PROTEST-ER: Retraining BERT for Protest Event Extraction

Tommaso Caselli[⋆], **Osman Mutlu**[†], **Angelo Basile**[◇], **Ali Hürriyetoğlu**[†]
[⋆]University of Groningen [†]Koç University [◇]Symanto Research
t.caselli@rug.nl, angelo.basile@symanto.com
{ahurriyetoglu,omutlu}@ku.edu.tr

Abstract

We analyze the effect of further pre-training BERT with different domain specific data as an unsupervised domain adaptation strategy for event extraction. Portability of event extraction models is particularly challenging, with large performance drops affecting data on the same text genres (e.g., news). We present PROTEST-ER, a retrained BERT model for protest event extraction. PROTEST-ER outperforms a corresponding generic BERT on out-of-domain data of 8.1 points. Our best performing models reach 51.91-46.39 F1 across both domains.

1 Introduction and Problem Statement

Events, i.e., things that happen in the world or states that hold true, play a central role in human lives. It is not a simplification to claim that our lives are nothing but a constant sequence of events. Nevertheless not all events are equally relevant, especially when the focus of attention and analysis moves away from individuals and touches upon societies. In this broader context, socio-political events are of particular interest since they directly impact and affect the lives of multiple individuals at the same time. Different actors (e.g., governments, multilateral organizations, NGOs, social movements) have various interests in collecting information and conducting analyses on this type of events. This, however, is a challenging task. The increasing availability and amount of data, thanks to the growth of the Web, calls for the development of automatic solutions based on Natural Language Processing (NLP).

Besides the good level of maturity reached by NLP systems in many areas, numerous challenges are still pending. **Portability** of systems, i.e., the reuse of previously trained systems for a specific task on different datasets, is one of them and it is far from being solved (Daumé III, 2007; Plank and van Noord, 2011; Axelrod et al., 2011; Ganin and Lempitsky, 2015; Alam et al., 2018; Xie et al., 2018; Zhao et al., 2019; Ben-David et al., 2020). As such, portability is a **domain adaptation** problem. Following Ramponi and Plank (2020), we consider a domain to be a *variety* where each corpus, or dataset, can be described as a multidimensional region including notions such as topics, genres, writing styles, years of publication, socio-demographic aspects, annotation bias, among other unknown factors. Every dataset belonging to a different variety poses a domain adaptation challenge.

Unsupervised domain adaptation has a long tradition in NLP (Blitzer et al., 2006; McClosky et al., 2006; Moore and Lewis, 2010; Ganin et al., 2016; Ruder and Plank, 2017; Guo et al., 2018; Miller, 2019; Nishida et al., 2020). The availability of large pre-trained transformer-based language models (TLMs), e.g., BERT (Devlin et al., 2019), has inspired a new trend in domain adaptation, namely **domain adaptive retraining** (DAR) (Xu et al., 2019; Han and Eisenstein, 2019; Rietzler et al., 2020; Gururangan et al., 2020). The idea behind DAR is as simple as effective: first, additional textual material matching the target domain is selected, then the masked language modeling (MLM) objective is used to further train an existing TLMs. The outcome is a new TLM whose representations are shifted to better suit the target domain. Fine-tuning domain adapted TLMs results in improved performance.

This contribution applies this approach to develop a portable system for **protest event extraction**. Our unsupervised domain adaptation setting investigates two related aspects. The first concerns the impact of the data used to adapt a generic TLM to a target domain (i.e., protest events). The second targets the portability in a zero-shot scenario of a domain-adapted TLMs across protest event datasets. Our experimental results provide additional evidence that further pretraining TLM on

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 12–19

domain-related data is a "cheap" and successful method in single-source single-target unsupervised domain adaptation settings. Furthermore, we show that fine-tuned retrained TLMs results in models with a better portability.

2 Task and Data

We focus on the protest event detection task following the 2019 CLEF ProtestNews Lab (Hürriyetoğlu et al., 2019).[1] Protest events are identified as politically motivated collective actions which lay outside the official mechanisms of political participation of the country in which the action takes place.

The lab is organised around three non-overlapping subtasks: (a.) document classification; (b.) sentence classification; and (c.) event extraction. Tasks (a.) and (b.) are text classification tasks, requiring systems to distinguish whether a document/sentence is referring to a protest event. The event extraction task is a sequence tagging problem requiring systems to identify event triggers and their corresponding arguments, similarly to other event extraction tasks, e.g., ACE (Linguistic Data Consortium, 2005).

The lab is designed to challenge models' portability in an unsupervised setting: systems receive a training and development data belonging to one variety and are asked to test both against a dataset from the same variety and a different one. We report in Table 1 the distribution of the markables (event triggers and arguments) for event extraction across the two varieties. We refer to the same variety (or source) distributions as India and to the different variety (or target) as China.

	India			China
Markable	**Train**	**Dev.**	**Test**	**Test**
Triggers	844	126	215	144
Arguments	1,895	288	552	295

Table 1: Distribution of event triggers and arguments. India is source. China is target.

The data are good examples of differences across factors characterising language varieties. For instance, although they belong to the same text genre (news articles), they describe protest events from two countries that have historical and cultural differences concerning what is worth protesting (e.g., caste protests are specific to India) and the type of protests (e.g., riots *vs*. petitions). Differences in the political systems entail differences in the actors of the protest events which is mirrored in the named entities describing person or organization names. Language is a further challenge. Both datasets are in English but they present dialectal and stylistic differences.

We quantified differences and similarities by comparing the training data (India$_{train}$) against the two test ones (India$_{test}$ and China$_{test}$) using the Jensen-Shannon (J-S) divergence and the out-of-vocabulary rate (OOV) that previous work has shown to be particularly useful for this purpose (Ruder and Plank, 2017). The figures in Table 2 better show how these data distributions occupy different regions in the variety space, with India$_{test}$ being closer to the training data than China$_{test}$. Tackling these similarities and differences is at the heart of our domain adaptation problem for event extraction.

	J-S		OOV	
↓Train / Test→	**India**	**China**	**India**	**China**
India	0.703	0.575	44.33%	53.82%

Table 2: J-S (Similarity) and OOV (Diversity) between train and test distributions for the event extraction task.

A further challenge is posed by the limited amount of training material. A comparison against the training portion of ACE shows that ProtestNews has 5 times less triggers and 4 times less arguments.[2] Unlike ACE, event triggers are not further classified into subtypes. However, seven argument types are annotated, namely *participant*, *organiser*, *target*, *etime* (event time), *place*, *fname* (facility name), and *loc* (location). The role set is inspired by ACE Attack and Demonstrate event types but they are more fine-grained. The markables are encoded in a BIO scheme (Beginning, Inside, Outside), resulting in different alphabets for triggers (e.g. B-*trigger*, I-*trigger* and O) and each of the arguments (e.g. O, B-*organiser*, I-*organiser*, B-*etime*, I-*etime*, etc.).

3 Continue Pre-training to Adapt

We applied DAR to English BERT `base-uncased` to fill a gap in language variety between BERT, trained on the BooksCorpus and Wikipedia, and the ProtestNews's data.

We collected two sets of domain related data from the TREC Washington Post Corpus version

[1] https://emw.ku.edu.tr/clef-protestnews-2019/

[2] The training portion of ACE has 4,312 triggers and 7,811 arguments.

Model	Input Format	Overall			Triggers			Arguments		
		P	R	F1	P	R	F1	P	R	F1
BERT	Document	$51.52_{4.20}$	$42.68_{4.98}$	$46.23_{1.98}$	$78.97_{4.32}$	$63.72_{4.76}$	$70.25_{1.87}$	$31.50_{17.54}$	$29.61_{16.09}$	$29.94_{16.49}$
NEWS-BERT	Document	$36.11_{3.77}$	$33.63_{7.79}$	$34.18_{3.48}$	$69.96_{5.18}$	$52.00_{10.32}$	$58.87_{5.41}$	$22.61_{4.69}$	$20.96_{9.62}$	$19.96_{6.95}$
PROTEST-ER	Document	$\textit{54.56}_{3.18}$	$48.47_{3.69}$	$51.11_{0.87}$	$70.48_{1.35}$	$67.90_{3.51}$	$69.08_{1.24}$	$37.59_{20.28}$	$40.20_{17.91}$	$37.86_{18.42}$
BERT	Sentence	$32.85_{6.27}$	$25.18_{6.61}$	$27.41_{4.19}$	$\textit{80.01}_{5.98}$	$29.30_{13.03}$	$41.16_{12.81}$	$18.95_{15.46}$	$22.79_{17.38}$	$19.74_{15.43}$
NEWS-BERT	Sentence	$52.86_{8.83}$	$10.76_{1.94}$	$17.67_{2.32}$	$\mathbf{92.92}_{1.84}$	$9.83_{3.08}$	$18.24_{5.90}$	$29.47_{6.16}$	$10.15_{1.12}$	$14.46_{0.85}$
PROTEST-ER	Sentence	$49.91_{1.99}$	$\textit{54.13}_{0.63}$	$\textit{51.91}_{0.97}$	$77.63_{1.41}$	$\textit{68.93}_{1.75}$	$\textit{72.99}_{0.80}$	$\textit{39.82}_{17.61}$	$\textit{46.13}_{17.86}$	$\textit{41.98}_{17.26}$
Best CLEF 2019	Sentence	**66.20**	**55.67**	**60.48**	79.79	**69.77**	**74.44**	**56.55**	**48.66**	**51.54**

Table 3: India data (source). Results for TLM are averaged over five runs. Standard deviation is reported in subscript. *Best* results correspond to the best system in the 2019 CLEF ProtestNews Lab tasks. Best scores are in bold. Second best scores are in italics.

Model	Input Format	Overall			Triggers			Arguments		
		P	R	F1	P	R	F1	P	R	F1
PROTEST-ER	Document	$\mathbf{64.48}_{5.01}$	$36.53_{2.76}$	$\textit{46.39}_{1.02}$	$74.07_{4.74}$	$\textit{69.30}_{5.66}$	$\textit{71.23}_{1.05}$	$\textit{42.70}_{18.68}$	$20.11_{14.83}$	$25.19_{14.71}$
PROTEST-ER	Sentence	$52.62_{5.34}$	$\textit{39.18}_{3.25}$	$44.62_{1.97}$	$\textit{74.08}_{3.20}$	$64.86_{7.44}$	$68.73_{2.75}$	$39.06_{16.03}$	$\textit{23.56}_{11.99}$	$\textit{27.02}_{11.81}$
Best CLEF 2019	Sentence	*62.65*	**46.24**	**53.21**	**77.27**	**70.83**	**73.91**	**49.64**	**33.57**	**39.56**

Table 4: China data (target). Results for TLM are averaged over five runs. Standard deviation is reported in subscript. *Best* results correspond to the best system in the 2019 CLEF ProtestNews Lab tasks. Best scores are in bold. Second best scores are in italics.

3[3] (WPC). The first collection (WPC-Gen) contains 100k random news articles. The second collection (WPC-Ev) contains all news articles related to an ongoing or past protest event for a total of 79,515 documents. The protest news articles have been automatically extracted with a specific BERT model for document classification trained and validated on an extended version of the document classification task from the ProtestNews Lab (Hürriyetoğlu et al., 2021). The model achieves an average F1-score of 90.15 on both India and China. We explicitly excluded as data for further pre-train BERT the CLEF 2019 India and China documents.

	J-S		OOV	
↓DAR / Test→	India	China	India	China
WPC-Gen	0.583	0.594	12.17%	4.38%
WPC-Ev	0.562	0.569	11.61%	4.46%

Table 5: J-S (Similarity) and OOV (Diversity) between the DAR datasets WPC-Gen and WPC-EV and the and test data distributions for the event extraction task.

We apply each data collection separately BERT `base-uncased` by further training for 100 epochs using the MLM objective. The outcomes are two pre-trained language models: NEWS-BERT and PROTEST-ER. The differences between the models are assumed to be minimal but yet relevant to assess the impact of the data used for DAR. To further support this claim we report in Table 5 an analysis of the similarities and differences of the DAR data materials against the India and China test data. As the figures show, the DAR datasets are equally different from the protest event extraction ones. Furthermore, we did not modify BERT original vocabulary by introducing new tokens. More details on the retraining parameters are reported in the Appendix A.1.

[3] https://trec.nist.gov/data/wapost/

4 Experiments and Results

Event extraction is framed as a token-level classification task. We adopt a joint strategy where triggers' and arguments' extent and labels are predicted at once (Nguyen et al., 2016). We used India_{test} to identify the best model (NEWS-BERT *vs.* PROTEST-ER) and system's input granularity. With respect to this latter point, we investigate whether processing data at document or sentence level could benefit the TLMs as a strategy to deal with limited training materials. We compare each configuration against a generic BERT counterpart. We fine-tune each model by training all the parameters simultaneously. All models are evaluated using the official script from the ProtestNews Lab. Triggers and arguments are correctly identified only if both the extent and the label are correct. We apply to China only the best model and input format.

India data Results for India are illustrated in Table 3. In general, PROTEST-ER obtains better results than BERT and NEWS-BERT. Sentence qualifies as the best input format for PROTEST-ER, while document works best for NEWS-BERT and

BERT.

The language variety of the data distributions used for DAR has a big impact on the performance of fine-tuned systems, with NEWS-BERT being the worst model. The extra training should have made this model more suited for working with news articles than the corresponding generic BERT. This indicates that selection of suitable data is an essential step for successfully applying DAR.

Globally, the results show that DAR has a positive effect on Precision, especially when sentences are used as input for fine tuning the models. Positive effects on Recall can only be observed for PROTEST-ER.

With the exclusion of NEWS-BERT, the systems achieve satisfying results for the trigger component. Argument detection, as expected, is more challenging, with no model reaching an F1-score above 50%. PROTEST-ER always performs better, especially when processing the data at sentence level. In numerical terms, PROTEST-ER provides an average gain of 11.74 points. [4] We observe a relationship between argument type frequency in the training data and models's performance where the most frequent arguments, i.e., *participant* (26.43%), *organizer* (18.31%), and *place* (14.45%), obtain the best results. However, PROTEST-ER improves performances also on the least frequent argument types, i.e., *loc* (6.49%) and *fname* (5.85) of, respectively, 12.00 and 5.38 points on average, when compared to BERT.

China data Results for China are reported in Table 4. We applied only PROTEST-ER keeping the distinction between document *vs*. sentence input. Although using sentences as input leads to the best results for India, we also observe that the results of the document input models are competitive, leaving open questions whether such a way of processing the input could be an effective strategy for model portability for event extraction. The results clearly indicate that PROTEST-ER is a competitive and pretty robust system. Interestingly, we observe that on the China data, the best results are obtained when processing data at document level.

Looking at the portability for the event components, it clearly appears that arguments are more difficult than triggers. Indeed, the absolute F1-score of the best models for triggers is in the same range of that for India. When focusing on the arguments, the drops in performances severely affect all argument types, except for *fname*. We also observe that the biggest drops are registered in those arguments that are most likely to express domain specific properties. For instance, the absolute F1-score difference between the best models for India and China for *place* is 39.79 points, 36.29 for *organizer*, and 27.11 for *etime*. On the contrary, only a drop of 9.84 points is observed for *participant*, suggesting that ways of indicating those who take part to a protest event (e.g. protesters, or rioters) are closer than expected.

[4] This figure has been obtained by grouping the scores of all models using the retrained version, regardless of the input format.

5 Discussion and Conclusions

Our results indicate that DAR is an effective strategy for unsupervised domain adaptation. However, we show that not every data distribution matching a potential target domain has the same impact. In our case, we measure improvements only when using data that more directly target the content of the task, i.e., protest events, possibly supplementing limitations in training materials. We have gathered interesting cues that processing data at document level can actually be an effective strategy also for a sequence labeling task with small training data. We think that this approach allows the TLMs to gain from processing longer sequences and acquire better knowledge. However, more experiments on different tasks (e.g., NER) and with different training sizes are needed to test this hypothesis.

A further positive aspect of DAR is that it requires less training material to boost system's performance, pointing to new directions for few-shot learning. We projected the learning curves of BERT and PROTEST-ER using increasing steps of the training data. PROTEST-ER achieves an overall F1-score $\sim$30% with only 10% of the training data, while BERT needs minimally 30% to achieve comparable performances (see Appendix A.3).

Disappointingly, PROTEST-ER falls way back the best model that participated in Protest-News. Skitalinskaya et al. (2019) propose a Bi-LSTM-CRF architecture using FLAIR contextualized word embeddings (Akbik et al., 2018). They also adopt a joint strategy for trigger and argument prediction. PROTEST-ER obtains a better Precision only on China for the overall evaluation and for trigger. Quite surprisingly, on India it is BERT that achieves better results on trigger, although the model appears to be quite unstable, as shown by the standard deviation. At this stage, it is still unclear whether these disappointing performances are due to the retraining (i.e., need to extend the number of documents used) or the small training corpus.

Future work will focus on two aspects. First, we will further investigate the impact of the size of the training data when using TLMs. This will require to experiment with different datasets and tasks. Secondly, we will explore solutions for multilingual extensions of PROTEST-ER.

Acknowledgments

The authors from Koc University were funded by the European Research Council (ERC) Starting Grant 714868 awarded to Dr. Erdem Yörük for his project Emerging Welfare.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, Savannah, GA. USENIX Association.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1077–1087, Melbourne, Australia. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

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

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

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

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019. Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 425–432, Cham. Springer International Publishing.

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.

Linguistic Data Consortium. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*, 5.4.3 2005.07.01 edition.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344, Sydney, Australia. Association for Computational Linguistics.

Timothy Miller. 2019. Simplified neural unsupervised domain adaptation. 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 414–419, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Kosuke Nishida, Kyosuke Nishida, Itsumi Saito, Hisako Asano, and Junji Tomita. 2020. Unsupervised domain adaptation of language models for reading comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5392–5399, Marseille, France. European Language Resources Association.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Gabriella Skitalinskaya, Jonas Klaff, and Maximilian Spliethöver. 2019. Clef protestnews lab 2019: Contextualized word embeddings for event sentence detection and event extraction. In *CLEF (Working Notes)*.

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.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. 2018. Learning semantic representations for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 5423–5432.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. 2019. Multi-source domain adaptation for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 32, pages 7287–7300. Curran Associates, Inc.

A Appendices

A.1 BERT-NEWS/PROTEST-ER Further Training

Preprocessing The unlabeled corpora of (protest related) news articles from the TREC Washington Post version 3 are minimally preprocessed prior to the language model retraining phase. We use the full text, including the title, of each news article. Document Creation Times are removed. We perform sentence splitting using `spaCy` (Honnibal et al., 2020).

Training details We further train the English BERT `base-uncased` for 100 epochs. We use a batch size of 64 through gradient accumulation. Other hyperparameters are illustrated in Table 6. Our TLM implementation uses the `HuggingFace` library (Wolf et al., 2020). The pretrainig experiment was performed on a single Nvidia V100 GPU and took 8 days.

Hyperparameter	Value
optimizer	adam
adam_epsilon	1e-08
learning rate	5e-05
logging steps	500
mlm probability	0.15
gradient accumulation steps	4
per gpu train batch size	16
max grad norm	1.0
pretrained model	bert-base-uncased
max-tokens	512
max epochs	100
random seed	42

Table 6: Hyperparameter configuration used for generating PROTEST-ER.

A.2 BERT/PROTEST-ER Fine-tuning

Table 7 shows the values of the hyperparameters used for fine-tuning BERT and PROTEST-ER. We used Tensorflow (Abadi et al., 2016) for the implementation and the `Huggingface` library (Wolf et al., 2020) for implementing the BERT embeddings and loading the data. We used the CRF implementation available from the Tensorflow Addons package.

The models are trained for a maximum of 100 epochs, using a constant learning rate of 2e-5; if the validation loss does not improve for 5 consecutive epochs, training is stopped. The best model is selected on the basis of the validation loss. We manually experimented with the learning rates 1e-5, 2e-5, 3e-5. No other hyperparameter optimization was performed.

Hyperparameter	Value
learning rate	2e-5
learning rate schedule	constant
clipnorm	1.0
optimizer	adam
dropout	0.1
max-tokens	512
max epochs	100
random seed	42

Table 7: Hyperparameter configuration used for task finetuning.

We used the original train, validation, and test splits of the event extraction task of the 2019 CLEF ProtestNews Lab.

We conducted all the experiments using the Google Colaboratory platform. The time required to run all the experiments on the free plan of Colaboratory is approximately 20 hours. Figure 1 graphically illustrates the base architecture.

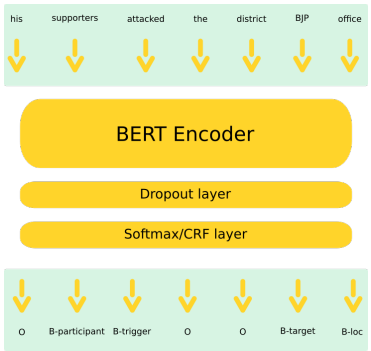

Figure 1: The base model architecture for the token classifier.

A.3 BERT/PROTEST-ER Learning Curves

In the following graphs we plot the learning curves of the BERT and PROTEST-ER model on the India and China dataset. In both cases, we observe that PROTEST-ER obtains competitive scores just using 10% of the training data, suggesting that the TLM's representations are already shifted towards the protest domain. To obtain the same results, the generic BERT models need minimally 30% of the training data, when using documents as input, and 70% of the training, when using sentences.

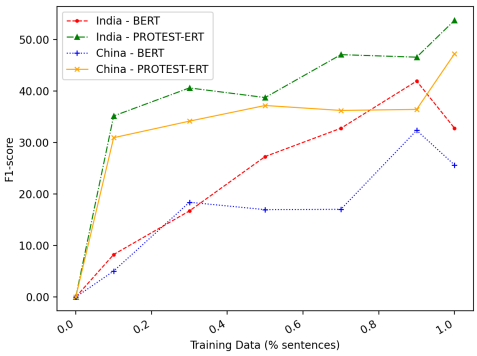

Figure 2: Learning curve for event extraction (triggers and arguments) for BERT and PROTEST-ER models on India and China, according to different portions (percentages) of the training materials (input granularity: **sentence**). Input data are randomly selected.

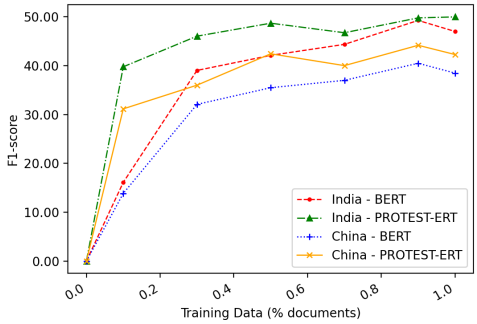

Figure 3: Learning curve for event extraction (triggers and arguments) for BERT and PROTEST-ER models on India and China, according to different portions (percentages) of the training materials (input granularity: **document**). Input data are randomly selected.