CamPros at CASE 2022 Task 1: Transformer-based Multilingual Protest News Detection

Kumari Neha[†] **Mrinal Anand**[†] **Tushar Mohan**[†]
Arun Balaji Buduru[†] **Ponnurangam Kumaraguru**[‡]
[†]Indraprastha Institute of Information Technology, Delhi
[‡]International Institute of Information Technology, Hyderabad
{nehak,mrinal20222,tushar19393,arunb}@iiitd.ac.in
{pk.guru}@iiit.ac.in

Abstract

Socio-political protests often lead to grave consequences when they occur. The early detection of such protests is very important for taking early precautionary measures. However, the main shortcoming of protest event detection is the scarcity of sufficient training data for specific language categories, which makes it difficult to train data-hungry deep learning models effectively. Therefore, cross-lingual and zero-shot learning models are needed to detect events in various low-resource languages. This paper proposes a multi-lingual cross-document level event detection approach using pre-trained transformer models developed for Shared Task 1 at CASE 2022. The shared task constituted four subtasks for event detection at different granularity levels, i.e., document level to token level, spread over multiple languages (English, Spanish, Portuguese, Turkish, Urdu, and Mandarin). Our system achieves an average F_1 score of 0.73 for document-level event detection tasks. Our approach secured 2^{nd} position for the Hindi language in subtask 1 with an F_1 score of 0.80. While for Spanish, we secure 4^{th} position with an F_1 score of 0.69. Our code is available at https://github.com/nehapspathak/campros/.

1 Introduction

The recent technological advancement has led to a continuous flow of information among users in online and offline ecosystems. Users' information may cover various social and political factors, often constituting information related to political violence, crisis, and protests, among others. The automatic detection of such socio-political protests/crisis events from news and social media has become crucial from a peaceful society perspective (Hürriyetoğlu et al., 2020, 2021). Not only does the early detection of such event helps in the deployment of early interventions, but it also helps understand people's perception of a socio-political event.

Event detection aims to identify and extract pertinent data from a text about specific categories of events. It is a crucial information extraction task that unearths and collects information about current and historical occurrences concealed in vast amounts of textual data. The CASE 2022 workshop focuses on detecting socio-political and crisis events in a multi-lingual setting at different granularity levels. This paper focuses on developing models and systems for "Multilingual Protest News Detection - Shared Task 1". In shared task 1, there are 4 subtasks. The aim of subtask 1 is to detect whether a news article contains event information. The news articles are in the form of documents. Hence the subtask looks at whether a given document contains event information. The second subtask focuses on detecting a sentence containing information about a past or ongoing event. The third task focuses on event sentence coreference identification, such as which event sentences in subtask 2 belong to the same event. The fourth and final subtask focuses on event extraction and aims to identify the event triggers and their arguments. We present our proposed system for subtask 1 in this paper.

Researchers have focused on Event extraction from various aspects in the past (Yadav et al., 2021; Lai et al., 2021a). The task presented by (Hürriyetoğlu et al., 2020) focused on event sentence co-reference identification. In the CASE 2021 socio-political and crisis event detection, the training dataset consisted of English, Spanish and Portuguese, while the test data were from English, Spanish, Portuguese, and Hindi (Hürriyetoğlu et al., 2021). In CASE 2022, however, new languages are introduced for multilingual document-level event detection. The workshop allows participants to create models for various subtasks and contrast related approaches. Subtask 1 consists of documents from English, Spanish, and Portuguese for training. For testing, the documents are available

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 169 - 174

in a zero-shot setting, including languages from low-resource languages; Hindi, Turkish, Urdu, and Mandarin. Identifying crisis and socio-political protest detection in a multi-lingual setting makes the Task very complex.

Our work mainly focuses on document-level (subtask 1) event detection in a multilingual setting. Our approach is based on pretrained transformer models and different learning strategies for making predictions. Since the tasks are designed for protests in a multilingual setting, we do not perform language-level pre-processing on our dataset. Our submission for subtask 1 achieved 2^{nd} position in zero-shot Hindi document-level event detection.

The rest of the paper is organized as follows. Section 2 describes the Related literature. The details of the Task and dataset are presented in Section 3. The proposed approach and experimental setup are described in Section 4. Results are described in Section 5, followed by Conclusion in Section 6. We intend to make our code public for further use by the community.

2 Related Work

In natural language processing (NLP), Event detection is a task that detects event triggers/mentions (i.e., the key terms that drive or express an event) and categorizes them into predefined event types (Lai et al., 2021b). The early detection of ongoing and past events exploited feature-based approaches to detect events (Li et al., 2013). However, the early data-driven (Hogenboom et al., 2011), knowledge-driven, and rule-based approaches missed the semantic relationship in the data (Danilova and Popova, 2014). Other early approaches for event detection include machine learning models such as SVM and decision trees (Schrodt et al., 2014). The recent deep learning approaches proposed in the literature (Ahmad et al., 2020) improve event detection; nonetheless, they are not generalizable for low-resource languages. To address the data scarcity problem for low-resource languages, researchers have recently used the pre-trained language model GPT-2 to generate training samples (Veyseh et al., 2021a).

Another less-discovered approach in the Event detection task is Cross-Lingual event detection which proposes model creation for effective performance over different languages (Guzman-Nateras et al., 2022). The work presented in (Lai et al., 2021b) utilizes knowledge from open-domain word sense disambiguation to transfer knowledge into few-shot learning models for Event detection, such that the model can generalize to new event types. To perform Event detection at the document level, the work in (Veyseh et al., 2021b) proposes a dynamic selection of relevant sentences in a document to create improved representation learning. Targeting the issues with scarce availability of low-resource languages, the CASE 2021 subtask introduced the multi-lingual crisis event detection dataset, which focuses on the zero-shot and few-shot detection of protest and crisis event (Hürriyetoğlu et al., 2021).

Language	**Label 1**	**Label 0**	**Total**
English (En)	1,912	7,412	9,324
Spanish (Es)	131	869	1,000
Portuguese (pt)	197	1,290	1,487

Table 1: Training Data available for training for Shared Task 1, subtask 1: Document-level crisis event prediction.

Language	**Documents**
English	3,871
Hindi	268
Mandarin	300
Spanish	400
Portuguese	671
Turkish	300
Urdu	299

Table 2: Test Data for testing for Shared Task 1, subtask 1: Document-level crisis event prediction.

3 Data

The dataset used in CASE 2022 has been created in the process presented in (Hürriyetoğlu et al., 2022). For subtask 1, the new data contains documents with and without protest events. The data provided for training are highly imbalanced and provided for only 3 languages. The testing data contains 7 languages, with documents from additional 4 languages apart from training data. Table 1 provides the details of the training data provided in the shared task. Table 2 presents the test data for the Task. Given that no training data is present for Hindi, Mandarin, Turkish and Urdu, the task of document event detection becomes a zero-shot classification problem.

Language	Model	macro-F1
English	mBERT+Softmax	0.76
	XLM-Roberta+LSTM	0.74
	XLM-Roberta+Sigmoid	**0.77**
	XLM-Roberta+Sigmoid (U)	0.72
Spanish	**mBERT+Softmax**	**0.69**
	XLM-Roberta+LSTM	0.63
	XLM-Roberta+Sigmoid	0.64
	XLM-Roberta+Sigmoid (U)	0.63
Portuguese	mBERT+Softmax	0.68
	XLM-Roberta+LSTM	0.71
	XLM-Roberta+Sigmoid	**0.76**
	XLM-Roberta+Sigmoid (U)	0.72

Table 3: Test results for English, Spanish and Portuguese documents, as reported in the shared task. The training data were present for the above 3 languages. U represents a model with under-sampled data.

3.1 Data proprocessing

Since we experiment with mBERT (cased) and other sentence-based embeddings, we do not lowercase our document corpus before training. We also do not conduct language-specific pre-processing to keep the preprocessing step language agnostic. However, we removed any URLs, and a single occurrence replaced repeated symbols. We also removed any extra spaces present in the data.

4 Methodology

The transformer-based models have recently gained success in various multilingual NLP tasks such as offensive content detection (Arango et al., 2022) and various zero-shot cross-lingual tasks (Kuo and Chen, 2022). We experiment with different multilingual models and analyze how the different models perform on the downstream task of document classification in subtask 1. We design the document classification problem as a sequence classification problem (Hettiarachchi et al., 2021; Gürel and Emin, 2021).

In our approach, we use different transformer models including XLM-Roberta (Conneau et al., 2020), mBERT (Devlin et al., 2018) and encoder-decoder based LASER (Artetxe and Schwenk, 2019) to generate embedding from the documents. We experiment with different layers on top of the multi-lingual sentence embedding. Our preliminary analysis found that transformer-based XLM-Roberta with a sigmoid layer outperformed other models in the macro-F1 score. Therefore, in our approach, we propose the XLM-Roberta model with a sigmoid classification layer for event prediction. XLM-Roberta is pre-trained on unlabeled Wikipedia text and CommonCrawl Corpus of 100 languages. The XLM-Roberta has a vocabulary size of $25,000$ and uses SentencePiece tokenizer (Kudo and Richardson, 2018). We fine-tuned the model for our task with the training data provided. The training data was highly imbalanced. However, oversampling and under-sampling methods did not provide any marginal improvement in the model's output as per our experiments.

4.1 XLM-Roberta Based Document Classification Models

XLM-Roberta belongs to an unsupervised representation learning framework as it does not use cross-lingual resources (Conneau et al., 2020). XLM-Roberta has L = 12 transformers, with H = 768 attention heads with A = 12, and 270M parameters. The maximum token size for input for XLM-Roberta is 512 tokens. The token size of 512 is less for creating document-level creation, as a lot of information might not be captured. However, breaking the sentences into 512-length tokens might lead to an incorrect labeling process for different sentence splits (Gürel and Emin, 2021). Due to the limitation of our system, our final approach uses a 256-length token for document embedding creation. The learning rate was $2.75e^{-05}$, the batch size for training was 32, and the training was done for 20 epochs. The total training time taken for the XLM-Robert-based model was approximately 2 hours. Since we use the Sigmoid layer on the top of XLM-Roberta, the final decision boundary for 0/1 was taken based on the probability of 0.6 for

Language	Model	macro-F1
Hindi	mBERT+Softmax	0.71
	XLM-Roberta+LSTM	0.75
	XLM-Roberta+Sigmoid	**0.80**
	XLM-Roberta+Sigmoid (U)	0.77
Turkish	mBERT+Softmax	0.69
	XLM-Roberta+LSTM	0.70
	XLM-Roberta+Sigmoid	**0.74**
	XLM-Roberta+Sigmoid (U)	0.69
Urdu	mBERT+Softmax	0.67
	XLM-Roberta+LSTM	0.72
	XLM-Roberta+Sigmoid	0.71
	XLM-Roberta+Sigmoid (U)	**0.73**
Mandarin	**mBERT+Softmax**	**0.75**
	XLM-Roberta+LSTM	0.71
	XLM-Roberta+Sigmoid	0.75
	XLM-Roberta+Sigmoid (U)	0.73

Table 4: Test results for Hindi, Mandarin, Turkish and Urdu documents, as reported in the shared task. Training data was not provided for the above language. Hence classification is done in a zero-shot setting.

all cases.

4.2 Experimental setup

For training all models, we use the Nvidia RTX 3090 GPU system with an installed Cuda version of 11.3. For training, we combined the training data from the 3 languages, English, Spanish, and Portuguese, as shown in Table 1. We performed at a 90:10 split for training and testing, respectively. The split was done randomly but stayed the same for all the experiments with models to obtain the result on the same set of datasets. The score we demonstrated for document-level classification was the F1-macro metric, which was selected as an evaluation metric for our models. We performed experiments with different epoch numbers and batch sizes with the same experimental setup.

4.3 Baselines

We experimented with different multilingual models such as XLM-Roberta (Conneau et al., 2020), mBERT (Devlin et al., 2018) and LASER (Artetxe and Schwenk, 2019) to obtain predictions. The performance for LASER was the worst in our case. Hence, we do not report the results from LASER-based models.

XLM-Roberta+Softmax (under-sampling): In this approach, before feeding the data into the model, we under-sample the majority class (i.e., a class with label 0 representing a no-event class) such that we have an equal number of documents for both label 0 and label 1 class. We under-sample the training data constituting the combination of documents from all the 3 languages. After this, we split the data into the ratio of 90:10 and fed it to the model with XLM-Roberta with softmax as the classification layer. The number of epochs for training is set to 20, and the batch size is taken as 32.

XLM-Roberta+LSTM: After we have created embedding using XLM-Roberta, we feed the embedding into long short-term memory (LSTM) layers to train the model. We use the sigmoid layer for the classification of events.

mBERT+Softmax: We also tried mBERT to create embedding, which is the multilingual BERT embedding for our experiment. The BERT tokenizer is based on wordpiece tokenizer. We used softmax as a classification layer and trained the model.

5 Results

In this section, we demonstrate and elaborate on the results from different models for each language. Table 3 shows the result for English, Spanish and Portuguese language, for which we had training data available. The best model for English came out to be XLM-Roberta+Sigmoid model, with a macro-F1 score of 0.77. The second best model for English was mBERT+Softmax model, with a macro-

F1 score of 0.76. While XLM-Roberta+LSTM showed macro-F1 score of 0.74, the undersampled majority class for XLM-Roberta+Sigmoid produced the worst result, with a macro-F1 score of 0.72. For Spanish, however, our proposed framework of XLM-Roberta+Sigmoid model was outperformed by mBERT+Softmax, with macro-F1 of 0.69. XLM-Roberta+Sigmoid remained the second best model with macro-F1 score of 0.64. The result for XLM-Roberta+LSTM and undersampled XLM-Roberta+Sigmoid came as 0.63. In Portuguese, our proposed framework outperformed all other baselines, with macro-F1 score of 0.76. The second best model for Portuguese was undersampled XLM-Roberta+Sigmoid, with macro-F1 score of 0.72. The macro-F1 score for XLM-Roberta+LSTM came as 0.71, while mBERT+Softmax performed worst for the Portuguese document classification task. Hence, we found that XLM-Roberta with the Sigmoid layer outperformed for English and Portuguese tasks; however, the best model for Spanish was multilingual BERT with the softmax layer.

Table 4 presents the results for the zero-shot classification for the respective languages. Our best model, the XLM-Roberta+Sigmoid model, obtained a macro-F1 score of 0.80 for Hindi and secured 2^{nd} in the shared task. The second best model for zero-shot Hindi document classification was undersampled XLM-Roberta+Sigmoid with a macro-F1 score of 0.77. The macro-F1 score for XLM-Roberta+LSTM model was 0.75. We found that for Hindi, mBERT+Softmax produced the worst results, with macro-F1 score of 0.71. For Turkish, the best model also came out as XLM-Roberta+Sigmoid, with macro-F1 as 0.74. Among the baselines, the XLM-Roberta+LSTM model produced a macro-F1 score of 0.70, while the macro-F1 score for both mBERT+Softmax and undersampled XLM-Roberta+Sigmoid came as 0.69. For the Urdu language, XLM-Roberta+LSTM marginally outperformed the proposed model, with a macro-F1 score of 0.72. The macro-F1 score for the proposed XLM-Roberta+Sigmoid came as 0.71. The worst model for Urdu was mBERT+Softmax, with a macro-F1 score of 0.67. In contrast, the best model for the Urdu language was the undersampled XLM-Roberta+Sigmoid model with a macro-F1 score of 0.73. For Mandarin, however, the best F1-score was obtained from both the mBERT+Softmax model and the proposed XLM-Roberta+Sigmoid model, with a marginal difference on the macro-F1 score of 0.75. The undersampled XLM-Roberta+Sigmoid produced a macro-F1 score of 0.73, while the XLM-Roberta+LSTM model produced a macro-F1 score of 0.71.

6 Conclusion

This paper describes our approaches for CASE@EMNLP 2022: Shared Task on Socio-political and Crisis Events Detection in multilingual settings. We explored various multilingual and zero-shot approaches and showed results across the languages in subtask 1. We propose XLM-Roberta with a Sigmoid layer for classifying crisis events in zero-shot and low-resource language settings. Our system achieved an average F1 score of 0.73. Among the given languages, our proposed approach was able to secure 2^{nd} place in the Hindi document event classification task. While comparing with our approach, the multilingual Bert with softmax layer obtained better results for Spanish and Mandarin, with the result for Spanish securing the 4^{th} spot in the shared task.

Acknowledgements

Thank you to the Precog Research Group for giving us constructive feedback, the IIIT Delhi and IIIT Hyderabad IT support teams for their assistance, and anonymous reviewers for their constructive feedback.

References

Faizan Ahmad, Ahmed Abbasi, Brent Kitchens, Donald A Adjeroh, and Daniel Zeng. 2020. Deep learning for adverse event detection from web search. *IEEE Transactions on Knowledge and Data Engineering*.

Aymé Arango, Jorge Pérez, Bárbara Poblete, Valentina Proust, and Magdalena Saldaña. 2022. Multilingual resources for offensive language detection. *WOAH 2022*, page 122.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Vera Danilova and Svetlana Popova. 2014. Socio-political event extraction using a rule-based approach. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 537–546. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alaeddin Gürel and Emre Emin. 2021. Alem at case 2021 task 1: Multilingual text classification on news articles. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 147–151.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2021. Daai at case 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 120–130.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

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, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (aespen): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Chia-Chih Kuo and Kuan-Yu Chen. 2022. Toward zero-shot and zero-resource multilingual question answering. *IEEE Access*.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021a. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400, Online. Association for Computational Linguistics.

Viet Dac Lai, Minh Van Nguyen, Thien Huu Nguyen, and Franck Dernoncourt. 2021b. Graph learning regularization and transfer learning for few-shot event detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2172–2176.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Philip A Schrodt, John Beieler, and Muhammed Idris. 2014. Three'sa charm?: Open event data coding with el: Diablo, petrarch, and the open event data alliance. In *ISA Annual Convention*. Citeseer.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021a. Unleash gpt-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6271–6282.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Ngo Trung, Bonan Min, and Thien Huu Nguyen. 2021b. Modeling document-level context for event detection via important context selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5403–5413.

Nishant Yadav, Nicholas Monath, Rico Angell, and Andrew McCallum. 2021. Event and entity coreference using trees to encode uncertainty in joint decisions. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.