MELODI at SemEval-2023 Task 3: In-domain Pre-training for Low-resource Classification of News Articles

Nicolas Devatine, Philippe Muller, Chloé Braud

MELODI at SemEval-2023 Task 3: In-domain Pre-training for Low-resource Classification of News Articles

Nicolas Devatine[1], **Philippe Muller**[1,3], **Chloé Braud**[2,3]
[1]IRIT, University of Toulouse
[2]IRIT, CNRS
[3]Artificial and Natural Intelligence Toulouse Institute (ANITI)

firstname.lastname@irit.fr

Abstract

This paper describes our approach to Subtask 1 "News Genre Categorization" of SemEval-2023 Task 3 "Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup", which aims to determine whether a given news article is an opinion piece, an objective report, or satirical. We fine-tuned the domain-specific language model POLITICS, which was pre-trained on a large-scale dataset of more than 3.6M English political news articles following ideology-driven pre-training objectives. In order to use it in the multilingual setup of the task, we added as a pre-processing step the translation of all documents into English. Our system ranked among the top systems overall in most language, and ranked 1st on the English dataset.

1 Introduction

The *News Genre Categorization* subtask of SemEval-2023 Task 3 (Piskorski et al., 2023) is a classification task that aims to identify the genre of a news article by considering three classes: opinion, reporting, satire. News genre categorization is important because it helps readers to quickly identify the type of content they are reading. It also helps to ensure that readers are not misled by the content they are consuming. The organizers provided articles in six languages for the training phase (English, French, German, Italian, Polish, and Russian) and three surprise languages were revealed for the evaluation on the test sets (Spanish, Greek and Georgian). It is therefore a multilingual task that includes small datasets during training and zero-shot classification for surprise languages.

Our main strategy relies on the fine-tuning of POLITICS (Liu et al., 2022), a state-of-the-art language model for political ideology prediction and stance detection, pre-trained on more than 3.6M English news articles. Based on this, a wide variety of strategies were tested, including a sliding window to cover long documents, pooling methods, adapted loss functions, data splitting and hyperparameters tuning. Moreover, as POLITICS is monolingual, and to take advantage of the large-scale pre-training in the multilingual setup, we propose as a pre-processing step to translate all articles from other languages into English. Other interesting strategies have also been considered and tested but have been less successful on the development set, including inherently multilingual models such as XLM-RoBERTa (Conneau et al., 2020) or long-sequence transformers (Beltagy et al., 2020).

We ranked 1st on the English test set with a Macro-F_1 score of 78.43 (+16.8 points above the second ranked system), but were less successful on the other languages with an average of 6th place (out of about 20 systems, depending on languages). We draw several observations from our participation in this task:

- the dominance of large-scale pre-trained and domain-specific language models over more adapted but not massively pre-trained approaches;
- the need to further develop multilingual language models and datasets rather than having to rely on low performing translation systems for complex tasks;
- the importance of considering the whole text in the case of long documents, whereas most language models are limited to the first 512 subtokens.

2 Background

News genre categorization is a multi-class, single-label classification problem that aims to predict the type of content of a news article according to 3 classes (Piskorski et al., 2023):

Opinion An article is an opinion piece when it expresses the writer's opinion on a topic. It is usually written in a persuasive style and is meant to influence the reader's opinion on the subject.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 108–113

Reporting News reporting is the genre that focuses on providing factual information. It aims to inform readers about the world around them and to provide an objective account of events.

Satire Satire is a form of writing that uses exaggeration, absurdity and obscenity to mock and ridicule people, organizations or events. It is not meant to deceive, but to expose and criticize behavior that is wrong or immoral. Satirical pieces often mimic real articles, using irony to provide humor.

2.1 Datasets

The input data for this task are news articles in plain text format, with their title being on the first line. During the first stage, a training split and a development split were provided for each of the 6 known languages: English (*en*), French (*fr*), German (*ge*), Italian (*it*), Polish (*po*), and Russian (*ru*). The annotations were given only for the training sets at first, then also for the development sets when the test sets were released. For the evaluation on the test sets, 3 surprise languages were revealed which involves zero-shot classification (no train or dev sets): Spanish (*es*), Greek (*gr*) and Georgian (*ka*). Given the limited number of news articles available per language, this is a kind of few-shot learning task. This problem is also characterized by a strong class imbalance, especially for the *satire* genre which is represented only a few dozen times across the datasets. Dataset statistics are summarized in Table 1. The distributions of classes vary a lot between corpora, between train, development, and probably test sets, given the observed performance differences between dev and test. This means there was a (blind) domain-adaptation aspect to the task.

2.2 Related Work

Substantial efforts have been made to address the problem of media analysis on a multitude of aspects such as bias analysis (Hamborg et al., 2019), summarization (Eyal et al., 2019), text categorization (Pérez-Rosas et al., 2018; Karimi and Tang, 2019) or propaganda techniques (Da San Martino et al., 2020). Recently, Liu et al. (2022) proposed POLITICS, a pre-trained language model over RoBERTa (Liu et al., 2019), fine-tuned on a large corpus of English news to address both political ideology prediction and stance detection. Their approach has proven to be the most efficient on a wide range of well known datasets for these tasks such as

	#Train	#Dev.	#Test.	#Total
en	433 (382;41;10)	83 (20;54;9)	54	570
fr	157 (103;43;11)	54 (35;15;4)	50	261
ge	132 (86;27;19)	45 (29;9;7)	50	227
it	226 (174;44;8)	77 (59;15;3)	61	364
po	144 (104;25;15)	50 (35;9;6)	47	241
ru	142 (93;41;8)	49 (32;14;3)	72	263
es	–	–	30	30
gr	–	–	64	64
ka	–	–	29	29
#Total	1234 (942;221;71)	358 (210;116;32)	457	2049

Table 1: Number of documents for each dataset. The class distribution is given in parentheses (#opinion;#reporting;#satire).

`Hyperpartisan` (Kiesel et al., 2019), `Allsides`[1] (Baly et al., 2020) or `BASIL` (Fan et al., 2019).

3 System Overview

Our main strategy is based on Liu et al. (2022)'s domain-specific pre-trained language model POLITICS which has recently demonstrated great success on a variety of news articles ideology and stance prediction datasets. It has several advantages for this task: (i) it was massively pre-trained on more than 3.6M English news articles, (ii) it relies on the comparison of articles on the same story written by media of different ideologies, (iii) it demonstrated its robustness in few-shot learning scenarios. Although predicting political ideology is not the same as detecting the genre of an article, our assumption is that these two notions overlap as in both cases there is a linguistic shift in the way information is conveyed. We can also observe similarities between the genre *reporting* and articles from the *Center* political class, or between the genre *opinion* and *Left* or *Right* leaning articles.

[1] `https://www.allsides.com`

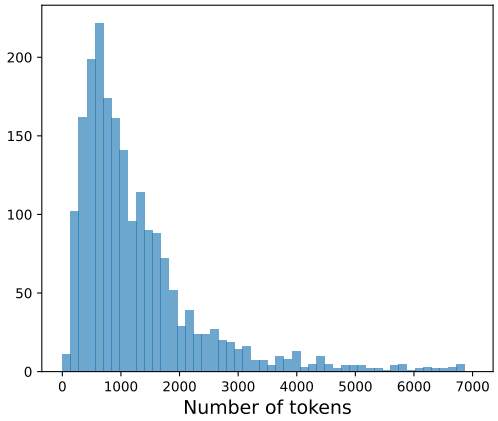

Figure 1: Distribution of the number of (POLITICS) tokens per article for the combination of all datasets.

Hyperparameter	
#Epochs	10
Learning Rate	$1e-5$ ($2e-5$)
Batch size	4
Loss Function	Cross Entropy
Optimizer	AdamW
Weight Decay	0.001
Classifier #Layers	2
Classifier Hidden Dim.	768
Classifier Dropout	0.2 (0.1)
Sliding window size	512
Sliding window overlap	64

Table 2: Hyperparameters used to fine-tune POLITICS and XLM-RoBERTa and Longformer. For Longformer, the values that differ are in parentheses and there is no sliding window.

3.1 Data Translation

POLITICS was trained solely on articles from English-language media, whereas in the multilingual configuration proposed for this task, we have to classify articles in 9 languages, 3 of which were unknown during training. Because of this strong constraint and the impossibility of re-training POLITICS in a multilingual configuration, we decided to resort to translation into English. By translating all texts into English, we end up with an augmented English training set that can be used with POLITICS. This solution represents an additional cost due to the pre-processing step, and a loss of information that depends on the quality of the translation system. Several translation models were compared based on performance, language coverage, and accessibility, including GoogleTranslate,[2] DeepL[3] and OPUS-MT (Tiedemann and Thottingal, 2020), resulting in choosing GoogleTranslate as the most appropriate one. DeepL was the best performing system but had accessibility limitations due to its pricing for handling large amounts of data, we had to fall back on GoogleTranslate which was the second best performing and freely available system.

3.2 Fine-tuning

For the classification model, we fine-tuned POLITICS on the English dataset augmented with translations. It works by adding a 2-layer perceptron on top of the standard `[CLS]` classification token representation to predict the distribution over class labels. POLITICS is based on RoBERTa and, like most language models, is limited with respect to the size of the input text, here at most 512 subtokens. News articles are on average much longer (Figure 1), thus truncating the first 512 subtokens would result in a significant loss of information. Rather than truncating the text, we use a sliding window of size 512 with an overlap of size 64, and aggregated the information by mean pooling. That is, we encode the first 512 tokens of the document, then the next 512 with an overlap of 64, and so on, until we reach the end of the document, then we calculate the mean of all these representations to obtain the final representation of the document.

3.3 Alternative Approaches

Other approaches of interest for this task were also considered, but proved to be less successful during the training phase on the development split. In particular, on English, POLITICS showed much higher performance than the other models considered and was on average better on the other languages. Thus, we have favored this model for the evaluation although a posteriori we can see that it is less efficient in certain cases (e.g. on surprise languages or on French/Russian, Table 3).

XLM-RoBERTa (Conneau et al., 2020) The multilingual version of RoBERTa (Liu et al., 2019), a state-of-the-art model in natural language processing, pre-trained on 2.5TB of CommonCrawl data containing 100 languages and including the 9 ones covered by this task. Similar to POLITICS, we used a sliding window to consider the entire

[2] https://translate.google.com
[3] https://www.deepl.com/

Model	*en*	*fr*	*ge*	*it*	*po*	*ru*	Avg	*es**	*gr**	*ka**	Avg*
Main strategy											
POLITICS	1st	7th	4th	4th	4th	5th	4th	5th	7th	9th	7th
	78.43	65.59	**77.88**	**58.66**	70.85	58.64	68.34	44.25	63.65	49.00	52.30
Baseline											
Bag-Of-Words+SVM	28.80	56.80	62.96	38.94	48.96	39.83	46.04	15.38	17.05	25.64	19.35
Control experiments											
POLITICS-512	69.43	**74.24**	76.98	53.97	66.34	58.51	66.57	48.42	**71.04**	**78.67**	**66.04**
Alt. approaches											
XLM-RoBERTa	59.42	72.60	68.05	57.05	**79.79**	57.64	65.75	40.98	60.71	51.44	51.04
Longformer-4096	66.51	73.82	75.62	57.69	75.56	**70.57**	**69.96**	**54.81**	56.66	55.84	55.77

Table 3: Macro-F_1 scores on the test sets for each language and different approaches. * indicates surprise languages (zero-shot). "512" means that only the first 512 subtokens of the inputs (no sliding window) were used to train the model. All results, except for the main strategy, were obtained when the submission platform reopened after the official submission deadline. We added the rank of the main system with respect to that score, according to the leaderboard published by the organisers. Note that the baseline and XLM-RoBERTa are the only models that have been trained on the original data (not translated).

article.

Longformer (Beltagy et al., 2020) Transformer-based model designed to process longer sequences, up to 4096 tokens, making it suitable for news article classification, and pre-trained on a variety of large-scale generic datasets, including Wikipedia, BooksCorpus, and CommonCrawl. No sliding window was introduced with this model, as the input length covers the vast majority of texts, cf. Figure 1.

4 Experimental Setup

First, the train, dev and test split articles for each language are translated using Google Translate. Next, we aggregate the train and dev splits for each language to create an augmented English dataset containing 1234 articles in the training split and 358 in the development split, on which we train a single English model.

We built on Liu et al. (2022)'s implementation,[4] and the HuggingFace transformers library (Wolf et al., 2020). Hyperparameters were set using grid search on the augmented development split for each of the models considered (Table 2). The hyperparameters for POLITICS and XLM-RoBERTa were found to be the same. We also conducted a control experiment on the sliding window to assess its effectiveness by evaluating the same model trained only on the first 512 tokens of the articles. Control experiments and alternative approaches (see table 3 were performed on the same set of hyperparameters. Since there is a strong class imbalance in the training set, we weighted the cross entropy on the class distribution, which turned out to give us better performance.

Regarding alternative approaches, we fine-tuned `xlm-roberta-large`[5] using a sliding window on the same set of hyperparameters as POLITICS (Table 2). Longformer was trained using `longformer-base-4096`[6] and the hyperparameters given in Table 2.

The official evaluation measure used is Macro-F_1. Micro-F_1 is also given as a secondary evaluation measure on the official leaderboard.

5 Results

Table 3 summarizes the results obtained by our main strategy, the control experiments, the baseline proposed by the task organizers and the alternative approaches. Our main strategy (POLITICS) obtains the best results for 3 of the 9 languages, with a special distinction for English on which it particularly stands out (+9 points than the second best approach we tested). This shows the benefits of large scale in-domain pre-training on English news articles, but also the limitations of translation. For French, the score is surprisingly low, which was not as pronounced in our experiments on the development split. From the control experiment, we can confirm the interest of the sliding window

[4] https://github.com/launchnlp/POLITICS

[5] https://huggingface.co/xlm-roberta-large

[6] https://huggingface.co/allenai/longformer-base-4096

rather than just truncating the article, with important gains on most training languages. Interestingly, for the surprise languages, removing the sliding window leads to much better results, which shows that in a zero-shot context, introducing too much unseen language-specific information confuses the model. Regarding the alternative approaches, the results obtained by Longformer confirm the previous observations on the importance of considering the whole article, with results close to or even better than POLITICS outside English and without any in-domain pre-training. Furthermore, the XLM-RoBERTa multilingual model performs well against POLITICS without the need for translation, but the lack of in-domain pre-training results in a significant performance drop, especially for English.

These results should be taken with a grain of salt for the following reasons: test sets are small (30-70 instances) hence the huge variance between models and across languages. This problem is compounded by the chosen evaluation : Macro-F_1 scores equalize the contributions of all classes to the final scores, meaning that one instance from the minority class classified one way or the other could swing the evaluation disproportionately. It would probably be informative to evaluate accuracy, and per class metrics, but gold labels were not provided for the test sets. This also makes it hard to estimate the distribution shift between train, development and test sets, although it is already apparent there are large differences between train and dev sets.

6 Conclusion

We propose to take advantage of in-domain pre-training for detecting the genre of news articles. We fine-tuned the large-scale English language model POLITICS using sliding windows on the multilingual corpus translated into English. Our various experiments have shown the effectiveness of in-domain pre-training as well as the importance of approaches adapted to the processing of long documents. We ranked 1st on the English dataset while being among the top systems overall.

Acknowledgements

Nicolas Devatine's work is supported by the SLANT project (ANR-19-CE23-0022). This work was partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program. This work is also partially supported by the AnDiaMO project (ANR-21-CE23-0020). Chloé Braud and Philippe Muller are part of the programme DesCartes and are also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. 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 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news ar-

ticles: an interdisciplinary literature review. *Int. J. Digit. Libr.*, 20(4):391–415.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. 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 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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. *CoRR*, abs/1907.11692.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. 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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

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: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.