HumVI: A Multilingual Dataset for Detecting Violent Incidents Impacting Humanitarian Aid

Hemank Lamba[1], **Anton Abilov**[1], **Ke Zhang**[1], **Elizabeth M. Olson**[1], **Henry K. Dambanemuya**[3] *, **João C. Bárcia**[2], **David S. Batista**[2], **Christina Wille**[2], **Aoife Cahill**[1], **Joel Tetreault**[1], **Alex Jaimes**[1]
[1] Dataminr, Inc., [2] Insecurity Insight, [3] Northwestern University
{hlamba,aabilov,kzhang,elizabeth.olson,acahill,jtetreault,ajaimes}@dataminr.com
hdambane@u.northwestern.edu, christina.wille@insecurityinsight.org
{joaobarcia, dsbatista}@gmail.com

Abstract

Humanitarian organizations can enhance their effectiveness by analyzing data to discover trends, gather aggregated insights, manage their security risks, support decision-making, and inform advocacy and funding proposals. However, data about violent incidents with direct impact and relevance for humanitarian aid operations is not readily available. An automatic data collection and NLP-backed classification framework aligned with humanitarian perspectives can help bridge this gap. In this paper, we present HumVI – a dataset comprising news articles in three languages (English, French, Arabic) containing instances of different types of violent incidents categorized by the humanitarian sector they impact, e.g., aid security, education, food security, health, and protection. Reliable labels were obtained for the dataset by partnering with a data-backed humanitarian organization, Insecurity Insight. We provide multiple benchmarks for the dataset, employing various deep learning architectures and techniques, including data augmentation and mask loss, to address different task-related challenges, e.g., domain expansion. The dataset is publicly available at https://github.com/dataminr-ai/humvi-dataset.

1 Introduction

Violent events that impact humanitarian efforts, such as the looting of aid trucks and the kidnapping of aid workers, frequently hinder the delivery of life-saving aid and protection efforts, with devastating consequences for conflict-affected populations. Conflicts also severely disrupt existing food systems, healthcare, and education structures (Gates et al., 2012), leading to food insecurity, malnutrition (Martin-Shields and Stojetz, 2019), increased mortality due to inadequate healthcare (Garry and Checchi, 2020), and significant educational disruptions for children (Kadir et al., 2019). The systematic collection and collation of information about such events from multilingual data sources are not just crucial, but urgent, to ensure that decision-makers have access to the right information to support humanitarian operations and adequately respond to local needs.

*Work done during author's internship at Dataminr.

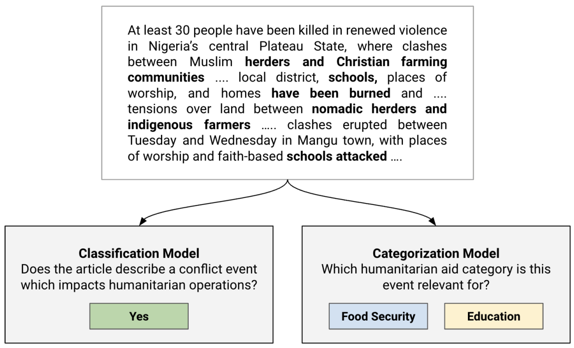

Figure 1: A sample input/output to the two NLP models for (i) relevance classification, followed by (ii) categorization of the event into humanitarian aid categories.

Aggregated information about violent events that impact humanitarian efforts can be operationalized by humanitarian organizations to systematically inform a wide range of different responses, including security risk management, program planning, and advocacy prioritization. Resources are severely limited during crisis response, and manual data collection in conflict-affected and highly insecure environments is rarely a priority. The use of NLP techniques can readily reduce the labor and resources required to maintain such a data collection pipeline. Importantly, it can also be carried out remotely, thus reducing the security exposure of those supporting a humanitarian response through data collection and analysis. Most NLP capacity in the humanitarian context has focused on English, limiting the use of models in conflict-affected countries where information is shared in French, Arabic, and other languages.

In order to process large amounts of disparate data to detect violent events, it is essential to build NLP models that can automatically detect such incidents. Next, these incidents should be tagged appropriately with the relevant type of humanitarian operation, e.g., an event where aid workers are threatened or harmed would be tagged as *aid security*.

We identify two primary gaps in the current publicly available datasets: (1) Most are focused on identifying a single type of event, such as disaster-related information (Alam et al., 2021, 2018), specific threats against human rights defenders (Ran et al., 2023), or civil unrest (Delucia et al., 2023). Even when datasets cover a wide range of events and associated tags (Raleigh et al., 2010), these tags often do not indicate the specific humanitarian sectors impacted by the event and lack information related to downstream humanitarian efforts, such as *healthcare* or *food security* (Trivedi et al., 2020). (2) Most of the humanitarian event detection datasets are disproportionately English and this has been identified as a key issue for improving adoption of NLP techniques in the humanitarian sector (Kreutzer et al., 2019; Rocca et al., 2023). The dataset presented in this paper is multilingual, and we hope it will support and encourage NLP practitioners to create multilingual models for humanitarian purposes.

To address these two issues, we introduce a new dataset HUMVI (HUManitarian Violent Incidents). This dataset comprises $17,497$ articles in three languages, with each article tagged for its relevance and, if relevant, categorized by the key humanitarian response sector where it is relevant, i.e., aid security, education, food security, health, and protection (a sample is shown in Figure 1).

There are three qualities of this dataset that ensure its value to the field.

Humanitarian Expert Verified. A key challenge of building such a dataset is obtaining labels from the end-users, i.e. humanitarian experts, who will consume the tagged dataset to provide aid-centred event monitoring and context analysis. To solve this, we partner with Insecurity Insight[1], a data-driven humanitarian to humanitarian (h2h) organization that has expertise in applying and consuming tagged news articles in order to support the aid sector with aid-focused information.

AI for Social Good. Recently, researchers have focused on multiple problems under the umbrella of AI for Social Good (AI4SG) (Shi et al., 2020). However, some of the SDGs (Sustainable Development Goals) are under-represented within AI community efforts (Gonzalez et al., 2023), including SDG 2: *Zero Hunger* (Fund, 2015). A focus of our work is *food security*, which aims to tag conflict events impacting food security. We believe that having a labeled dataset directly related to this SDG will encourage more NLP researchers to develop technology to benefit organizations that rely on understanding the impact of this SDG.

Domain Expansion. The dataset-building task was carried out in conjunction with Insecurity Insight. While the organization had previously developed tagging systems for health, education, protection, and aid security events, expanding to a new category (food security) and new languages (French and Arabic) was an obstacle for them, given their limited staffing. This mirrors a challenging problem in the NLP space: domain expansion, *i.e.*, expanding a machine learning model to work with new classes and on new input sources. In this paper, we provide two versions of the dataset: (a) HUMVI-core and (b) HUMVI-expansion. The core dataset focuses on four possible event types and is available only in English. However, the expansion dataset extends the core dataset by adding two new languages (French and Arabic) and one new category, food security. We believe that the expansion dataset will enable NLP researchers to benchmark their solutions for the real-world challenges frequently encountered by humanitarian organizations while working in a capacity-constrained setting.

To summarize, our contributions are follows:

[C1] Humanitarian Violent Event Dataset. We provide a comprehensive multilingual dataset of violent events, labeled by humanitarian experts for relevancy to multiple key humanitarian aid sectors.

[C2] AI for Good. To the best of our knowledge, this is the only dataset that provides instances of conflict events that might lead to food insecurity, an under-represented research area.

[C3] Domain Expansion. HUMVI-expansion provides a dataset that is representative of a challenging real world problem, *i.e.*, resource-constrained domain expansion (Yang et al., 2022).

[C4] Baseline Experiments. We show leading NLP models perform on this dataset, thus setting a baseline for future work while showcasing the complexities and challenges of adapting NLP techniques to real world scenarios.

We make the dataset and associated repository

[1] https://insecurityinsight.org/

public at https://github.com/dataminr-ai/humvi-dataset.

2 Related Work

2.1 Event Detection

The development of large-scale geographically and temporally disaggregated conflict-event datasets emerged in the early 2010s, in large part to facilitate research on conflict dynamics. Among the earliest and best known are the Armed Conflict Location and Event Dataset (ACLED) (Raleigh et al., 2010), Social Conflict Analysis Database (SCAD) (Salehyan et al., 2012), the Uppsala Conflict Data Program (UCDP) (Sundberg and Melander, 2013), and GDELT (Global Database of Events, Language and Tone) (Leetaru and Schrodt, 2013). These datasets are widely leveraged by humanitarian practitioners for conflict research, early warning, and crisis response (Donnay et al., 2019; Sur et al., 2019; Hegre et al., 2019; Penson et al., 2024; GRANIT, 2024). Most existing datasets are generic and extensive, covering events from various geographies. In contrast, HUMVI focuses on events with direct humanitarian impact and relevance to humanitarian aid operations. The applied tags indicate the specific humanitarian aid sector(s) where the event is relevant.

Additionally, another category of event detection dataset exists that is more specific and task-oriented. Recent work include datasets for detecting attacks against human right defenders (Ran et al., 2023), monitoring infrastructure construction that threatens environmental conservation (Keh et al., 2023), detecting gun violence related attacks (Pavlick et al., 2016), classifying types of crime in social media messages (Jarquín-Vásquez et al., 2023), detecting human rights violations (Pilankar et al., 2022), and tracking migrant deaths and disappearances (Brian and Laczko, 2014). Though these datasets are carefully curated, they are limited to detecting events that are related to a single specific event type. Our dataset, on the other hand, provides relevance labels for several types of violent events that are key for contextualizing and informing the timely delivery of humanitarian aid in conflict settings.

The real-time nature of social media has made it a popular data source for detecting events in real time (Fedoryszak et al., 2019). Most of the work has been centered on detecting crisis events with the goal of informing appropriate humanitarian response. Alam et al. (2021) proposed *CrisisBench*, a dataset comprising 310K tweets annotated for informativeness and humanitarian response for multiple different types of crisis events, e.g., flood, earthquake, etc. Similarly, Parraga-Alava et al. (2021) created a dataset comprising 25K Spanish tweets indicating emergency or non-emergency related content in Ecuador (Parraga-Alava et al., 2021). Beyond classification, multiple new tasks like summarizing the crisis related tweets for effective humanitarian response have also been proposed (Yela-Bello et al., 2021; Faghihi et al., 2022). Though useful, event detection on social media platforms often face challenges associated with veracity, (*i.e.*, able to filter out potentially unverified information or disinformation) and velocity, (*i.e.*, the ability to handle rapid streams of messages at scale) (Panagiotou et al., 2016), which are very different from the challenges associated with our dataset.

Datasets	Properties			
	News Domain	Usecase Tags	Multi Lingual	Manual Labels
GDELT	✔	✘	✔	✘
ACLED	✔	✘	✔	✔
CrisisBench	✘	✘	✘	✔
HumSet	✘	✔	✔	✔
HUMVI	✔	✔	✔	✔

Table 1: A brief overview of the datasets in this space in comparison to the proposed dataset.

Our work is closest to HumSet (Fekih et al., 2022) which is a multilingual dataset of humanitarian response documents where each entry has been annotated for different sectors, and its impact and needs. HumSet is associated with humanitarian response documents which are typically reports generated by actors in the humanitarian sector (NGOS, UN agencies, etc), whereas HUMVI is focused on news articles and can be used to inform the humanitarian response reports. We summarize major differences in Table 1.

2.2 NLP for Social Good

Advancements in NLP technologies have broadened their application scope across education (Kasneci et al., 2023; Madnani and Cahill, 2018), assistants (de Barcelos Silva et al., 2020), legal fields (Martinez-Gil, 2023), and more. With these advancements, researchers are increasingly prioritizing NLP applications for societal benefit (Cowls et al., 2021). Significant positive impacts have been achieved in healthcare (Ghassemi et al., 2020), disaster response (Arachie et al., 2020; Liu et al., 2021; Alam et al., 2021), climate change (Luo et al., 2020; Diggelmann et al., 2020), and mental health (Pérez-Rosas et al., 2019; Evensen et al., 2019), aligned with UN SDGs (Gonzalez et al., 2023). Despite being rated as important (Yang et al., 2020), some SDGs like SDG1 (No Poverty), SDG2 (Zero Hunger), SDG6 (Clean Water), and SDG13 (Climate Action) receive little attention from NLP researchers. Our dataset includes labeled instances relevant to humanitarian concerns, addressing issues such as food insecurity linked directly to SDG2 (Zero Hunger). This dataset can support NLP researchers exploring these critical areas.

2.3 Domain Expansion

A key challenge in NLP research is to develop multilingual models. Early methods included cross-lingual transfer of monolingual embeddings, aligned using unsupervised techniques (Lample et al., 2017; Conneau et al., 2017). A natural extension was to create multilingual embeddings that can be used directly (Ruder et al., 2019). Following the success of transformer models, pre-trained multilingual models were created (Devlin et al., 2018; Ebrahimi and Kann, 2021). With the availability of large-scale data and compute, massive encoder-decoder models (Liu et al., 2020; Xue et al., 2020) and then eventually decoder-only models (Achiam et al., 2023) were trained. These models can be leveraged in zero-shot or few-shot settings and still be performant without requiring any specific fine-tuning. Another set of approaches includes augmenting the training dataset by translating it into the target language (Edunov et al., 2018).

Another key challenge is to how to extend the model to new categories, where getting high-quality labeled data for training is expensive or unavailable. Data augmentation is one of the most popular techniques that transforms existing data in a reliable class-preserving manner to create new data (Chen et al., 2023). Self-training (Du et al., 2020; Scudder, 1965) is a promising technique that enables leveraging unlabeled datasets to create pseudo-labels which can be used to further improve the model. Other techniques like multi-task learning and consistency regularization have also been used to tackle this issue. However, more recently, the most popular methods have been to scale up language models and leverage their zero-shot or few-shot capabilities to achieve strong performance (Brown et al., 2020). The dataset presented in this paper poses multilinguality and limited data learning as two of its main challenges. As part of our process of benchmarking the dataset, we apply the methods mentioned above and demonstrate their performance.

Figure 2: A schematic representation of Insecurity Insight's data processing system.

3 Data Collection

3.1 Data Collection Overview

As mentioned in Section 1, we worked closely with Insecurity Insight to create HumVI. Insecurity Insight already had an established workflow for sourcing relevant articles, after which humanitarian experts classify them for relevance, and then tag them with the appropriate categories that capture the downstream humanitarian concern (Figure 2). They were also leveraging multiple ways to identify links to news articles, which we list below:

NewsAPI. Researchers have extensively used NewsAPI (Lisivick, 2018) to collect news articles for various purposes (Keh et al., 2023; Jain et al., 2024). Similarly, for this dataset, NewsAPI was used to search for English articles about violent conflict incidents using a curated list of domain-specific keywords, detailed in Appendix A.1.

OSAC. The Overseas Security Advisory Council[2] is a US Department of State organization whose primary purpose is to share critical information related to security. As part of this effort, at regular intervals, they share English articles selected by OSAC analysts that are highly relevant to global security. This list is collected by Insecurity Insight, and added to the queue to be reviewed by their experts.

Manual Collection. Besides classifying articles as relevant and tagging them with appropriate categories, experts at Insecurity Insight also manually upload news articles to the database. These articles may have been missed by automatic scrapers, and

[2] https://www.osac.gov/Content/Browse/News

Category	Description	Example (Title with URL in bracket)
Aid Security *(aid)*	Harms or threats towards aid agencies or aid workers.	Aid group says Israeli strike kills 7 of its workers in Gaza, including foreigners. [https://tinyurl.com/4szy5d6d]
Education *(edu)*	Harms or threats towards education providers, education workers, education related infrastructure and students.	Nigeria: Gunmen Kidnap Nine Students in Delta. [https://tinyurl.com/vduwuxyd]
Food Security *(food)*	Harm or threats towards entities involved in food production, processing, and distribution, as well as damage to transport and energy infrastructure that supports the food supply chain.	Benue buries 17 persons killed by 'herders' as governor suggests solution. [https://tinyurl.com/2jvjyxva]
Health *(hlth)*	Harm or threats towards health providers or health workers.	FCT NMA raises concern over kidnapping of members. [https://tinyurl.com/4uv49u7f]
Protection *(prtc)*	Harm or threats towards internally displaced persons and refugees.	Israeli Airstrike Kills 36 Palestinians During Suhoor Meal in Nuseirat Refugee Camp. [https://tinyurl.com/5abcbaz7]

Table 2: Description of different categories along with examples. The shorthand used for the categories is in brackets. In particular, Food Security is the new category being added as part of expansion dataset

hence need to be added to ensure more comprehensive coverage.

3.2 Source Expansion

One of the key challenges that motivated the collection of HumVI was the need to expand the current data collection and tagging process to include a new category (food security) and new languages (French and Arabic), which were rarely covered in the original sources. To enhance sourcing, we added GDELT (Leetaru and Schrodt, 2013), a large open-source database of multilingual news article links. We query the GDELT Event Database for articles which are regionally tied to either Burkina Faso, Cameroon, the Central African Republic, the Democratic Republic of the Congo, Palestine, Haiti, Mali, Niger, Nigeria, Somalia, Syria or Yemen. We scrape the full article content and detect the article language, and English, French and Arabic articles are added to the database.

3.3 Labeling Process

Once the potentially relevant news articles are identified, the title and full text of the news article is collected. The human expert first annotates the relevance of the article (a binary classification). The relevance is determined by assessing whether the article describes a conflict event and the event belongs to one or more of the pre-defined humanitarian categories (aid security, education, food security, health and protection). For a category to be applied the article should describe reports of conflict events happening at a particular time and location, typically described as being carried out by a *perpetrator* against a given *person who fulfills a specific humanitarian function* (*e.g.* an aid or health worker) or *essential civilian infrastructure* (*e.g.* aid convoys, hospitals, bakeries or schools) [3]. A brief overview of the categories is given in Table 2 (detailed description in Appendix B).

We obtained the labels in two ways: (1) On-the-job labeling (OTJ) and (2) Offline labeling (OFL). Data collected via OTJ originates from the partnering organization's established production workflows and was collected live from the English-only data sources News API and OSAC. The scraped article title and content is reviewed by the expert annotators in their internal annotation tool before they determine whether the article is relevant and assign the event categories. For OFL, we worked with 7 humanitarian experts from Insecurity Insight to expand the data collection to include articles from GDELT in English, French, and Arabic. We follow similar annotation guidelines as in OTJ. The annotators performed the offline task in spreadsheets. They reviewed the scraped article title and full text before assigning one or more event categories as per the guidance. If none of the event categories are assigned, the sample is marked as non-relevant. HumVI contains OTJ labels for most training data and OFL labels for the test data.

3.3.1 Quality Control

Both methods of soliciting labels are carried out by a team of domain experts trained by Insecurity Insight. For OTJ labeling, the labeling correctness is ensured by significant training efforts, well defined guidelines, and the periodic review and correction

[3]Detailed definitions are provided here: https://insecurityinsight.org/methodology-and-definitions

by a supervising expert. OFL labeling was carried out by experts who have been trained similarly and were asked to follow similar guidelines.

Label	Language English A=3, N=100	French A=7 N=75	Arabic A=7 N=75
Relevance	0.89	0.77	0.86
Aid	0.91	0.89	0.66
Education	0.95	0.67	0.64
Health	0.88	0.80	0.75
Protection	0.96	0.78	0.88
Food	0.81	0.59	0.69

Table 3: Average pair-wise Kappa score for N Samples in each language across A annotators.

Source	Language English	French	Arabic
News API	7,486	0	0
OSAC	7,069	0	0
GDELT*	867	894	900
Manual*	209	35	37
Total	15,631	929	937
Relevant	3,049	233	358
Unlabeled	41,589	4,689	8,977

Table 4: Source and Language Distribution. * indicates the labels for data source were collected in OFL manner.

We also measured inter-annotator agreement rates in English, French and Arabic. As shown in Table 3, the average annotator agreement rate is high (indicating substantial to near perfect agreement (Landis and Koch, 1977)), and only slightly worse for some annotators on the new category being introduced (food security). Extended annotation guidelines and details are provided in Appendix B.

We removed articles from the collected data if the crawl failed to retrieve the text or if the article contained no text. Additionally, we eliminated duplicate articles with identical lowercase text or identical URLs.

3.4 Data Description

HumVI includes $17,497$ labeled articles in English, French, and Arabic. The high-level statistics are listed in Table 4. English articles are predominantly sourced from NewsAPI and OSAC, whereas French and Arabic articles were either sourced from GDELT or collected manually as shown in Table 4. Note that the NewsAPI, OSAC and English articles are disproportionately high in our dataset because the majority of those instances were part of the standard workflow at Insecurity Insight when data collection began. Labels for NewsAPI and OSAC articles have been obtained through OTJ labeling, whereas all GDELT and manually collected articles are labelled through OFL.

4 Dataset Split and Variants

We provide two different variants of the dataset described in the previous section.

HumVI-core. The core dataset is English only, and contains labels for four categories relevant to humanitarian work – it does not include any label for food security. The dataset consists of $15,631$ news articles.

HumVI-expansion. The expansion dataset is centered on the real world challenge of expanding to a new category (food security) and new languages (French, Arabic). This dataset variant is a superset of HumVI-core, and contains $17,497$ articles. This dataset poses multiple interesting challenges for an NLP practitioner - (i) how to expand to new modeling classes and languages (ii) how to build performant models in a resource-constrained manner.

For both variants of the dataset, we temporally split the data on 2024-02-08 such that all of the articles before this date are part of the train set. The test set is collected exclusively through the OFL labeling process and consists of news articles collected after 2024-03-15. All news articles collected between 2024-02-08 (train cutoff) and 2024-03-15 (test start date) are part of an unlabeled dataset and are also released along with the paper.[4] We present detailed numbers on the dataset split in Table 5. Note that in the training dataset there are very few instances of French and Arabic, however, in the test dataset, there are a similar number of articles for all three languages. Similarly, the presence of the food security tag is under-represented in the training dataset compared to other categories, but equally represented in the test dataset.

Along with HumVI-expansion, we also release a large collection of unlabeled data that can be used to improve the model's performance through active learning (Settles, 2009) or through semi-supervised learning (Learning, 2006).

[4]The unlabeled dataset only contains articles from GDELT. We have found GDELT to be a superset of query based News-API access and OSAC.

lang	Train Dataset aid	edu	prtc	hlth	food	total	Test Dataset aid	edu	prtc	hlth	food	total
en	**518**	**1,101**	**256**	**900**	241	14,593	**124**	**70**	**77**	**149**	134	1,038
fr	14	1	7	9	44	133	58	60	12	69	52	796
ar	10	2	22	20	47	137	90	57	91	106	72	800

Table 5: Language and class distribution across train test split of the dataset. The table represents HUMVI-expansion, the **bolded** entries are a part of HUMVI- core.

5 Models

We experiment with multiple different model types to establish benchmarks for two tasks on HUMVI: (1) relevance classification (i.e., predict if the given news article is relevant or not); and (2) category classification (i.e., predict one or more categories where news article might be relevant).

5.1 Base Architectures

Fine-tuned Models. In our experiments, we train separate models for each task. However, they could also be modeled jointly by training a single model to predict both relevance and category labels - which we leave as an open research direction. We experiment with multiple different model architectures to benchmark their performance on HUMVI.

We experiment with fine-tuning multiple different transformer models, specifically monolingual (English) and multilingual variants of `DistilBERT` (Sanh et al., 2019), `BERT` (Devlin et al., 2018) and `RoBERTa` (Liu et al., 2019; Conneau et al., 2019).

LLM. We benchmark five zero-shot LLM models on the dataset: two open-source models, `DBRX` (Databricks, 2023) and `LLaMa3-70b` (AI@Meta, 2024), and three proprietary models, `GPT-4` (Achiam et al., 2023), `GPT-4o` (OpenAI, 2024), and `Mistral Large` (AI, 2024). All models use $temperature = 0$ and the same chain-of-thought-style prompt, derived from the annotation protocol with minimal tuning.[5] The LLM experiments are set up as a single *implicit* category classification task: if an article is labeled with a category, it is considered relevant.

5.2 Domain Expansion

To handle domain expansion we experiment with the following two approaches.

Translation Augmentation. In order to adapt to new languages, we create synthetically labeled data by translating English samples into French and Arabic (Edunov et al., 2018).

Model	relev.	aid	edu	hlth	prtc
`DistilBERT`	0.84	0.86	0.86	0.83	0.83
`BERT`	0.84	**0.87**	0.83	0.84	**0.84**
`RoBERTa`	**0.86**	**0.87**	**0.87**	**0.85**	**0.84**
`DBRX`	0.67	0.41	0.26	0.54	0.31
`Llama-3-70B`	0.72	0.60	0.66	0.70	0.37
`Mistral-Large`	0.79	0.75	0.69	0.76	0.42
`GPT-4-Turbo`	0.80	0.76	0.69	0.75	0.47
`GPT-4-o`	0.81	0.76	0.75	0.72	0.53

Table 6: F1 Scores of various models on HUMVI-core for both relevance and category classification.

Masking Loss. Since labeling for the food security tag started post the train cutoff date in OFL manner, most instances in the HUMVI-expansion training dataset lack this label. To prevent learning from partially labeled data, we further use label loss masking (Duarte et al., 2021) in category classification. The Binary Cross Entropy (BCE) loss is average pooled only over labeled categories, excluding food security.

Model	English	French	Arabic
`DistilBERT`	0.82	0.74	0.79
`BERT`	**0.83**	0.76	0.80
`XLM-RoBERTa`	0.82	**0.77**	**0.83**
`DBRX`	0.71	0.62	0.66
`Llama-3-70B`	0.73	0.60	0.67
`Mistral-Large`	0.78	0.76	0.74
`GPT-4-Turbo`	0.78	0.74	0.71
`GPT-4-o`	0.80	0.69	0.76

Table 7: F1-Scores on HUMVI-expansion for relevance classification.

6 Results

Metric of Choice. While there is a range of metrics that could evaluate model performance, we settled on the following use-case-specific metrics together with Insecurity Insight. For the relevance classifier, we compute precision at a minimum recall of 0.8 to surface more relevant content. For the category

[5]The full prompt is available here.

Model	English					French					Arabic				
	aid	edu	hlth	prtc	food	aid	edu	hlth	prtc	food	aid	edu	hlth	prtc	food
Distil-BERT	0.83	0.87	0.79	0.81	0.45	0.79	0.87	0.80	0.67	0.34	0.79	0.82	0.76	0.85	0.32
BERT	**0.85**	**0.88**	0.82	0.82	0.51	**0.84**	**0.88**	**0.85**	**0.73**	0.44	0.80	**0.86**	0.79	**0.84**	0.28
XLM-R	**0.85**	**0.88**	**0.82**	**0.84**	**0.58**	**0.84**	**0.88**	0.82	0.68	**0.64**	**0.82**	**0.86**	**0.83**	**0.84**	**0.41**
DBRX	0.41	0.26	0.54	0.31	0.37	0.31	0.46	0.59	0.08	0.27	0.27	0.24	0.51	0.35	0.20
Llama3	0.60	0.66	0.70	0.37	0.41	0.53	0.76	0.69	0.09	0.33	0.46	0.51	0.65	0.39	0.26
Mistral	0.75	0.69	0.76	0.42	0.44	0.66	0.80	0.78	0.19	0.41	0.60	0.55	0.67	0.46	0.30
GPT-4T	0.76	0.69	0.75	0.47	0.29	0.62	0.80	0.77	0.24	0.35	0.70	0.56	0.76	0.24	0.18
GPT-4o	0.76	0.75	0.72	0.53	0.34	0.67	0.82	0.76	0.15	0.35	0.70	0.66	0.69	0.42	0.24

Table 8: F1-Scores on HUMVI-expansion for category classification.

Lang	Mode	DistilBERT	BERT	XLM-RoBERTa
English	Base	0.81	0.82	0.83
	wAug	0.82	0.83	0.82
French	Base	0.69	0.76	0.81
	wAug	0.74	0.76	0.77
Arabic	Base	0.69	0.75	0.81
	wAug	0.79	0.80	0.83

Table 9: Relevance classification F1 scores of finetuned models on HUMVI-expansion before (**Base**) and after (**wAug**) translation augmentation.

classifier, we optimize for precision and compute recall at a minimum precision of 0.8. We report the harmonic mean F1 score of these metrics averaged across five runs. For LLMs, we compute the F1 score directly based on the predictions without applying a threshold.

Results on HUMVI-core. We report results in Table 6. The fine-tuned transformers-based methods perform best for both relevance and categorization tasks. The large language models did not perform as well for either the relevance or category classification tasks. We believe that the relatively low performance of LLMs can be attributed to the prompt being too large and occupying a lot of context length. Future work could include reducing the prompt length, and then running few shot inference to try and improve performance.

Results on HUMVI-expansion. We note that XLM-RoBERTa performs the best in adapting to the new category and new languages with limited labeled data for both relevance classification (Table 7) and categorization (Table 8). Unsurprisingly, all models struggle with performance on the new category (food security). The challenge of adapting to new languages (French, Arabic) is however manageable by leveraging multilingual pretrained transformers. LLMs do not perform as well as the finetuned models especially for the French and Arabic protection category. For LLMs, the proprietary models outperform the open-source ones, sometimes by a large margin. For example, the difference in performance between Llama-3 and GPT-4o is 1.2x for the English aid security category, and up to 1.5x for the Arabic aid security category.

Effect of Translation Augmentation. We report the effect of translation augmentation on relevance classification performance for fine-tuned models for HUMVI-expansion in Table 9. The results for categorization are presented in Table 10. We note that, in general, translation augmentation improves model performance for new languages introduced, particularly Arabic. Similar trends are also observed for other fine-tuned transformer models for the category classification task.

Effect of Masking Loss. In general, we see that the application of masking loss improves the performance of the fine-tuned models (Table 10) in certain cases. However, for food security, the performance increases for English, but decreases considerably for French and Arabic.

For future modeling work, we report some misclassification themes (Appendix D)

7 Conclusions

In this paper, we presented HUMVI, a multilingual dataset of $17K$ news articles classified by their relevance to humanitarian aid efforts and tagged with appropriate categories. Developed in collaboration with Insecurity Insight, this dataset addresses real-world challenges in optimizing data workflows with limited resources using NLP techniques. We trained and evaluated multiple NLP models for relevance and event classification, showing that both fine-tuned models and zero-shot LLMs can perform well, leaving room for improvement.

		English					French					Arabic				
Architecture	Mode	Aid	Edu	Hlth	Prtc	Food	Aid	Edu	Hlth	Prtc	Food	Aid	Edu	Hlth	Prtc	Food
distilBERT	Base	0.82	0.87	0.81	0.83	0.28	0.70	0.82	0.73	0.16	0.13	0.64	0.70	0.69	0.16	0.13
	wAug	0.84	0.87	0.79	0.81	0.33	0.84	0.87	0.82	0.58	0.46	0.78	0.82	0.76	0.85	0.26
	wMask	0.83	0.86	0.80	0.82	0.41	0.76	0.83	0.76	0.15	0.05	0.71	0.72	0.72	0.33	0.19
	wBoth	0.83	0.87	0.79	0.81	0.45	0.79	0.87	0.80	0.67	0.34	0.79	0.82	0.76	0.85	0.32
BERT	Base	0.85	0.88	0.82	0.82	0.41	0.79	0.88	0.81	0.44	0.52	0.77	0.81	0.78	0.83	0.31
	wAug	0.85	0.87	0.83	0.81	0.46	0.85	0.88	0.85	0.40	0.52	0.80	0.85	0.80	0.83	0.10
	wMask	0.85	0.88	0.83	0.83	0.53	0.82	0.88	0.81	0.49	0.42	0.77	0.81	0.79	0.84	0.25
	wBoth	0.85	0.88	0.82	0.82	0.51	0.84	0.88	0.85	0.73	0.44	0.80	0.86	0.79	0.84	0.28
XLM-RoBERTa	Base	0.85	0.87	0.83	0.83	0.48	0.83	0.87	0.83	0.45	0.65	0.81	0.83	0.82	0.72	0.52
	wAug	0.86	0.88	0.82	0.82	0.46	0.84	0.88	0.83	0.68	0.65	0.83	0.86	0.84	0.83	0.29
	wMask	0.86	0.87	0.82	0.83	0.52	0.83	0.88	0.81	0.62	0.50	0.80	0.83	0.82	0.80	0.45
	wBoth	0.85	0.88	0.82	0.84	0.58	0.84	0.88	0.82	0.68	0.64	0.82	0.86	0.83	0.84	0.41

Table 10: F1 Categorization Performance for HUMVI-expansion. Different variations of the fine-tuned models with Augmentation (wAug), Mask Loss (wMask) and both (wBoth) are shown.

We believe that *humanitarian organizations* can utilize this dataset to classify real-time news articles based on relevance and tag them with specific humanitarian aid sector relevant categories. The results can help determine whether LLMs or finetuned transformer-based models are suitable for their processes. Additionally, organizations needing extra tagging for humanitarian events associated with the text can benefit from the models trained using the provided data and code.

For *NLP researchers* we hope that they will find this dataset a valuable resource for addressing real-world problems in the AI4SG space, particularly for developing techniques in resource-limited scenarios. The dataset presents interesting challenges, with significant potential for improving baseline performance.

8 Limitations

Like any data-based study, the dataset presented here is subject to multiple limitations. We take a critical approach for the dataset we have created and list the limitations and biases that could exist at each stage of the process followed. For *data collection*, our dataset comes from a limited set of countries and hence might not be readily extensible to different parts of the world that we have not covered. Additionally, the differences in event reporting and content might need to be analyzed before being applied to a new geographic area. Similarly, we leverage three different ways of obtaining news article links. However, these may be systematically biased towards certain types of content, and they may ignore other types of content, e.g., hyperlocal news reporting in local languages that these link curation systems have not included. We rely heavily on leveraging GDELT to retrieve news articles in different languages from varied geographies. However, GDELT's coverage might not be equitable across different languages and regions (Leetaru and Schrodt, 2013), resulting in under-representation of certain geographies. For *annotation*, we leverage the labels that have been defined by Insecurity Insight. Though the labels were created with due deliberation and discussion with humanitarian experts, they are still tuned for Insecurity Insight's needs. Therefore, if another organization uses the dataset, they need to align with the label definitions provided here or relabel relevant subsets of the dataset for their needs if the definitions do not align. We hope that this disclosure will be helpful to any user of this dataset in the future.

Ethical Considerations

The dataset is created using publicly available news articles, and does not breach any contract for obtaining the data. We have ensured that the web scraper only accesses publicly available data and excludes any data behind a paywall. Furthermore, we release only the links to the article, along with a scraper code. Therefore, if any source website changes its web scraping policies in the future, that change will be reflected when retrieving the articles. Although we cannot guarantee that PII (personally identifiable information) is not included in any of the news articles, we only source from published, publicly available materials. For the annotation process, we leveraged internal humanitarian experts at the partnering organization Insecurity Insight, who were duly compensated for their services during the course of their professional, paid employment.

Acknowledgements

We thank our colleagues Sirene Abou-Chakra and Jessie End from Dataminr's AI for Social Good Program for facilitating this research, Timothy Bishop, Helen Buck and domain experts from Insecurity Insight for coordinating and performing the data labeling process, and the anonymous reviewers for their constructive comments and suggestions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mistral AI. 2024. Au large: Mistral large. Accessed: 2024-06-13.

AI@Meta. 2024. Llama 3 model card.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*, volume 15, pages 923–932.

Chidubem Arachie, Manas Gaur, Sam Anzaroot, William Groves, Ke Zhang, and Alejandro Jaimes. 2020. Unsupervised detection of sub-events in large scale disasters. In *Proceedings Of The AAAI conference on artificial intelligence*, volume 34, pages 354–361.

Tara Brian and Frank Laczko. 2014. Migrant deaths: An international overview. *Fatal Journeys*, page 15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Josh Cowls, Andreas Tsamados, Mariarosaria Taddeo, and Luciano Floridi. 2021. A definition, benchmark and database of ai for social good initiatives. *Nature Machine Intelligence*, 3(2):111–115.

Databricks. 2023. Introducing dbrx: A new state-of-the-art open llm. Accessed: 2024-06-13.

Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, 147:113193.

Alexandra Delucia, Mark Dredze, and Anna L Buczak. 2023. A multi-instance learning approach to civil unrest event detection on twitter. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 18–33.

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

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Karsten Donnay, Eric T. Dunford, Erin C. McGrath, David Backer, and David E. Cunningham. 2019. Integrating conflict event data. *Journal of Conflict Resolution*, 63(5):1337–1364.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *arXiv preprint arXiv:2010.02194*.

Kevin Duarte, Yogesh Rawat, and Mubarak Shah. 2021. Plm: Partial label masking for imbalanced multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2739–2748.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. *arXiv preprint arXiv:2106.02124*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sara Evensen, Yoshihiko Suhara, Alon Halevy, Vivian Li, Wang-Chiew Tan, and Saran Mumick. 2019. Happiness entailment: Automating suggestions for well-being. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 62–68. IEEE.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. Crisisltlsum: A benchmark for local crisis event timeline extraction and summarization. *arXiv preprint arXiv:2210.14190*.

Mateusz Fedoryszak, Brent Frederick, Vijay Rajaram, and Changtao Zhong. 2019. Real-time event detection on social data streams. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2774–2782.

Selim Fekih, Nicolò Tamagnone, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, and Navid Rekabsaz. 2022. Humset: Dataset of multilingual information extraction and classification for humanitarian crisis response. *arXiv preprint arXiv:2210.04573*.

S Fund. 2015. Sustainable development goals. *Available at this link: https://www. un. org/sustainabledevelopment/inequality*.

Sylvia Garry and Francesco Checchi. 2020. Armed conflict and public health: into the 21st century. *Journal of Public Health*, 42(3):e287–e298.

Scott Gates, Håvard Hegre, Håvard Mokleiv Nygård, and Håvard Strand. 2012. Development consequences of armed conflict. *World Development*, 40(9):1713–1722.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. 2020. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191.

Fernando Gonzalez, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. Beyond good intentions: Reporting the research landscape of nlp for social good. *arXiv preprint arXiv:2305.05471*.

GRANIT. 2024. Multi-hazard surveillance tool for west african coastal countries: Benin - côte d'ivoire - ghana - togo (granit). Technical report, UN Office for the Coordination of Humanitarian Affairs, International Organization for Migration, REACH.

Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina von Uexkull, and Jonas Vestby. 2019. Views: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174.

Sameer Jain, Sedrick Scott Keh, Shova Chhetri, Karun Dewan, Pablo Izquierdo, Johanna Prussmann, Pooja Shrestha, César Suárez, Zheyuan Ryan Shi, Lei Li, et al. 2024. Where it really matters: Few-shot environmental conservation media monitoring for low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22141–22149.

Horacio Jesús Jarquín-Vásquez, Delia Irazú Hernández Farías, Luis Joaquín Arellano, Hugo Jair Escalante, Luis Villaseñor Pineda, Manuel Montes-y-Gómez, and Fernando Sánchez-Vega. 2023. Overview of DA-VINCIS at iberlef 2023: Detection of aggressive and violent incidents from social media in spanish. *Proces. del Leng. Natural*, 71:351–360.

Ayesha Kadir, Sherry Shenoda, and Jeffrey Goldhagen. 2019. Effects of armed conflict on child health and development: a systematic review. *PloS one*, 14(1):e0210071.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Sedrick Scott Keh, Zheyuan Ryan Shi, David J Patterson, Nirmal Bhagabati, Karun Dewan, Areendran Gopala, Pablo Izquierdo, Debojyoti Mallick, Ambika Sharma, Pooja Shrestha, et al. 2023. Newspanda: Media monitoring for timely conservation action. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15528–15536.

Tino Kreutzer, Patrick Vinck, Phuong N Pham, Aijun An, Lora Appel, Eric DeLuca, Grace Tang, Muath Alzghool, Kusum Hachhethu, Bobi Morris, et al. 2019. Improving humanitarian needs assessments through natural language processing. *IBM Journal of Research and Development*, 64(1/2):9–1.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Semi-Supervised Learning. 2006. Semi-supervised learning. *CSZ2006. html*, 5.

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

Matt Lisivick. 2018. Newsapi python library.

Junhua Liu, Trisha Singhal, Lucienne TM Blessing, Kristin L Wood, and Kwan Hui Lim. 2021. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM conference on hypertext and social media*, pages 133–141.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Nitin Madnani and Aoife Cahill. 2018. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109.

Charles P Martin-Shields and Wolfgang Stojetz. 2019. Food security and conflict: Empirical challenges and future opportunities for research and policy making on food security and conflict. *World development*, 119:150–164.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

OpenAI. 2024. Hello gpt-4o. Accessed: 2024-06-13.

Nikolaos Panagiotou, Ioannis Katakis, and Dimitrios Gunopulos. 2016. Detecting events in online social networks: Definitions, trends and challenges. *Solving Large Scale Learning Tasks. Challenges and Algorithms: Essays Dedicated to Katharina Morik on the Occasion of Her 60th Birthday*, pages 42–84.

Jorge Parraga-Alava, Roberth Alcivar-Cevallos, Leticia Vaca-Cardenas, and Jaime Meza. 2021. Urbangency: An emergency events dataset based on citizen sensors for monitoring urban scenarios in ecuador. *Data in Brief*, 34:106693.

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for nlp. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024.

Steve Penson, Mathijs Lomme, Zacharey Carmichael, Alemu Manni, Sudeep Shrestha, and Bo Pieter Johannes Andrée. 2024. A data-driven approach for early detection of food insecurity in yemen's humanitarian crisis. Technical report, The World Bank.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935.

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.

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.

Shihao Ran, Di Lu, Joel Tetreault, Aoife Cahill, and Alejandro Jaimes. 2023. A new task and dataset on detecting attacks on human rights defenders. *arXiv preprint arXiv:2306.17695*.

Roberta Rocca, Nicolò Tamagnone, Selim Fekih, Ximena Contla, and Navid Rekabsaz. 2023. Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian nlp. *Frontiers in big Data*, 6:1082787.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Idean Salehyan, Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. Social conflict in africa: A new database. *International Interactions*, 38(4):503–511.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

H Scudder. 1965. Adaptive communication receivers. *IEEE Transactions on Information Theory*, 11(2):167–174.

Burr Settles. 2009. Active learning literature survey.

Zheyuan Ryan Shi, Claire Wang, and Fei Fang. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*.

Ralph Sundberg and Erik Melander. 2013. Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Rahul Sur, Rakib Hossain, Nabina Panday, Peter Dixon, and Fazliddin Samandarov. 2019. Evaluation of the effectiveness of human rights monitoring, reporting and follow-up in the United Nations multidimensional peacekeeping operations. Program Evaluation IED-19-004, OHCHR, DPO, DMSPC.

Anusua Trivedi, Kate Keator, Avirishu Verma, Rahul Dodhia, and Juan Lavista Ferres. 2020. Eco: Using ai for everyday armed conflict analysis. In *AI for Good Workshop, co-located with IJCAI 2020*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Diyi Yang, Ankur Parikh, and Colin Raffel. 2022. Learning with limited text data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 28–31.

Siqi Yang, Wenwu Zhao, Yanxu Liu, Francesco Cherubini, Bojie Fu, and Paulo Pereira. 2020. Prioritizing sustainable development goals and linking them to ecosystem services: A global expert's knowledge evaluation. *Geography and Sustainability*, 1(4):321–330.

Jenny Paola Yela-Bello, Ewan Oglethorpe, and Navid Rekabsaz. 2021. Multihumes: Multilingual humanitarian dataset for extractive summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1713–1717.

A Data Collection Details

A.1 NewsAPI

News articles are collected from NewsAPI by combining and querying a list of keywords on a daily basis. Keyword examples are listed in Table 11.

Nouns	Verbs
aid worker	abducted
cattle	assaulted
doctor	attacked
doctor	bombed
farm	burned
grain	damaged
health facility	destroyed
market	injured
mobile clinic	killed
NGO staff member	looted
refugee	robbed
school	shelled
shepherd	threatened

Table 11: NewsAPI query keywords for identifying violent incidents. A full list is available here

A.2 GDELT

We collect all articles between December 1, 2023 and April 15, 2024 from the GDELT Events Database (version 2) using the GDELT Python library. We identify the countries associated with an article considering the following GDELT country codes: *Actor1CountryCode, Actor2CountryCode, Actor1Geo_CountryCode, Actor2Geo_CountryCode, ActionGeo_CountryCode*. An article's language is identified by scraping the article content and applying the ELD Python library. The complete GDELT data collection script is shared here.

A.3 Unlabeled Dataset

Along with HUMVI-expansion, we also release a large collection of unlabeled data. We believe that unlabeled data can be used to improve the model's performance through active learning (Settles, 2009) or through semi-supervised learning (Learning, 2006).

B Annotation Guidance

All annotators who participated in both the on the job labeling (**OTJ**) and offline labeling (**OFL**) were hired and trained directly by Insecurity Insight. Tables 12, 13, and 14 show the extended category descriptions that are used to train the annotators to perform the labeling task.

C Modeling Details

C.1 Model Hyperparameters

For BERT, DistilBERT and RoBERTa we leverage the transformers library and use the same training setup: the text is preprocessed with max 512 tokens. For training, we use the AdamW optimizer with $1 \cdot 10^{-5}$ learning rate, 16 batch size, 10 epochs, and 0.01 weight decay. All models were trained with the `transformers` library (Wolf et al., 2019) on the `NVIDIA A10G` machine. Each run lasted 30 minutes to 3 hours. Multiple training runs were conducted, and we report the average performance. We detail the size of each of the models in Table 15. We use the following checkpoints:

- HUMVI-core: bert-base-cased, distilbert-base-cased and roberta-base.
- HUMVI-expansion: bert-base-multilingual-cased, distilbert-base-multilingual-cased and xlm-roberta-base.

C.2 LLM Prompting

The prompt used in all LLM experiments is available here. We experimented with multiple different prompts, and selected the final prompt due to its higher performance on a small subset and its comprehensiveness in capturing the guidelines given by Insecurity Insight to their experts.

The size of the LLMs could be variable, and is directly related to their performance. We provide number of parameters for each open model in Table 15.

Prompt Limitations One of the drawbacks in our experimentation is the prompt used for LLMs. We note that in aligning the LLM prompt with the guidelines set by Insecurity Insight for their human annotators, we might have overburdened the model with context that might be distracting and might have led to sub-par zero-shot performance. We believe distilling the prompt to a smaller length, while maintaining essential guidelines could help in improving the performance. A possible direction here is to summarize the prompt, and then use it. Another possible direction is to provide examples in the prompt and leverage use few-shot learning instead of zero-shot.

<table>
<thead>
<tr><th>Category</th><th>Description</th></tr>
</thead>
<tbody>
<tr><td>Aid Security</td><td>Harms or threats towards aid agencies or old workers. Such as:

An aid worker or staff member being killed, wounded, kidnapped, arrested, threatened, sexually harassed, discriminated, tortured, expelled, robbed or denied passage;
and/or property, vehicle, cash, equipment, or supplies belonging to an aid agency is attacked, denied access, set on fire, taken over, invaded, broken into, affected by a demonstration, have weapons installed in it, used to launch attacks, robbed of cash, equipment, or information
and/or when aid agencies are forced to close or temporarily suspend operations, denied access, accused and/or investigated and/or fined, faced with demonstrations or rioting, denied a visa, deprived of utilities, faced with diversions of aid, have staff expelled, operate in a context of active fighting instability, affected by new laws or indirect government action, affected by mining, affected by strikes, be threatened;

which may result in the aid agency to proactively or reactively increase vigilance, change security measures, restrict movement, relocate staff, move assets, close, hibernate, or operate remotely.</td></tr>
<tr><td>Education</td><td>Harms or threats towards education providers, education workers, education-related infrastructure, and students. Such as:

An educator was killed, wounded, kidnapped, arrested, threatened, sexually harassed, discriminated, tortured, expelled, robbed or denied passage;
and/or an education facility (school, university, any building serving to provide education services) is attacked, Set on fire, taken over, invaded, broken into, affected by a demonstration, have weapons installed in it, used to launch attacks; robbed of cash, equipment, or information
or when an education provider is forced to close or temporarily suspend operations, denied access, accused and/or investigated and/or fined, faced with demonstrations or rioting, denied a visa, deprived of utilities, faced with diversions of aid, have staff expelled, operate in a context of active fighting instability, affected by new laws or indirect government action, affected by mining, affected by strikes, threatened

which may result in the education provider to proactively or reactively increase vigilance, change security measures, restrict movement, relocate staff, move assets, close, hibernate, or operate remotely.

The following events are excluded:

Events affecting retired educators, including teachers, academics, education support and transport staff.
Threats or violence towards students in public/open spaces where it is clear they were not en route to or from school. E.g. kidnapped while at a market on a Saturday.
Demonstrations involving students in open spaces, such as town halls and streets.
</td></tr>
</tbody>
</table>

Table 12: Extended description of categories

Category	Description
Health	Harm or threats towards health providers or health workers. Such as: • A health worker (doctor, nurse, ambulance driver, paramedic, military medic, physiotherapist, vaccination worker, or any other health staff member) was killed, wounded, kidnapped, arrested, threatened, sexually harassed, discriminated, tortured, expelled, robbed or denied passage; • and/or Hospital, health centre, mobile health unit, pharmacy, ambulance, or other building/vehicle used for health purposes, or cash, equipment, or supplies belonging to a health provider is, attacked, denied access, set on fire, taken over, invaded, broken into, affected by a demonstration, have weapons installed in it, used to launch attacks; robbed of cash, equipment, or information • or when a health provider is forced to close or temporarily suspend operations, denied access, accused and/or investigated and/or fined, faced with demonstrations or rioting, denied a visa, deprived of utilities, faced with diversions of aid, have staff expelled, operate in a context of active fighting instability, is affected by new laws or indirect government action, is affected by mining, is affected by strikes, is threatened which may result in the the health provider to proactively or reactively increase vigilance, change security measures, restrict movement, relocate staff, move assets, close, hibernate, or operate remotely.

Protection	Harm or threats towards internally displaced persons and refugees.
	
	• Incidents where camp settlements (both informal and formal) and its infrastructure (incl. Education and health facilities inside the camp) are affected and/or residents inside these camps are affected.
	  – An IDP/displaced/refugee camp was attacked, denied access, set on fire, taken over, invaded, broken into, affected by a demonstration, have weapons installed in it or armed entry into the camps, used to launch attacks, robbed of cash, equipment, or information/ looted, raided by security/military forces, forced to close or be dismantled, affected by close proximity to conflict/instability
	  – An IDP/displaced/refugee camp resident was killed, wounded, kidnapped, arrested, threatened, sexually harassed, discriminated, tortured, expelled, robbed or denied passage; forced out of camps
	
	• Events which include a disruption of aid (incl. supplies/food/security) to the camp and its residents. This includes obstruction of aid; the forced closure or dismantling of camps; inability for aid organizations to provide aid to the camps etc.
	  – and/or when aid agencies: are forced to close or temporarily suspend operations, denied access, accused and/or investigated and/or fined faced with demonstrations or rioting, denied a visa, deprived of utilities, faced with diversions of aid, have staff expelled, operate in a context of active fighting instability, are affected by new laws or indirect government action, are affected by mining, are affected by strikes, are threatened.
	
	The following events are *excluded*:
	
	• Events where IDP/refugees/displaced persons are affected outside of the camp structure (e.g. shipwrecks; deaths on the migration routes).
	
	• administrative and government actions which are aimed towards curbing migration (e.g. mass deportations; changes in government policies etc..)

Table 13: Extended description of categories (continued)

Category	Description
Food Security	Harm or threats towards objects and actors directly involved in the production, processing and distribution of food; damaging or destroying transport and energy infrastructure which facilitates the food supply and distribution. • Includes events affecting objects or actors directly involved in the production, processing and distribution of food: – The setting on fire, striking with explosive and other weapons (e.g. firearms) and damage or destruction resulting from other conflict actions to crops and grazing land, farms, food or livestock markets, food production or processing factories, food storage warehouses, granaries, farm buildings (e.g. barns and wheat storage silos, bakeries and other food shops, tractors and combine harvesters, water infrastructure (e.g. pumps, wells, pipelines and transport tanks) – Looting or robbery of crops, food (including food aid), tractors, combine harvesters, agricultural equipment (e.g. ploughs, irrigators) or livestock – Abductions or kidnappings, arrests and killings or injuries of: farmers and other farm workers, pastoralists or herders, fishers, food aid workers, food production workers (e.g. individuals who work in food production factories) or food distribution workers (e.g. drivers of food transport trucks) – Physically violent clashes between farmers and pastoralists (including events which are both lethal and non-lethal). – Physical violence occurring at food or livestock markets (e.g. shootings using firearms, detonations of IEDs) which does not necessarily result in damage or destruction of the market itself. – Taking over, forced entry into, denial of access or invasion of farms and farm buildings, food production or processing factories, bakeries and other food shops; or food storage warehouses • Includes events which affect food security in a more indirect way by damaging or destroying transport and energy infrastructure which facilitates the continuation of food supplies and distributions. – The setting on fire, striking with explosive weapons and conflict actions resulting in the damage or destruction of bridges, roads, ports, airports, rail lines; gas, oil and electricity pipelines and plants; or stations The following events are *excluded*: • protests (violent and non-violent) voicing concerns about food insecurity; non-violent political or public policies pursued by governments in relation to food insecurity (e.g. sanctions imposed on individuals, governments or private corporations); attacks on civilians except for those involved in the production of food or its distribution (e.g. farmers, pastoralists, food production factory workers).

Table 14: Extended description of categories (continued)

D Error Analysis

D.1 Misclassification Themes

We identified two main classes of inputs for which most of the models did not perform well.

Missing Context. In some cases, the *victim* entity is directly related to the humanitarian aid response category and it is hard for models to understand the context. For instance, in Example D.1.1 , the victims are employees of World Central Kitchen, an organization that provides meals to those affected by conflict. However, the models categorized them merely as aid workers, overlooking the crucial context that targeting workers of a meal-providing organization directly impacts food security.

Indirect Impact. For certain examples, the model focuses on the first category and ignores the rest of the text that could be detrimental for predicting other different categories that could be added for direct or indirect impact. An example is shown in Example D.1.2. In the example, since it was an attack on UNICEF supplies, the model was able to identify the context of aid workers or their aid being targeted, but not the indirect impact of that aid being blocked which is lack of crucial health supplies and food supplies.

D.1.1 Example of Missing Context Misclassification

News Article: US Secretary of State Antony Blinken said on Tuesday that Washington has urged Israel to conduct a swift, thorough and impartial investigation into Monday night's air strike that killed seven aid workers with the World Central Kitchen charity in Gaza. ..... he said of the NGO workers killed in the strike. .... over the accidental deaths of seven World Central Kitchen (WCK) employees in Monday's strike. .... take immediate steps to protect aid workers and facilitate vital humanitarian operations in Gaza....
Ground Truth Label: food-sec, aid-sec
Predicted Label: aid-sec

Architecture	Num Params
BERT	110M
DistilBERT	66M
RoBERTa	110M
XLM-RoBERTa	125M
DBRX	132B
llama-3-70B	70B
Mistral-Large	Proprietary
GPT-4-Turbo	Proprietary
GPT-4o	Proprietary

Table 15: Model Size (by number of parameters) of the models we used in experiments.

D.1.2 Example of Indirect Impact Misclassification

News Article: UNICEF said one of their containers carrying essential supplies was looted by gangs at Haiti's main port. .... The United Nations Children's Fund (UNICEF) said Saturday that one of its 17 aid containers at Haiti's main port was looted. .... The container was carrying "essential items for maternal, neonatal, and child survival, as well as critical supplies for early childhood development and education, water equipment, and others," the agency said. ..... "Looting of supplies that are essential for life saving support for children must end immediately," .... Gang violence has spiked throughout the country in recent days. .... Some hospitals in the city have been forced to close over safety concerns.... Shortages of electricity, fuel and medical supplies have affected hospitals .... is supported by the Caribbean regional body CARICOM, the United Nations and the United States.
Ground Truth Label: food-sec, aid-sec, health
Predicted Label: aid-sec