Multilingual Protest News Detection - Shared Task 1, CASE 2021

Ali Hürriyetoğlu[⋆], Osman Mutlu[⋆], Erdem Yörük[⋆], Farhana Ferdousi Liza[†], Ritesh Kumar[◇], Shyam Ratan[◇]
[⋆]Koç University [†]University of Essex [◇]Dr. Bhimrao Ambedkar University, Agra
{ahurriyetoglu,omutlu,eryoruk}@ku.edu.tr
farhana.ferdousi.liza@essex.ac.uk
ritesh78_llh@jnu.ac.in, shyamratan2907@gmail.com

Abstract

Benchmarking state-of-the-art text classification and information extraction systems in multilingual, cross-lingual, few-shot, and zero-shot settings for socio-political event information collection is achieved in the scope of the shared task Socio-political and Crisis Events Detection at the workshop CASE @ ACL-IJCNLP 2021. Socio-political event data is utilized for national and international policy- and decision-making. Therefore, the reliability and validity of such datasets are of utmost importance. We split the shared task into three parts to address the three aspects of data collection (Task 1), fine-grained semantic classification (Task 2), and evaluation (Task 3). Task 1, which is the focus of this report, is on multilingual protest news detection and comprises four subtasks that are document classification (subtask 1), sentence classification (subtask 2), event sentence coreference identification (subtask 3), and event extraction (subtask 4). All subtasks have English, Portuguese, and Spanish for both training and evaluation data. Data in Hindi language is available only for the evaluation of subtask 1. The majority of the submissions, which are 238 in total, are created using multi- and cross-lingual approaches. Best scores are between 77.27 and 84.55 F1-macro for subtask 1, between 85.32 and 88.61 F1-macro for subtask 2, between 84.23 and 93.03 CoNLL 2012 average score for subtask 3, and between 66.20 and 78.11 F1-macro for subtask 4 in all evaluation settings. The performance of the best system for subtask 4 is above 66.20 F1 for all available languages. Although there is still a significant room for improvement in cross-lingual and zero-shot settings, the best submissions for each evaluation scenario yield remarkable results. Monolingual models outperformed the multilingual models in a few evaluation scenarios, in which there is relatively much training data.

1 Introduction

Every day across the globe, hundreds of different socio-political protest events against various decisions taken by the respective governments or authorities take place. These events are of interest to political scientists, policy makers, democracy watchdogs and other stakeholders for multiple reasons including analysing the nature, scope and extent of such events, forming public opinion about various causes, gauging the state of freedom and democracy across different nations and others. However, manually keeping track of such events at a national level itself is a very challenging task and it is more so if we are trying to get a sense of these events across the globe. Given this, automated methods of collecting and, possibly, processing protest news events from multiple countries and locations gain great significance. But the automated identification and collection of such events in multiple languages also comes with its own set of significant challenges. This task was designed to address some of these challenges..

The task of event information detection, in general, could be divided into multiple subsequent steps and the efficiency at each of these steps could drastically affect the quality of the resultant event database. Thus, we believe one must consider a complete pipeline including the following steps i) classification of documents and sentences as relevant or not (in the sense that whether they describe an event or not - in this specific case event is a protest event); ii) identification of the sentences that provide information about the same event; and iii) extraction of event information. Finally the resultant database of the events should be tested against a manually created list of events to evaluate the performance of the state-of-the-art systems on this task. We have formulated these different steps into three inter-dependent tasks - Task 1 is a Multilingual Protest News Detection task, Task 2 complements the first task with fine-grained semantic event classification (Haneczok et al., 2021) using data reported by Piskorski et al. (2020) and Task 3 evaluates the performance of the systems developed for Task 1 on a real-world scenario, in this it specifically evaluates the system for the task

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 79–91

of identifying the events surrounding Black Lives Matter movement, using data from Twitter and New York Times (Giorgi et al., 2021).

In order to benchmark the state-of-the-art in these three tasks, we organized the shared task Socio-political and Crisis Events Detection[1]. The shared task is held in the scope of the workshop Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)[2] (Hürriyetoğlu et al., 2021) that is held at the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021).[3] We report results of the Task 1 that is Multilingual Protest News Detection in this report.

Task 1 follows Extracting Protests from News (ProtestNews) and Event Sentence Coreference Identification (ESCI) tasks that were organized at Conference and Labs of the Evaluation Forum (CLEF 2019) (Hürriyetoğlu et al., 2019a,b) and Automated Extraction of Socio-political Events from News (AESPEN 2020) at Language Resources and Evaluation Conference (LREC) (Hürriyetoğlu et al., 2020) respectively. The ProtestNews and ESCI were monolingual tasks comprising only English data from various countries and evaluated for cross-context generalization of automated text processing systems across texts collected from different countries. This edition of the shared task series focuses on language generalization of the event information collection systems in four languages viz. English, Hindi, Portuguese, and Spanish. The Task 1 we present in this report follows all the steps we find essential for event information collection in a multilingual setting. It is divided into the following subtasks.

Subtask 1; Document classification:
The first subtask aims to identify if a news article contains information about a past or ongoing socio-political event.

Subtask 2; Sentence classification:
The second subtask asks the question if a sentence contains information about a past or ongoing event.

Subtask 3; Event sentence coreference identification:
The third sub-task is about identifying which event sentences (per definition provided in subtasks 1 and 2) are about the same event. The event sentences in question are from the same document.

Subtask 4; Event Extraction:
The final subtask is the extraction of event entity spans such as triggers and event arguments.

We particularly focus on events that are in the scope of contentious politics and characterized by riots and social movements, i.e., the repertoire of contention (Giugni, 1998; Tarrow, 1994). We utilize an extended version of the GLOCON Gold standard dataset that is created based on this definition in this task (Hürriyetoğlu et al., 2021). The languages in scope for all of the subtasks are English, Spanish, and Portuguese. The subtask 1 comprises test data in Hindi as well. This setting creates a total of 13 evaluation scenarios such as subtask 1 English, Subtask 4 Portuguese, etc. Participants had access to training data for all the subtasks and in all languages. There is no training data in Hindi language and its test data is available only for subtask 1. Moreover, training data in Spanish and Portuguese are relatively small in comparison to data in English.

This report discusses relevant work in Section 2, annotation of the data set utilized in the benchmark in Section 3, task and data descriptions in Sections 4 and 5. We provide results of baseline systems we developed for subtasks 1 and 2 and participant submissions in sections 7 and 8. We conclude the report in section 9.

2 Related Work

Automated socio-political event information collection has a long history (Hutter, 2014; Schrodt and Yonamine, 2013). Many event ontologies such as IDEA (Bond et al., 2003), CAMEO (Gerner et al., 2002), ACLED (Raleigh et al., 2010), and PLOVER[4] have been proposed in this domain. These ontologies facilitated development of automated event information collection tools such as MPEDS (Hanna, 2017), PETRARCH (Norris et al., 2017), TABARI, and BBN Accent, EMBERS (Saraf and Ramakrishnan, 2016). The databases that are created using automated methods at various levels are GDELT (Leetaru and Schrodt, 2013), ICEWS (O'Brien, 2010), MMAD (Weidmann and Rød, 2019), PHOENIX, POLDEM (Kriesi et al., 2019), SPEED (Nardulli

[1] https://github.com/emerging-welfare/case-2021-shared-task, accessed on May 26, 2021. The repository contains sample data, evalution scripts, and samples of submission files.

[2] https://emw.ku.edu.tr/case-2021/, accessed on May 26, 2021.

[3] https://2021.aclweb.org/, accessed on May 26, 2021.

[4] https://github.com/openeventdata/PLOVER, accessed on May 30, 2021.

et al., 2015), TERRIER (Liang et al., 2018), and UCDP (Sundberg et al., 2012). Although majority of this work is on western countries and English language, there are considerable number of similar studies on collecting socio-political event information from text originated from countries other than western countries and in languages other than English (Sönmez et al., 2016; Danilova, 2015). The main data source of the event information has been text of news articles. But the use of social media posts has gradually increased in recent times (Zhang and Pan, 2019; Sech et al., 2020).

The application of state-of-the-art automation using machine learning and computational linguistics techniques requires gold standard annotated corpora that can be utilized for the task and benchmarks that facilitate comparison of the proposed methods for protest event information collection (Wang et al., 2016; Lorenzini et al., 2016). However, there are only a few corpora shared for research purposes in this domain (Makarov et al., 2016; Sönmez et al., 2016; Sech et al., 2020) and to the best of our knowledge, there is no benchmark available. Our efforts via this task establishes a common ground for comparison and benchmarking in a multilingual setting.

The multilingual text processing has become a critical target in computational linguistics and machine learning. Tackling this task enables us to collect information about global events that are reported and to trace occurrence of similar events in many languages. Moreover, this technology facilitates event information collection from local sources, which provide detailed information about events. New benchmark data sets such as XTREME (Hu et al., 2020) and system proposals such as mBERT (Devlin et al., 2019a), XLM (Lample and Conneau, 2019), mBART (Liu et al., 2020), and XLM-R (Conneau et al., 2020) have demonstrated promising results on various tasks (Hakala and Pyysalo, 2019). Multilingual embedding creation is the other major research line, in which the approaches such as LASER (Artetxe and Schwenk, 2019a) and LaBSE (Feng et al., 2020) have been proposed. These methodological advancements extend the exploration space for detecting event information. Consequently, this technology contributes to the resolution of the popularity or ideological bias of the sources toward popular and mainstream events both at global and local levels.

In general, it is not an optimum decision to work with a single language due to biases, absence of event information in a single source or international sources etc. We must invest in generalizability and multilinguality of the event information collection systems and therefore in the current task we incorporate these aspects as well. By design, zero- or few-shot learning is required to tackle some subtask and language combinations (Pires et al., 2019) in this task since the released data set contained relatively small training data in Spanish and Portuguese and no training data in Hindi. Thus the final evaluation provides some insights into these approaches for contentious socio-political event data collection and classification task.

3 Annotation

The multilingual version of the corpus GLOCON Gold (Hürriyetoğlu et al., 2021), which was reported as containing data only in English, is utilized in this task. This corpus is created by random sampling from news archives and double annotation (Yörük et al., 2021) for the data in English, Spanish, and Portuguese. There are document, sentence, and token level annotations that are performed on the whole news articles. The quality of the annotations are ensured by a detailed annotation manual[5], adjudications, spot-checks, and semi-automated quality checks before the next level of annotation starts. A cascaded annotation workflow is applied. For instance, quality of the document level annotations is ensured before the sentence level annotation starts. The inter-annotator aggrements (IAA) that are measured using Krippendorf's alpha (Krippendorff et al., 2016) are .75 and .65 in average for document- and sentence-level annotations. The token level IAA is between .35 and .60 for the information types in scope. All disaggrements are resolved by the annotation supervisor. Moreover, spot-checks and semi-automated error corrections have fixed 10% of the annotation errors in total (Hürriyetoğlu et al., 2021). The document and sentence level annotations yielded the data for subtasks 1 and 2 respectively. The token level annotations produced the data for subtasks 3 and 4.

The data in Hindi is prepared applying a slightly different methodology but using the same annotation manual. A native graduate student from India has annotated these articles at the document level. Twenty Hindi newspapers and periodicals available on the web are used as sources for this data set. This data set contains all possible articles and editorials related to ongoing farmer protest in India against the three farm bills (1.Bill on agri market, 2.Bill on contract farming, and 3.Bill relating to commodities) passed by the government of India in August, 2020. [6] The current annotated data set cov-

[5] https://github.com/emerging-welfare/general_info/tree/master/annotation-manuals, accessed on May 29, 2021.

[6] https://en.wikipedia.org/wiki/2020%

ers equal proportion of articles from each source, which are twenty except the periodical Panchjanya, which has only 19 articles. All articles are searched and collected manually from web pages of each newspaper and periodical with the metadata date, date of article retrieval, URL, location of incident, and location of newspaper.

Overall, the news articles used in this task are obtained from China and South Africa in English, from Brazil in Portuguese, from Argentine in Spanish, and from India in English and Hindi. The annotation team consists of graduate students in social and political sciences. Students from Turkey, Brazil, and India have annotated text in English, Spanish and Portuguese, and Hindi respectively. These students are trained on contentious politics of their target country and annotation methodology before they started the annotation. News reports that are not related to a target country are excluded from the token level annotations in order to improve precision of the annotations.

4 Task Description

Task 1 consists of four subtasks that are at document, sentence, and token levels. The subtasks are as follows.

Subtask 1 aims at classifying news articles. If the document reports an event that has happened or is ongoing, it should be labelled as relevant. Scheduled events, speculations, and anything else should be marked as irrelevant. Subtask 1 is a binary classification problem.

Subtask 2 has the same aim as subtask 1 but for sentences of a document. [7] A sentence should have some token(s) that qualify as event trigger or a reference to an event trigger in another sentence.

Subtask 3 is about determining event sentences that provide information about the same event. All event sentences in a document are clustered according to the events they report.

Subtask 4 marks all tokens in an event sentence based on the information they hold[7]. The event trigger and its arguments such as participant, place, target, organizer, time, and facility name are annotated. The event trigger can be a coreferent of a trigger in another sentence.

The subtasks are multilingual by means of comprising data in English, Portuguese, and Spanish languages both for training and evaluation of the automated text processing systems. Moreover, the tasks are a few-shot scenario since Portuguese and Spanish training data is significantly less than English data. Finally, the subtask 1 includes a zero-shot setting in which participants do not have access to data in Hindi language, but they should predict documents in Hindi language.

Hürriyetoğlu et al. (2021) have showed that, although, event information collection could be performed utilizing systems developed only for subtask 4 with potential contribution of the systems developed for subtask 3 in principle, this setting is not possible in practice due to challenge of reliable annotation of event information at token level for development and evaluation of event extraction systems. Document and sentence annotation significantly facilitates reliable annotation of event information at token level. Moreover, authors have demonstrated a considerable increase in F1 in case document and sentence classification systems are applied before token level event extraction. Thus, we consider application of these subtasks in this order indispensable for reliable collection of event information.

5 Data Description

We share text data in English, Spanish and Portuguese for training and evaluation. Also, there is data in Hindi language for evaluation of the subtask 1. Finally, participants are free to use any additional data they may think that will help to improve their systems.

This section provides details on the format, size, and preparation of the data shared with the participants. Moreover, we describe how data across subtasks depend on each other and how we deal with copyright issues in the subsection on the data preparation.

5.1 Data Format

Listing 1: A training sample from subtask 1.

```
{
 "id":100187,
 "text":"Hall of fame\nResults -
    Pyeongchang 2018 Winter
    Olympic Games\nSee the full
    results from th",
 "label":0
}
```

All of our data is shared in JSON files except for subtask 4 which is shared as plain text files. The subtasks 1 and 2 are both text classification tasks, so their format, which can be seen in Listing 1, are the same, differing only in JSON field names. The "label" is the correct label assigned to the article/sentence and "text" is the article/sentence's text.

E2%80%932021_Indian_farmers%27_protest, accessed on June 9, 2021.

[7] The annotators see the whole document during the annotation

"text" field is named "sentence" for subtask 2 data. The "label" field is not shared for test data.

Listing 2: A training sample from subtask 3.

```
{
 "id":55471,
  "sentences": [
    "Lt-Col Andre Traut said the
       teenager laid the
       complaint at the Robertson
        police station following
       a farmworkers' protest in
       the area.",
    "Table grape harvesters
       started protesting about
       their working conditions
       in De Doorns last month.",
    "The protests spread to 15
       other towns and resulted
       in two deaths and the
       destruction of property.",
    "The farmworkers' strike
       resumed on Tuesday when
       their demands were not met
       ."
  ],
  "sentence_no":[2,5,7,8],
  "event_clusters":[[5,7,8],[2]]
}
```

As shown in Listing 2, fields for subtask 3 consist of positive sentences of an article ("sentences"), the ordering of these sentences in the article ("sentence_no") and correct clustering of these sentences ("event_clusters"). The "event_clusters" field is not shared for test data. Finally, for subtask 4, we share text files in BIO format, which is the standard for information extraction tasks (Ramshaw and Marcus, 1995). Below in b we provide a sample in BIO format.[8] The sample in human readable format is demonstrated in a. The bold face indicates the event trigger and the underlined tokens specify the arguments of the event trigger.

a. The recruits, at Valluvar Kottam **shouted slogans** including, "HCL lend us your ears, give us back our two years" while undertaking the day-long **fast**.

b. The$_{O}$ recruits$_{B\text{-participant}}$,$_{O}$ at$_{B\text{-fname}}$ Valluvar$_{I\text{-fname}}$ Kottam$_{I\text{-fname}}$ shouted$_{B\text{-trigger}}$ slogans$_{I\text{-trigger}}$ including$_{O}$,$_{O}$ "$_{O}$ HCL$_{B\text{-target}}$ lend$_{O}$ us$_{O}$ your$_{O}$ ears$_{O}$,$_{O}$ give$_{O}$ us$_{O}$ back$_{O}$ our$_{O}$ two$_{O}$ years$_{O}$ "$_{O}$ while$_{O}$ undertaking$_{O}$ the$_{O}$ day-long$_{O}$ fast$_{B\text{-trigger}}$.$_{O}$

[8]The participants receive this in vertical format.

5.2 Data Size

Total size[9] of the shared data for all languages and subtasks can be seen at Table *1*. The distribution of labels for training data for each subtask are as follows:

- Positive sample ratio for subtask 1 is .21, .13 and .13 for English, Portuguese and Spanish respectively.
- Positive sample ratio for subtask 2 is .19, .24 and .16 for English, Portuguese and Spanish respectively.
- For subtask 3, number of clusters in a sample in percentages can be found at Table 2
- The number of spans/entities for subtask 4 are shown in Table 3.

The sample size should be the same for the subtasks 3 and 4 in principle since they both are annotated when a news article is positive. However, it can be observed in Table 1 that subtask 3 has significantly less data than subtask 4. This is due to the exclusion of the articles with single positive sentence from subtask 3, as they only have one possible clustering solution.

		Subtask 1	Subtask 2	Subtask3	Subtask 4
English	Train	9,324	22,825	596	808
	Test	2,971	1,290	100	179
Portuguese	Train	1,487	1,182	21	33
	Test	372	1,445	40	50
Spanish	Train	1,000	2,741	11	30
	Test	250	686	40	50
Hindi	Train	-	-	-	-
	Test	268	-	-	-

Table 1: Sample[9] counts for all subtasks in all languages.

	1	2	3	4+
English	.62	.27	.06	.05
Portuguese	.57	.33	.05	.05
Spanish	.73	.27	.0	.0

Table 2: Number of clusters (events) in a sample in percentages in subtask 3 in all languages.

5.3 Data Preparation

Before preparing the data we had to consider the data shared in previous editions of the shared task, copyright issues and possible inference between data of separate subtasks.

Some portion of the data in English was shared with academic community in previous shared

[9]A sample is denoted as an article for subtasks 1, 3 and 4, and a sentence for subtask 2.

	English	Portuguese	Spanish
trigger	4,595	122	157
participant	2,663	73	88
place	1,570	61	15
target	1,470	32	64
organizer	1,261	19	25
etime	1,209	41	40
fname	1,201	48	49

Table 3: Number of spans in subtask 4 training data in all languages.

tasks (Hürriyetoğlu et al., 2019b, 2020) and publications as sample data (Hürriyetoğlu et al., 2021). On the one hand, a sample previously shared in training data should not be placed in test data since its correct answer is known. On the other hand, a sample previously shared in test data should not be shared in training data since that would make previous shared task obsolete.

We respect the copyright of the news sources. We never share the whole text of a news article. To further prevent possible copyright issues, we share only one third of the text starting from the beginning of the document in subtask 1, scramble the sentences in subtask 2, and use only positively labeled sentences in subtasks 3 and 4.

The final data preparation step is about avoiding inference of the labels of the data for a subtask from data of the other subtasks. As it is described in Section 3, our annotation process happens in a cascaded manner: sentence level depending on document level, token and sentence coreference depending on sentence and document levels. These dependencies between levels create the possibility to infer an upper level's label using a lower level's data (upmost level being document level). For example, for a sample in document level test data, one can easily confirm this sample is positive by checking to see if any of its sentences are shared in sentence level data. So when we prepare our data, we make sure there are no overlaps between levels that have these dependencies. This exclusion applies in the following cases:

- From our subtask 3 and 4 data, we exclude samples whose sentence(s) are in subtask 1's test data.
- From our subtask 3 and 4 data, we exclude samples that are in subtask 1's test data.
- From our subtask 2 data, we exclude sentences that belong to articles that are in subtask 1's test data.

As these cases show, the overlaps are handled in a top-down manner. Handling them in bottom-up manner, meaning excluding samples from upper levels (moving samples from test to training data), would disrupt the positive sample ratio and possibly create a bias in the data. Since sentence coreference and token level data are not dependent on each other, this process of sampling and exclusion is not carried out in this case. Data for these subtasks is derived from the same documents by respecting the training and evaluation splits.

6 Evaluation

Although the subtasks form a coherent flow, task participants can focus on one or more of them. Therefore, participants can choose the tasks or subtask(s) they would like to participate in. Participants have access to all of the data for all tasks and subtasks. Any combination of these resources to achieve high performance for any of the tasks is allowed. For instance, Task 1 data could be used to potentially improve the performance on Task 2 and vice versa.

Participants had access to the test data for a week and could submit up to five submissions for each subtask and language combination. The best score of each team is reflected to the leaderboard.[10] Additional submissions are allowed (after the competition ended) on a separate Codalab page[11] in case participating teams would like to run additional experiments or create multiple submissions of the same system for measuring standard deviation of their systems. However, the additional submission page allows only one submission for each language and subtask combination per day.

F1-macro is calculated on the predictions on the test data for the subtasks 1 and 2. We use a python implementation[12] of the original[13] conlleval evaluation script for subtask 4. The subtask 3 is evaluated using scorch - a python implementation of CoNLL-2012 average score for the test data (Pradhan et al., 2014). [14]. We carry out separate evaluation for each subtask using the test data for each language separately.

7 Baseline Systems

We created baseline models for the subtasks 1 and 2 in English, Portuguese, and Spanish. Document

[10] https://competitions.codalab.org/competitions/31247#results, accessed on June 9, 2021.

[11] https://competitions.codalab.org/competitions/31639, accessed on June 9, 2021.

[12] https://github.com/sighsmile/conlleval, accessed on June 6, 2021.

[13] www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt, accessed on June 11, 2021.

[14] https://github.com/LoicGrobol/scorch, accessed on June 6, 2021.

classification is a challenging task. For simplicity, we have done document classification on the summaries of documents, which are the most important sentence in the document generated using the LexRank extractive summarization method (Erkan and Radev, 2004). Thus document summarization task was converted into an important part of the sentence classification task pipeline. As such, the input text for the document classification is a sentence rather than a set of sentences.

We have used an Attention (i.e. Transformer) (Devlin et al., 2019b) based Neural Network model for feature representation (Minaee et al., 2021) and multilingual sentence representations (Reimers and Gurevych, 2020) for the subtasks 1 and 2 with three languages — English, Spanish and Portuguese. Among available approaches (Schwenk and Douze, 2017; Artetxe and Schwenk, 2019b; Reimers and Gurevych, 2020), Reimers and Gurevych (2020) provide efficient representation for sentences for 50+ languages from various language families. The main motivation of using the multilingual approach is to learn efficient representation for the low resource (Non-English) languages. Specifically, we have used 'distiluse-base-multilingual-cased' for learning sentence representation of the three languages. We have also used language-specific sentence representation (Reimers and Gurevych, 2019) for the English language. Specifically, for the experiment, we used 'paraphrase-distilroberta-base-v1', which is a 'DistilBERT-base-uncased' model fine-tuned on a large dataset of paraphrase sentences. We apply a Linear Support Vector Machine (SVM) classifier trained on these features. The multilingual representation yields competitive results in our experiments for the language-specific representation in English language classification.

We have used 70% of the training data to train the model and 30% of the data to validate the models for subtasks 1 and 2. For the document classification task, the validation scores are 74.85 for the English language, 49.27 for the Spanish language, and 56.67 for the Portuguese language. The test scores are 76.78, 64.45, 64.13 for English, Portuguese and Spanish respectively. For the sentence classification task (subtask 2) the F1-macro on the validation data is 79.67 with the language-specific representation of the English language. With multilingual representation, the validation F1-macro is 76.90 for the English language, 73.93 for the Portuguese, and 73.42 for the Spanish language. The score on the test data is 67.08, 67.42 , and 66.75, for English, Portuguese and Spanish respectively.

8 Results

43 people that form around 30 teams were registered for Task 1. In total 238 submissions are prepared for the different subtasks and language combinations by 13 teams. The scores of the submissions are calculated on a Codalab page.[15] The teams that have participated are ALEM (Gürel and Emin, 2021), AMU-EuraNova (Bouscarrat et al., 2021), DAAI (Hettiarachchi et al., 2021), DaDeFrTi (Re et al., 2021), FKIE_itf_2021 (Becker and Krumbiegel, 2021), Handshakes AI Research (HSAIR) (Kalyan et al., 2021b), IBM MNLP IE (Awasthy et al., 2021), SU-NLP (Çelik et al., 2021), NoConflict (Hu and Stoehr, 2021) IIITT (Kalyan et al., 2021a), and NUS-IDS (Tan et al., 2021). Two participants that has the user names Jitin, and jiawei1998 on the Codalab page of the task did not write any description paper. [16]

We provide details of the results and submissions of the participating teams for each subtask in the following subsections.

Team	English	Hindi	Portuguese	Spanish
ALEM	80.82_{4}	N/A	72.98_{5}	46.47_{7}
AMU-EuraNova	53.46_{9}	29.66_{7}	46.47_{8}	46.47_{7}
DAAI	84.55_{1}	77.07_{3}	82.43_{2}	69.31_{4}
DaDeFrTi	80.69_{5}	78.77_{1}	77.22_{4}	73.01_{2}
FKIE_itf_2021	73.90_{7}	54.24_{6}	62.39_{6}	68.20_{5}
HSAIR	77.58_{6}	59.55_{5}	81.21_{3}	69.84_{3}
IBM MNLP IE	83.93_{2}	78.53_{2}	84.00_{1}	77.27_{1}
SU-NLP	81.75_{3}	N/A	N/A	N/A
NoConflict	51.94_{10}	N/A	N/A	N/A
jitin	67.39_{8}	70.49_{4}	52.23_{7}	62.05_{6}

Table 4: The performance of the submissions in terms of F1-macro and their ranks as a subscript for each language and each team participating in subtask 1.

8.1 Subtask 1

Subtask 1 results are provided in Table 4. The team "DAAI" has submitted the best results for English test data using a Big-Bird-RoBERTa. Team "DaDeFrTi" obtained the best score on Hindi data, which is a zero-shot cross-lingual setting by training a multilingual XLM-RoBERTa (XLM-R) based classification model with additional data either acquired from external data sets, collected from the web or translated from the original data. Finally, the "IBM MNLP IE" has ranked first for

[15] https://competitions.codalab.org/competitions/31247#results, accessed on May 26, 2021.

[16] The mapping between the team names and the Codalab user names is as follows: ALEM: alaeddin, AMU-EuraNova: lbouscarrat, DAAI: hansih, DaDeFrTi: davegh, FKIE_itf_2021: skent, Handshakes AI Research (HSAIR): vivekkalyanHS, IBM MNLP IE: kjbarker, SU-NLP:fcelik, NoConflict: pitehu, IIITT: AdeepH, and NUS-IDS: tanfiona

Team	English	Portuguese	Spanish
ALEM	79.67_{5}	42.79_{10}	45.30_{10}
AMU-EuraNova	75.64_{9}	81.61_{6}	76.39_{6}
DaDeFrTi	79.28_{6}	86.62_{3}	85.17_{2}
FKIE_itf_2021	64.96_{11}	75.81_{8}	70.49_{9}
HSAIR	78.50_{7}	85.06_{4}	83.25_{3}
IBM MNLP IE	84.56_{2}	88.47_{1}	88.61_{1}
IIITT	82.91_{4}	79.51_{7}	75.78_{7}
SU-NLP	83.05_{3}	N/A	N/A
NoConflict	85.32_{1}	87.00_{2}	79.97_{5}
jiawei1998	76.14_{8}	84.67_{5}	83.05_{4}
jitin	66.96_{10}	69.02_{9}	72.94_{8}

Table 5: The performance of the submissions in terms of F1-macro and their ranks as a subscript for each language and each team participating in subtask 2.

	Scores		
Team	English	Portuguese	Spanish
DAAI	80.40_{3}	90.23_{5}	81.83_{5}
FKIE_itf_2021	77.05_{6}	91.33_{3}	82.52_{3}
Handshakes AI Research	79.01_{4}	90.61_{4}	81.95_{4}
IBM MNLP IE	84.44_{1}	92.84_{2}	84.23_{1}
NUS-IDS	81.20_{2}	93.03_{1}	83.15_{2}
SU-NLP	78.67_{5}	N/A	N/A

Table 6: The performance of the submissions in terms of CoNLL-2012 average score Pradhan et al. (2014) and their ranks as a subscript for each language and each team participating in subtask 3.

	Scores		
Team	English	Portuguese	Spanish
AMU-EuraNova	69.96_{3}	61.87_{4}	56.64_{4}
Handshakes AI Research	73.53_{2}	68.15_{2}	62.21_{2}
IBM MNLP IE	78.11_{1}	73.24_{1}	66.20_{1}
SU-NLP	2.58_{5}	N/A	N/A
jitin	66.43_{4}	64.19_{3}	58.35_{3}

Table 7: The performance of the submissions in terms of F1 score based on CoNLL 2003 (Tjong Kim Sang and De Meulder, 2003) and their ranks as a subscript for each language and each team participating in subtask 4.

Portuguese and Spanish. The team trains three XLM-R based classification models that consist of ensemble of multiple models with various configurations.

Team "ALEM" compares mono-lingual and multilingual BERT based models, opting for monolingual models for English and Portuguese, and multilingual model for Spanish test data. Team "AMU-EuraNova" divides the given text into chunks with small overlaps and generates a prediction for each chunk in order to solve the length issue that multilingual BERT faces, but it receives a poor score due to the way they reduce multiple chunks' predictions into a final one. Team "FKIE_itf_2021" uses frozen multilingual BERT embeddings to train 100 small neural nets and ensemble them via majority voting. Team "Handshakes AI Research" trains a classification model with LaBSE embeddings. Team "SU-NLP" makes use of vanilla RoBERTa. Team "NoConflict" uses the same model they trained for subtask 2 to test for subtask 1 English data.

8.2 Subtask 2

Subtask 2 results are demonstrated in Table 5. Team "NoConflict" does extra pre-training of English only RoBERTa model on political news articles before finetuning on English training data to achieve first place for English test data. The best scores for Portuguese and Spanish were submitted by "IBM MNLP IE" by applying the same approach, which is multilingual training, they followed for subtask 1.

Team "DaDeFrTi" trains a multilingual XLM-R based classification model with additional data either acquired from external data sets, collected from the web or translated from the original data. Team "ALEM" compares mono-lingual and multilingual BERT based models, opting for monolingual models for all languages. Team "AMU-EuraNova" uses the same model as their subtask 1 solution, but it achieves reasonable scores this time due to majority of samples being smaller than their chunking size. Team "FKIE_itf_2021" uses frozen multilingual BERT embeddings to train a single small MLP. Team "Handshakes AI Research" trains a multilingual XLM-R based classification model. Team "SU-NLP" uses an ensemble of vanilla RoBERTa and a CNN model that's fed stemmed text as an extra channel. Team "IIITT" uses an ensemble of 3 classification models based on multilingual BERT, multilingual Distill BERT and English-only RoBERTa.

8.3 Subtask 3

The results of subtask 3 are reported in Table 6. The Team "IBM MNLP IE" submitted the best results for the test data in English and Spanish. This team applies agglomerative clustering with scores of pairs of sentences obtained by a XLM-R based model. Team "NUS-IDS" uses the clustering algorithm employed by Örs et al. (2020) with scores of pairs of sentences obtained by BERT based LSTM model with extra semantic features. Their multilingual model achieves first place for Portuguese and second place for Spanish test data. Their English-only model achieved second place for English test data.

Team "DAAI" uses different sentence transformers as a pairwise scorer and applies hierarchical

clustering algorithm, fine-tuning or training them from scratch. Team "FKIE_itf_2021" uses frozen multilingual BERT embeddings to train a pairwise scorer and applies a greedy clustering algorithm. Team "Handshakes AI Research" uses multilingual BERT embeddings to train a pairwise scorer and applies a greedy clustering algorithm. Moreover, they use extra data for English, and translations from English for Portuguese and Spanish data. Team "SU-NLP" uses an ensemble of 3 transformer based models as a pairwise scorer and applies the clustering algorithm proposed by Örs et al. (2020).

8.4 Subtask 4

We provide the results of the subtask 4 in Table 7. Results of the team "IBM MNLP IE" are by far the best for all languages. This team approaches this subtask as sequence labelling problem and fine-tunes a pre-trained language model (XLM-R large) with the data provided. The model they created using the training data for all languages ranked first for the test data in Portuguese and in Spanish. Their ensemble model that comprises five different English-only models performed the best for the test data in English.

Team "Handshakes AI Research" also considers subtask 4 as a sequence labelling problem and fine-tunes XLM-R multilingual using the Viterbi algorithm for the final classification. They use a previously defined technique to produce translations from English data to the rest of the languages, trying to mitigate the issue of smaller data size for Portuguese and Spanish. They achieved second place for English, Portuguese and Spanish test sets with this model. Team "AMU-EuraNova" uses the same chunking method with mBERT as their subtask 1 solution, but with extra stability experiments and behavioural fine-tuning with additional named entity data sets. Team "SU-NLP" trains a bidirectional LSTM on top of RoBERTa's contextualized word embeddings with conditional random fields.

9 Conclusion and Future Work

This shared task shows that multilingual and cross-lingual approaches perform surprisingly well for subtasks of protest event information collection. We observed that merging the training data from multiple languages improves the performance. Moreover, the performance of the first and second submissions, which are prepared by two different teams, for the cross-lingual zero-shot setting for subtask 1 in Hindi language are 78.77 and 78.53 in terms of F1-macro, thereby demonstrating the promising suitability of the approach for zero-shot multilingual setting. Another significant outcome is that "IBM MNLP IE" has outperformed all other teams by more than 4 points of F1-macro in all languages in subtask 4, which is the most challenging subtask.

Monolingual models outperforms multilingual models in case sufficient training data (Subtask 1, English) or additional further pre-training data is available (Subtask 2, English). These conditions are satisfied mostly for evaluation scenarios pertaining to English language. Although, multilingual models yield best performance in some scenarios, the monolingual models ranked second or third place.

Automated event information collection approaches are prone to major issues like bias toward majority class and popular content and limited generalizability that affect reliability and validity of them (Leins et al., 2020; Bhatia et al., 2020; Chang et al., 2019; Wang et al., 2016; Eck, 2021; Lorenzini et al., 2016; Schrodt, 2020; Raleigh, 2020; Boschee, 2021). We consider this benchmark as the first step to obtain comparable results across various automated approaches in a multilingual setting. We form a basis for increasing variety of the data that can be utilized for developing and evaluating event information collection systems by extending the language data that has various levels of availability such as few-shot and zero-shot settings. Furthermore, this benchmark allows determination of the most suitable text processing approaches for this task by identifying the performance levels that can be achieved applying recent technology. Last but not least, the random sampling of the corpus utilized in the shared task enables realistic recall quantification that has been challenging to measure to date (Hürriyetoğlu et al., 2021; Yörük et al., 2021).

We will be extending available training data and include additional data in different languages in the future iterations of this benchmark.

Acknowledgments

The authors from Koc University are funded by the European Research Council (ERC) Starting Grant 714868 awarded to Dr. Erdem Yörük for his project Emerging Welfare. Farhana Ferdousi Liza would like to acknowledge the support of the Business and Local Government Data Research Centre (ES/S007156/1) funded by the Economic and Social Research Council (ESRC) for undertaking this work.

References

Mikel Artetxe and Holger Schwenk. 2019a. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transac-*

tions of the Association for Computational Linguistics, 7:597–610.

Mikel Artetxe and Holger Schwenk. 2019b. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 task 1: Multi-granular 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)*, online. Association for Computational Linguistics (ACL).

Nils Becker and Theresa Krumbiegel. 2021. FKIE_itf_2021 at CASE 2021 Task 1: Using Small Densely Fully Connected Neural Nets for Event Detection and Clustering. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2020. You are right. i am alarmed–but by climate change counter movement. *arXiv preprint arXiv:2004.14907*.

Doug Bond, Joe Bond, Churl Oh, J. Craig Jenkins, and Charles Lewis Taylor. 2003. Integrated data for events analysis (idea): An event typology for automated events data development. *Journal of Peace Research*, 40(6):733–745.

Elizabeth Boschee. 2021. Keynote Abstract: Events on a Global Scale: Towards Language-Agnostic Event Extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Lo Bouscarrat, Antoine Bonnefoy, Ccile Capponi, and Carlos Ramisch. 2021. AMU-EURANOVA at CASE 2021 Task 1: Assessing the stability of multilingual BERT. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Furkan Çelik, Tuğberk Dalkılıç, Fatih Beyhan, and Reyyan Yeniterzi. 2021. SU-NLP at CASE 2021 Task 1: Protest News Detection for English. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

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. 2015. A pipeline for multilingual protest event selection and annotation. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 309–313.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Kristine Eck. 2021. Keynote Abstract: Machine Learning in Conflict Studies: Reflections on Ethics, Collaboration, and Ongoing Challenges. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457479.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, Martin Andrews, Hu Tiancheng, Niklas Stoehr, Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, Brenda Curtis, and Ali Hürriyetoğlu.

2021. Discovering Black Lives Matter events in the United States - Shared Task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Marco G. Giugni. 1998. Was It Worth the Effort? The Outcomes and Consequences of Social Movements. *Annual Review of Sociology*, 24:371–393.

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)*, online. Association for Computational Linguistics (ACL).

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China. Association for Computational Linguistics.

Jacek Haneczok, Guillaume Jacquet, Jakub Piskorski, and Nicolas Stefanovitch. 2021. Fine-grained event classification in news-like text snippets shared task 2, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Alex Hanna. 2017. MPEDS: Automating the Generation of Protest Event Data.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed 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)*, online. Association for Computational Linguistics (ACL).

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Erdem Yörük, Osman Mutlu, Deniz Yüret, and Aline Villavicencio. 2021. Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021): Workshop and Shared Task Report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, and Osman Mutlu. 2019a. A task set proposal for automatic protest information collection across multiple countries. In *Advances in Information Retrieval*, pages 316–323, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019b. 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, 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, Marseille, France. European Language Resources Association (ELRA).

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*, 3(2):308–335.

Swen Hutter. 2014. Protest event analysis and its offspring. In Donatella della Porta, editor, *Methodological Practices in Social Movement Research*, pages 335–367. Oxford: Oxford University Press, Oxford.

Pawan Kalyan, Duddukunta Reddy, Adeep Hande, Ruba Priyadharshini, Ratnasingam Sakuntharaj, and Bharathi Raja Chakravarthi. 2021a. IIITT at CASE 2021 Task 1: Leveraging Pretrained Language Models for Multilingual Protest Detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Sureshkumar Vivek Kalyan, Tan Paul, Tan Shaun, and Martin Andrews. 2021b. Shared Task 1 System Description : Exploring different approaches for multilingual tasks. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Hanspeter Kriesi, Bruno Wüest, Jasmine Lorenzini, Peter Makarov, Matthias Enggist, Klaus Rothenhäusler,

Thomas Kurer, Silja Häusermann, and Altiparmakis Patrice Wangen. 2019. Poldem–protest event dataset 30.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality and quantity*, 50(6):2347–2364.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Yan Liang, Khaled Jabr, Christan Grant, Jill Irvine, and Andrew Halterman. 2018. New techniques for coding political events across languages. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 88–93.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Jasmine Lorenzini, Peter Makarov, Hanspeter Kriesi, and Bruno Wueest. 2016. Towards a Dataset of Automatically Coded Protest Events from English-language Newswire Documents. In *Paper presented at the Amsterdam Text Analysis Conference*.

Peter Makarov, Jasmine Lorenzini, and Hanspeter Kriesi. 2016. Constructing an annotated corpus for protest event mining. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 102–107, Austin, Texas. Association for Computational Linguistics.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Peter F. Nardulli, Scott L. Althaus, and Matthew Hayes. 2015. A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data. *Sociological Methodology*, 45(1):148–183.

Clayton Norris, Philip Schrodt, and John Beieler. 2017. PETRARCH2: Another event coding program. *The Journal of Open Source Software*, 2(9).

Sean P. O'Brien. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12(1):87–104.

Faik Kerem Örs, Süveyda Yeniterzi, and Reyyan Yeniterzi. 2020. Event clustering within news articles. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 63–68, Marseille, France. European Language Resources Association (ELRA).

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Jacek Haneczok, and Guillaume Jacquet. 2020. New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6663–6678, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Clionadh Raleigh. 2020. Keynote abstract: Too soon? the limitations of AI for event data. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, page 7, Marseille, France. European Language Resources Association (ELRA).

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Francesco Re, Daniel Vegh, Dennis Atzenhofer, and Niklas Stoehr. 2021. Team "DaDeFrNi" at CASE 2021 Task 1: Document and Sentence Classification for Protest Event Detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Parang Saraf and Naren Ramakrishnan. 2016. Embers autogsr: Automated coding of civil unrest events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 599608, New York, NY, USA. Association for Computing Machinery.

Philip Schrodt and f Jay Yonamine. 2013. A guide to event data: Past, present, and future. *All Azimuth: A Journal of Foreign Policy and Peace*, 2:5 – 22.

Philip A. Schrodt. 2020. Keynote abstract: Current open questions for operational event data. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, page 8, Marseille, France. European Language Resources Association (ELRA).

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. 2020. Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221, Online. Association for Computational Linguistics.

Çağıl Sönmez, Arzucan Özgür, and Erdem Yörük. 2016. Towards building a political protest database to explain changes in the welfare state. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 106–110. Association for Computational Linguistics.

Ralph Sundberg, Kristine Eck, and Joakim Kreutz. 2012. Introducing the ucdp non-state conflict dataset. *Journal of Peace Research*, 49(2):351–362.

Fiona An Ting 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)*, online. Association for Computational Linguistics (ACL).

S. Tarrow. 1994. *Power in Movement: Social Movements, Collective Action and Politics*. Cambridge Studies in Comparative Politics. Cambridge University Press.

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 - Volume 4*, CONLL 03, page 142147, USA. Association for Computational Linguistics.

Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503.

Nils B. Weidmann and Espen Geelmuyden Rød. 2019. *The Internet and Political Protest in Autocracies*, chapter Coding Protest Events in Autocracies. Oxford Studies in Digital Politics, Oxford.

Erdem Yörük, Ali Hürriyetoğlu, Çağrı Yoltar, and Fırat Duruşan. 2021. Random Sampling in Corpus Design: Cross-Context Generalizability in Automated Multicountry Protest Event Collection. *American Behavioral Scientist*, 0(0):00027642211021630.

Han Zhang and Jennifer Pan. 2019. Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57.