Comparing Geospatiality of Topics between Geotag- and Geoparsing-based Geolocations

Johannes Mast[1*], Richard Lemoine-Rodríguez[1,3], Vanessa Rittlinger[1], Christian Geiß[1,4], and Hannes Taubenböck[1,2,3]

[1] *German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Münchener Straße 20, 82234 Weßling, Germany;*

[2] *Institute of Geography and Geology, Chair of Global Urbanization and Remote Sensing, University of Würzburg, Am Hubland, 97074 Würzburg, Germany;*

[3] *Geolingual Studies Team, University of Würzburg, Am Hubland, 97074 Würzburg, Germany;*

[4] *Department of Geography, Chair of Georisk Research with Remote Sensing Methods, University of Bonn, Meckenheimer Allee 166, 53115 Bonn, Germany*

Abstract

Geolocated social media data offers the opportunity to analyze text data spatially in a wide variety of contexts. Previous work has identified that the likelihood of texts to contain mentions of locations varies between topics, indicating differences in their geospatiality. Social media posts can be linked to geographic locations through two main approaches: geoparsing, which extracts geographic information for places mentioned in the text, and geotags, corresponding to geographic coordinates explicitly attached to posts. In this study, we examine a curated data set of both geotagged and non-geotagged tweets for several thousand of Nigerian Twitter users, to explore differences between geotagging-based and geoparsing-based geolocation approaches in topic representation, controlling for the effects of users and time. Our findings indicate that the two approaches yield data with similar proportions of location mentions, but the interaction between topic and geospatiality varies substantially for some topics. We conclude that the method chosen to geolocate social media data can impact the number of geolocated posts differently across topics. This should be considered in research involving the identification of geolocations from social media posts.

Keywords

geographic information extraction, geoparsing, nlp, social media

1. Introduction

Geolocated text data enables the application of spatial analysis methods and it is valuable in the study of several topics across multiple scientific fields [1], [2], [3], [4], [5]. One form of geolocated text data are geotagged social media texts. The explicitly attached geocoordinates allow researchers to directly query geographic text data via APIs. However, geotags are only available on some platforms, such as Instagram or Twitter (now X), which only represent a small part of web data and whose APIs are often restricted and not free to use. As an alternative, geoparsing allows to geolocate texts based on mentions of locations within them [6], [7]. Alongside efforts to create free and open web indices, such as pursued by the OpenSearch Initiative [8], [9], [10], [11], geoparsing approaches have the potential to unlock a much larger variety of text data sources and contribute substantially to open and reproducible research on geographic topics [5]. However, previous work [11] has shown that the likelihood of texts to contain geoparsable geoinformation (their "geospatiality") varies depending on the texts' topic. This affects geodata availability and can introduce biases. While previous work analyzed this based on a geoparsing approach [6], [12], differences in the geospatiality of topics discussed in posts including geotags (i.e., geographic coordinates) have not yet been explored. Therefore, it is not well understood to what extend geoparsing and geotagging approaches yield similar geolocated datasets, especially regarding the topics they contain. In this study, we aim to address this research gap and analyze whether topical geospatiality differs between geoparsed and geotagged text data. Concretely, using Twitter posts (tweets), we seek to answer the following research questions:

GeoExT 2025: Third International Workshop on Geographic Information Extraction from Texts at ECIR 2025, April 10, 2025, Lucca, Italy

* Corresponding Author

Johannes.mast@dlr.de (Johannes Mast) Richard.lemoine-rodriguez@uni-wuerzburg.de (Richard Lemoine-Rodríguez) Vanessa.rittlinger@dlr.de (Vanessa Rittlinger) Christian.geiss@dlr.de (Christian Geiß) Hannes.taubenboeck@dlr.de (Hannes Taubenböck)

RQ1: Do geotagged and non-geotagged tweets differ regarding the frequency of geoparsable locations within their texts, and to what extent is this affected by the topic of the texts?

RQ2: Does the effect of topical geospatiality vary between geoparsing-based and geotag-based approaches to data selection?

2. Data and methods

2.1. Data

We used data from Twitter (now X), a microblog platform whose salience and accessibility have made it a popular data source in a wide variety of research fields [1]. Twitter offers a geotagging feature which allows users to explicitly link their posts to a location, including geographic coordinates [13]. In a previous study [4], several thousands of stationary Twitter users (i.e., users who did not relocate to another country) from Nigeria were identified based on their timelines of geolocated tweets. We queried and selected both geotagged and non-geotagged tweets posted by those users from official Twitter clients (excluding third-party applications) for 48 distinct and randomly spaced one-week intervals between 2015 and 2019. For every user and week, their tweets were used only if the user produced both geolocated and non-geolocated tweets during a given week. By focusing on users which were stationary in Nigeria, we ensured also for their non-geolocated tweets that they were likely produced in the same country. The geotagged and non-geotagged datasets are thus comparable.

2.2. Topic classification and geoparsing

To assign tweets to topics, we trained a transformer-based text classification model on data from a Nigerian web forum Nairaland using the domain adaptation approach described in [4]. Tweets shorter than 10 tokens were excluded due to their potential limited thematic context. We merged the 42 topics, which were derived from the structure of Nairaland, into 17 overarching supertopics (see Figure 2) as well as an *Other* category capturing generic or unassigned content.

To identify geolocations within texts of both geotagged and non-geotagged tweets, we used an ensemble of four named entity recognition (NER) models to identify entities of the GPE (geopolitical entities), LOC (non-GPE locations), and FAC (facilities) types, and used the Geonames API for geocoding [14]. We considered a text geolocated if at least two models detected a spatial entity in it that could be geocoded to a real geographic location. For the ensemble, we included widely-used state of the art NER models: flair-ner-english-ontonotes-large [9], SpacyNER [15], and bert-base-NER [7], as well as masakhaNER, a NER model which was optimized for several languages of Africa [16].

2.3. Analytical approach

Based on our filtered and classified data, we performed two experiments on our geotagged and non-geotagged set of tweets: Firstly, for each dataset we quantified the frequency of tweets containing at least one geoparsed entity $Frac_{Geo}$ across topics and datasets.

Secondly, we modeled the probability of geoinformation as a function of the topic and geolocation type, controlling for effects of user, time, and text length in a mixed modeling approach with the presence of geolocation (geoparsed or geotagged) as the binary response variable and using the *Other* category as a reference class. This yielded coefficients in the form of log-odds ratios for each topic which can be interpreted as indicators for the topics' geospatiality.

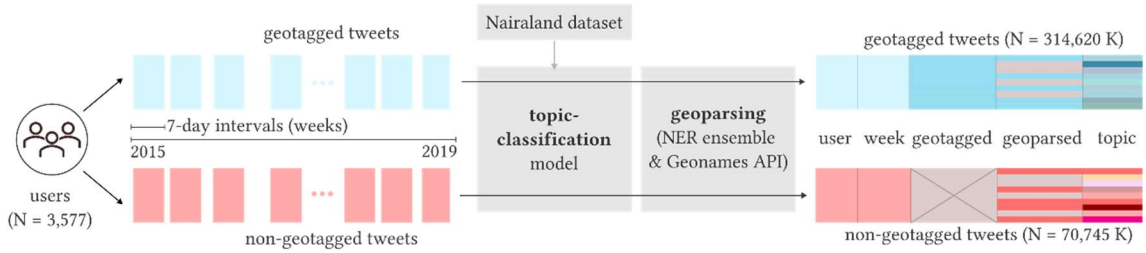

Figure 1: Derivation of the two datasets of geotagged and non-geotagged tweets.

3. Results

3.1. Frequencies of geoparsed entities in geotagged and non-geotagged tweets

The fraction of tweets that contained geoparsed locations was slightly higher in the geotagged tweets (12.5% or 39,346) than in the non-geotagged tweets (11.2%, 7,891). Looking at the former, $Frac_{Geo}$ varied strongly depending on the topic, from 33% for *International Politics* to 3% for *Private Life, Family & Relationships* (Figure 2). Altogether, $Frac_{Geo}$ was similar between the two datasets for most topics, but with two notable exceptions: Geotagged *Advert* tweets contained far more textual spatial references than the non-geotagged *Advert* tweets, and the inverse was found for *Travel, Tourism & Migration*: Here, geotagged tweets contained relatively fewer spatial references.

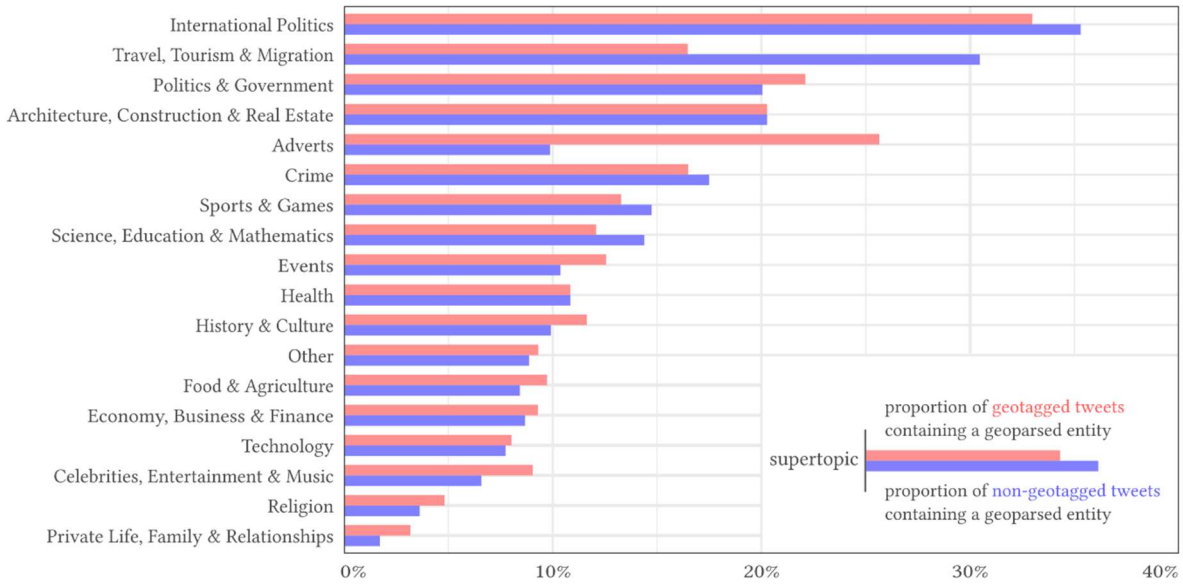

Figure 2: Comparison of $Frac_{Geo}$ between geotagged tweets (indicated by red bars, N=314,620) and non-geotagged tweets (indicated by blue bars, N=70,745) for the 17 supertopics.

3.2. Differences between topical geospatiality in geocoding and geotagging approaches to geolocation

We modeled the geospatiality of topics for both a geoparsing and geotagging approach to geolocation. As a first observation, the geoparsing approach yielded much higher coefficients for most topics than the geotagging approach (Figure 3). While for the geotagging approach, the effect was still significant ($p<0.05$) for 11 of the 17 topics, the modeled log-odds were generally lower than in the geocoding approach with the notable exception of *Travel, Tourism and Migration*, which showed strong geospatiality in both approaches. Notably, the two approaches did not only show differences in magnitude, but also in sign: *International Politics* showed a strong positive effect on the likelihood of geocodeable entities $Frac_{Geo}$, but a slight negative effect on geotagging frequency.

The correlation between the two rankings was moderate and non-significant (spearman ρ = 0.45, p-value = 0.073).

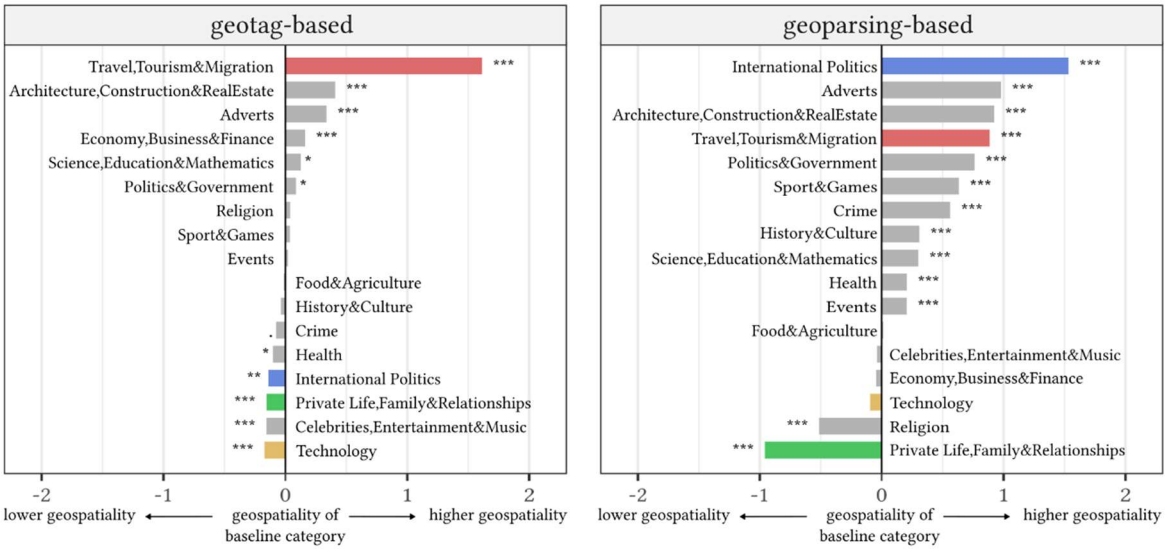

Figure 3: Modeled effect (log odds ratio) of topic on the likelihood of containing geotags (left) and geoparsed entities (right). Positive numbers indicate higher odds compared to the baseline category.

4. Discussion and outlook

Our results indicate that geocodeable entities appeared in similar frequency in geotagged and non-geotagged tweets. However, our findings suggests that this cannot be presumed to be true for every topic. For example, in the topic *Travel, Tourism and Migration*, texts contained place names more rarely in geotagged than in non-geotagged tweets. There could be a possibility that this topic covers a complex and diverse semantic field, and that geotagged tweets cover a different subset of this field (e.g., posts about visiting prestigious locations without mentioning them) than non-geotagged tweets (e.g., travel plans).

Results of our mixed modeling analysis suggest that the geospatiality of topics differs between geotagging and geoparsing approaches. The effect of topics is much higher in the geoparsing approach – plausible, since in this approach, location and topic are both derived from the same (usually short) texts. Compared to this, the lower modeled coefficients in the geotagging approach seem to indicate some detachment between topic and identified location, although topical effects remain and for some topics even show different results than the geoparsing approach. In conjunction, our findings suggest that different approaches to geographic information extraction lead to different representations of topics within the extracted information.

Consequently, researchers using topics as a means to structure and analyze data should consider the impact of their geolocation method. Future assessments should expand to include other text data types, such as news media articles and web forums. This contributes to emerging initiatives aiming to effectively integrate diverse text sources for applications where identifying geolocation is key [8] and advances openness, transparency, and reproducibility in the associated scientific disciplines.

Acknowledgements

This study was conducted as part of the project MIGRAWARE (Grant No. 01LG2082C), funded by the German Federal Ministry of Education and Research program WASCAL WRAP 2.0 and partially funded by the projects OpenSearch@DLR phase II (internal DLR project) and “A New Focus in English Linguistics: Geolingual Studies”, funded by the Volkswagen Foundation (Grant No. 98 662).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Karami, M. Lundy, F. Webb, and Y. K. Dwivedi, 'Twitter and research: A systematic literature review through text mining', *IEEE access*, vol. 8, pp. 67698–67717, 2020.

[2] R. Lemoine-Rodríguez, C. Biewer, and H. Taubenböck, 'Can Social Media Data Help to Understand the Socio-spatial Heterogeneity of the Interests and Concerns of Urban Citizens? A Twitter Data Assessment for Mexico City', in *Recent Developments in Geospatial Information Sciences*, H. Carlos-Martinez, R. Tapia-McClung, D. A. Moctezuma-Ochoa, and A. J. Alegre-Mondragón, Eds., Cham: Springer Nature Switzerland, 2024, pp. 119–133. doi: 10.1007/978-3-031-61440-8_10.

[3] R. Lemoine-Rodríguez, J. Mast, M. Mühlbauer, N. Mandery, C. Biewer, and H. Taubenböck, 'The voices of the displaced: Mobility and Twitter conversations of migrants of Ukraine in 2022', *Information Processing & Management*, vol. 61, p. 103670, Jan. 2024, doi: 10.1016/j.ipm.2024.103670.

[4] J. Mast, M. Sapena, M. Mühlbauer, C. Biewer, and H. Taubenböck, 'The migrant perspective: Measuring migrants' movements and interests using geolocated tweets', *Population Space and Place*, vol. 30, no. 2, p. e2732, Mar. 2024, doi: 10.1002/psp.2732.

[5] H. Senaratne *et al.*, 'The Unseen—an investigative analysis of thematic and spatial coverage of news on the ongoing refugee crisis in West Africa', *ISPRS International Journal of Geo-Information*, vol. 12, no. 4, p. 175, 2023.

[6] X. Hu *et al.*, 'Location Reference Recognition from Texts: A Survey and Comparison', *ACM Comput. Surv.*, vol. 56, no. 5, pp. 1–37, 2023, doi: 10.1145/3625819.

[7] Y. Hu and R.-Q. Wang, 'Understanding the removal of precise geotagging in tweets', *Nat Hum Behav*, vol. 4, no. 12, pp. 1219–1221, Sep. 2020, doi: 10.1038/s41562-020-00949-x.

[8] Open Search Foundation e.V., 'Home', Open Search Foundation. Accessed: Jan. 25, 2025. [Online]. Available: https://opensearchfoundation.org/en/

[9] M. Granitzer *et al.*, 'Impact and development of an Open Web Index for open web search', *Journal of the Association for Information Science and Technology*, vol. 75, no. 5, pp. 512–520, 2024, doi: 10.1002/asi.24818.

[10] O. W. I. Open Search Foundation e.V., 'The Open Web Index Dashboard'. Accessed: Jan. 28, 2025. [Online]. Available: https://openwebindex.eu/%PUBLIC_URL%

[11] Open Search Foundation e.V. and C. Plote, 'Welcome', Open Web Search – Promoting Europe's Independence in Web Search. Accessed: Jan. 28, 2025. [Online]. Available: https://openwebsearch.eu/welcome/

[12] J. Mast *et al.*, 'Geospatiality: The effect of topics on the presence of geolocation in English text data', *International Journal of Geographical Information Science*, 2025, doi: http://dx.doi.org/10.1080/13658816.2025.2460051.

[13] X. X. Zhu *et al.*, 'Geoinformation Harvesting From Social Media Data: A community remote sensing approach', *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 4, pp. 150–180, Dec. 2022, doi: 10.1109/MGRS.2022.3219584.

[14] 'GeoNames'. Accessed: Nov. 09, 2022. [Online]. Available: https://www.geonames.org/

[15] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, 'spaCy: Industrial-strength Natural Language Processing in Python', 2020, doi: 10.5281/zenodo.1212303.

[16] D. I. Adelani *et al.*, 'MasakhaNER: Named Entity Recognition for African Languages', *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1116–1131, 2021, doi: 10.1162/tacl_a_00416.