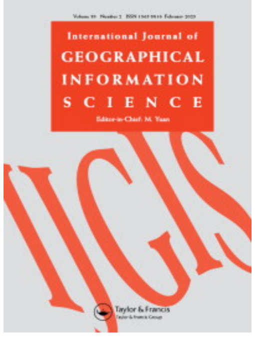


International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Geospatiality: the effect of topics on the presence of geolocation in English text data

Johannes Mast, Richard Lemoine-Rodríguez, Vanessa Rittlinger, Martin Mühlbauer, Carolin Biewer, Christian Geiß & Hannes Taubenböck

To cite this article: Johannes Mast, Richard Lemoine-Rodríguez, Vanessa Rittlinger, Martin Mühlbauer, Carolin Biewer, Christian Geiß & Hannes Taubenböck (14 Feb 2025): Geospatiality: the effect of topics on the presence of geolocation in English text data, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2460051

To link to this article: https://doi.org/10.1080/13658816.2025.2460051

Published online: 14 Feb 2025.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2460051

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

Geospatiality: the effect of topics on the presence of geolocation in English text data

Johannes Mast[a], Richard Lemoine-Rodríguez[b,c], Vanessa Rittlinger[a], Martin Mühlbauer[a], Carolin Biewer[c,d], Christian Geiß[a,e] and Hannes Taubenböck[a,b,c]

[a]German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany; [b]Institute of Geography and Geology, Chair of Global Urbanization and Remote Sensing, University of Würzburg, Würzburg, Germany; [c]Geolingual Studies Team, University of Würzburg, Würzburg, Germany; [d]Department of English and American Studies, Chair of English Linguistics, University of Würzburg, Würzburg, Germany; [e]Department of Geography, Chair of Georisk Research with Remote Sensing Methods, University of Bonn, Bonn, Germany

ABSTRACT

Geolocated text data are a promising data source for spatial analyses in many fields, from disease surveillance to the spatial humanities. This study investigates the relationship between texts' thematic categories and their likelihood of containing usable geolocation information by quantifying and modelling this relationship across seven diverse English text datasets of different types, including web forums, microblogs, news, and magazines. We find that the likelihood of geoinformation is highly variant, being high for the category 'Travel, Tourism & Migration' and low for 'Private Life, Family & Relationships'. The rank-correlation of this likelihood between datasets is moderate to strong. These findings indicate that the topic plays a significant role in determining the frequency of geospatial references within the text, and that the effect is not entirely dataset-specific. This contributes to the empirical study of the concept of spatiality and provides valuable insights for bias mitigation in the increasing use of text as data for spatial analyses.

ARTICLE HISTORY
Received 14 October 2024
Accepted 24 January 2025

KEYWORDS
Geocoding; NLP; geography; web data; NER

1. Introduction

Over the past decades, many scientific disciplines have experienced an increased interest in the role of space and place in what is often referred to as a 'spatial turn' (Tally 2012). Spatial methods are increasingly used to study the geographic distribution of physical and social phenomena. The empirical side of the spatial turn is supported by an increasing amount of algorithmic power, advanced spatial analysis techniques, and, perhaps most importantly, digital data (Zhu *et al.* 2022). Of the available data, a

CONTACT Johannes Mast Johannes.mast@dlr.de

substantial amount are texts from online communication and media. When these texts contain references to places on earth, analysts can link text content and geographic location in a process called geoparsing (Hu *et al.* 2023) and enable spatial approaches in diverse fields of research and various applications (Karami *et al.* 2021, Zhu *et al.* 2022, Hu *et al.* 2023). Text data that have been geospatially referenced in this way can be used to complement other data sources in geographic information databases (Zhu *et al.* 2022), allow to identify hotspots and coldspots of activity (Taubenböck *et al.* 2018), and analyse which topics are prevalent in a district (Lansley and Longley 2016, Lemoine-Rodríguez *et al.* 2024). Advances in machine learning and natural language processing may enable additional applications for geospatially located text data.

However, not all data can be geospatially located. In other words, geospatially located datasets miss a part of data. But are these data missing at random (Rubin 1976)? Or are some pieces of content less likely to be missing than others, because they have a stronger connection to the geospatial world, a connection that could be called their *geospatiality*?

It is intuitive that some activities, like reading, can be performed virtually anywhere, while others, such as air travel, involve specific locations and even movement between them. One could describe this as a difference in geospatiality: Reading may be an inherently less spatial activity than travel. In other words: Our lives are spatial, but not every aspect of life is equally spatial. Such variability in geospatiality would likely be reflected in written communication, in the form that, for example, texts about travel contain more mentions of geographic locations than texts about literature. This would lead to varying suitability of the texts for spatial analyses, as spatial methods can only be applied to data which contains some geospatial reference. Therefore, variability in geospatiality would lead to a variability in data suitability and could affect analyses as a form of bias. Looking through the lens of geospatially located web data, we might mistakenly believe a place to be abandoned, when it is actually a thriving place of activity where activity is merely not geospatial. An example could be a library which many people visit to read and share quotes and insights from their favourite books online without mentioning the library itself. A tourist attraction, on the other hand, will be mentioned by many of the people who visit it. Consequently, studies relying on georeferenced text data to inform about land use may be blind towards certain land use types and hotspot analyses of digital activity within a city may underestimate the importance of stores compared to stadiums and libraries compared to landmarks. When known, such effects can be accounted for.

However, while differences in geospatiality are intuitive, their existence in web data and their effect on topic-based analyses have not been specifically studied.

Therefore, we seek to provide an exploration of variability in geospatiality across topics and web data sources such as web forums, news sites, and microblogs. In particular, we focus on thematic geospatiality, the affinity of topics to contain references to places on Earth, and examine text data in the English language; with its being used on over 43% of all websites, English remains the most widespread language on the web (Common Crawl 2024).

2. Background

Geospatially referenced text data have been analysed in a wide range of fields (Karami *et al.* 2021, Zhu *et al.* 2022). Key to the application of geospatial techniques to data is some form of geolocation, i.e. coordinates. In some sources of digital data, such as Twitter, geolocations can be explicitly attached in the form of geotags, which can refer to an area or even precise geocoordinates (Zhu *et al.* 2022). Alternatively, it can be inferred from locations (e.g. place names or addresses) mentioned in the text, a process typically referred to as geoparsing.

Regardless of the method, studies have shown that the text data which can be geolocated are only a fraction of all text data (Olteanu *et al.* 2019, Zhu *et al.* 2022). It is all the more important to understand how well this subset represents text data and online discourse as a whole, and to what degree it is affected by various biases. Olteanu *et al.* (2019) define biases as systematic distortions in sampled data that compromises their representativeness. Knowledge about such biases improves our understanding of the data's validity and enables us to make more accurate comparisons with other datasets or even post-hoc adjustments (Sen *et al.* 2021). Consequently, demographic biases in geolocated data have been the focus of several studies. For Twitter (now 'X') there seems to be a consensus that geolocated Twitter data do not equally represent all demographic groups (Longley *et al.* 2015, Malik *et al.* 2015, Pavalanathan and Eisenstein 2015) or all geographic areas (Hecht and Stephens 2014, Malik *et al.* 2015, Jiang *et al.* 2019). In addition to such demographic biases in the Twitter user base, Karami *et al.* (2021) also find differences between those users who use geolocation features and those who do not. Notably, the two groups show differences in linguistic choices and topics of interests.

Altogether, biases of geolocated data have been studied at the level of users and population groups. What remains underexplored as a bias factor, however, is the content of the texts themselves. Are certain types of content underrepresented in text data sources? To the best of our knowledge, only the study by J. Jiang *et al.* (2023) investigated a direct relation between content and geolocation for Tweets, and measured that compared to a random sample of Tweets, geolocated Tweets exhibit higher use of first-person pronouns and focus on positive events. Although some of these findings are in seeming contrast with earlier findings at user level (Karami *et al.* 2021), they suggest that positivity and collectivism might be overrepresented in geolocated Tweets. So far unexplored are topics, which are a useful explorative and analytical tool for structuring data semantically. In natural language processing, topics are usually measured as clusters of thematically similar content which are often marked with a representative label, such as 'sports', 'economy', 'disasters', or 'philosophy'. Some datasets are already pre-structured (such as web-forums, Mast *et al.* 2024) while others can be structured with machine learning or rule-based classification approaches (e.g. Grootendorst 2022). Acknowledging the usefulness of topics for structuring data, we suspect that variation in their geospatiality leads to unequal representation in georeferenced text data.

For instance, natural disasters have a clear association with real locations where disasters have occurred or where disaster response is coordinated (Kersten and Klan 2020). The same is not true for the field of epistemology, a branch of philosophy which concerns itself with location-independent theories about the nature of knowledge (Steup and Neta 2005). In practice, this might mean that content about location-independent topics is less likely to contain geographical reference, whether explicitly (e.g. geotags) or implicitly (mentions of locations). Such content will be underrepresented in datasets which are compiled or filtered based on geolocations. Consequently, differences in geospatiality would constitute an additional content production bias which needs to be considered in the analysis of geolocated text data. The existence of such a phenomenon is also indicated by scientific work on optimization of web searches, where studies found links between the thematic content of texts and their geographical component (Gan *et al.* 2008, Jiang *et al.* 2019).

Motivated by the potential of spatial analyses of text data and following the intuition that differences in geospatiality exist between different aspects of human life, we aim to extend the current state of research and focus on topical areas as a source of bias. In other words, we aim to analyse variation in geospatiality according to topic.

We define geospatiality as the likelihood to contain an identifiable geospatial reference. By 'identifiable' we mean that current approaches in geoparsing can recognize them as referring to real place on Earth that can be described by geographic coordinates. Therefore, this definition does not consider geotags and is practical rather than formal. In other words, we relay the decision of what constitutes a valid geospatial reference to the creators of the gazetteers (geographical dictionaries), of the models, and the data used to train them. Defined in this way, geospatiality can be seen as a subset of spatiality, which includes spatial references in the wider sense, such as real places not on earth (e.g. Deimos, a moon of Mars), fictional places (e.g. 'Hogwarts'), relative spatial references such as deixis (e.g. 'over there'), spatial terms used in other contexts (e.g. 'political left'), demonyms (e.g. 'Germans'), organisations linked to locations (e.g. 'Indian National Team'), and things named after places (e.g. 'French Toast').

While both intuition and the existing work suggest that topics vary in geospatiality, it has, to our knowledge, not yet been analysed systematically or across text types. Further, it has, to our knowledge, not been subject of an analysis at the topic-level. This work seeks to close this research gap by analysing the relation between content and geospatiality across a variety of text types from various platforms.

Concretely, we sought to answer the following research questions:

RQ 1: Does the observed frequency of geolocations vary between topics in typical sources of English text web data?

RQ 2: To what degree can this variation in georeferencing frequency be attributed to the topics themselves?

RQ 3: Are there differences between various forms of web data?

To answer these questions, we analysed seven datasets of English web texts and identified both topics and spatial references, wherever possible. We computed the fraction of spatial references in documents from each topic, and firstly present descriptive statistics that allow for an answer to RQ 1. Secondly, to answer RQ 2, we identified the contribution of topics by controlling for effects of authorship, time, and text length in a mixed modelling approach. Finally, we address RQ 3 by comparing the rank-correlation of topics between platforms.

3. Methods

3.1. Data

Web data are vast, varied and constantly evolving. Instead of attempting to create a representative sample of all web data, we adopt a practical approach by selecting data from various corpora that were previously used in research: Reddit, Nairaland, Twitter, Stackexchange, and The Global Database of Events, Language, and Tone Project's collections of web news (GDELT) and of American texts from the Internet Archive (IA-Americana). Each corpus consists of a different type of text: character-limited texts on Twitter, forum comments on Nairaland, Reddit, and Stackexchange, articles on news websites from the GDELT project, and magazines and official documents in IA-Americana. We selected content in English, the most widely used and analysed language on the web (Common Crawl 2024). Thus, our sample, while not a representative sample of web data as a whole, exemplifies data that are practically available for use by researchers and it allows an examination of the consistency of geospatiality effects across text types.

We collected control variables to control for population biases which might affect the measured frequency of geolocated content. For instance, cosmopolitan and wealthy users might be more active than average users, use more geolocation and post more frequently about their travels. Likewise, there might be temporal biases (e.g. shifts in user behaviour and platform functionality over time). We acknowledge them as a natural effect in our descriptive approach to RQ 1. However, when attempting to extract the underlying contribution of the topic (RQ 2) we need to account and control for such effects. Thus, we collected for each observation also their author (e.g. username, web domain), creation time (discrete intervals, either calendar month or year), and text length. In addition, all documents were assigned (A) a topic which described the content and (B) a binary label that indicated whether the document contained geolocation information. Figure 1 illustrates the workflow of the study and Table 1 illustrates the structure of the preprocessed data as a fictive example.

3.2. Topic taxonomy

To analyse and compare the effect of topics across platforms, we structured all datasets into the same 18 topics, developing a generic topic taxonomy in an inductive coding process that was based on text content and pre-existing categories within the datasets. This taxonomy is just one among an infinity of possible ways to structure content and is intended to represent the thematic variety inherent in the datasets. Typically, studies define topics to suit their respective application, and it is not our goal to derive a universally useful taxonomy. What is important for the purposes of the present analysis, however, is that it is human-interpretable and consistent across data sources. We proceeded as follows: First, a topic-taxonomy was defined by the

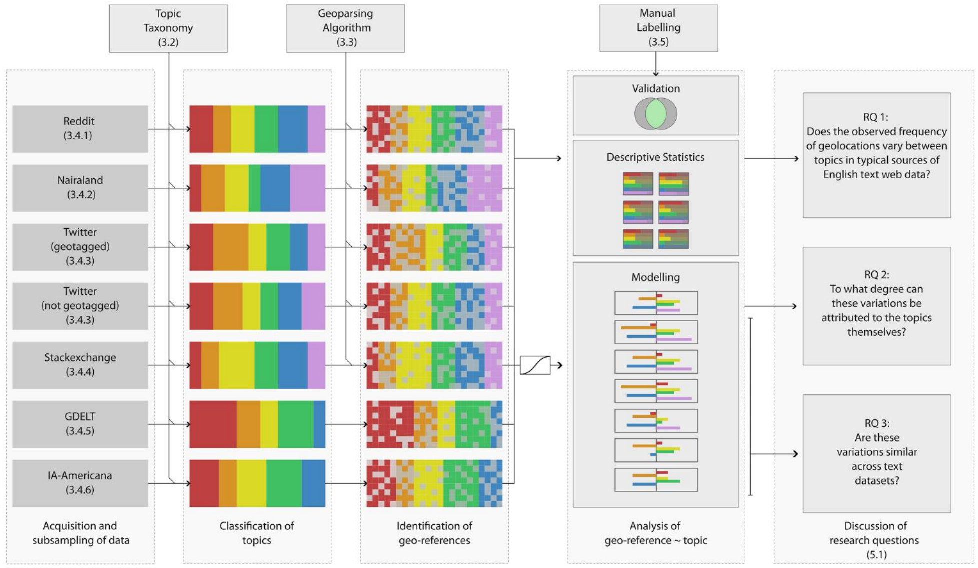

Topic Taxonomy (3.2)
Geoparsing Algorithm (3.3)
Manual Labelling (3.5)
Reddit (3.4.1)
Nairaland (3.4.2)
Twitter (geotagged) (3.4.3)
Twitter (not geotagged) (3.4.3)
Stackexchange (3.4.4)
GDELT (3.4.5)
IA-Americana (3.4.6)
Acquisition and subsampling of data
Classification of topics
Identification of geo-references
Validation
Descriptive Statistics
Modelling
Analysis of geo-reference ~ topic
RQ 1: Does the observed frequency of geolocations vary between topics in typical sources of English text web data?
RQ 2: To what degree can these variations be attributed to the topics themselves?
RQ 3: Are these variations similar across text datasets?
Discussion of research questions (5.1)

Figure 1. Workflow of the study.

Table 1. Fictive example of data structure.

ID	Dataset	Author	Time	Text length	Geolocated	Topic
GD27	GDELT	abcd.com	month 2017 08	271	FALSE	Health
GD28	GDELT	abcd.com	month 2017 08	824	FALSE	Health
GD29	GDELT	abcd.com	month 2017 09	434	TRUE	Events
GD30	GDELT	abcd.com	month 2017 09	3006	FALSE	Religion
GD31	GDELT	zyxw.de	month 2015 01	579	TRUE	Religion
…	…	…	…	…	…	…
TW731	Twitter_no_tag	Johndoe1990	month 2016 02	131	FALSE	Events

main authors with the aim of capturing the major pre-existing categories from each dataset (Figure 2(a) and Appendix Table A8). This resulted in a taxonomy containing 18 topics: *Adverts*; *Architecture, Construction & Real Estate*; *Celebrities, Entertainment & Music*; *Crime*; *Economy, Business & Finance*; *Events; Food & Agriculture*; *Health*; *History & Culture*; *International Politics*; *Natural Disasters and Hazards; Politics & Government; Private Life*; *Family & Relationships*; *Religion*; *Science, Education & Mathematics*; *Sports & Games*; *Technology*; *Travel, Tourism & Migration*. Additionally, a category ‘*Other*’ captured unassigned, general or thematically ambiguous content (e.g., jokes) and served as a reference class. We assigned each document to one of these topics, building on site-specific category structure where one existed (such as subreddits on Reddit, Figure 2(b)), meaning different datasets were processed differently. The assignment of dataset-specific categories (subreddits, Stackexchange sites, GDELT themes, Nairaland subforums, see Section 3.4. for details for each dataset) to topics was proposed by the main author and subsequently checked for plausibility and consistency by three otherwise uninvolved researchers (graduate and postgraduate researchers from the fields of Geography, Digital Humanities, and English language studies). Based on this manual qualitative assessment, the taxonomy was modified if two reviewers or one reviewer and the main author agreed on a change. Not all sources contained all topics. For

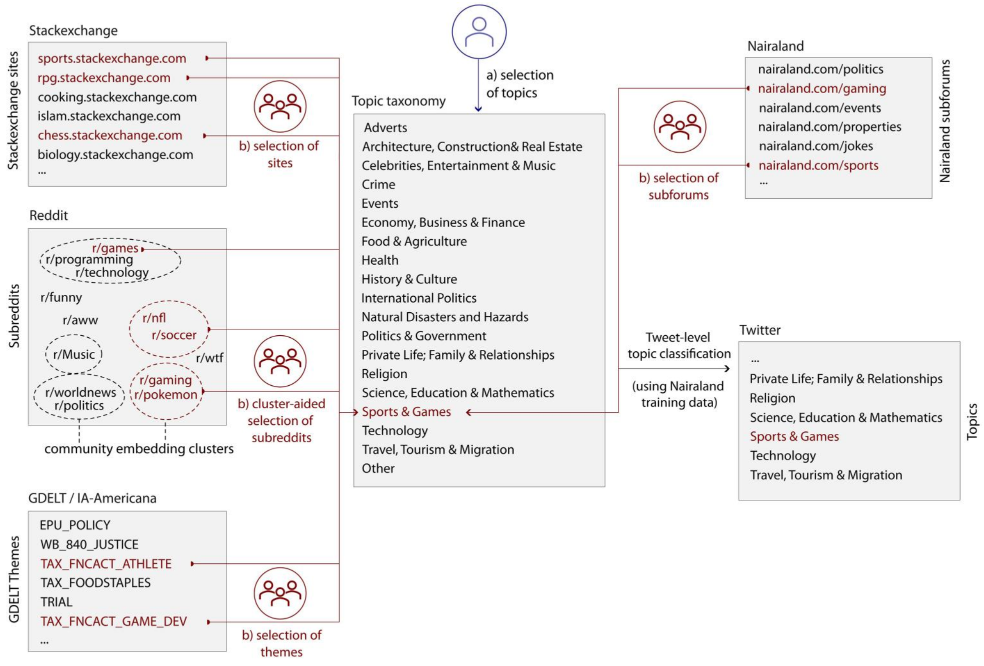


Figure 2. Development of the topic-taxonomy. First, topics were chosen (a), then dataset-specific categories from the data sources (e.g. GDELT themes from GDELT) were allocated (b) to the topics.

example, the GDELT project explicitly filters strictly entertainment or economic content and there is no advert-related Stackexchange site (although self-promotion is acceptable as long as it is related to the topic[1]). A description of the taxonomy can be found in the supplementary material to this study.[2]

3.3. Geolocation

For all datasets except GDELT and IA-Americana, to identify geolocations within the posts, 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. We considered a text geolocated if at least two models detected a spatial entity within them that could be geocoded to a real place on earth.

For the ensemble, we included widely used state of the art NER models: bert-base-NER (Devlin *et al.* 2018), SpacyNER (Honnibal *et al.* 2020), and flair-ner-english-ontonotes-large (Schweter and Akbik 2021). As three of our datasets are geographically focused on Nigeria, we also included masakhaNER, a NER model which was optimized for the African context (Adelani *et al.* 2021), to identify locations within the comments.

For geocoding the identified entities, we queried the Geonames API to check for matching coordinates. We did not perform location disambiguation or attempt to precisely locate the texts as we were merely interested in whether they contained any valid geolocation at all.

Table 2. Dataset statistics, including types of observations, sources of information used to assign topics and whether the observation contains location information, and units for author and time.

Dataset	Document type	Topic source	Location source	Author unit	Time unit
WEBIS Reddit	forum comment *($N = 522{,}353$)*	subreddit	ensemble geoparser	username *($N = 113{,}512$)*	- none -
Nairaland	forum comment *($N = 2{,}153{,}680$)*	subforum	ensemble geoparser	username *($N = 99{,}045$)*	month *($N = 80$)*
Twitter (geotagged)	Tweet *($N = 314{,}620$)*	text classification	ensemble geoparser	author id *($N = 3{,}577$)*	month *($N = 35$)*
Twitter (non-geotagged)	Tweet *($N = 70{,}745$)*	text classification	ensemble geoparser	author id *($N = 3{,}577$)*	month *($N = 35$)*
Stackexchange	forum comment *($N = 5{,}671{,}182$)*	site/subdomain	ensemble geoparser	author id *($N = 147{,}895$)*	month *($N = 188$)*
GDELT	news article *($N = 26{,}018{,}682$)*	GDELT-tagger	GDELT geoparser	web address *($N = 10{,}000$)*	month *($N = 35$)*
IA-Americana	text section *($N = 4{,}146{,}089$)*	GDELT-tagger	GDELT geoparser	book author *($N = 2{,}726$)*	year *($N = 10$)*

The data structure is illustrated by the fictive example data in Table 1, while descriptive statistics for the compiled datasets are presented in Table 2.

3.4. Datasets

In this section, we introduce the different datasets. Due to the datasets' heterogeneity, each required different preprocessing steps which we devised to achieve comparability in the key variables (topic and geolocation) while maintaining the characteristics of the data sources. For example, we allocated content from all datasets to the same topic taxonomy, but used pre-existing structure where possible (e.g. subreddits on Reddit), rather than applying a single topic classification algorithm to all datasets. Likewise, we did not enforce a consistent text length within our datasets, for example, by excluding news articles whose length exceeds those of Tweets (280 characters at most). Instead, we include text length as a variable in our models. Time was discretized to either calendar months or years. Because these time intervals were used as a control variable and not analysed themselves, this coarse granularity of intervals ensured sufficient observations per interval. Of course, our choices in preprocessing represent only one among many approaches. The datasets and the preprocessing will now be described in detail for each dataset.

3.4.1. WEBIS Reddit corpus (Reddit)

Reddit is a news-aggregator and discussion website that is structured into many community-moderated subforums called subreddits, which are typically dedicated to certain topics, types of content, communities, or locations (e.g. Boston). Due to its large and diverse content, it has been extensively studied and used as a data source for research and machine learning, as well as for geographic information (Fox *et al.* 2021, Berragan *et al.* 2022). In this study, we use the *WEBIStldr-17* Reddit corpus, which was compiled by Völske *et al.* (2017) for the purpose of training models for automatic text summarization. This corpus has the advantage of containing relatively long and content-rich comments. We assigned comments to topics based on the subreddit they are posted in, a challenging task due to their large number. Therefore, we supported the manual assignment by using the Reddit community embeddings produced by Partridge *et al.* (2024) which clustered subreddits into clusters based on the connectedness of users posting within them. We built on this work by assigning clusters to topics, manually adding or removing subreddits from clusters to improve semantic consistency.

3.4.2. Nairaland

Nairaland is a text-based Nigerian Web Forum which has been studied in the context of politics (Nwachukwu 2015), cybercrime (Lamidi 2020), online humor (Lamidi 2016), discourse on terrorism (Chiluwa and Odebunmi 2016), health (Oyebode and Orji 2019), and migration (Mast *et al.* 2024). Nairaland is intended as a platform for Nigerian users and the predominant language on Nairaland is the Nigerian variety of English. In the context of this study, Nairaland is an example of a geographically focused community with distinct linguistic and cultural elements that might affect the way thematic geospatiality is expressed. Nairaland is structured into several dozens of subforums dedicated to topics such as Politics, Religion, and Travel. By deciding which subforum to post their comment in, users implicitly choose a thematic label for their texts. Therefore, the collective Nairaland corpus can be considered to be structured by the community into community-defined clusters. Randomly sampling threads created on this website from 2014 to 2021, we obtained several millions of comments, including timestamps, usernames and subforum names (Mast *et al.* 2024). We group the subforums into the 19 topics of the taxonomy.

3.4.3. Twitter – geotagged and non-geotagged

Twitter (now 'X') is a microblog platform whose salience and accessibility have made it a popular data source in a wide variety of research fields (Karami *et al.* 2020). Twitter offers a geotagging feature which allows users to explicitly attach a location to their texts (Zhu *et al.* 2022). In a previous study (Mast *et al.* 2024), distinguished stationary and migratory Twitter users from Nigeria based on their timelines of geolocated Tweets. For several thousands of these users, Tweets were queried for 48 distinct and randomly spaced one-week intervals between 2015 and 2019. From this dataset, we selected Tweets without geolocations which were posted via official apps by stationary users whose residence was within Nigeria for the entire studied timespan (Twitter non-geotagged). Additionally, we queried geotagged Tweets which were posted within Nigeria by the same users via official apps during the same 48 weeks (Twitter geotagged). For every user and week, their Tweets were used only if the user produced both geolocated tweets and non-geolocated tweets during the week. By focusing on users which were stationary in Nigeria we had confidence that almost all of their non-geolocated Tweets were produced in Nigeria, i.e. in the same geographical context as the geolocated Tweets. The geotagged and non-geotagged datasets are thus highly comparable. However, we do not consider them a single dataset as previous research by Serere and Resch (2024) found differences in the use of named entities between geotagged and non-geotagged tweets.

To assign Tweets to topics, we trained a transformer-based (Devlin *et al.* 2018) topic-classification model on the Nairaland dataset (Section 3.2.2) using the domain adaptation approach described in Mast *et al.* (2024) and the *twhin-bert* model (Zhang *et al.* 2023) as a baseline. Tweets were then classified into topics by applying the trained topic model. Tweets shorter than 10 tokens were excluded because classification accuracies of extremely short texts are low.

The mean comment length in these Twitter datasets increases slightly from roughly 105 characters to around 133 over the course of 2018, when the character limit was increased from 140 to 280. The geotagged dataset is substantially larger than the non-geotagged one, despite being derived for the same users and time, indicating that the former represents users who frequently use the geotagging feature.

3.4.4. Stackexchange

Stackexchange (Stack Exchange Inc 2024) is a network of 170+ community question answering sites on topics in diverse fields, such as cooking, music, robotics, and travel. Stackexchange releases its data in the form of dumps (Stack Exchange, Inc 2024) and has been used a data source for social network analysis, algorithmic development, and training in natural language processing (Firouzjaei 2024, Stack Exchange Inc. 2024). We downloaded the dump from April 2024, which includes posts from August 2008 to March 2024, and selected those sites for the analysis which semantically matched topics of the taxonomy. From these sites, we extracted all posts. To limit data volume, we excluded authors who posted less than 5 times, more than 10,000 times, or exclusively on one topic. Stackexchange differs from the other datasets because not all topics are available for the entire timespan. Starting from the initial Stackoverflow site, other Stackexchanges were added over time.

3.4.5. Global database of events, language, and Tone project (GDELT)

The Global Database of Events, Language, and Tone Project (GDELT) is a database containing news articles from a large number of countries of the world and in more than one hundred languages (Leetaru and Schrodt 2013). It constantly monitors a large number of news sources and ingests and processes published articles to extract information about their content, such as locations and themes. The full texts are not made available, but the derived locations and themes are. These themes are thematic labels which may be unique to GDELT or sourced from other taxonomies, for instance, the World Bank's topical taxonomy (World Bank 2015), CrisisLex (Olteanu *et al.* 2014), or the Conflict and Mediation Event Observations Event and Actor Codebook (CAMEO, Gerner *et al.* 2002). While GDELT does not provide details on the algorithm used to identify locations and themes, nor a comprehensive taxonomy of themes (Williams 2020), the themes have informative names and typically a prefix which informs the researcher about the nature and source of the theme (Blanqué *et al.* 2022). For example, the theme name *WB 470 EDUCATION* indicates that it is derived from the World Bank's taxonomy and concerns education. GDELT has been used in studies on conflict (Qiao *et al.* 2017, Blanqué *et al.* 2022, Senaratne *et al.* 2023), public opinion (Bodas-Sagi and Labeaga 2016), urban branding (Zheng 2020), and disasters (Owuor *et al.* 2020), among other research topics (Buckingham *et al.* 2020) and is a valuable resource for computational journalism (El Ouadi and Beskow 2024).

For each topic, we selected matching GDELT themes until all topics were represented by several themes. Not all topics could be represented by GDELT themes because some topics, such as sports and entertainment, are intentionally excluded from the database (Leetaru and Schrodt 2013). An article often contained several themes, making this a multilabel dataset. GDELT also identifies locations in the text, which we used to identify geolocated articles. Due to the immense volume of the dataset, we queried only a sample of news articles, that is, only those published during the 48 weeks covered by the Twitter data. We then selected 10,000 sources (web domain names) for which we had 10 different articles covering at least 10 different months and 10 different topics. From these sources, we used all articles as documents in the analysis. While the true text length is not reported by GDELT, we approximated it by using the highest possible character offset from any GDELT field.

3.4.6. Internet Archive Americana (IA-Americana)

Many documents are openly available in language public domain collections. The GDELT project processed all American books and documents in the Internet Archive's American Libraries collection for which English-language text was available and made the results available on the Google BigQuery platform (Leetaru, 2015). Within all these documents, locations and themes were also identified by the GDELT project as in the GDELT dataset, although with somewhat lower accuracy due to digitization errors and metadata issues (Leetaru 2015). We queried all available documents for the years 2005 to 2014. For this timespan, the available documents mostly consist mostly of periodicals, magazines, and documents issued by government agencies. The substantial length of these documents makes them likely to contain a large number of different themes and locations, that are unlikely to be related over the entire span of the document. Thus, we derived smaller observation units from the documents: For every occurrence of a theme, we checked whether a location had been detected within a 100-character window around it. In other words, the 200 characters around a theme were considered to be an observation which was either geolocated (if a sufficiently precise location was detected) or not. We note that the author field in this data source contains entries with multiple authors (e.g. 'Marx, Karl, 1818–1883; Engels, Friedrich, 1820–1895') and does not necessarily have a clear relationship with an author as a person.

3.5. Manual labelling

We manually labelled a subset of the data to examine how the topics and geolocation assigned by our approach compared to those assigned manually by a human. For this, we selected 1,075 text documents, stratified by datasets and topics, and applied manual interpretation and web research to label them. We report overall accuracies and Cohen's Kappa (κ) (Cohen 1960) as a measure of how the assignment agreed with those of a human expert. Unfortunately, this manual labelling could not be done systematically for IA-Americana; while the full texts are available online, the themes extracted by GDELT are affected by a varying character offset that is not provided in the metadata. Thus, the text window corresponding to each instance of a GDELT theme could not be systematically reproduced for review.

3.6. Analysis and modelling

We analysed the data entered into the analysis (Table 2) first with a descriptive approach and secondly with a modelling approach. In the descriptive approach, we quantified the frequency of geolocations within

1. each dataset and topic,
2. each dataset and timestep,
3. each dataset and author,
4. each dataset and text length interval,

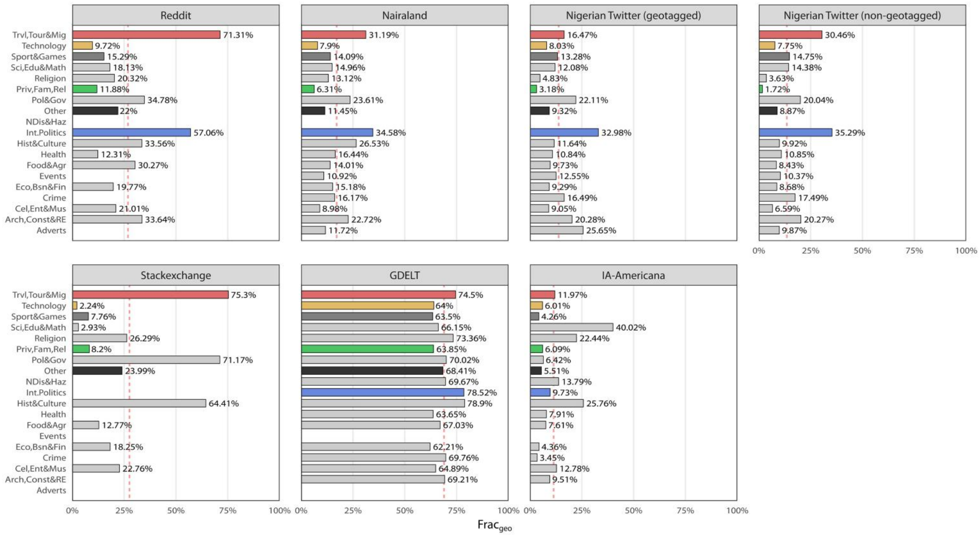


Figure 3. Observed *Frac*$_{geo}$ within the datasets. The dotted vertical line indicates the mean across topics. Several topics are coloured for visualisation purposes.

Table 3. Summary statistics of observed *Frac*$_{geo}$ across topics within each dataset.

Dataset	Min	Max	Mean	σ	Range
GDELT	0.622	0.789	0.687	0.051	0.167
IA-Americana	0.035	0.400	0.116	0.096	0.365
Nairaland	0.063	0.346	0.167	0.079	0.283
Twitter (not geotagged)	0.017	0.353	0.133	0.087	0.336
Twitter (geotagged)	0.032	0.330	0.138	0.075	0.298
Reddit	0.097	0.713	0.274	0.172	0.616
Stackexchange	0.022	0.753	0.280	0.268	0.731

σ: standard deviation over the frequency of all topics. range: differences between the highest frequency topic and lowest frequency topic.

5. by relating the number of geolocated documents to the number of all documents within the group:

$$Frac_{geo} = \frac{N_{geolocated}}{N_{geolocated} + N_{notgeolocated}}$$

This measure allowed us to infer observed differences between topics (RQ 1) as well as describe other effects at the level of datasets, authors, times, and text length. As all these effects partially overlap, the descriptive approach is insufficient to extract topic-specific effects on geolocation frequency.

To disentangle these effects and answer RQ 2, we applied a modelling approach. We based the statistical inference on a mixed model (Gries 2015) with random intercepts for the effects of time and author. We estimated one generalized linear mixed model (GLMM) for each dataset using the glmmTMB package (Brooks *et al.* 2017) in the R programming language (R Core Team 2024). We did not fit a single model over all datasets due to differences in their processing. Instead, we consider the models to be distinct, but comparable, experiments on geospatiality in their respective domains.

All models followed the same basic formula, with presence of geolocation as the binary response variable, fixed effects for text length and the topic as the categorical predictor, with the *Other* category as the reference class. Random intercepts were included to account for the influence of timestep and author. The notable exceptions were the Reddit model, where no effect for timestep was included due to a lack of timestamps in the source data, and the IA-Americana data where we did not include a fixed effect for text length as the text documents were based on a fixed window of 200 characters. The fixed-effects estimates and p-values for each topic and dataset allow an answer to RQ 2. For statistical analyses of the data and for visualizing the results, we used the tidyverse suite of packages (Wickham *et al.* 2019), also implemented in R.

To identify agreement between datasets (RQ 3), we manually interpreted the ranking of geospatiality, using only the pairwise Spearman's rank correlation coefficient ρ (Spearman 1904) as an ancillary statistic. The statistical modelling of dataset-specific effects was not feasible because of the substantial differences in how the datasets were processed and what the topics represent. Therefore, we relied on a manual interpretation of the ranking, considering the properties of the datasets.

4. Results

4.1. Observed frequency of geolocated texts within the datasets

The observed frequency of geolocated documents among all documents gave a first indication of differences between topics (Figure 3). We observed substantial variations within and between datasets. The highest frequencies were found for the GDELT data with 67.9% on average and the topics *History & Culture* and *International Politics* leading with 78.9 and 78.5% respectively. On the other extreme, in the IA-Americana dataset, only 13.2% of text documents contained geolocations. It must be considered that text windows in IA-Americana are, on average, much shorter than articles in GDELT. They are, however, comparable in length to the two Twitter datasets which also show similar overall frequencies (means of 13.3% and 13.8% respectively).

Key to the research questions on geospatiality were the differences between topics within datasets (Table 3). We found these to be highest in the Stackexchange dataset (range 73.1%, $\sigma = 26.8\%$). For the *Technology* topic on Stackexchange, which included with Stackoverflow the firstand largest of the sites, $Frac_{Geo}$ was at 2.27% compared to 75.3% for *Travel, Tourism & Migration* and 71.2% for *Politics & Government*. The Reddit dataset was only a bit lower in variability (range of 61.6%, $\sigma = 17.2\%$). All the other corpora were measured with much smaller variability, with σ from 6.0 to 8.7% and ranges from 24.3 to 33.6%.

Although time units differed between the datasets, it could be seen that $Frac_{Geo}$ was more stable in time (top in Figure 4) than in the thematic domain. However, the Stackexchange site showed a different pattern, starting from almost no documents with geospatial references in the earlier years, and substantially increasing frequency as more Stackexchange sites were added to the network over time. Further, the Twitter datasets peak in $Frac_{Geo}$ in June and July 2018, around the time of the FIFA Football World Cup.

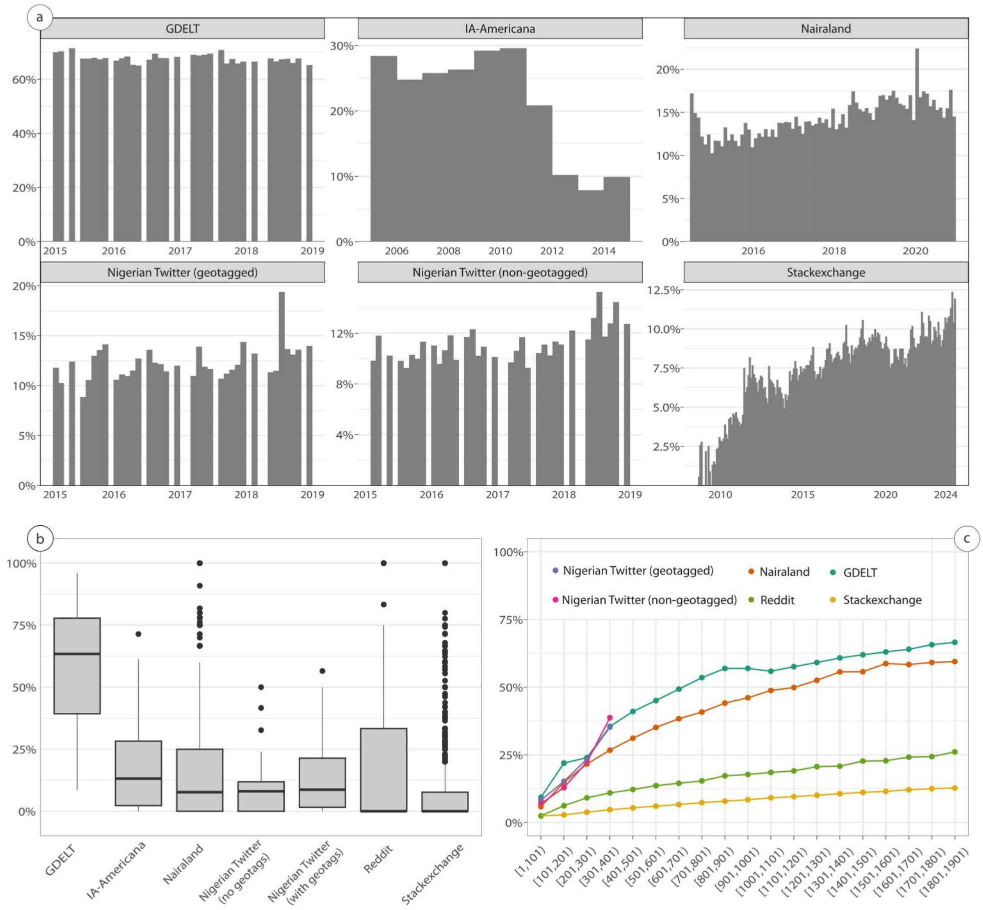


Figure 4. Observed $Frac_{geo}$ across timesteps (a), authors (b), and text length (c).

Next, we examined how $Frac_{Geo}$ varied between authors (lower left in Figure 4). Variability across authors was substantial. As measured by the coefficients of variation (CV), GDELT (0.36) is the only dataset where standard deviation is lower than the mean, unlike for IA-Americana (CV: 1.10), Nairaland (CV: 1.35), untagged (CV: 1.53) and tagged (CV: 1.07) Nigerian Twitter, Reddit (CV: 1.29), and Stackexchange (CV: 2.00). Notably for Reddit and Stackexchange, the median of $Frac_{Geo}$ was zero. In other words, more than half of the authors used no geospatial references in the texts we recorded.

Finally, text length (lower right in Figure 4) shows a clear association with the likelihood of containing spatial references. The magnitude of the relationship and the shape of the curve differ between the datasets, with the Twitter datasets increasing most strongly.

Altogether, we observe substantial between-topic variability in $Frac_{Geo}$, but also variability along text length, timesteps, and especially across authors.

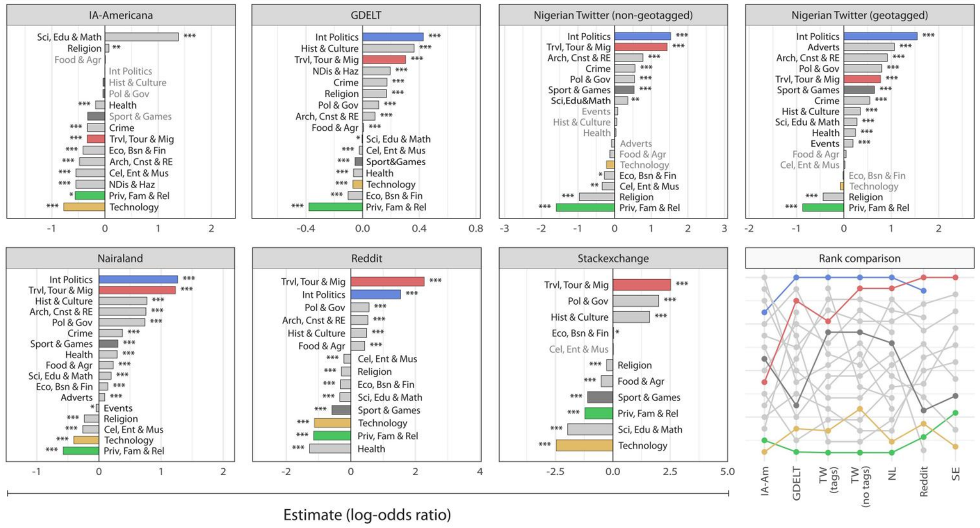


Figure 5. Estimates of each topic's effect on odds of a document to contain geoinformation, per dataset.

4.2. Modelled geospatiality of topics

Accounting for these variabilities, the separate mixed models fitted for each dataset found significant effect for topics in all datasets. The fixed-effect estimates for each topic within its respective dataset are presented in Figure 5, along with their significance. In all datasets, most topics' effects differed significantly ($p < 0.001$) from the baseline category *Other*. The highest positive effect on $\mathrm{Frac}_{\mathrm{Geo}}$ was found for *Travel, Tourism & Migration* in the Stackexchange dataset, with a coefficient of 2.52. The strongest negative effect was also found for Stackexchange, where the coefficient for *Technology* was estimated at −2.47. For example, the impact of *Religion* in the non-geotagged Twitter dataset was estimated with a log-odds ratio of 1, meaning that the odds of texts in that topic to contain a georeference were around 2.7 times lower as for a comment in the baseline *Other* category.

The fixed effects of text length ranged from 0.0003 in the Stackexchange dataset to 0.0084 in the not-geotagged Twitter dataset. The full model estimates, including for random effects, are presented in the Appendix Tables A1–A7.

4.3. Agreement of ranking over datasets

The relative ranking of topics based on their geospatiality was mostly, but not perfectly consistent across datasets. For example, *Travel, Tourism & Migration* and International Politics ranked high in most datasets. With similar consistency, *Technology* and *Private Life, Family & Relationships* ranked low.

However, this pattern is not true for all datasets. In the geotagged Twitter data, *Travel, Tourism & Migration* only ranks fifth. On Reddit, the least spatial topic is *Health*, which in most other datasets is close to the baseline topic *Other*. And highest ranking

Table 4. Observed spearman rank correlation between the rankings on each dataset.

Dataset	N. Twitter (non-geotagged)	N. Twitter (geotagged)	Nairaland	Reddit	Stack-exchange	GDELT	IA-Americana
Nigerian Twitter (non-geotagged)		0.85***	0.86***	0.64*	0.39	0.65**	0.21
Nigerian Twitter (geotagged)	0.85***		0.79***	0.68**	0.55	0.64**	0.23
Nairaland	0.86***	0.79***		0.72**	0.67*	0.79***	0.34
Reddit	0.64*	0.68**	0.72**		0.87***	0.84***	0.25
Stackexchange	0.39	0.55	0.67*	0.87***		0.61*	0.11
GDELT	0.65**	0.64**	0.79***	0.84***	0.61*		0.4
IA-Americana	0.21	0.23	0.34	0.25	0.11	0.4	

Asterisks encode significance as *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Note that for these pairwise correlations, no correction for multiple tests was applied.

Table 5. Validation accuracies.

Dataset	Accuracy	Kappa	Precision	Recall	FP	FN	TP	TN	*n*
GDELT	0.70	0.27	0.92	0.70	12	64	146	30	252
Nairaland	0.97	0.88	0.88	0.91	3	2	21	154	180
Twitter (geotagged)	0.94	0.76	0.88	0.73	3	8	22	147	180
Twitter (non-geotagged)	0.93	0.74	0.92	0.68	2	11	23	144	180
Stackexchange	0.94	0.83	0.79	0.96	6	1	22	91	120
WEBIS Reddit	0.97	0.92	0.88	1.00	4	0	29	115	148

FP = false positives; FN = false negatives, TP = true Positives; TN = true negatives.

on IA-Americana is *Science, Education & Mathematics*, rather *than International Politics*, which has been relegated to sixth place.

The rank-correlation coefficient (ρ) quantifies the high agreement between the three datasets from the Nigerian context (Table 4). The non-geotagged Twitter dataset is rank-correlated with the Nairaland data at $\rho = 0.86$ ($p < 0.001$) and the geotagged Twitter dataset at $\rho = 0.85$ ($p < 0.001$). Note that these three can be presumed to have very similar topic assignment, as the algorithm for labelling Tweets was trained on Nairaland data.

But even between very different platforms, there are similarities. The topic ranking on Reddit is similar to Stackexchange ($\rho = 0.87$ with $p < 0.001$), but also articles from GDELT ($\rho = 0.84$ with $p < 0.001$) and, although with lower significance, the Twitter datasets ($\rho = 0.64$ with $p < 0.05$ and $\rho = 0.68$ with $p < 0.01$, respectively). GDELT, which uses multi-label approach, is correlated significantly with all single-labelled datasets per text. The same is not the case for IA-Americana (which was processed with the same algorithm as GDELT) but has no significant correlation with any of the other datasets, whose texts were authored around a century later. Other than that, the only datasets between which the ranking is not correlated are Stackexchange and the Twitter data. Of the correlated datasets, correlations range from 0.64 to 0.87, which can be interpreted as moderate to very strong (Akoglu 2018). Overall, most rankings show imperfect, but significant correlations, which are unlikely to result from chance.

4.4. Validation of georeferences

The validation compared our semi-algorithmic approach to a human annotator working directly on the texts. We used accuracy metrics to quantify the agreement between the two references. It is important to mention that, due to ambiguities and the subjectivity in thematic labelling, we consider neither to be a true ground truth.

For the presence of georeferences in texts, algorithm and expert agreed in most cases, with accuracies between 93% and 97% and κ values ranging from 0.74 to 0.92 (Table 5). The notable exception is the GDELT dataset, where agreement was much lower with 69% accuracy and κ of 0.25.

4.5. Validation of topic classification

For the topic classification, accuracies ranged from 32 to 67% and κ of 0.28 to 0.64. A notable exception was the Stackexchange corpus, where agreement was measured with an accuracy of 85% and κ of 0.84. There, and for the Twitter and Nairaland datasets and the ratio of false positives to false negatives indicates underestimation, for the other datasets overestimation. Notably, most cases of disagreement involved the *Other* category. The full confusion matrices can be found in Appendix Figure A1.

5. Discussion

5.1. Research questions

In this study, we analysed whether the observed frequency of geolocations varies between topics in sources of English text web data (RQ 1), to what extent these variations can be attributed to the texts' topics (RQ 2) and to what extent this effect of topics is consistent across datasets (RQ 3). The results provide strong evidence that the observed frequency of geolocations varies between topics in the observed sources of English text web data. Therefore, the answer to RQ 1 is positive, with practical implications. Spatial methods will not be equally straightforward to implement in all fields of study. Research on people's private lives, personal relationships, and mental health, which have been studied on social media in particular (e.g. Chancellor and De Choudhury 2020), will find it harder to acquire the necessary data for incorporating spatial methods, since these topics lack geospatiality. On the other hand, studies of mobility and international politics may find more abundant data.

Concerning RQ 2, all models exhibit significant differences in the effect of topics on the likelihood of a text to contain geolocation. This indicates that within our analysed datasets, topics indeed vary in their inherent geospatiality. We also can get an idea of the magnitude of geospatiality effects: On Stackexchange, *Travel, Tourism & Migration* is measured with a log-odds ratio of around 2.5 compared to the baseline, while for *Technology* this value is roughly −2.5. In other words, the odds of a text document to be geolocated with current methods might vary by a factor of almost 150 between these two topics. Granted, this is the most extreme effect we measured, but it illustrates how misleading the assumption of equal geospatiality between topics can potentially be. The random intercepts we modelled show substantial variation in geolocation use among users, confirming previous findings by Karami *et al.* (2021).

Concerning RQ 3, the results indicate that the differences are frequently correlated between most pairs of datasets. We measured significant and considerable rank-correlations of geospatiality between most datasets, even those that were very different web mediums. For example, the news pages collected by GDELT and the conversational Nigerian web forum Nairaland lead to geospatiality-rankings that are correlated at 0.79. The highest correlations of 0.79 to 0.86 were found for Nairaland and the Twitter datasets, which is plausible considering their matching geographic context. However, even the two Twitter datasets are not perfectly correlated. This supports and extends the observation made by Serere and Resch (2024) that geotagged and non-geotagged tweets are overall similar but should not be assumed to be identical.

No dataset's ranking is correlated with that of IA-Americana, which contains periodicals and official documents. This indicates that text type and intentions of the writer also influence topical geospatiality. IA-Americana was not correlated even with GDELT, which was processed with an identical algorithm. The strong correlations between the other datasets are all the more noteworthy and do not rule out that geospatiality might be seen as an inherent property of specific topics.

Altogether, these findings imply that topics can indeed be a source of bias in geospatial data suitability that adds to the influences of text type, time, and author-specific effects. Specifically, comparative studies that rely solely on raw observation counts to map spatial activity patterns or compare prominence between topics may be affected. These studies should consider geospatiality as a source of bias and not generalize from geolocated data to data as a whole. To mitigate this bias, analysts could consider identifying topic categories within their text data and apply within-topic normalization to the observation counts to mitigate the effect of thematic geospatiality. Beyond that, our study provides empirical evidence of topic-specific geospatiality in contemporary web data, which can contribute to the understanding of spatiality as a concept in Linguistics and Literature studies (Tally 2012).

5.2. Limitations

There are several limitations that define the scope of confidence for our findings. Here, we discuss in turn issues about the analysed data, the topics, and the geolocation.

Firstly, our data can only depict a small portion of the volume and the variety of all available web text data. As datasets were not selected to be representative of all web text sources, we make no claim concerning geospatiality as a general phenomenon of text data, much less the English language as a whole. The diversity of the selected datasets provides a broad perspective but limits our ability to discern whether differences (e.g. between the results on IA-Americana and the other datasets) result from differences in text type, medium, or author demographics. A more diverse set of metadata-rich global social media data could enable more detailed comparisons, but no such dataset is, to our knowledge, currently openly available. On the level of topics, the topic-taxonomy we applied, informed by our inspection of the data sources and involving a substantial degree of human judgment, is certainly biased by our own background and our previous usage of web text data.

Second, data sources differ in a multitude of ways, and matching topics across datasets is always imperfect due to limited semantic overlap. For example, much content that was labelled as *Technology* on Stackexchange deals with computer programming issues, which has very limited overlap with the tech-related news that account for much of the *Technology* in the GDELT corpus. Political discussions on the web forum Nairaland have far more potential to be interactive than political news captured by GDELT. While both contain content that we consider political within the reference frame of the respective platform, the discourse takes a different form. Further, it is clear that representing topics as distinct and clearly separable is a strong abstraction of the reality. For example, almost any topic, from *sports* to *health*, can potentially also be a *political* question. This fuzziness is reflected by the high rate of disagreement we measured between human expert and the semi-algorithmic approach. Thus, caution should be applied in the interpretation of our results for individual topics. Nevertheless, many cases of disagreement are plausible and do not indicate a systematic issue that could cast doubt onto our overall conclusion. Rather, they indicate an inherent level of noisiness and ambiguity within the data. Frequently, the human expert assigned the *Other* category to uncertain cases that were algorithmically allocated to one of the topics. This suggests that the used data is frequently ambiguous and thematically heterogeneous. For instance, comments in the politics forum do not always directly relate to politics. The exception that seems to prove this rule is the high agreement on the Stackexchange data, a comparatively strongly moderated platform that enforces thematic homogeneity. It is likely that most forms of web content do not exhibit the same degree of thematic homogeneity. Especially in conversations, off-topic remarks and overlapping conversations mean that a clear assignment of texts to a single topic is the exception rather than the rule. However, we argue that some abstraction is always necessary in extracting human-interpretable information from immense datasets, and in this study, we aimed to reflect that. The presence of strong patterns despite the aforementioned ambiguities and abstractions underlines the strength, and therefore, the relevance of the geospatiality effect.

A similar perspective can be applied to the notion of geolocation. Treating geolocated-ness as binary state is also an abstraction and does not consider different degrees and granularities of being spatial. A qualitative analysis of our validation data revealed that many false positive classifications were caused by non-geographic places (e.g. 'Narnia',' Jupiter'), place names used in non-geographic context (e.g. 'Virgin America'), metonyms, and demonyms. These can be seen as technical errors, but in some fields and application contexts, may be useful spatial information. Thus, more nuanced analyses are desirable, which consider different perspectives and degrees of spatiality. Of course, the algorithmic identification of characteristics like metonymy is still a technical challenge, and limitation to our study. We expect that advances in large language models and artificial intelligence will reduce these limitations in the future.

As our definition of geospatiality as the likelihood to contain an identifiable geospatial reference is methodological, rather than formal, we hope that it will be relevant for data users. Of course, in this practical approach, what is identifiable depends on the capability of the applied methods. Therefore, our findings will need to be re-evaluated as geoparsing methods change and improve.

We believe that the topic-specific effects will appear however geospatiality is defined, but also expect that they will vary in some form. For example, whether demonyms (e.g. 'Germans', 'Romans', 'Punjabis') are considered geospatial references will have a substantial impact on the measured geospatiality of a topic like *History and Culture*.

5.3. Future directions

We found significant evidence that topical geospatiality affects the datasets we analysed. This provides sufficient grounds for future efforts to identify dimensions along which geospatiality varies. We hope that future studies with a narrower set of comparable data sources can improve on this research by fitting models across several datasets to identify effects at dataset-level. Likewise, the number of observations was not sufficient to identify – and control for – differences of thematic geospatiality at the level of author and time (which, in the mixed modelling framework, could be implemented as random slopes). For example, *Health* might not be a topic with particularly strong geospatial ties, but during the covid19 pandemic health-related news, measures, and indicators became widely reported and discussed in spatial terms, while other aspects of life became less geospatial. In our study, we observe that on Nairaland and Twitter, geospatiality peaks during events, such as the 2018 FIFA Football world cup. For this specific event and these platforms, the proportion of geolocated tweets actually increased for the topic *Sports & Games*. This, of course, is anecdotal evidence, and could result from several overlapping effects. A systematic analysis of event-specific changes in geospatiality seems warranted, also due to their relevance for the study of events (Senaratne *et al.* 2014). Likewise in space, variation is likely. For example, *Politics* might be more spatial in larger countries with spatially manifested ethnic divides, such as Nigeria, than in smaller, homogeneous countries like Lesotho.

A second major area that could be investigated spatial granularity and precision of geolocations: Do some topics primarily focus on states and regions, while others mention precise locations and landmarks (consider *International Politics* vs. *Tourism*)? Do some topics have a tendency towards mentioning a set of distant locations while others focus on proximal locations (consider *International Politics* vs. *Natural Disasters & Hazards*)? Are some geoparsing techniques better suited for certain topics? At a practical level, comparing the performance of geoparsing techniques for different topics will allow researchers to choose geoparsing approaches that suit their subject area best.

Future research could also draw upon concepts of spatiality from diverse fields (see Tally 2012) to formalize definitions of (geo-)spatial reference with regards to spatial granularity and test to what extent this influences the measured thematic geospatiality in data. We believe that geospatiality is an interesting and intuitive concept that can serve as a meeting ground for theoretical and practical fields of research. Its empirical study holds both theoretical and practical value.

6. Conclusions

We found clear evidence indicating that the frequency of georeferenced texts can vary between topics in some widely used web text data sources. After controlling for effects at the levels of author, time, and text length, we found significant evidence that the topics themselves are affecting the frequency of geolocations, which indicates, at least to some extent, topic-specific geospatiality effects. Accepting some limitations and depending on the composition of the data sets, these effects were moderately to strongly similar between most analysed datasets. We recommend that studies using web text data for spatial analyses should consider the relationship between the content they analyse and frequency of georeferences, in order to correctly gauge the representativeness of their data. We do not claim that our findings can be generalized to web-based text data on the whole. We rather urge researchers to consider in future studies the diverging goals of communication on individual websites and social media platforms and the very different forms of text composition and production one therefore finds on the internet. The current study should be seen as the starting point for a larger inquiry. The ubiquity of text data and improvements in geoparsing are promising great opportunities for geographic applications, and we hope that a better understanding of geospatiality can contribute to the quality of such research.

Notes

1. meta.stackexchange.com/questions/7931/.
2. https://doi.org/10.5281/zenodo.13941044.

Acknowledgment

The authors would like to thank Sebastian Düker and Melissa Auer for their assistance in improving the topic taxonomy.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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)- programme WASCAL WRAP 2.0. The study was partially supported by the project Open Search @DLR phase II (internal DLR project) and the project 'A New Focus in English Linguistics: Geolingual Studies', funded by the Volkswagen Foundation (Grant No. 98 662).

Notes on contributors

Johannes Mast is a Ph.D. student in the department 'Geo-Risks and Civil Security' at the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR). His research interest is the joint use of remote sensing and text data in the study of migration processes and urban environments. He contributed to conceptualization, methodology, software, validation, formal analysis, investigation, writing, editing, and visualization.

Richard Lemoine-Rodríguez holds a Ph.D. (Dr. rer. nat.) in Geography from the Ruhr-Universität Bochum, Germany, and is a postdoctoral researcher at the Geolingual Studies research and teaching unit at the University of Würzburg, Germany. His research integrates concepts and methods from urban ecology, geoinformatics and natural language processing to generate evidence-based knowledge to contribute to understand cities. He contributed to conceptualisation, resources, writing, reviewing, and editing.

Vanessa Rittlinger is a researcher in the department 'Geo-Risks and Civil Security' at the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR). Her current research focuses on the development of efficient workflows for information extraction in the context of natural disasters. She contributed to conceptualisation, validation, data curation, writing, reviewing, and editing.

Martin Mühlbauer is a permanent researcher in the department 'Geo-Risks and Civil Security' at the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR). His research focuses on the provision and analysis of remote sensing and ancillary data sources, particularly in the context of disaster management and civil security. He contributed to conceptualisation and resources.

Carolin Biewer is a full professor and the holder of the Chair of English Linguistics at the University of Würzburg, Germany, and a co-founder of Geolingual Studies, a research and teaching unit at the University of Würzburg, which brings together scholars from linguistics, remote sensing, and data sciences to better understand the interrelation between physical space and socially constructed space as place. She contributed to the supervision, funding acquisition, writing, reviewing, and editing.

Christian Geiß is a full professor and the leader of the research group for Georisk Research using Remote Sensing Methods at the University of Bonn, Germany, and the Team Georisks in the department 'Geo-Risks and Civil Security' at the German Remote Sensing Data Center (DFD). His research interests include the development of machine learning methods for the interpretation of earth observation data and exposure and vulnerability assessment in the context of natural hazards. He contributed to conceptualization, supervision, funding acquisition, writing, reviewing, and editing.

Hannes Taubenböck is head of the department 'Geo-Risks and Civil Security' at the German Remote Sensing Data Center (DFD) of the German Aerospace Center (DLR), he is full professor and the holder of the Chair for Global Urbanisation and Remote Sensing at the University of Würzburg, and co-founder of Geolingual Studies at the University of Würzburg. His current research interests include value adding to remote-sensing-based classification products in combination with other geodata and the development of algorithms for information extraction in the context of urban geography. He contributed to conceptualization, supervision, funding acquisition, writing, reviewing, and editing.

Data and codes availability statement

The data and code that support the findings of this study are available on zenodo at https://doi.org/10.5281/zenodo.13941044. These data were derived from the following resources available in the public domain:

WEBIS Reddit corpus is available at https://doi.org/10.5281/zenodo.1043504. The Stackexchange data are available at: archive.org/details/stackexchange. The GDELT and IA-Americana datasets are provided by the GDELT project at: gdeltproject.org. This study used third party data made available under licence that the author does not have permission to share. Requests to access the data should be directed to X Corp. at developer.x.com.

References

Adelani, D.I., *et al.*, 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9, 1116–1131.

Akoglu, H., 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18 (3), 91–93.

Berragan, C., *et al.*, 2022. Geoparsing comments from Reddit to extract mental place connectivity within the United Kingdom. UC Santa Barbara: Center for Spatial Studies. http://dx.doi.org/10.25436/E28C7R.

Blanqué, P., *et al.*, 2022. Monitoring narratives: An application to the equity market. *SSRN*. https://doi.org/10.2139/ssrn.4081958.

Bodas-Sagi, D., and Labeaga, J., 2016. Using GDELT data to evaluate the confidence on the Spanish government energy policy. *International Journal of Interactive Multimedia and Artificial Intelligence*, 3, 38–43.

Brooks, M.E., *et al.*, 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9 (2), 378–400.

Buckingham, K., *et al.*, 2020. The untapped potential of mining news media events for understanding environmental change. *Current Opinion in Environmental Sustainability*, 45, 92–99.

Chancellor, S., and De Choudhury, M., 2020. Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine*, 3 (1), 43.

Chiluwa, I., and Odebunmi, A., 2016. On terrorist attacks in Nigeria: Stance and engagement in conversations on *Nairaland*. *Communication and the Public*, 1 (1), 91–109.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), 37–46.

Common Crawl 2024., Distribution of Languages [online]. Available from: https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html [Accessed 24 Sep 2024].

Devlin, J., *et al.*, 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

El Ouadi, A., and Beskow, D., 2024. Comparison of Common Crawl News & GDELT. In *2024 IEEE International Systems Conference (SysCon)*. IEEE, 1–3.

Firouzjaei, H.A., 2024. A deep learning-based approach for identifying unresolved questions on Stack Exchange Q & A communities through graph-based communication modelling. *International Journal of Data Science and Analytics*, 18 (2), 205–218.

Fox, N., *et al.*, 2021. Reddit: A novel data source for cultural ecosystem service studies. *Ecosystem Services*, 50, 101331.

Gan, Q., *et al.*, 2008. Analysis of geographic queries in a search engine log. *In*: *Proceedings of the first international workshop on Location and the web*. Presented at the WWW '08: The 17th International World Wide Web Conference, Beijing China: ACM, 49–56.

Gerner, D.J., et al., 2002. *Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions*. New Orleans: International Studies Association.

Gries, S., 2015. The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10 (1), 95–125.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv Preprint*, *arXiv:2203.05794.* https://maartengr.github.io/BERTopic/index.html#citation.

Hecht, B., and Stephens, M., 2014. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International AAAI Conference on Web and Social Media*, 8 (1), 197–205.

Honnibal, M. et al., 2020. spaCy: Industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303.

Hu, X., *et al.*, 2023. Location reference recognition from texts: A survey and comparison. *ACM Computing Surveys*, 56 (5), 1–37.

Jiang, Y., Li, Z., and Ye, X., 2019. Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. *Cartography and Geographic Information Science*, 46 (3), 228–242.

Jiang, J., *et al.*, 2023. Geolocated social media posts are happier: Understanding the characteristics of check-in posts on Twitter. *In*: *Proceedings of the 15th ACM Web Science Conference 2023*. Presented at the WebSci '23: 15th ACM Web Science Conference 2023, Austin TX USA: ACM, 136–146.

Karami, A., *et al.*, 2020. Twitter and research: A systematic literature review through text mining. *IEEE Access*, 8, 67698–67717.

Karami, A., *et al.*, 2021. Analysis of geotagging behavior: Do geotagged users represent the twitter population? *ISPRS International Journal of Geo-Information*, 10 (6), 373.

Kersten, J., and Klan, F., 2020. What happens where during disasters? A Workflow for the multifaceted characterization of crisis events based on Twitter data. *Journal of Contingencies and Crisis Management*, 28 (3), 262–280.

Lamidi, I.M., 2016. *Humour markers and their interpretations in the Nairaland virtual community*. Ibadan: University of Ibadan.

Lamidi, M.T., 2020. Investigating cybercrime in Nigeria. In *Encyclopedia of Criminal Activities and the Deep Web*. Hershey, USA: IGI Global, 1018–1033.

Lansley, G., and Longley, P.A., 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85–96.

Leetaru, K.H., 2015. Mining libraries: Lessons learned from 20 years of massive computing on the world's information. *Information Services & Use*, 35 (1–2), 31–50.

Leetaru, K., and Schrodt, P.A., 2013. Gdelt: Global data on events, location, and tone, 1979–2012. *In*: *ISA annual convention.* Storrs, USA: Citeseer, 1–49.

Lemoine-Rodríguez, R., Biewer, C., and Taubenböck, H., 2024. 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*: H. Carlos-Martinez, R. Tapia-McClung, D.A. Moctezuma-Ochoa, and A.J. Alegre-Mondragón, eds. *Recent Developments in Geospatial Information Sciences*. Cham, Switzerland: Springer Nature, 119–133.

Longley, P.A., Adnan, M., and Lansley, G., 2015. The Geotemporal demographics of Twitter usage. *Environment and Planning A: Economy and Space*, 47 (2), 465–484.

Malik, M., *et al.*, 2015. Population bias in geotagged tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 9 (4), 18–27.

Mast, J., *et al.*, 2024. The migrant perspective: Measuring migrants' movements and interests using geolocated tweets. *Population, Space and Place*, 30 (2), e2732.

Nwachukwu, E., 2015. Framing the #Occupy Nigeria Protests in Newspapers and Social Media. *Open Access Library Journal*, 02 (05), 1.

Olteanu, A., *et al.*, 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *Proceedings of the International AAAI Conference on Web and Social Media*, 8 (1), 376–385.

Olteanu, A., *et al.*, 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 13.

Owuor, I., Hochmair, H.H., and Cvetojevic, S., 2020. Tracking hurricane Dorian in GDELT and twitter. *AGILE: GIScience Series*, 1, 1–18.

Oyebode, O., and Orji, R., 2019. Detecting factors responsible for diabetes prevalence in Nigeria using social media and machine learning. *In*: *2019 15th International Conference on Network and Service Management (CNSM)*. Presented at the 2019 15th International Conference on Network and Service Management (CNSM), Halifax, NS, Canada: IEEE, 1–4.

Partridge, V., *et al.*, 2024. Here be livestreams: Trade-offs in creating temporal maps of Reddit. *In*: *ACM Web Science Conference*. Presented at the Websci '24: 16th ACM Web Science Conference, Stuttgart Germany: ACM, 81–91.

Pavalanathan, U., and Eisenstein, J., 2015. Confounds and Consequences in Geotagged Twitter Data. *In*: *Proceedings of the 2015 conference on empirical methods in natural language processing,* Lisbon. Association for Computational Linguistics, 2138–2148.

Qiao, F., *et al.*, 2017. Predicting social unrest events with hidden Markov models using GDELT. *Discrete Dynamics in Nature and Society*, 2017, 1–13.

R Core Team 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D.B., 1976. Inference and missing data. *Biometrika*, 63 (3), 581–592.

Schweter, S., and Akbik, A., 2021. FLERT: Document-Level Features for Named Entity Recognition. https://doi.org/10.48550/arXiv.2011.06993.

Sen, I., *et al.*, 2021. A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85 (S1), 399–422.

Senaratne, H., *et al.*, 2014. Moving on Twitter: using episodic hotspot and drift analysis to detect and characterise spatial trajectories. *In*: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN '14. Presented at the the 7th ACM SIGSPATIAL International Workshop*, Dallas/Fort Worth, Texas: ACM Press, 23–30.

Senaratne, H., *et al.*, 2023. 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*, 12 (4), 175.

Serere, H.N., and Resch, B., 2024. Understanding the impact of geotagging on location inference models for accurate generalization to non-geotagged datasets. *Geomatica*, 76 (1), 100004.

Spearman, C., 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15 (1), 72–101.

Stack Exchange Inc 2024. Stack exchange, [online]. Available from: https://stackexchange.com/ [Accessed 14 June 2024].

Stack Exchange Inc. 2024. Stack Overflow and OpenAI Partner to Strengthen the World's Most Popular Large Language Models - Press release - Stack Overflow [online]. Available from: https://stackoverflow.co/company/press/archive/openai-partnership/ [Accessed 29 September 2024].

Stack Exchange, Inc 2024. Stack Exchange Data Dump [online]. *Internet Archive*. Available from: https://archive.org/details/stackexchange [Accessed 14 June 2024].

Steup, M., and Neta, R., 2005. Epistemology. *In*: *Stanford encyclopedia of philosophy*. Stanford, USA: Stanford University.

Tally, R., 2012. *Spatiality*. London: Routledge.

Taubenböck, H., *et al.*, 2018. Are the poor digitally left behind? indications of urban divides based on remote sensing and Twitter data. *ISPRS International Journal of Geo-Information*, 7 (8), 304.

Völske, M., *et al.*, 2017. Mining Reddit to learn automatic summarization. *In*: *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.

Wickham, H., *et al.*, 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4 (43), 1686.

Williams, S., 2020. Appendix on: Global Database of Events, Language and Tone (GDELT). Newport, UK: Office for National Statistics.

World Bank, W., 2015. WBG Topical Taxonomy [online]. https://vocabularyserver.com/worldbank/taxonomy/. Available from: https://vocabularyserver.com/worldbank/taxonomy/ [Accessed 18 October 2023].

Zhang, X., *et al.*, 2023. TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations at Twitter. *In*: *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. New York: ACM, 5597–5607. https://dynamo.cs.ucsb.edu/publications/1525.

Zheng, C., 2020. Comparisons of the city brand influence of global cities: Word-embedding based semantic mining and clustering analysis on the big data of GDELT global news knowledge graph. *Sustainability*, 12 (16), 6294.

Zhu, X.X., *et al.*, 2022. Geoinformation Harvesting From Social Media Data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine*, 10 (4), 150–180.

Appendix

Table A1. Estimated coefficients for model on: GDELT.

Term	Estimate	SE	Statistic	*p*
(Intercept)	−0.4559	0.0138	−33.1484	0
Topic: Architecture, Construction & Real Estate	0.0876	0.0028	31.754	0
Topic: Celebrities, Entertainment & Music	−0.0234	0.0019	−12.4614	0
Topic: Crime	0.1715	0.0018	96.3453	0
Topic: Economy, Business & Finance	−0.1027	0.0017	−58.8909	0
Topic: Food & Agriculture	0.0068	0.002	3.4003	0.0007
Topic: Health	−0.064	0.0018	−35.4632	0
Topic: History & Culture	0.3635	0.0079	46.0347	0
Topic: International Politics	0.4293	0.002	211.045	0
Topic: Natural Disasters and Hazards	0.1943	0.0024	82.2403	0
Topic: Politics & Government	0.1134	0.0017	65.7901	0
Topic: Religion	0.1686	0.0026	63.6862	0
Topic: Science, Education & Mathematics	−0.0044	0.0018	−2.5008	0.0124
Topic: Sports & Games	−0.0544	0.0103	−5.2788	0
Topic: Technology	−0.0688	0.0019	−37.1373	0
Topic: Travel, Tourism & Migration	0.3042	0.002	149.6827	0
Topic: Private Life, Family & Relationships	−0.3798	0.01	−37.9742	0
text length	0.0003	0	2021.5574	0
random effect (intercept σ) for author	1.0563			
random effect (intercept σ) for timestep	0.0504			

Table A2. Estimated coefficients for model on: IA-AMERICANA.

Term	Estimate	SE	Statistic	*p*
(Intercept)	−2.1197	0.0463	−45.7472	0
Topic: Architecture, Construction & Real Estate	−0.4787	0.0352	−13.6182	0
Topic: Celebrities, Entertainment & Music	−0.5457	0.039	−14.0026	0
Topic: Crime	−0.332	0.0292	−11.3696	0
Topic: Economy, Business & Finance	−0.4123	0.0263	−15.6805	0
Topic: Food & Agriculture	0.0071	0.0266	0.2661	0.7902
Topic: Health	−0.1762	0.0258	−6.8248	0
Topic: History & Culture	−0.0343	0.0864	−0.3968	0.6915
Topic: International Politics	−0.0025	0.0329	−0.0748	0.9404
Topic: Natural Disasters and Hazards	−0.548	0.0364	−15.0385	0
Topic: Politics & Government	−0.0362	0.0261	−1.3891	0.1648
Topic: Private Life, Family & Relationships	−0.5622	0.2475	−2.2714	0.0231
Topic: Religion	0.0737	0.0277	2.6621	0.0078
Topic: Science, Education & Mathematics	1.3805	0.0254	54.2891	0
Topic: Sports & Games	−0.3269	0.2043	−1.6001	0.1096
Topic: Technology	−0.7739	0.0389	−19.8736	0
Topic: Travel, Tourism & Migration	−0.3334	0.0308	−10.8253	0
random effect (intercept σ) for author	1.499			
random effect (intercept σ) for timestep	0.0757			

Table A3. Estimated coefficients for model on: Nairaland.

Term	Estimate	SE	Statistic	*p*
(Intercept)	−2.4976	0.0131	−190.7583	0
Topic: Adverts	0.0942	0.0186	5.0752	0
Topic: Architecture, Construction & Real Estate	0.7574	0.0273	27.7695	0
Topic: Celebrities, Entertainment & Music	−0.2612	0.0142	−18.4131	0
Topic: Crime	0.3776	0.014	27.054	0
Topic: Economy, Business & Finance	0.1424	0.0131	10.8763	0
Topic: Events	−0.0437	0.0187	−2.342	0.0192
Topic: Food & Agriculture	0.2302	0.0147	15.651	0
Topic: Health	0.2942	0.017	17.2783	0
Topic: History & Culture	0.7703	0.0202	38.1262	0
Topic: International Politics	1.2668	0.0133	95.2783	0
Topic: Politics & Government	0.7385	0.0132	55.9232	0
Topic: Private Life, Family & Relationships	−0.5774	0.0131	−44.0802	0
Topic: Religion	−0.2365	0.0146	−16.1884	0
Topic: Science, Education & Mathematics	0.1949	0.0147	13.2414	0
Topic: Sports & Games	0.3051	0.0146	20.8603	0
Topic: Technology	−0.4059	0.0139	−29.1121	0
Topic: Travel, Tourism & Migration	1.2313	0.0127	96.7454	0
Text length	0.0014	0	284.3321	0
Random effect (intercept σ) for author	0.7743			
Random effect (intercept σ) for timestep	0.0638			

Table A4. Estimated coefficients for model on: NigerianTwitter (geotagged).

Term	Estimate	SE	Statistic	*p*
(Intercept)	−3.1566	0.0535	−58.9897	0
Topic: Adverts	1.0645	0.0539	19.7329	0
Topic: Architecture, Construction & Real Estate	0.9177	0.0601	15.2599	0
Topic: Celebrities, Entertainment & Music	0.0202	0.0506	0.3982	0.6905
Topic: Crime	0.5449	0.0578	9.4291	0
Topic: Economy, Business & Finance	−0.0188	0.0548	−0.3422	0.7322
Topic: Events	0.1909	0.0488	3.9146	0.0001
Topic: Food & Agriculture	0.0442	0.0471	0.9385	0.348
Topic: Health	0.2439	0.0624	3.9115	0.0001
Topic: History & Culture	0.3451	0.0559	6.1683	0
Topic: International Politics	1.5474	0.0527	29.375	0
Topic: Politics & Government	0.7973	0.0483	16.5159	0
Topic: Private Life, Family & Relationships	−0.8712	0.0592	−14.7202	0
Topic: Religion	−0.441	0.061	−7.2251	0
Topic: Science, Education & Mathematics	0.2735	0.0644	4.245	0
Topic: Sports & Games	0.645	0.0505	12.7782	0
Topic: Technology	−0.0744	0.067	−1.1093	0.2673
Topic: Travel, Tourism & Migration	0.773	0.0478	16.1757	0
Text length	0.0064	0.0001	50.2389	0
Random effect (intercept σ) for author	0.9083			
Random effect (intercept σ) for timestep	0.1046			

Table A5. Estimated coefficients for model on: NigerianTwitter (non-geotagged).

Term	Estimate	SE	Statistic	*p*
(Intercept)	−3.5309	0.1077	−32.7953	0
Topic: Adverts	−0.0852	0.134	−0.6363	0.5246
Topic: Architecture, Construction & Real Estate	0.7731	0.1407	5.4963	0
Topic: Celebrities, Entertainment & Music	−0.3402	0.1169	−2.911	0.0036
Topic: Crime	0.5572	0.122	4.5688	0
Topic: Economy, Business & Finance	−0.2825	0.1243	−2.2721	0.0231
Topic: Events	0.0873	0.1073	0.8131	0.4162
Topic: Food & Agriculture	−0.1297	0.105	−1.2348	0.2169
Topic: Health	0.0389	0.1358	0.2863	0.7747
Topic: History & Culture	0.0582	0.1275	0.4563	0.6482
Topic: International Politics	1.5314	0.1148	13.3344	0
Topic: Politics & Government	0.5413	0.1089	4.9692	0
Topic: Private Life, Family & Relationships	−1.5965	0.1505	−10.6046	0
Topic: Religion	−0.9646	0.1429	−6.7518	0
Topic: Science, Education & Mathematics	0.3596	0.1391	2.5847	0.0097
Topic: Sports & Games	0.5328	0.1124	4.7414	0
Topic: Technology	−0.2203	0.1413	−1.5598	0.1188
Topic: Travel, Tourism & Migration	1.4351	0.1127	12.7333	0
Text length	0.0084	0.0003	31.8551	0
Random effect (intercept σ) for author	0.8015			
Random effect (intercept σ) for timestep	0.0287			

Table A6. Estimated coefficients for model on: Stackexchange.

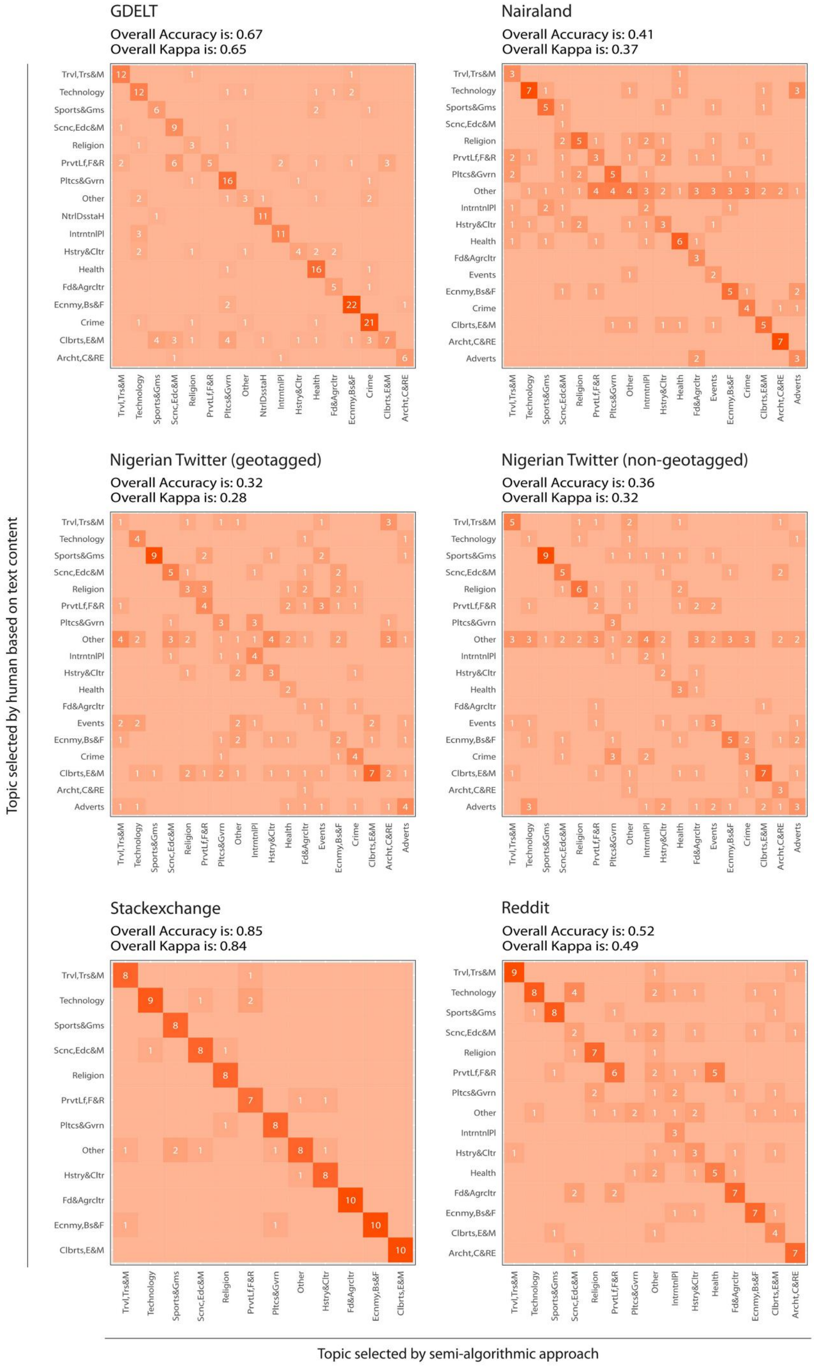
Term	Estimate	SE	Statistic	*p*
(Intercept)	−2.0019	0.0112	−178.289	0
Topic: Celebrities, Entertainment & Music	0.0185	0.0168	1.1011	0.2708
Topic: Economy, Business & Finance	0.0299	0.014	2.14	0.0324
Topic: Food & Agriculture	−0.5047	0.0171	−29.5464	0
Topic: History & Culture	1.6023	0.0168	95.2781	0
Topic: Politics & Government	1.9984	0.0164	121.4918	0
Topic: Private Life, Family & Relationships	−1.2213	0.0204	−59.9909	0
Topic: Religion	−0.2665	0.0144	−18.4749	0
Topic: Science, Education & Mathematics	−1.9674	0.0116	−169.1926	0
Topic: Sports & Games	−1.1027	0.0128	−86.4369	0
Topic: Technology	−2.4744	0.0121	−204.3273	0
Topic: Travel, Tourism & Migration	2.5225	0.0155	162.8395	0
Text length	0.0003	0	161.1036	0
Random effect (intercept σ) for author	0.8669			
Random effect (intercept σ) for timestep	0.0162			

Table A7. Estimated coefficients for model on: WEBIS Reddit.

Term	Estimate	SE	Statistic	*p*
(Intercept)	−1.8796	0.0112	−168.5564	0
Topic: Architecture, Construction & Real Estate	0.5279	0.1369	3.8555	0.0001
Topic: Celebrities, Entertainment & Music	−0.2166	0.0175	−12.3564	0
Topic: Economy, Business & Finance	−0.3224	0.0254	−12.684	0
Topic: Food & Agriculture	0.4294	0.0473	9.0788	0
Topic: Health	−1.278	0.0337	−37.8816	0
Topic: History & Culture	0.49	0.0519	9.4402	0
Topic: International Politics	1.5405	0.0199	77.4335	0
Topic: Politics & Government	0.5612	0.0158	35.5922	0
Topic: Private Life, Family & Relationships	−1.1533	0.0181	−63.7781	0
Topic: Religion	−0.2912	0.019	−15.338	0
Topic: Science, Education & Mathematics	−0.3394	0.0198	−17.1455	0
Topic: Sports & Games	−0.589	0.0124	−47.4515	0
Topic: Technology	−1.1282	0.0181	−62.3831	0
Topic: Travel, Tourism & Migration	2.2714	0.062	36.6099	0
Text length	0.0004	0	91.7287	0
Random effect (intercept σ) for author	0.7338			

Table A8. Description of topics that form the topic taxonomy.

Topic	Description
Health	Physical and mental health (excluding physical exercise for performance, which should rather be assigned to Sports & Games).
History & Culture	Humanities, performing and creative arts (generally not including crafts such as woodworking), history (excluding very recent history), cultural heritage.
Science, Education & Mathematics	Natural sciences, mathematics, statistics, academia, schools, education, research.
Economy, Business & Finance	Both personal finance and economics.
Sports & Games	Video games, card & board games, professional sports and casual sports, fitness, other competitions of physical activity.
Technology	Including electronics, software engineering, development of software (also games). Practically applied data science and machine learning, but excluding theoretical statistics and data science (those should rather go to Science, Education & Mathematics).
Adverts	Self-promotion, typically of commercial content, or job offers.
Private Life, Family & Relationships	Also dating, pets & housework.
Religion	Religion, spirituality, philosophy.
Politics & Government	Domestic politics (any country). Party politics, elections.
Travel, Tourism & Migration	Including politics of migration. Means of travel and transport, travel suggestions. Including topics about refugees and traffic infrastructure.
Other	A mixed category. Things that are by their nature about a wide range of different topics, such as: jokes, literature, advice.
International Politics	Relations between countries, geopolitics, diplomacy, cross-border military operations.
Celebrities, Entertainment & Music	Including movies, TV Series, anime, cartoons, and comics. Celebrities when they are from the entertainment industry. Famous people who are active in a different field (e.g. religious leaders or politicians in office) are rather assigned to a category fitting their field, if possible, unless the text only deals with their private life. Contemporary and classical music and musicians, also including theory of music and instruments.
Food & Agriculture	Cooking, both professional and casual farming (excluding drugs), gardening, vegetarianism/veganism.
Events	Birthdays, Weddings, similar celebrations, reunions, and other organized gatherings that do not clearly fit another category.
Architecture, Construction & Real Estate	Including urban planning.
Crime	Excluding manmade disasters at larger scale.
Natural Disasters and Hazards	Including smaller-scale disasters and accidents affecting only a few people. Not including manmade disasters such as pollution, oil spills, etc.



GDELT

Overall Accuracy is: 0.67
Overall Kappa is: 0.65

Nairaland

Overall Accuracy is: 0.41
Overall Kappa is: 0.37

Nigerian Twitter (geotagged)

Overall Accuracy is: 0.32
Overall Kappa is: 0.28

Nigerian Twitter (non-geotagged)

Overall Accuracy is: 0.36
Overall Kappa is: 0.32

Stackexchange

Overall Accuracy is: 0.85
Overall Kappa is: 0.84

Reddit

Overall Accuracy is: 0.52
Overall Kappa is: 0.49

The number indicates the number of cases for each combination of topics resulting from the annotations of the human and semi-algorithmic approach.

Figure A1. Confusion matrices of the topic classification.