INNOVATIVE SEARCH APPLICATIONS, OPEN SEARCH USE CASES AND CHALLENGES

#OSSYM2022

Searching and Structuring the Twitter Stream for Crisis Response: A flexible Concept to Support Research and Practice

Jan Bongard, Institute of Data Science, 11 Oct. 22

Civil Security

Threats:

- Climate change, pandemics, military conflicts
- Complex and rapidly changing threats

Research Objectives:

- Integration of Web- & Social Media data as additional data source
- Provision of a comprehensive information structure
- Enabling reactive & proactive decisions

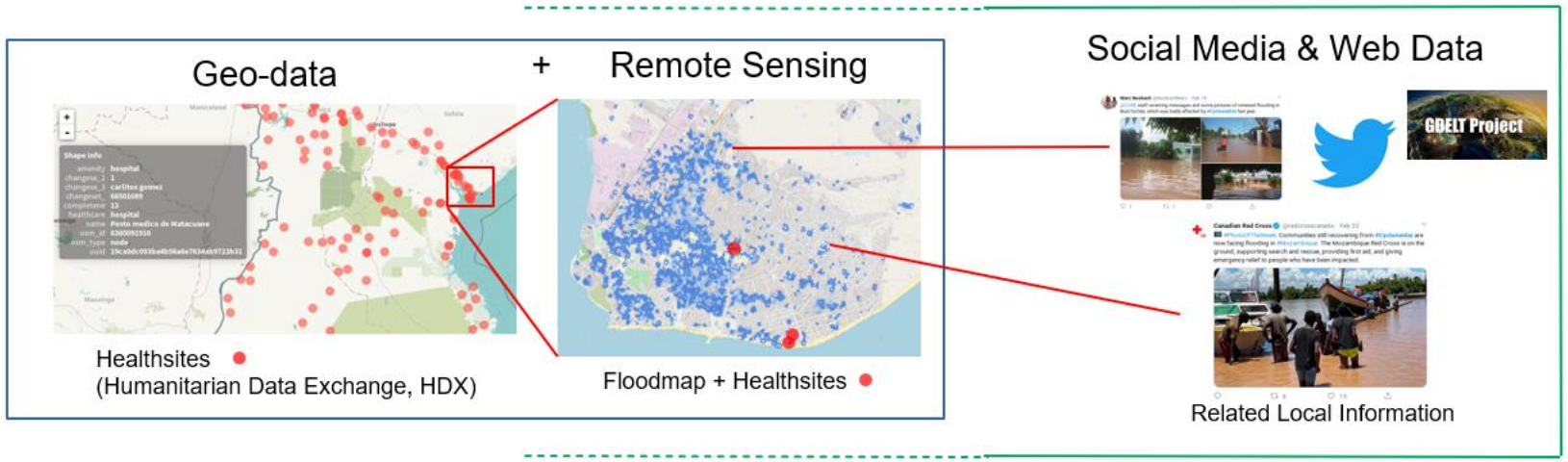
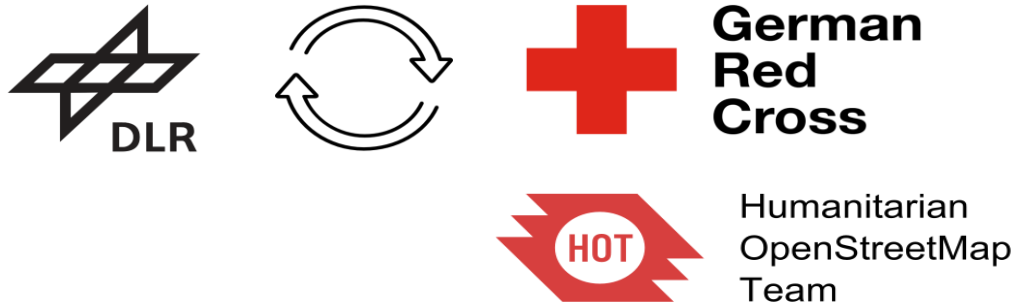


Figure 1: Established workflows & new web sources

Applied Research

- Gap between research and practice
- Unknown:
  - Infrastructure
  - Frameworks
  - Workflows
  - Topics of interest during decision making



- Solution: Early and continuous transfer of technology and knowledge

Decision Maker Requirements

- Real-time & continuous data availability
- Location is key
- Globally applicable workflow
- Flexible methods
- Data bias & trustworthiness
- User friendly visualization

Backbone: Modular Processing System

1 Data Acquisition	2 Pre-Processing	3 Information extraction	4 Classification
- Twitter Stream - Twitter Full-Archive	- Translation - Word count - Text cleaning	- NER (Flair) - GazPNE2 [1]	- Relevance [3] - Humanitarian categories [3] - Information type [4] - Information source [4]
Extensions		- Place name Disambiguation [2] - Geocoding	- Individual frameworks - Image Classification - Misinformation - Neural networks [5] - Gaussian Processes [6] - Few-shot models [7]

Backbone: Modular Processing System

1 Data Acquisition

- Twitter Stream
- Twitter Full-Archive

2 Pre-Processing

- Translation
- Word count
- Text cleaning

3 Information extraction

- NER (Flair)
- GazPNE2 [1]

4 Classification

- Relevance [3]
- Humanitarian categories [3]
- Information type [4]
- Information source [4]

Extensions

- Place name Disambiguation [2]
- Geocoding

- Individual frameworks
- Image Classification
- Misinformation
- Neural networks [5]
- Gaussian Processes [6]
- Few-shot models [7]

[1] Hu, X., Zhou, Z., Sun, Y., Kersten, J., Klan, F., Fan, H., & Wiegmann, M. (2022). GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models. IEEE Internet of Things Journal.
[2] Hu, X., Sun, Y., Kersten, J. & Klan, F (2022). How can voting mechanisms improve the robustness of individual toponym resolution approaches? (Accepted at: 17th International Conference on Location Based Services).
[3] Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. In: *ICWSM* (pp. 933-942).
[4] Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In: 8th ICWSM (pp. 376-385).
[5] Wiegmann, M., Kersten, J., Klan, F., Potthast, M., & Stein, B. (2020). Analysis of Filtering Models for Disaster-Related Tweets. In: Proceedings of the 17th ISCRAM, May 2020. ISCRAM. Blacksburg, Virginia (USA).
[6] Kersten, J., Bongard, J., & Klan, F. (2022). Gaussian Processes for One-class and Binary Classification of Crisis-related Tweets. In: Proceedings of the 19th ISCRAM, May 2022. Tarbes (France). (in print).
[7] Kruspe, A., Kersten, J., & Klan, F. (2019). Detecting event-related tweets by example using few-shot models. In: Proceedings of the 16th ISCRAM, May 2019. Valencia, (Spain).

Backbone: Modular Processing System

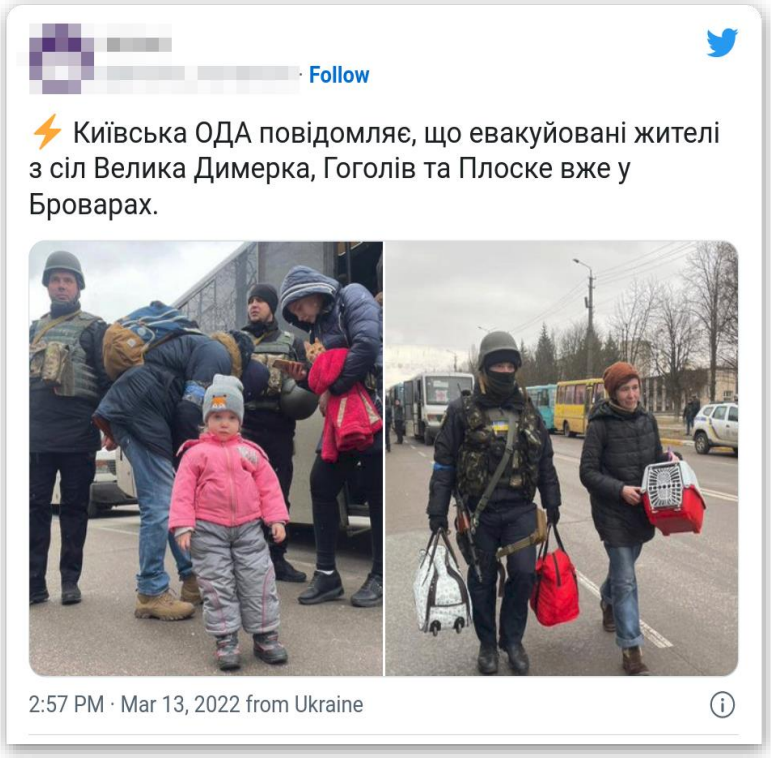


Figure 2: Tweet example

Translation:

'⚡Kyiv Regional State Administration (ORG) reports that evacuated residents from the villages of Velyka Dymerka (LOC), Gogoliv (LOC) and Ploske (LOC) are already in Brovary (LOC). https://t.co/...'

Classification:

Relevance: 0.99
Humaid: Displaced people and evacuations
Type: Affected individuals
Source: Not labeled

Frontend: Interactive Dashboard

1 User-defined queries:
- On relevance, humanitarian categories, keywords & time

2 Heat map:
- Define region of interest

3 Detailed content analysis:
- Feature space: Universal Sentence Encoder + UMAP Dim. Reduction
- Interactive semantic clustering
- Cluster selection & export functionality

4 Hover information:
- Tweet & cluster summarization: NER, tf-idf

5 Temporal development of selected clusters

6 Media carousel of selected tweets

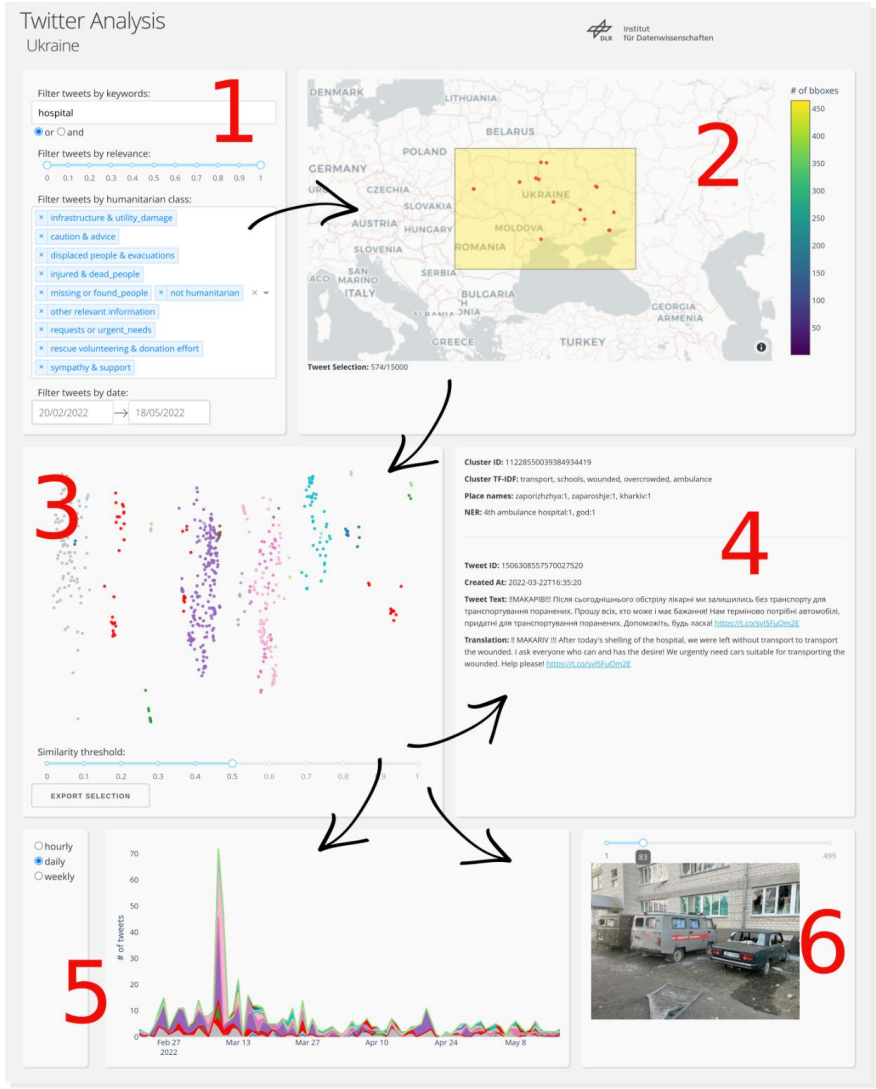


Figure 3: Dashboard Ukraine war (Twitter Full Archive Feb – May 2022)

First practical deployment

German Red Cross:

- Missing first hand information about medical infrastructure
- 526 hospital related tweets containing place names:
  - Attacks on hospitals: 25 cities
  - Functioning / partially operating: 33 cities
  - Completely destroyed: 9 cities

- Volume of relevant content dependent on virality of certain topics

Future Research

- Diversification of web sources
- Dashboard development & adaptation to practitioners‘ workflows
- Fusion with GIS Data & spatiotemporal analysis
- Information aggregation & summarization via clustering

Thank you!