ARC-NLP at CASE 2022 Task 1: Ensemble Learning for Multilingual Protest Event Detection

Umitcan Sahin, Oguzhan Ozcelik, Izzet Emre Kucukkaya, Cagri Toraman
{ucsahin, ogozcelik, ekucukkaya, ctoraman}@aselsan.com.tr
Aselsan Research Center, Ankara, Turkey

Abstract

Automated socio-political protest event detection is a challenging task when multiple languages are considered. In CASE 2022 Task 1, we propose ensemble learning methods for multilingual protest event detection in four subtasks with different granularity levels from document-level to entity-level. We develop an ensemble of fine-tuned Transformer-based language models, along with a post-processing step to regularize the predictions of our ensembles. Our approach places the first place in 6 out of 16 leaderboards organized in seven languages including English, Mandarin, and Turkish.

1 Introduction

Socio-political protest events are organized to protest against various decision and policy makers. An example is the social movement of Arab Springs and Internet hacktivism. The detection of socio-political protest events in news articles is a challenging task when news are reported in multiple languages.

The shared task of Multilingual Protest News Detection (Hürriyetoğlu et al., 2022; Hürriyetoğlu et al., 2020), organized in the workshop of Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) that is held at the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), targets automated detection of protest events considering language generalization of the event information collection systems. The shared task includes four subtasks:

Subtask 1, Protest Document Classification: The subtask aims to detect if news articles contain past or ongoing protest events. There are three source languages; English, Spanish, and Portuguese. In addition, there are seven target languages including English, Turkish, and Mandarin. The granularity of classification is document-level. The prediction output is binary (protest exists or not).

Subtask 2, Protest Sentence Classification: The subtask aims to detect if the news sentences contain protest events. There are three source and target languages; English, Spanish, and Portuguese. The granularity of classification is sentence-level. The prediction output is binary.

Subtask 3, Protest Event Sentence Coreference Identification: The subtask aims to identify which protest sentences are about the same event. There are three source and target languages; English, Spanish, and Portuguese. The granularity of grouping is sentence-level. The prediction output is clusters of protest event sentences.

Subtask 4, Protest Event Extraction: The subtask aims to extract or label protest entity spans such as triggers and participants. There are three source and target languages; English, Spanish, and Portuguese. The granularity of classification is word span-level. The prediction output is entity labels.

The ARC-NLP team participated in all subtasks of Multilingual Protest News Detection. Our main approach for all subtasks is based on two factors. First, we utilize Transformer-based language models that are pretrained on specific languages, e.g. RoBERTa (Liu et al., 2019), and also multilingual corpus, e.g. mDeBERTa (He et al., 2021a). Second, we apply ensemble learning and post-processing methods to obtain better and smoother predictions, considering that large language models are stochastic (Bender et al., 2021). Besides, we apply customized methods for each subtask according to the subtask's definition and requirements. Our approach places the first place in 6 out of 16 leaderboards organized in seven languages including English, Mandarin, and Turkish. In the following sections, we present our detailed solutions and leaderboard results for all subtasks of multilingual protest event detection.

Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE), pages 175 - 183
December 7-8, 2022 ©2022 Association for Computational Linguistics

Language	Train	Test
English (EN)	9,324	3,871
Spanish (ES)	1,000	400
Portuguese (PR)	1,487	671
Hindi (HI)	-	268
Turkish (TR)	-	300
Mandarin (MA)	-	300
Urdu (UR)	-	299

Table 1: The number of instances in **Subtask 1.**

2 Subtask 1: Protest Document Classification

2.1 Dataset

The dataset in Subtask 1 consists of news documents collected in various languages, and corresponding protest labels (positive or negative). The collection and annotation processes are described in (Hürriyetoğlu et al., 2021) for the 2021 data, and in (Hürriyetoğlu et al., 2020) for the 2022 data. The number of instances is given for 2022 in Table 1. While English, Spanish, and Portuguese have training samples that are labeled, the other languages only have unlabeled test samples (i.e. zero-shot evaluation). In Subtask 1, the class labels are unbalanced, that is, there are more negative samples (no past or ongoing event in document) than positive ones.

2.2 Methods

We focus on ensemble learning of multilingual or monolingual language models. We also use data processing techniques, such as data translation to improve our models further. In Table 2, we share our best performing three submissions for each language for Subtask 1 (S1), which are based on four methods[1]:

Ensemble of multilingual language models (S1-multi): English, Spanish, and Portuguese have labeled data that can be used in training models, but not other languages. Therefore, we combine the labeled samples from English, Spanish, and Portuguese to construct the training data (i.e., source). We rely on a Transformer-based multilingual model, namely mDeBERTa (He et al., 2021c), which is the multilingual version of DeBERTa. It is pre-trained with the 2.5T CC100 multilingual dataset. In Subtask 1, we use the mDeBERTa V3 base model that has 12 layers and a hidden size of 768. We use the HuggingFace's Pytorch implementation of this model (He et al., 2021b), the corresponding tokenizer with max length 512, extra padding and truncation. We set epoch number to 5 and use constant learning rate $2e-5$.

We train five *split* mDeBERTa models, each with 80% of the training data randomly selected from the entire training data with replacement. Furthermore, we train a single *full* mDeBERTa model using the entire training data. While **S1-multi-5** in Table 2 uses the predictions of the five split models, **S1-multi-6** uses the predictions of the five split models and one full model together. Moreover, we follow two approaches to ensemble the models' predictions into final test labels. First, we take the majority voting of the five split models, called **M1**. Second, we compute the average softmax probabilities of the five split models and one full model for each class in test samples, called **M2**. The classes with the highest probabilities are selected for final test labels.

Ensemble of monolingual language models (S1-mono): We use Transformer-based monolingual models, namely RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021c) for English[2], BETO (Cañete et al., 2020) for Spanish, and BERTimbau (Souza et al., 2020) for Portuguese. All monolingual models are their base versions, and HuggingFace's Pytorch implementations are used. We fine-tune these models with the samples from their respective languages for document classification. Other notations (ensemble size, majority method, and hyperparameters) are the same as in multilingual models.

Ensemble of monolingual language models with Target Translation (S1-mono-TT): For zero-shot evaluation, we translate each target test language with no training instances (Spanish, Hindi, Turkish, Mandarin, and Urdu) to a source language (English) using Google's translation[3]. **mono-TT-5** in Table 2 consists of five DeBERTa models (trained with 80% of train data). The predictions are computed from the translated test data and ensembled together using M1 majority voting. In addition, **mono-TT-6** consists of five DeBERTa models and one full DeBERTa model whose predictions are computed on the translated test data using M2 majority voting. We use the same hyperparameters and settings as in previous setups.

[1] We did not submit all versions of the following methods for each language. Instead, we submitted best performing three models in our internal experiments for each language.

[2] We mostly observe that DeBERTa and mDeBERTa have better performances than RoBERTa and XLM-R in our internal experiments.

[3] https://translate.google.com

Method	Target Lang.	Train data (Source)	Backbone Models	Majority	Score
S1-mono-5	EN	EN	RoBERTa (x5 split)	M1	**80.74**
S1-mono-6		EN	RoBERTa (x5 split + x1 full)	M2	80.67
S1-multi-6		EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	80.03
S1-multi-5	PR	EN+ES+PR	mDeBERTa (x5 split)	M1	**79.85**
S1-multi-6		EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	78.73
S1-mono-6		PR	BERTimbau (x5 split + x1 full)	M2	77.96
S1-mono-TT-6	ES$^{EN}_{trans}$	EN	DeBERTa (x5 split + x1 full)	M2	**69.44**
S1-mono-6	ES	ES	BETO (x5 split + x1 full)	M2	68.75
S1-multi-6		EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	67.74
S1-multi-5	HI	EN+ES+PR	mDeBERTa (x5 split)	M1	**80.08**
S1-multi-6		EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	78.96
S1-mono-TT-6	HI$^{EN}_{trans}$	EN	DeBERTa (x5 split + x1 full)	M2	75.63
S1-mono-ST-6	TR	EN$^{TR}_{trans}$	BERTurk-128k (x5 split + x1 full)	M2	**84.06**
S1-mono-ST-5			BERTurk-128k (x5 split)	M1	83.27
S1-mono-TT-6	TR$^{EN}_{trans}$	EN	DeBERTa (x5 split + x1 full)	M2	82.89
S1-mono-TT-5	MA$^{EN}_{trans}$	EN	DeBERTa (x5 split)	M1	**83.39**
S1-mono-TT-6		EN	DeBERTa (x5 split + x1 full)	M2	83.23
S1-multi-6	MA	EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	83.06
S1-mono-TT-5	UR$^{EN}_{trans}$	EN	DeBERTa (x5 split)	M1	**77.99**
S1-mono-TT-6		EN	DeBERTa (x5 split + x1 full)	M2	77.48
S1-multi-6	UR	EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	76.15

Table 2: Our submitted models for **Subtask 1 (S1), Document Classification**. L$1^{L_2}_{trans}$ means that language L_1 is translated to language L_2. Split models are trained with randomly selected 80% of the train data and full models with all train data. Highest F1-macro scores are given in bold.

Ensemble of monolingual language models with Source Translation (S1-mono-ST): We translate a source language (English training samples) to a target language with no training data (Turkish) using Google's translation tool. We use Transformer-based monolingual BERTurk[4], which is trained with translated Turkish data. **mono-ST-5** in Table 2 consist of five split BERTurk models (trained with 80% of the training data) and final test labels are computed on the original Turkish test data using M1 majority voting. Similarly, **mono-ST-6** consists of five split BERTurk models and one full BERTurk model (trained with the entire training data) together, and final test labels are computed on the original Turkish test data using M2 majority voting. We use the same hyper-parameter and tokenizer settings as in previous setups.

2.3 Leaderboard Results

Our best performing model for each language in Subtask 1 is given in Table 3 along with best competitor scores in 2022 and our rankings. We rank the first place in Turkish and Mandarin, second place in Portuguese, and third place in Urdu.

3 Subtask 2: Protest Sentence Classification

3.1 Dataset

The dataset in Subtask 2 consists of news sentences and corresponding protest labels (positive or negative) in English, Spanish, and Portuguese. The collection and annotations are described in (Hürriyetoğlu et al., 2021) for the 2021 data. There is no new data provided in 2022. The number of examples for each language are given in Table 4. The problem of unbalanced class label distributions is also present in this task.

3.2 Methods

We mainly focus on multilingual and monolingual language models as in Subtask 1. In Table 5, we share our best performing two methods for each language for Subtask 2 (S2), which are based on two methods[5]:

Ensemble of multilingual language models (S2-multi): We combine the labeled instances from English, Spanish, and Portuguese to construct the training data (source). We utilize multilingual language models, namely mDeBERTa (He et al., 2021c), in the subtask. We use the corresponding tokenizer with max length 128, extra padding and truncation. We set epoch number to 5 and use

[4] https://huggingface.co/dbmdz/bert-base-turkish-128k-cased

[5] We follow a similar approach to Subtask 1 in our internal experiments for Subtask 2.

Lang.	Method	Our score	Best Competitor Score 2022	Rank 2022
EN	S1-mono-5	80.74	**82.49**	4
PR	S1-multi-5	79.85	**80.07**	2
ES	S1-mono-TT-6	69.44	**74.96**	5
HI	S1-multi-5	80.08	**80.78**	4
TR	S1-mono-ST-6	**84.06**	82.91	1
MA	S1-mono-TT-5	**83.39**	83.06	1
UR	S1-mono-TT-5	77.99	**79.71**	3

Table 3: The 2022 leaderboard scores for **Subtask 1, Document Classification**.

Language	Train	Test
English (EN)	22,825	1,290
Spanish (ES)	2,741	686
Portuguese (PR)	1,182	1,445

Table 4: The number of instances in **Subtask 2**.

constant learning rate $2e-5$ throughout the training. We train five *split* mDeBERTa models, each with 80% of the training data randomly selected from the entire training data with replacement. Furthermore, we train a single *full* mDeBERTa model using the entire training data. While *S2-multi-5* uses the predictions of the five split models, *S2-multi-6* uses the predictions of the five split models and one full model together. The meanings of M1 and M2 are also the same as in Subtask 1.

Ensemble of monolingual language models (S2-mono): Our second method utilizes monolingual language models. In Subtask 2, we use RoBERTa (Liu et al., 2019) for English. The model is the base version. We use HuggingFace's pytorch implementation, the corresponding tokenizers with max length 128, extra padding and truncation. We set epoch number to 10 and use constant learning rate $2e-5$. *S2-mono-6* includes five split models (trained with the 80% of training data) and one full model (trained with the entire training data) together, whose predictions are ensembled using the M2 majority voting.

3.3 Leaderboard Results

Our best performing model for each language in Subtask 2 is reported in Table 6 along with best competitor scores in both 2021 and 2022, and our rankings. We rank the third place in English and Spanish in 2022.

4 Subtask 3: Event Sentence Coreference Identification

4.1 Dataset

The dataset in Subtask 3 consists of news sentences and corresponding clusters in three different languages (English, Spanish, and Portuguese). The statistics of the dataset are given in Table 7. The numbers of instances are smaller than those of previous subtasks. The number of instances in English is significantly higher than those of other languages. The number of clusters also varies in the dataset.

4.2 Methods

Our approach for Subtask 3 is based on ensemble learning of hierarchical clustering. In order to cluster the sentences, we calculate the distance between two sentences, and then feed this distance score to a hierarchical clustering algorithm.

We construct pairs of sentences from the dataset by labeling them according to existing clustering labels. For instance, assume that there are three sentences with numbers 20, 21, and 22 in two clusters as [[20],[21, 22]]. We then construct the sentence pairs (21, 22) as positive, and (20, 21) and (20, 22) as negative pairs. We calculate the Cosine distance similarity between these sentence pairs for obtaining training instances. The training of sentence pairs is a binary classification task (positive or negative) with binary cross-entropy loss. The output softmax probability is used as distance score.

After training and obtaining a distance similarity model, we apply hierarchical or agglomerative clustering algorithm using the distance scores. For linking two clusters, we use single linkage, where the distance between nearest points in two clusters is considered.

Based on this clustering approach, we apply ensemble learning as in previous subtasks. Since there are very small number of training instances in Spanish and Portuguese training datasets, we exploit translating target languages to English, and merging the instances of all languages in multilingual models. In Table 8, we share our best performing submissions for each language for Subtask 3 (S3), which are based on four methods[6]:

[6] We also tried different methods such as BERTopic (Grootendorst, 2022) and SBERT (Reimers and Gurevych, 2019),

Method	Target Lang.	Train data (Source)	Backbone Models	Majority	Score
S2-multi-6	EN	EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	**83.77**
S2-mono-6		EN	RoBERTa (x5 split + x1 full)	M2	80.68
S2-multi-5	PR	EN+ES+PR	mDeBERTa (x5 split)	M1	**86.53**
S2-multi-6		EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	86.11
S2-multi-6	ES	EN+ES+PR	mDeBERTa (x5 split + x1 full)	M2	**87.20**
S2-multi-5		EN+ES+PR	mDeBERTa (x5 split)	M1	85.16

Table 5: Our submitted models for **Subtask 2 (S2), Sentence Classification**. Split models are trained with randomly selected 80% of the train data and full models with all train data. Highest F1-macro scores are given in bold.

Lang.	Method	Our score	Best Competitor Score		Rank	
			2021	2022	2021	2022
EN	S2-multi-6	83.77	85.32(Hu and Stoehr, 2021)	**85.93**	3	3
PR	S2-multi-5	86.53	88.47(Awasthy et al., 2021)	**89.67**	4	4
ES	S2-multi-6	87.20	88.61(Awasthy et al., 2021)	**88.78**	2	3

Table 6: The 2021 and 2022 leaderboard scores for **Subtask 2, Sentence Classification**.

Language	Train	Test
English (EN)	596	100
Spanish (ES)	21	40
Portuguese (PR)	11	40

Table 7: The number of instances in **Subtask 3.**

Multilingual language model with hierarchical clustering (S3-multi-1): We merge the original instances from English, Spanish, and Portuguese to construct the training data. We apply distance model and hierarchical clustering as explained above. For distance model, we rely on a Transformer-based multilingual model, namely XLM-R (Conneau et al., 2020). In Subtask 3 (S3), we train only a single (1) multilingual model without using ensembles (*S3-multi-1*). We use the XLM-R base model that has 12 layers and a hidden size of 768. We use the HuggingFace's Pytorch implementation of this model (He et al., 2021b), the corresponding tokenizer with max length 512, extra padding and truncation. We set epoch number to 20 and use constant learning rate $2e-5$. We use the SciPy implementation for hierarchical clustering. We set the hierarchical clustering threshold as 0.65.

Ensemble of monolingual language models with hierarchical clustering and Source Translation (S3-mono-ST): We translate Spanish and Portuguese to English, and then merge all instances. The test data is also translated to English. We apply distance model and hierarchical clustering as explained above. For distance model, we train a monolingual language model, namely RoBERTa (Liu et al., 2019). We use the RoBERTa base model that has 12 layers and a hidden size of 768. The hyperparameters and other settings are the same as in the previous method.

We train five *split* RoBERTa models, each with 80% of the training data randomly selected from the entire training data with replacement. Furthermore, we train a single *full* RoBERTa model using the entire training data. **S3-mono-ST-6** in Table 8 uses the predictions of the five split models and one full model together. Moreover, we apply the following approach to ensemble the models' predictions into final test labels. The algorithm we are using is based on the getting connected components on a graph after getting rid of the low probability connections. To do so, the binary similarity matrix that is symmetric is calculated based on the pairs in clusters for each clustering model. After that, we get element-wise average of the similarity matrices to get a single matrix of probabilities. A pre-determined threshold (0.60) is then applied to remove the low probability scores, so that we obtain a final similarity matrix that contains binary decisions for sentence pairs.

Ensemble of monolingual language models with hierarchical clustering and Target Translated (S3-mono-TT): We use only English instances for training a monolingual language model. However, we translate the target languages (Spanish and Portuguese) to English, since they have very small number of training instances. We apply distance model and hierarchical clustering as explained above. For distance model, we train RoBERTa (Liu et al., 2019) base model. The hyperparameters and other settings are the same as in the previous

however we did not achieve better performances. We did not submit all versions of the reported methods for each language. Instead, we submitted best performing models in our internal experiments for each language.

Method	Target Lang.	Train data	Backbone Models	Score
S3-multi-1	EN	EN + ES + PR	XLM-RoBERTa Base	79.44
S3-mono-ST-6	EN	EN + ES_{trans}^{EN} + PR_{trans}^{EN}	RoBERTa (x5 split + x1 full)	84.24
S3-mono-TT-6	EN	EN	RoBERTa (x5 split +x1 full)	84.26
S3-mono-TT-16	EN	EN	RoBERTa (3x5 split (15) + x1 full)	**85.11**
S3-multi-1	ES	EN + ES + PR	XLM-RoBERTa Base	82.68
S3-mono-ST-6	ES_{trans}^{EN}	EN + ES_{trans}^{EN} + PR_{trans}^{EN}	RoBERTa (x5 split + x1 full)	**85.25**
S3-mono-TT-6	ES_{trans}^{EN}	EN	RoBERTa (x5 split +x1 full)	*
S3-mono-TT-16	ES_{trans}^{EN}	EN	RoBERTa (3x5 split (15) + x1 full)	83.70
S3-multi-1	PR	EN + ES + PR	XLM-RoBERTa Base	88.88
S3-mono-ST-6	PR_{trans}^{EN}	EN + ES_{trans}^{EN} + PR_{trans}^{EN}	RoBERTa (x5 split + x1 full)	92.04
S3-mono-TT-6	PR_{trans}^{EN}	EN	RoBERTa (x5 split +x1 full)	91.21
S3-mono-TT-16	PR_{trans}^{EN}	EN	RoBERTa (3x5 split (15) + x1 full)	**93.00**

Table 8: Our submitted models for **Subtask 3 (S3), Event Sentence Coreference Identification**. All methods are based on hierarchical clustering with single linkage. $\text{L}_{1}{}_{trans}^{L_2}$ means that language L_1 is translated to language L_2. Split models are trained with randomly selected 80% of the train data and full models with all train data. Highest CoNLL-2012 average (Pradhan et al., 2014) scores are given in bold. (*) means that the submission score is not produced by the leaderboard system.

Lang.	Methods	Our Score	Best Competitor Score		Rank	
			2021	2022	2021	2022
EN	S3-mono-TT-16	**85.11**	84.44 (Awasthy et al., 2021)	-	1	1
ES	S3-mono-ST-6	**85.25**	84.23 (Awasthy et al., 2021)	-	1	1
PR	S3-mono-TT-16	93.00	**93.03** (Tan et al., 2021)	-	2	1

Table 9: The 2021 and 2022 leaderboard scores for **Subtask 3, Event Sentence Coreference Identification**. Highest CoNLL-2012 average (Pradhan et al., 2014) scores are given in bold.

method.

We train five *split* RoBERTa models, each with 80% of the training data randomly selected from the entire training data with replacement. Furthermore, we train a single *full* RoBERTa model using the entire training data. **S3-mono-TT-6** in Table 8 uses the predictions of the five split models and one full model together. Besides, we construct a bigger ensemble to reflect more aspects from different models, such that we repeat five splits three times to get 15 different models and one full model together (**S3-mono-TT-16**). We apply the same approach to ensemble the models' predictions into final test labels as in the previous method.

4.3 Leaderboard Results

In Subtask 3, the scoring metric is CoNLL-2012 average score (Pradhan et al., 2014). The leaderboard and our ranking among 2021 and 2022 submissions can be seen in Table 9. In 2022, we accomplished the first place in all languages. In 2021 leaderboard, we get the first place in English and Spanish and we get the second place in Portuguese.

5 Subtask 4: Protest Event Extraction

5.1 Dataset

The dataset in Subtask 4 consists of entity spans in news sequences for three languages (English, Spanish, and Portuguese). Event entity types are event time, facility name, organizer, participant, place, target, and trigger. The number of sequences are highly imbalanced for English compared to Spanish and Portuguese. We provide a detailed statistics of the dataset in Table 10.

	Language	English		Spanish		Portuguese	
	Data split	Train	Test	Train	Test	Train	Test
Entity	Facility	1,201	-	49	-	48	-
	Organizer	1,261	-	25	-	19	-
	Participant	2,663	-	88	-	73	-
	Target	1,470	-	64	-	32	-
	Trigger	4,595	-	157	-	122	-
	Place	1,570	-	15	-	61	-
	Time	1,209	-	40	-	41	-
Total	Sequences	808	88	30	50	33	50
	Word count	103,327	11,334	3,712	7,852	2,780	6,280
	Vocab. size	12,841	3,160	1,379	2,424	1,034	2,046

Table 10: The number of instances and entity types in **Subtask 4**.

5.2 Methods

We utilize monolingual and multilingual models in ensemble learning of token classification with a specific focus on post-processing predictions. We preprocess the input data since there are very long sequences that do not fit the input layer of the models, where the maximum sequence length is 512. We, therefore, split sequences, whose sequence

Method	Target Lang.	Train data	Backbone Models	Majority	Post-Proc	Score
S4-multi-10	EN	EN	XLM-R (x5) + XLM-R-CRF (x5)	✓	✗	75.70
S4-multi-PP-10			XLM-R (x5) + XLM-R-CRF (x5)	✓	✓	75.90
S4-mono-PP-10-v2			DeBERTa (x5) + DeBERTa-CRF (x5)	✓	✓	77.46
S4-mono-PP-10-v3			DeBERTa-CRF (x10)	✓	✓	**77.84**
S4-multi-10	PR	EN+ES+PR	XLM-R (x5) + XLM-R-CRF (x5)	✓	✗	70.89
S4-multi-PP-10			XLM-R (x5) + XLM-R-CRF (x5)	✓	✓	71.50
S4-multi-PP-10-v2			mDeBERTa (x5) + mDeBERTa-CRF (x5)	✓	✓	**73.84**
S4-multi-PP-10-v3			mDeBERTa-CRF (x10)	✓	✓	**73.84**
S4-multi-10	ES	EN+ES+PR	XLM-R (x5) + XLM-R-CRF (x5)	✓	✗	66.08
S4-multi-PP-10			XLM-R (x5) + XLM-R-CRF (x5)	✓	✓	66.46
S4-multi-PP-10-v2			mDeBERTa (x5) + mDeBERTa-CRF (x5)	✓	✓	**68.00**
S4-multi-PP-10-v3			mDeBERTa-CRF (x10)	✓	✓	67.91

Table 11: Our submitted models for **Subtask 4 (S4), Event Extraction**. Highest CoNLL (Tjong Kim Sang and De Meulder, 2003) macro F1 scores are given in bold.

length is greater than 512 tokens, with a window size of 200. For instance, we split a sequence having 654 words as four groups having 200, 200, 200, and 54 words. We do not use data translation due to the granularity of classification (i.e. translated word spans may not match the original sequence). In Table 11, we share our best performing three submissions for each language for Subtask 4 (S4), which are based on four methods[7]:

Ensemble of multilingual language models (S4-multi): We have more number of instances in English compared to Spanish and Portuguese. Having less data in a language complicates our task, since the granularity of the task is word span-level. We use a multilingual model, XLM-R (Conneau et al., 2020). We also use XLM-R-CRF, which is a hybrid model of Transformer-based language model and Conditional Random Fields (CRF) (Lafferty et al., 2001). The motivation behind using the CRF on top of Transformer-based language model is that the hybrid model can achieve promising results for the long named entities (Ozcelik and Toraman, 2022). In Subtask 4, we use the XLM-R base cased model that has 12 layers and a hidden size of 768. We use the HuggingFace's Pytorch implementation of this model (He et al., 2021b), the corresponding tokenizer with max length 512, extra padding and truncation. We set epoch number to 20 and use constant learning rate $5e-5$.

We train five XLM-R and five XLM-R-CRF models, fine-tuned with different seeds on full train data (**S4-multi-10** in Table 11). Majority voting is applied after training of 10 models. During majority voting, instead of choosing the most frequent classes, we use a task-specific algorithm to choose best label. We first create a label transition dictionary, where possible transitions have positive weights while transition errors have negative weights. For instance, `B-etime` → `I-etime` have positive weight, but `O` → `I-{entity type}` have negative weights since `O` label cannot be followed by any type of `I-{entity type}`.

Ensemble of multilingual language models with Post-Processing (S4-multi-PP): In this method, we apply the same approach and ensemble models as in the previous method. The only differences are that we use an additional multilingual language model, mDeBERTa (He et al., 2021c) (**S4-multi-PP-10-v2** and **S4-multi-PP-10-v3** in Table 11), and we apply a post-processing step on the prediction labels of ensemble members as follows. Post-processing is applied after the majority voting step, since there still occurs transition errors for the predictions, e.g., prediction of `O` label just before `I-{entity type}`. We, thereby, automatically fill the entity chunks when transition error occurs. For instance, an entity chunk having three labels `B-target I-target I-target` is corrected if it is predicted as `B-target O I-target`.

Ensemble of monolingual language models with Post-Processing (S4-mono-PP): In this method, we apply the same approach, ensemble models, and post-processing method as in the previous method. The only difference is that we use a monolingual language model on English, namely DeBERTa (He et al., 2021a) (**S4-mono-PP-10-v2** and **S4-mono-PP-10-v3**). This method is not applied for Spanish and Portuguese since the number of training instances are very small and we do not have translations.

[7] We did not submit all versions of the following methods for each language. Instead, we submitted best performing models in our internal experiments for each language.

Language	Model	Our score	Best Competitor Score		Our Rank	
			2021	2022	2021	2022
EN	S4-mono-PP-10-v3	**77.84**	78.11 (Awasthy et al., 2021)	76.49	2	1
PR	S4-multi-PP-10-v3	73.84	73.24 (Awasthy et al., 2021)	**74.57**	1	2
ES	S4-multi-PP-10-v2	68.00	66.20 (Awasthy et al., 2021)	**69.87**	1	2

Table 12: The 2021 and 2022 leaderboard scores for **Subtask 4, Event Extraction**.

5.3 Leaderboard Results

In Subtask 4, the evaluation metric is CoNLL (Tjong Kim Sang and De Meulder, 2003) macro F1 score. The leaderboard and our ranking among 2021 and 2022 submissions can be seen in Table 12. We get the first place in Portuguese and Spanish in 2021, and English in 2022. We achieve promising improvement in our scores for all languages when majority and post-processing are applied. Thus, we believe that our methods can generalize to many languages in token classification tasks.

6 Discussion and Conclusion

In this study, we summarize our solutions for multilingual protest event detection under four subtasks that have different granularities from document to word span-level. Our overall approach is based on ensemble learning and post-processing, which places the first place in 6 out of 16 leaderboards organized in seven languages including English, Mandarin, and Turkish.

Based on the experiments and leaderboard results, we have the following observations.

- We argue that post-processing predictions benefit the predictions of ensemble models due to the fact that large language models are stochastic (Bender et al., 2021). Specifically, post-processing predictions have significant benefits in the performances of our ensemble models in Subtask 3 and Subtask 4.
- When zero-shot evaluation (i.e. no available training data) is considered such as Turkish in this task, we observe that Transformer-based language models pretrained on a target language perform better in ensemble learning compared to multilingual models. Furthermore, we observe that for languages such as Spanish, Mandarin, and Urdu, monolingual Transformer-based language models pretrained on English perform better than multilingual language models. For fine-tuning, we translate the training data in source languages, such as English, to a target language, such as Turkish.

We plan to extend our experiments to different data collections, such as tweets, in different languages, specifically the languages used in Eastern Europe and Middle East countries.

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, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

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.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021b. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021c. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Tiancheng Hu and Niklas Stoehr. 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, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Osman Mutlu, Çağrı Yoltar, Fırat Duruşan, and Burak Gürel. 2020. Cross-context news corpus for protest events related knowledge base construction.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

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.

Oguzhan Ozcelik and Cagri Toraman. 2022. Named entity recognition in turkish: A comparative study with detailed error analysis. *Information Processing & Management*, 59(6):103065.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Fiona Anting Tan, Sujatha Das Gollapalli, and See-Kiong Ng. 2021. NUS-IDS at CASE 2021 task 1: Improving multilingual event sentence coreference identification with linguistic information. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 105–112, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.