage I models in Stage II yields better results than the vanilla Stage I models.

The breakdown of the performance based on how each model was pretrained in Stage I is also shown in the Figure 2. The results are mixed as to which Stage I pretraining paradigms were useful to the Stage II downstream performance. In German, French and Italian, cross-lingual pretraining tends to be more beneficial than multi-task pretraining. In English, Polish and Russian, multi-task pretraining tends to be more beneficial. Interestingly, the combination of the both was never the best option in any language.

We analyzed the effect of incorporating external, balanced datasets for English subtask 1 (Figure 3). When directly fine-tuning PLM in Stage II, we see that models using either of external datasets tend to be considerably better than ones trained on the original training data. For both conditions, we can see that Augmented (large) tend to perform better than Augmented (small). This shows that obtaining a balanced dataset is important for news

	Team	micro	macro
1	MarsEclipse	71.12	66.05
2	QCRITeam	66.02	60.56
3	vera	65.25	60.14
4	TeamAmpa	63.22	57.27
5	**Hitachi**	**62.91**	**56.73**
6	PolarIce	62.22	56.44

(a) German (18 teams)

	Team	micro	macro
1	vera	57.89	53.90
2	TeamAmpa	56.70	50.96
3	MarsEclipse	56.23	49.05
4	**Hitachi**	**54.26**	**47.16**
5	PolarIce	53.53	48.17
6	QUST	51.31	46.21

(b) English (22 teams)

	Team	micro	macro
1	MarsEclipse	55.28	53.68
2	BERTastic	53.69	52.02
3	vera	53.42	52.03
4	**Hitachi**	**51.41**	**48.83**
5	TeamAmpa	50.56	47.89
6	TheSyllogist	48.57	46.16

(c) French (18 teams)

	Team	micro	macro
1	MarsEclipse	61.73	54.46
2	QCRITeam	59.91	47.95
3	**Hitachi**	**59.77**	**51.51**
4	TeamAmpa	59.67	48.27
5	PolarIce	58.41	46.88
6	UMUTeam	57.63	44.67

(d) Italian (18 teams)

	Team	micro	macro
1	MarsEclipse	67.31	63.84
2	vera	64.52	60.27
3	QCRITeam	64.19	59.87
4	UMUTeam	64.18	59.31
5	**Hitachi**	**63.40**	**58.40**
6	SATLab	62.02	56.99

(e) Polish (18 teams)

	Team	micro	macro
1	MarsEclipse	44.98	30.33
2	vera	44.14	35.59
	⋮	⋮	⋮
8	UMUTeam	38.49	28.84
9	**Hitachi**	**37.00**	**32.59**
10	Riga	31.51	22.19

(f) Russian (17 teams)

Table 3: The official leaderboard for subtask 2 showing the rank, micro F1 and macro F1 score

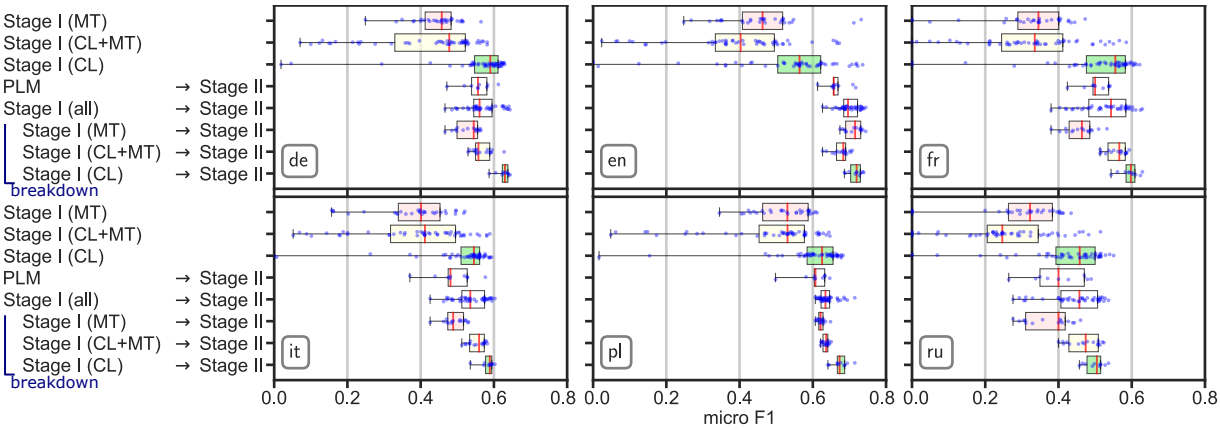

Figure 4: Comparison of subtask 2 micro-F1 under different training paradigms (CL: cross-lingual/MT: multi-task)

	Team	micro F1	macro F1
1	MarsEclipse	59.44	52.90
2	vera	57.05	51.85
3	QCRITeam	55.47	48.29
4	TeamAmpa	55.41	48.22
5	**Hitachi**	**54.79**	**49.20**
6	PolarIce	53.60	47.74

Table 4: An unofficial subtask 2 leaderboard by the mean micro-F1 in six languages

genre categorization.

6.2 Subtask 2: Framing Detection

Excerpts from the official leaderboards (Piskorski et al., 2023) for subtask 2 are shown in Table 3. We were third to fifth places in all but Russian where we obtained the ninth place.

In Figure 4, we show micro-F1 for the development dataset of all the models considered for the ensemble construction. As in subtask 1, the models fine-tuned from cross-lingual and/or multi-lingual pretraining tend to perform better than the single language/task models fine-tuned directly from PLM. This suggests that cross-lingual multi-task training was also useful for framing detection.

In all languages in subtask 2, cross-lingual Stage II pretraining tends to result in better micro-F1 than multi-task models (Figure 4). We suspects that this lies in the difference in the linguistic nature of two subtasks; Framing can be determined by lexical semantic to some extent, hence transfers well across different languages with multilingual transformers. On the other hand, distinguishing the genre requires capturing language-specific pragmatics which may be the reason why it did not transfer between languages as effectively as subtask 2.

7 Conclusion

In our participation to SemEval-2023 Task 3, we investigated different strategies for multilingual genre and framing detection. Through the extensive experiments, we found that collecting an external balanced dataset can help genre categorization. We also find that cross-lingual and multi-task training can help both genre and framing detection and found that cross-lingual training is more beneficial for framing detection. We constructed ensemble models from the results and achieved the highest macro-F1 in Italian and Russian genre detection.

For future work, we will investigate the effect of cross-lingual multi-task training on zero-shot language transfer (Greek, Spanish and Georgian subtasks that we did not participate), as well as the effect on and benefit from training models on persuasion techniques detection (subtask 3).

Acknowledgements

We used computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST) for the experiments.

References

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. 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*.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking Embedding Coupling in Pre-trained Language Models. In *International Conference on Learning Representations*.

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*.

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*.

Arianna D'Ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake News Detection: a Survey of Evaluation Datasets. *PeerJ Computer Science*, 7:e518.

Benjamin Horne and Sibel Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire Than Real News. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

A Appendix

A.1 Data Analysis

Numbers of articles in each subtask and their overlaps are summarized in Table 5. We can see that the numbers of articles are limited for the relative large label space and there exist considerable overlaps of articles between subtask 1 and 2.

Readers should refer (Piskorski et al., 2023) for the details on the datasets.

A.2 Details of Hyperparameter Search

As described in Section 5, we carried out the extensive experiments as a series of hyperparameter searches. In this section, we will list and describe all the hyperparamter search spaces of Stage I and II training.

The hyperparameter search spaces of Stage I training are listed as in the following:

- Cross-lingual multi-task: Table 6
- Cross-lingual Table 7 and 8
- Multi-task: Table 9 and 10

The hyperparameter search spaces of Stage II training are listed in the following:

- Subtask 1 in English: Table 11
- Subtask 1 in all other languages: Table 12
- Subtask 2 in English: Table 13
- Subtask 2 in all other languages: Table 14

We introduced a loss weighting technique and introduced it as an additional hyperparameters.

Language	Subtask 1	Subtask 2	Overlap
Germany (de)	132	132	97
English (en)	433	433	433
French (fr)	157	158	119
Italian (it)	226	227	170
Polish (pl)	144	145	106
Russiun (ru)	142	143	107

Table 5: Numbers of articles in subtask 1 and 2 training data, and their overlaps

Hyperparameter	Values
Base model	XLM-RoBERTa large, RemBERT
Dataset	All, All but English
Classwise training	No
Max steps	100, 200, 300
Learning rate	30, 20, 15, 10, 8, 5 ($\times 10^{-6}$)
Batch size	32, 64
Weight decay	0.02, 0.01, 0.001
Loss scaling	Yes, No
Loss scale threshold	N/A, 5, 10,000
Gradient clipping	1.0
Warmup ratio	0.2

Table 6: Hyperparamter search space for Stage I cross-lingual multi-task training

Hyperparameter	Values
Base model	XLM-RoBERTa large, RemBERT
Dataset	All, All but English
Max steps	100, 200, 300
Learning rate	30, 20, 15, 10, 8, 5 ($\times 10^{-6}$)
Batch size	32, 64
Weight decay	0.02, 0.01, 0.001
Loss scaling	Yes, No
Gradient clipping	1.0
Warmup ratio	0.2

Table 7: Hyperparamter search space for Stage I cross-lingual training subtask 1

Hyperparameter	Values
Base model	XLM-RoBERTa large, RemBERT
Dataset	All, All but English
Classwise training	No
Max steps	100, 200, 300
Learning rate	30, 20, 15, 10, 8, 5 ($\times 10^{-6}$)
Batch size	32, 64
Weight decay	0.02, 0.01, 0.001
Loss scale threshold	N/A, 5, 10,000
Gradient clipping	1.0
Warmup ratio	0.2

Table 8: Hyperparamter search space for Stage I cross-lingual training of subtask 2

Since label distributions are highly skewed in subtask 1, we weight losses for each label $\mathcal{L}_l$ ($l \in \{satire, opinion, reporting\}$) by w_l (i.e., $\mathcal{L}'_l = w_l \cdot \mathcal{L}_l$) such that they are inversely proportional to the count of each label c_l (i.e., $w_l \propto 1/c_l$) while adding up to 1 (i.e., $\sum_l w_l = 1$).

$$w_l = \frac{\text{hmean}(c_{satire} + c_{opinion} + c_{reporting})}{c_l},$$

where hmean is harmonic mean.

For subtask 2, we carried out the classification in both a single multi-label classification and multiple, separate binary classifications. We have also regarded the choice of the classification method as a hyperparamter and incorporated this into the random search in Stage II ("classwise").

A.3 Details of Ensemble Construction

We created an ensemble for each language-subtask pair from the results of hyperparameter search. As outlined in Section 5.3, we implemented multiple ensemble methods and manually chose the best one for each language-subtask pair. Here, we show the details of ensemble construction and selection on each subtask.

Hyperparameter	Values
Base model	RoBERTa large
Dataset	Aug. (large) + official subtask 2 dataset, Aug. (small) + official subtask 2 dataset
Max steps	80, 120, 160, 200
Learning rate	80, 6, 5, 4, 2 ($\times 10^{-6}$)
Batch size	32, 64
Weight decay	0.02, 0.01, 0.001
Loss scaling	Yes, No
Loss scale threshold	N/A, 5, 10,000
Gradient clipping	1.0
Warmup ratio	0.2

Table 9: Hyperparamter search space for Stage I multi-task training of English subtask 1 and 2

Hyperparameter	Values
Base model	XLM-RoBERTa large, RemBERT
Dataset	Official subtask 1 and 2 datasets in each language
Max steps	80, 120, 160, 200
Learning rate	15, 12, 10, 8, 6, 4 ($\times 10^{-6}$)
Batch size	16, 32
Weight decay	0.02, 0.01, 0.001
Loss scaling	Yes, No
Loss scale threshold	N/A, 5, 10,000
Gradient clipping	1.0
Warmup ratio	0.2

Table 10: Hyperparamter search space for Stage I multi-task training of a single language (German, French, Italian, Polish or Russian)

A.3.1 Ensembles for Subtask 1

For subtask 1, we implemented three ensemble methods:

Top one We choose the best model with the best macro-F1 in the development dataset.

Top 3 average We picked three models based on the macro-F1 score in the development dataset. We take an average of the output probabilities (i.e., scores after softmax).

Bootstrap bagging We greedily add models to average ensemble with replacement until the score no longer improves or the ensemble size reaches five. We use the minimum F1 score of all the classes. This idea of trying to improve the worst-class performance was inspired by distributionally robust optimization.

Due the scarcity of the development data, the results tend to be unstable. Hence, we manually chose the best ensemble type for each language with following criteria while monitoring leave-one-out metrics on the development dataset.

- We try to choose model with a good class score balance (i.e., good macro-F1) and good gen-

Hyperparameter	Values
Base model	RoBERTa large, Stage I models
Dataset	Aug. (small), Aug. (large)
Max steps	
if PLM	100, 150, 200
if Stage I	30, 50, 80, 100, 150
Learning rate	
if PLM	20, 15, 10, 8, 6, 4, 3, 2 ($\times 10^{-6}$)
if Stage I	10, 8, 6, 5, 4, 2 ($\times 10^{-6}$)
Batch size	16, 32
Weight decay	0.02, 0.01, 0.001
Gradient clipping	1.0
Warmup ratio	0.2

Table 11: Hyperparamter search space for Stage II training of English subtask 1

Hyperparameter	Values
Base model	XLM-RoBERTa large, RemBERT, Stage 1 models
Dataset	Official dataset for each language
Max steps	
if PLM	160, 200, 240
if Stage I	30, 50, 80, 100, 150
Learning rate	
if PLM	15, 12, 10, 8, 5 ($\times 10^{-6}$)
if Stage I	10, 8, 6, 5, 4, 2 ($\times 10^{-6}$)
Batch size	16, 32
Weight decay	0.02, 0.01, 0.001
Loss scaling)	Yes, No
Gradient clipping	1.0
Warmup ratio	0.2

Table 12: Hyperparamter search space for Stage II training of German, French, Italian, Polish and Russian subtask 1

eral classification abilities (good mAP and ROC-AUC).

- Unless the difference is unbearably large, we tried to avoid top one model as it can be unstable.

After choosing the best ensemble method, we recreated the ensemble using the whole development dataset.

A.3.2 Ensembles for Subtask 2

For subtask 2, we implemented nine ensemble methods:

Top one We choose the best model with the best macro-F1 in the development dataset.

Top n average We picked n models based on the score in the development dataset. We take average of the output probabilities (i.e., score after the sigmoid). We adopted ranking by (1) F1 score with $n = 3$, (2) average precision score with $n = 3$, (3) ROC-AUC score with $n = 3$, (4) F1 score with $n = 5$, (5) average precision score with $n = 5$, and (6) ROC-AUC score with $n = 5$.

Bootstrap bagging Same as subtask 1 but we optimize F1 score.

Stacking ensemble We fit lasso regression classifier on the development dataset ($C = 1.0$), regarding probability from each model as a feature.

Hyperparameter	Values
Base model	RoBERTa large, Stage I models
Dataset	Official dataset
Classwise training	Yes, No
Max steps	
if PLM	100, 150, 200
if Stage I w/ classwise	80, 100, 120, 140
if Stage I w/o classwise	80, 100, 120, 140, 180
Learning rate	
if for PLM	20, 15, 12, 10, 8, 6 ($\times 10^{-6}$)
if Stage I	15, 12, 10, 8, 6, 4, 1 ($\times 10^{-6}$)
Batch size	16, 32
Weight decay	0.02, 0.01, 0.001
Loss scale threshold	N/A, 5, 10,000
Gradient clipping	1.0
Warmup ratio	0.2

Table 13: Hyperparamter search space for Stage II training of English subtask 2

Hyperparameter	Values
Base model	XLM-RoBERTa large, RemBERT, Stage I models
Dataset	Official dataset
Classwise training	Yes, No
Max steps	
if PLM	160, 200, 240, 280
if Stage I w/ classwise	80, 100, 120, 140
if Stage I w/o classwise	80, 100, 120, 140, 180
Learning rate	
if for PLM	15, 12, 10, 8, 6, 4 ($\times 10^{-6}$)
if Stage I	15, 12, 10, 8, 6, 4, 1 ($\times 10^{-6}$)
Batch size	16, 32
Weight decay	0.02, 0.01, 0.001
Loss scale threshold	N/A, 5, 10,000
Gradient clipping	1.0
Warmup ratio	0.2

Table 14: Hyperparamter search space for Stage II training of German, French, Italian, Polish and Russian subtask 2

We manually chose the best ensemble type for each language-*label* pair in the same way as subtask 1.