SPARTA at CASE 2021 Task 1: Evaluating Different Techniques to Improve Event Extraction

Arthur Müller
University of the Bundeswehr Munich
85577 Neubiberg, Germany
arthur.mueller@unibw.de

Andreas Dafnos
University of the Bundeswehr Munich
85577 Neubiberg, Germany
andreas.dafnos@unibw.de

Abstract

We participated in the Shared Task 1 at CASE 2021, Subtask 4 on protest event extraction from news articles (Hürriyetoğlu et al., 2022) and examined different techniques aimed at improving the performance of the winning system from the last competition round (Hürriyetoğlu et al., 2021). We evaluated in-domain pre-training, task-specific pre-fine-tuning, alternative loss function, translation of the English training dataset into other target languages (i.e., Portuguese, Spanish, and Hindi) for the token classification task, and a simple data augmentation technique by random sentence reordering. This paper summarizes the results, showing that random sentence reordering leads to a consistent improvement of the model performance.

1 Introduction

The generation of protest event datasets over the last decades has allowed social movement scholars to study the dynamics and evolution of collective action in contemporary societies. The collection of relevant events is usually based on the systematic, manual analysis of news articles, which provide information about the variables of interest such as the location, date, and main protagonists of protest demonstrations (Hutter, 2014).

It has been noted, however, that the manual coding of news articles is time and labor-consuming, and, as a result, comparative and longitudinal studies that rely on multiple news sources may not be feasible (Lorenzini et al., 2022). Recent work on approaches that automatically retrieve protest information is promising and may address this challenge.

CASE 2021 Task 1: Multilingual protest news detection (Hürriyetoğlu et al., 2021) constitutes a collaborative project that attempts to map the features of political contention through the automated analysis of news articles at different data levels. We participate in Subtask 4, which focuses on identifying event triggers and their arguments and involves detecting protest events in three languages: English, Portuguese, and Spanish.

The paper proceeds as follows: Section 2 discusses related work in the field of computational social science, whereas section 3 defines the task of event extraction. Section 4 describes the architecture of our approach. Section 5 provides details about the experiments we conducted. Finally, in section 6, we summarize and discuss the results.

2 Related Work

The use of automated tools for the identification and coding of political event data spans a period of more than 30 years (Hanna, 2017), and, for this task, several methodological approaches have been developed and tested. Initial attempts to automatically parse text and produce structured data were based on the Kansas Event Data System (KEDS) (Schrodt et al., 1994), which, along with its successors programs such as TABARI (Schrodt, 2009) and PETRARCH (Norris, 2016), was designed to provide information about different types of political action and also their source and target actors.

In the field of contentious politics, that is mainly interested in the activities of social movements and protest groups, the standard approach involved for a long time the manual coding of text. However, half-automated techniques have also been introduced. For instance, Lorenzini et al. (2022) have developed several filters (e.g., a location-based filter) and document and event-trigger classifiers to select newspaper articles that contain protest-related information. In the final step of their procedure, the authors create samples of relevant articles and manually extract the features of protest events.

Taking advantage of recent advances in machine learning methods, other scholars have turned their attention to approaches that automatically detect and classify protest information. However, unlike coding systems such as KEDS and its successors programs that make use of actor and verb dictio-

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 189 - 194
December 7-8, 2022

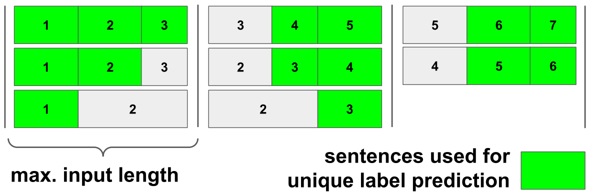

Figure 1: Sentence splitting into overlapping sequences.

naries, the new techniques primarily rely on pre-trained transformer-based language models (Liu et al., 2021), such as BERT (Devlin et al., 2018). CASE 2021 and 2022 Task 1 (Hürriyetoğlu et al., 2021, 2022) are such research projects—organized as shared tasks—that focus on the generation of multilingual protest event data and involve four subtasks: 1. Document classification; 2. Sentence classification; 3. Event sentence coreference identification; and 4. Event extraction.

In the following sections, we focus on subtask 4 and discuss techniques that improve over the baseline multilingual model XLM-RoBERTa (Conneau et al., 2019).

3 Event Extraction Task

The event extraction task consists of identifying text spans in given news article sentences and classifying them into entity types such as *trigger*, *participant*, *place* etc. Given $S = (w_1, .., w_n)$ a sentence and $T = \{t_1, .., t_m\}$ a set of entity types, the task consists of identifying spans $s = (w_b, .., w_e)$ such that $typeof(s) \in T$. This task can be reformulated as the token classification task, where IOB2 labels (Sang and Veenstra, 1999) are assigned to tokens in sentences to form spans. Hereby, the first token w_b within the span s is assigned the label B_{type} and the rest of the tokens the label I_{type}, where $type \in T$. All tokens outside of any identified spans are assigned the token O.

4 Architecture

The objective of the conducted evaluations was to show possible improvement compared to the winning system from last year's participation at CASE 2021 (Hürriyetoğlu et al., 2021) by the *IBM* team (Awasthy et al., 2021). The authors trained variants of the multilingual model XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2019) on news article sentences to predict IOB2 labels for event extraction. Therefore, all experiments in our paper used the same base model and similar training settings.

In contrast to *IBM* team's approach, we did not provide an ensemble variant of the model but relied only on a single multilingual model. Another significant architectural difference was how the inputs were provided to the model; instead of splitting the news articles into single sentences, we used the maximum possible input length of 512 tokens and fed as many full sentences as possible to the model, providing as a result more context. If the news article exceeded the maximum input length, it was split into overlapping sentence sequences as shown in Figure 1. Thus, some sentences were presented to the model multiple times during the fine-tuning procedure with different preceding or following contexts. However, the final predicted token labels during the test procedure were derived only from the reconstructed non-overlapping sequence of sentences, leading to unique predictions. In both procedures, we removed the concatenating separator token `[SEP]` from the input. We should also note that the predicted token labels correspond to the IOB2 labels.

5 Experiments

Starting from the base model, several techniques were evaluated after fine-tuning the model on the provided dataset for Subtask 4 (Hürriyetoğlu et al., 2021). Similar to the *IBM* team, we used only 10% of the English dataset as a development set. Thus, the influence of the employed techniques on other languages was mainly inferred from the testing results in the provided Codalab page. The best models for submission were selected according to the highest CoNLL F1 score and lowest mean validation loss on the development set. The best values of F1 achieved 80.06% and 80.86%. Models were fine-tuned for 20 epochs using hyperparameters as shown in Table 1. The fine-tuning was conducted on four NVIDIA A100 GPUs each with 40GB RAM leveraging the Distributed Data Parallel (DP) paradigm (Li et al., 2020).

5.1 Further Pre-Training

The current literature suggests that further pre-training of models on in-domain data can produce promising results, especially when the target language has a different—and yet unknown—token distribution for the pre-trained model. For instance, in the case of the language used on Twitter, further pre-training of the XLM-R models led to significant improvements in the task of stance detection

Parameter	Pre-Training	Fine-Tuning
Input Length	512	512
Batch Size	1280	20
AdamW_{lr}	1e-5	2e-5
$\text{AdamW}_{\text{beta}}$	(0.9, 0.999)	(0.9, 0.999)
$\text{AdamW}_{\text{eps}}$	1e-6	1e-8
Weight Decay	0	0.001
Linear Warmup	0	0.1
Dice Loss Parameter		**Fine-Tuning**
Smooth		0.5
Square Denominator		true
Using Logits		true
Ohem Ratio		0.0
Alpha		0.0
Reduction		mean
Index Label Position		true

Table 1: Parameters for pre-training and fine-tuning.

Datasets	en	es	pr	hi
Count Love	38k			
$\text{Count Love}_{\text{t}}$		38k	38k	38k
POLUSA	21k			
POLUSA_{t}		21k	21k	21k
GDELT 2.0	177k	40k	8.3k	0.5k
$\text{GDELT 2.0}_{\text{t}}$		177k	177k	177k
Sum per lang	236k	276k	244.3k	236.5k
Sum total				992.8k

Table 2: Sizes of collected, filtered, and translated datasets for further pre-training. The index *t* indicates the datasets translated from English.

(Müller et al., 2022). *NoConflict* team used further pre-training for subtasks 1 and 2 at CASE 2021 (Hu and Stöhr, 2021). It was also employed with success for the task of event extraction on a dataset that was based on online news archives from India (Caselli et al., 2021). The approach used BERT (Devlin et al., 2018) as the base model.

In this paper, our objective was to evaluate whether further pre-training on protest-specific news articles can integrate more—yet unknown—token distributions into the model. Therefore, we collected, filtered, and translated multiple datasets for four languages: English, Portuguese, Spanish, and Hindi. We used the Hindi language for pre-training, although a dataset for Hindi is not provided for subtask 4.

The Count Love dataset (Leung and Perkins, 2021) consists of semi-automated collected protest news articles in English. We used the provided crawler to recollect data and removed missing articles collecting 81,500 articles, of which ca. 38,000 were labeled as protest-related news. To filter missing articles, we used the content length of 150 characters and expressions that indicated missing or restricted web pages during the crawling process, such as *"Unfortunately, our website is currently unavailable"* and *"Please whitelist us to continue reading"*. Some web pages were not accessible due to necessary subscriptions or legal geographic restrictions. The collected English dataset was translated into Portuguese, Spanish, and Hindi using the Argos Translate library. We reused the provided labels in order to train a binary classifier based on the $\text{XLM-RoBERTa}_{\text{base}}$ (Conneau et al., 2019) and identify protest-related news for each of the four languages with an F1 score of ca. 85%, which was used to filter articles in the following datasets:

The POLUSA dataset (Gebhard and Hamborg, 2020) consists of ca. 0.9 mio political news articles in English. It was also used by the previously mentioned *NoConflict* team at CASE 2021 for Subtasks 1 and 2 (Hu and Stöhr, 2021). The authors provided us with the full dataset, and we used the previously trained binary English-based classifier to filter protest-related news; a process which resulted in ca. 21,000 articles. We translated them into the three languages mentioned above.

GDELT 2.0 Event Database is a large-scale news database that monitors different types of events in 65 languages. We downloaded the files containing links to articles beginning from February 2015 to July 2022 and filtered them to obtain protest-related news using codes 140–149 according to the CAMEO codebook. Additionally, we applied the binary classifier to filter protest-related articles. Those consisted of ca. 4% for Hindi and ca. 11% for English, Spanish, and Portuguese. Finally, we translated English texts into these three languages.

As can be seen from the overview of collected and translated dataset sizes in Table 2, even the originally multilingual GDELT dataset resulted in very low amounts of items for non-English languages. Therefore, the translation procedure we employed was driven by the idea that translated texts could create more diversity in the token distribution regarding the different ways protests are described.

Model	Loss	en	pr	es
IBM's S1	cross	75.95	<u>73.24</u>	<u>66.20</u>
PT_1	dice	75.70	**74.57**	69.08
PT_2	cross	**76.49**	73.11	69.58
$FT_{es\text{-}1}$	cross	75.72	74.45	**69.87**
$FT_{es\text{-}2}$	cross	75.28	73.33	69.35

Table 3: Summary of the best models as CoNLL F1 score. *PT* indicates models with further pre-training on the multilingual dataset. *FT* models were previously fine-tuned on the Spanish part of the CoNNL 2002 task. The loss functions *dice* and *cross* correspond to Dice Loss and Cross-Entropy. The underlined numbers are the best results from the previous competition round at CASE 2021. The bold numbers show our best values.

Model	Data	en	pr	es
$TR_{en+es+pr}$	en+pr+es +pr-pseudo +es-pseudo	75.66	67.23	62.18
TR_{es}	pr+es +es-pseudo		71.59	63.94
TR_{pr}	pr+es +pr-pseudo		69.68	66.01

Table 4: Summary of the best models as CoNLL F1 score for dataset translation. The data labels *en*, *pr*, *es* indicate the usage of original parts of the training dataset. The parts *pr-pseudo* and *es-pseudo* are translated from the English dataset into Portuguese and Spanish.

The pre-training of the base model was conducted using the full multilingual collected dataset with hyperparameters according to Table 1. It was repeated up to 7 epochs on the same but randomly ordered articles. In contrast to the fine-tuning procedure, we did not split sentences. Instead, the first 512 tokens were fed into the model, assuming that the most important information is available at the beginning of the article. All pre-trained models for each epoch and parameter combination were fine-tuned and the best model was selected for evaluation on the Codalab page. The pre-training was conducted on an NVIDIA DGX V100 machine with 16 GPUs each with 32 GB RAM. We used the Fully Shared Data Parallel (FSDP) paradigm (Baines et al., 2021). To achieve the high batch size of 1280, the technique of gradient accumulation was additionally leveraged.

5.2 Pre-Fine-Tuning on Similar Tasks

Learning similar or related tasks is known to be beneficial for model performance (Ruder, 2017). Therefore, we evaluated fine-tuned models that were trained on the Spanish part of the CoNLL 2002 dataset (Tjong Kim Sang, 2002) and are available on HuggingFace (Wolf et al., 2020):

1. *xlm-roberta-large-finetuned-conll02-spanish*
2. *MMG/xlm-roberta-large-ner-spanish*

5.3 Dice Loss Function

As an alternative to classic cross-entropy loss for fine-tuning, we used the Dice Loss (Li et al., 2019), which has been shown to be beneficial for tasks with imbalanced class distributions. This is true for token classification tasks, where most tokens are labeled using the IOB2 label O. Also, other annotated entity types are highly imbalanced in the data provided for Subtask 4.

5.4 Translating the Training Dataset

Translating the training dataset for the token classification task and transferring corresponding IOB2 labels to translated tokens has already been explored by the *Handshakes* team at CASE 2021 (Kalyan et al., 2021). Their approach was based on translating sentences word-by-word using auxiliary embedding mapping. Here we explored an alternative technique suggested for Named Entity Entity Recognition in the clinical domain (Schäfer et al., 2022). We used a trained model for Neural Machine Translation, the multilingual $BART_{50}$ model (Tang et al., 2020), to first translate the original English text into the target languages. Next, embeddings from an auxiliary model were used to map every word of the source sentence to one or multiple tokens in the translated sentence. For this task, we employed the multilingual $BERT_{base\text{-}cased}$ model (Devlin et al., 2018).

5.5 Augmentation by Sentence Reordering

Since we used sentence sequences as the input to our models, it was possible to randomly reorder them as a simple data augmentation technique. For every article with more than one sentence, we added up to three random combinations to the training fold. This technique was initially employed by default for all experiments.

6 Final Results and Discussion

The final results on testing datasets for the approaches of pre-training and pre-fine-tuning are summarized in Table 3. We compare the results to IBM's S1 multilingual model as the baseline, which was trained on the same multilingual dataset. IBM's S1 achieved the best results for Portuguese and Spanish languages in the last CASE 2021 competition. At least one of our models achieved better results for each of the three languages; however, the most pronounced difference is for Spanish—between 2.88 and 3.67 points. The further pre-trained model PT_1 and the pre-fine-tuned model $FT_{es\text{-}1}$ achieved nearly the same results for Portuguese.

The numbers indicate that conducting an expensive pre-training procedure on additional protest-related data does not have the expected boosting effect for the model performance. This suggests that the XLM-R models already integrate sufficient knowledge about the type of language used to describe protests. Comparable results can be achieved using a pre-fine-tuned model on a similar task. Furthermore, the usage of the Dice Loss does not lead either to very different results compared to the classical Cross-Entropy loss on this task.

It is important to mention that models in Table 3 were trained using the simple data augmentation technique. We argue that at least part of the performance increase was caused by this technique. To evaluate its influence, we retrained 10 models using different parameters but without augmentation, including the best models. There was a consistent increase measured on the English development set due to data augmentation on average by 0.70 points. On testing datasets, the average improvement resulted in 0.73 points for English, 1.03 for Portuguese, and 0.70 for Spanish.

Finally, we evaluated the translation technique, which resulted in performance drops. Table 4 summarizes the results of these three models. In the first model, the original dataset parts for the three languages were used, and the English part was further translated into Portuguese and Spanish. The following two models used the Portuguese and Spanish datasets and a translated part into one of these languages. Compared to IBM's S1, the performance dropped especially for those target languages in which datasets were extended by additional translated parts. Apparently, this approach introduced lots of noise. Manual evaluation of the Spanish translation showed that in many cases the conjunctions and articles within entity spans—such as *de, del, la, etc.*—were missing the appropriate labels.

7 Conclusion

In this paper, we presented the models developed for the Shared Task 1 Subtask 4 at CASE 2021. We explored different techniques to improve the baseline multilingual model. The best result was achieved by improving on the Spanish test data by 3.67 points of CoNLL F1 score over the winner of the previous competition round. Our submissions ranked 1st for Portuguese and Spanish and 2nd for English in the current competition round.

Acknowledgments

This research is funded by dtec.bw—Digitalization and Technology Research Center of the Bundeswehr [project SPARTA]. Furthermore, the authors gratefully acknowledge the computing time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at the Bundeswehr University Munich.

References

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 Task 1: Multigranular and multilingual event detection on protest news. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 138–146.

Mandeep Baines, Shruti Bhosale, Vittorio Caggiano, Naman Goyal, Siddharth Goyal, Myle Ott, Benjamin Lefaudeux, Vitaliy Liptchinsky, Mike Rabbat, Sam Sheiffer, et al. 2021. Fairscale: A general purpose modular pytorch library for high performance and large scale training.

Tommaso Caselli, Osman Mutlu, Angelo Basile, and Ali Hürriyetoğlu. 2021. PROTEST-ER: Retraining BERT for Protest Event Extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 12–19.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Lukas Gebhard and Felix Hamborg. 2020. The POLUSA dataset: 0.9 M political news articles balanced by time and outlet popularity. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 467–468.

Alex Hanna. 2017. MPEDS: Automating the generation of protest event data.

Tiancheng Hu and Niklas Werner Stöhr. 2021. Team "NoConflict" at CASE 2021 Task 1: Pretraining for Sentence-Level Protest Event Detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 152–160. Association for Computational Linguistics.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Selçuk Gürel, Benjamin Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. Extended Multilingual protest news detection - Shared Task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2022)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection-shared Task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91.

Swen Hutter. 2014. Protest event analysis and its offspring. In *Methodological practices in social movement research*, edited by Donatella Della Porta, pages 33–67. OUP Oxford.

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, pages 1–28.

Vivek Kalyan, Paul Tan, Shaun Tan, and Martin Andrews. 2021. Handshakes AI Research at CASSE 2021 Task 1: Exploring different approaches for multilingual tasks. *arXiv preprint arXiv:2110.15599*.

Tommy Leung and L Nathan Perkins. 2021. Counting Protests in News Articles: A Dataset and Semi-Automated Data Collection Pipeline. *arXiv preprint arXiv:2102.00917*.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv:1911.02855*.

Jiangwei Liu, Liangyu Min, and Xiaohong Huang. 2021. An overview of event extraction and its applications. *arXiv preprint arXiv:2111.03212*.

Jasmine Lorenzini, Hanspeter Kriesi, Peter Makarov, and Bruno Wüest. 2022. Protest event analysis: Developing a semiautomated NLP approach. *American Behavioral Scientist*, 66(5):555–577.

Arthur Müller, Jasmin Riedl, and Wiebke Drews. 2022. Real-Time Stance Detection and Issue Analysis of the 2021 German Federal Election Campaign on Twitter. In *International Conference on Electronic Government*, pages 125–146. Springer.

Clayton Norris. 2016. PETRARCH 2: PETRARCHer. *arXiv preprint arXiv:1602.07236*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. *arXiv preprint cs/9907006*.

Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. 2022. Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation with Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62.

Philip A Schrodt. 2009. TABARI: Textual analysis by augmented replacement instructions. *Dept. of Political Science, University of Kansas, Blake Hall, Version 0.7. 3B3*, pages 1–137.

Philip A Schrodt, Shannon G Davis, and Judith L Weddle. 1994. Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.