Combining Remote Sensing with Webdata and Machine Learning to Support Humanitarian Relief Work

Jens Kersten[1](jens.kersten@dlr.de), Malin Kopitzsch[1], Jan Bongard[1], and Friederike Klan[1]
[1]German Aerospace Center (DLR), Jena, Germany

Research Context

DLR-Project Data4Human: Demand-driven data services for humanitarian aid [1]

Involved partners:

German Red Cross · World Food Programme · Humanitarian OpenStreetMap Team

Scope

- Current data collection and analysis mainly based on remote sensing, geospatial and in-situ data
- Webdata and social media potentially adds value in monitoring natural and man-made disasters [2,3,4]
- General issues: Limited resources, data overload and bias, missing or uncertain location information, trustworthiness, ….

Addressed questions (Fig. 1):
- What are the current information gaps and user needs?
- How can these gaps be complemented by webdata and corresponding ML methods?
- What are the potentials, benefits and weaknesses, especially for social media data?

Scenarios and Data

Focus Area: Mozambique, Africa (Fig. 2)
Incident types:
- Natural disasters: Cyclones Idai & Kenneth, March-May 2019
- Conflicts, humanitarian crises & security-related incidents

Data sources:
1) GEDELT [5]:
   - Live, global event database
   - Broadcast, print, online news
   - 60+ languages
2) Twitter:
   - Worldwide & immediate platform
   - 1 % live stream + full history archive

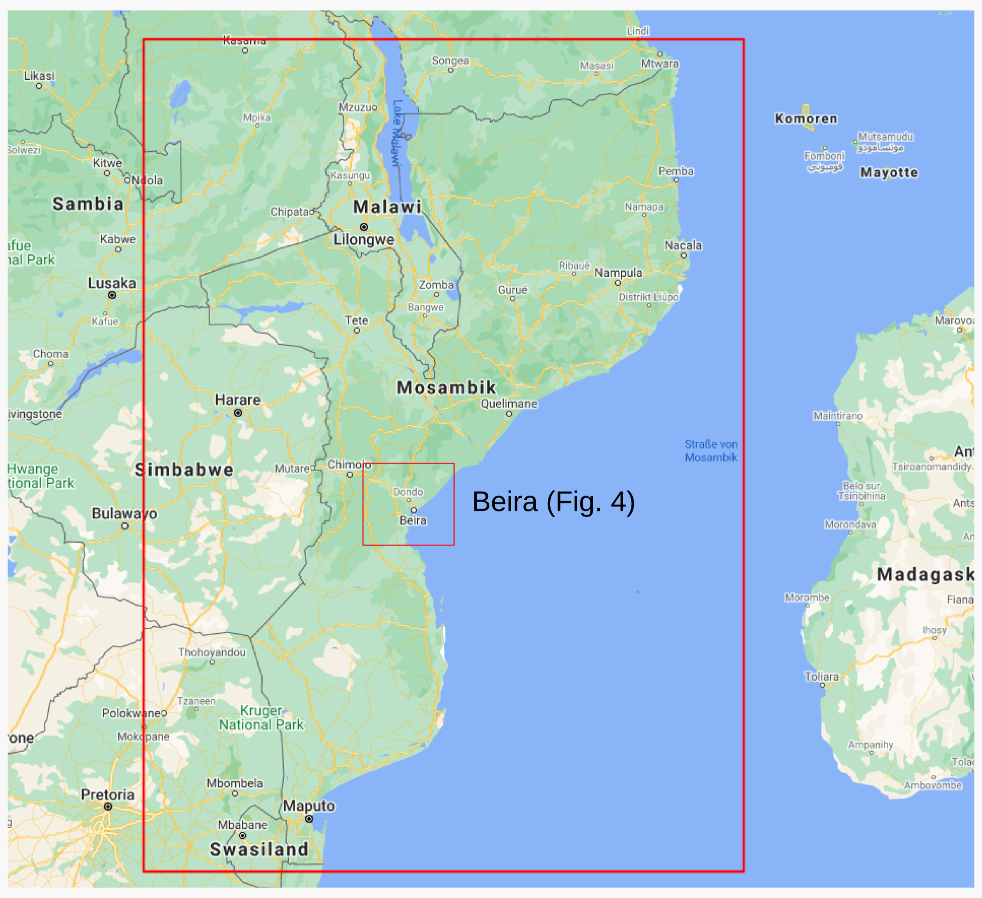

Fig. 2: Map and AOI of the study area Mozambique.

Needs Assessment Results

- Natural disasters already well covered, but information gaps possible
- Desired: A better description of conflicts and human behavior

- Relevant and helpful: Local information
  - Impact, needs & damage assessment
  - Health-related content
  - Information on injuries and death toll
  - Detection and monitoring of security-related incidents, social domino effects, armed clashes, ……

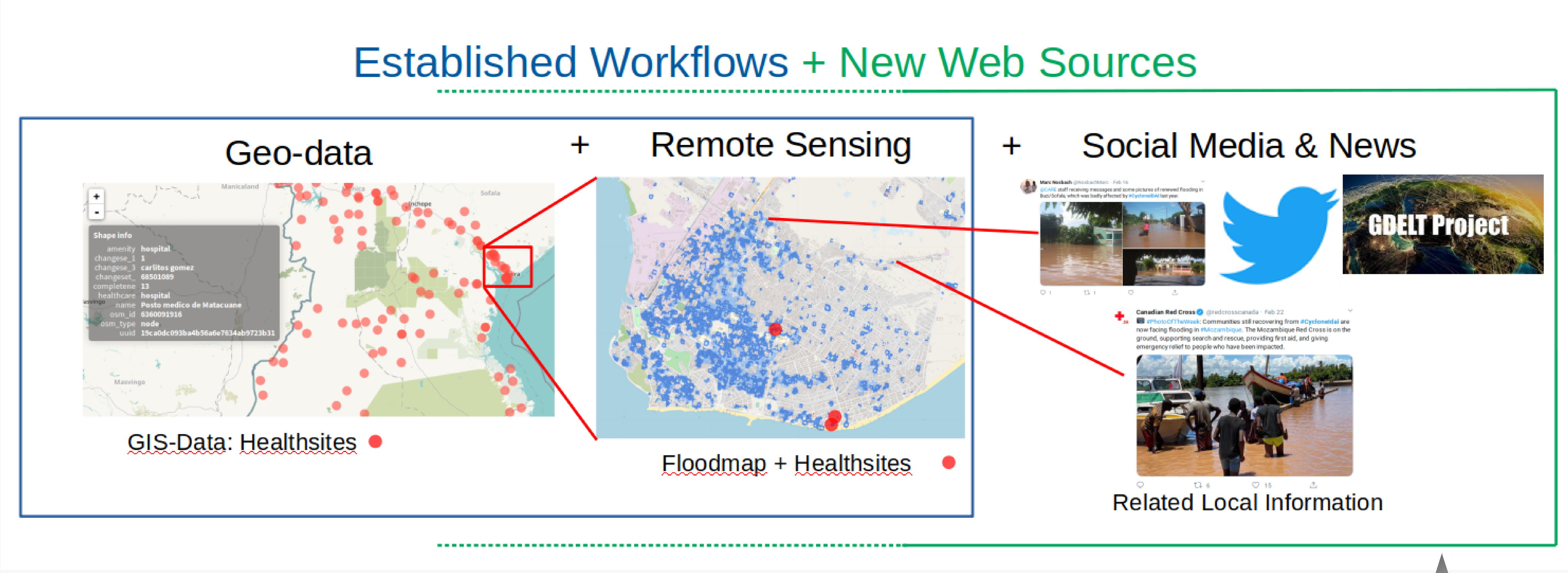

Fig. 1: Main question of this work: How to complement, enrich and fill gaps of existing workflows?

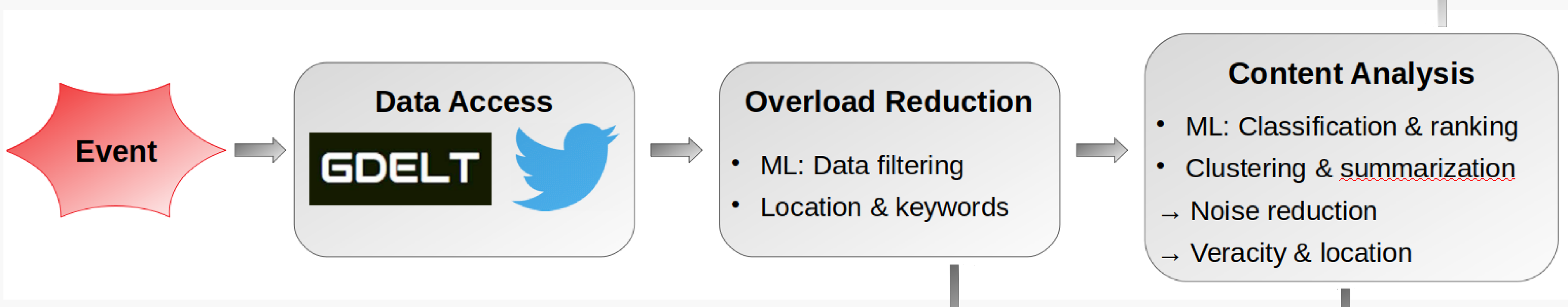

Fig. 3: Workflow for news article and tweet analyses.

Data Analysis Methods

We utilize state-of-the-art ML methods to address common practical issues, like overload reduction, content analysis, and data aggregation/contextualization (Fig. 3):

1) Overload Reduction
- Binary classification models [6]
- Location and keywords

2) Classification
- Deep Learning Models [7]
- Information type, source, priority
- Multi-class, -label and -task models

3) Stream Analysis [8]
- Semantic clustering of Twitter stream data
- Embeddings-based text representation
- Merging and linking of clusters
- Topic detection and tracking
- Topic and stream summarization

First Results

Collected data (March 15 – May 15, 2019):
- ~2.3 GB/day (1% worldwide Twitter stream)
- ~200k news articles/day (GDELT)

1) Twitter
- Filtering by geolocation (Fig. 2)
- F1-score ~ 0.83, but expected to be lower in case of new event types not yet covered in training data [6]
- After overload reduction (binary classification): ~7,000 potentially relevant Tweets identified → to be investigated

2) GDELT
- Filtering by keywords & locations (Fig. 2)
- Web-scaping of news articles
- Application of binary ML model to identify potentially relevant news articles
- F1-score ~ 0.83
- Result products: daily/weekly maps (Fig. 4)

- Despite data sparsity, Twitter tends to add more value (impact assessment)
- Information overload reduction is definitely required, but just the starting point for further in depth-analyses that allow to gain valid and localized information from unstructured stream data
- Besides the type of information contained in a tweet (e.g. help needed), also the source type (e.g. first party observation or governmental) is crucial

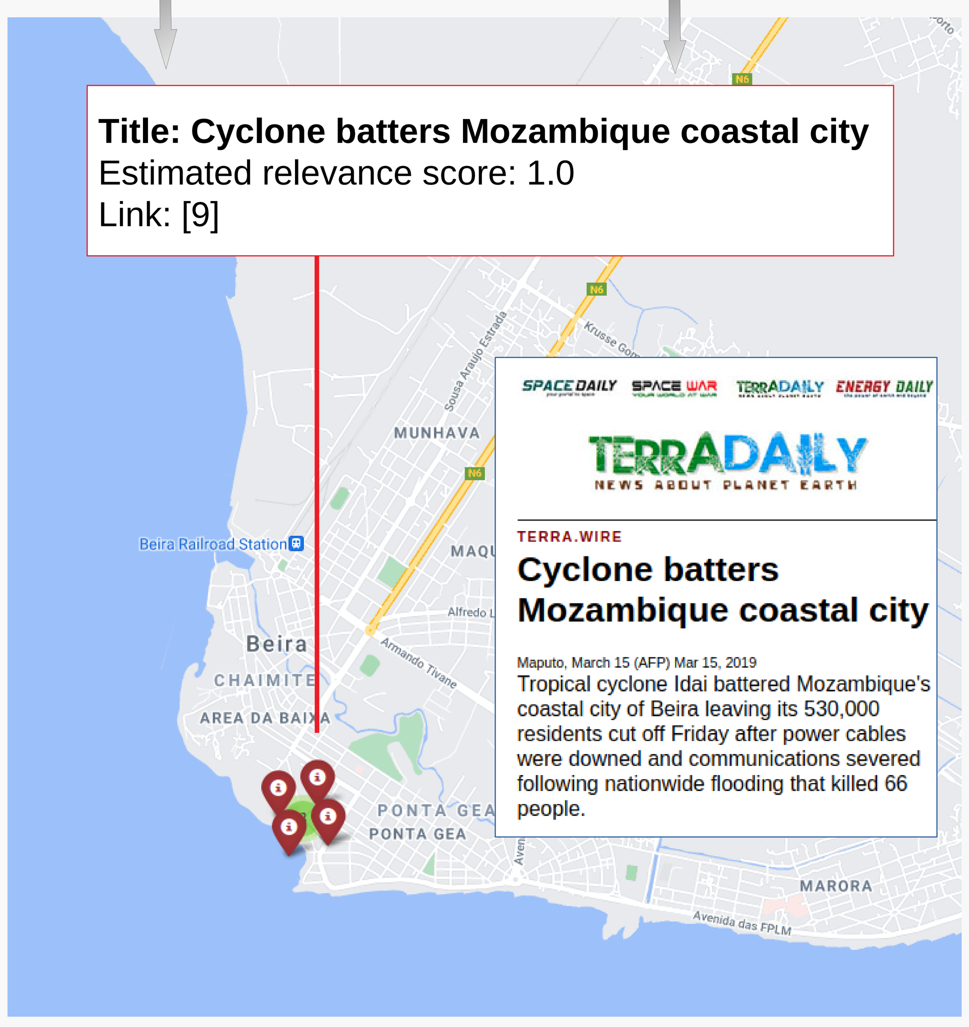

Fig. 4: Interactive example map representing four crisis-related news articles (GDELT) published March 15, 2019, Beira, Mozambique.

Conclusions and Outlook

- Current ML models are well suited to analyze the content of microblogs and news articles
- Each NGO addresses different questions → flexible methods required
- Adaptive stream analysis, clustering and summarization is required, in order to better understand single information snippets → noise reduction and validation
- Future work will focus on tweet classification and Twitter stream analysis (quantitative experiments)
- A prototypical workflow according to Fig. 3 will be implemented in order to analyze historical and live stream data

References

[1] https://reliefweb.int/report/world/data4human-aid-organisations-get-helping-hand-german-aerospace-center

[2] Wiegmann, et al.: Opportunities and Risks of Disaster Data from Social Media: A Systematic Review of Incident Information, Nat. Hazards Earth Syst. Sci. Discuss., in press, 2020.

[3] Kruspe et al.: Review article: Detection of informative tweets in crisis events, Nat. Hazards Earth Syst. Sci. Discuss., in review, 2020.

[4] Kersten and Klan: What happens where during disasters? A Workflow for the multifaceted characterization of crisis events based on Twitter data. JCCM, 28(3), 2020.

[5] https://www.gdeltproject.org/

[6] Wiegmann et al.: Analysis of Detection Models for Disaster-Related Tweets, ISCRAM 2020.

[7] Kruspe, et al., Classification of Incident-related Tweets: Exploiting Word and Sentence Embeddings, TREC 2019.

[8] Kerstenet al.: Combining Supervised and Unsupervised Learning to Detect and Semantically Aggregate Crisis-Related Twitter Content, ISCRAM 2021, in review.

[9] https://www.terradaily.com/afp/190315122324.yz9weul6.html

Deutsches Zentrum für Luft- und Raumfahrt

DLR

EGU General Assembly 2021