Orsanigo et al. *EPJ Data Science* (2025) 14:86
https://doi.org/10.1140/epjds/s13688-025-00597-z

EPJ.org

EPJ Data Science

RESEARCH

Open Access

The geography of interest in international conversations on Twitter

Veronica Orsanigo[1,2], Sebastiano Bontorin[1], Thomas Louf[1], Elisa Leonardelli[1], Alessio Palmero Aprosio[1,2], Pierluigi Sacco[3], Sara Tonelli[1] and Riccardo Gallotti[1*]

Handling Editor: Luca Rossi

*Correspondence: rgallotti@fbk.eu
[1]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy
Full list of author information is available at the end of the article

Abstract

Current discourse on online social networks provides a valuable window into the dynamics, opinions, and behaviors of real-world social networks. This study leverages geolocated Twitter data and location references in tweets to examine patterns of international and interstate attention, evaluated through the number of mentions across three regions: Europe, South America and the United States. We construct directed networks where source nodes represent geographical areas (countries or states) tweets originate from, target nodes represent areas of locations mentioned in tweet text, and edges are weighted by mention frequency, a measure we define as "interest". Our analysis reveals that these networks are remarkably dense, with nearly all areas mentioning most others, although with significant asymmetries in attention distribution. Within each region, there are some states that consistently receive or generate disproportionate amounts of mentions. To explain these patterns, we develop augmented gravity models that incorporate economic, geographic, linguistic, and demographic factors. These models demonstrate that the interest is primarily shaped by the GDP of both source and target areas, the geographic distance between them, and migration flows. The effect of distance varies notably across regions, with European discourse showing substantially weaker geographic constraints compared to the Americas, suggesting that regional integration may reduce spatial friction in international and interstate attention. Through topic modeling, we further identify both common discourse domains (tourism, sports, culture) present across all regions and distinctive regional preoccupations that reflect specific historical and political contexts. These findings illuminate how digital communication both reflects and potentially reshapes traditional patterns of international attention, offering insights into the evolving nature of global discourse in the digital age.

Keywords: Data science; Geographical Information System; Locations extraction; Spatial network; International attention

1 Introduction

Social media platforms have fundamentally transformed the landscape of international communication, enabling people from different countries to interact and overcome physical boundaries in radically novel ways. The advent of digital platforms and social media

Springer

has ushered in an era of global conversation unprecedented in human history [1–3]. For the first time, billions of individuals can engage in direct, instantaneous cross-border communication, transcending geographical, political, and cultural boundaries that have traditionally constrained international discourse. This technological transformation has not only accelerated existing patterns of transnational communication; it has fundamentally reconfigured the architecture of global information exchange [4]. Social media platforms in particular have democratized participation in international discussions, enabling ordinary citizens, not just diplomats, journalists, or business elites, to express interest in distant places, comment on foreign events, and develop awareness of other societies [5]. The geographical scope and scale of these conversations is without historical parallels: a Malaysian teenager can now directly engage with Canadian peers about shared interests, a Brazilian journalist can instantly report on European events to a global audience, and diaspora communities can maintain real-time connections with their homelands. Yet, even as digital technology dissolves certain spatial constraints, questions remain about whether traditional determinants of international connections, such as economic relationships, geographic proximity, linguistic commonality, and migration patterns [6, 7], continue to shape attention and interest in this new communicative landscape.

The analysis of transnational communication on social media offers rich insights into how international discourse is structured and about the drivers that shape patterns of cross-border attention and interest [8, 9]. The formation of transnational communication networks is influenced by various factors. Previous research [6] has demonstrated that social ties on Twitter are shaped by geographical distance, language similarities, national boundaries, and frequency of air travel between regions. However, most existing studies focus on direct user-to-user interactions by analyzing hashtags or retweets networks [10, 11], while comparatively less attention has been devoted to understanding patterns of interest between countries at a macro level.

Location-related information is fundamental to analyze transnational interactions. Locations can be extracted from social media data through various approaches. While geotagged data provide the most straightforward geographical information, only approximately 1-2% of tweets include precise geo-coordinates [12], necessitating more sophisticated techniques for geographical inference [13]. [14] developed methods to extract geographical information from user meta-fields such as profile descriptions and self-declared locations. [15] demonstrated that users' physical locations can be inferred from the content of their posts, while [16] leveraged social network structures to derive location information. [17] combined Named Entity Recognition with knowledge bases to map location mentions to geographic entities. In a study using Facebook pages, [18] employed the page-like graph to predict location information by adopting Breadth-First Search (BFS).

Several previous studies have used geotagged data for location inference and have analyzed cross-border communication and information flows on social media. [19] combined Twitter data and data from mobile phone providers as a proxy for the position of people, to infer the size of a crowd in a football stadium. [7] examined Twitter mentions between users from different countries, modeling international communication patterns using a gravity approach combined with social, economic, and cultural variables. [10] studied the retweet network between users located in Africa and Europe to investigate the online diffusion of migration-related information. The international migration network was investigated also in [20] to understand its structure and key factors determining its evolution over

time. [21] examined the attractiveness of 20 of the most popular tourist sites in the world through geolocated tweets as a proxy for human mobility. [22] identified regions of strong internal connectivity in the global Twitter mentions network, correlating these clusters with shared languages, borders, and historical ties. [23] explored how language groups connect within multilingual Twitter communities, demonstrating how linguistic factors shape communication networks. [24] analyzed mobility patterns of Twitter users across countries, comparing Twitter-based human travel networks with gravity model predictions. The gravity model [25], originally developed to explain human migration patterns and later applied to trade studies [26], provides a powerful framework for understanding collective flows [25, 27, 28] between geographical entities. In its basic form, it predicts that the interaction between two locations is proportional to their "masses" (typically population or economic size) and inversely proportional to the distance between them. [29] extended this approach to incorporate additional variables such as social and cultural factors in modeling trade relationships.

In our research we focus on international and interstate communication in specific regions of the world i.e. Europe, South America and the United States, that were identified through geotagged data. We chose these three regions because they are characterized by numerous internal boundaries and relatively free movement of people, and provide interesting contrasts in terms of linguistic diversity, economic development, historical ties, and political integration. Europe represents a multilingual context with high economic integration but persistent cultural and linguistic divides. South America offers a case of greater linguistic homogeneity (predominantly Spanish and Portuguese) but more variable levels of economic development and integration. The United States, while officially monolingual at the federal level, provides an interesting counterpoint as a federation of states with significant internal differences in economic development, politics, and culture. By comparing patterns across these three regions, we can identify both common dynamics and context-specific factors shaping international interest networks.

Our approach differs from previous studies in several important ways. First, rather than examining direct communication between users from different countries, we focus on the latent interest between countries as revealed through location mentions in tweet content. This approach captures patterns of international attention that may not involve direct cross-border communication but nevertheless reflect how countries relate to each other in social media discourse. Second, for each region we perform a network analysis of location mentions and combine it with both gravity modeling and topic modeling, providing a multifaceted understanding of both the structural patterns and content dimensions of international interest networks. Third, we employ a comparative approach, analyzing three distinct world regions with different linguistic, cultural, and geographic characteristics. This comparative dimension allows us to distinguish common patterns from region-specific phenomena.

Our research addresses four key questions.

(RQ1) Do similar patterns of intraregional (international or interstate) interest emerge across different world regions, or are there distinct dynamics?

(RQ2) Is intraregional discourse on social media balanced across countries (or states), or concentrated around certain key nodes?

(RQ3) What drives the pattern of interest between countries (or states)?

(RQ4) What topics characterize the discourse about different countries (or states) in different contexts?

By answering these questions, we contribute to understanding how international relations are reflected and potentially shaped by patterns of attention and discourse on social media platforms.

A remark is needed here. In this manuscript, we correctly use the terms "countries", "nations", "international interest" for Europe and South America. However, to avoid redundancy and make the reading smoother, since our approach, analysis and results also apply to the United States, in the rest of the manuscript we do not specify the terms "states" and "interstate interest" when we talk about the US or our results in general.

Our results show that the observed networks reveal the patterns of interest across the three regions. All networks show high connectivity, with most countries mentioning most others, but with significant variations in the strength of such connections. Some countries emerge as central nodes, being mentioned far more frequently than others, while some demonstrate a higher propensity to mention foreign locations. The gravity models help explain these patterns, revealing the influence of economic factors, geographical proximity, linguistic ties, and migration flows on international interest patterns. The topic modeling results complement these structural analyses by revealing the content dimensions of international discourse, identifying both common and region-specific themes (such as travel, sports, and politics).

In the subsequent sections, we detail our datasets (Sect. 2), present the network analysis (Sect. 3), develop and evaluate gravity models to explain observed patterns (Sect. 4), present the topic modeling approach and findings (Sect. 5), discuss the implications of our results for understanding international discourse patterns in the digital age (Sect. 6), describe in details the data collection and preprocessing, and the methodologies for location extraction and topic modeling (Sect. 7).

2 Data

Our study is based on three distinct datasets, which contain tweets posted between January 1, 2019, and December 1, 2022 respectively geolocated in Europe, South America, and the US. For each region (namely for each dataset), we selected the tweets containing a location in their text which belong to the corresponding region (e.g.: for Europe we kept tweets mentioning places located in Europe). We therefore obtained three filtered datasets: *Europe*, *South America*, and *US*. Details about the identification of these tweets and the locations mentioned in their text can be found in Sect. 7.2 and 7.3. The analysis presented in this manuscript have been performed on these final subsets of tweets. Details about the number of tweets collected for each region of the world and tweets containing locations in their text are presented in Table 1. The full data collection procedure is explained in Sect. 7.1.

The choice of these regions was made because of a mix of considerations related to both our research inquiry and practical constraints. First of all, our methodological approach (see Sect. 7.3) benefits from the fact that English, Spanish, and other high-resource languages, for which enough data is available online to train the main location extraction models, are predominant in these areas. Moreover, we sought to include both multilingual regions (Europe) and areas where only one or two languages across different countries or states are widely spoken outside of local ethnic communities (South America and the US),

Table 1 Geographical areas and volume of collected tweets by region. Number of countries considered (or states for US) in the three regions. We also report the total number of collected tweets between 01-01-2019 and 01-12-2022 for each region. In the last column, we show the number of tweets, per region, that contain locations belonging to that region in their text

World region	Countries/States	Tweets	Tweets mentioning locations in the region
Europe	45	1,350,000	55,907
South America	10	300,000	16,560
United States	49	4,337,352	233,150

as this allows us to examine whether the number of languages commonly used in a region has an impact on our findings. Each region also features numerous internal boundaries with relatively free movement of people. In addition, many countries and states in each region share some common culture and history, have a common market, and are part of a different bigger system (a supranational political and economic union, a continent, a federal country). These characteristics make these three regions comparable in terms of international attention patterns between countries.

3 Who talks about whom? The network of international interests

3.1 Creation of the network

For each dataset (*Europe*, *South America*, and the *United States*), we construct a network representing the shared interests between countries. In these networks, we define the *source* as the country where the tweet is geo-tagged and the *target* as the country of the locations mentioned in the text of the tweet. We quantify the *interest per day*, Φ_{XY}, between a source X and a target Y as the daily frequency with which the source mentions the target. To compute it, we divide the total number of mentions from the source X to the target Y (N_{XY}) by the number of minutes needed to reach the quota for the tweets geolocated in X, t_X, and we multiply it by the number of minutes in one day, 1440. Details about the data collection procedure can be found in Sect. 7.1. Given the time t_X to collect the tweets for country X, the equation yielding the interest per day Φ_{XY} between X and Y follows:

$$\Phi_{XY} = 1440 \cdot \frac{N_{XY}}{t_X} \tag{1}$$

With each dataset, we then create a network of international attention following four main steps: i) determine the corresponding country for each location using a geocoding system; ii) establish directed links between source and target countries; iii) calculate the frequency of mentions per day between each country pair; and iv) construct the complete network of international interests.

Since our study focuses on intra-region communication and interest, we only consider mentions between countries or states belonging to the same region, which leads to the creation of three distinct networks with no overlap, as shown in Fig. 1B. Figure 1A illustrates the network creation pipeline using the Netherlands as the source country and Belgium as the target country, while the complete methodology is detailed in Sect. 7.3. Prior to implementing this approach, we evaluated several alternative methods for location identification and geocoding, as described in Sect. 7.2

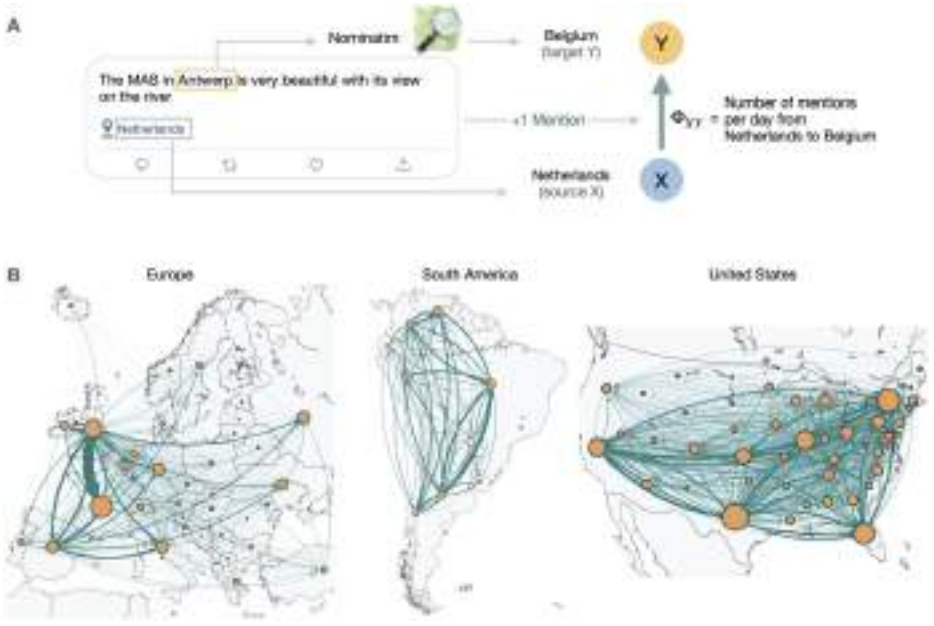

Figure 1 Location extraction and Networks of international interests. **(A)** Example of location extraction from a dummy tweet. The tweet is geolocated in the Netherlands, which identifies the source node (X). The location string "Antwerp" mentioned in the text is sent to Nominatim [30] and the respective country (Belgium) is identified as the target node (Y). A mention is then added in the edge from the Netherlands to Belgium. The total number of mentions between the Netherlands and Belgium (N_{XY}) is computed, and finally the number of mentions per day between the two (Φ_{XY}) is estimated. **(B)** We extract the mentions per day between countries Φ_{XY} in Europe and South America and between states in the United States. The resulting networks of international interests are shown as a network of flows. Nodes shown are sized by in-strength and edges are sized by weights. Data for the boundaries was extracted from the GSHHG dataset [31]

3.2 Analysis of the network

The European weighted network of international interests exhibits remarkably high connectivity, with a density of 0.735. We observe notable asymmetries in certain countries such as Russia, Ukraine, and France, which display high in-strength but low out-strength values. This suggests that these countries are frequently discussed by others while they themselves mention other countries less frequently. Several Western European countries, such as the United Kingdom, Germany, France, Spain, and Italy, receive particularly high mention rates. Regarding out-strength, the United Kingdom and Spain demonstrate the highest propensity to mention other countries.

Some European countries, particularly the smaller ones, exhibit distinctive characteristics in our analysis. Vatican City presents an especially interesting case study. Among tweets geolocated in Vatican City, approximately 24% contain at least one location reference in their text, significantly higher than other European countries where this percentage ranges between 0.5% and 10%. As shown in Sect. 5.1.2, topics related to the Pope and the Vatican Museums occur with remarkable frequency in Vatican City tweets. This distinctive pattern likely reflects the city-state's status as a major tourist and religious destination, with most users tweeting from there being visitors and pilgrims mentioning Vatican City, Rome, and Italy in their posts.

The South American network also displays nearly complete connectivity, with a density of 0.978. Here we observe more balanced in-strength and out-strength distributions

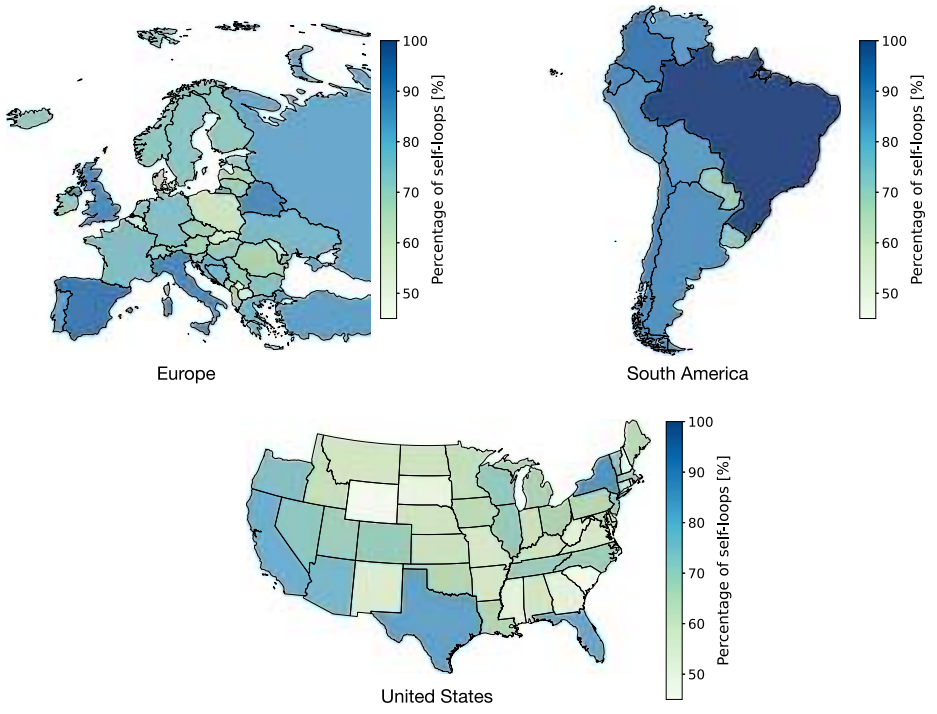

Figure 2 Self-loops in Europe, South America, and the United States. How often, as a percentage of the total number of mentions, each country mentions itself

compared to Europe, with less pronounced inequality between different countries, though larger countries still maintain dominant positions in the network.

Similarly, the United States network shows high connectivity with a density of 0.922. Within the US network, we detect significant variations in in-strength values, with more populated and economically prosperous states such as California, Texas, New York, and Florida receiving substantially more mentions than others.

An intriguing pattern emerges when examining self-loops, i.e., instances where countries mention themselves. As expected, each country predominantly references itself in its social media discourse. Figure 2 illustrates the percentage of self-loops across the three regions, revealing that Spain and the United Kingdom in Europe, Brazil in South America, and New York, California, Florida, and Texas in the United States exhibit the highest proportions of self-references.

Figure 3 presents ego-networks for three representative cases: Germany, Colombia, and Texas, showing both their in-mentions and out-mentions patterns. Germany receives the most mentions from Austria and the Czech Republic, while it predominantly mentions the United Kingdom and France. Colombia receives most mentions from neighboring Venezuela and Ecuador, while it primarily references Venezuela, Brazil, and Argentina. Texas receives the most mentions from nearby states such as Oklahoma, Arkansas, and Louisiana, while it predominantly mentions the most populated states: California, New York, and Florida. Across all ego-networks, we generally observe that countries tend to mention their geographical neighbors, countries sharing linguistic similarities, and countries with major economic and cultural influence in the international landscape [32].

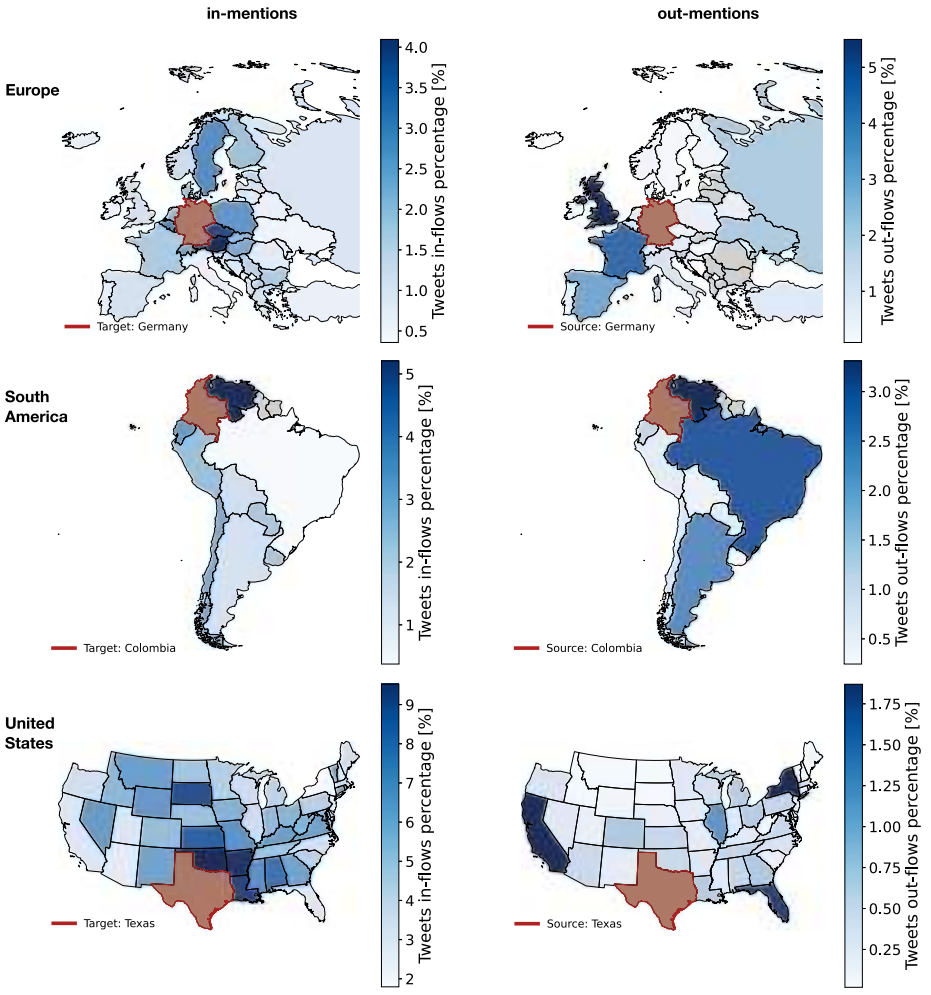

Figure 3 Ego-networks. How many times, as a percentage of the total number of mentions, Germany, Colombia, and Texas are mentioned by other countries or states (in-mentions) or mention them (out-mentions). The percentage of self-loops is not reported here

4 What are the drivers of the mentioning behavior?

To understand what drives the mentioning behaviors shown above, we first perform a preliminary exploratory analysis of the relationship between GDP per capita and the estimated number of tweets produced per day per person in each country. To compute the daily production rate of each country, we divide the number of collected tweets T_X for country X by the number of minutes needed to reach the sample quota, t_X, and we multiply it by the number of minutes in one day, 1440. Given the time t_X to collect the tweets for country X, the equation that computes the daily production rate TD_X for the country X follows:

$$TD_X = 1440 \cdot \frac{T_X}{t_X} \tag{2}$$

Then, we divide this quantity by the population of the country X to gauge the relationship with the GDP per capita. From the analysis, we find a Spearman's rank correlation

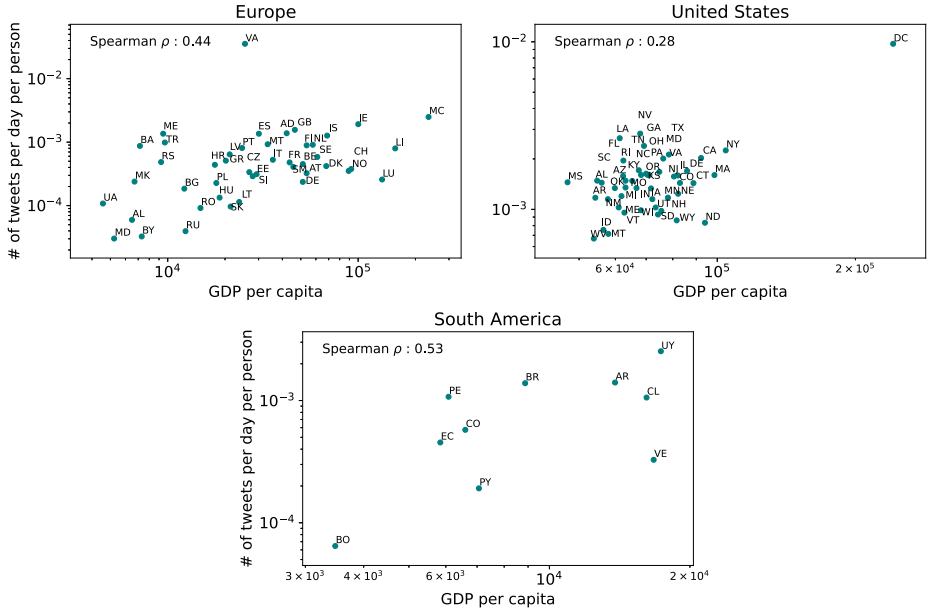

Figure 4 GDP per capita against number of tweets per day per person. We report for Europe, South America, and the United States in log-log scale the GDP per capita of the country (state for the US) versus the average number of tweets generated by an individual in a day in that country. Spearman's rank coefficient is reported

coefficient of 0.44 (p-value 0.002) for Europe, 0.53 (p-value 0.117) for South America, and 0.28 (p-value 0.051) for the United States. This result, as already shown in [24], indicates a modest positive correlation between GDP per capita and Twitter activity, suggesting that citizens of wealthier countries (the ones with higher GDP or socioeconomic development per capita) tend to be somewhat more active on the platform [33]. Figure 4 illustrates this result.

4.1 Augmented gravity model

To further explain the observed patterns in our networks of international interest, we employ a gravity model framework. Following the aforementioned results and insights from previous research [26, 29], we develop a modified gravity model starting from population, GDP, and geographical distance, and then incorporate additional variables to better predict mention frequencies between countries.

As already seen in [7] and in [34], cultural, social, and economic predictors, such as common language, religion, migration, or trade intensity can be added to enrich gravity models. We therefore integrate in our model some additional variables (language and migration flows) that might be leading factors driving the interest between different countries.

We define as Φ_{XY} the interest between countries X and Y, measured as the number of mentions per day from X to Y.

We start from a traditional gravity model that considers as extensive variables either population (p_X and p_Y) or GDP (g_X and g_Y) of the source and target countries, along with their geographical distance, computed as the distance between their centroids (r_{XY}). Based on our findings in Sect. 4 suggesting a relationship between GDP per capita and Twitter

activity, we incorporate GDP as a key predictor rather than population, and we actually observe a better R^2 in this case.

We then develop more sophisticated models by introducing additional predictors: a binary variable indicating whether the two countries share, respectively for Europe and South America, a language family or a specific language ($l_X = l_Y$), and migration flows both outgoing ($m_{XY} : X \to Y$) and incoming ($m_{YX} : Y \to X$) [35, 36]. We use the language predictor for Europe and South America, namely the regions with different official languages across countries, while we do not use it in the United States. In particular, we consider the language families present in Europe (Albanian, Romance, Germanic, Slavic, Finno-Ugric, Maltese, Turkish, Greek) and we set $l_X = l_Y$ if countries X and Y share at least one language family among their official languages. The choice of considering language families was made because it allows us to group countries that share a similar cultural and historical background, such as the Balkan countries or Lithuania and Latvia, or Finland and Estonia. Just checking whether two countries speak exactly the same language would not capture these commonalities, which are important in Europe. For instance, international mobility choices of European students are clearly influenced by language proximity [37]. For South America, instead, we consider Spanish and Portuguese, and we set $l_X = l_Y$ if countries X and Y speak the same language. Since Portuguese is the official language only in Brazil, this variable essentially distinguishes it from other countries.

The basic form of our gravity model using GDP as the mass variable is expressed as:

$$\Phi_{XY} = k\frac{g_X^{\alpha 1} g_Y^{\alpha 2}}{r_{XY}^{\gamma}} \tag{3}$$

where Φ_{XY} represents the estimated interest between source and target, g_X and g_Y are the GDPs of the source and target countries, r_{XY} is the distance between them, and k (normalization factor), α_1, α_2, γ are coefficients to be estimated.

When incorporating additional predictors such as shared language family and migration flows, the model becomes:

$$\Phi_{XY} = k\frac{g_X^{\alpha_1} g_Y^{\alpha_2} \beta(l_X, l_Y) m_{YX}^{\delta}}{r_{XY}^{\gamma}} \tag{4}$$

where

$$\beta(l_X, l_Y) = \begin{cases} \beta^* \text{ if } l_X = l_Y \\ 1 \text{ else} \end{cases} \tag{5}$$

We then consider the logarithmic form and estimate the parameters, as shown in Sect. 4.2.

We also consider the gravity model's version with GDP per capita and population instead of total GDP, but since R^2 does not improve significantly and the two models have equivalent Akaike Information Criterion (AIC) [38] values, we stick with the model presented above. More details about this can be found in Supplementary Materials S5, where we also show that population plays an important role, which, in our model, is captured by total GDP.

Table 2 Gravity model. Coefficients of the predictors in the complete model, which considers the GDP of the source country, the GDP of the target country, the distance between the countries, the migration flow from the target country to the source, and whether the two countries' languages belong to the same language family (or share the same language) or not

	Intercept (K)	Source GDP (α_1)	Target GDP (α_2)	Distance (γ)	Migration $Y \to X$ (δ)	Common language (β^*)
Europe	–12.010	0.636	0.337	0.065	0.242	1.132
South America	–13.748	0.748	0.606	0.309	0.103	1.690
United States	–15.913	0.750	0.706	0.382	0.184	

When examining the topics (see Sect. 5), we observe that even if our data collection was carried out to avoid biases in the topics, some major international events characterize them – for instance, the Russo-Ukrainian war as well as the UK-Europe negotiations and the finalization of Brexit. For this reason, some countries could be more present on the scene than they would have been if we considered other time windows for the data collection. To understand whether the gravity model accurately describes mentioning behavior, we also ran it by removing the tweets, and consequently the mentions between countries, whose identified topic was *war*, *Brexit* or *Vatican City*-related (we also removed this one because of the particular features of Vatican City-related tweets, where we find a disproportionate number of mentions compared to other countries). As shown in Supplementary Materials S6, we do not observe significant changes in R^2, meaning that the gravity model describes the mentioning behaviors well in any case.

4.2 Gravity model: analysis of the predictors

Our analysis was extended systematically from the simplest model (using only population or GDP and distance) to increasingly complex models, incorporating predictors incrementally (language and migration flows) to identify the most explanatory combination. Table 2 presents the coefficients for each predictor in the models that achieve the highest R^2 values across our three study regions. For the United States, language is not included as a predictor since all states share the same official language. As explained in 4.1, we choose the models with no decoupling of GDP and population, which have AIC values respectively equal to 1709 for Europe, 26 for South America, and 1367 for the United States. Moreover, to check our assumptions about language families in Europe 4.1, we try to run both the model with language families and the one with exact language matching. We observe that, in the second case, R^2 is slightly lower (0.6672 instead of 0.6675). For this reason, we prefer to rely on language families rather than exact language matching.

For both Europe and South America, the model with the greatest explanatory power includes GDP, geographic distance, language family, and incoming migration (from target Y to source X). For the United States, the optimal model similarly includes GDP, distance, and incoming migration. The direction of migration effects suggests that people in destination countries tend to discuss the countries from which they originate.

The coefficients in our gravity models can be interpreted as elasticities, providing clear insights into how each factor influences international interest. For instance, the source GDP coefficient (α_1) of 0.636 for Europe indicates that a 1% increase in a country's GDP is associated with approximately 0.64% more mentions of other countries, reflecting how economic capacity enables outward attention. Similarly, the target GDP coefficient (α_2)

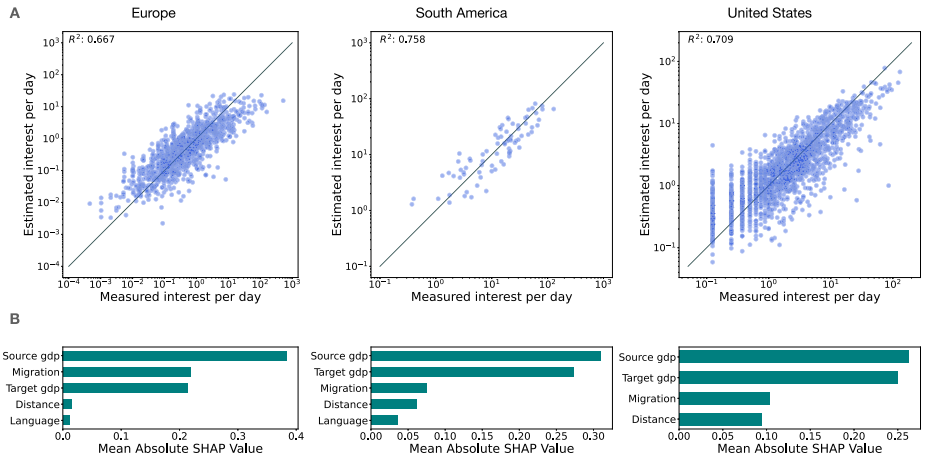

Figure 5 Gravity model. **A)** The first row shows the measured interest per day versus the estimated interest per day from the models with the highest R^2 for each region. **B)** The second row shows the importance of the used predictors for each model through SHAP

of 0.337 suggests that a 1% increase in a country's GDP corresponds to receiving about 0.34% more mentions from other countries, quantifying the relationship between economic prominence and international visibility. The distance coefficient (γ) deserves particular attention for its notable regional variation: the value of 0.065 for Europe is substantially smaller in magnitude than the corresponding values for South America (0.309) and the United States (0.382). This difference means that doubling the distance between two European countries reduces mentions by only about 4.6% ($2^{0.065} \approx 1.046$), while the same distance increase in South America or the United States reduces mentions by approximately 24% ($2^{0.309} \approx 1.239$) and 30.3% ($2^{0.382} \approx 1.303$), respectively. This striking contrast quantifies our observation that geographic proximity plays a much weaker role in determining patterns of attention within Europe compared to the Americas. The migration coefficient (δ) values similarly provide quantitative measures of how population movements shape discursive connections, with the larger coefficient for Europe (0.242) compared to South America (0.103) and the US (0.184) suggesting that migration flows have a stronger influence on international discourse patterns in the European context.

Figure 5 illustrates the relationship between measured interest per day and estimated interest per day derived from our best-performing gravity models for each region. We also present SHAP (SHapley Additive exPlanations) [39] values to quantify the relative importance of each predictor. Across all three regions, the GDP of both source and target countries consistently emerges as the most influential predictor. Migration from target to source also demonstrates substantial importance, particularly for Europe and to a lesser extent for South America and the United States. Geographical distance exerts a stronger influence in the United States and South America models compared to Europe. The language factor has some importance in both linguistically diverse regions but plays a relatively minor role compared to economic and migration factors. As mentioned above, the language factor distinguishes Brazil from the other countries. Even if on a relatively small scale, it negatively influences the interest between Brazil and others. The fact that it is the country that refers to itself the most, as shown in Fig. 2, is another indicator of its relative isolation in this regard.

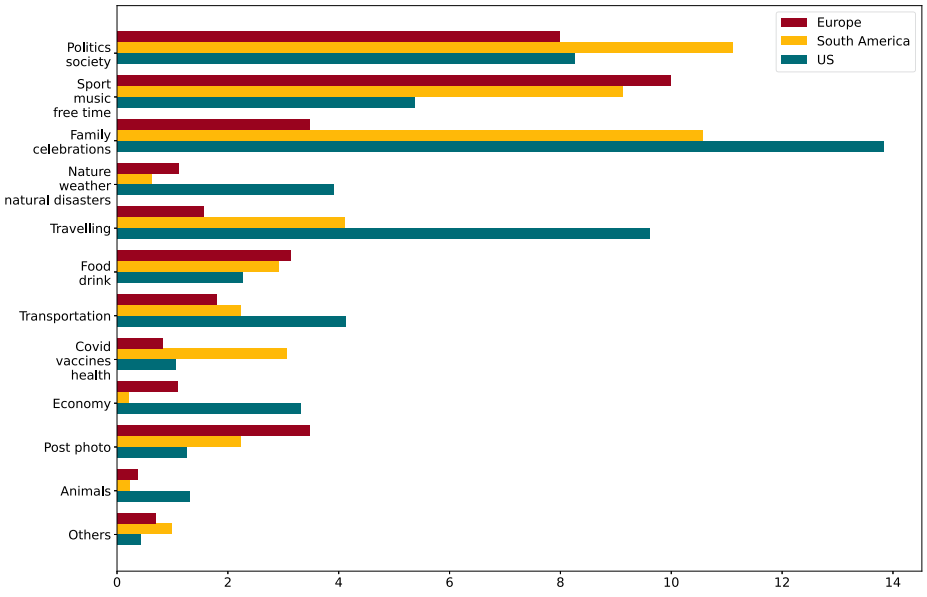

Figure 6 Topic modeling. 11 macro-categories in common among Europe, South America, and the United States are identified. The other minor topics are classified as "Others". For each category, its percentage of tweets for each region is reported

5 Topic modeling

Understanding the content of conversations about locations provides the context in which network structures and interest patterns develop. Previous research explored topics in the international discussion on social media [10, 40]. In our work, we study this dimension by comparing topics in our three datasets. To systematically analyze the textual content characterizing our datasets, we implement topic modeling on tweets containing location mentions in each of our three regional datasets. This task presents unique challenges, as we are working with a diverse corpus of multilingual tweets collected without keyword filtering, making the identification of coherent thematic patterns potentially difficult. In the following subsections, we present a qualitative overview of the content of our datasets.

5.1 Topics description

Our topic analysis reveals both notable similarities and differences in discourse patterns across the three regions. We identify 25 macro-topics per region and 11 macro-areas that appear consistently across all regions, that we define as "common topics", alongside distinctive regional themes that reflect specific cultural, political, and social contexts. Details about the methodology adopted are reported in Sect. 7.4.

Figure 6 displays the distribution of the topics percentages across the three regions. The table in Supplementary Materials S7 presents the distribution of major topic categories across the three regions, including the percentage of tweets belonging to each category and topics labels.

5.1.1 Common topics

Political discourse varies substantially in content while remaining a common topic category. European tweets frequently mention the Russo-Ukrainian war, Brexit negotiations,

and migration and refugee debates. South American tweets often refer to governmental changes, women's rights, and education. US tweets commonly discuss social issues, elections, and political polarization. Variations in political content relate to distinctive regional governance structures and concerns. However, the common presence of political topics demonstrates how governance and policy issues remain central to discussions about places.

Leisure-related discourse constitutes another significant common category, though with regional variations in emphasis. Among sport-related tweets, football/soccer dominates in Europe and South America, while American football is featured prominently in the US. European users mention Formula 1 racing. American users commonly reference baseball, basketball, and golf. The prevalence of sports-related content across all regions underscores the importance of sporting events in creating place associations and driving international interest. A similar finding was observed for football-related events also in [10]. As to cultural activities, we find that concerts, museums, art exhibitions, books feature across all regional datasets. Online shopping is also present as a leisure activity.

Family and celebrations are recurrent themes in the three datasets, where we can find content related to birthdays, Christmas, and other religious events. European tweets frequently mention the Pope and Vatican events. South American tweets present more spiritual contents, using words like "dios" *(God)*, "benediciones" *(Blessings)*, and "misericordia" *(mercy)*. US tweets often contain evangelical Christian terminology, church activities, and religious gatherings.

Environmental and nature topics appear across all regions, with variations reflecting local climate conditions and concerns. European tweets mention snow, rain, earthquakes, and climate change. South American tweets frequently reference wildfires, floods, and earthquakes. US tweets often mention extreme weather events, particularly hurricanes, tornadoes, and wildfires. These regional variations reflect distinctive environmental challenges while highlighting how climate and weather conditions fundamentally shape perceptions and discussions of places.

Travel-related discourse is present across the three regions with a particular interest for art and heritage cities in Europe, landscapes in South America, and road trips in the US. Moreover, photography is often related to holiday experiences.

Culinary discourse constitutes another cross-regional theme, albeit with distinctive regional features. European tweets frequently mention wine, beer, coffee, and restaurant experiences, often in the context of tourism. South American tweets commonly reference restaurants and local production, such as coffee and cocoa. US tweets frequently mention restaurants and coffee chains. The strong presence of food-related content across all regions underlines the major contribution of the culinary dimension to place identity and cross-border interest.

Economy-related topics are present in all three datasets. In Europe the main themes are blockchains and cryptocurrency, in South America commercial activities and enterprises, whereas in the US the focus is on job offers, real estate, and banks.

Transportation references appear consistently throughout the corpus, with users discussing airports, flights, roads, transportation infrastructure and delays. European tweets more frequently reference train travel and cross-border mobility, reflecting the region's integrated rail system. South American tweets mention gasoline and public transports. US tweets disproportionately reference car travel, traffic, and car accidents, in line with

US main transportation modes. These differences highlight the characteristic mobility infrastructures of each region, as well as the importance of transportation in defining relationships between places.

Pandemic-related discourse is featured prominently in tweets across all regions, though with varying emphasis. European tweets more frequently mention vaccination campaigns and health protocols. South American tweets often reference hospital resources and number of cases. US tweets show greater polarization regarding health measures and vaccination. These variations reflect different pandemic experiences while demonstrating how public health crises reshape international discourse about places.

In all three regions, we find location mentions in the context of *photo and video sharing*. Users frequently post messages such as "just posted a photo from Paris" or "sharing a video from Rio," typically with links to Instagram or other image-sharing platforms. This pattern highlights the role of location as a context marker for social media content, particularly for users documenting travel experiences.

Animal-related content shows remarkable consistency across regions, and is typically centered on domestic pets and their company, plus occasional references to wildlife.

5.1.2 Regional topics

In addition to these common categories, we identify distinctive regional topics. European discourse features several distinctive thematic clusters not prominent in other regions.

The Pope and Vatican appear in European tweets relatively often, particularly those geolocated or mentioning Vatican City. As reported in Sect. 5.1, some countries are mentioned in a larger number of tweets, and as observed in Sect. 3.2, the Vatican City is an outstanding example with its large number of location-mentioning tweets. Eurovision Song Contest references are another distinctively European topic, with significant temporal concentration around the annual competition. Brexit and UK-EU relations form a prominent European discourse domain, particularly in tweets from the UK and neighboring countries. Refugee movements and immigration policies also play a major role in European discourse, especially regarding Mediterranean crossings and integration debates. The Russia-Ukraine conflict also emerged as a significant topic in European tweets, especially after the 2022 developments.

South American discourse also reveals several regional themes. References to crime, drug trafficking, and narcotics enforcement appear with remarkable frequency in South American tweets. Justice and feminist movements are also often featured.

The US discourse presents several nationally distinctive topics. Domestic politics, particularly regarding Trump, is a prominent discourse domain distinct from international references to American politics. Racial justice and Black Lives Matter discussions feature prominently in US location-mentioning tweets, especially with reference to specific cities and protest sites. Gun violence and mass shootings are covered pretty often as well. Evangelical Christianity emerges as a distinctive religious discourse in the US tweets compared to the more Catholic-oriented religious references in European and South American content. Cannabis legalization and consumption are another distinctively American discourse domain, particularly regarding states with different regulatory approaches. Abortion and LGBTQ+ rights are another characteristic theme of US tweets.

6 Discussion

In this study, we have analyzed international interest patterns as reflected in social media discourse, constructing networks of location mentions from geo-tagged tweets across three world regions. Our integrated methodology, combining network analysis, gravity modeling, and topic modeling, provides novel insights into the patterns of reciprocal interest of countries in the digital public sphere. Here, we discuss the broader implications of our findings, acknowledge the limitations of our approach, and suggest promising directions for future research.

6.1 Key findings and implications

(RQ1) Our comparative study displays similarities across different world regions, as shown by the network analysis, the gravity model, and the topic modeling results. On the other hand, each region also presents some peculiarities.

(RQ2) The network analysis reveals several consistent patterns in the online mentioning behavior. All networks display high connectivity (edge densities > 0.70), indicating that social media discourse generally covers most countries (or states for the US) within each region rather than exclusively focusing on few selected ones. However, the networks also exhibit pronounced asymmetries in centrality, with certain nodes receiving disproportionate attention.

(RQ3) This asymmetry largely aligns with economic prominence, as our gravity models confirm GDP as the strongest predictor of both incoming and outgoing interest. The substantial influence of economic factors suggests that, despite the seemingly democratizing potential of social media, traditional power structures continue to shape patterns of international attention in digital discourse [41, 42].

The effect of distance on international interest varies markedly across regions, with European discourse demonstrating substantially weaker distance constraints ($\gamma = 0.065$) compared to South America ($\gamma = 0.309$) and the United States ($\gamma = 0.382$). This finding has significant implications for understanding how regional integration affects international discourse. Europe's extensive transportation infrastructure, economic integration, and institutional interconnectedness appear to have effectively reduced the friction of distance in shaping attention patterns [43]. This suggests that as regions become more integrated economically, infrastructurally, politically, the geographic constraints on international discourse may be weakened [44, 45]. The influence of migration on international interest offers particularly noteworthy information on diaspora communities and transnational identities [46–48]. Our models consistently find that migration flows from the target country to the source country positively influence mention frequencies. This pattern suggests that migration creates lasting discursive connections between countries of origin and destination, with immigrant communities maintaining attention to their homelands and potentially serving as cultural mediators in shaping perceptions of their countries of origin [49, 50]. This finding aligns with research on the roles of diaspora communities in cultural diplomacy and transnational public spheres [51, 52], while providing quantitative evidence of these dynamics at scale.

The relatively modest effect of shared language family on interest patterns ($\beta^* = 1.132$ for Europe and $\beta^* = 1.690$ for South America) suggests that linguistic barriers may be less strong in the digital age than traditionally assumed. This could reflect the increasing prevalence of machine translation technologies, the growth of multilingualism among social media users, or the emergence of English as a global lingua franca [53–55].

(RQ4) Our topic modeling results reveal both the universality of certain discourse domains (travel, sports, culture) and the distinctive regional preoccupations that characterize international attention in different contexts. The predominance of tourism-related content across all regions underscores how leisure mobility, with all its internal differentiations, continues to shape international awareness and interest [56]. At the same time, the distinctive political topics in each region (EU integration in Europe, governmental instability in South America, political polarization in the US) highlight how regional political cultures fundamentally shape the substance of international discourse. These content patterns reflect the multifaceted nature of international interest, encompassing both common human concerns and historically specific regional dynamics.

The integration of structural and content analyses enables us to develop a more nuanced understanding of the factors driving international attention patterns. While our gravity models identify some major structural determinants of interest (GDP, distance, migration), our topic models offer insight into the substantive dimension of these structural patterns. Such a dual perspective reveals how economic prominence translates into specific forms of visibility (coverage of cultural events, political developments, tourism destinations) and how migration creates specific discursive connections (discussions of cultural practices, political developments in countries of origin, etc.).

6.2 Methodological contributions

In addition to our substantive findings, this study makes several methodological contributions to the analysis of international communication patterns. Our approach demonstrates the value of integrating geo-tagged social media data with location extraction from text to construct networks of international interest. This methodology captures patterns of attention and discourse that may not involve direct cross-border communication but nonetheless reflect how countries relate to each other in public consciousness.

The application of augmented gravity models to international interest networks provides a useful framework for systematically testing multiple explanatory factors simultaneously. By adapting a modeling approach traditionally adopted for trade and migration to patterns of digital discourse, we show how concepts from spatial interaction modeling can improve our understanding of communication flows in the digital age. This quantitative framework enables precise comparisons of how different factors shape international attention in varying economic, social, and cultural contexts.

Our implementation of multilingual topic modeling using BERTopic illustrates the potential of transformer-based approaches for analyzing thematic patterns in diverse, multilingual corpora. Using contextual embeddings rather than traditional bag-of-words representations, we were able to identify coherent topics across languages and contexts.

6.3 Limitations

Several limitations of our study need to be recognized. First, Twitter (now $\mathbb{X}$) users represent a nonrandom sample of the general population, with notable demographic skews toward younger, more educated, and more urban individuals [57]. This sampling limitation affects the generalizability of our findings to broader patterns of international interest beyond specific forms of social media discourse. Additionally, our focus on geo-tagged tweets (approximately 1-2% of all tweets) introduces potential selection biases, as users who enable location services may differ systematically from those who do not.

Our data collection strategy helped mitigate the influence of big events, in particular Brexit, COVID, and Russo-Ukrainian war, in the international discourse, but they still play an important role as revealed by the topic modeling results. This can still lead to potential specific biases in our dataset, to the detriment of external validity.

Our location extraction methodology, while systematically evaluated and optimized, inevitably misses some location references and incorrectly identifies others. The performance of named entity recognition varies across languages, potentially introducing systematic biases in our detection of locations in less commonly spoken languages. These technical limitations affect the comprehensiveness and precision of our interest networks.

Moreover, our gravity models, while providing substantial explanatory power, necessarily simplify the complex factors shaping international discourse. Cultural affinities, historical relationships, media systems, and geopolitical considerations likely influence international interest in ways not fully captured by our models. The unexplained variance in our models suggests the presence of additional factors beyond the socioeconomic, geographic, linguistic, and migration variables that we included.

Furthermore, our cross-sectional approach cannot capture temporal dynamics in international interest patterns. Major events, evolving geopolitical relationships, and changes in migration patterns, among others, likely cause shifts in interest networks over time. Without a longitudinal analysis, we cannot distinguish stable structural patterns from temporally specific fluctuations in international attention.

Finally, our focus on three world regions, while enabling valuable comparative analysis, cannot ensure that our findings also make sense in different contexts. Patterns of international interest in Africa, Asia, and the Middle East might reveal different structural features and dynamics due to distinctive historical relationships, linguistic landscapes, and geopolitical configurations.

6.4 Future research directions

The above-listed limitations suggest several promising directions for future research. Longitudinal studies of international interest networks are needed to identify how major events reshape attention patterns and to sort out stable structural relationships from temporal fluctuations. Such research could also investigate whether international crises temporarily reduce or amplify the importance of factors such as distance and GDP in shaping cross-border attention.

Expanding the geographical scope to include additional world regions is crucial to solidify the comparative dimension of this research. Studying interest networks in regions with different linguistic landscapes, colonial histories, and geopolitical configurations will help us better appreciate how these contextual factors affect international attention patterns. Particular value would lie in examining regions with distinctive characteristics not present in our current sample, such as Asia's combination of linguistic diversity and rapid economic development. Another research line could shift the focus by also including inter-regional mentions to explore countries' attention towards other regions of the world.

Future studies should moreover incorporate additional predictors into gravity models of international interest, potentially including media system variables, diplomatic relationship metrics, colonial history indicators, or cultural distance measures, among others. Such expanded models are likely to explain at least in part the variance currently not accounted for by the socioeconomic, geographic, linguistic, and migration factors in our approach.

Some additional methodological innovations are called to strengthen our research agenda. Integrating sentiment analysis with location extraction will allow to distinguish positive, negative, and neutral mentions, providing a richer understanding of the mechanisms of international attention. Natural language inference techniques will help identify the semantic relations between locations, distinguishing comparative mentions from simple co-occurrences. Finally, network analysis techniques such as community detection and role analysis will likely reveal more complex and elusive structural patterns in international interest networks.

Finally, combining social media data with other sources of international discourse (news media coverage, web search patterns, diplomatic statements, etc.) will provide a more comprehensive picture of international attention patterns across communication channels. Such multi-platform analysis is essential to understand how different discourse arenas potentially construct different geographies of attention and interest.

6.5 Concluding remarks

This study has introduced a novel approach to understanding international attention patterns through the lens of location mentions in social media discourse. By constructing networks of interest between countries based on geo-tagged tweets, we have shown how digital communication reflects traditional geographies of attention in the international system. The interactions we observe between economic factors, geographic proximity, linguistic similarity, and migration flows suggest that the social mechanics of digital discourse sit at the intersection of established structural forces and emerging communication practices.

Looking forward, the study of international interest networks offers promising avenues for broader theoretical inquiry. First, these networks provide a window into what might be termed "discursive globalization", namely how digital communication creates new forms of awareness and attention across borders that may complement or challenge traditional diplomatic, economic, and cultural exchanges. Second, the asymmetries in these networks raise important questions about discursive power and visibility in the international system, suggesting that digital platforms may be reproducing or even amplifying existing inequalities in global attention rather than democratizing international discourse [58] [59].

The methodological framework we have developed, that combines location extraction, network analysis, gravity modeling, and topic modeling, offers a template for future research examining how attention flows between geographical entities across various digital platforms and communication contexts. By extending this approach to different platforms, time periods, and world regions, researchers can build a more comprehensive understanding of how digital discourse shapes, and is shaped by, international relations.

As communication technologies continue to evolve and global connectivity deepens, the patterns of international interest identified in this study will likely be transformed as well. Tracking these transformations will be essential for understanding how digital communication is reshaping the geography of global attention and, potentially, the conduct of international relations more broadly. Digital discourse does not merely reflect the world: it increasingly becomes a sphere in which relationships between nations are enacted, contested, and re-imagined. The development of more sophisticated approaches to the analysis of these discursive relationships will enhance our understanding of the complex interplay between digital communication and the evolving international order.

The global conversations mapped in our study also offer valuable insights into emerging forms of digitally enabled collective intelligence. As diverse participants across geographical boundaries contribute their local knowledge, perspectives, and experiences to shared discursive spaces, they potentially generate distributed understanding of complex phenomena that no single observer or institution could achieve alone. The patterns of international interest we have identified may serve as indicators of how information flows through this global cognitive system, highlighting which nodes function as key information sources or attention hubs, which regions are more isolated or connected, and how economic and demographic factors condition these knowledge exchanges [60] [61]. Understanding these patterns could prove invaluable for tracking the evolution of major socially challenging topics. For instance, asymmetrical patterns of attention to political events reveal how polarization may spread across borders, and the variation in discourse about social conflicts indicates which kinds of tensions generate broader international concern. Developing more sophisticated methods to analyze these digital conversation networks is key for researchers and policymakers to better anticipate emerging social challenges, identify cross-border solidarity developing around particular issues, and understand how collective awareness of global problems develops across different regional and economic contexts. This approach recognizes social media not merely as a dataset for academic inquiry but as an evolving system of distributed cognition with significant implications for how societies identify, understand, and potentially address shared challenges.

In conclusion, the integration of computational methods with theoretical perspectives from international relations, human geography, and communication studies offers a particularly fertile ground for future interdisciplinary research. Such integration can help illuminate not only the structural patterns of international discourse but also their implications for global governance, cultural exchange, and political developments across borders. In an era where digital communication increasingly mediates our understanding of the world beyond our immediate experience, mapping the flows of international attention provides crucial insight into how the global sphere is imagined, discussed, and ultimately shaped through everyday discourse.

7 Methods

7.1 Data collection

The data collection was performed at the beginning of 2023. For each dataset, we collected a random sample of tweets posted between January 1, 2019, and December 1, 2022, using the Academic Twitter API [62], when it was still freely available. Three big events occurred in this time frame, as noted in Sect. 5: Brexit negotiations, COVID, and the Russo-Ukrainian war. To reduce potential biases from these and other major events that might have dominated users' interest during specific time periods, we divided our four-year timeframe into small 2-minute intervals and sequentially selected intervals at random from which we collected tweets until our desired sample quota was reached. Thus, a larger number of intervals has been queried for small countries or countries where Twitter is less used. During the data collection, we queried for a blank space (the Twitter API's query required the keyword field to be not empty) to retrieve tweets unrelated to specific keywords in order to capture the general discourse without focusing on specific topics. For the European dataset, we considered 45 countries and collected 30,000 geo-tagged tweets for

each of them, using the *place_country* field to specify the country in our API queries. For South America, we followed the same approach, collecting 30,000 geo-tagged tweets for each of the 10 countries in the region. For South America, we removed French Guyana since it is actually part of the French territory, and we also decided not to include Guyana and Suriname (two ex-colonies respectively of Great Britain and the Netherlands) in order to focus on the more homogeneous set of the more populous Latin countries. Finally, we considered only the contiguous United States (48 US states sharing a border and the District of Columbia) to simplify the spatial analysis, as distances on land and across the sea might have different weights. The full list of selected countries and states is available in Section S4 of Supplementary Materials. We collected a total of 4,337,352 tweets geolocated within the country in 49 US states (eventually excluding Washington State due to disambiguation issues, as explained in Sect. 7.3.1). Due to the limitations of the Twitter API, we could not query directly by state, so we collected tweets geolocated in the US and subsequently assigned them to the appropriate states based on the *place_name* field.

7.2 Locations extraction and geocoding: tools evaluation

Before implementing our full analytical pipeline, we conducted a systematic evaluation of available tools and approaches to identify the most effective methods for our study. We first assessed three geocoding systems: Nominatim [30], Geonames [63], and ArcGis [64], to determine which one most accurately identified the correct country for a given location. Using a manually annotated dataset containing 363 global locations, namely strings representing places such as "Paris" or "The Big Apple", with their country code, we evaluated each system's performance by means of the score $(C - I)/T$, where C represents correct matches for the countries, I represents incorrect matches, and T represents the total number of locations in the dataset. Nominatim demonstrated superior comparative performance with a score of 0.808, making it our preferred geocoding solution.

Next, we evaluated different methods for extracting location entities from tweets. We first created a gold standard of 371 tweets, where we manually identified and annotated all location references, recording the correct matching places in Nominatim's output format to facilitate automated comparison. We then compared three tools to perform location entity extraction: Stanza [65], TweetNLP [66], and Twitter's native Named Entity Recognition tool, whose results are reported in the subfield *Place* of the *annotations* field of the tweet [67, 68]. For each method, we extracted potential location entities from the text, geocoded them using Nominatim, and compared the results against our reference dataset using the same scoring metric. Twitter's built-in entity extraction achieved the highest performance with a score of 0.887, making it our primary method for location identification, supplemented by Stanza[1] for languages not supported by Twitter's algorithm. While Stanza enables the identification of locations in tweets written in other languages, its performance is somewhat inferior to Twitter's tool since Stanza's models were not specifically trained on the unique linguistic characteristics of tweets, which feature short sentences and limited contextual information. However, it helps identify locations for specific languages.

[1] https://stanfordnlp.github.io/stanza/ner_models.html.

7.3 Processing pipeline

Our methodology employs a systematic pipeline to extract and analyze location mentions in tweets. For each geo-tagged tweet, we first determine whether it contains location references by using the strategy presented before in Sect. 7.2, which examines the subfield *Place* of the *annotations* field of the tweets and supplements Stanza for the European dataset, which contains a greater variety of languages not supported by Twitter's tool. Details about the language of the tweets containing locations in their text and about the improvement of the identification of locations by adding Stanza can be found in Supplementary Materials S1-S2. The majority of tweets containing locations are in English or Spanish, both because these are predominant languages across our three regions of interest and because Twitter's entity detection performs optimally for these languages.

After extracting location strings from tweet text, we submit them to Nominatim to retrieve their geographical information, particularly focusing on the country (or state for the US) associated with each location. Having identified both the source country (where the tweet originated) and the target countries (mentioned in the tweet), we establish directed links between them. We then quantify these connections by counting how many times each source country X mentions each target country Y, and compute the interest Φ_{XY} by dividing the number of mentions (N_{XY}) by the time needed to reach the tweets quota from X, as detailed in Eq. (1). This procedure is carried out to normalize the collected data to get a comparable measure across all the countries. This process yields a directed and weighted country-to-country network where nodes represent countries and edge weights Φ_{XY} correspond to the interest of source country X for target country Y. For analytical clarity, we exclude self-loops (where source and target countries coincide) from this network.

Figure 1A illustrates this methodology using an example with the Netherlands as the source country and Belgium as the target country.

7.3.1 Anomalies in location identification

During our analysis, we identified and addressed several anomalies in the location extraction process. In the European dataset, we observed that the string "EU" was being incorrectly mapped to "Eu", a small town in France, rather than the European Union. By examining the tweet content, we determined that users were clearly referencing the European Union, not the French town. We systematically identified these instances and removed them from our analysis as they did not reference a specific country but rather the supranational organization.

A second issue concerned Washington State versus Washington DC. After constructing the US network, we noticed an unusually strong connection between Washington State and Washington DC. A manual inspection revealed that when "Washington" appeared in the tweet text, Nominatim consistently identified it as Washington DC, regardless of whether the user was referring to the Western state or the Capital District. To address this ambiguity, we adopted a context-based approach: when "Washington" appeared alongside another location (e.g., "Seattle, Washington"), we submitted both locations together to Nominatim, which typically resolved the ambiguity correctly. However, when contextual disambiguation proved impossible, we removed these tweets from our analysis. To prevent potential biases, we ultimately excluded Washington State from our analysis entirely, reducing our US tweet corpus from 4,337,352 to 4,256,084 tweets. Importantly, this

exclusion did not significantly alter our findings; the general network characteristics and gravity model outcomes remained consistent.

7.4 Topic modeling

For each region (Europe, South America, United States), we focused exclusively on tweets containing location references within the respective geographical area. The data preprocessing consisted of removing the locations themselves (to avoid having topics identified by the locations, which would have made our analysis not informative), URLs, emojis, numbers, user mentions, and hashtag symbols while preserving the text. Before performing topic modeling we also removed stopwords according to the languages present in each corpus. So, we got three refined subsets: 46,065 European tweets, 14,947 South American tweets, and 192,202 US tweets. We then carried out topic modeling separately on each regional corpus.

The number of tweets per country is not equally distributed since the number of tweets containing locations is different for different countries (see the details in Supplementary Materials S3). This means that the topics do not represent every country in equal measure, but they have to be considered as an overall representation of the three macro-regions (Europe, South America, and the United States).

We employed BERTopic [69], a topic modeling approach that leverages transformer-based language models and clustering techniques. Unlike traditional topic modeling methods such as LDA (Latent Dirichlet Allocation), which rely on bag-of-words representations, BERTopic captures semantic relationships through contextual embeddings. This is particularly advantageous for our multilingual corpus with its short-form content since it helps solve ambiguities in meaning and context such as words with multiple meanings or limited contextual clues. We extracted embeddings using the `paraphrase-multilingual-mpnet-base-v2` pretrained model, which effectively handles multiple languages and was specifically fine-tuned for semantic similarity tasks. We then reduced the dimensionality of these embeddings using UMAP (Uniform Manifold Approximation and Projection) [70] and clustered them with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) [71]. In configuring HDBSCAN, we calibrated the minimum topic size parameter, i.e., the minimum number of documents (tweets in our case) required to form a distinct topic, according to each dataset's size: 70 for Europe, 30 for South America, and 150 for the United States.

Then, we reduced the number of topics by setting the *nr_topics* parameter, which allows us to automatically find the most similar topics and merge them into macro topics. Finally, since the topics are assigned numbers rather than names, and are therefore not easily interpretable, we took the 30 most relevant words for each topic and used them to prompt[2] a multilingual large language model (LLM), `Aya Expanse 32B`, to generate interpretable labels. We tried setting different numbers of topics, and selected 25 because the related result was the most interpretable. These values were carefully selected to balance granularity with interpretability, allowing us to identify a manageable number of meaningful topics while preserving informative clusters.

[2]Prompt used: "The following words come from a topic cluster generated using BERTopic on a dataset of tweet posts. Please generate a descriptive label in English that summarizes the main theme of the cluster. The label should be concise (4–5 words max) and accurately reflect the topic. Return only one potential name."

We manually checked and compared the 25 macro topics for each region and systematically categorized them into broader thematic areas to facilitate cross-regional comparison, as shown in Sect. 5.1. This final categorization was made to group topics into meaningful macro areas, where semantically far topics could also be inserted if belonging to the same umbrella category (eg: Brexit and Russo-Ukrainian war represent semantically far concepts, but we can consider both of them as social/political issues). The detailed labels assigned by the LLM and their manual categorization into macro-areas are visible in Section S7 of Supplementary Material.

Abbreviations
NLP, Natural Language Processing; NER, Named Entity Recognition; GDP, Grosss Domestic Product; AIC, Akaike Information Criteria; SHAP, SHapley Additive exPlanations; LDA, Latent Dirichlet Allocation; UMAP, Uniform Manifold Approximation and Projection; HDBSCAN, Hierarchical Density-Based Spatial Clustering of Applications with Noise; LLM, Large Language Model.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-025-00597-z.

Additional file 1. (PDF 2.0 MB)

Acknowledgements
The authors thank Maurizio Napolitano and Gian Maria Campedelli for the insightful discussions.

Author contributions
Conceptualization, T.L. and R.G.; methodology, V.O., S.B., T.L. and E.L.; data analysis, V.O.; data curation, V.O.; writing—original draft preparation, V.O. and P.S. writing—review and editing, S.B., T.L., E.L., A.P.A., S.T., R.G.; visualization, V.O. and S.B.; supervision, A.P.A, S.T. and R.G. All authors have read and agreed to the published version of the manuscript.

Funding information
S.B., T.L. and R.G. acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU.

Data availability
The datasets supporting the conclusions of this article are available in the Tweet2Geo repository, https://github.com/vorsanigo/Tweet2Geo.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy. [2]University of Trento, Trento, Italy. [3]University of Chieti-Pescara, Chieti, Italy.

Received: 15 May 2025 Accepted: 26 October 2025 [illegible]

References

1. Zhang X, Yang Q, Albaradei S, Lyu X, Alamro H, Salhi A, Ma C, Alshehri M Jaber II Tifratene F, et al (2021) Rise and fall of the global conversation and shifting sentiments during the covid-19 pandemic. Humanit Soc Sci Commun 8(1):1–10
2. Barnett GA, Xu WW, Chu J, Jiang K, Huh C, Park JY, Park HW (2017) Measuring international relations in social media conversations. Gov Inf Q 34(1):37–44
3. Kouloukoui D, de Marcellis-Warin N, da Silva Gomes SM, Warin T (2023) Mapping global conversations on Twitter about environmental, social, and governance topics through natural language processing. J Clean Prod 414:137369
4. Acemoglu D, Bimpikis K, Ozdaglar A (2014) Dynamics of information exchange in endogenous social networks. Theor Econ 9(1):41–97

5. Carpentier N, Dahlgren P, Pasquali F (2013) Waves of media democratization: a brief history of contemporary participatory practices in the media sphere. Convergence 19(3):287–294
6. Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. Soc Netw 34(1):73–81
7. García-Gavilanes R, Mejova Y, Quercia D (2014) Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication. In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing, pp 1511–1522
8. Hänska M, Bauchowitz S (2019) Can social media facilitate a European public sphere? Transnational communication and the europeanization of Twitter during the eurozone crisis. Soc Media Soc 5(3):2056305119854686
9. Sinanan J, Horst HA (2022) Communications technologies and transnational networks. In: Handbook on transnationalism. Edward Elgar Publishing, pp 371–387
10. Leonardelli E, Tonelli S (2024) The geography of information diffusion in online discourse on Europe and migration. In: Proceedings of the international AAAI conference on web and social media, vol 18, pp 904–916
11. Alieva I, Moffitt JD, Carley KM (2022) How disinformation operations against Russian opposition leader alexei navalny influence the international audience on Twitter. Soc Netw Anal Min 12(1):80
12. Twitter geo data. https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data
13. Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E (2013) Mapping the global Twitter heartbeat: the geography of Twitter. First Monday
14. Zohar M (2021) Geolocating tweets via spatial inspection of information inferred from tweet meta-fields. Int J Appl Earth Obs Geoinf 105:102593
15. Han B, Cook P, Baldwin T (2014) Text-based Twitter user geolocation prediction. J Artif Intell Res 49:451–500
16. Zheng X, Han J, Sun A (2018) A survey of location prediction on Twitter. IEEE Trans Knowl Data Eng 30(9):1652–1671
17. Hoang TBN, Mothe J (2018) Location extraction from tweets. Inf Process Manag 54(2):129–144
18. Lin Y-C, Lai C-M, Chapman JW, Wu SF, Barnett GA (2018) Geo-location identification of Facebook pages. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE, pp 441–446
19. Botta F, Moat HS, Preis T (2015) Quantifying crowd size with mobile phone and Twitter data. R Soc Open Sci 2(5):150162
20. Barnett GA, Nam Y (2024) A network analysis of international migration: longitudinal trends and antecedent factors predicting migration. Glob Netw 24(2):e12455
21. Bassolas A, Lenormand M, Tugores A, Gonçalves B, Ramasco JJ (2016) Touristic site attractiveness seen through Twitter. EPJ Data Sci 5(1):12
22. Hedayatifar L, Morales AJ, Bar-Yam Y (2020) Geographical fragmentation of the global network of Twitter communications. Chaos: interdiscip J Nonlinear Sci 30(7)
23. Eleta I, Golbeck J (2014) Multilingual use of Twitter: social networks at the language frontier. Comput Hum Behav 41:424–432
24. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271
25. Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: models and applications. Phys Rep 734:1–74
26. Crymble A (2019) Introduction to gravity models of migration & trade. The Programming Historian
27. Bontorin S, Centellegher S, Gallotti R, Pappalardo L, Lepri B, Luca M (2025) Mixing individual and collective behaviors to predict out-of-routine mobility. Proc Natl Acad Sci USA 122(17):e2414848122
28. Wilson AG (1971) A family of spatial interaction models, and associated developments. Environ Plan A 3(1):1–32
29. Coşkuner Ç, Sogah R (2023) Augmented gravity model of trade with social network analysis. Sustainability 15(19):14085
30. Nominatim. https://nominatim.org/
31. GSHHG dataset. https://www.soest.hawaii.edu/pwessel/gshhg/
32. Wu HD, Groshek J, Elasmar MG (2016) Which countries does the world talk about? An examination of factors that shape country presence on Twitter. Int J Commun
33. Sohail SS, Khan MM, Madsen DØ, Alam MA, Irshad RR (2025) Geospatial and linguistic analysis of Twitter behavioral trends: examining the impact of socioeconomic development on social media use. Hum Behav Emerg Technol 2025(1):1376983
34. Wang L, Chen K (2025) The impact of trade liberalization on China–asean trade relations along the belt and road: an augmented gravity model analysis. Finance Res Lett 71:106418
35. Ortiz-Ospina E, Roser M, Ritchie H, Spooner F, Gerber M (2022) Migration. Our World in Data. https://ourworldindata.org/migration
36. US migration flows. (2022). https://www.census.gov/data/tables/time-series/demo/geographic-mobility/state-to-state-migration.html
37. Ovchinnikova E, Van Mol C, Jones E (2023) The role of language proximity in shaping international student mobility flows. Glob Soc Educ 21(4):563–574
38. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control. 19(6):716–723. Math Rev. 423716
39. Shapley LS Notes on the N-Person Game — II: the Value of an N-Person Game
40. Ji J, Robbins M, Featherstone JD, Calabrese C, Barnett GA (2022) Comparison of public discussions of gene editing on social media between the United States and China. PLoS ONE 17(5):e0267406
41. Couldry N (2015) The myth of 'us': digital networks, political change and the production of collectivity. Inf Commun Soc 18(6):608–626
42. Entman RM, Usher N (2018) Framing in a fractured democracy: impacts of digital technology on ideology, power and cascading network activation. J Commun 68(2):298–308
43. Kleinschmidt C (2010) Infrastructure, networks, (large) technical systems: the 'hidden integration'of Europe. Contemp Eur Hist 19(3):275–284
44. Meyer CO (2005) The europeanization of media discourse: a study of quality press coverage of economic policy co-ordination since Amsterdam. J Common Mark Stud 43(1):121–148

45. Koopmans R, Statham P (2010) The making of a European public sphere: media discourse and political contention. Cambridge University Press, Cambridge
46. Kreis R (2017) # refugeesnotwelcome: anti-refugee discourse on Twitter. Discourse Commun 11(5):498–514
47. Goodman S, Kirkwood S (2019) Political and media discourses about integrating refugees in the uk. Eur J Soc Psychol 49(7):1456–1470
48. Mustafa-Awad Z, Kirner-Ludwig M (2021) Syrian refugees in digital news discourse: depictions and reflections in Germany. Discourse Commun 15(1):74–97
49. Andersson K (2019) Digital diaspora: an overview of the research areas of migration and new media through a narrative literature review. Hum Technol 15(2):142–180
50. Ponzanesi S (2020) Digital diasporas: postcoloniality, media and affect. Interventions 22(8):977–993
51. Stone D, Douglas E (2018) Advance diaspora diplomacy in a networked world. Int J Cult Policy 24(6):710–723
52. Collins N, Bekenova K (2020) European cultural diplomacy: diaspora relations with Kazakhstan. In: Cultural diplomacy and international cultural relations: volume I. Routledge, London, pp 74–92
53. Lenihan A (2014) Investigating language policy in social media: translation practices on Facebook. In: The language of social media: identity and community on the Internet. Springer, Berlin, pp 208–227
54. Dewey M, Jenkins J (2010) English as a lingua franca in the global context: interconnectedness, variation and change. Contending Glob World Engl 9:72–92
55. Cabrera L (2024) Babel fish democracy? Prospects for addressing democratic language barriers through machine translation and interpretation. Am J Polit Sci 68(2):767–782
56. Strömblad E (2024) Identifying mobility segments for leisure travel: a cluster analysis based on a one-month travel survey. Transp Res, Part A, Policy Pract, 181:104001
57. Carrascosa JM, Cuevas R, Gonzalez R, Azcorra A, Garcia D (2015) Quantifying the economic and cultural biases of social media through trending topics. PLoS ONE 10(7):e0134407
58. Blank G (2017) The digital divide among Twitter users and its implications for social research. Soc Sci Comput Rev 35(6):679–697
59. Messias J, Vikatos P, Benevenuto F (2017) White, man, and highly followed: gender and race inequalities in Twitter. In: Proceedings of the international conference on web intelligence, pp 266–274
60. Makowsky MD, Rubin J (2013) An agent-based model of centralized institutions, social network technology, and revolution. PLoS ONE 8(11):e80380
61. Sacco PL, Gallotti R, Pilati F, Castaldo N, De Domenico M (2021) Emergence of knowledge communities and information centralization during the covid-19 pandemic. Soc Sci Med 285:114215
62. Twitter API. https://developer.twitter.com/en/docs/twitter-api
63. Geonames. https://www.geonames.org/
64. ArcGis. https://www.arcgis.com/index.html
65. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD (2020) Stanza: a python natural language processing toolkit for many human languages. arXiv preprint. arXiv:2003.07082
66. Camacho-Collados J, Rezaee K, Riahi T, Ushio A, Loureiro D, Antypas D, Boisson J, Espinosa-Anke L, Liu F, Martínez-Cámara E, et al (2022) Tweetnlp: cutting-edge natural language processing for social media. arXiv preprint. arXiv:2206.14774
67. Twitter annotations. https://developer.twitter.com/en/docs/twitter-api/annotations/overview
68. Twitter annotations FAQ. https://developer.twitter.com/en/docs/twitter-api/annotations/faq
69. Grootendorst M (2022) Bertopic: neural topic modeling with a class-based tf-idf procedure. arXiv preprint. arXiv:2203.05794
70. McInnes L, Healy J, Melville J (2018) Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint. arXiv:1802.03426
71. Campello RJ, Moulavi D, Sander J (2013) Density-based clustering based on hierarchical density estimates. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, pp 160–172

Publisher's note