**Abstract**

Timely and effective response to humanitarian crises requires quick and accurate analysis of large amounts of text data—a process that can highly benefit from expert-assisted NLP systems trained on validated and annotated data in the humanitarian response domain. To enable creation of such NLP systems, we introduce and release HUMSET, a novel and rich multilingual dataset of humanitarian response documents annotated by experts in the humanitarian response community. The dataset provides documents in three languages (English, French, Spanish) and covers a variety of humanitarian crises from 2018 to 2021 across the globe. For each document, HUMSET provides selected snippets (entries) as well as assigned classes to each entry annotated using common humanitarian information analysis frameworks. HUMSET also provides novel and challenging entry extraction and multi-label entry classification tasks. In this paper, we take a first step towards approaching these tasks and conduct a set of experiments on Pre-trained Language Models (PLM) to establish strong baselines for future research in this domain. The dataset is available at [https://blog.thedeep.io/humset/](https://blog.thedeep.io/humset/).

1 Introduction

During humanitarian crises caused by reasons ranging from natural disasters, wars, or epidemics such as COVID-19, a timely and effective humanitarian response highly depends on fast and accurate analysis of relevant data to yield key information. Early in the response phase, namely in the first 72 hours after a disaster strikes, the humanitarian response analysts in international organizations review large amounts of data loosely or strongly relevant to the crisis to gain situational awareness.

---

**HUMSET: Dataset of Multilingual Information Extraction and Classification for Humanitarian Crisis Response**

Selim Fekih¹, Nicolò Tamagnone¹, Benjamin Minixhofer³, Ranjan Shrestha², Ximena Contla¹, Ewan Oglethorpe¹, Navid Rekabsaz³

¹Data Friendly Space ²ToggleCorp Solutions ³Johannes Kepler University Linz, LIT AI Lab, Austria

{selim, nico, ximena, ewan}@datafriendlyspace.org  
ranjan.shrestha@togglecorp.com  
{benjamin.minixhofer, navid.rekabsaz}@jku.at

A large portion of this data appears in the form of secondary data sources i.e. reports, news, and other forms of text data, and is integral in revealing which type of relief activities to undertake. Analysis in this phase involves extracting key information and organizing it according to sets of pre-defined domain-specific structures and guidelines, referred to as humanitarian analysis frameworks.

While typically only small workforces are available to analyze such information, an automatic document processing system can significantly help analysts save time in the overall humanitarian response cycle. To facilitate such systems, we introduce and release HUMSET, a unique and rich dataset of document analysis in the humanitarian response domain. HUMSET is curated by humanitarian analysts and covers various disasters around the globe that occurred from 2018 to 2021 in 46 humanitarian response projects. The dataset consists of approximately 17K annotated documents in three languages of English, French, and Spanish, originally taken from publicly-available resources. For each document, analysts have identified informative snippets (entries) with respect to common humanitarian frameworks and assigned one or many classes to each entry (details in §2).

HUMSET provides a large dataset for the training and evaluation of entry extraction and classification models, enabling the research and development of further NLP systems in the humanitarian response domain. We take the first step in this direction, by studying the performance of a set of strong baseline models (details in §3). Our released dataset expands the previously provided collection by Yela-Bello et al. (2021) with a more recent and comprehensive set of projects, as well as additional classification labels. Other similar datasets in the humanitarian domain, Imran et al. (2016) present human-annotated Twitter corpora collected during 19 dif-
ferent crises between 2013 and 2015, Alam et al. (2021) provide a combination of various social-media crisis-related existing datasets, and Adel and Wang (2020) and later Alharbi and Lee (2021) publish Arabic Twitter classification datasets for crisis events. HUMSET, in contrast to the current resources which mostly originated from social media, is created by humanitarian experts through an annotation process on official documents and news from the most recognized humanitarian agencies, conferring high reliability, continuous updating, and accurate geolocation information.

2 HUMSET Dataset

The collection originated from a multi-organizational platform called the Data Entry and Exploration Platform (DEEP), developed and maintained by Data Friendly Space (DFS). The platform facilitates classifying primarily qualitative information with respect to analysis frameworks and allows for collaborative classification and annotation of secondary data. The dataset is available at https://blog.thedeep.io/humset/.

2.1 Dataset Overview

HUMSET consists of data used to inform 46 humanitarian response operations across the globe. 24 responses were in Central/South America, 14 in Africa, and 8 in Asia (detailed countries can be found in Table 6 in Appendix).

For each project, documents, referred to as leads, related to a particular humanitarian crisis are collected, analyzed, and annotated. The annotated documents in the dataset mostly consist of recently released information, with 79% of the documents being released in 2020 and 2021 (Table 5 in Appendix), and 90% of all documents being sourced from websites (see Table 4 in Appendix for the most commonly used platforms). Documents are selected from different sources, ranging from official reports by humanitarian organizations to international and national media articles. Overall, documents consist of files in PDF format (70.4%) and HTML pages (29.6%) with an average length of ~2K words. The number of documents analyzed per project varies, ranging from 2 to 2,266.

The relevant snippets of texts, referred to as entries, in each document are annotated by humanitarian experts. The dataset provides an average of ~10 entries per document, and an average length of ~65 words per entry. Overall, HUMSET is composed of 148,621 tagged entries, selected from 16,857 documents, and in three languages: English (61.3%), French (20.4%) and Spanish (18.3%). The list of projects as well as the number of documents and annotated entries per project is reported in Table 7 in Appendix. Figure 1 shows the distribution of the number of tagged documents per project, as well as the number of tokens per document and entry.

2.2 Humanitarian Analysis Frameworks and Data Annotation Process

The concept of analytical frameworks originated in the social sciences (Ragin and Amoroso, 2011), but can be considered foundational and indispensable in numerous research fields. An analytical framework is a set of methodologies and guidelines to facilitate data collection, collation, and analysis, helping to understand what information will be useful and what can be discarded.

In the humanitarian domain, an analytical framework (or analysis framework) not only assists decision-makers to speed up humanitarian response.

Figure 1: (a) Distribution of documents per project. (b) Log-scale distribution of tokens per document. (c) Log-scale distribution of tokens per excerpt.
Table 1: Overview of humanitarian analysis framework.

| Categories   | #   | Tags                                                                 |
|--------------|-----|----------------------------------------------------------------------|
| Sectors      | 11  | Agriculture, Cross-sector, Education, Food Security, Health, Livelihoods, Logistics, Nutrition, Protection, Shelter, WASH (Water, Sanitation & Hygiene) |
| Pillars 1D   | 7   | Context, COVID-19, Displacement, Humanitarian Access, Information & Communication, Casualties, Shock/Event |
| Subpillars 1D| 33  | Details in Table 8 in Appendix                                        |
| Pillars 2D   | 6   | Capacities & Response, Humanitarian Conditions, Impact, At Risk, Priority Needs, Priority Interventions |
| Subpillars 2D| 18  | Details in Table 9 in Appendix                                        |

For each project, documents relevant to understanding the situation, unmet needs, and underlying factors are captured and uploaded to the DEEP platform. From these sources, entries of text are selected and categorized into an analysis framework. Humanitarian annotators are trained in specific projects to follow analytical standards and thinking to review secondary data.

This process eventually results in annotating and organizing the data according to the humanitarian analysis framework. As the HUMSET dataset is created in a real-world scenario, the distribution of annotated entries is skewed, with 33 tags being present in less than 2% of data.

Tables 10, 11, and 12 in Appendix show the detailed number and proportions of the annotated entries in Sectors, Subpillars 1D, and 2D, respectively. Figure 2 in Appendix reports the distribution of tags in dataset.

2.3 NLP Tasks

Entry Extraction Task. The first step for humanitarian taggers in analyzing a document is finding entries containing relevant information. A piece of text or information is considered relevant if it meaningfully contains at least one tag present in the given humanitarian analytical framework. Since documents often contain a large amount of information (Figure 1), it is extremely beneficial to automate the process of entry identification, and this is the first task of this research. This can be seen as an extractive summarization task i.e. selecting a subset of passages that contain relevant information from the given document. However, the entries do not necessarily follow the common units of text such as sentence and paragraph and can appear in various lengths. In fact, only 38.8% of entries consist of full sentences, and the rest are snippets that are shorter or longer than sentences. This limits the direct applicability of prior approaches to extractive summarization (Liu and Lapata, 2019; Zhou et al., 2018), and makes the task particularly challenging for NLP research.

Multi-label Entry Classification Task. After selecting the most relevant entries within a document, the next step is to categorize them according to the humanitarian analysis framework (Table 1). An automatic suggestion on which tag to choose from a large number of possibilities can be decisive in speeding up the annotation process. For each category, more than one tag can be assigned to an entry. Hence, we can view this task as multi-label classification.

3 Experiments and Results

To conduct a set of baseline experiments on HUMSET according to the mentioned tasks, we split the data into training, validation, and test sets for all our experiments (80%, 10%, and 10%, respectively). We apply stratified splitting (Szymanski and Kajdanowicz, 2017) to maintain the same distribution of labels for each set. Implementation details of Entry Extraction (Section 3.1) and Entry Classification (Section 3.2) are available at https://github.com/the-deep/humset.
3.1 Entry Extraction

We evaluate the performance of the entry extraction task using ROUGE-1, ROUGE-2, and ROUGE-L F1 score (Lin, 2004). The target text (ground truth) is a concatenation of all relevant entries, and the predicted text is a concatenation of all entries predicted as relevant. We consider a simple heuristic method (LEAD4), as well as Transformer-based (Vaswani et al., 2017) pre-trained language models (PLM) with a multilingual backbone as our baselines as explained in the following:

**LEAD4**: LEAD-n is a simple baseline where the first n sentences are predicted as being relevant entries. Consistent with prior work (Yela-Bello et al., 2021), we choose n = 4. **Transformers**: to approach the task using Transformer-based PLM, we formulate the task as a token classification problem. The objective is to distinguish between tokens that are part of relevant entries and tokens which are not. For simplicity, we fine-tune the entire model and do binary classification using a two-layer prediction head on top of the contextualized representation of each token. We conduct our experiments on XtremeDistil_{l6−h256} (Mukherjee and Hassan Awadallah, 2020) and XLM-R_{Base} (Conneau et al., 2019) as the underlying PLM.

The evaluation results of the mentioned methods are reported in Table 3. Among our baselines, the model with XLM-R_{Base} shows the best overall performance. However, we should consider these experiments as starting points, and improvements on this task are expected by increasing model capacity and architectural variations.

### Table 3: Entry extraction results for ROUGE F1 (R).

| Model           | R-1 | R-2 | R-L |
|-----------------|-----|-----|-----|
| LEAD4           | 0.32| 0.24| 0.31|
| XtremeDistil_{l6−h256} | 0.33| 0.25| 0.33|
| XLM-R_{Base}    | **0.42** | **0.35** | **0.41** |

3.2 Entry Classification

We test different multi-label sequence classification models applied to our five categories. We use the Precision and F1-score metrics to assess the performance of the models on each subcategory. We report macro-averages of the metrics, as the tags are unbalanced (see Table 2) and macro-averaging can provide a more nuanced view of the performance especially by supporting the more sparse classes. Finally, we perform threshold tuning of the classification decision boundary with respect to macro-average F1-scores for each label of each category (Pillai et al., 2013). Tuning the threshold is done by finding the optimal results on the validation set, used to make classifications on the test set. We conduct experiments using fastText (Joulin et al., 2016), as well as Transformer-based PLMs as explained below.

**fastText**: is an Open Source library for text representation and classification that consists of a bag of n-grams representation and a linear classifier. fastText classification is language-agnostic and does not need language-specific pre-trained word vectors, allowing us to train a multilingual classifier as a simple baseline. To handle multiple labels, we trained independent binary classifiers for each label. **Transformers**: For consistency with the previous task, we fine-tune the same multilingual PLMs and add a dense layer on top for multi-label classification.

Table 2 reports the evaluation results on the mentioned baseline models. For comparison, a random baseline is also reported. The random baseline is a stratified random classification, created based on the distribution of the classes in the training set. Similar to the entry extraction task, the XLM-R_{Base} outperforms other baselines. Although overall promising results are obtained, we highlight the shortcoming of the models on the categories with many tags (Subpillars 1D and Subpillars 2D), suggesting future research directions for addressing these challenges.
4 Conclusion

We presented HUMSET, a new dataset of annotated humanitarian data, containing 148,621 entries with a total of 62 different tags. We have shown two NLP-based tasks that can be applied to it, providing initial experiments of its applications. HUMSET is a multilingual human-annotated humanitarian text dataset not composed of social network data, providing a valuable and highly reliable resource for the development of automation tools regarding crisis response and humanitarian aid activities.

5 Limitations

HUMSET is composed of an aggregation of 46 different projects, each with a different contribution in terms of data quantity and topics (Table 7). This can introduce an implicit bias due to the different goals and themes of each project and on respective analysis framework understanding and interpretation by humanitarian annotators. (Röttger et al., 2021) refer to it as persistent subjectivity. This is a complex and challenging limitation and, for example, (Geva et al., 2019) show how this kind of bias can be monitored using annotator identifiers as features in NLP models training when data is produced by crowdsourcing project (Sheng and Zhang, 2019). Since HUMSET is an extension of a real-case application and not the result of crowdsourcing, a more structured analysis on these aspects is needed.

Another complexity lies in the raw data sources. Lead text is the result of a text extraction process from PDF and HTML files (c.f. Section 2.1). In both cases, converting visually-rich graphical text representation into plain text involves errors and limitations. There are several works proposing solutions for digital documents layout-aware text extraction (Ramakrishnan et al., 2012; Zhu and Cole, 2022) but they are often domain-specific, applying only to specific types of documents. (Xu et al., 2020) propose a Transformer-based multi-modal architecture for documents understanding using text, layout, and image data as features. Improvement in document processing could produce better data quality and subsequently improve performance on the entry extraction task (Section 3.1).

Finally, we should point out that HUMSET might contain societal biases and stereotypes and/or overrepresent particular demographics or entities. This case is observed and studied in several other data resources and scenarios (Bolukbasi et al., 2016; Krieg et al., 2022a; Rekabsaz et al., 2021b), which can lead to reflecting or even exaggerating societal biases in the system’s output (Melchiorre et al., 2021; Rekabsaz and Schedl, 2020), and may negatively affect users’ perception and interaction behavior (Krieg et al., 2022b). Hence, when using the dataset (particularly for real-world applications), we strongly recommend first defining and monitoring such potential biases (De-Arteaga et al., 2019; Rekabsaz et al., 2021c), and then mitigating them using the proposed methods in literature (Elazar and Goldberg, 2018; Zmigrod et al., 2019; Rekabsaz et al., 2021a; Zerveas et al., 2022; Ganhör et al., 2022).

6 Acknowledgement

We want to thank the humanitarian community users of the Data Entry and Exploration Platform (DEEP) for their openness and interest in sharing their data for this research and to trust that the NLP community can help them make their work better. We want to extend our gratitude, especially to the people working in the USAID’s Bureau for Humanitarian Assistance (BHA) for entrusting Data Friendly Space (DFS) with a grant to create the DEEP and to work on this paper; to the DEEP users and taggers for their work that make this set possible; to the project owners in DEEP that allowed us to use the data to create this set and their organizations; to the DEEP board members Internal Displacement Monitoring Centre (IDMC), International Federation of the Red Cross, iMMAP, Office of the High Commissioner for Human Rights, Okular Analytics, United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA), United Nations High Commissioner for Refugees (UNHCR), United Nations Children’s Fund (UNICEF), United Nations Development Coordination Office (UNDGO), and the Danish Refugee Council (DRC). To the whole DFS team around the world, the ToggleCorp team, our partner institutions ISI Foundation and Johannes Kepler University Linz for their continuous support in making the usage of NLP possible in the humanitarian community.
References

Ghadah Adel and Yuping Wang. 2020. Detecting and classifying humanitarian crisis in arabic tweets. In 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), pages 269–274.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Olli. 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.

Alaa Alharbi and Mark Lee. 2021. KawaiiTh: an Arabic Twitter corpus for crisis events. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 42–52, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in Neural Information Processing Systems.

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.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 120–128.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 11–21.

Christian Ganhörr, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. Mitigating consumer biases in recommendations with adversarial training. In Proceedings of the 45th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2022. ACM.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. arXiv preprint arXiv:1605.05894.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2022a. Grep-biasir: A dataset for investigating gender representation-bias in information retrieval results. arXiv preprint arXiv:2201.07754.

Klara Krieg, Emilia Parada-Cabaleiro, Markus Schedl, and Navid Rekabsaz. 2022b. Do perceived gender biases in retrieval results affect relevance judgements? In Proceedings of the European Conference on Information Retrieval, Workshop on Algorithmic Bias in Search and Recommendation (ECIR-BIAS 2022).

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. Information Processing Management (IP&M), 58(5):102666.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. XtremeDistil: Multi-stage distillation for massive multilingual models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2221–2234, Online. Association for Computational Linguistics.

Ignazio Pillai, Giorgio Fumera, and Fabio Roli. 2013. Threshold optimisation for multi-label classifiers. Pattern Recognition, 46(7):2055–2065.

Charles C Ragan and Lisa M Amoroso. 2011. Constructing social research: The unity and diversity of method. Pine Forge Press.

Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. 2012. Layout-aware text extraction from full-text pdf of scientific articles. Source code for biology and medicine, 7(1):1–10.
Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021a. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of BERT rankers. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 306–316.

Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021b. TripClick: The Log Files of a Large Health Web Search Engine. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2507–2513. Association for Computing Machinery, New York, NY, USA.

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2065–2068.

Navid Rekabsaz, Robert West, James Henderson, and Allan Hanbury. 2021c. Measuring societal biases in text corpora via first-order co-occurrence. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective NLP tasks. CoRR, abs/2112.07475.

Victor S Sheng and Jing Zhang. 2019. Machine learning with crowdsourcing: A brief summary of the past research and future directions. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 9837–9843.

Piotr Szymanski and Tomasz Kajdanowicz. 2017. A scikit-based python environment for performing multi-label classification. CoRR, abs/1702.01460.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740.

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, Online. Association for Computational Linguistics.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through set-based document reranking and neutrality regularization. In Proceedings of the 45th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2022. ACM.

Dongsong Zhang, Lina Zhou, and Jay F Nunamaker Jr. 2002. A knowledge management framework for the support of decision making in humanitarian assistance/disaster relief. Knowledge and Information Systems, 4(3):370–385.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Miao Zhu and Jacqueline M Cole. 2022. Pdftaextractor: A tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. Journal of Chemical Information and Modeling, 62(7):1633–1643.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.
A  Additional Statistics

| Website             | Number | Prop. (%) |
|---------------------|--------|-----------|
| reliefweb.int       | 5,669  | 33.6      |
| dhakatribune.com    | 834    | 4.9       |
| redhum.org          | 635    | 3.8       |
| humanitarianresponse.info | 612 | 3.6       |
| unb.com.bd          | 380    | 2.3       |
| **Sum**             | **8,130** | **48.2** |

Table 4: The most frequently sourced websites by number and proportion of documents.

| Year       | Number | Proportion (%) |
|------------|--------|----------------|
| Before 2018 | 68     | 0.4            |
| 2018       | 834    | 4.9            |
| 2019       | 2,639  | 15.7           |
| 2020       | 6,087  | 36.1           |
| 2021       | 7,229  | 42.9           |
| **Sum**    | **16,857** | **100.0** |

Table 5: Publishing year of documents.

| Region          | Countries                                                                 |
|-----------------|---------------------------------------------------------------------------|
| **Africa**      | Burkina Faso, Cameroon, Chad, Libya, Niger, Nigeria, DRC, Somalia, South Sudan, Sudan |
| **Asia & Middle East** | Afghanistan, Bangladesh, Lebanon, Syria, Yemen                          |
| **Central/South America** | Argentina, Aruba, Bolivia, Chile, Colombia, Costa Rica, Curacao, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Honduras, Mexico, Panama, Paraguay, Peru, The Bahamas, Trinidad and Tobago, Uruguay, Venezuela |

Table 6: Countries of projects per region.

Figure 2: Proportion of tags in dataset. This figure shows the unbalanced nature of the dataset. Each bar represents a different tag. The y-axis shows the proportion of entries that contain each tag. The horizontal line added is the 2% occurrence line. It is used to visualize the relatively high number of tags with occurrence inferior to 2%.
| Project                              | # Leads | # Entries |
|-------------------------------------|---------|-----------|
| 2020 DFS Libya                      | 354     | 1,581     |
| 2020 DFS Nigeria                    | 496     | 2,238     |
| COVID-19 Americas Region            | 76      | 607       |
| Central America - Dengue Outbreak 2019 | 38    | 218       |
| Central America: Hurricanes Eta and Iota | 48   | 363       |
| GIMAC Afghanistan                   | 248     | 7,617     |
| GIMAC Cameroon                      | 134     | 5,253     |
| GIMAC Chad                          | 269     | 4,945     |
| GIMAC Niger                         | 125     | 3,180     |
| GIMAC Somalia                       | 157     | 5,720     |
| GIMAC South Sudan                   | 208     | 7,212     |
| GIMAC Sudan                         | 99      | 3,169     |
| IMMARP/DFS Bangladesh              | 2,266   | 14,342    |
| IMMARP/DFS Burkina Faso            | 1,279   | 13,443    |
| IMMARP/DFS Colombia                 | 1,411   | 10,760    |
| IMMARP/DFS Niger                    | 1,443   | 9,620     |
| IMMARP/DFS RDC                      | 1,586   | 13,065    |
| IMMARP/DFS Syria                    | 1,736   | 12,043    |
| Lebanon Situation Analysis          | 22      | 268       |
| Libya Situation Analysis (OA)       | 681     | 2,868     |
| Nigeria Situation Analysis (OA)     | 651     | 3,496     |
| Situation Analysis Generic Libya    | 437     | 2,355     |
| Situation Analysis Generic Yemen    | 371     | 2,256     |
| The Bahamas - Hurricane Dorian - Early Recovery Assessment | 17 | 191 |
| UNHCR Argentina                     | 160     | 1,430     |
| UNHCR Aruba                         | 23      | 106       |
| UNHCR Bolivia                       | 9       | 101       |
| UNHCR Chile                         | 341     | 2,415     |
| UNHCR Colombia                      | 649     | 6,349     |
| UNHCR Costa Rica                    | 85      | 603       |
| UNHCR Curacao                       | 20      | 111       |
| UNHCR Dominican Republic            | 66      | 411       |
| UNHCR Ecuador                       | 190     | 1,648     |
| UNHCR El Salvador                   | 64      | 349       |
| UNHCR Guatemala                     | 65      | 348       |
| UNHCR Guyana                        | 28      | 352       |
| UNHCR Honduras                      | 57      | 469       |
| UNHCR Mexico                        | 16      | 96        |
| UNHCR Panama                        | 35      | 295       |
| UNHCR Paraguay                      | 2       | 27        |
| UNHCR Peru                          | 247     | 2,936     |
| UNHCR Trinidad and Tobago           | 67      | 574       |
| UNHCR Uruguay                       | 29      | 272       |
| UNHCR Venezuela                     | 293     | 1,805     |
| Venezuela crisis 2019               | 176     | 714       |
| Yemen Situation Analysis (OA)       | 83      | 381       |

| Total                      | 16,857 | 148,602 |

Table 7: Key statistics per project.
| Pillar 2D | Subpillars 2D | # | Label |
|----------|---------------|---|-------|
| Capabilities & Response | International Response, Local Response, National Response, Number of People Reached / Response Gaps | 4 | |
| Humanitarian Conditions | Coping Mechanisms, Living Standards, Physical And Mental Well Being, Number of People In Need | 4 | |
| Impact | Driver/Aggravating Factors, Impact on People, Impact on Systems, Services and Networks, Number of People Affected | 4 | |
| At Risk | Risk And Vulnerabilities, Number of People at Risk | 2 | |
| Priority Needs | Expressed by Humanitarian Staff, Expressed by Population | 2 | |
| Priority Interventions | Expressed by Humanitarian Staff, Expressed by Population | 2 | |

Table 9: Pillars and subpillars 2D in humanitarian framework.

| Sectors | Data Points Number | Proportion (%) |
|---------|--------------------|----------------|
| Agriculture | 2,816 | 1.9 |
| Cross | 24,447 | 16.4 |
| Education | 9,630 | 6.5 |
| Food Security | 14,898 | 10.0 |
| Health | 32,284 | 21.7 |
| Livelihoods | 15,494 | 10.4 |
| Logistics | 2,422 | 1.6 |
| Nutrition | 5,011 | 3.4 |
| Protection | 25,986 | 17.5 |
| Shelter | 8,975 | 6.0 |
| WASH | 10,588 | 7.1 |
| Count: | 115,176 | 77.5 |

Table 10: Proportion of sectors in the dataset.

| Pillar 1D | Subpillar 1D | Data Points Number | Prop. (%) |
|----------|--------------|-------------------|-----------|
| Demography | | 3,041 | 2.0 |
| Economy | | 4,969 | 3.3 |
| Environment | | 1,124 | 0.8 |
| Legal & Policy | | 2,637 | 1.8 |
| Politics | | 1,871 | 1.3 |
| Security & Stability | | 7,615 | 5.1 |
| Sociocultural | | 1,610 | 1.1 |
| International Response, Local Response, National Response, Number of People Reached / Response Gaps | | | |
| Cases | | 5,501 | 3.7 |
| Contact Tracing | | 627 | 0.4 |
| Deaths | | 3,095 | 2.1 |
| Hospitalization & Care | | 510 | 0.3 |
| Restriction Measures | | 5,006 | 3.4 |
| Testing | | 1,621 | 1.1 |
| Vaccination | | 2,764 | 1.9 |
| Humani- | | | |
| tarian Conditions | | | |
| COPVID- | | | |
| Inten- | | | |
| tions | | | |
| Local Integration | | 1,657 | 1.1 |
| Pull Factors | | 342 | 0.2 |
| Push Factors | | 2,047 | 1.4 |
| Type / Numbers / Movements | | 8,280 | 5.6 |
| Number Of People Facing Humanitarian Con- | | 658 | 0.4 |
| straints/Humanitarian Access Gaps | | | |
| Access Constraints/Humanitarian | | | |
| Access Gaps | | | |
| Physical Constraints | | 1,483 | 1.0 |
| Population to Relief | | 204 | 0.1 |
| Relief to Population | | 863 | 0.6 |
| Humanitarian Access | | | |
| Information and Communication | | | |
| Communication Means and Preferences Information | | 150 | 0.1 |
| Challenges and Barriers Knowledge and Info Gaps (Hum) | | 100 | 0.1 |
| Knowledge and Info Gaps (Pop) | | 573 | 0.4 |
| Number Of People Facing Humanitarian Con- | | 218 | 0.1 |
| straints/Humanitarian Access Gaps | | | |
| Number Of People Facing Humanitarian Con- | | | |
| straints/Humanitarian Access Gaps | | | |
| Information and Communication | | | |
| Communication Means and Preferences Information | | | |
| Challenges and Barriers Knowledge and Info Gaps (Hum) | | | |
| Knowledge and Info Gaps (Pop) | | | |
| Overall Count: | | 993 | 0.7 |

Table 11: Proportion of each pillar and subpillar 1D in the dataset.
Table 12: Proportion of each pillar and subpillar 2D in the dataset.

| Pillar 2D | Subpillars 2D | Data Points | Number | Prop. (%) |
|-----------|--------------|-------------|--------|----------|
| Capacities & Response | International Response | 15,422 | 10.4 |
| | Local Response | 211 | 0.1 |
| | National Response | 6,596 | 4.4 |
| | Number Of People Reached / Response Gaps | 3,316 | 2.2 |
| | **Pillar 2D Count:** | **20,272** | **13.6** |
| Humanitarian Conditions | Coping Mechanisms | 4,994 | 3.4 |
| | Living Standards Number of People In Need | 25,922 | 17.4 |
| | Physical and Mental Well Being | 15,845 | 10.7 |
| | **Pillar 2D Count:** | **43,301** | **29.1** |
| Impact | Driver/Aggravating Factors | 12,029 | 8.1 |
| | Impact on People | 12,319 | 8.3 |
| | Impact on Systems, Services And Networks | 14,335 | 9.6 |
| | Number of People Affected | 2,293 | 1.5 |
| | **Pillar 2D Count:** | **33,699** | **22.7** |
| At Risk | Risk and Vulnerabilities Number of People Aa Risk | 9,625 | 6.5 |
| | **Pillar 2D Count:** | **9,625** | **6.5** |
| Priority Needs | Expressed by Humanitarian Staff | 1,664 | 1.1 |
| | Expressed by Population | 1,669 | 1.1 |
| | **Pillar 2D Count:** | **3,272** | **2.2** |
| Priority Interventions | Expressed by Humanitarian Staff | 5,575 | 3.8 |
| | Expressed by Population | 287 | 0.2 |
| | **Pillar 2D Count:** | **5,844** | **3.9** |
| | **Overall Count:** | **94,429** | **63.5** |