

Operationalizing AI for Good: Spotlight on Deployment and Integration of AI Models in Humanitarian Work

Anton Abilov, Ke Zhang, Hemank Lamba,
Elizabeth M. Olson, Joel Tetreault, Alex Jaimes

Dataminr, Inc.

{aabilov,kzhang,hlamba,elizabeth.olson,jtetreault,ajaimes}
@dataminr.com

Abstract

Publications in the AI for Good space have tended to focus on the research and model development that can support high-impact applications. However, very few AI for Good papers discuss the process of deploying and collaborating with the partner organization, and the resulting real-world impact. In this work, we share details about the close collaboration with a humanitarian-to-humanitarian (H2H) organization and how to not only deploy the AI model in a resource-constrained environment, but also how to maintain it for continuous performance updates, and share key takeaways for practitioners.

1 Introduction

The last ten years have seen a surge in AI and Natural Language Processing research to address real world problems that have a social good impact (Adauto et al., 2023). Many of these problems align with the United Nations Sustainable Development Goals (UNSDG)¹. This has also led to a surge in publications in this space to the point that even prominent AI research conferences have special tracks and themes related to social good (ie. AAI, ACL-IJCNLP in 2021 (Zong et al., 2021)) and many targeted venues to tackle this topic such as the NLP for Positive Impact workshop series².

Jin et al. (2021) describe four different stages of AI for Good tasks: 1. Fundamental theories, 2. Building block tools, 3. Applicable tools and 4. Deployed applications. While there have been a lot of publications in this space (for example Adauto et al. (2023) found that just over 13% of all papers in the ACL Anthology map to one of the UNSDGs), most published AI for Good work has tended to focus more on the first three stages: specifically on analysis of the problem area, building a dataset, or

building a model. However, there is comparatively very little published work on the fourth stage: on how these models fare when deployed in the real world and how they align with the expectations of the social good organization. In fact, for the ACL-IJCNLP 2021 special theme of "NLP for Social Good", only one of the twelve accepted papers mentioned deployment.

In addition, there has been very little work that discusses the collaboration process between a humanitarian organization and AI practitioners where a model is built to be used by the partner organization. The closest works are Tomašev et al. (2020) and Kshirsagar et al. (2021), which highlight how AI teams should approach and undertake AI4SG projects - but do not mention any details about development and deployment process.

In this short paper, we present our experience with working with Insecurity Insight³, a humanitarian-to-humanitarian organization (H2H), to bring an NLP model into the real world and provide impact to that organization and the aid community it supports. This work builds upon our previous research (Lamba et al., 2024), in which we developed a multilingual dataset of news articles in English, French, and Arabic, annotated with various types of violent incidents categorized by the humanitarian sectors they affect—such as aid security, education, food security, health, and protection. We also evaluated a range of deep learning architectures and techniques to tackle the associated task-specific challenges. In this paper, we take the next step by addressing the critical final stage: model deployment. In particular, we discuss not only the technical and process aspects of deploying a model in a resource-constrained environment, but also how to maintain it for continuous performance updates. We conclude with key takeaways and best practices for both AI model developers

¹<https://sdgs.un.org/goals>

²<https://sites.google.com/view/nlp4positiveimpact>

³<https://insecurityinsight.org/>

and humanitarian experts around technical topics, collaboration and processes. While this is just one example of a deployment, we hope this paper will encourage others to share their experiences and lessons learned.

2 Partnership Case Study

2.1 Partner Details

Insecurity Insight is a data-based H2H organization. Their aim is to support the work of aid agencies, healthcare providers, and other civil organizations by providing data-driven intelligence reports that can be used by these organizations for efficient resource allocation, humanitarian response, fund raising, advocacy, among others. Before our collaboration, Insecurity Insight collected news articles from select data sources (i.e. NewsAPI (Lisivick, 2018), OSAC⁴, and through manual uploading of news articles by humanitarian experts. These articles were then passed to an SVM model for relevance classification and category classification (categories defined on downstream humanitarian impact - education, aid, health and protection). Once classified and tagged, they were reviewed and summarized by humanitarian experts. However, this workflow had two drawbacks: (1) it was limited to existing downstream humanitarian categories and (2) it focused only on English articles.

2.2 Problem Scope

For our partnership with Insecurity Insight, we identified the following three shared goals. The plan was to develop NLP models which could address these goals and then deploy them in their workflow.

Goal 1. Improve the existing workflow to identify and classify more relevant news events.

Goal 2. Expand to new domain of food security.

Goal 3. Expand to French and Arabic articles.

2.3 Resource Constraints

A key challenge of AI4SG collaborations is that often the organization that uses AI might not have many resources to dedicate to the development, hosting, and maintenance of AI models. Our partner organization also faced similar challenges. Working in resource-constrained environments produces interesting challenges for AI developers. We list some of them below:

Labeling Resources: Our partner had a limited number of humanitarian experts on staff, leading to a constrained article review capacity in the live production workflow, as well as limited time for completing separate offline annotation tasks, which were crucial for model development.

Low Compute Environment: The model was intended to be deployed within the existing infrastructure to avoid incurring additional costs for the partner organization. The deployment infrastructure consists of Heroku Basic dyno (1 vCPU, 512MB memory) for running scheduled crawling jobs, a dedicated VPS machine (4 vCPUs, 8GB memory) for hosting the classifier API and a MongoDB database (2GB storage). There is no real-time latency requirement for the model inference, however it is critical for the throughput rate of the scheduled crawling and classification jobs to keep up with the influx of new articles.

Maintenance: The partner had minimal engineering staff so it was crucial to deliver a solution that was robust and easily maintained.

3 Implementation and Deployment

Following standard ML Ops practices (Shankar et al., 2024) we split the model development into three stages: offline experimentation, staging deployment calibration and deployment monitoring (as presented in Figure 1).

3.1 Offline Experimentation

GDELT Source Expansion: Two of the key goals are to expand the current workflow so that it can tag in new domains and expand to articles in French and Arabic. To address both, we augment the current data sourcing with GDELT (Leetaru and Schrodt, 2013), a large real-time open-source database of multilingual news articles.

Data Labeling: To collect labeled data for the new input distribution, we established an offline spreadsheet labeling process with 7 humanitarian experts from Insecurity Insight using annotation guidelines similar to their established live workflow. Expert annotators reviewed the title and content of the scraped article before determining whether the article is relevant and assigning the event categories. To ensure high quality labels, we used annotator deliberation to improve high inter-annotator agreement rates. Given the limited annotation resources, we tried to get annotation for a sample of data ensuring that it was diverse in lan-

⁴<https://www.osac.gov/Content/Browse/News>

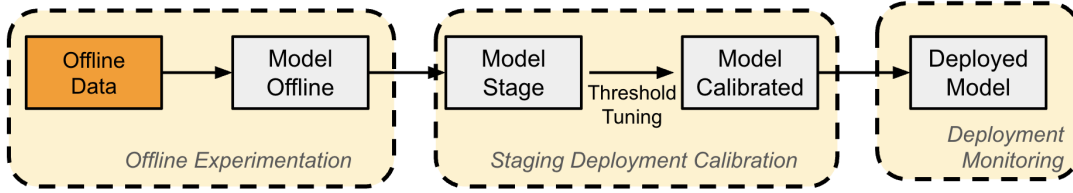


Figure 1: Stages of our model lifecycle

guage, categories, and a base model confidence’s score. The dataset and associated repository are published at <https://github.com/dataminr-ai/humvi-dataset>. More details on the data collection and quality control can be found in our previous work (Lamba et al., 2024).

Model Development and Selection: We trained two models - (1) Relevance Model for identifying relevant news articles, and (2) Categorization Model for tagging relevant articles with proper downstream humanitarian categories. In order to detect food security events, the category classification model is expanded to five output classes. During training, we translated English data to French and Arabic to augment initial training samples, and used label loss masking (Duarte et al., 2021) to account for the new category label. We focused on evaluating three smaller-sized multilingual transformer models - BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019; Conneau et al., 2019), and DistilBERT (Sanh et al., 2019), all of which could be deployed given the compute and latency constraints. We temporally split the labeled data to establish offline relevance and category classification performance on a held-out test set. XLM-RoBERTa performed best in expanding to the new input domain and languages (Relevance F1 scores ranged from 0.81 to 0.83 for the three languages) and thus was selected for deployment ($\mathcal{M}_{stage}$); ensuring that new workflow can source new types of articles with higher coverage and can tag them for new languages and new categories.

3.2 Staging Deployment Calibration

Though the deployed model performed well on the offline dataset, the main test was whether those scores would hold when deployed in the real world setting and bring value to Insecurity Insight. We envisioned the model performance could be lower due to (1) content drift (Elwell and Polikar, 2011) given the offline test set was collected a few months earlier; (2) possible mismatch between offline and online computing environments; and (3) the in-

creased volume of articles could overwhelm the human review system given limited staffing.

Offline Test Setup: To minimize the risks above, we worked closely with our partner to conduct a pre-deployment test in a staging environment. We integrated the GDELT data source and deployed the model $\mathcal{M}_{stage}$ and ran it in parallel to the existing production system for 2 weeks. To evaluate the “live” performance on data from GDELT, we sampled 1,000 examples using stratified sampling by discretized model confidence scores. For existing sources NewsAPI and OSAC, we re-use the labels from the production SVM-based system.

Model Threshold Tuning: We tuned relevance classification thresholds for each language given the annotated data sampled from the live staging environment. Table 1 presents the recall, precision, and estimated volume of weekly articles to review given different threshold options for English. Table 2 further presents the estimated volume of articles to review (i.e., articles predicted as relevant) across three different sources: NewsAPI, OSAC and GDELT. After the source expansion, around 90% of the ingested data came from GDELT.

Option	Threshold	Recall	Precision	Volume
Baseline	0.184	0.85	0.785	951 (20x)
Option 1	0.646	0.790	0.802	803 (17x)
Option 2	0.943	0.532	0.854	484 (10x)
*Option 3	0.951	0.405	0.903	367 (8x)

Table 1: Volume (number of articles to review per week) and quality (precision & recall) impact given different proposed thresholds for relevance classification for English. *=Model Selected

Per the initial requirement from our partner, the baseline model threshold (0.184) was tuned with max precision at minimum recall 0.85 to minimize missing potentially relevant articles. With the inclusion of GDELT this approach would lead to a **20x** estimated increase (from 46 to 951 weekly) in articles to review. We discussed this recall-volume trade-off with Insecurity Insight and decided to move forward with Option 3 (henceforth $\mathcal{M}_{prod}$)

at minimum 0.90 precision to reduce the expected labeling burden increase to **8x**. We perform a similar analysis for Arabic and French (see results in Tables 5 and 6 in Appendix A.1), and select a threshold at a lower minimum precision (0.80 for Arabic and 0.62 for French) due to the smaller number of articles crawled.

Threshold	Recall	Precision	Source	Volume
0.184	0.85	0.785	NewsAPI	80
			OSAC	21
			GDELT	850
0.646	0.790	0.802	NewsAPI	67
			OSAC	16
			GDELT	720
0.943	0.532	0.854	NewsAPI	36
			OSAC	8
			GDELT	440
*0.951	0.405	0.903	NewsAPI	22
			OSAC	5
			GDELT	340

Table 2: Volume (number of articles to review per week) and quality (precision & recall) impact given different proposed thresholds for relevance classification for English. The volume is broken down by source. Most articles came from the expanded source GDELT. *=Model Selected

For category classification, we set one threshold across all languages for each category. We tune it to minimum precision ≥ 0.8 in line with the baseline system.

3.3 Post-Deployment Analysis

To assess the deployment we compared data from the live system 4 months after the final model $\mathcal{M}_{prod}$ deployment with the baseline system performance in 2024. Table 3 shows the impact of the deployment in terms of article volume across each stage of the system. Overall, we surfaced **3.6x** more confirmed relevant articles compared to the baseline system with a **3.2x** increase in manual labeling effort. The precision of the system had improved from the 0.80 baseline and is closely aligned with the estimated precision from the pre-deployment threshold tuning stage (0.92 for English, 0.82 for French and 0.82 for Arabic). The GDELT source expansion led to a **23x** increase in crawled articles per week, and the updated classifier predicted **9x** more articles as relevant. A significant number of confirmed relevant articles were surfaced in French and Arabic (42% of the total baseline volume).

Food Security: We expected an **8x** volume increase but only marginally improved the system’s

Pipeline Stage	Baseline	Deployment
Crawled	450	10,550
Predicted Relevant	54	496
– English	54	326
– French	0	41
– Arabic	0	129
Confirmed Relevant	43/54	154/171
– English	43/54	131/142
– French	0	9/11
– Arabic	0	14/17

Table 3: Volume (number of articles per week) across each stage of the system before and after the model deployment.

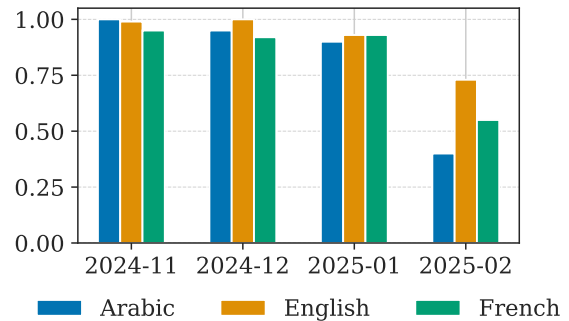


Figure 2: Relevance classifier precision over time by source language.

ability to surface more articles of this category (from 1 to 3 per week). The F1 Score for this class significantly drops between offline evaluation ($F1 = 0.679$) and product deployment ($F1 = 0.014$) for English articles. And there were even no articles in Arabic or French labeled. Full results per category are presented in Table 4 in Appendix A.2. Upon further review, we determined that there were missing labels due to annotation inconsistencies, which were traced back to unclear annotator guidance and poor calibration. This highlights the importance of performing regular data quality checks.

Performance Over Time: Figure 2 shows the relevance model performance over time. Notably there was a performance drop in the last month of collected data across all languages. This showed that there was a risk of model performance degradation due to shifts in the live data distribution. We addressed this drop by providing the partner with workflows for continuously monitoring the live model performance and a recipe for retraining

the model artifact based on new labeled data.

4 Discussion

Developing and deploying AI strategies for “AI for Good” projects presents unique opportunities and challenges for AI practitioners and NGOs. Ensuring a sustainable and impactful deployment requires a collaborative approach that bridges technical expertise and domain-specific knowledge. Below we outline key takeaways from our collaboration with Insecurity Insight.

T1. Understanding the Problem: Before developing AI models, practitioners must deeply understand the problem they want to address. This requires a thorough stakeholder engagement, data assessment, and problem scoping. During the early phase of the project, we gathered crucial domain knowledge from domain experts and engineers in Insecurity Insight to get a deep understanding of their current service and system, impact measurement, specific needs with priorities, resource and operational constraints, data availability and technical stack. This helps inform our key decisions in the steps of data collection, model selection and deployment.

T2. Data Availability and Quality: Both parties must assess the availability, reliability and bias of data sources. Available data may be noisy or limited in scope, thus requiring new data collection methods or annotation. Data quality could be an ongoing issue, and thus it is important to start early and iterate: practitioners should work with domain experts to come up with clear annotation guidelines. In this particular study, we found it is essential to be mindful of the domain expert’s time (operational cost). This requires both teams to setup realistic and meaningful plans and schedules.

T3. Capacity Building: For AI solutions to be sustainable, partner organizations must have the capacity to use and maintain them. It is important to keep the partner in the loop throughout the development process and establish support mechanisms for model updates, debugging, and continuous improvement.

T4. Model Performance Mismatch Awareness: Both parties should be aware of potential discrepancies between offline evaluations and real-world AI performance (as we saw with our food security results). Establishing a staged testing environment helps validate and refine AI solutions before deployment, reducing unexpected behaviors in pro-

duction. Both parties should be flexible in adjusting metrics to better fit real-world needs (e.g., optimize for precision instead of recall).

T5. Impact Assessment and Continuous Monitoring: It is important to establish clear metrics to measure success. Once deployed, AI solutions should be regularly evaluated for performance drift (as shown in Figure 2). While automated monitoring pipelines can track key metrics in real-world use, continuous calibration of labeling quality is integral to informing robust metrics. Retraining with fresh data and adjusting decision thresholds helps maintain accuracy and thwart content drift.

In short, this paper details our experience of developing and deploying a model to assist a humanitarian organization in a resource-constrained setting. The implementation process and takeaways may be useful for practitioners that are seeking to operationalize AI models in low-resource settings. This “final stage” is often quite challenging, and we hope other practitioners will publish their process and impacts as well.

5 Limitations

We acknowledge that this is just one example of an AI deployment in a humanitarian setting. Ideally, we would present several examples of such deployments to paint a more robust picture of the different decisions partners can make, and the associated challenges. However, that is outside the scope of this short paper. We hope that by going into the details of this deployment process and showing the real-world impact will encourage others to publish their findings as well.

Another aspect we want to acknowledge is that there are many different types of AI for Good projects and deployments. A group of AI scientists partnering with a humanitarian organization is just one configuration.

6 Ethical Considerations

The dataset is constructed from publicly available news articles, ensuring that no contractual agreements were violated in the data acquisition process. Our web scraper strictly accessed openly available content, excluding any material behind paywalls. For the annotation process, we engaged internal humanitarian experts from the partnering organization. These experts were fairly compensated as part of their professional, paid employment.

References

- Fernando Adauto, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. [Beyond good intentions: Reporting the research landscape of NLP for social good](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 415–438, Singapore. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- 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*.
- 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.
- Ryan Elwell and Robi Polikar. 2011. [Incremental learning of concept drift in nonstationary environments](#). *IEEE Transactions on Neural Networks*, 22(10):1517–1531.
- Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. [How good is NLP? a sober look at NLP tasks through the lens of social impact](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.
- Meghana Kshirsagar, Caleb Robinson, Siyu Yang, Shahrzad Gholami, Ivan Klyuzhin, Sumit Mukherjee, Md Nasir, Anthony Ortiz, Felipe Oviedo, Darren Tanner, Anusua Trivedi, Yixi Xu, Ming Zhong, Bistra Dilkina, Rahul Dodhia, and Juan M. Lavista Ferres. 2021. [Becoming good at ai for good](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 664–673, New York, NY, USA. Association for Computing Machinery.
- Hemank Lamba, Anton Abilov, Ke Zhang, Elizabeth M Olson, Henry Kudzanai Dambanemuya, João Cordovil Bácia, David S. Batista, Christina Wille, Aoife Cahill, Joel R. Tetreault, and Alejandro Jaimes. 2024. [HumVI: A multilingual dataset for detecting violent incidents impacting humanitarian aid](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12705–12722, Miami, Florida, USA. Association for Computational Linguistics.
- Kalev Leetaru and Philip A Schrod. 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.
- 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*.
- 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*.
- Shreya Shankar, Rolando Garcia, Joseph M. Hellerstein, and Aditya G. Parameswaran. 2024. [“we have no idea how models will behave in production until production”: How engineers operationalize machine learning](#). *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–34.
- Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, and 1 others. 2020. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):2468.
- Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors. 2021. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online.

A Appendix

A.1 Threshold Tuning across 3 Languages

As we tune the thresholds per language, Table 5 and 6 presents the quality and volume impact under different thresholds. Arabic shows a good volume of articles, which meets well with our initial goal of expanding to collecting articles from Arabic-speaking local geographical areas. Although we were not able to surface a good number of French articles, this is still a good start for Insecurity Insight.

A.2 Categorization Model Performance

Table 4 compares the metrics of categorization model between using the offline test set and using the live labeled data in production. The metrics across most event category and languages align well before and after deployment. However, we observed significant metric discrepancy for Food Security across all languages, and for Aid Security in Arabic. This could be attributed to multiple reasons: (1) model degenerates due to content drifts and poor model generalization; (2) There was just

Category	Old Model	New Model (Offline Test Set)			New Model (Live data)		
	English	English	French	Arabic	English	French	Arabic
Food Security	Not supported	0.679	0.491	0.661	0.014	No labels	No labels
Aid Security	0.560	0.729	0.745	0.688	0.672	0.947	0.362
Education	0.245	0.773	0.563	0.571	0.669	0.671	0.772
Health	0.365	0.681	0.792	0.629	0.758	0.680	0.664
Protection	0.357	0.708	0.775	0.888	0.908	0.655	0.764

Table 4: The performance of category classification using the offline test set versus using the live labeled data in production system. There observed as huge discrepancy of performance metrics for Food Security across the languages, and Aid Security in Arabic language.

Option	Threshold	Recall	Precision	Volume
Baseline	NA	NA	NA	0
Option 1	0.125	0.676	0.50	63
*Option 2	0.881	0.432	0.615	39
Option 3	0.942	0.324	0.706	26

Table 5: Volume (number of articles to review per week) and quality (precision & recall) impact given different proposed thresholds for relevance classification for French, which was crawled only from GDELT source.
*=Model Selected

Option	Threshold	Recall	Precision	Volume
Baseline	NA	NA	NA	0
Option 1	0.361	0.793	0.605	230
Option 2	0.824	0.690	0.714	211
*Option 3	0.952	0.414	0.8	150

Table 6: Volume (number of articles to review per week) and quality (precision & recall) impact given different proposed thresholds for relevance classification for Arabic, which was crawled only from GDELT source.
*=Model Selected

not many Food Security event happened during the time when the live data was collected; (3) The labelers who reviewed Food Security articles did not perform as guided. Through reviewing samples with high food security category classification score we determined that there are missing labels due to improper annotator guidance and calibration. This highlights the importance of performing regular data quality checks.