A New Task and Dataset on Detecting Attacks on Human Rights Defenders

Shihao Ran   Di Lu
Joel Tetreault   Aoife Cahill   Alejandro Jaimes
Dataminr Inc.
{sran,dlu,jtetreault,
acahill,ajaimes}@dataminr.com

Abstract

The ability to conduct retrospective analyses of attacks on human rights defenders over time and by location is important for humanitarian organizations to better understand historical or ongoing human rights violations and thus better manage the global impact of such events. We hypothesize that NLP can support such efforts by quickly processing large collections of news articles to detect and summarize the characteristics of attacks on human rights defenders. To that end, we propose a new dataset for detecting **Attack**s on **H**uman **R**ights **D**efenders (HRDsAttack) consisting of crowdsourced annotations on 500 online news articles. The annotations include fine-grained information about the type and location of the attacks, as well as information about the victim(s). We demonstrate the usefulness of the dataset by using it to train and evaluate baseline models on several sub-tasks to predict the annotated characteristics.

1 Introduction

It is essential for human rights organizations to track, analyze and summarize attacks on human rights defenders over time and across locations for better personnel protection and situational analysis. To do so, multiple event attributes denoting different aspects of the attacking event need to be extracted from textual sources. However, this would be a time-consuming process if done manually. Figure 1 gives an example of the kinds of information that such organizations need to extract.

In order to train and evaluate an NLP model to extract this information automatically, a relevant dataset is necessary. The ideal dataset requires accurate annotations for both the breadth (the number of extracted event attributes) and depth (the levels of granularity for each event attribute) of the events. However, all existing Event Extraction (EE) datasets (e.g. ACE05 (Doddington et al., 2004), ERE (Song et al., 2015), ACE05-E (Wadden et al.,
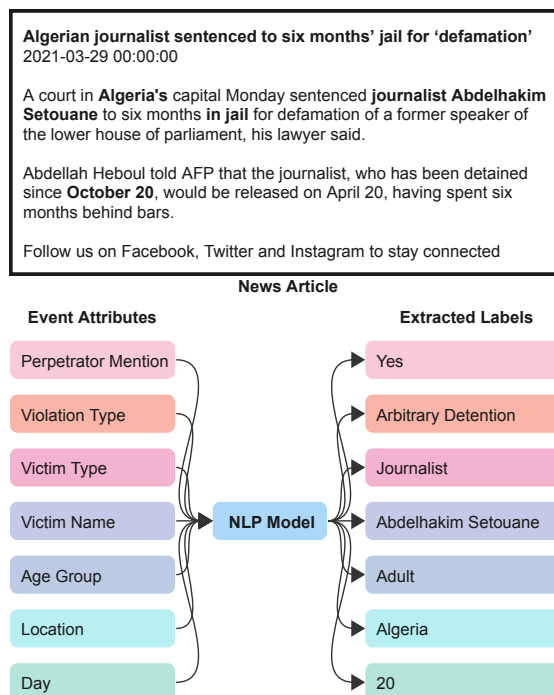

Figure 1: An example of the input/output to an NLP model for extracting event attributes about an attacking event on human rights defenders.

2019), ACE05-E+ (Lin et al., 2020)) do not contain annotations at a sufficiently fine-grained level. Although some existing ontologies and datasets do include annotations related to attacking events, e.g. the ATTACK event type in the ACE05 dataset along with the associated AGENT attribute, they are incomplete with respect to many of the details of interest to human rights organizations and do not contain annotations relevant to victim characteristics or the time/location of the attacking event. As a result, existing open-source EE models trained on these datasets (Honnibal et al., 2020; Wadden et al., 2019; He et al., 2019) are unable to predict the complete set of relevant information.

To mitigate the gap in existing resources, we present HRDsAttack, a new dataset containing

crowdsourced annotations on 500 online news articles (including article title, article body text, and publication time). Each news article is annotated with 13 different event attributes to capture critical information about attacks on human rights defenders, including the type and location of the attacks, as well as information about the victim(s) and the perpetrator. With HRDsAttack, we hope to support more research opportunities for including NLP in applications related to human rights, as well as for broader AI for Social Good (AI4SG) efforts.

To summarize, our contributions are threefold:

1. We present a new dataset (HRDsAttack) that includes annotations for fine-grained event details on attacks on human rights defenders. By focusing on expanding the breadth and depth of the attacking event relative to existing EE ontologies, we aim to address the limited scope of existing NLP resources. The complete ontology for our dataset is shown in Table 1;

2. We propose a new NLP task to extract fine-grained event details on attacks on human rights defenders.

3. We demonstrate the usefulness of HRDsAttack with a strong baseline model based on Question Answering (QA) using the T5 model (Raffel et al., 2020) as the backbone in a multi-task setting.

The HRDsAttack dataset along with the code for model training and evaluation is available at https://github.com/dataminr-ai/HRDsAttack.

2 Related Work

2.1 Event Extraction

Event Extraction (EE) is an NLP task that aims to extract key information such as *who, what, where, and when* from a text. The most commonly used dataset for EE is the ACE05 English corpus (Doddington et al., 2004) which consists of 33 event types and 22 event argument roles across 599 documents from newswires, web blogs, and broadcast conversations. While the ACE ontology covers a large range of event types, only two of them are related to attacking events: the LIFE.INJURE event and the CONFLICT.ATTACK event. Some of the other datasets that focus on extracting event triggers or event arguments are based on the ACE05

ontology (Wadden et al., 2019; Lin et al., 2020), and only cover limited aspects of the information that HRDsAttack covers, e.g. the ATTACKER and TARGET attributes in the LIFE.INJURE and CONFLICT.ATTACK events. The Armed Conflict Location and Event Data (ACLED) dataset (Raleigh et al., 2010) covers political violence and protest events with annotations for event type, actors and targets, but it does not cover victim-dependent attributes. In comparison, HRDsAttack focuses on attacking events on human rights defenders and provides more event attributes for the attacks, along with more granular information regarding each event attribute.

In terms of modeling approaches, early work on EE formulated the task as a token-based classification problem which leveraged different types of features (Ahn, 2006; Liao and Grishman, 2010a,b; Li et al., 2013). More recent approaches focus on applying neural models to EE tasks, such as CNNs (Chen et al., 2015), RNNs (Liu et al., 2019), and other advanced model structures (Nguyen and Nguyen, 2019; Zhang et al., 2019).

2.2 NLP Research for Human Rights

Existing NLP research resources around event detection and extraction related to Human Rights are extremely limited. Previous work has focused on identifying potential human rights abuse incidents from social media posts (Alhelbawy et al., 2020; Pilankar et al., 2022), alongside more general applications such as detecting abusive language (Golbeck et al., 2017; Djuric et al., 2015; Aroyo et al., 2019), or procedure-focused applications (e.g. data modeling processes for human rights data (Miller et al., 2013; Fariss et al., 2015)), or predicting judicial decisions of the European Court of Human Rights using NLP (O'Sullivan and Beel, 2019). To our knowledge, there are no event extraction datasets which target human rights issues, which makes HRDsAttack a first in this research area.

3 Dataset

In this section, we describe the construction of the HRDsAttack dataset, which contains 500 annotated news articles, including article title, article body text, and publication time. We select news articles as the data source rather than other data sources (such as social media posts) since online news articles generally have higher accessibility, better trustworthiness of the source, and longer content

Category	Event Attribute	Labels	Label Definitions
Perpetrator	Perpetrator Mention	Yes	There is one or more explicit mention of the perpetrator in the news article.
		No	There is no explicit mention of the perpetrator in the news article.
	Perpetrator Type	State Security Forces	Anyone employed by or representing a state institution.
		Other State Actors	Other actors that are a part of the state or other non-military authorities of a state.
		Other non-state actors	Other actors /Private actors that are not a part of the state and act without the state's permission, support, or acquiescence.
		Other actors with permissions	Armed actors that are not a part of the state but act with the state's permission, support or acquiescence.
		Other actors without permissions	Other actors that are not a part of the state.
		Regional Organizations	Person or group working for a regional or international organization.
		Insufficient Information	There is insufficient information available to determine one of the categories described above.
		None	Not applicable, when Perpetrator mention is No.
Violation	Violation Type	Arbitrary Detention	Arrest or detention not in accordance with national laws.
		Enforced Disappearance	Unlawful deprivation of liberty enforced or authorized by the state, that is not acknowledged by the state or the location of the victim is kept secret.
		Killing	Unlawful death inflicted upon a person with the intent to cause death or serious injury.
		Kidnapping	Deprivation of liberty that is not enforced or authorized by the state.
		Torture	The action or practice of inflicting severe pain or suffering on someone as a punishment or in order to force them to do or say something.
		Other	Sexual violence or other acts causing or intending to cause harm, such as coercion or discrimination.
		Unknown	No harmful acts were conducted or there is insufficient information to determine the harmful acts.
Victim	Victim Name	-	Name of the victim.
	Victim Type	Human Rights Defender	A person exercising their right, to promote and strive for the protection and realization of human rights and fundamental freedoms.
		Trade Unionist	A person exercising their right to form and join trade unions to protect their interests.
		Journalist	A person observing events, statements, policies, etc. that can affect society, with the purpose of systematizing such information to inform society.
		Insufficient Information	There is insufficient information available to make select one of the categories described above.
	Victim Population Type	Individual	A named individual victim.
		Multiple	Multiple unnamed individuals.
	Victim Age Group	Adult	Age >= 18.
		Child	Age <17.
		Other	A mixture of age groups, when Victim Population Type is Multiple.
		Unknown	There is insufficient information available to determine the age group.
	Victim Sex Group	Man	Male.
		Woman	Female.
		Other	Other gender types.
		Unknown	There is insufficient information available to determine the sex group.
Location	Country	-	Country in which the attack occurred
	Region	-	Region in which the attack occurred, such as a state or a province
	City	-	City in which the attack occurred
Time	Year	-	Year the attacking event occurred
	Month	January, ..., December	Month the attacking event occurred
	Day	1, 2, 3, ..., 31	Day (of the month) the attacking event occurred

Table 1: Labeling ontology of HRDsAttack.

length.

In our work, we sample online news articles from the GDELT database[1], which we discuss in more detail in Section 3.2.

3.1 Annotation Labels

To ensure the comprehensiveness of the annotations regarding capturing event details, we first identify the event attributes or labels required for annotation. As shown in Table 1, according to the UN Human Rights SDG 16.10.1 Guidance Note[2], we identify the following 5 categories of attributes: PERPE-TRATOR, VIOLATION, VICTIM, LOCATION, and TIME. Each category has one or more associated event attributes, all denoting key information about the primary event described in the original article[3]. If there are multiple events mentioned in the article, only the primary event (i.e. the event that happened closest to the publication time) is annotated. We also specify that the VICTIM category could have multiple entries per article, while other categories can only have one entry per article (i.e. only one entry for the primary attack event). The ontology

[3] All label values for each event attribute are prescribed by the SDG 16.10.1 Guidance Note.

for the annotation labels is shown in Table 1.

3.2 Data Sampling

To build HRDsAttack, we first scrape 80,112 online news articles in the time range of 2019/09/01 to 2022/05/01 from the GDELT database following the CAMEO codebook (Schrodt, 2012), a standard framework for coding event data (Yuan, 2016). These scraped news articles are identified as relevant to human rights defenders by an existing human rights monitoring workflow.

During our pilot studies, we identified a data imbalance issue from the annotations under random sampling. Specifically, we observed significantly skewed label distributions in event attributes VIOLATION TYPE and VICTIM TYPE, the minority classes being TORTURE and KIDNAPPING for VIOLATION TYPE, and HUMAN RIGHTS DEFENDERS and TRADE UNIONISTS for VICTIM TYPE. To address this issue, we apply keyword filtering and targeted sampling to ensure HRDsAttack is well-balanced across classes in each event attribute.

To include more samples with a higher probability of containing events associated with these minority attributes, we first reduce the original 80,112 samples into four smaller, targeted sample sets. Each targeted sample set corresponds to the articles that contain the keyword for each of the minority classes. We then randomly sample 25 articles from each targeted sample set to form a batch of 100 samples for each round of full annotation. Table 2 shows the keywords used for minority class targeted sampling.

Minority Event Attribute	Keyword
Torture	torture
Kidnapping	kidnapping
Human Rights Defenders	human right
Trade Unionists	trade union

Table 2: Keyword for each minority class used in keyword filtering and targeted sampling.

3.3 Annotation Process

The annotation is done by qualified workers (Turkers) on Amazon Mechanical Turk (AMT). We design and implement a separate qualification task to recruit top-performing Turkers, and we only release the full annotation tasks to the Turkers that surpass a predefined performance bar based on the qualification tasks.

3.3.1 Qualification Tasks

For the qualification task, all US-based Turkers that have a HIT (Human Intelligence Task [4]) approval rate greater than 90% and a total number of HITs approved greater than 500 are able to participate. In the qualification task, we sample three different news articles and ask all participant Turkers to annotate every event attribute for each news article through three questionnaires (each HIT contains three questionnaires, one for each news article). We then evaluate their performance on this annotation task. All three news articles are also annotated by domain experts, and we use their annotations as the ground truth answers for calculating the Turker accuracy. We only recruit Turkers who have 75% or higher average accuracy across all three news articles. We launched three rounds of qualification tasks with 50 assignments in total, and ten Turkers passed the qualification tasks.

The instructions and the task interface for the qualification tasks are shown in Figures 4 to 11 in Appendix A.

3.3.2 Full tasks

In the full task, each HIT only contains a single news article. The instructions and the annotation interface are identical to the qualification task. We launched all 500 samples in 5 batches, each batch containing 100 HITs. During our pilot studies, we did not observe a significant quality improvement with replication factor 3 due to relatively high agreement scores between the Turkers (Table 8 in Appendix C). We hypothesize that this is because the annotation task itself is highly objective. Therefore, we did not apply replication factors during the full task.

We compensate each Turker with $7.50 per assignment in the qualification task (three news articles per assignment) and $2.00 per assignment in the full task (one news article per assignment). We also provide an additional bonus to all participant Turkers of $0.5 per assignment. The final pay rate is $15.00 per hour, which is over the US national minimal wage of $7.50[5].

The annotation instructions and the task interface for the full tasks are shown in Figures 12 to 15 in Appendix A.

---

[4]A HIT represents a single, self-contained, virtual task that a Turker can work on, submit an answer, and collect a reward for completing.

[5]https://www.dol.gov/general/topic/wages/minimumwage

3.4 Data Statistics

To create a benchmark dataset from HRDsAttack, we randomly split the 500 annotated samples into train, dev, and test set with a 3:1:1 ratio. Table 3 shows the statistics of the splits. A breakdown of the label-level statistics for each event attribute can be found in Table 7 in Appendix B.

	Train	Dev	Test	Total
No. of Articles	300	100	100	500
Total No. of Tokens	287,911	97,038	124,658	509,607
Avg. No. of Tokens	959.70	970.38	1,246.58	1,019.21
Total No. of Victims	687	272	204	1,163
Avg. No. of Victims	2.29	2.72	2.04	2.33

Table 3: Textual statistics of HRDsAttack splits. The average number of tokens and victims is averaged per news article.

4 Our Model

With the construction of HRDsAttack, we now turn to developing a model for the task. We noted earlier that existing state-of-the-art EE models are not suitable as baselines, as they rely on extensive human annotations based on token-level annotations, hence cannot easily be re-trained and evaluated on this dataset. For instance, AMR-IE (Zhang and Ji, 2021) and GraphIE (Qian et al., 2018) are trained on the ACE05 dataset and ERE dataset. Some recent research casts the EE task as QA tasks or Seq2seq tasks, such as RCEE_ER (Liu et al., 2020) and Text2Event (Lu et al., 2021). In this section, we propose a new model for extracting fine-grained details regarding attacks on human rights defenders.

4.1 Overall Framework

Given the limited amount of training data and the range and variety of event attributes, we propose using a single Seq2Seq Question Answering (QA) model. Training a unified model has the advantageous property that it shares the training data across all the sub-tasks thus potentially leading to better performance for each sub-task. Figure 2 shows the overall framework of our proposed baseline model.

We formulate all of the subtasks as a generation task following T5 (Raffel et al., 2020), which proposes reframing all NLP tasks into a unified text-to-text format. The input to the T5 model is a natural language sentence composed of (1) a task prefix (e.g. *'extract victims'*), (2) an attribute-oriented question (e.g. *'Who is the victim of the*

violation?'), and (3) a context which is the original article. The output is a text string which explicitly refers to the value of the concerned event attribute (e.g. *'Abdelhakim Setouane'*).

4.2 Input-Output Design

We group the event attributes into three categories: general article-dependent attributes, victim-dependent attributes, and publication time-dependent attributes, and we design input and output formats for them respectively. For all of the three categories, the output is a text string that explicitly refers to the value of the relevant event attribute, e.g. *'Yes'* for PERPETRATOR MENTION, or *'state security forces'* for PERPETRATOR TYPE. The input formats for the three categories have minor differences [6]:

- **General Article-dependent Attributes:** Most of the event attributes depend on the general information contained within the article (i.e. do not rely on additional input other than article's body text). These include PERPETRATOR MENTION, PERPETRATOR TYPE, and VIOLATION TYPE. For these attributes, the input is the concatenation of a task prefix, an attribute-oriented question, and the original article (e.g. the top three examples in Figure 2).

- **Victim-dependent Attributes:** Some event attributes, such as VICTIM SEX TYPE, depend on the information related to a specific victim. Thus we incorporate the victim name into the input question, as exemplified in the fourth and fifth examples in Figure 2.

- **Publication Time-dependent Attributes:** In some cases, the YEAR, MONTH, and DAY attributes related to the attack event are not explicitly present in the article, and we need to infer them based on a combination of the article publication time and the relevant time mentioned in the article (e.g. *last month, two weeks ago, yesterday*). The article publication time is available as metadata in the GDELT dataset (e.g. *2021-03-29 00:00:00*). For these attributes, we add publication time information into the input, as shown in the last example of Figure 2.

---

[6]The complete lists of input and output formats are provided in Table 9 in Appendix D.
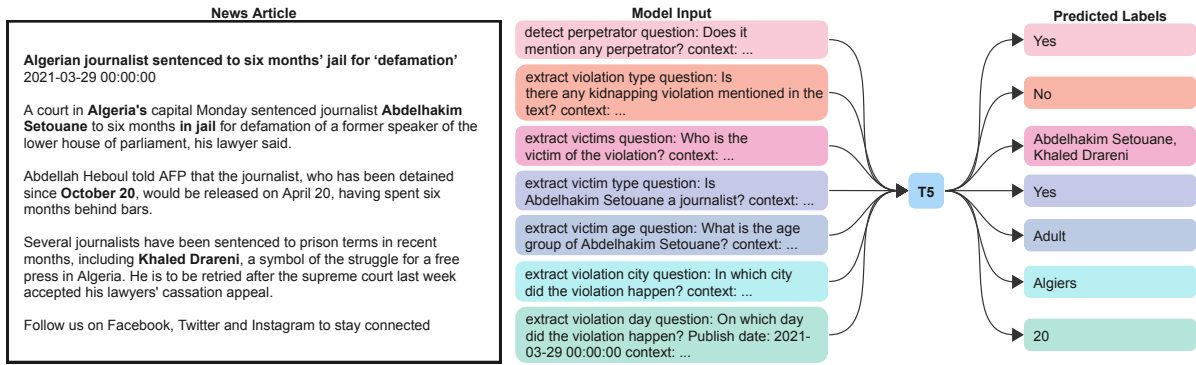
Figure 2: Overall framework of the proposed Sequence-to-Sequence Question-Answering model.

Task Prefix. Following the multi-task setting in the original T5 work, we add a task prefix at the beginning of the input text. The task prefix is used to instruct the T5 model to perform a particular task. It could be any arbitrary text. In our work, we use a brief task description as the task prefix for each event attribute, e.g. *'detect perpetrator'* for PERPETRATOR MENTION or *'extract violation type'* for VIOLATION TYPE (Figure 2). The complete list of all the task prefixes is shown in Table 10 in Appendix D.

4.3 Long Document Resolution

The maximum input length allowed by the T5 model is 512 tokens, but around 75% of the articles from the GDELT dataset exceed that length limit. We explore two options to deal with articles with more than 512 tokens: **Truncation** and **Knowledge Fusion**. Additional methods for handling long documents are discussed in Appendix E.

Truncation. We only use the first 512 tokens of the input text. The articles from GDELT are news articles, and the first several sentences from a news article usually contain the most important information. Thus a simple solution is to truncate the article and ignore the cut content.

Knowledge Fusion. To mitigate the information loss in the Truncation method, we adopt a split-fuse approach (Figure 3) by (1) splitting the documents into short paragraphs using the spaCy (Honnibal et al., 2020) tokenizer[7]; (2) applying the model to each of the paragraphs; and then (3) merging the results from each paragraph to obtain the final results for the

Figure 3: Knowledge Fusion approach.

original article. For event attributes that allow more than one value (e.g. VICTIM NAMES), we keep all of the unique results, and for other attributes, we only keep the one with the highest confidence score (beam search score).

5 Experiments

5.1 Evaluation Metrics

We consider the following metrics for evaluating different event attributes:

- **Precision, Recall, and F1 Score**: we use Precision, Recall, and F1 score to evaluate the

---

[7]We use the en_core_web_sm spaCy pipeline.

model performance on PERPETRATOR MENTION and VIOLATION TYPE.

- **Accuracy**: we use accuracy (i.e. percentage correct) to evaluate the model performance on PERPETRATOR TYPE, VICTIM TYPE, VICTIM SEX TYPE, VICTIM AGE GROUP, COUNTRY, REGION, CITY, YEAR, MONTH, and DATE.

- **Fuzzy Match Precision, Recall, and F1 Score**: For the VICTIM NAME attribute, we use precision, recall, and F1 score based on exact matching and fuzzy matching, respectively. For exact matching, one predicted victim name is counted as correct only if it exactly matches with a victim name in the ground truth. For fuzzy matching, one predicted victim name is counted as correct if it has overlapping tokens with a victim name in the ground truth. For example, a predicted victim name *Jordan* is counted as correct when it matches with a ground truth name *Michael Jordan*.

5.2 Baseline Models

We consider the following models in our evaluation:

- **DyGIE++** ([Wadden et al., 2019](#)): a joint Information Extraction (IE) model and we use the checkpoint trained on the ACE05 dataset. It requires mapping from the ACE event ontology[8] to HRDsAttack. As a result, it only covers two attributes: PERPETRATOR MENTION and VICTIM NAME as there is no available mapping for the other event attributes in HRDsAttack.

- **T5 w/ Truncation**: our proposed T5-based model with truncation.

- **T5 w/ Knowledge Fusion**: our proposed T5-based model with knowledge fusion.

- **Hybrid (final model)**: a hybrid model based on T5 w/ Truncation and T5 w/ Knowledge Fusion. The model only applies knowledge fusion to PERPETRATOR MENTION, VICTIM NAME, and VICTIM AGE GROUP attributes. This hybrid strategy is decided based on the evaluation results on the dev set.

[8]The ACE ontology covers event types such as ATTACK and INJURE.

We recognize that it would be ideal to have more baseline models for comparison, such as a retrained version of DyGIE++ on HRDsAttack. However, many existing EE models are trained on token-level annotations and are not designed for the additional event attributes that HRDsAttack covers (e.g. VICTIM TYPES). Therefore, we had to design a specialized model for this task. We plan to benchmark more Sequence-to-Sequence based models on HRDsAttack in future work.

5.3 Training Implementation

We use the T5-large checkpoint [9] provided by Huggingface ([Romero, 2021](#)) to initialize the model and all experiments are run on a single AWS g5.xlarge instance. The AWS g5.xlarge instance is equipped with a single NVIDIA A10G GPU with 24 GB of GPU memory. Table 4 shows the hyperparameters we use to train the model.

Hyperparameter	Value
Learning rate	1e-4
Learning rate decay	1e-5
Epoch	20
Batch size	4
Gradient accumulation steps	16

Table 4: Hyperparameter settings for model training.

5.4 Overall Performance

Table 5 shows the performance of the four models on the test set: the DyGIE++ baseline, T5 w/ Truncation, T5 w/ Knowledge Fusion, and the Hybrid model. Both T5-based models significantly outperform the DyGIE++ baseline, except for the precision of PERPETRATOR MENTION. In addition, we get further improvement from the Knowledge Fusion method for the PERPETRATOR MENTION, VICTIM NAME, and YEAR attributes. For other attributes, we get results that are slightly worse than those without Knowledge Fusion. This aligns with our assumption that violation events may be elaborated in the later parts of the news articles with specific victim names and violation types. So by applying the Knowledge Fusion method, we can significantly improve the recall of some event attributes. But for other information such as violation time and location, they usually appear in the first several sentences of the news article. The

[9]https://huggingface.co/t5-large

Event Attribute	Metric	DyGIE++	T5 w/ Truncation	T5 w/ Knowledge Fusion	Hybrid
Perpetrator Mention	Precision	**100.00**	93.68	93.81	93.81
	Recall	36.54	97.80	**100.00**	100.00
	F1	53.52	95.70	**96.81**	96.81
Perpetrator Type	Accuracy	-	**62.00**	60.00	62.00
Victim Name	Exact Match Precision	9.41	**75.61**	59.30	59.30
	Exact Match Recall	9.19	24.03	**39.53**	39.53
	Exact Match F1	9.30	36.47	**47.44**	47.44
	Fuzzy Match Precision	17.65	**85.37**	63.95	63.95
	Fuzzy Match Recall	17.24	27.13	**42.64**	42.64
	Fuzzy Match F1	17.44	41.18	**51.16**	51.16
Victim Type	Accuracy	-	**72.41**	71.67	72.41
Victim Sex Type	Accuracy	-	**89.66**	86.67	89.66
Victim Age Group	Accuracy	-	**93.10**	92.50	92.50
Violation Type	Precision	-	**67.91**	61.24	67.91
	Recall	-	75.26	**81.44**	75.26
	F1	-	**71.39**	69.91	71.39
Country	Accuracy	-	**66.00**	65.00	66.00
Region	Accuracy	-	**3.00**	2.00	3.00
City	Accuracy	-	**23.00**	12.00	23.00
Year	Accuracy	-	46.00	**50.00**	46.00
Month	Accuracy	-	**33.00**	29.00	33.00
Day	Accuracy	-	**14.00**	8.00	14.00

Table 5: Overall performance of the baseline models on HRDsAttack test set (%). All experiments are based on a single run with a preset random seed.

time and location information appearing in the later parts may not be related to the primary attacking event. So based on the evaluation results on the dev set (Table 11 in Appendix F), we propose a hybrid model as our final baseline model. The hybrid model only applies Knowledge Fusion to PERPETRATOR MENTION, VICTIM NAME, and VICTIM AGE GROUP attributes. We notice that the hybrid model designed based on the dev set does not achieve the best performance for VICTIM AGE GROUP and YEAR attributes on the test set. It might be the fact that the hybrid strategy is overfitted on the dev set. And we leave the optimization of the hybrid model as future work.

While the hybrid model outperforms the DyGIE++ baseline in almost all of the event attributes and unlocks the extraction of new attributes, we do see a relatively lower model performance in attributes such as REGION and DAY. We hypothesize that the ambiguity in REGION labels and the large number of classes in DAY labels introduce additional challenges to the model, especially with a limited amount of training data. For instance, some annotators mistakenly put *London* under REGION instead of CITY. We acknowledge that the annotation instructions could be further improved to address this issue.

Event Attribute	Metric	Hybrid
Victim Type	F1	22.89
Victim Sex Type	F1	33.33
Victim Age Group	F1	46.01

Table 6: End-to-end performance of the Hybrid model on HRDsAttack (%) for victim-dependent attributes with model predicted victim names. All experiments are based on a single run with a preset random seed.

We also evaluate the end-to-end performance on the victim-dependent attributes with the model-predicted victim names (Table 6). And we use F1 scores as the evaluation metric. One victim-dependent attribute is counted as correct only when both the predicted victim name and the predicted attribute value match with the ground truth.

6 Conclusion

In this paper, we present a new dataset that supports extracting detailed information about attacks on human rights defenders under a new task setting. Compared with existing event extraction resources, we focus on the human rights domain and expand to more event attributes for capturing event details more comprehensively. Our new dataset (HRDsAt-

tack) contains 500 human-annotated news articles with 13 different event attributes regarding the victim(s), the type of perpetrator and violation(s), as well as the time and location of the attacks. We demonstrate the usefulness of the dataset by developing a Sequence-to-Sequence-based Question Answering model tailored for this task. While it achieves decent performance on some event attributes, there are many where there is much room for improvement. We view this model as a strong baseline for future work. We believe models trained with HRDsAttack could be generalized to detect attacking events in other domains or targeting a different population. And we hope that this work encourages additional research on the development of new AI4SG NLP resources in the future.

Acknowledgements

Limitations

While HRDsAttack is, to the best of our knowledge, the first dataset on extracting attacks on human rights defenders, there are some limitations. For one, while being the first corpus of its kind, our dataset is English-only. Second, the number of documents is limited. While the sample size of HRDsAttack (500) is on par with some of the other EE datasets, such as ACE05 (599), we do see more samples being beneficial to subsequent model training and supporting other future studies. In addition, despite the effort to balance the class labels in the event attributes, some of the labels still remain imbalanced, such as PERPETRATOR TYPE.

Ethics Statement

The construction of HRDsAttack involves human annotations on AMT. The Turkers are provided with clear annotation instructions and are informed of the conditions where they would be qualified or disqualified. We compensate the Turkers with a final paid rate of $15.00 per hour which is over the US national minimal wage of $7.50.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*.

Ayman Alhelbawy, Mark Lattimer, Udo Kruschwitz, Chris Fox, and Massimo Poesio. 2020. An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, 153:113365.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion proceedings of the 2019 world wide web conference*, pages 1100–1105.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Christopher J Fariss, Fridolin J Linder, Zachary M Jones, Charles D Crabtree, Megan A Biek, Ana-Sophia M Ross, Taranamol Kaur, and Michael Tsai. 2015. Human rights texts: converting human rights primary source documents into data. *PloS one*, 10(9):e0138935.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing bert-based transformer architectures for long document summarization. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: Main volume*, pages 1792–1810.

Xinyu He, Lishuang Li, Xingchen Song, Degen Huang, and Fuji Ren. 2019. Multi-level attention based blstm

neural network for biomedical event extraction. *IEICE TRANSACTIONS on Information and Systems*, 102(9):1842–1850.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. *none*.

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

Shasha Liao and Ralph Grishman. 2010a. Filtered ranking for bootstrapping in event extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*.

Shasha Liao and Ralph Grishman. 2010b. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009.

Jian Liu, Yubo Chen, and Kang Liu. 2019. Exploiting the ground-truth: An adversarial imitation based knowledge distillation approach for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*.

Ben Miller, Ayush Shrestha, Jason Derby, Jennifer Olive, Karthikeyan Umapathy, Fuxin Li, and Yanjun Zhao. 2013. Digging into human rights violations: Data modelling and collective memory. In *2013 IEEE international conference on big data*, pages 37–45. IEEE.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Conor O'Sullivan and Joeran Beel. 2019. Predicting the outcome of judicial decisions made by the european court of human rights. *arXiv preprint arXiv:1912.10819*.

Yash Pilankar, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes, and Pramod Pathak. 2022. Detecting violation of human rights via social media. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 40–45.

Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay. 2018. Graphie: A graph-based framework for information extraction. *arXiv preprint arXiv:1810.13083*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Manuel Romero. 2021. T5 (base) fine-tuned on squad for qg via ap. https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap.

P Schrodt. 2012. Conflict and mediation event observations event and actor codebook v. 1.1 b3.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Yihong Yuan. 2016. Modeling inter-country connection from geotagged news reports: A time-series analysis. *arXiv preprint arXiv:1604.03647*.

Junchi Zhang, Yanxia Qin, Yue Zhang, Mengchi Liu, and Donghong Ji. 2019. Extracting entities and events as a single task using a transition-based neural model. In *IJCAI*.

Zixuan Zhang and Heng Ji. 2021. Abstract meaning representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49.

A Annotation Interface

We list the screenshots of our annotation interface from Figure 4 through Figure 15, including the annotation guideline (instructions) page for both the qualification tasks and the full tasks, the example page, as well as the task pages.

Annotation Instructions

This is a qualification task for another larger task. In this task, you will read **3** news articles. Each news article is about one or more attack events targeting human rights defenders or a similar population. You will then fill out a questionnaire for each news article using the information from the article.

Some important definitions:

- **Perpetrator**: A person who carries out a harmful, illegal, or immoral act. If one or more explicit mentions of perpetrators exist in the news article, you need to select **one** option from the following five types of Perpetrators for the **primary** perpetrator:
    1. **State Actor**: Actor that is part of the government, or public servants working for the government.
    2. **Other actors acting with the State's permission, support or acquiescence**: Actors that are not a part of the state but act with the state's permission, support or acquiescence.
    3. **Private Actors/ Actors not acting with the State's permission, support or acquiescence**: Other actors / Private actors that are not a part of the state and act without the state's permission, support or acquiescence.
    4. **Other actors (State's permission, support or acquiescence status unknown)**: Actors that are not a part of the state, there is not sufficient information about whether they act with or without the state's permission, support or acquiescence.
    5. **Insufficient information**: There is not sufficient information available to select one of the categories described above.
- **Violation**: The harmful, illegal, or immoral act carried by the perpetrator. You need to select **all** options that apply from the following six types of Violations:
    1. **Killing**: Unlawful death inflicted upon a person with the intent to cause death or serious injury.
    2. **Enforced disappearance**: Unlawful deprivation of liberty enforced or authorized by the state, that is not acknowledged by the state or the location of the victim is kept secret.
    3. **Torture**: The action or practice of inflicting severe pain or suffering on someone as a punishment or in order to force them to do or say something.
    4. **Arbitrary detention**: Arrest or detention not in accordance with national laws.
    5. **Kidnapping**: Deprivation of liberty that is not enforced or authorized by the state.
    6. **Other harmful acts**: Sexual violence or other acts causing or intending to cause harm, such as coercion or discrimination.

Figure 4: Screenshot of the Qualification Task Instructions (1/3).

- **Victim**: A person or a group of people harmed, injured, or killed as a result of a crime, accident, or other event or action.

  You then need to select **all** options that apply from the following four types of Victims:

  1. **Journalist**: A person observing events, statements, policies, etc. that can affect society, with the purpose of systematizing such information to inform society, including support staff, as well as community media workers and "citizen journalists" when they momentarily play that role.
  2. **Trade Unionist**: A person exercising their right to form and to join trade unions for the protection of their interests. A trade union is an association of workers organized to protect and promote their common interests.
  3. **Human Rights Defender**: A person exercising their right, to promote and to strive for the protection and realization of human rights and fundamental freedoms, including some journalists and trade unionists.
  4. **Insufficient information**: There is not sufficient information available to make select one of the categories described above.

You can check all these definitions later on in the questionnaire interface by hovering your mouse cursor on the info icon ❶ next to each item.

About the questionnaire:

- **You must answer every question in the questionnaire.** If you cannot find the answer to a specific question, either select the answer **Insufficient Information, Unknown, Other** for selection questions, or type in **N/A** in the text box. You will not be able to submit the HIT unless you answered all the questions.

- For **Year, Month, Day, Location** related questions, if the event happened across multiple days and/or happened at multiple locations, please fill in with the starting time and/or the starting location.

- When answering about victims, you need to first identify if the victim mentioned in the article refers to **named individuals (someone with a name)** or **a group of unnamed individuals (without explicit names, such as "a group of students")**. In the latter case, if the group has mixed genders and age groups, select "Other" for Victim Sex Type and Victim Age Group questions.

- For each victim that has a name mentioned in the article, please select **"Named Individual"** in the questionnaire then fill in their name individually. If multiple names are mentioned, click the **"Add Another Victim"** button to add all victims with their corresponding names. Only select **"Multiple (i.e. Group of unnamed individuals)"** when multiple victims are mentioned in the article without their specific names.

Figure 5: Screenshot of the Qualification Task Instructions (2/3).

4. **Insufficient information**: There is not sufficient information available to make select one of the categories described above.

You can check all these definitions later on in the questionnaire interface by hovering your mouse cursor on the info icon ⓘ next to each item.

About the questionnaire:

- **You must answer every question in the questionnaire.** If you cannot find the answer to a specific question, either select the answer **Insufficient Information, Unknown, Other** for selection questions, or type in **N/A** in the text box. You will not be able to submit the HIT unless you answered all the questions.

- For **Year, Month, Day, Location** related questions, if the event happened across multiple days and/or happened at multiple locations, please fill in with the starting time and/or the starting location.

- When answering about victims, you need to first identify if the victim mentioned in the article refers to **named individuals (someone with a name)** or **a group of unnamed individuals (without explicit names, such as "a group of students")**. In the latter case, if the group has mixed genders and age groups, select "Other" for Victim Sex Type and Victim Age Group questions.

- For each victim that has a name mentioned in the article, please select **"Named Individual"** in the questionnaire then fill in their name individually. If multiple names are mentioned, click the **"Add Another Victim"** button to add all victims with their corresponding names. Only select **"Multiple (i.e. Group of unnamed individuals)"** when multiple victims are mentioned in the article without their specific names.

Read an example:

On the next page, you will see one example article and one questionnaire with correct answers already filled in. Please read through the example carefully.

You will be asked to copy a randomly generated code into one text box in the middle of the example questionnaire, you only need to do this once on the example page.

NOTE: Your submission will be REJECTED if the code you pasted in the box does not match the one shown to you.

Next

Figure 6: Screenshot of the Qualification Task Instructions (3/3).

Note: This is an example.
Read the news article below then answer the questions on the right hand side of your screen.

The Execution of Mujtaba al-Sweikat

02-May-2019

A Saudi student who gained admission to Western Michigan University was beheaded last week for charges associated with his participation in a pro-democracy protest.

He was 17 years old, at the airport to catch a flight to the U.S., where he planned to visit colleges and hoped to attend Western Michigan University, where he subsequently gained admission. But he was arrested for crimes related to participation in a pro-democracy protest before he could board the plane.

Last week Saudi Arabia announced Mujtaba al-Sweikat was one of 37 people executed for terrorism-related crimes. The human rights group Reprieve says al-Sweikat, who was arrested in December 2012, was convicted on the basis of a confession obtained by torture.

CNN reported on court documents it obtained regarding al-Sweikat's prosecution. According to the documents obtained by the broadcaster, al-Sweikat's father, who served as his lawyer, portrayed his son as a diligent student, loyal to the kingdom of Saudi Arabia, who completed his final high school exams with a score of 94 percent. Although al-Sweikat confessed to throwing Molotov cocktails at security forces and running a chat group to help organize demonstrations, his father claimed in reality he joined the demonstrations just twice, for five minutes each time.

"He was subjected to psychological and physical abuse, which drained his strength," his father reportedly said. "The interrogator dictated the confession to Sweikat and forced him to sign it so that the torture would stop. He signed it."

United Nations human rights officials had previously written to Saudi officials regarding al-Sweikat's case, using an alternative spelling of his name. In November 2016, they wrote in regard to information they had received that he "was routinely subjected to torture including suspension from his hands and feet, sleep deprivation, severe beatings with cables and shoes, cigarettes burns and pouring of cold water on his body during winter. He was put in solitary confinement for three months. As a result, Mr. Suwaiket suffers from a broken shoulder, sustained pain in back and knees and blood deficiency due to insufficient nutrition. He has been deprived of any medical care. Mr. Suwaiket was reportedly subjected to acts of torture until he confessed to armed disobedience against the king and to attacking, shooting and injuring security forces, civilians and passers-by … On 1 June 2016, after several hearings, Mr. Suwaiket was convicted and sentenced to death by the [Specialized Criminal Court], on the sole basis of the confession extracted under torture."

Perpetrator ⓘ :
Is there explicit mentions of the perpetrator?
◉ Yes   ○ No
If Yes, please select the perpetrator type:
◉ State Actor   ○ Other Actors with State Permissions/Supports
○ Other Actors without State Permissions/Supports, Private Actors
○ Other actors (State's permission, support or acquiescence status unknown)
○ Insufficient information

Violation type ⓘ :
☑ Killing   ☐ Enforced disappearance   ☑ Torture   ☐ Arbitrary Detention
☐ Kidnapping   ☐ Other Harmful Acts

Victim Population:
◉ Named Individual   ○ Multiple (i.e. Group of unnamed individuals)

Victim Name:
Mujtaba al-Sweikat

Victim Keywords:
student,executed

Victim Type ⓘ :

Figure 7: Screenshot of the Qualification Task Example Page (1/2).

7101

The Saudi mission wrote, "The lower court judgment sentenced Mojtaba Suwaiket to death after convicting him of committing crimes such as: manufacturing firebombs (Molotov cocktails) and supplying them to others for use against law enforcement officers; throwing firebombs at law enforcement officers and their vehicles; involvement in the targeting of a security patrol by opening fire on it; monitoring police officers, their location and their movements and transmitting the information to members of another cell who used it to implement one of their operations; concealment of wanted individuals charged with opening fire on security patrols and persons charged with launching attacks on private property; concealment of persons charged with setting fire to a security vehicle and stealing a machine gun and bulletproof vest from it."

In a subsequent communication , the Saudi mission said al-Sweikat's death sentence was upheld on appeal by the Supreme Court, rendering it "final and enforceable."

Western Michigan previously confirmed al-Sweikat was admitted in 2013 but never enrolled. A university spokeswoman, Paula Davis, said Monday he was an applicant for pre-finance studies and English language. "We learned about his shocking death last week and grieve this tragic loss of a young life full of potential," she said.

The American Federation of Teachers, which represents Western Michigan faculty, was involved in campaigning for al-Sweikat's release.

"Saudi Arabia's sickening criminal beheading of a young student, after he was tortured and held in solitary confinement for years, is a despicable violation of international law and basic humanity," Randi Weingarten, the AFT president, said in a statement last week. "Condemned at 17, Mujtaba al-Sweikat was planning to attend Western Michigan University when he was arrested after attending a peaceful protest rally. Today, we discover this young man has been executed, along with more than 30 others, in a ghastly display of state brutality."

"If it was not already clear, Saudi Arabia, under the leadership of Crown Prince Mohammed bin Salman, has moved to the top tier of the bloodiest regimes in world history," he said. "We demand the U.S. government immediately condemn, in every way and with every means, this disgusting and outrageous crime."

⚑ | Report an error, an omission, a typo; suggest a story or a new angle to an existing story; submit a piece, a comment; recommend a resource; contact the webmaster, contact us: deathpenaltynews@gmail.com

Opposed to Capital Punishment? Help us keep this blog up and running! DONATE!

"One is absolutely sickened, not by the crimes that the wicked have committed,

but by the punishments that the good have inflicted." -- Oscar Wilde

Year of the Incident:

2019

Month of the Incident:

April

Day of the Incident:

N/A

Checkpoint:

Please copy and paste the following code into the text box:

0ac1d9b7-547a-4492-8878-8c7042cdffae

Copy/Paste the code above.

Location of the Incident:

N/A

N/A

Saudi Arabia

Back   Next

Figure 8: Screenshot of the Qualification Task Example Page (2/2).

Article 1 / 3

Instructions: Read the news article below then answer the questions.

Fijian union leaders detained as dispute with Water authority of Fiji intensifies

01-May-2019

"The New Zealand Council of Trade Unions is calling for the Government to step in over the ""appalling"" detainment of members of the Fijian trade union.

The problems stem from disagreements between the Water Authority of Fiji and its workers over the use of temporary work contracts.

Fijian Trade Union Congress secretary general Felix Anthony and other officials have been detained, a statement from the New Zealand Council of Trade Unions (CTU) claims.

Hundreds of Water Authority of Fiji workers have picketed the authority's depots in Wailoku, Lautoka, Kinoya, Wailoku and Labasa. Further protests were planned for later in the week.

READ MORE:

* Three trade unionists questioned by Fiji police for allegedly breaching the law

Earlier on Monday, National Union of Workers information relations officer Shiu Lingam, general secretary of Fijian Teachers Association Paula Manumanuitoga and Fiji Nurses Association general secretary Salanieta Matiavi were taken in for questioning by Fijian police.

CTU president Richard Wagstaff said the detainments were ""an appalling response from the government to Fijians exercising their democratic right; to organise and to protest"".

""This is an attack on the rights of working people in Fiji and is both outrageous and shocking. We stand in solidarity with Fijian working people and their trade unions.""

He called on the New Zealand Government to take action and request the immediate release of the union members.

The CTU plans a protest outside the Fijian Embassy on Pipitea St, Wellington, at 5.15pm on Thursday."

Perpetrator ❶ :

Is there explicit mentions of the perpetrator?
○ Yes   ○ No

Violation type ❶ :

Violation: The action of break or fail to comply with (a rule or formal agreement). You need to select all options that apply from the following three types of Violations:

1. Killing: Unlawful death inflicted upon a person with the intent to cause death or serious injury.
2. Enforced disappearance: Unlawful deprivation of liberty enforced or authorized by the state, that is not acknowledged by the state or the location of the victim is kept secret.
3. Torture: The action or practice of inflicting severe pain or suffering on someone as a punishment or in order to force them to do or say something.
4. Arbitrary detention: Arrest or detention not in accordance with national laws.
5. Kidnapping: Deprivation of liberty that is not enforced or authorized by the state.
6. Other harmful acts: Sexual violence or other acts causing or intending to cause harm, such as coercion or discrimination.

☐ Killing   ☐ Enforced disappearance   ☐ Torture   ☐ Arbitrary Detention

☐ Kidnapping   ☐ Other Harmful Acts

Victim Population:

○ Named Individual   ○ Multiple (i.e. Group of unnamed individuals)

Victim Keywords:

Enter any keywords about the victim, separated by ','

Victim Type ❶ :

☐ Journalist   ☐ Trade Unionist   ☐ Human Rights Defender

Figure 9: Screenshot of the Qualification Task Articles (1/3). By hovering over the information icon next to **Violation Type**, Turkers can check the definitions of all violation types on this page.

Instructions: Read the news article below then answer the questions.

Cuban police detain and beat journalist Roberto Jesús Quiñones in Guantánamo

24-April-2019

""The ongoing pattern of detentions of independent journalists like Roberto Jesús Quiñones shows that recent political reforms have not improved the situation for the press,"" said CPJ.

This statement was originally published on cpj.org on 24 April 2019.

Cuban authorities should immediately release journalist Roberto Jesús Quiñones, the Committee to Protect Journalists said today.

On April 22, at around 2:00 p.m., Cuban police agents detained Quiñones, a contributor to the news website CubaNet, as he was standing outside of the Guantánamo Municipal Tribunal, according to CubaNet and the Association for Press Freedom (APLP), a Cuban press freedom organization. At the time of his detention, Quiñones was covering the trial of two Cuban evangelical pastors facing charges for homeschooling their children, CubaNet reported.

While being transported in the police car, agents beat Quiñones, injuring his mouth, tongue, and right thumb and causing an inflammation in his right ear, his wife told APLP. In an audio recording of a phone interview with a journalist from U.S. government-funded Radio Martí recorded before police took his phone away – posted on Radio Martí's website and reviewed by CPJ – Quiñones said, "They broke my mouth, my shirt is covered in blood and I am detained here, I do not know why, because I was just here in the Municipal Tribunal. My mouth is fractured. They are about to take away my phone." CubaNet Executive Director Hugo Landa and Editor Roberto Hechavarría Pilia confirmed to CPJ that the voice in the audio is that of Quiñones.

Hechavarría Pilia told CPJ today that Quiñones remains in detention in a Guantánamo police station. Police in Havana detained another Cuban journalist, Augusto César San Martín, last week, according to CPJ reporting.

"The ongoing pattern of detentions of independent journalists like Roberto Jesús Quiñones shows that recent political reforms have not improved the situation for the press," said CPJ Central and South America Program Coordinator Natalie Southwick. "If President Miguel Díaz-Canel wants to show the world that his government is committed to positive change

Perpetrator ❶ :

Is there explicit mentions of the perpetrator?

○ Yes   ○ No

Violation type ❶ :

☐ Killing   ☐ Enforced disappearance   ☐ Torture   ☐ Arbitrary Detention

☐ Kidnapping   ☐ Other Harmful Acts

Victim Population:

○ Named Individual   ○ Multiple (i.e. Group of unnamed individuals)

Victim Keywords:

_____
Enter any keywords about the victim, separated by ','

Victim Type ❶ :

> **Victim**: A person harmed, injured, or killed as a result of a crime, accident, or other event or action. You need to select all options that apply from the following three types of Victims:
>
> 1. Journalist: A person observing events, statements, policies, etc. that can affect society, with the purpose of systematizing such information to inform society, including support staff, as well as community media workers and "citizen journalists" when they momentarily play that role.
> 2. Trade Unionist: A person exercising their right to form and to join trade unions for the protection of their interests. A trade union is an association of workers organized to protect and promote their common interests.
> 3. Human Rights Defender: A person exercising their right, to promote and to strive for the protection and realization of human rights and fundamental freedoms, including some journalists and trade unionists.
> 4. Insufficient information: There is not sufficient information available to make select one of the categories described above.

☐ Journalist   ☐ Trade Unionist   ☐ Human Rights Defender

☐ Insufficient Information

Figure 10: Screenshot of the Qualification Task Articles (2/3). By hovering over the information icon next to **Victim Type**, Turkers can check the definitions of all victim types on this page.

Instructions: Read the news article below then answer the questions.

Sweden grants asylum to two members of the Russian Pussy Riot collective

30-April-2019

Sweden has granted political asylum to two activist members of the Russian protest collective Pussy Riot.

Swedish public television SVT reported that the pair had feared being arrested if they had returned to Moscow.

Activists of the protest group Pussy Riot face the media after a recent arrest. Source: EPA

Advertisement

Lusine Djanyan and Alexey Knedlyakovsky initially lodged asylum claims in 2017, citing harassment and death threats in Russia as a result of their direct action and anti-government concerts.

These claims were rejected by Sweden's immigration office, but the couple appealed the decision.

"I am very happy that my children will be able to grow up in security, in particular in view of what has been going on (in Russia)" Ms Djanyan told SVT after the immigration tribunal granted their asylum appeal.

The collective - often decked out in neon balaclavas and tights - made its name with politically-charged performances, lambasting everything from the Russian church to the persecution of the country's gay community.

In 2016, Ms Knedlyakovsky was jailed for 15 days after hanging a cross on a statue depicting the head of the Russian FSB security service.

In August 2012, three members were sentenced to two years in a prison colony for "hooliganism and religious hatred" for performing a song protesting against Russian President Vladimir Putin in a central Moscow church.

Perpetrator ❶ :

Is there explicit mentions of the perpetrator?

○ Yes   ○ No

Violation type ❶ :

☐ Killing   ☐ Enforced disappearance   ☐ Torture   ☐ Arbitrary Detention

☐ Kidnapping   ☐ Other Harmful Acts

Victim Population:

○ Named Individual   ○ Multiple (i.e. Group of unnamed individuals)

Victim Keywords:

_____
Enter any keywords about the victim, separated by ','

Victim Type ❶ :

☐ Journalist   ☐ Trade Unionist   ☐ Human Rights Defender

☐ Insufficient Information

Victim Sex Type:

○ Male   ○ Female   ○ Unknown

Victim Age Group:

○ Adult (18 and above)   ○ Child (below 18)   ○ Unknown

Figure 11: Screenshot of the Qualification Task Articles (3/3).

Annotation Instructions

In this task, you will read **1** news article. The news article is about one or more attack events targeting human rights defenders or a similar population. You will then fill out a questionnaire using the information from the article.

Some important definitions:

- **Perpetrator**: A person who carries out a harmful, illegal, or immoral act. If one or more explicit mentions of perpetrators exist in the news article, you need to select **one** option from the following seven types of Perpetrators for the **primary** perpetrator:

  1. **State Security Forces**: Anyone employed by or representing a state institution, including government, courts, parliament, police, military.

  2. **Other State Actors**: Other actors that are a part of the state, e.g. civilian authorities such as Ministry of Interior, Ministry of Defence, or other non-military authority of a state, such as a President or Prime Minister, other ministers, regional governors, and associated staff / civilian administrators.

  3. **Armed actor on behalf of the State**: Armed actors that are not a part of the state but act with the state's permission, support or acquiescence, e.g. mercenaries where there is an agreement in place that they are working for the State.

  4. **Non-State Armed Actor**: Other actors / Private actors that are not a part of the state and act without the state's permission, support or acquiescence. E.g. armed groups who may have control over parts of a State's territory, "terrorist" groups, militias, etc.

  5. **Other Non-State Actor**: Other actors that are not a part of the state, e.g. private security companies, private enterprises.

  6. **Regional Or International Organizations**: Person or group working for a regional or international organisation, e.g. African Union, UN peacekeeping.

  7. **Insufficient information**: There is not sufficient information available to select one of the categories described above.

- **Violation**: The harmful, illegal, or immoral act carried by the perpetrator. You need to select **all** options that apply from the following seven types of Violations:

  1. **Killing**: Unlawful death inflicted upon a person with the intent to cause death or serious injury.

  2. **Enforced disappearance**: Unlawful deprivation of liberty enforced or authorized by the state, that is not acknowledged by the state or the location of the victim is kept secret.

  3. **Torture**: The action or practice of inflicting severe pain or suffering on someone as a punishment or in order to force them to do or say something.

Figure 12: Screenshot of the Full Task Instructions (1/2).

3. **Human Rights Defender**: A person exercising their right, to promote and to strive for the protection and realization of human rights and fundamental freedoms, including some journalists and trade unionists.

4. **Insufficient information**: There is not sufficient information available to make select one of the categories described above.

You can check all these definitions later on in the questionnaire interface by hovering your mouse cursor on the info icon ❶ next to each item.

About the questionnaire:

- **You must answer every question in the questionnaire.** If you cannot find the answer to a specific question, either select the answer **Insufficient Information, Unknown, Other** for selection questions, or type in **N/A** in the text box. You will not be able to submit the HIT unless you answered all the questions.

- For **Year, Month, Day, Location** related questions, if the event happened across multiple days and/or happened at multiple locations, please fill in with the starting time and/or the starting location.

- When answering about victims, you need to first identify if the victim mentioned in the article refers to **named individuals (someone with a name)** or **a group of unnamed individuals (without explicit names, such as "a group of students")**. In the latter case, if the group has mixed genders and age groups, select "Other" for Victim Sex Type and Victim Age Group questions.

- For each victim that has a name mentioned in the article, please select **"Named Individual"** in the questionnaire then fill in their name individually. If multiple names are mentioned, click the **"Add Another Victim"** button to add all victims with their corresponding names. Only select **"Multiple (i.e. Group of unnamed individuals)"** when multiple victims are mentioned in the article without their specific names.

- If the news article is not about a event targeting human rights defenders, please select the last option for selection questions and put N/A in the text boxes. For example, select "No/Unknown Harmful Acts" for "Violation type" question.

Read an example:

On the next page, you will see one example article and one questionnaire with correct answers already filled in. Please read through the example carefully.

You will be asked to copy a randomly generated code into one text box in the middle of the example questionnaire, you only need to do this once on the example page.

NOTE: Your submission will be REJECTED if the code you pasted in the box does not match the one shown to you.

<div align="center">

Next

</div>

Figure 13: Screenshot of the Full Task Instructions (2/2).

Instructions: Read the news article below then answer the questions.

${title}

${publish_date}

${article_interface}

Perpetrator ❶ :

Perpetrator: A person who carries out a harmful, illegal, or immoral act. You need to select one option from the following seven types of Perpetrators:

1. State Security Forces: Anyone employed by or representing a state institution, including government, courts, parliament, police, military.
2. Other State Actors: Other actors that are a part of the state, e.g. civilian authorities such as Ministry of Interior, Ministry of Defence, or other non-military authority of a state, such as a President or Prime Minister, other ministers, regional governors, and associated staff / civilian administrators.
3. Armed actor on behalf of the State: Armed actors that are not a part of the state but act with the state's permission, support or acquiescence, e.g. mercenaries where there is an agreement in place that they are working for the State.
4. Non-State Armed Actor: Other actors / Private actors that are not a part of the state and act without the state's permission, support or acquiescence. E.g. armed groups who may have control over parts of a State's territory, "terrorist" groups, militias, etc.
5. Other Non-State Actor: Other actors that are not a part of the state, e.g. private security companies, private enterprises.
6. Regional Or International Organizations: Person or group working for a regional or international organisation, e.g. African Union, UN peacekeeping.
7. Insufficient information: There is not sufficient information available to select one of the categories described above.

Is there explicit mentions of the perpetrator?

○ Yes   ○ No

Violation type ❶ :

☐ Killing   ☐ Enforced disappearance   ☐ Torture   ☐ Arbitrary Detention

☐ Kidnapping   ☐ Other Harmful Acts   ☐ No/Unknown Harmful Acts

Victim Population:

○ Named Individual   ○ Multiple (i.e. Group of unnamed individuals)

Victim Keywords:

Figure 14: Screenshot of the Full Task Annotation Page (1/2). By hovering over the information icon next to **Perpetrator**, Turkers can check the definitions of all perpetrator types on this page.

${title}

${publish_date}

${article_interface}

Victim Age Group:

○ Adult (18 and above)   ○ Child (below 18)   ○ Other   ○ Unknown

Add Another Victim

Year of the Incident:

Enter the Year of the incident in numbers

Month of the Incident:

Enter the Month of the incident in English, e.g. January, February

Day of the Incident:

Enter the day of the month of the incident in numbers

Location of the Incident:

Town / City

Region / State / Province

Country

Back   Submit

Figure 15: Screenshot of the Full Task Annotation Page (2/2).

B Label Statistics

In Table 7, we list the statistics of all the labels in HRDsAttack, as well as their distributions in the train, dev, and test set.

C Cohen-Kappa Scores

We calculate the average pair-wise Cohen-Kappa scores for each qualified Turker during our pilot study using 100 Hits with replication factor 3. While we did our due diligence to make our annotation instructions as comprehensive as possible, some of the concepts regarding Human Rights were sometimes challenging to distinguish for the Turkers. The relatively lower weighted average of Cohen-Kappa scores for some event attributes (PERPETRATOR MENTION: 0.40, PERPETRATOR TYPE: 0.41) are also potentially due to the imbalanced distributions of these attributes. The weighted averages of Cohen-Kappa scores for other attributes are all higher than 0.61 for violation and victim-related classes (VIOLATION TYPE: 0.67, VICTIM POPULATION TYPE: 0.64, VICTIM TYPE: 0.62), which are generally considered as substantially agree.

D Input-Output Design

Table 9 shows the input questions and answers for all the event attributes covered in HRDsAttack. And Table 10 shows all the task prefixes that we add to the beginning of the input text in our multi-task training regime.

E Long Document Solutions

Besides the two solutions we evaluate in the paper (Truncation and Knowledge Fusion), there are two other possible solutions that we describe here. First, some work proposes splitting a long document into shorter sequences, then using a transformer to generate sequence representations for each of them (Grail et al., 2021). Then those sequence representations are fed into another network to generate the final document representation. But in this case, a large number of training examples is required to learn the parameters of the network layers which generate the document representation. There is also the long transformer (Longformer (Beltagy et al., 2020)) approach proposed to handle long documents. But in contrast to the T5 model which is pretrained on many pretraining tasks, it is difficult to reframe all of the subtasks as a unified Sequence-to-Sequence task based on those long transformers. In comparison, the approaches we proposed are all post-processing steps that are less expensive than the aforementioned methods.

F Model Performances on the Development Set

Table 11 shows the performance of the models on the dev set of HRDsAttack. The best baseline model (T5 Hybrid) is chosen based on the model performance on the dev set.

Category	Event Attribute	Labels	Train	Dev	Test	Total
Perpetrator	Perpetrator Mention	Yes	272	95	91	458
		No	28	5	9	42
	Perpetrator Type	State Security Forces	149	60	56	265
		Other State Actors	25	6	10	41
		Other non-state actors	34	11	9	54
		Other actors with permissions	10	5	3	18
		Other actors without permissions	41	10	10	61
		Regional Organizations	4	1	1	6
		Insufficient Information	9	1	1	11
		None	28	6	10	44
Violation	Violation Type	Arbitrary Detention	138	53	55	246
		Enforced Disappearance	27	8	8	43
		Killing	109	33	34	176
		Kidnapping	76	21	16	113
		Torture	56	19	24	99
		Other	131	46	50	227
		Unknown	20	5	10	35
Victim	Victim Name	-	463	198	130	791
	Victim Type	Human Rights Defender	145	32	42	219
		Trade Unionist	59	25	9	93
		Journalist	195	104	66	365
		Insufficient Information	356	120	113	589
	Victim Population Type	Individual	463	198	130	791
		Multiple	224	74	74	372
	Victim Age Group	Adult	491	217	131	839
		Child	19	5	9	33
		Other	34	4	17	55
		Unknown	143	46	47	236
	Victim Sex Group	Man	274	115	90	479
		Woman	115	35	33	183
		Other	76	15	29	120
		Unknown	222	107	52	381
Location	Country	-	279	89	89	457
	Region	-	69	26	12	107
	City	-	163	53	45	261
Time	Year	-	268	83	80	431
	Month	January, ..., December	183	60	48	291
	Day	1, 2, 3, ..., 31	109	35	35	179

Table 7: Label statistics of HRDsAttack.

Worker	No. of HITs	Average Pair-wise Cohen-Kappa Score					Worker
Index	Finished	Perpetrator Mention	Perpetrator Type	Violation Type	Victim Population Type	Victim Type	Average
1	85	0.48	0.41	0.74	0.65	0.66	0.59
2	51	0.22	0.43	0.63	0.53	0.61	0.48
3	51	0.60	0.53	0.74	0.82	0.66	0.67
4	47	-0.04	0.31	0.72	0.74	0.57	0.46
5	38	0.45	0.37	0.70	0.46	0.63	0.52
6	15	0.71	0.49	0.45	0.71	0.38	0.55
7	9	1.00	0.24	0.32	0.40	0.80	0.55
8	2	0.50	0.50	0.00	1.00	0.00	0.40
9	1	0.00	0.00	0.00	0.00	0.00	0.00
10	1	0.00	0.00	0.00	0.00	0.00	0.00
Weighted Average		0.40	0.41	0.67	0.64	0.62	-

Table 8: Turker agreement scores for some of the event attributes calculated during the pilot study with 100 HITs, replication factor 3.

Category	Event Attribute	Input Question	Output Answer
Gen	Perpetrator Mention	Does it mention any perpetrator?	One of {Yes, No}
	Perpetrator Type	What is the type of the perpetrator?	One of {state security forces, regional organizations, other actors with permissions, other actors without permissions, other state actors, other non-state actors, insufficient info}
	Violation Type	Is there any arbitrary detention violation mentioned in the text? Is there any enforced disappearance violation mentioned in the text? Is there any kidnapping violation mentioned in the text? Is there any killing violation mentioned in the text? Is there any torture violation mentioned in the text? Is there any other violation mentioned in the text?	One of {Yes, No}
	Victim Name	Who is the victim of the violation?	{VICTIM_NAME1, VICTIM_NAME2, … }
	Country	In which country did the violation happen?	{COUNTRY_NAME}
	Region	In which region did the violation happen?	{REGION_NAME}
	City	In which city did the violation happen?	{CITY_NAME}
Vic	Victim Sex Type	What is the sex of {VICTIM_NAME}?	One of {woman, man, other, unknown}
	Victim Age Group	What is the age group of {VICTIM_NAME}?	One of {adult, child, other, unknown}
	Victim Population Type	What is the population type of {VICTIM_NAME}?	One of {Individual, multiple}
	Victim type	Is {VICTIM_NAME} a trade unionist? Is {VICTIM_NAME} a journalist? Is {VICTIM_NAME} a human rights defender?	One of {Yes, No}
Tim	Year	In which year did the violation happen?	Year (YYYY)
	Month	In which month did the violation happen?	Month (month name)
	Day	On which day did the violation happen?	Day (D with no leading zeros)
	Victim type	Is {VICTIM_NAME} a trade unionist? Is {VICTIM_NAME} a journalist? Is {VICTIM_NAME} a human rights defender?	One of {Yes, No}

Table 9: Summary of the predefined questions and answers for event attributes. **Gen** stands for the category of general article-dependent attributes, **Vic** stands for the category of victim-dependent attributes, and **Tim** stands for the category of publication time-dependent attributes.

Class	Task-prefix
Perpetrator Mention	detect perpetrator
Perpetrator Type	extract perpetrator type
Violation Type	extract violation type
Victim Name	extract victims
Victim Sex Type	extract victim sex
Victim Age Group	extract victim age
Victim Population Type	extract victim population type
Victim Type	extract victim type
Country	extract violation country
Region	extract violation region
City	extract violation city
Year	extract violation year
Month	extract violation month
Day	extract violation day

Table 10: Task Prefix for each event attribute.

Event Attribute	Metric	DyGIE++	T5 w/ Truncation	T5 w/ Knowledge Fusion	Hybrid
	Precision	**97.30**	95.88	95.96	95.96
Perpetrator Mention	Recall	37.89	97.89	**100.00**	100.00
	F1	54.55	96.88	**97.94**	97.94
Perpetrator Type	Accuracy	-	**68.00**	**68.00**	68.00
	Exact Match Precision	10.37	**85.07**	65.08	65.08
	Exact Match Recall	7.14	29.08	**41.84**	41.84
Victim Name	Exact Match F1	8.46	43.35	**50.93**	50.93
	Fuzzy Match Precision	19.26	**85.07**	73.02	73.02
	Fuzzy Match Recall	13.27	29.08	**46.94**	46.94
	Fuzzy Match F1	15.71	43.35	**57.14**	57.14
Victim Type	Accuracy	-	**81.25**	66.12	81.25
Victim Sex Type	Accuracy	-	**85.42**	80.87	85.42
Victim Age Group	Accuracy	-	97.92	**98.36**	98.36
	Precision	-	**64.14**	57.09	64.14
Violation Type	Recall	-	68.65	**76.22**	68.65
	F1	-	**66.32**	65.28	66.32
Country	Accuracy	-	**62.00**	59.00	62.00
Region	Accuracy	-	**9.00**	2.00	9.00
City	Accuracy	-	**20.00**	16.00	20.00
Year	Accuracy	-	**52.00**	46.00	52.00
Month	Accuracy	-	**32.00**	32.00	32.00
Day	Accuracy	-	**18.00**	10.00	18.00

Table 11: Overall performance of the baseline models on HRDsAttack dev set (%). All experiments are based on a single run with a preset random seed.

A   For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitations

☑ A2. Did you discuss any potential risks of your work?
Section Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B   ☑ Did you use or create scientific artifacts?

Section 5 Experiments

☑ B1. Did you cite the creators of artifacts you used?
Section 5 Experiments

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We plan on releasing our dataset under the MIT license as well pending legal approval. The LICENSE will be provided alongside the dataset as a text file on GitHub when the paper is published.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
The Term of Use for GDELT is: all datasets released by the GDELT Project are available for unlimited and unrestricted use for any academic, commercial, or governmental use of any kind without fee. We plan on releasing our dataset under the MIT license as well pending legal approval. The LICENSE will be provided alongside the dataset as a text file on GitHub when the paper is published.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We included the worker ID for each annotation in the dataset, which is anonymized.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.3 Annotation Process

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.5 Data Statistics

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ **Did you run computational experiments?**

Section 5 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5.3 Training Implementation

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.3 Training Implementation

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.4 Results and Table 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.4 Long Document Resolution

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Section 3.3 Annotation Process

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3.3 Annotation Process

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3.3 Annotation Process

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The Term of Use for GDELT is: all datasets released by the GDELT Project are available for unlimited and unrestricted use for any academic, commercial, or governmental use of any kind without fee.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We used a review process internal to our organization with HCI research scientists.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.3 Annotation Process