SinaAI at SemEval-2023 Task 3: A Multilingual Transformer Language Model-based Approach for the Detection of News Genre, Framing and Persuasion Techniques

Aryan Sadeghi[◇], **Reza Alipour**[◇], **Kamyar Taeb**[◇],
Parimehr Morassafar[◇], **Nima Salemahim**[◇], **Ehsaneddin Asgari**[§]
[◇] DH-NLP Lab, Computer Engineering Department, Sharif University of Technology
[§] AI Innovation Center, Data:Lab, Volkswagen AG, Munich, Germany

Abstract

This paper describes SinaAI's participation in SemEval-2023 Task 3, which involves detecting propaganda in news articles across multiple languages. The task comprises three subtasks: (i) genre detection, (ii) news framing, and (iii) persuasion technique identification. The employed dataset includes news articles in nine languages and domains, including English, French, Italian, German, Polish, Russian, Georgian, Greek, and Spanish, with labeled instances of news framing, genre, and persuasion techniques. Our approach combines fine-tuning multilingual language models such as XLM, LaBSE, and mBERT with data augmentation techniques. Our experimental results show that XLM outperforms other models in terms of F1-Micro in and F1-Macro, and the ensemble of XLM and LaBSE achieved the best performance. Our study highlights the effectiveness of multilingual sentence embedding models in multilingual propaganda detection. Our models achieved highest score for two languages (Greek and Italian) in sub-task 1 and one language (Russian) for sub-task 2.

1 Introduction

SemEval 2023 Task 3 aims to evaluate the ability of computational models to detect three important aspects of news articles: genre, news framing, and persuasion techniques. The provided dataset includes news articles from various sources, such as newspapers and online media, which have been annotated with labels indicating the framing, genre, and persuasion technique. Participants are required to submit models that can accurately identify these labels for each article. This task is significant as it provides a means to automatically analyze and classify news articles, which can be useful in various applications such as content filtering and recommendation systems (Park, 2019; Gena et al., 2019).

1. The genre detection sub-task involves identifying the genre of a news article, such as whether it is an opinion piece, a news report, or a satire piece. The task involves a large dataset of texts in different genres and their corresponding genre labels, divided into training, development, and test sets. The evaluation metric is Macro F1-score. Earlier research has tackled the challenge of automatically detecting satire in English using handcrafted features. As demonstrated by Burfoot and Baldwin (2009), one such feature involves checking the accuracy of entity mentions within the context, and McHardy et al. (2019) proposes an adversarial training method to improve the accuracy of satire detection in news articles. The method addresses the challenge of confounding variables that can influence the classification of a text as satire or non-satire.

2. The news framing sub-task focuses on identifying how a news article frames a topic, such as by emphasizing certain aspects or presenting it from a particular perspective. The evaluation metric is Micro F1-score. The Media Frames Corpus, introduced by Card et al. (2015), initially introduced news framing to computational linguistics. This corpus focused on three topics, including immigration, smoking, and same-sex marriage, and asked annotators to recognize any of the 15 framing dimensions present in Boydstun et al. (2019). In this regard, Some methods emerged to identify the frames within news headlines; for instance, in Liu et al. (2019), the authors propose an effective way of identifying frames within news headlines using the BERT Devlin et al. (2018) model. However, their method only identifies one single frame. Therefore, in Akyürek et al. (2021), the authors aimed to enhance the approach by fine-tuning MultiBERT, a multilingual version of pretrained BERT, and developed a new approach for news framing analysis that is multi-label and multilingual and introduced a new dataset for this pur-

[◇]Equal contribution, listed randomly.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2168–2173
July 13-14, 2023 ©2023 Association for Computational Linguistics

pose. Their system uses a deep learning model that achieved state-of-the-art performance on the task of news framing analysis in different languages. The paper contributes to the multilingual and cross-cultural analysis of news framing and can help media practitioners seeking to understand how news articles are framed in multiple languages. Overall, their work has important implications for improving our understanding of how news is framed in different languages and cultures.

3. The persuasion technique sub-task involves identifying the techniques used in a news article to persuade the reader, such as by appealing to emotions or using rhetorical devices. The evaluation metric is Micro F1-score. This subtask is related to SemEval-2021 Task 6 on Detection of Persuasion Techniques in Texts and Images by Zampieri et al. (2021), which focuses on news articles. They described the various approaches and techniques used by the participating teams to tackle the task, including rule-based methods, machine learning models, and deep learning models. The article also discusses the potential applications and implications of the task, including its relevance for improving natural language understanding and the detection of fake news and propaganda. Gupta and Sharma (2021) used a RoBERTa-based model that leverages pre-trained language models and fine-tuning to detect persuasion techniques in the textual content. They also used data augmentation techniques to increase training data and improve the generalization ability of their model.

A better understanding of how news articles are framed, what genre they belong to, and what persuasion techniques are used can improve our ability to analyze and interpret news content. This has important implications for the field of media studies and society as a whole, as it can help us better understand the role of the media in shaping public opinion Piskorski et al. (2023).

We approached the news article classification task by fine-tuning multilingual language models such as XLM, LaBSE, and mBERT, using ensembling and data augmentation techniques. Our experimental results show that XLM outperforms the other models in terms of F1-Micro and F1-Macro. Moreover, the ensemble of XLM and LaBSE achieved the best performance. Our models achieved the highest score in two languages (Greek and Italian) for sub-task 1, and in one language (Russian) for sub-task 2.

2 Approach

2.1 Models

Given the multilingual nature of the task, this study mainly focuses on supervised fine-tuning of pre-trained multilingual language models using news articles. This approach allows us to perform reasonably well on other languages with only supervised training on a few languages, even for unseen labeled instances. In the following section, we provide a summary of the pre-trained language models that we used in this study.

Multilingual BERT (mBERT): Multilingual BERT is a pre-trained language model that can be fine-tuned for various natural language processing (NLP) tasks in multiple languages. It is based on the popular BERT architecture and trained on a large corpus of text from 104 languages. Its strength lies in its ability to handle multiple languagesWang et al. (2019).

XLM is used for cross-lingual news classification, where the goal is to classify news articles in multiple languages into predefined categories. XLM-R was created to overcome language barriers and promote cross-lingual communication. It utilizes large volumes of multilingual news data. The XLM-R is able to transfer knowledge between languages, allowing it to comprehend and produce text in unfamiliar languages Conneau and Lample (2019); Conneau et al. (2020).

Language-agnostic BERT Sentence Embedding (LaBSE) is a transformer-based model that can encode sentence meanings in a language-independent manner. LaBSE is trained on parallel sentences and produces shared sentence-level embeddings for more than 100 languages Feng et al. (2020).

2.2 System Overview

In this section, we describe our models for the three sub-tasks, which are considered a multi-class problem for sub-task 1 and a multi-label problem for sub-tasks 2 and 3. We mainly use the mBERT, XLM-R, and LaBSE multilingual models. LaBSE and XLM-R are trained on a relatively larger dataset. XLM-R and mBERT have shown great strength in handling complex tasks Conneau and Lample (2019); Feng et al. (2020); Wang et al. (2019).

Data cleaning: We use the multilingual SpaCy model[1] for the text cleaning step, which involves tokenizing the text, removing stop words and punctuation. We keep the newline characters in the articles to maintain their structure.

Data augmentation: Since the datasets for subtasks 1 and 2 are not sufficient for proper language model fine-tuning, we perform data augmentation by creating synthetic news through sentence resampling from the original documents. Specifically, we select 30% of the sentences from each document to create new ones, and this selection is done with replacement five times for each news article.

2.3 Dataset

The SemEval 2023 Task 3 dataset contains news data from nine different languages, of which six have a train set, validation (dev) set, and test set, while the remaining three only have a test set. The languages with train and validation sets are English, French, Italian, German, Polish, and Russian. Georgian, Greek, and Spanish only have a test set. The average number of sentences per article is approximately 30 for the train, dev, and test sets. The number of labels and languages for each subtask is shown in Table 1.

Table 1: Number of labels and languages for each subtasl

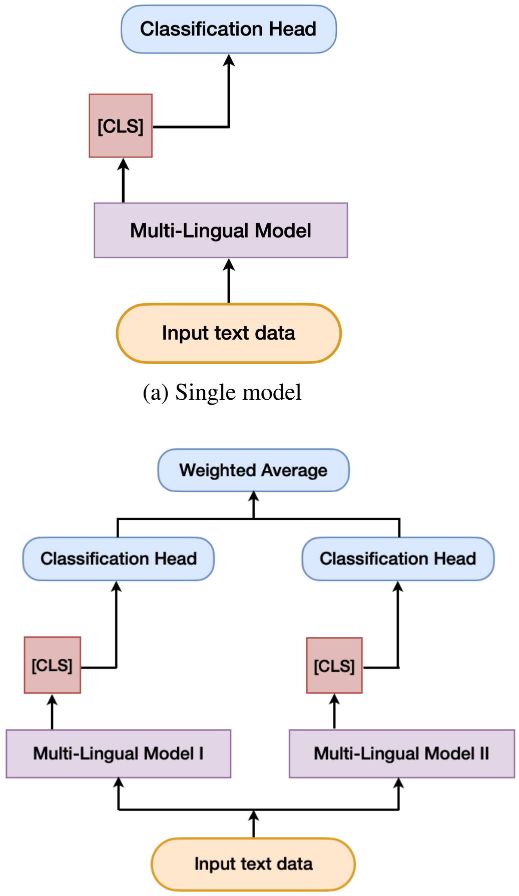
Subtask	# Labels	# Languages
1. Genre detection	3	9
2. News framing	14	9
3. Persuasion tech.	23	9

2.4 Multilingual News Classification sub-tasks

As mentioned before, the task has three sub-tasks. We provide two different approaches: using a single multilingual model with a classifier head or using two multilingual models with a classifier head for each and ensembling the output, Figure 1.

2.4.1 Single model

The single language model-mode is illustrated in Figure 1a (a). We fine-tune the XLM-RoBERTa-Large (XRL) Li et al. (2021) and LaBSE model (only for sub-task 1) for the sub-tasks separately. We use two linear layers before the output layer in the classification head. Between the two linear layers, we use dropout and Tanh activation functions, and there is a dropout between the first linear layer and the multilingual (XLR or LaBSE) model.



(a) Single model

(b) Ensemble model

Figure 1: **Single model (a)** uses only one multilingual model to extract features from the text, and the classification head is fine-tuned in a supervised setting to predict the label. **Ensemble model (b)**, on the other hand, benefits from ensembling of two multilingual models, where their outputs are combined in a weighted average scenario.

2.4.2 Ensemble

The ensemble model is illustrated in Figure 1b. Two multilingual models extract features from the same input text. Then, two different classifier heads transform the features into output labels, and the final output is computed as the weighted average of these two logits. The average weights are tuned in the model training process to create an ensemble between the two language models. The normalized weight parameter (α) is multiplied by the output of the first classification head, and $1 - \alpha$ is multiplied by the output of the other one to calculate the weighted average of both model outputs. We use combinations of the XLM model with both mBERT and LaBSE.

[1] https://spacy.io/

Table 2: The performance results of the proposed models in sub-task 1: genre classification, over the validation set.

Model	Augmentation	F1-Macro	F1-Micro
LaBSE + Classifier	N	0.59	0.63
XRL + Classifier	N	0.62	0.68
XRL + Classifier	Y	0.65	0.72
XRL + LaBSE + Classifier	N	0.56	0.67
XRL + LaBSE + Classifier	Y	0.59	0.69
XRL + mBERT + Classifier	N	0.58	0.70
XRL + mBERT + Classifier	Y	0.63	0.71

Table 3: The performance results of the proposed models in sub-task 2: news framing over the validation set.

Model	Augmentation	F1-Macro	F1-Micro
XRL + Classifier	N	0.60	0.52
XRL + mBERT + Classifier	Y	0.50	0.59
XRL + LaBSE + Classifier	Y	0.54	0.61
XRL + LaBSE + Classifier	N	0.55	0.61

Table 4: The performance results of the proposed models in sub-task 3: persuasion detection, over the validation set.

Model	Augmentation	F1-Macro	F1-Micro
XRL + Classifier	N	0.13	0.40

Experimental Setup

We explored different language model instances for both single and ensemble models in sub-tasks 1 and 2. However, due to time and resource limitations, we only used the single method for sub-task 3. We chose the best-performing model in sub-tasks 1 and 2 and used their associated hyperparameters for fine-tuning in task 3.

We divided the given dev set into validation and test sets to use the dev set for tuning the hyperparameters. To obtain separate test and validation sets for each language, we randomly split the dev dataset, allocating 50% of each language's samples to the test and dev sets. Weighted cross-entropy was used as the loss function, and AdamW Loshchilov and Hutter (2019) was employed as the optimizer with a linear learning rate scheduler that decreased the learning rate through training epochs. The models were trained for 30 epochs in sub-tasks 1 and 2, and 5 epochs in sub-task 3, with a learning rate set to 1e-5 and a batch size of 3 for a single model and 1 for the ensemble model. Finally, the best model was saved and used for prediction on the test set.

Evaluation metrics: F1-Micro and F1-Macro are measured and reported for all three tasks. In sub-task 1, F1-Macro was used as an official metric, and in sub-tasks 2 and 3 F1-Micro was chosen as an official metric.

Table 5: Official leaderboard and scores for subtask 1: genre detection. The results presented in the table were generated using XRL + Classifier model without data augmentation.

Language	Test F1-Macro	Place
Greek	0.805	1
German	0.782	3
Polish	0.663	6
French	0.637	8
English	0.506	9
Italian	0.502	9
Russian	0.442	10
Georgian	0.467	11
Spanish	0.322	11

3 Results

The results of the models for sub-tasks 1, 2, and 3 on the validation set are presented in Tables 2, 3, and 4, respectively. In sub-task 1, the XRL model with data augmentation achieved the highest Macro-F1 of 0.65 and a Micro-F1 of 0.72. In sub-task 2, the ensemble model of XRL and LaBSE outperformed other models based on the validation set. In sub-task 3, the XLM single model achieved

Table 6: The performance and post leaderbord of using XRL + LaBSE + Classifier (XLC) and XRL + mBERT + Classifier (XmC) models in sub-task 2 over the test set for each languages. All results are based on F1-Micro.

Language	XLC with Aug.	XLC w/o Aug.	XmC with Aug.	Place
Russian	0.447	**0.456**	0.384	1
Italian	0.605	**0.606**	0.584	2
Georgian	0.564	0.593	**0.629**	3
Spanish	0.493	0.493	**0.554**	3
German	0.611	0.638	**0.650**	4
Greek	0.463	**0.537**	0.505	4
Polish	0.564	**0.640**	0.634	5
French	0.458	**0.472**	0.453	10
English	0.493	0.471	**0.506**	11

Table 7: Post leaderboard and scores for subtask 1. The data in the table was obtained using XRL + Classifier model with augmentation.

Language	Test F1-Macro	Place
Greek	0.826	1
Italian	0.782	1
German	0.782	3
French	0.767	3
Georgian	0.857	4
Russian	0.655	5
Polish	0.663	6
Spanish	0.531	6
English	0.506	9

Table 8: Post leaderboard and scores for subtask 3, persuasion technique detection.

Language	Test F1-Micro	Place
Georgian	0.432	2
Greek	0.234	7
Spanish	0.296	8
Polish	0.358	9
German	0.492	10
Italian	0.477	10
Russian	0.303	11
French	0.343	13
English	0.268	18

a Micro-F1 of 0.60 and a Macro-F1 of 0.13.

The performance of the models in all sub-tasks over the official test set is reported in Tables 5, 6, 7, and 8. However, the official results of sub-tasks 2 and 3 were not reported due to a minor error in output formatting at the submission time. In sub-task 1, XRL with augmentation achieved the best score among all teams for two languages, Greek and Italian. In sub-task 2, the performance of three ensemble models was reported since their performance on the validation set was close to each other. The ensemble of XRL and LaBSE without augmentation outperformed other models in the Russian language.

4 Discussion and Conclusions

SemEval 2023 subtask 3 aimed to classify news articles in multiple languages in three different sub-tasks: news framing, genre detection, and persuasion technique identification. Our dataset comprised news articles from various domains and nine different languages, with labeled instances of news framing, genre, and persuasion techniques. We showed that XLM outperformed the other language models in F1-Micro and F1-Macro scores in sub-task 1 when used alone but performed even better when combined with mBERT or LaBSE in an ensemble approach in task 2. Consequently, we used the most effective models to train for different tasks, analyzed their performance across languages and propaganda techniques, and explored data augmentation techniques to improve their performances. Our study highlights the effectiveness of multilingual sentence embedding models in detecting multilingual propaganda and provides crucial insights into different models and data augmentation techniques. However, due to limited computational resources, we had to limit the number of experiments we ran for hyper-parameter optimization which may have affected our performances.

References

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry T. Wijaya. 2021. Multi-label and multilingual news framing analysis.

In *Proceedings of the 3rd Workshop on Fact Extraction and Verification (FEVER)*.

Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2019. Tracking the development of media frames within and across policy issues.

Clinton Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 161–164. The Association for Computer Linguistics.

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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*. The Association for Computer 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *CoRR*, abs/2007.01852.

Cristina Gena, Pierluigi Grillo, Antonio Lieto, Claudio Mattutino, and Fabiana Vernero. 2019. When personalization is not an option: an in-the-wild study on persuasive news recommendation. *Information*, 10(10):300.

Vansh Gupta and Raksha Sharma. 2021. Nlpiitr at semeval-2021 task 6: Roberta model with data augmentation for persuasion techniques detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1061–1067.

Jialu Li, Hao Tan, and Mohit Bansal. 2021. Improving cross-modal alignment in vision language navigation via syntactic information. *CoRR*, abs/2104.09580.

Siyi Liu, Lei Guo, Kate K. Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. 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 2019, Hong Kong, China, November 3-4, 2019*. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 660–665. Association for Computational Linguistics.

Chang Sup Park. 2019. Does too much news on social media discourage news seeking? mediating role of news efficacy between perceived news overload and news avoidance on social media. *Social Media+ Society*, 5(3):2056305119872956.

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.

Alex Wang, David Yang, Yu Wei, Weizhu Liu, Xiaodong Chen, and Haisheng Li. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Maram Hasanain, Saif Mohammad Hasan, Giovanni Da San Martino, Hamdy Mubarak, and Chris van der Lee. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, pages 376–391.