Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction

Ali Hürriyetoğlu†, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret & Burak Gürel

Koç University, Rumelifeneri yolu, Sariyer, Istanbul 34450, Turkey

Keywords: Event extraction; Text classification; Political science; Social science; News; Contentious politics; Protests; Event coreference resolution

Citation: Hürriyetoğlu, A., et al.: Cross-context news corpus for protest event-related knowledge base construction. Data Intelligence 3(2), 308-335 (2021). doi: 10.1162/dint_a_00092
Received: November 30, 2020; Revised: February 26, 2021; Accepted: March 3, 2021

ABSTRACT

We describe a gold standard corpus of protest events that comprise various local and international English language sources from various countries. The corpus contains document-, sentence-, and token-level annotations. This corpus facilitates creating machine learning models that automatically classify news articles and extract protest event-related information, constructing knowledge bases that enable comparative social and political science studies. For each news source, the annotation starts with random samples of news articles and continues with samples drawn using active learning. Each batch of samples is annotated by two social and political scientists, adjudicated by an annotation supervisor, and improved by identifying annotation errors semi-automatically. We found that the corpus possesses the variety and quality that are necessary to develop and benchmark text classification and event extraction systems in a cross-context setting, contributing to the generalizability and robustness of automated text processing systems. This corpus and the reported results will establish a common foundation in automated protest event collection studies, which is currently lacking in the literature.

1. INTRODUCTION

Socio-political event knowledge bases enable comparative social and political studies. The spatiotemporal distribution of these events sheds light on the causes and effects of government policies and political discourses that resonate in society. Protest event databases have been of particular prominence in social

† Corresponding author: Ali Hürriyetoğlu (Email: ahurriyetoglu@ku.edu.tr; ORCID: 0000-0003-3003-1783).
and political science in this respect, and this is the type of event we focus on. We define protest events within the scope of contentious politics (CP), referred to as a “repertoire of contention” by Tilly et al. [1, 2], which includes demonstrations, riots, strikes, and many other types of collective action.

Since news media provide a continuous flow of data over time and enable researchers to determine the significance of events that are reported, social and political scientists turn to news data to create knowledge bases of protest events [3, 4, 5]. Protest event data collection has been carried out manually [6], semi-automatically [7], and automatically [8, 9, 10, 11]. While manual approaches tend to be too expensive, automatic methods have proven limited in the information they are able to collect [12, 13, 14]. Moreover, there has not been any common ground across projects that would enable a comparison of the results of distinct studies [15]. Therefore, as members of the Emerging Welfare (EMW) project, we took on the challenge of creating a common foundation in terms of the required high-quality data, and state-of-the-art tools for fully automating the creation of reliable and valid protest knowledge bases. This foundation would serve as a benchmark and enable protest event collection studies to benefit. This effort has yielded a gold standard corpus (GSC) that will also serve the machine learning (ML) and computational linguistics communities to study text processing tool development for constructing knowledge bases of protest events.

The Corpus consists of English news articles from various international sources and local sources\textsuperscript{\textregistered} from India, China, and South Africa. The variety of the sources has enabled researchers to study cross-context robustness and generalizability, which are critical requirements of the ML models, by addressing changes in style and content across sources. The annotations have been applied at document-, sentence-, and token-levels in a consecutive manner. Each individual level is annotated by two people, and adjudicated by the annotation supervisor. Moreover, detailed manual and semi-automatic quality checks and error analyses were applied to the annotations.

The Corpus contains more than 12,000 news articles labeled as protest and non-protest. Over 1,000 articles which contain protest events have undergone sentence-level annotation whereby each sentence in the article was labeled to see whether it contains event information. Finally, these articles were annotated at the token level for detailed event information such as the trigger(s), trigger semantic category, place, time and actor(s) of events, semantic category of the actor(s), and event coreference information. The corpus has enabled the development of a pipeline of ML models that extract a knowledge base of protest events from archives of news articles. Moreover, some parts of the GSC have enabled shared tasks for cross-context document and sentence classification as well as token extraction [16, 17]\textsuperscript{\textregistered} and event sentence coreference identification [18].

The contributions of this paper include 1) an effective methodology for creating a corpus that enables the creation of robust ML-based text processing tools, 2) insights from applying this methodology to news archives across contexts, 3) a corpus that contains data from multiple contexts and annotations at various:

\textsuperscript{\textregistered} International sources were filtered based on meta-information to focus on the case countries.

levels of granularity, 4) results of a pipeline consisting of automated tools that are created using the corpus, and 5) first recall quantification on unfiltered raw data as opposed to recall measurements on settings limited by keyword-filtered data, which is common in event knowledge base projects.

We describe the context of our work in reference to recent related work in Section 2. Next we introduce our methodology, the manuals we prepared, and the corpus we have created in Sections 3, 4, and 5, respectively. Section 6 reports the results of the ML tools that were created using the corpus. Finally, Section 7 concludes this report by presenting overall results and pointing to future steps we are planning to pursue in the future.

2. RELEVANT WORK

The quality of the automated socio-political event collection is determined by language resources, automated tools that exploit these resources, the assumptions made to complete the design of an event collection system, and data sets that are inputs or outputs of these systems. The existing language resources are rare and there are few accessible tools. Moreover, the assumptions made in delivering a resulting data set are not examined in diverse settings.

Automated tools for event information collection are designed in terms of pipelines that receive news articles from one or more news sources and yield records of event information. Each tool is inherently limited to the language resource it utilizes for development and the setting within which it is validated. Therefore, when an automated tool is used to analyze different sources, i.e., cross-context performance of these pipelines, quality of the result is rarely evaluated. The first step of these pipelines, which is discriminating between relevant and irrelevant documents, has been extensively studied by Croicu and Weidmann [19] and Hanna [20]. Keyword lists and labeled documents aid in determining which news reports contain relevant events. These studies provide their own keyword list and describe the way they use it. Moreover, labeled documents are presented as their URLs or document IDs in proprietary collections such as Lexis Nexis without their content [21]. Accessing the data set with such limited information, and the necessity of purchasing subscriptions to these databases are significant limitations. Our approach and database is novel in the sense that it does not restrict itself with keywords, and applies state-of-the-art ML models to tackle selecting documents that contain event information, which is known as the report selection problem in this field [4].

Once protest event-related documents are determined, what remains as a task is to extract event information on the token level. There are several established event ontologies and relevant language resources that can be exploited for this task, the most prominent of which are ACE [22], TAC-KBP [23], and CAMEO [24]. All three frameworks include wide ranges of event types and sub-types which can serve the needs of diverse domains. However, in terms of comparative social and political studies that focus on contentious politics, data based on ACE and TAC-KBP (Rich ERE) annotation frameworks are limited in size and scope. Both frameworks include ATTACK and DEMONSTRATE event categories as those relevant to the collection of CP events. The ATTACK category does not discriminate between authors of the actions
enumerated under ATTACK, i.e., attack, clash, and bomb, and thus includes state actions that are excluded from contentious politics. In other words, these event ontologies adhere only to syntactic rules, which do not allow the semantics of the event triggers and their arguments to affect the annotation. On the other hand, the ACE and Rich ERE category of DEMONSTRATE in itself, is too restrictive to be applicable in terms of a broad understanding of CP for two reasons. First, as it limits the scope of this event type to spontaneous (unorganized) gatherings of people, it excludes certain actions of political and/or grassroots organizations such as political parties and non-governmental organizations (NGOs). Protest actions of such organizations sometimes do not involve mass participation despite aiming at challenging authorities, raising their political agendas, or issuing certain demands. Putting up posters, distributing brochures, and holding press declarations in public spaces are examples of such protest events. Second, the requirement of mass participation in a public area leaves many protest actions such as online mass petitions and boycotts, which are not necessarily tied to specific locations where people gather, and actions of individuals or small groups such as hunger strikes and self-immolation. Unlike ACE and Rich ERE, the CAMEO framework seems to offer event categories better suited to the collection of CP events as it is directed towards the domain of international politics and governmental actions. CAMEO and its simplified version PLOVER includes the PROTEST category which includes most but not all CP event categories. Important elements of CP action, such as violent clashes, group confrontations, and armed attacks are found in other categories which include state actions as well as non-state, that is, CP events. In fine, the multi-purpose and comprehensive nature of existing language resources appears as shortcomings and complications in terms of a database that focuses exclusively on CP events.

Consequently, protest-event specific annotation schemas and data sets were proposed for creating automated or semi-automated event knowledge bases [4, 11, 24, 25]. In their corpus, Makarov et al. [26] used the ontology of CP events which is very similar to ours and identified ten event types. They coded event actors and issues that correspond to each event. Our definition is broader in the sense that we include election rallies as a form of CP events while they exclude it on the basis of being related to institutional politics. The Mass Mobilization in Autocracies Database (MMAD), on the other hand, defines protest events as specifically anti-regime political actions, thereby excluding industrial actions (strikes, etc.) and conflicts between social groups. They also exclude protest events with less than 25 participants and the systematic use of armed force. Apart from the differences of these projects from ours in terms of event ontology, these resources are mainly created using keywords for a single context, and it is a challenge to obtain the data sets based on the limited information shared. We follow the detailed protest event information tradition as proposed by Lorenzini et al. [15] and Gerner et al. [24], working on data unrestricted with keywords and making our data available to all researchers with shared tasks and sufficient information.

To sum up, recent studies on event data often assume one or more of the following 1) analyzing thousands or millions of sources will compensate for low recall performance of the tools, 2) a news report contains information about a single event, 3) analyzing a sentence individually is sufficient for extracting relevant information about an event, and 4) tool performance on a new source will be comparable to the performance on the validation setting [27, 28, 29]. Quantifying the effect of these assumptions is not a simple task; therefore, they are rarely tested. We make a point to provide observations that shed light on the effect of these assumptions, or we refrain from making assumptions at all.
3. METHODOLOGY

A GSC of protest events that can enable large-scale, multi-source socio-political studies should be representative of the content it aims to capture. Moreover, it should enable quantification of the automation performance across contexts; therefore, using available corpora such as English Gigaword [30] is not an option for this setting. In order to satisfy these requirements, our methodology is designed to incorporate multiple sources and countries and to apply a detailed annotation methodology without reducing the content quality.

We collected online local news articles and international sources when local sources were not accessible, from India, China, and South Africa. We first downloaded URLs of the freely accessible parts of online news archives of daily newspapers that include Indian Express (IEX), New Indian Express (NIEX), The Hindu (TH), Times of India (ToI), South China Morning Post (SCMP), and People’s Daily (PD). Then, for each source, we took a random sample of these URLs and downloaded their content for labeling and annotation.

The random sampling approach made the task challenging, as it contained few relevant documents compared to keyword filtered samples, yet it approximated the real data universe more accurately. Keyword lists run the risk of excluding events that are reported without the use of common protest terms (the phrase “classrooms remained empty” can be used to refer to a teachers’ strike, for instance) [16, 31, 32]. Moreover, lexical variance across contexts cannot always be captured using keywords. For instance, the phenomena such as “bandh” and “dharna” are protest event types that are specific to India, and thus, are not covered by any general-purpose protest keywords list. Our evaluation of four keyword lists, which are utilized by Huang et al. [25], Wang et al. [12], Weidmann and Rød [4], and Makarov et al. [26], yielded .68 and .80 precision and recall on our randomly sampled batches at best.

The annotation process of this randomly sampled raw data is based on an annotation manual that was created by an expert, and applied to document-, sentence- and token-level annotations for each particular target context while the annotation team was continuously monitoring the annotations to achieve a high inter-annotator agreement (IAA). The same manuals were applied to data collected from different sources and countries, enabling the attainment of comparable measures of automatic tool performance across contexts. Finally, in order to eliminate the risk of incorrect labeling due to lack of knowledge about a country, a domain expert in the politics of the target country instructed the annotators before they started the annotation.

The annotation team consisted of a supervisor, a social scientist responsible for maintaining the annotation manuals and resolving annotator disagreements. The annotators, who worked in pairs were master students.

Downloaded from http://direct.mit.edu/dint/article-pdf/3/2/308/1921154/dint_a_00092.pdf by guest on 30 June 2021
or Ph.D. candidates in social or political sciences. Throughout the annotation, the overlap ratio of annotated articles between pairs was 100%. The annotation started by labeling whether a news article mentioned a past or ongoing protest. Then, the sentence(s) that contained protest information was identified. Finally, protest information such as participants, place, and time was detected in the protest-related sentences at the token level. The three levels of annotation were separate but integrated in the sense that they formed a pipeline in which each batch of documents went through each step, which would build upon the result of the previous step. The aim here was to maximize time and resource efficiency and performance by utilizing the feedback of each level of annotation for the whole process. The lack of clear boundaries between these levels at the beginning of the annotation project caused a relatively lower IAA, and more time was spent on the quality check and correction of the data set. In order to ameliorate this, we added an additional step, which was sentence level annotation to the main steps of protest event pipelines. This order of tasks enabled the possibility for error analysis and optimization during annotation and tool development efforts.

Each batch for document and sentence levels has been corrected in terms of:

1) **Spot-checks**: 10% of the agreements were checked by the annotation supervisor

2) **ML-internal**: 80% of the batch was used to create an ML model. Next, the remaining 20% was predicted using this model. This procedure was repeated until all instances were used at least once in training and once in test data.

3) **ML-external**: The annotated and corrected data from previous batches were used to create an automated classifier, which then classified the newly annotated batch.

The disagreements between the classifiers and the annotations were checked manually for ML-internal and ML-external. Annotations that were checked based on spot-checks, ML-internal, and ML-external were found to be incorrectly annotated wrongly at around 2%, 50%, and 10%, respectively. In total, around 10% of the annotations were corrected using these measures.

In order to increase the time efficiency of annotation via increasing the number of relevant documents, we applied a recall-optimized active learning (AL) sampling from news archives by utilizing the initial ML-based classifiers (three or more) already trained on random samples. We followed this procedure in cases where we needed to improve the performance of a tool on a source that had already been covered, trained the tool on a new source from a country that was already reviewed, or adapted the tools to a different period. For these AL samples, we first trained multiple ML-based classifiers (three or more) on the available corpus, and then predicted a random batch from the new context. To achieve elevated recall scores, we took the logical OR of all classifiers as a final prediction, and selected positive samples to be annotated. Although the recall decreased from 100% to 97% in such a sample, the precision increased from around 5% to around 70% in comparison to a random sample where AL significantly decreased the

---

* This performance was measured on an AL sample that was predicted as positive and 200 news articles that were excluded from annotation as they were predicted as negative at this sampling operation. The training data consisted of around 4,000 news articles that were randomly sampled and annotated from the same country of the resulted AL batch.
effort the annotators needed to spend on annotation [33]. Since annotators encountered many more positive samples in this setting, we expected the decrease of recall to be a minor issue at the moment.

4. Annotation Manuals

Annotation tasks on document-, sentence- and token-levels each have their respective annotation manuals that define the task and establish the rules of annotation, enumerate cases that might be confusing, and elaborate on the rules in the context of concrete situations and examples.

**EMW Document-Level Protest Annotation Manual (DOLPAM)** has been created for document-level annotation. This manual lays out the protest event ontology; that is, the protest event definition which specifies the range of CP events that are included in the scope of the project. Also, it contains the rules by which the news articles are identified as containing CP events. In brief, CP events cover any politically motivated collective action that lies outside the official mechanisms of political participation associated with formal government institutions. This broad event definition is further developed in two sections. The first section identifies three abstract categories of collective action, namely, political mobilizations, social protests, and group confrontations to define the broad range of socio-political events that the project simply refers to as protest events. Next, five specific categories of CP events are identified as concrete manifestations of the types of collective action already defined. Demonstrations, industrial actions, group clashes, political violence, armed militancy, and electoral mobilization events are the concrete types of events that our event ontology encompasses. Once the event definition is laid out, the manual establishes certain criteria for determining which news stories that report protest events can be classified as protest news articles. These criteria include the necessity of civilian actors, and the existence of concrete time and place information which confirms that the event(s) the report mentions has indeed taken place. Only the news reports that mention events that have taken place in the past or are taking place at the time of writing are labeled as protest news articles. The references to the future (i.e., planned, threatened, announced or expected) events are not labeled as protest with the exception of threats of violent events or attempts to carry them out.

**EMW Sentence Level Protest Annotation Manual (SELPAM)** establishes rules for determining the classification of sentences of news reports as protest event sentences and non-event sentences. Similar to document-level annotation, sentences which contain references to protest events are labeled as protest event sentences. These are defined as sentences which give information about an event in the news report also containing at least one direct reference to a protest event; that is, all event sentences must contain an expression that denotes the event.

**EMW Token Level Protest Annotation Manual (TOLPAM)** acts as the guide to the annotation on the token level. The TOLPAM defines all variables and pieces of information about protest events that the EMW

---

The annotation manuals can be found on https://github.com/emerging-welfare/general_info/tree/master/annotation-manuals.

Although planned events and protest threats could have a role in our analysis [24], they are neither relevant in the CP context nor their prevalence, which is below 0.5% of a random sample according to our observations, allows their automated analysis.
project aims at extracting, and it establishes the rules according to which expressions in the event sentences are annotated using tags. Event arguments that exist within the text are only tagged in event sentences. This is to ensure that arguments belong to their respective events unambiguously, which improves IAA significantly. There are general rules which apply to all tags, as well as specific rules which apply to individual tags.

There are two main categories of tags: syntactic and semantic tags. Syntactic tags label expressions according to whether they are triggers (events) or event arguments. They are grouped under event, participant, organizer and target characteristics. Event characteristics contain the trigger tags, i.e., event expressions which either directly denote the event (event anchor) or refer to it (event mention) as well as tags for time, place, facility, and centrality (i.e., whether it is urban or rural) arguments of events. Each event can have only single anchor and zero or more event mentions that refer to the event anchor. Actor arguments of events are labeled under two categories: participants and organizers. Participants are personal actors (i.e., individuals or groups) who actively engage in the protest action. Organizers are organizations (political parties, NGOs, unions, etc.) that hold or take part in the protest events. In some cases, influential individuals or leaders might be the organizers of the protest events. Persons are annotated as organizers only in special cases where the article designates them explicitly as organizers or leaders of the protests. Each actor argument is labeled with participant (or organizer) type, name, ideology, religious, ethnic and caste identity, and socioeconomic status labels. Finally, target arguments of events are annotated with target type and target name labels. These labels also designate the possible antagonists of the protest events including governments, officials, leaders, political organizations or other social groups in the case of group clashes.

Semantic tags classify events and participant and organizer arguments into sub-types identified per the requirements of the EMW project. Every event trigger, participant actor, and organizer actor is labeled with a syntactic tag and one of the semantic sub-type tags. For events, semantic tags correspond to types of collective action. Demonstrations (rallies, marches, sit-ins, slogan shouting, gatherings, etc.), industrial actions (strikes, slow-downs, picket lines, gheraos, etc.), group clashes (fights, clashes, lynching, etc.), armed militancy (attacks, bombings, assassinations, etc.) and electoral politics (election rallies) are the sub-types. Participant expressions are categorized semantically into peasant (people who work in agriculture and/or live in rural areas), worker (any kind of public and private sector worker, blue and white collar), small producer (owners of small shops, small traders, artisans including transport owners like 4 owner taxi drivers), employer/executive (owners and managers of medium and large-sized businesses), professional (university-educated professionals such as physicians, lawyers, academics, and journalists, who work in a private or public sector), student (students from all levels of education), people (general categories which refer to citizens like women, residents, religious or ethnic community members, expressions such as mob and crowds), activist (ordinary members of political parties, grassroots organizations, NGOs), politician (members of political parties who are members of legislative organs and/or executive branches of government), and militant (members of armed political organizations such as Islamic fundamentalist militants and members of armed revolutionary organizations) sub-types. Any category of participants who cannot be placed into any of the above categories is marked as ‘other.’ Organizer sub-types are political party, NGO, trade union, armed/militant organization, and chamber of professionals.
Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction

Below are examples of event information annotated at the token level. The bold tokens are the event triggers. The underlined tokens are event arguments which are listed in order as event time (sentence 2); event time, organizer name, participant type, facility type, target name, event time, facility name, and participant type (sentence 3); as well as event place, participant count, participant type, and facility type (sentence 4). In the first sentence, the event triggers are also labeled with the semantic sub-type label “group clash” as the event is a communal clash. It should be noted that an event that has not taken place, the rally, in the second sentence is not annotated. The tokens that refer to the event that took place are also labeled with semantic sub-type “demonstration.” In the third sentence, there are two event references. ‘Gathered’ and ‘shouted slogans’ are the triggers for the first event, which are also labeled with the “demonstration” semantic tag. The triggers of the second event reference are “attack” and “killed,” which are labeled with the “armed militancy” semantic sub-type tag®. The actor arguments in the third sentence are labeled with semantic sub-types “political party” (‘BJP’), “activist” (‘workers’), and “militant” (‘militants’). In the fourth sentence, event triggers are labeled with the “demonstration” semantic tag, while the participant actor token (‘workers’) is labeled with the “worker” semantic tag.

1. It took a communal turn that had resulted in stone-pelting, arson and loot.
2. The Bhim Army and other Dalit groups were refused permission to organize a rally against atrocities on May 9, sparking off violence and vandalism, with several vehicles and buses burnt.
3. At noon, BJP workers gathered in the square and shouted slogans, condemning the failure of the Union Government in delivering justice to the victims of last year’s terror attack at the train station where armed militants killed 25 people.
4. In Bangalore, hundreds of workers participated in the rally in front of the collectorate.

5. CORPUS

We have annotated the corpus at three levels: document, sentence, and token. The document level refers to what a reader sees on a news article. It consists of a title, publication time, and the article text. The sentence level refers to a text unit that ends with a sentence-completing punctuation mark. The token level refers to a punctuation mark or sequence of alphanumeric characters that is characterized as a word in English. The following subsections describe the number of instances for each level, the quality of the annotations, and the storage and release of the corpus.

5.1 Data Statistics

The document counts for each document level batch are reported in Table 1. Each batch is named after the source it was sampled from. In cases where we annotate data from a corpus such as EventStatus [25]

® We treat the event triggers and any other expressions that have a hyphen between them as a single token, e.g., ‘stone-pelting’. But, when there is not a hyphen between words, which is the case for ‘shouting slogans’, the expression consists of two tokens. The first token is annotated as B-trigger and the following token(s) are annotated as I-trigger.
Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction

(ES) and RCV1 [34] that are readily released, we use their names as batch names. Suffixes are added to distinguish between different batches from the same source. For instance, SCMP1 and SCMP2 differ in terms of the period they cover, which is 2000–2002 and 2000–2017, respectively.

Table 1. Document label statistics and sampling method.

<table>
<thead>
<tr>
<th></th>
<th>ES</th>
<th>INT</th>
<th>IEX</th>
<th>NIEX</th>
<th>PD</th>
<th>RCV1</th>
<th>SCMP1</th>
<th>SCMP2</th>
<th>TH</th>
<th>Tol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protest</td>
<td>151</td>
<td>262</td>
<td>296</td>
<td>71</td>
<td>69</td>
<td>802</td>
<td>17</td>
<td>19</td>
<td>264</td>
<td>481</td>
</tr>
<tr>
<td>Non-Protest</td>
<td>149</td>
<td>738</td>
<td>265</td>
<td>630</td>
<td>732</td>
<td>367</td>
<td>985</td>
<td>483</td>
<td>782</td>
<td>1,985</td>
</tr>
<tr>
<td>Sampling</td>
<td>K</td>
<td>AL</td>
<td>AL</td>
<td>R</td>
<td>R</td>
<td>AL</td>
<td>R</td>
<td>R</td>
<td>AL</td>
<td>R&amp;AL</td>
</tr>
</tbody>
</table>

Note: K, R, and AL indicate Keyword, Random, and Active Learning, respectively. ES, INT, and RCV1 are EventStatus, International, and Reuters which is filtered for China using meta-information.

Active learning was applied to creating three batches of articles, which are INT2 (Guardian), SCMP3, and NIEX2 (New Indian Express) that were annotated at the sentence level and which are reported in Table 2. We sampled the full documents and annotated their sentences. The high number of non-protest sentence annotations stems from documents that do not contain any protest information.

Table 2. Sentence level statistics.

<table>
<thead>
<tr>
<th></th>
<th>INT2</th>
<th>SCMP3</th>
<th>NIEX2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protest</td>
<td>1,658</td>
<td>511</td>
<td>1,299</td>
</tr>
<tr>
<td>Non-protest</td>
<td>9,045</td>
<td>2,847</td>
<td>7,083</td>
</tr>
</tbody>
</table>

Note: The total number of sentences and their annotations as protest and non-protest is reported.

Sentences of a subset of the positive documents were annotated at the token level. The number of the information types in the annotated documents, which are 704 and 135 from India and China, respectively, are reported in Table 3. A news article was annotated at the token level only if the event happened or was happening in the same country of the source or the country under focus for international sources because protest event characteristics that are different across countries can affect the quality of the annotation.

Table 3. Token level statistics and IAA in terms of Krippendorf’s alpha.

<table>
<thead>
<tr>
<th>Tag name</th>
<th>Time</th>
<th>Trigger</th>
<th>Place</th>
<th>Facility</th>
<th>Participant</th>
<th>Organizer</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>822</td>
<td>1,378</td>
<td>645</td>
<td>392</td>
<td>2,283</td>
<td>1,260</td>
<td>1,453</td>
</tr>
<tr>
<td>China</td>
<td>144</td>
<td>142</td>
<td>82</td>
<td>52</td>
<td>272</td>
<td>88</td>
<td>109</td>
</tr>
<tr>
<td>IAA</td>
<td>60.07</td>
<td>50.02</td>
<td>41.82</td>
<td>39.10</td>
<td>39.50</td>
<td>47.44</td>
<td>34.38</td>
</tr>
</tbody>
</table>

Around 10% of the positively annotated documents at the document level in a random sample reports a protest event that does not occur in a country under focus.
The country-specific total count of the documents, the separate events in these documents, and the triggers referring to these events are presented in Table 4. The number of triggers that refer to multiple events is provided as well. Each document contains 1.8 events and each event is described using 3.1 triggers on average across the countries in the scope. The overall percentages of events in the documents, triggers in events, and tokens in triggers are demonstrated in Figure 1. The figure shows that 41%, 65%, and 18% of the documents, events, and triggers contain more than one event, trigger, and token, respectively. These statistics are comparable across data collected from each country and jointly illustrate the first quantification of the phenomenon of reporting multiple events in a news article, using multiple triggers to describe an event, and bringing together multiple tokens that denote a trigger in the socio-political events domain. This information indicates the importance of coreference information for event information collection tasks and shows that the assumption that news articles contain event information only in the title, leading sentences, or in an indexed summary inherently leads to disregarding a significant amount of event information. Examples of this assumption were made for instance by Tanev et al. [28] and Jenkins et al. [35]. This was also observed by Johnson et al. [29] using extrinsic evaluation of protest event databases.

Table 4. Number of events in a document.

<table>
<thead>
<tr>
<th>Country</th>
<th>#documents</th>
<th>#events</th>
<th>#triggers</th>
<th>#multi event triggers</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>639</td>
<td>1,171</td>
<td>3,577</td>
<td>71</td>
</tr>
<tr>
<td>China</td>
<td>73</td>
<td>123</td>
<td>416</td>
<td>11</td>
</tr>
<tr>
<td>South Africa</td>
<td>184</td>
<td>357</td>
<td>1,060</td>
<td>16</td>
</tr>
</tbody>
</table>

Figure 1. The overall percentages of events, triggers, and tokens in documents, events, and triggers, respectively.

The distribution of the semantic event categories for each country in scope and the overall percentage of the categories in the data set are provided in Table 5. Most of the events fall into the demonstration and armed militancy categories. The remaining categories comprise only 28% of the events. The event category incidence ratios are imbalanced both in individual countries and across countries. This difference is one potential reason for the performance gap of the automated tools for event information collection in cross-context settings [17].
Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction

Table 5. Number and percentage of demonstrations, political violence and armed militancy, group clashes, industrial actions, and electoral mobilization of events in a document.

<table>
<thead>
<tr>
<th></th>
<th>#demonstration</th>
<th>#armed militancy</th>
<th>#group clash</th>
<th>#industrial act</th>
<th>#electoral politics</th>
<th>#other</th>
</tr>
</thead>
<tbody>
<tr>
<td>India</td>
<td>1,177</td>
<td>787</td>
<td>656</td>
<td>366</td>
<td>62</td>
<td>20</td>
</tr>
<tr>
<td>China</td>
<td>358</td>
<td>64</td>
<td>50</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>South Africa</td>
<td>269</td>
<td>39</td>
<td>57</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Overall ratio</td>
<td>53%</td>
<td>19%</td>
<td>17%</td>
<td>8%</td>
<td>2%</td>
<td>1%</td>
</tr>
</tbody>
</table>

The semantic categories of the actors, which are participants and organizers, are annotated at the token-level for data collected from India. The annotated participant categories and their counts are Activist in 416, People in 416, Militant in 275, Worker in 249, Professional in 153, Student in 146, Politician in 118, Peasant in 54, and Small Producer in 44 cases. The organizer counts are 518 for Political Party, 267 for NGO, 139 for Labor Union, 123 for Militant Organization, and 47 for Chamber of Professionals. There are 409 participant and 161 organizer annotations, mostly first and last names, that were not semantically labeled, because they do not carry any semantic information per our annotation schema. Finally, the Employer/Executive category labels have been removed since they emerged only four times in the whole data set.

5.2 Corpus Quality

The inter-annotator agreement (IAA) for document and sentence levels is above .75 and .65 Krippendorf’s alpha [36] on average, respectively. The IAA for the token level is less consistent than other levels, which can be seen in Table 3 for each information type. We interpret these scores as an indication of the difficulty of the task and the extent to which the annotation manuals can facilitate consistent annotation.

The semantic information for event triggers, participants, and organizers were annotated after the token level annotations were completed and adjudicated for most of the documents. The annotations for the semantic category of the event trigger were applied following the double annotation and quality control approach at the token level. The IAA for the semantic category of the event trigger is .86 and .85 for data from India and China, respectively. However, the semantic annotations for participants and organizers were annotated at the token level by only one experienced annotator by resolving any issue with the annotation supervisor. Hence, these were the only annotations for which we could not provide an IAA score.

We selected the evaluation method proposed by Denis and Baldridge [37] and used in the CoNLL-2012 shared task [38] for measuring the quality of the event coreference annotation at the token level. We assumed the annotations of one annotator as the gold standard truth and annotations of the other annotator as predictions. We repeated this procedure for each annotator, and calculated the average of these scores. The average is .77, which we believe to be sufficient for modeling purposes.

* An event sentence may contain more than one participant or organizer. Therefore, the event sentence and annotation counts do not match.
The quality of all annotations in the corpus was improved semi-automatically. Adjudications, spot-checks, and manual analysis of the system predictions allowed us to fix at least 10% of the annotations for each task. The agreement scores should be perceived as an indication of the complexity of the concepts we attempted to process automatically.

In short, we suggest serious consideration of the following factors for creating a high-quality cross-context corpus that reliably captures a certain phenomenon, in a way that it can be modelled using machine learning models:

1. Variety of sources: Each different data source enriches the representation of the phenomena in the corpus. Use of the automated tools in any context that is not represented in the corpus will yield outputs that are less reliable and less valid. Therefore, each target context must be represented with at least one source in the corpus.

2. Random sampling: Raw data demonstrate how the target phenomenon is presented. Therefore, working with random samples is critical until an operational setting that can quantify performance of any other method of sampling the documents to be annotated, such as keyword selection or active learning, can be established.

3. Supervisor: Consistency is the key element of automated tool development. Therefore, an expert, preferably the same person throughout the whole corpus preparation task, should prepare and maintain the annotation manual, train the annotators, adjudicate and spot-check the annotations, and double-check the incompatibilities between automated predictions and manual annotations.

4. Annotation manual: The first version of an annotation manual should be a minimally viable product, consisting of a generic description of the target phenomena and basic instructions for the annotators. It should be updated as more data are observed throughout the annotation. In case of backward incompatible updates, the previous annotations should be updated semi-automatically using data annotated with the new manual.

5. Annotating at various levels: The text should be annotated at multiple levels, e.g., document, sentence, and token, consecutively in a way that ensures the quality of the annotations at each individual level. Any annotation error detected in the preceding level should be corrected. For instance, a result of sentence level annotation, e.g., lack of relevant sentences, can indicate the error committed on document level annotation, e.g., relevant label for the document.

6. Tracking irrelevant Information: Identifying relevant and irrelevant information is equally important for a gold standard corpus. This information will enhance the automated tool performance and enable quantification of recall. If any update is required for the annotation manual, the irrelevant documents should be considered to be relevant only if they satisfy the new conditions.

7. Training of the annotators: Each context should be understood as well as possible by the people who will analyze data and make decisions. Therefore, annotators should be trained on the contexts within the scope. This practice increases consistency and decreases time spent on supervision of the annotators and adjudication of the annotations.

8. Double annotation: The formal, abstract description of the phenomena in the annotation manual may differ drastically from how it is referred to within the text. Annotators’ respective backgrounds may also affect their interpretations. Therefore, each annotation should be performed by at least two
annotators, the annotators in a pair should be changed frequently, and different pairs of annotators should work on individual levels of annotation for each batch of documents.

9). Multiple expert annotators: Cross-context work on a non-trivial phenomenon requires the attention of multiple experts. Therefore, an annotation team must consist of more than two annotators, which would prevent systematically erroneous agreements. Multiple annotation teams would further necessitate an expert annotation supervisor to ensure consistency and compliance with the manuals. Regular and consistent communication between annotators and the supervisor is critical for assuring consistency and reducing time and effort spent on retrospective corrections.

10). Adjudications and spot-checks: Annotation disagreements between annotators should be adjudicated by the annotation supervisor. Moreover, some portion of the annotation agreements should be double-checked as spot-checks. Each disagreement or incorrect annotation provide insights to the supervisor about the data, the annotation manual, and the target phenomena. Errors and disagreements should be considered feedback and serve to inform changes to the annotation manuals during their maintenance. Annotation manuals should be updated recurrently in light of feedback from annotators, though not in a way that would reverse existing rules or create inconsistencies.

11). Inter-annotator agreement (IAA): Keeping track of the IAA scores serves as a guide for determining points to be improved in the annotation manual, revealing differences in approaches and viewpoints of annotators, improving the rules of annotation, and even adjusting the annotation tasks. These scores also serve as information sources about the performance expected when automated tools that handle the target phenomena are created. High IAA will enable good performance.

12). Semi-automated error correction: Batches of documents should be used to train and test ML models as soon as their annotation and quality assurance is complete. Any inconsistencies between the model predictions and the annotations must be double-checked by the annotation supervisor. This step frequently enables the identification of annotation errors caused by disagreement between annotators.

13). Determination of the baseline performance: State-of-the-art ML models should be trained and validated on the corpus. The performance scores on a held-out test data and leave-one-context-out setting offers information about the usability of the corpus for the contexts in the scope of the study. We suggest following all of these essential steps particularly at the beginning of a text annotation study. Some steps may be altered as the project progresses. For instance, keywords or active learning can be used to select documents that will be annotated only after testing them on the randomly sampled and annotated data. Another example is that if the annotation manuals become complete and stable after first iterations of the annotation, the training of the annotators could be relatively simpler in next iterations.

5.3 Corpus Release

The document and sentence level annotations are stored in JSON and token level data are stored in FoLiA [39] formats. We have distributed the corpus in a way that does not violate the copyright of the news sources. This involves only sharing information that is needed to reproduce the corpus from the source in
cases news article distribution may be problematic. To this end, the document and sentence level data can be downloaded using software we developed and packaged in a Docker image. These software tools download, clean, and align text and are provided in a Docker image in order to facilitate ease of use and reproducibility. The validation of this software was performed during the aforementioned shared tasks.

Protest database building demands special attention to ethical dimensions of research because of the risks associated with political dissent. We have developed a data sharing protocol to minimize these risks and prevent the use of our data for malicious purposes against activists. Our data sharing policy is shaped by our belief in the power of scientific collaboration and research transparency for theory and method development as much as our concern about the well-being of precarious groups and/or individuals that are engaged in political activities for social change. In order to balance these sometimes-conflicting priorities, we have embraced a two-tier data sharing method. At the first layer, we share a processed and limited version of our data in a visualized form on an open-access website of the GLOCON Project. The protest data visualized on country maps on this website provide only macro-level information about the main protest categories, such as year, event-place (province or city), event-type (five major aggregated event-type semantic categories we have produced: demonstrations, industrial actions, group clashes, armed militancy, and electoral politics), urban/rural locations, ethnicity, and ideology. At the second layer, we share our detailed protest data sets (including detailed event, organizer, participant, place, facility and time information) and our GSC and computational tools only with researchers and parties (including social movements themselves) who comply with ethical standards of social movements research and norms for the protection of research subjects. In line with this policy, we do not share our detailed data with government or law enforcement agencies or researchers who collaborate or receive funding from intelligence or defense agencies.

6. EVALUATION

We have exploited the data from India to train ML-based models using BERT [40] for document and sentence classification and token extraction in various scenarios. We fine-tuned the pre-trained BERT-based model with our data. The hyperparameters are the same as those of the original authors of each model for each level except for our sentence classifier which restricts maximum sequence length to 128 instead of 512. Table 6 provides F1-macro scores of the document and sentence classification and an F1-score that is based on the CoNLL 2003 evaluation script for token extraction models. The test data for India are a held-out sample because the training data are from India and all data for China and international data in the corpus. The row indicating “Token” is only the trigger detection performance in this table.

* Please follow the instructions on the Global Contentious Politics Gold Standard Data (GLOCON GOLD) repository to obtain the corpus: https://github.com/emerging-welfare/glocongold.
* https://glocon.ku.edu.tr/, accessed on April 7, 2021.
* For our document model, in inference time, if we split the text in subparts smaller than 512 tokens and take the logical or of each subpart’s prediction as that document’s prediction, the performance increases 2-3 F1 macro points in comparison to just using the first 512 tokens in a document for prediction.
* The international data are our Guardian sample that is filtered using active learning for China.
Table 6. F1-macro of document and sentence classification and F1 for trigger detection.

<table>
<thead>
<tr>
<th></th>
<th>India</th>
<th>China</th>
<th>Int-China</th>
<th>South Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>.89</td>
<td>.82</td>
<td>.83</td>
<td>.85</td>
</tr>
<tr>
<td>Sentence</td>
<td>.85</td>
<td>.79</td>
<td>.83</td>
<td>.85</td>
</tr>
<tr>
<td>Token</td>
<td>.74</td>
<td>.67</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The token level scores are based on the BERT-based model fine-tuned on our GSC and are generated using a held-out part of it (see Table 7). Additionally, we fine-tuned the Flair NER model [41], which is trained on CoNLL 2003 NER [42] with our data by mapping our place, participant, and organizer tags to “LOC”, “PER”, and “ORG” in CoNLL data, respectively. In comparison to the BERT-based model, this model yielded significantly better results of .780, .697, and .652 for the place, participant, and organizer types, respectively. Finally, we ran the same test data through an event extraction model, which was also a BERT-based model trained on ACE event extraction data. We measured the trigger detection performance of this model based on its CONFLICT category predictions. The F1 scores of the CONFLICT type are .543 for its own data and .479 for our new data. The difference between the scores obtained using ACE and our training data show that our efforts significantly contribute to the protest event collection studies.

Table 7. Token level information extraction scores per information type.

<table>
<thead>
<tr>
<th></th>
<th>Trigger</th>
<th>Time</th>
<th>Place</th>
<th>Facility</th>
<th>Participant</th>
<th>Organizer</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>.756</td>
<td>.663</td>
<td>.724</td>
<td>.436</td>
<td>.649</td>
<td>.568</td>
<td>.497</td>
</tr>
<tr>
<td>Recall</td>
<td>.691</td>
<td>.704</td>
<td>.646</td>
<td>.436</td>
<td>.564</td>
<td>.619</td>
<td>.485</td>
</tr>
<tr>
<td>F1</td>
<td>.722</td>
<td>.683</td>
<td>.683</td>
<td>.436</td>
<td>.604</td>
<td>.593</td>
<td>.491</td>
</tr>
</tbody>
</table>

We integrate the tools reported in Table 6 and report their performance in Table 8 on a separate data set of 200 news articles, which consist of 100 positively and 100 negatively predicted documents at the document level from India*. Doc, Sent, and Tok correspond to the tool applied in the order the tool name is mentioned in the configuration name. The highest precision, recall, and F1-macro was yielded by Doc+Sent+Tok, Tok, and Doc+Tok, respectively. The event trigger detection score is the reported one for the Tok. The performance of the trigger detection is lower than the one reported in Table 7 since this evaluation setting contains non-protest documents. The obvious result is that each additional component improves precision but decreases the recall. The interesting result here is that integrating only the document classification tool enhances precision and a slight decrease in recall in comparison to other configurations.

Table 8. Trigger detection performances in various configurations of a pipeline.

<table>
<thead>
<tr>
<th></th>
<th>Tok</th>
<th>Sent+Tok</th>
<th>Doc+Tok</th>
<th>Doc+Sent+Tok</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>.624</td>
<td>.696</td>
<td>.660</td>
<td>.701</td>
</tr>
<tr>
<td>Recall</td>
<td>.663</td>
<td>.561</td>
<td>.647</td>
<td>.547</td>
</tr>
<tr>
<td>F1</td>
<td>.643</td>
<td>.621</td>
<td>.653</td>
<td>.614</td>
</tr>
</tbody>
</table>

* We exclude 15 documents that contain events not related to India.
Additional results that were yielded using some parts of this corpus can be found in various publications. The results obtained in a shared task for the cross-context document and sentence classification and token extraction were reported in the overview paper of the ProtestNews Lab [17], which was held in the scope of Conference and Labs of the Evaluation Forum (CLEF 2019). Participants of this shared task reported comparable results to the performance reported in this paper. Moreover, the data set was facilitated in the event sentence coreference identification task in which the participants developed systems to identify sentences about the same event in the scope of the workshop Automated Extraction of Socio-political Events from News (AESPEN) at Language Resources and Evaluation Conference (LREC 2020) [18].

6.1 Performance of Semantic Classification Models

We created three sentence classification models exploiting data from India for the semantic categories of the event trigger, participant, and organizer. The sentence-level labels were inferred from token level labels. A BERT-based model with the maximum sequence length of 128 was fine-tuned for each of these tasks using the same BERT settings reported above.

The data exploited for the creation and evaluation of the event trigger semantic classification model is reported in Table 9. The sentences that contain event triggers that belong to multiple semantic categories were excluded from all subsets of the data, which are train, development, and test.

<table>
<thead>
<tr>
<th>Class</th>
<th>Train</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armed Militancy</td>
<td>361</td>
<td>48</td>
<td>72</td>
</tr>
<tr>
<td>Demonstration</td>
<td>847</td>
<td>113</td>
<td>169</td>
</tr>
<tr>
<td>Electoral Politics</td>
<td>31</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Group Clash</td>
<td>278</td>
<td>37</td>
<td>56</td>
</tr>
<tr>
<td>Industrial Action</td>
<td>202</td>
<td>27</td>
<td>41</td>
</tr>
<tr>
<td>Other</td>
<td>11</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

The performance scores of the model for trigger semantic categorization are provided in Table 10. The data points labeled as “Electoral Politics” and “Other” are changed to “Demonstration” since their size is relatively small and the model that we trained by including these classes yielded an F1-score of only .60. The final model yielded relatively low scores for the Group Clash category. Our investigation via a confusion matrix showed that the model most frequently confuses the Group Clash category with the Demonstration category. We observed that the demonstration and group clashes simultaneously occur frequently in the data.

Next, we prepared the data for the participant and organizer semantic classification models. The “No” label, which is maintained as a separate class, is attached to the event sentences that do not contain any semantic information pertaining to participants or organizers. There are three kinds of cases where the “No” label is used. First, participant or organizer annotation may not occur in an event sentence, as is the case
in 1,018 and 1,776 instances, respectively. Second, some first and last names do not indicate any semantic
information, as in 99 and 173 cases, respectively. Third, some participants or organizers may carry semantic
information that is not covered in our annotation schema. For instance, managers and employers, which
occur only four times in the data set, are not covered in the participant categories. The remaining sentence
counts were 1,336 for participant and 1,094 for organizer categories. We observed 139 sentences that
contained multiple participant semantic categories and 38 that contained multiple organizer semantic
categories. We consider these sentences to be conflicting and have thus excluded them from training and
development sets when we created an ML model using these data. However, we included these data in
the test by repeatedly adding the same sentence for each semantic type that occurs for a fair evaluation.
Consequently, there are 387 and 377 unique sentences out of 422 and out of 385 samples in participant
and organizer test sets, respectively.

Table 10. The precision, recall, and F1-macro for the semantic event trigger classification at the sentence level.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armed Militancy</td>
<td>.85</td>
<td>.95</td>
<td>.90</td>
</tr>
<tr>
<td>Demonstration</td>
<td>.82</td>
<td>.90</td>
<td>.86</td>
</tr>
<tr>
<td>Group Clash</td>
<td>.74</td>
<td>.51</td>
<td>.61</td>
</tr>
<tr>
<td>Industrial Action</td>
<td>.90</td>
<td>.73</td>
<td>.81</td>
</tr>
<tr>
<td>Overall average</td>
<td>.83</td>
<td>.77</td>
<td>.79</td>
</tr>
</tbody>
</table>

Table 11 demonstrates the data sizes of each semantic participant class and the performance scores in
terms of precision, recall, and F1-macro. The average score is .60 for all of these metrics. The performance
of the model is the best for the militant and activist categories and the worst for the politician and small producer categories. Although the data size for a class affects the performance, we observe that the peasant category predicted better than the politician category, which has more data. Our analysis using a confusion matrix shows that six out of 14 instances of politician annotations are predicted as activist. Thus, any further work on this task should focus on both increasing data for each class and revising the annotation schema.

Table 11. Data size and performance scores for the semantic participant classification at sentence level.

<table>
<thead>
<tr>
<th></th>
<th>Data set</th>
<th>Performance</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Development</td>
<td>Test</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>People</td>
<td>234</td>
<td>31</td>
<td>59</td>
<td>.60</td>
<td>.61</td>
</tr>
<tr>
<td>Militant</td>
<td>167</td>
<td>22</td>
<td>39</td>
<td>.83</td>
<td>.76</td>
</tr>
<tr>
<td>Activist</td>
<td>236</td>
<td>32</td>
<td>62</td>
<td>.61</td>
<td>.70</td>
</tr>
<tr>
<td>Peasant</td>
<td>33</td>
<td>4</td>
<td>7</td>
<td>.44</td>
<td>.57</td>
</tr>
<tr>
<td>Student</td>
<td>69</td>
<td>9</td>
<td>18</td>
<td>.70</td>
<td>.66</td>
</tr>
<tr>
<td>Politician</td>
<td>52</td>
<td>7</td>
<td>14</td>
<td>.26</td>
<td>.28</td>
</tr>
<tr>
<td>Professional</td>
<td>64</td>
<td>8</td>
<td>17</td>
<td>.55</td>
<td>.64</td>
</tr>
<tr>
<td>Worker</td>
<td>126</td>
<td>17</td>
<td>35</td>
<td>.72</td>
<td>.51</td>
</tr>
<tr>
<td>Small Producer</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>.50</td>
<td>.33</td>
</tr>
<tr>
<td>No</td>
<td>838</td>
<td>112</td>
<td>168</td>
<td>.86</td>
<td>.85</td>
</tr>
<tr>
<td>Overall average</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.60</td>
<td>.59</td>
</tr>
</tbody>
</table>
The data size for each class and the performance for the organizer semantic classification task is provided in Table 12. The data for this task enable the model to obtain relatively high scores for most of the classes. The average performance scores are around .70 for precision, recall, and F1-macro. The predictions for the classes militant organization and labor union are the most accurate. Conversely, the classes chamber of professionals and NGO are predicted more poorly.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Train</th>
<th>Development</th>
<th>Test</th>
<th>Performance</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Militant Organization</td>
<td>55</td>
<td>8</td>
<td>12</td>
<td>.90</td>
<td>.75</td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td>Political Party</td>
<td>234</td>
<td>31</td>
<td>52</td>
<td>.68</td>
<td>.71</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Chamber of Professionals</td>
<td>23</td>
<td>3</td>
<td>5</td>
<td>.25</td>
<td>.40</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>Labor Union</td>
<td>60</td>
<td>8</td>
<td>14</td>
<td>.78</td>
<td>.78</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>NGO</td>
<td>119</td>
<td>16</td>
<td>27</td>
<td>.68</td>
<td>.55</td>
<td>.61</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1,369</td>
<td>182</td>
<td>275</td>
<td>.93</td>
<td>.94</td>
<td>.93</td>
<td></td>
</tr>
<tr>
<td>Overall average</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.70</td>
<td>.69</td>
<td>.69</td>
<td></td>
</tr>
</tbody>
</table>

The former category does not show any pattern in the confusion matrix. But the latter category is mostly confused with the Political Party and Chamber of Professionals.

### 6.2 Insights from Benchmarks

Various parts of this corpus were used in shared tasks that were open to any research team to participate in 2019 and 2020. This subsection presents a summary of the insights pertaining to the quality of the corpus reported by the participants.

The Lab ProtestNews [16,17] was organized in the scope of Conference and Labs of the Evaluation Forum (CLEF 2019). The evaluation setting required participants to achieve unsupervised domain adaptation from India to China on English data. The Lab consisted of three subtasks that were document, sentence, and token classification. Although deep learning approaches outperformed traditional ML techniques, the gap between the system performances on India and China was still significant. The difference between the data distributions from India and China was measured using Jensen-Shannon (J-S) divergence and out-of-vocabulary rate of token (OOV) ratio by the team ProTestA [43]. The team ProTestA observed a positive significant correlation between the J-S similarity scores and the system performances across the three tasks. Moreover, the team DeepNEAT [44] reported nontrivial differences between the longest sentences in training and test data, which are 440 and 643, respectively.

The shared task Event Sentence Coreference Identification (ESCI) was organized in the scope of the workshop Automated Extraction of Socio-political Events from News [18]. The ESCI task required participants

---

*https://emw.ku.edu.tr/clef-protestnews-2019/


to group event sentences extracted from the same document in case they are about the same event. One of the task participants reported that the title of the news articles co-occurred with some meta-data such as publication time and place in the same sentence. This team remedied the issue using regular expressions to remove the meta-data in these sentences [45].

6.3 Applying the Tools on News

Finally, as supplementary information to Table 8, we report the effect of these configurations on a random sample from four news archives from India. We have aggregated results of the tools on randomly sampled 20,000 news articles from each source. Figure 2 provides the temporal event distribution from July 2002 to January 2018 for each configuration provided in Table 8\textsuperscript{a}. We observe a comparable pattern across the sub-figures. However, the intensity of the events as measured in terms of event count differs across these configurations. Moreover, the Tok and Doc+Sent+Tok, which yield the highest recall and precision, respectively, have the potential not to be comparable. An evaluation against real event count, which has not been performed in the scope of our study, should direct us towards the right path to follow.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{Temporal distribution of news from four sources from India. Note: Each figure is based on a configuration that is noted under it.}
\end{figure}

\textsuperscript{a} This is only to temporally compare results of each configuration. Any additional comment on these figures requires further investigation.
7. CONCLUSION AND FUTURE WORK

We introduced a gold standard corpus (GSC) that enables benchmarking and the creation of automated tools for contentious political event-related information collection across contexts. The methodology we developed to ensure the quality of the corpus, our observations during the application of this methodology, and the results obtained via the automated tools created using the corpus have been reported in detail. The clear performance drop when the test data differ from training data, known as the domain or covariate shift problem [46], shows how critical it is to incorporate cross-context aspects to the corpus creation, tool development, and evaluation. Handling each context separately at least in the evaluation phase is an indispensable part of the solution to measure and improve reliability of the performance scores. This is achieved by testing the models on data collected from the target contexts. Our recent ML models were mainly created using training data from only a single country. The following steps should incorporate data from multiple contexts and engage with domain adaptation techniques at the model creation phase [47, 48].

We have kept track of what is included and excluded at each level to better automatize the task and allow quantification of the recall, which has been a notable gap in this field. Restricting data sets by using keywords or basing a protest knowledge base on a subset of a source due to practical reasons was harming the validity and reliability of the resulting data sets. Starting with a random sample and proceeding with recall-optimized active learning during the creation of the gold standard corpus ensures that training data will improve the quality of the final gold standard data set. We obtained 97% recall and 60% precision when we used recall-optimized ML models to extend the corpus.

Our approach based on random sampling from multiple contexts has enabled us to create a test-bed for reliably measuring the performance of the state-of-the-art ML methodology on the target tasks and settings. We have invested in the line of research that shows improvement when data are prepared using random samples and when deep learning methodologies such as BERT are facilitated in comparison to using keyword filtered data sets and traditional ML in cross-context settings [49]. However, Yörük et al. [49] reported on the performance of the random samples only relevant to the document classification task. We will continue this investigation on sentence classification and event extraction tasks.

The immediate step that will follow this study is to evaluate models trained with English data from India, China, and South Africa. We have already extended the GSC with news sources in Portuguese and Spanish and the cross-lingual extension and validation of the models will follow shortly. Other extensions will be on semantic categories such as violent vs. non-violent, urban vs. rural, and economic demands vs. non-economic demands. Furthermore, we will handle documents and sentences that contain multiple events. The recent models assume there is a separate event for each event trigger identified by the token extractor [27]. However, our observations directed us to identify and link the triggers that denote the same event [50, 51]. We will be developing tools for linking the event triggers about the same event in our pipeline [52, 18]. We will work on news sources in Portuguese and Spanish to cover at least Brazil, Mexico, and Argentina. This work will facilitate cross-lingual ML models to be created. Consequently, the data from one language are expected to enhance the work in a new language [53]. Any error found will be corrected and the corpus will be extended with new data to avoid reducing the corpus to a temporally restricted and biased collection.
ACKNOWLEDGEMENTS

This study is funded by the European Research Council (ERC) Starting Grant 714868 awarded to Dr. Erdem Yörük for his project Emerging Welfare.

AUTHOR CONTRIBUTIONS

A. Hürriyetoğlu (ahurriyetoglu@ku.edu.tr) led the study, wrote and submitted the paper, developed text processing tools, and created machine learning models reported in this study. E. Yörük (eryoruk@ku.edu.tr) determined the need for this effort, and together with Ç. Yoltar (cyoltar@ku.edu.tr) and B. Gürel (bgurel@ku.edu.tr) operationalized the concept of the contentious politics to be applied for this study, identified sources used for collecting data, and ensured quality. O. Mutlu (omutlu@ku.edu.tr) wrote the results of the paper with A. Hürriyetoğlu, collected data, maintained an online annotation platform, ensured quality of the annotations, developed text processing tools, and created deep learning models reported in this study. F. Duruṣan (fdurusan@ku.edu.tr) wrote the annotation manual and the annotation setting related parts in the paper, proofread the paper, and supervised the annotation team. D. Yüret (dyuret@ku.edu.tr) advised the team on the annotation methodology and state-of-the-art machine learning model creation.

DATA AVAILABILITY STATEMENT

The data sets generated and/or analyzed during the current study are not publicly available and we explained how researchers should contact us to obtain the whole data at section 5.3. Supplemented data including the annotation manuals are available in the Science Data Bank repository, 10.11922/sciencedb.j00104.00092, under an Attribution 4.0 International (CC BY 4.0). The detailed protest data sets are available from the corresponding author on reasonable request.

REFERENCES

Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction


Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction


AUTHOR BIOGRAPHY

Ali Hürriyetoğlu is a postdoctoral research fellow at Koç University, Turkey, in the ERC projects “Emerging Welfare” and “Social ComQuant: Excelling in Computational and Quantitative Social Sciences in Turkey”. He is performing research and coordinating the work-package related to automatically collecting a multi-country and multilingual protest (contentious politics) events database from local news sources. In addition to his research-related duties, he is teaching Social Sciences and Computing course at the Social Sciences department. Dr. Hürriyetoğlu is a computational linguist who has a background in computer and cognitive sciences. He performed research on extracting actionable information from social media in the scope of his Ph.D. studies. He has been working in industrial, governmental, and academic settings to process news and social media text in various domains throughout his career. Recently, he is active in natural language processing and information retrieval communities in terms of suggesting novel data sets and tasks that are tackled by many teams all over the world. His recent research focus is on the robustness and the generalizability of text processing systems across contexts. ORCID: 0000-0003-3003-1783

Erdem Yörük is an Associate Professor in the Department of Sociology at Koç University of Turkey and an Associate Member in the Department of Social Policy and Intervention at the University of Oxford, UK. He is the principal investigator of the ERC-funded project “Emerging Welfare” (The New Politics of Welfare: Towards an “Emerging Markets” Welfare State Regime) (emw.ku.edu.tr). He is also a member of the Young Academy of Europe and an associate editor of European Review. He holds a Ph.D. from the Department of Sociology at Johns Hopkins University (2012), an MA in Sociology and a BSc in Electrical and Electronics Engineering from Boğaziçi University, Turkey. His work focuses on social welfare and social policy, social movements, political sociology, and comparative and historical sociology. His work has been supported by the National Science Foundation (NSF), Ford Foundation, European Commission Marie Curie CIG, European Research Council, European Commission Twinning Action Grants (socialcomquant.ku.edu.tr), and the Science Academy of Turkey BAGEP. His articles have appeared in World Development, Governance, Politics and Society, New Left Review, Current Sociology, South Atlantic Quarterly, International Journal of Communication, Social Policy and Administration, Social Indicators Research, among others. ORCID: 0000-0002-4882-0812
Osman Mutlu is a graduate student currently in the MS program in Computer Science and Engineering at the Graduate School of Science and Engineering, Koç University, Turkey. He has worked in Emerging Welfare Project at Koç University as a Project Engineer. He played a primary role in the creation of a multilingual and multi-level gold standard corpus, text processing tools, and deep learning models for event information collection. He is one of the main organizers of the shared task series that started with CLEF 2019 ProtestNews Lab and continued with the Event Sentence Coreference Identification task at AESPEN @ LREC 2020, and with the multilingual event collection challenge at CASE @ ACL-IJCNLP 2021.

ORCID: 0000-0001-6144-5685

Fırat Duruşan is a Ph.D. candidate in the Department of Political Science and Public Administration, Ankara University, Turkey. His ongoing dissertation research involves a comparative political-economic study of social policy in Argentina and Turkey in the post-2000 period with a particular focus on cash transfer programs. He worked as a research assistant in Ankara University and Hitit University between 2010 and 2017. He was a visiting researcher in Facultad Latinoamericana de Ciencias Sociales (FLACSO) in Buenos Aires, Argentina between 2014 and 2015. He holds a BA (2005) and MA (2008) degrees from the Department of Political Science and Public Administration in Middle East Technical University. He has worked in the Emerging Welfare Project at Koç University as a research assistant. He continues to work as a researcher in the Twinning Action project, Social ComQuant at Koç University.

Çağrı Yoltar is a post-doctoral fellow and the project leader of the “Emerging Welfare” project at Koç University, Turkey. She received her Ph.D. in Cultural Anthropology from Duke University. She also holds an MA degree in Sociology from Bogazici University, and an MA in Science, Technology and Society Studies and BSc in Management Science and Engineering from Istanbul Technical University. Her scholarly interests lie at the intersection of political and economic anthropology and social policy. Her current research, which was supported by Social Science Research Council, the Wenner-Gren Foundation for Anthropological Research, and various Duke University fellowships addresses the politics of care in the context of ethnic conflict by linking contestations over sovereignty to anthropological notions of kinship, reciprocity, debt, and value. Her articles have appeared in Governance, Political and Legal Anthropological Review (PoLAR), Cultural Anthropology, Middle Eastern Studies and Jadaliyya.
Deniz Yüret is a full-time professor in the Computer Engineering Department, Koç University, Istanbul, Turkey, and the founding director of the Koç University - İş Bank (KUIS) AI Center. Dr. Yüret received his BS, MS, and Ph.D. from MIT, where he was a member of MIT AI Lab (now CSAIL) from 1988 to 1998. Before joining Koç University in 2002, he co-founded the question answering start-up Inquira, Inc. which was later acquired by Oracle. His research is focused on artificial intelligence, machine learning, and natural language processing.

ORCID: 0000-0002-7039-0046


ORCID: 0000-0002-1666-8748