Geospatiality of Topics in English Text Data

Deutsches Fernerkundungs-datenzentrum

Johannes Mast[1], Richard Lemoine-Rodríguez[2,3], Vanessa Rittlinger[1], Martin Mühlbauer[1], Carolin Biewer[3,4], Christian Geiß[1,5] and Hannes Taubenböck[1,2,3]

[1]German Aerospace Center, German Remote Sensing Data Center, Weßling, Germany; [2]Institute of Geography and Geology, Chair of Global Urbanization and Remote Sensing, University of Würzburg, Germany; [3]Geolingual Studies Team, University of Würzburg, Germany; [4]Department of English and American Studies, Chair of English Linguistics, University of Würzburg, Germany; [5] Department of Geography, Chair of Georisk Research with Remote Sensing Methods, University of Bonn, Germany

MOTIVATION

Spatially mapping content from Web Text Data – Potentials and Challenges

Many scientific disciplines have experienced an increased interest in the role of geographic location in what is often referred to as a "spatial turn" [1].

Web texts that mention locations can be analyzed spatially, making them a promising data source. However, not all texts are equally likely to mention locations.

The topic matters – but how much? [2]

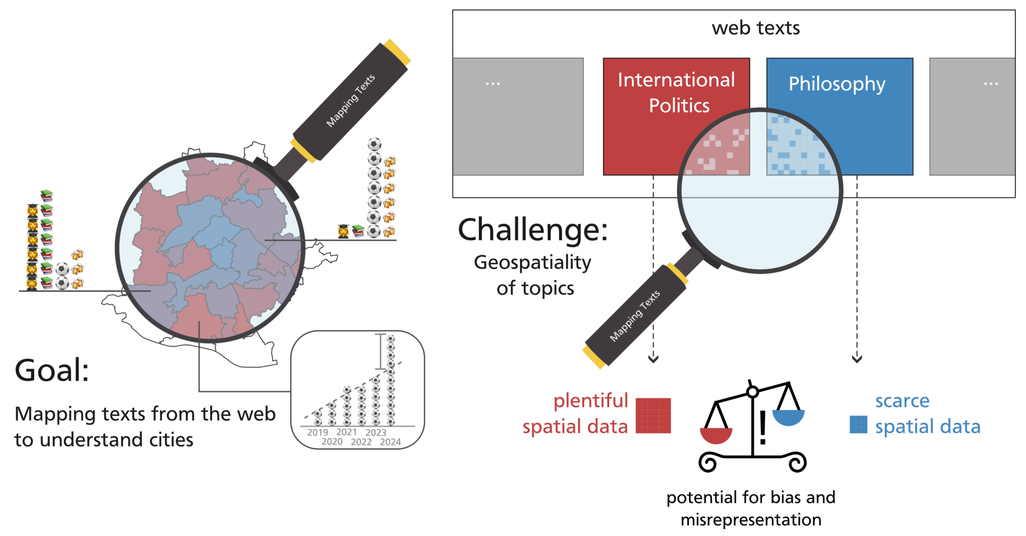

Fig.1 Graphical Outline of the Study

APPROACH

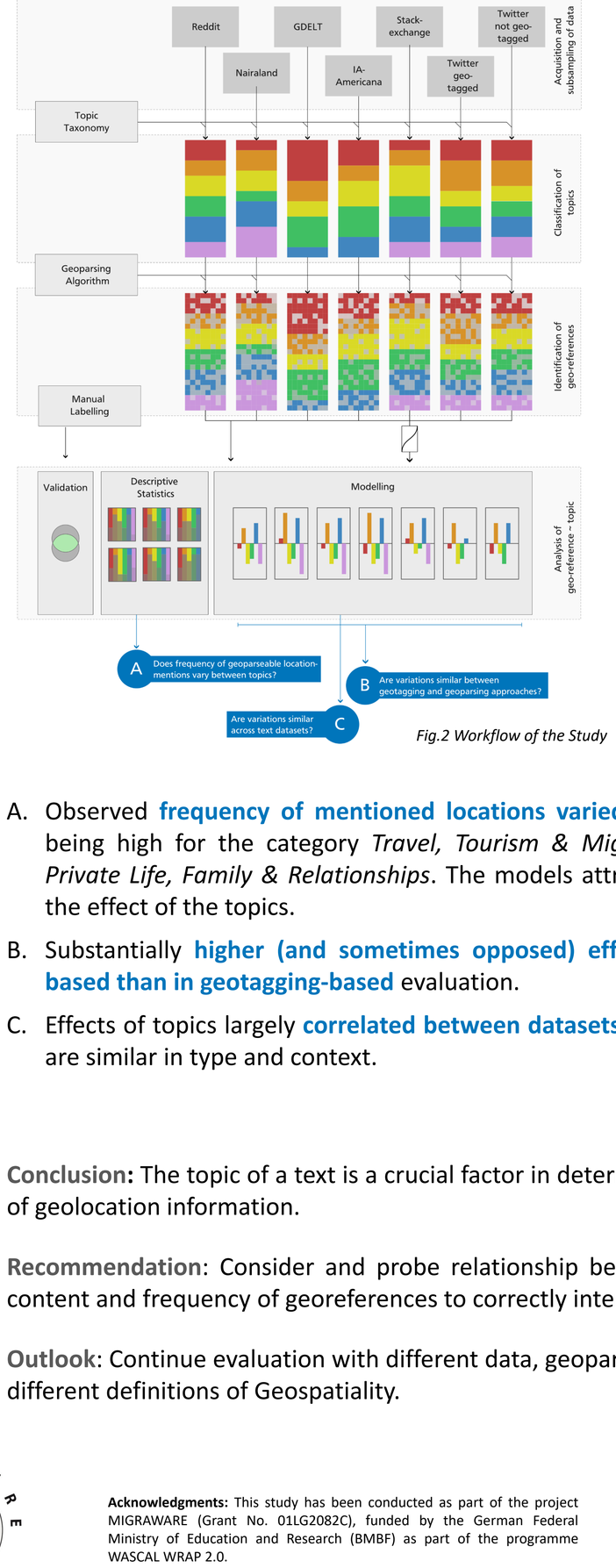

Fig.2 Workflow of the Study

Modeling Topics' Effect on Likelihood of Location Mentions

Across several millions of texts from different web platforms and text types, we algorithmically (A) detected whether they contained an identifiable spatial location and (B) identified the general topic they contained, following a custom taxonomy of 19 topics derived in an inductive coding approach.

We modeled the likelihood of geoinformation versus topic in a mixed-modeling approach, controlling for effects of author, time, and text length.

A. **Topic-specific effects on geospatiality** were assessed via the topics' fixed effects in the model.

B. **Similarity between geotagging and geoparsing approaches** was analyzed using data from Twitter, where we compared the results based on geoparsed location mentions (non-geotagged) to results based on geotags.

C. **Similarity across datasets** was measured using Spearman's rank correlation coefficient.

RESULTS

A. Observed **frequency of mentioned locations varied strongly by topic**, being high for the category *Travel, Tourism & Migration* and low for *Private Life, Family & Relationships*. The models attribute this largely to the effect of the topics.

B. Substantially **higher (and sometimes opposed) effects in geoparsing-based than in geotagging-based** evaluation.

C. Effects of topics largely **correlated between datasets**, moreso if datasets are similar in type and context.

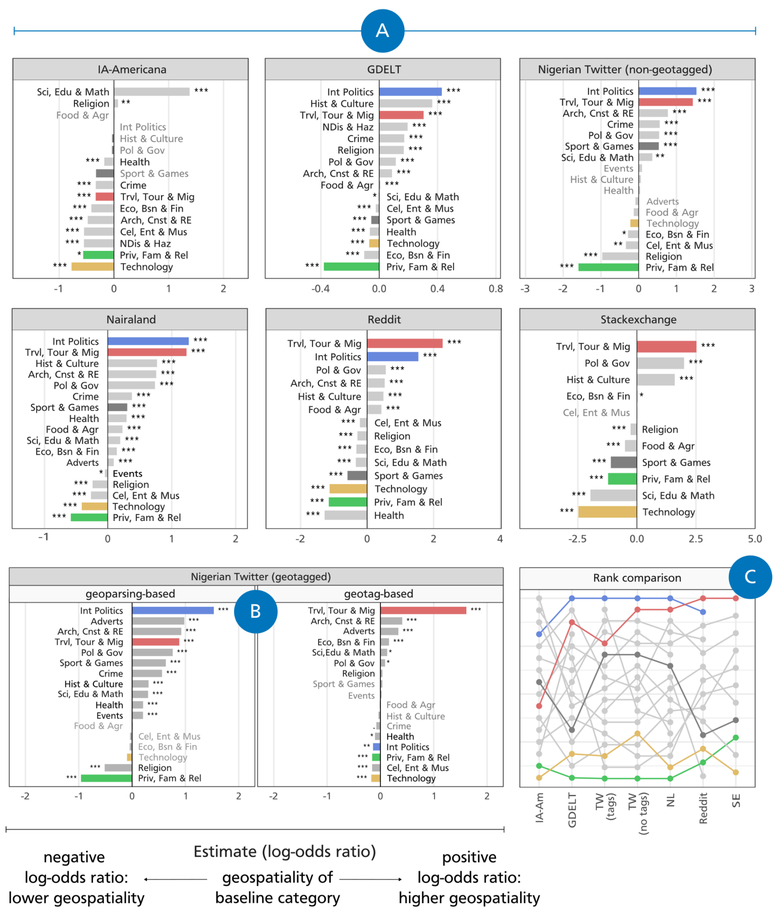

Fig.3 Modeled geospatiality effect of topics (log-odds ratio compared to a baseline topic other) within the datasets.

Conclusion: The topic of a text is a crucial factor in determining the presence of geolocation information.

Recommendation: Consider and probe relationship between the analyzed content and frequency of georeferences to correctly interpret results.

Outlook: Continue evaluation with different data, geoparsing methods, and different definitions of Geospatiality.

SPONSORED BY THE Federal Ministry of Education and Research

MIGRAWARE

Acknowledgments: This study has been conducted as part of the project MIGRAWARE (Grant No. 01LG2082C), funded by the German Federal Ministry of Education and Research (BMBF) as part of the programme WASCAL WRAP 2.0.

[1] Robert Tally. Spatiality. Routledge, 2012. URL https://doi.org/10.4324/9780203082881
[2] Mast, J., Lemoine-Rodríguez, R., Rittlinger, V., Mühlbauer, M., Biewer, C., Geiß, C., & Taubenböck, H. (2025). Geospatiality: the effect of topics on the presence of geolocation in English text data. International Journal of Geographical Information Science, 1–32. https://doi.org/10.1080/13658816.2025.2460051

Link to our article

DLR Deutsches Zentrum für Luft- und Raumfahrt