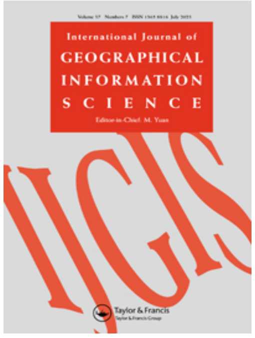

Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

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

Content-location relationships: a framework to explore correlations between space-based and place-based user-generated content

Vicente Tang & Marco Painho

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2023.2213869

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

Content-location relationships: a framework to explore correlations between space-based and place-based user-generated content

Vicente Tang and Marco Painho

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Lisbon, Portugal

ABSTRACT

The use of social media and location-based networks through GPS-enabled devices provides geospatial data for a plethora of applications in urban studies. However, the extent to which information found in geo-tagged social media activity corresponds to the spatial context is still a topic of debate. In this article, we developed a framework aimed at retrieving the thematic and spatial relationships between content originated from space-based (Twitter) and place-based (Google Places and OSM) sources of geographic user-generated content based on topics identified by the embedding-based BERTopic model. The contribution of the framework lies on the combination of methods that were selected to improve previous works focused on content-location relationships. Using the city of Lisbon (Portugal) to test our methodology, we first applied the embedding-based topic model to aggregated textual data coming from each source. Results of the analysis evidenced the complexity of content-location relationships, which are mostly based on thematic profiles. Nonetheless, the framework can be employed in other cities and extended with other metrics to enrich the research aimed at exploring the correlation between online discourse and geography.

ARTICLE HISTORY
Received 3 March 2022
Accepted 4 May 2023

KEYWORDS
Content-location relationships; UGC; geo-tagged activity; topic modeling

1. Introduction

Cities are multi-layered systems, hosting complex human-environment interactions in the form of activities, functions, flows, places and meanings embedded into the surrounding urban landscape (Gao *et al.* 2017, Iranmanesh *et al.* 2022). Today, the widespread use of online platforms through mobile phones and location-based services provides fast and voluminous georeferenced data in urban areas. Geospatial big data from user-generated content (UGC) are a major data source for urban studies, with applications such as identifying regions of interest (Shang *et al.* 2016), examining urban

CONTACT Vicente Tang vtang@novaims.unl.pt

perception and functional structure (Hu *et al.* 2021), unraveling mobility patterns, mapping sentiments, among many others (Belcastro *et al.* 2021, Gao *et al.* 2021). When location is attached to published content, users are regarded as social sensors and their footprints are often used as a spatial proxy for obtaining place-based information (Goodchild 2007, Papadakis *et al.* 2020). Therefore, exploring the degree of correlation between space-based thematic information and the surrounding place-based information is crucial when geo-tagged UGC is an ubiquitous source of data in the literature.

Platforms such as Twitter, with approximately 200 million daily active users and more than 500 million tweets per day (Jay 2022), provide extensive location-based data in densely populated areas. Textual information linked to geographic coordinates, however, do not necessarily reflect the thematic signatures associated with the geographic context from where the user has posted (McKenzie and Adams 2017). Another widely exploited type of location-based UGC data in the literature is represented by points of interest (POI), sourced from platforms such as OpenStreetMap (OSM), Foursquare and Yelp (Niu and Silva 2020). Attributes such as thematic tags, user reviews and number of check-ins are more enriched in place-based information and better mirror the encompassing spatial context, portrayed by place names, functions and affordances (Psyllidis *et al.* 2022). The discrepancies between content and location are naturally more evident in geo-tagged activity from social networks (eg Twitter and Instagram), where relationships between what is said and where it is said are not obvious.

Previous research has addressed the content-location relationships through different lenses. One such avenue is the inference of geography in both non-georeferenced and geo-tagged textual content by extracting topononyms, geocoding as well as investigating 'geo-indicativeness' – the degree to which lexica semantically indicates geographic features (Adams and Janowicz 2021, Melo and Martins 2017, Qiu *et al.* 2022). In place semantics, natural language processing (NLP) methods have been applied to geo-tagged UGC to obtain thematic and cognitive dimensions of places (Hu 2018a, 2021). In NLP, efforts to incorporate location into topic modeling algorithms are examples of how researchers have acknowledged that location is not just another attribute, but is often intertwined with content (Bo and Martin 2013, Wang *et al.* 2020). However, studies that focus specifically on addressing the extent to which location and content are related in geo-tagged social media activity are still scarce. In addition, most of the works employing topic modeling rely on the latent Dirichlet allocation (LDA), which has been outperformed by more recent algorithms, making room for improvements regarding the NLP methods of choice (Egger and Yu 2022).

The extent to which everyday conversation in social networks is geo-indicative may vary depending on temporal and spatial scales, as well as the thematic signatures of the text (Gao *et al.* 2017, Fu *et al.* 2018). As these data sources continue to support research in urban studies, we need to outline reproducible and straightforward steps aimed at assessing the correlation between text and the urban environment for a given city (de Oliveira and Painho 2021). Limitations found in previous works include employing outdated topic models, relying on manual classification steps, performing content analysis based on individual point-based short-text activity and restricting context information to place types (Hahmann *et al.* 2014, Herfort *et al.* 2014, McKenzie and Adams 2017). Furthermore, traditional bag of words topic models, such as LDA, do not consider the syntactic and semantic relationships between words within a document, but recent algorithms are supported by methods that enable the contextualized representation of words (Yang *et al.* 2023). As content-location correlations are the bridge between spatial context and the content of online activity, the efforts to investigate these relationships should invest in up-to-date topic modeling techniques.

In this paper, we introduce a framework for modeling and comparing similar thematic signatures derived from space-based and place-based online activity. The content-location relationships are better represented as the relations between topics originated from geo-tagged social media text and those from POI reviews and tags. Since georeferenced social media data reveals information that is attached to space while not necessarily being thematically related to it, we refer to these sources as space-based. In contrast, POI information and reviews are considered place-based sources as they are better at disclosing urban functions, affordances and perceptions that describe and are related to space. Our contribution relies on providing a methodological framework that can be employed in other cities to enhance the content-location discussions and that is based on more recent methods for topic modeling which have not been applied for this task, more specifically the Bidirectional Encoder Representations from Transformers topic model (BERTopic, Grootendorst (2022)). We also attempt to improve previous efforts by aggregating textual content based on a grid, extracting statistically significant thematic regions, using metrics to objectively assess spatial and thematic similarity, as well as using place reviews as our proxy of the urban landscape. The framework is employed using geo-tagged Twitter posts as our space-based source and reviews and names from Google Places and OpenStreetMap as our place-based reference. All platforms provide large datasets from extensive activity in the majority of urban regions across the globe, including the city of Lisbon (Portugal), where we chose to test the framework.

The remainder of this paper is organized as follows. In Section 2, we present the literature that covers theoretical and methodological aspects of our study. Sections 3 and 4 bring forward our data and methods respectively, from which we obtained our results, found in Section 5. Section 6 is dedicated to our interpretation and discussion of the findings, and lastly, we present our concluding remarks in Section 7.

2. Background and related work

2.1. The relationships between content and location in social networks

Natural language in explicitly geo-tagged social media activity can either disclose information *about* a place or merely *from* a place (Hu 2018b). In both cases, content may be influenced or caused by features and events from users' origins at different scales, such as the locale, neighborhood, city and country. This is particularly exploited in previous works that analyze geo-tagged social media data for situational awareness and emergency response in natural disasters including floods, earthquakes and typhoons (Herfort *et al* 2014, Huang and Xiao 2015, Suwaileh *et al.* 2022). Although extreme circumstances might generate a higher correlation between content and location, geo-tagged user-generated content can also reflect everyday urban life. In GIScience, these have become common sources for spatially assessing urban thematic characteristics derived from websites, digital gazetteers, social media (Twitter, Foursquare, Instagram, Flickr), Wikipedia, among others (Hobel *et al.* 2015, 2016, Chen *et al.* 2019, Twaroch *et al.* 2019, Belcastro *et al.* 2021, Gao *et al.* 2021). In social media, textual content can act as location-based proxies for urban life in regard to activities (eg shopping, working, eating out, recreation) and functions (eg commercial, transportation, residential) that cities can support in different places and regions (Gao *et al.* 2017). However, we need to be aware of the limitations in relying on social media posts with coordinates, as its attachment to space might not necessarily indicate correlation with the neighboring settings (Fu *et al.* 2018).

The vast number of works that explore urban dynamics from geo-text data is evidence that correlation between content and location is generally assumed to be high. Using tweets and POI classes, Hahmann *et al.* (2014) demonstrated that content-location correlation is often low and varies according to place types, arguing that studies should acknowledge this relationship in their applications while also discussing the need to critically consider the link between a piece of information to a pair of coordinates. With that in mind, McKenzie and Adams (2017) used place labels from Foursquare in a supervised topic modeling of geo-text data from social media platforms, showing that content related to built-up places seem to have a lower correlation while content characterized by physiographic features exhibit a higher alignment between data sources. Their theoretical underpinnings stem from the discussion between space and place, which is in fact fundamental in content-location relationships. Other similar examples in the literature seem to focus on the space-based aspects, such as extracting user positions based on tweet meta-data and matching to correspondent locations found in GeoNames and OSM data (Zohar 2021). More place-oriented approaches for discussing the relationships between content and location are timid: while Lansley and Longley (2016) revealed the influence of land-use and urban activities on the content of tweets, Yu *et al.* (2022) standpoint was to consider POI reviews as adequate spatial proxies of place-based information. Therefore, content-location relationships must be seen through an extended perspective, where comparisons are based not only on positions but also on meanings, functions, activities and affordances of the urban landscape.

2.2. Natural language processing and geo-tagged user-generated content

NLP consists in several techniques that aims at structuring, extracting information and making sense of human natural language (Lamurias and Couto 2019). As geo-tagged UGC carries information on people's in-space activities, opinions and experiences, it provides discursive information that can be used to explore different thematic attributes related to the urban landscape (Dunkel 2015, Martí *et al.* 2019). According to Twaroch *et al.* (2019), UGC does reflect people's experiences, focus, opinions and interests to a significant degree, and therefore NLP is a crucial tool to find relevant patterns in unstructured text data. The most prevalent NLP method found in the literature is topic modeling, which is able to reduce the complexity of massive geo-text datasets to extract thematic signatures linked to places, activities and perceptions (Fu *et al.* 2018).

From the wide range of available topic models, the LDA algorithm became pervasive in the literature (Liu *et al.* 2019). LDA is an unsupervised probabilistic model based on word co-occurrences (Blei *et al.* 2003, Jenkins *et al.* 2016). Some of the countless examples include extracting cognitive regions of Northern and Southern California (Gao *et al.* 2017); identifying urban functional regions in cities with check-in information (Gao *et al.* 2017); estimating geographic regions from unstructured text (Adams and Janowicz 2021); as well as the previously mentioned works of Lansley and Longley (2016), McKenzie and Adams (2017) and Yu *et al.* (2022). Nonetheless, research on topic modeling methods has empirically demonstrated the disadvantages of LDA, including careful tuning of hyper-parameters for generating cohesive topics, the requirement of detailed assumptions, overlapping topics, user-defined number of topics and restrictions in assessing the correlation between topics as word correlations are ignored (Egger and Yu 2022).

Although LDA has been one of the best-known and widely used models, other methods for text representation have been developed in the last years. In particular, algorithms that use word or sentence embeddings have been applied in more recent topic models such as the Top2Vec (Egger 2022). Word embeddings are vector representations of text data that enable semantic properties to be encoded whereby similar pieces of text information are nearer in vector space (Naseem *et al.* 2021). Therefore, by embedding words in a continuous vector space, words with similar semantic and syntactic meaning can be mapped to nearby points (Comber and Arribas-Bel 2019). Embeddings have been used within GIScience for tasks such as address geocoding, fine-scale land-use identification from POI data and even for building algorithms aimed at reasoning the complex spatial semantics of place types (Place2Vec), among others (Yao *et al.* 2017, Yan *et al.* 2017, Zhang *et al.* 2022). However, works in the field that employ topic models supported by word embeddings are still not commonplace, especially for exploring location-content relationships.

As embedding-based models are able to generate contextual representations, relationships that emerge in the vector space might be related to context emerging from the geographic space. Therefore, even without inserting spatial variables, the use of embedding-based topic modeling is more effective in unraveling latent geographic topics of interest and in the separation of geographic and non-geographic clusters (Yang *et al.* 2023). Among recent algorithms, Grootendorst (2022) has developed the BERTopic, a model that combines BERT embeddings (Bidirectional Enconder Representations From Transformers, developed by Devlin *et al.* (2019)) and other methods that enable higher flexibility for different use cases. The model works by first creating embeddings that use a pretrained language model, followed by reducing the dimensionality of documents and grouping semantically similar documents into clusters that represent distinct topics. Lastly, the model employs a class-based TF-IDF (term frequency-inverse document frequency) to compare the importance of terms and retrieve the most representative words per topic (Grootendorst 2022, Egger and Yu 2022).

BERTopic has been employed in social media text analysis such as investigating public sentiments regarding the monkeypox outbreak (Ng *et al.* 2022) and detecting cognitively distorted thinking patterns in Twitter (Alhaj *et al.* 2022). BERT embeddings have also been implemented in methods aimed at extracting geospatial information and toponyms in unstructured text (Chu *et al.* 2022, Berragan *et al.* 2023). In addition to outperforming other topic models, BERTopic is able to generate more interpretable topics, allows multilingual analysis and automatically finds the number of topics (Egger and Yu 2022, Egger 2022). In this paper, we have opted for implementing the model not only because the vector space might reflect the spatial context better than traditional approaches such as LDA, but also because the use of BERTopic in exploring location-content relationships in UGC has not been carried out in the literature.

3. Data and preprocessing

Using the Twitter Search API, we retrieved all georeferenced tweets posted roughly within the metropolitan area of the city of Lisbon, Portugal. Our search query collected tweets that lay within a 40 km radius around the centroid of Lisbon's municipality without time constraints. The following filtering and selection are exemplified in Figure 1.

First, tweets without explicit coordinate-based geo-tagging were removed to best represent users' active location sharing. Then, we selected those tweets whose assigned language field was Portuguese as the high number of tourists in the city might influence the data distribution. The next steps were to remove tweets with duplicated text entries to reduce contamination of spams, followed by clipping the data to Lisbon's municipality extent. Based on a 200 m-spaced hexagonal grid, we filtered user contribution in space by allowing up to 10 tweets per user per cell with the objective of reducing users that might skew data distribution in specific locations. The chosen spatial unit of analysis has an area of approximately 0.03 km^2, which is able to embed most city blocks but not enough to cover neighborhoods. We believe that this resolution is reasonable for our analysis based on the city's urban fabric and similar grid-based implementations (Andrade *et al.* 2020). As for limiting user contribution, our goal was to reduce the effects of potential dominating spatial bias from the

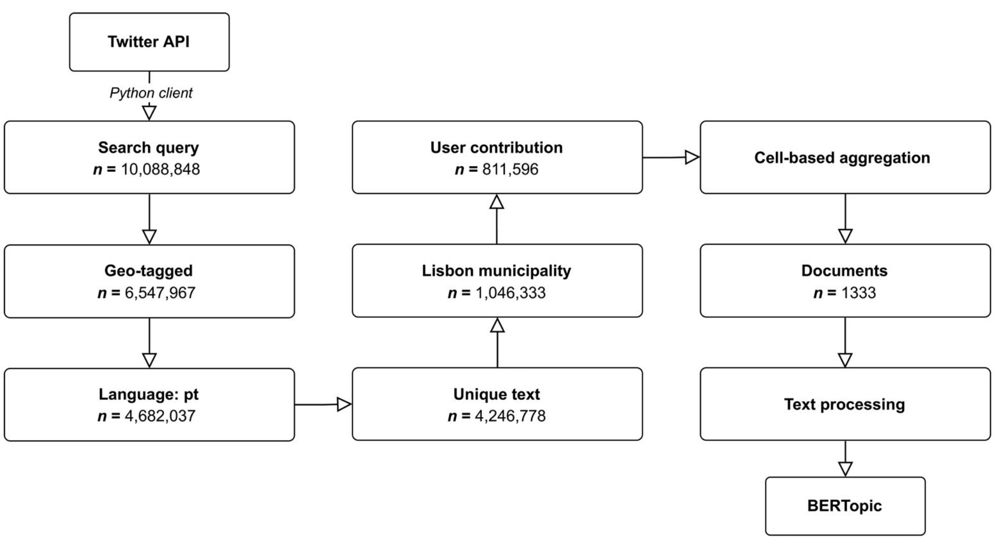

Figure 1. Twitter data filtering and preprocessing steps prior to topic modeling.

most active users (Gao *et al.* 2017). We also removed the cells containing less than the median value of tweets per cell across the study area. In combination, the previous tasks were aimed at both spatially leveraging user contribution and ensuring that areas with reduced user activity would not contribute to the topic modeling. Since there are no standards on these procedures, our choice of thresholds per cell was done arbitrarily both for the cell removal and for limiting user contribution.

The publishing dates of filtered tweets ranged from 2010 to 2021 and thus we assume that more than a decade of space-based online activity might have substantially contributed to shaping thematic information regarding different aspects of the city. After obtaining the final tweet distribution, we spatially aggregated their textual within each cell of the hexagonal grid covering the city. Therefore, each hexagonal cell represented a document in our topic modeling analysis. Throughout the paper, we will use the word 'document', 'hexagon' and 'cell' interchangeably depending on the context, although they are the same in our analysis. Lastly, we processed the text for the model by removing unwanted text such as special characters, emojis, urls and stop words.

Representing the thematic place-based counterpart, we sourced data from Google Places API and OSM (Figure 2). We gathered all POI from Google Places within the city, as well as POI and building centroids across Lisbon from OSM. Features extracted from Google Places consisted of user reviews and place names, whereas we retrieved non-empty place names from OSM. We opted not to use place type tags from both sources as we intended to mainly focus on textual information generated by users (place reviews) and place names that act as specific information linked to places. Place type tags not only might not represent specific locations in the city but also are not necessarily defined by users. Following feature extraction, we aggregated the text-based data based on the previous hexagonal grid, succeeded by the same text processing prior to topic modeling. Most text data originated from users' reviews on Google Places, where publishing dates ranged from 2011 to 2022.

By having similar temporal distributions, both datasets from Twitter and Google may thematically reflect consolidated place-based urban dimensions, even though POI might appear or cease to exist. Figure 3 shows the data distribution of instances from the space-based and place-based data sources prior to cell-based aggregation.

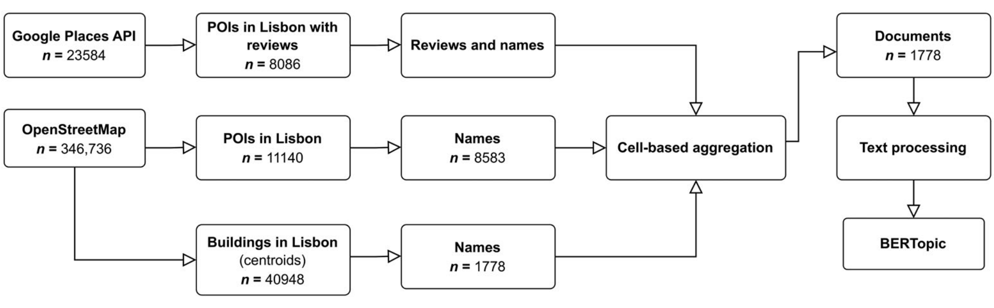

Figure 2. Google Places and OSM data filtering and preprocessing steps prior to topic modeling.

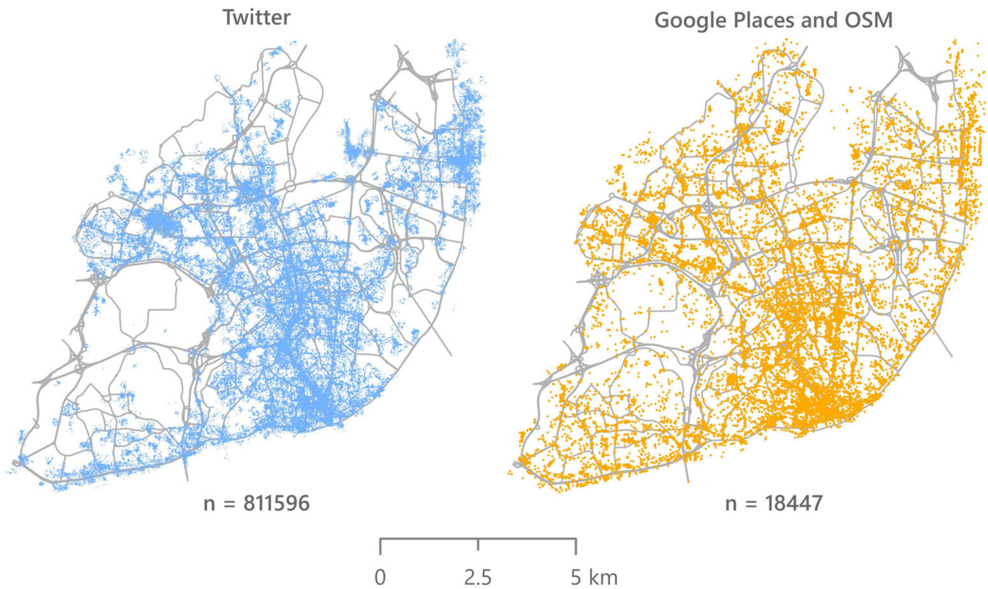

Figure 3. Point locations of data instances from Twitter, Google Places and OSM prior to hexagonal cell aggregation.

4. Methods

The framework we introduce is visually described in Figure 4. Our spatial unit of analysis are the cells that compose the hexagon-based grid across the city. The main components of the framework include: setting the aggregated textual data from Twitter and place-based sources (Google and OSM); employing the BERTopic transformer-based topic modeling for each source; comparing topics emerged from each source using the cosine similarity metric; carrying out Getis-Ord G_i^* hotspot analysis for retrieving statistically representative topic-based cells; applying the Jaccard similarity index aimed at ultimately comparing thematic and spatial similarities that support the discussion on content-location relationships for the case study.

4.1. Topic modeling

In order to extract thematic signatures from our space-based and place-based sources of textual information, we applied the BERTopic algorithm developed by Grootendorst (2022). Each cell of our hexagonal grid covering the city of Lisbon contained aggregated text-based data, acting as our documents for topic extraction. The BERTopic algorithm uses pre-trained transformer models, rooted in neural network architectures and able to encode words in vector-based representations (Saidi et al. 2022). In addition, it merges machine learning approaches to both reduce dimensionality through UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) and cluster similar embeddings for topic identification through HDBSCAN (Hierarchical Clustering and Density-Based Spatial Clustering of Applications with Noise).

We employed the BERTopic model for each data source independently, although we have set the same hyper-parameters to reduce any model-driven variations in the

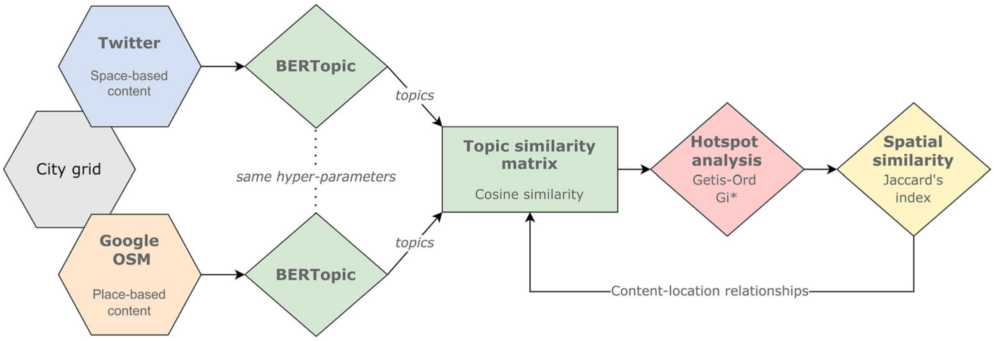

Figure 4. Methodological framework developed in the study.

originated topics. Whereas most parameters were kept as default given the lack of similar frameworks that use BERTopic, we did modify others for our implementation. For the HDBSCAN hyper-parameters, we set the minimum cluster size to 10 documents while keeping the minimum number of samples as 5 to potentially minimize the number of outliers (Grootendorst 2022). Since topics are generated through density-based clustering, documents are not forced to fit clusters and hence those that fail to belong to a topic are considered outliers, which helps reducing noise and generating more meaningful topics (Egger and Yu 2022). In addition, this also means that although hyper-parameters can be tuned to reduce outliers or change the minimum number of documents for topic generation, there is no prior selection regarding the number of topics. The embedding-based model reduces the dimensions and clusters documents into an optimal number of topics given the input parameters and data.

As for the UMAP, we set the number of neighboring sample points to 5 to constrain local neighborhood size and focus on local as opposed to global patterns. Increasing the number of neighbors provides a more global view of the embedding structure whereas lower values output a more local perspective (Grootendorst 2022). As UMAP is stochastic in nature, we also set a random state to guarantee the reproducibility of the model. For each topic, we retrieved the top 15 words that contributed the most in representing the information for the topic cluster. Lastly, we chose a multilingual embedding model, as not only Google Places reviews might be in different languages, but also in tweets, as languages assigned by Twitter are not always accurate. For each data source, the final output is the topic probability distributions across the grid cells, which are the input of the hotspot analysis, whereas word probability distributions for each topic are compared using the cosine similarity between topics.

4.2. Cosine similarity

To objectively compare the topics identified in the model between data sources, we used the cosine similarity metric. The similarity metric represents the angle between vectors. As the output topic information from BERTopic consists of the 15 most important words that form the topic cluster and their respective values of importance or probability, we treated topics as 15-dimensional vectors. The smaller the angle between vectors, the more similar the topics are in the vector space (Liu *et al.* 2019). The cosine similarity is defined as follows:

$$\text{similarity} = \cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}|| \, ||\mathbf{B}||} \tag{1}$$

Where A and B are vectors, and similarity is given by calculating the product between vectors divided by the cross product of their lengths (Fu *et al.* 2018). With values ranging from 0 to 1, we computed the cosine similarity for all topics retrieved from Twitter against those from the place-based sources. We then filtered the output pairwise matrix to select the highest values for each Twitter-based topic, showing the most similar corresponding place-based topics within the vector space. Following the selection, we assess the spatial relationship between corresponding topics to characterize and visualize the content-location relationships.

4.3. Getis-Ord G_i^*

The last two steps consist in assessing the spatial relationship between the space-based and place-based topics across the city, ultimately aimed at providing insights regarding the content-location relationships. In the first stage, we carried out a hot-spot analysis to retrieve statistically significant cells in regard to topic distributions, represented by the probability values assigned to documents or hexagons of belonging to each topic retrieved by the algorithm. For this task, we chose to calculate the Getis-Ord Gi* statistic, part of the G family of statistics developed by Getis and Ord (2010) aimed at characterizing pronounced local clusters of high and low values. In a study area with n points and $X = [x_1, \ldots, x_n]$ measurements, and assuming weights $w_{i,j}$ to be defined between all pairs of points i and j (for all $i, j \in \{1, \ldots, n\}$), the Getis-Ord G_i^* is denoted as:

$$rG_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2}{n-1}}} \tag{2}$$

Where $\bar{X}$ is the mean of all measurements and S is the standard deviation of all measurements (Kumar and Parida 2021). In our implementation, we ran the hotspot analysis for all topics extracted in the previous stage and identified cells with z-scores higher than 1.65, which are samples with standard deviations that have 90% or higher confidence or significance in regard to not responding to a random spatial distribution (Rossi and Becker 2019).

4.4. Jaccard index

In the second stage, we computed the Jaccard index metric for the identified hotspot areas corresponding to the pairwise comparison of similar topics derived from space-based and place-based sources. In other words, after selecting significant spatial distributions of the topics that yielded higher values of cosine similarity between sources, we computed the spatial similarity between these distributions. The metric is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

Where the result corresponds to the intersection divided by the unions of two sets A and B. Ranging from 0 to 1, the metric measures the similarity of two sets, represented in our case by the hotspot areas computed previously, similar to the approach of Heikinheimo *et al.* (2020). More precisely, sets A and B are the areal extent composed by cells identified as hotspots with z-scores higher than 1.65 from each source respectively. We then discussed the spatial and thematic similarities between the topics derived from Twitter and the corresponding ones derived from POI names and reviews.

5. Results

5.1. Topics

As described in the previous section, we ran the BERTopic algorithm using the same hyper-parameters for the cell-based documents derived from our space-based (Twitter) and place-based (Google and OSM) data sources. While Twitter data yielded 31 topics with 376 outlier documents, Google and OSM data yielded 35 topics with 381 outliers. We present a selection of interesting topics, their words and probabilities as well as word translations from the space-based and place-based sources in Tables 1 and 2 respectively. We have listed all identified topics and their information in the appendix (Tables A1–A4), including the ones we do not mention or discuss throughout the paper. The topic order is based on the descending count of documents (hexagonal cells) that were assigned by the algorithm as belonging to the topics. Topic belonging corresponds to the dominant topic of each hexagon or the topic with the highest probability for the document, as each cell yielded probability values ranging from 0 to 1 for all topics. In addition, topic information also includes the corresponding number of instances that were originally aggregated in the documents: tweets, OSM features and Google Places POI. In total, the 31 Twitter topics were modeled based on the aggregated text of 610,593 tweets and the 35 topics of place-based sources originated from 11,392 Google and OSM instances.

Alfama, a historic neighborhood in Lisbon known for Fado – a famous style of Portuguese folk music (Cocola-Gant and Gago 2021) – was the theme identified in Topic 4. The words thematically characterize the neighborhood as common in-situ activities include concerts (fado music) and dining out. Words that build the football topic (Topic 2) include mostly references to the two largest stadiums in Lisbon and their respective football teams, Benfica and Sporting (Borges 2019). The location-specific airport topic (Topic 22) mainly consists of references to Lisbon's airport, whereas the university topic (Topic 28) contains words that are both generally related to higher education as well as specific faculties of the University of Lisbon. Overall, interpretable topics emerged from the social media network yielded thematic profiles

Table 1. Selected sample of interesting topics from Twitter.

Football (Topic 2) Hexagons: 49/tweets: 41079			Alfama (Topic 4) Hexagons: 43/tweets: 26754		
Word	Translation	Prob.	Word	Translation	Prob.
Estádio	Stadium	0.1462	Duetos	Duet	0.0698
Benfica	Benfica football team	0.1183	Bar	Bar	0.0472
Sport	Benfica football team	0.0995	Amp	Organization in Alfama	0.0434
Alvalade	José Alvalade stadium	0.0604	Alfama	Alfama neighborhood	0.0420
José	José Alvalade stadium	0.0568	Gastronomia	Gastronomy	0.0400
Slbenfica	Benfica football team	0.0470	Restaurant	Restaurant	0.0375
xxi	–	0.0414	Café	Café	0.0341
Sporting	Sporting football team	0.0374	Praça	Plaza/square	0.0299
Luz	Luz stadium	0.0301	Mercado	Market	0.0297
Carregabenfica	Benfica football team	0.0267	Ribeira	Area in Lisbon	0.0277
Campo	field	0.0266	Fado	Fado music	0.0269
Bairro	Neighborhood	0.0235	Sobremesa	Dessert	0.0268
Alto	Tall/high	0.0232	Comércio	Business	0.0260
Slb	Benfica football team	0.0214	Música	Music	0.0252
sportingcp	Sporting football team	0.0204	Concerto	Concert	0.0248
Airport (Topic 22) Hexagons: 16/tweets: 14991			**University (Topic 28) Hexagons: 11/tweets: 3580**		
Word	Translation	Prob.	Word	Translation	Prob.
Aeroporto	Airport	0.3283	Faculdade	Faculty/university	0.1310
Lis	Lisbon airport	0.1976	Ciências	Sciences	0.1254
Others	–	0.0559	Universidade	University	0.1218
Delgado	Lisbon airport	0.0467	Colombo	Colombo mall	0.0557
Humberto	Lisbon airport	0.0466	Cinemas	–	0.0552
Chegadas	arrivals	0.0446	Campus	–	0.0530
Airport	–	0.0432	Justiça	Justice	0.0505
Arrivals	–	0.0431	FCUL	University of Lisbon	0.0381
Terminal	–	0.0423	Medicina	Medicine	0.0355
Departures	–	0.0367	Dentaria	Dental	0.0343
Partidas	Departures	0.0367	Holmes	Local gym chain (Holme's place)	0.0265
Comunidades	Communities	0.0332	Campo	Field	0.0264
Lisboalis	Lisbon airport	0.0321	Place	Local gym chain (Holme's place)	0.0208
Internacional	International	0.0318	Filme	Film	0.0168
Portuguesas	Portuguese	0.0312	IMAX	IMAX cinema	0.0160

mostly related to neighborhoods, locations and areas of interest, rather than general place-mediated activities.

Interesting topics from the place-based perspective included health (Topic 4), education (Topic 7), shopping mall (Topic 18) and sports (Topic 30). Topics originated from documents based on Google Places and OSM also contained words related to specific places and areas within the city, yet overall to a lesser extent in comparison with topics from Twitter.

5.2. Topic similarity

We computed the cosine similarity for all Twitter topics against those originated from Google and OSM data. For each topic, we selected the highest value of similarity using the pairwise matrix to obtain the most similar corresponding place-based topic. In Table 3, we listed each Twitter topic and their matching topics alongside their cosine similarity values.

Table 2. Selected sample of interesting topics from Google Places and OSM.

Health (Topic 4) Hexagons: 41/OSM: 198/Google POI: 232			Education (Topic 7) Hexagons: 35/OSM: 140/Google POI: 96		
Word	Translation	Prob.	Word	Translation	Prob.
Good	–	0.0145	Escola	School	0.0800
Atendimento	Service/treatment	0.0144	School	–	0.0652
Excelente	Excellent	0.0140	University	–	0.0312
Farmácia	Pharmacy	0.0133	Faculdade	Faculty/university	0.0218
Service	–	0.0129	Universidade	University	0.0216
Saúde	Health	0.0127	Azulejos	Portuguese tiles	0.0197
Clínica	Clinic	0.0124	Registo	Registration	0.0181
Centro	Center	0.0114	Teachers	–	0.0169
Hospital	–	0.0112	José	José Fontana square	0.0166
Simpatia	sympathy	0.0103	Ensino	Teaching/education	0.0164
Café	–	0.0103	professores	Professors	0.0154
Great	–	0.0103	Fontana	José Fontana square	0.0152
lda	company	0.0102	Faculty	–	0.0142
Appointment	–	0.0092	Superior	Higher (education)	0.0137
Rua	Street	0.0091	campus	–	0.0131
Shopping mall (Topic 18) Hexagons: 20/OSM: 107/Google POI: 156			Sports (Topic 30) Hexagons: 12/OSM: 23/Google POI: 26		
Word	Translation	Prob.	Word	Translation	Prob.
Atendimento	Customer service	0.0215	Futebol	Football	0.0554
Colombo	Colombo mall	0.0161	Campo	Field	0.0516
Serviço	–	0.0147	Musgueira	Musgueira sports complex	0.0451
Good	–	0.0144	Bandeiras	Flags	0.0451
lda	Company	0.0139	Desportivo	Sports	0.0449
loja	Store	0.0131	Ténis	Tennis	0.0407
Empresa	Company	0.0130	Park	–	0.0387
Centro	Center	0.0127	Universitário	University	0.0368
Excelente	Excellent	0.0119	Tennis	–	0.0346
Really	–	0.0111	Clube	Club	0.0327
Service	–	0.0099	Ferreira	Portuguese surname	0.0311
Profissionalismo	Professionalism	0.0096	Condições	Conditions	0.0311
Preço	Price	0.0095	Desportiva	Sports	0.0306
Telheiras	Telheiras neighborhood	0.0094	Court	–	0.0302
Equipa	Team/staff	0.0093	Amaral	Portuguese surname	0.0283

In Table 3, we highlighted the highest values of similarity that represent the most similar comparisons in the vector space. The most similar topics (0.9527) were represented by the football topic from Twitter (Topic 2) and Topic 19 from Google and OSM, whose thematic signatures are related to football as well as Lisbon-based football teams and stadiums. The second highest similarity (0.9451) did not yield easy interpretations regarding the topics' semantic relationships. While Topic 3 from Twitter is mostly related to landmarks and POI located in downtown Lisbon, the corresponding Topic 29 thematic signatures are characterized by shopping-related activities. However, the topic from Twitter has 'comércio' (business or commerce) as its first representative word, although likely related to a main landmark in the city named 'Praça do Comércio' (Comércio plaza).

The third highest similarity (0.9396) was measured for the comparison between Topic 19 from Twitter, which refers to landmarks in two different neighborhoods, and Topic 6, from which words did not point towards a discernible thematic profile. Apart from the outlier, we noticed that three particular place-based topics were associated

Table 3. Most similar topic pairs based on the highest cosine similarity values yielded when comparing Twitter topics against those from Google Places and OSM.

Topics Twitter	Google and OSM	Cosine similarity	Topics Twitter	Google and OSM	Cosine similarity
−1 (outlier)	5	0.9367	15	5	0.9136
0	5	0.9367	16	20	0.8646
1	29	0.9162	17	6	0.927
2	19	**0.9527**	18	21	0.9163
3	29	**0.9451**	19	6	**0.9396**
4	14	0.9338	20	29	0.9286
5	29	0.8268	21	5	0.9038
6	6	0.9018	22	15	0.9275
7	6	0.8898	23	5	0.9218
8	5	0.9225	24	5	0.9221
9	21	0.8486	25	6	0.8911
10	26	0.915	26	14	0.8936
11	5	0.9105	27	29	0.9338
12	28	0.9242	28	7	0.906
13	29	0.9153	29	6	0.9213
14	20	0.8813	30	10	0.9039

with Twitter topics in six different comparisons. Topic 6, with no specific thematic profile; Topic 29 (shopping activities); and Topic 5, which is vaguely related to general services in the city.

5.3. Spatial similarity

Based on the previous identified topics, we ran the Getis-Ord G_i^* hotspot analysis to seek the local high values of the topic distribution. For each output, we selected the cells with z-scores denoting 90% confidence or higher. Cell-based hotspot areas are better at depicting the relevant regions in regard to the original distributions of topics, which oftentimes are spread across the city. Then, for each topic pair we computed the Jaccard index based on the distribution of the selected cells as shown in Table 4. The three highest outputs are highlighted.

With their thematic profiles linked to Lisbon's airport, the Jaccard index between the Topic 22 (Twitter) and Topic 15 (Google Places and OSM) scored the highest value (0.18). Terms in the topics include 'departures', 'arrivals' and 'taxi' as well as references to the name of the airport. Another instance of similar themes in the geographic space is represented by the football topic pair, which yielded the third-highest Jaccard index (0.15).

The second highest measurement was yielded by the Topic 5 and 29 pair (0.16). By itself, the topic from Twitter does not point towards a specific thematic profile, however, the place-based topic is strongly related to shopping activities in the city. Therefore, the high spatial relationship suggest that the uncertain thematic profile might also be linked to shopping, even though the topic is polluted with noise.

5.4. Content-location relationships

The core of this study lies at providing a framework to extract thematic and spatial relationships between content generated from space-based and place-based sources,

Table 4. Jaccard indices between selected hotspot areas of similar topic pairs from space-based and place-based sources.

Topics Twitter	Google and OSM	Jaccard index	Topica Twitter	Google and OSM	Jaccard index
−1 (outlier)	5	–	15	5	0.05
0	5	~0	16	20	0.07
1	29	0.01	17	6	0.05
2	19	**0.15**	18	21	~0
3	29	0	19	6	~0
4	14	0.03	20	29	0.03
5	29	**0.16**	21	5	0.04
6	6	~0	22	15	**0.18**
7	6	~0	23	5	0.06
8	5	0.11	24	5	0.02
9	21	0.01	25	6	0.02
10	26	0.04	26	14	0.12
11	5	0.03	27	29	0
12	28	0.04	28	7	~0
13	29	0.06	29	6	0.03
14	20	0.07	30	10	0.05

ultimately enriching the discussion on content-location relationships within a given city. Since it is not feasible to discuss about all relationships in regard to comparisons between the topics' vector and geographic space, we brought forward visualizations of topic pairs selected on the basis of their spatial and thematic similarities. In Figures 5 and 6, we display the high-value hotspot distributions from the two most similar topic pairs according to cosine similarity and Jaccard index, respectively.

The football theme is represented by the topic pair with the highest similarity in the vector space as well as significant spatial overlap. Visual inspection allow us to observe their similar hotspot distribution. The output suggests the content-location correlation for this thematic profile is high. This is not the case for the second most similar topic pair, which had no spatial overlap whatsoever. The topic from the social network mainly revealed landmarks of Lisbon's downtown, while it also contained words linked to the 'FIL' exhibition center. Although identified as part of the same topic, these two thematic signatures are related to distinct regions. The lack of spatial correlation indicate that despite having high cosine similarity, their thematic profiles are distinct, as its corresponding place-based topic consisted of shopping-related words.

As for the most spatially similar topic pair, the airport thematic profile evidences a high content-location correlation between Twitter and the place-based counterparts. This might suggest that when users geo-tag content related to airports, they are most likely engaging in activities afforded by the airport location. However, content-location relationships become blurry when comparing the distribution of the second-highest spatially similar topic pair. Contaminated with noise, the topic from Twitter does not indicate a clear thematic profile, yet the comparison with the corresponding place-based shopping theme shows a significant spatial correlation. We selected place-based topics based on the highest cosine similarity against each Twitter topic, yet this pair had yielded the lowest value from all topic pairs. In Figure 7 we present two final examples of topic pairs to complement our discussion.

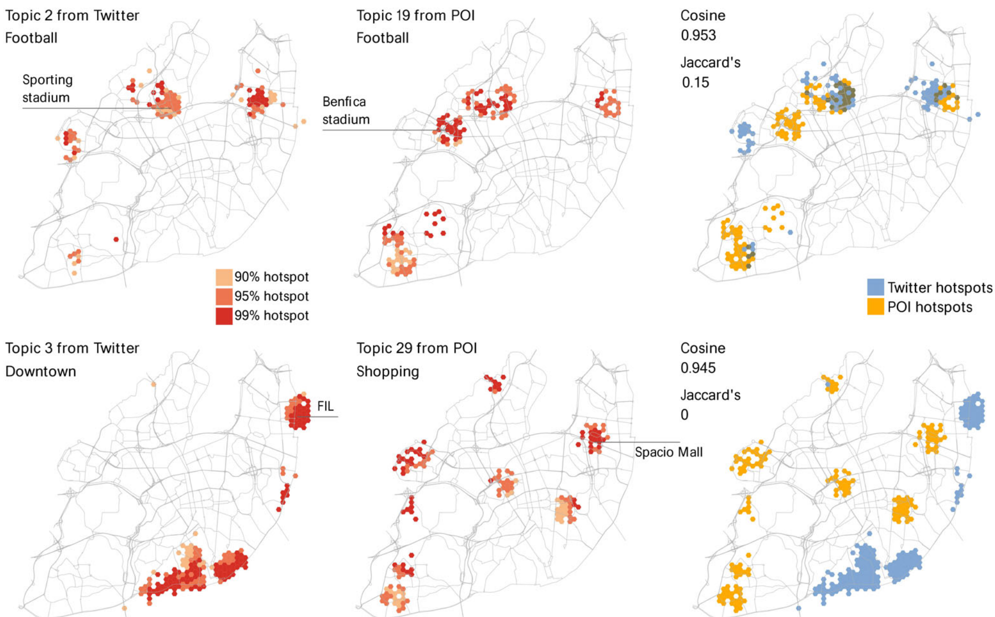

Figure 5. Hotspot distribution and Jaccard index of the two most similar topic pairs (top and bottom) based on the cosine similarity.

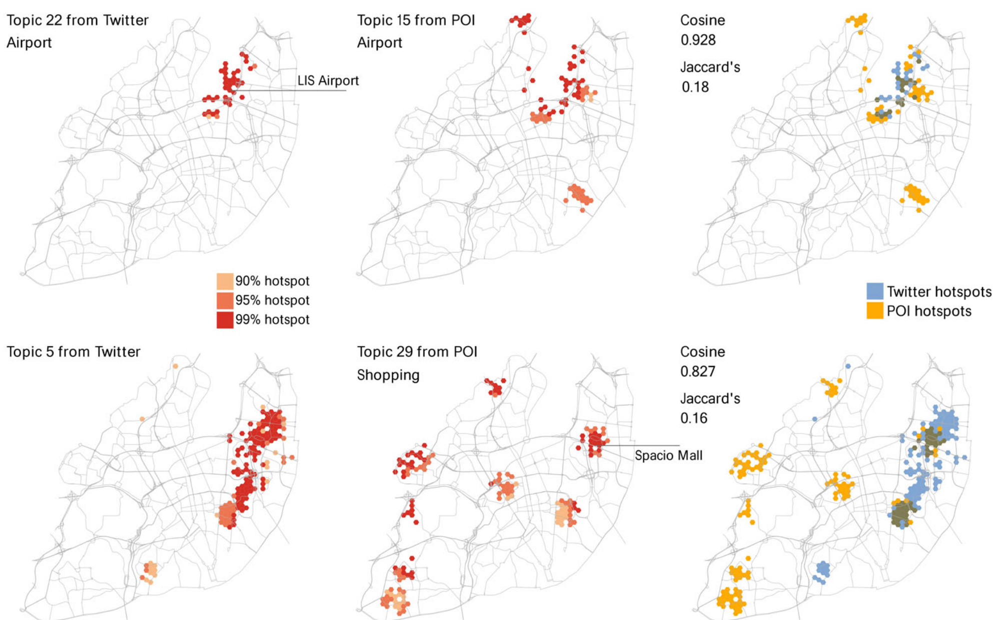

Figure 6. Hotspot distribution and Jaccard index of the two most spatially similar topic pairs (top and bottom) based on the Jaccard index.

Similar thematic profiles regarding education showed negligible spatial overlap across the city. The topic from Twitter was mainly linked to instances related to higher education and universities, whereas the place-based topic also contained words

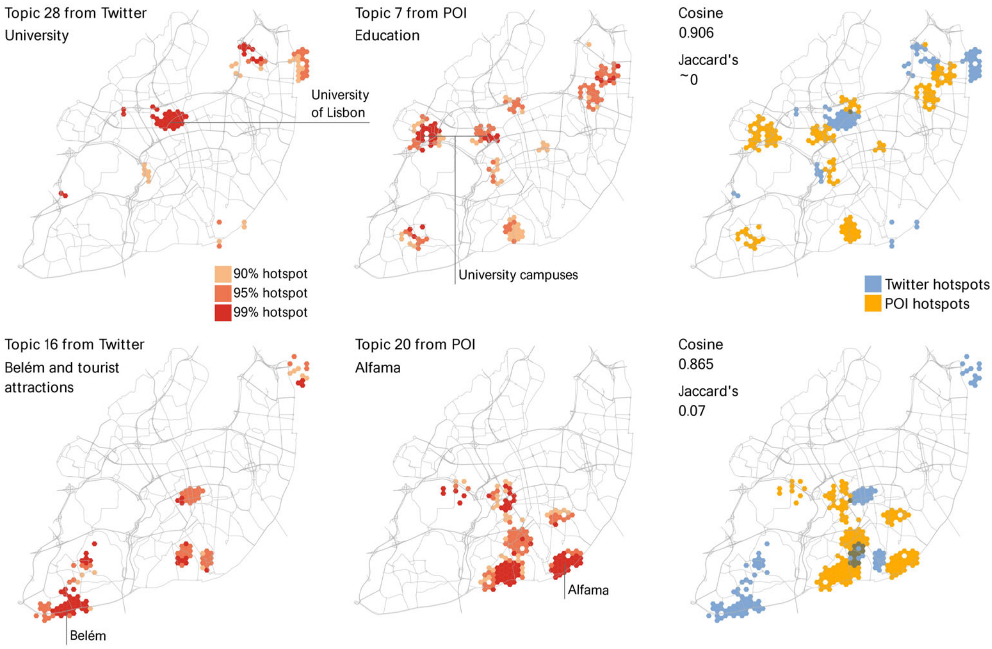

Figure 7. Hotspot distribution and Jaccard index of two topic pairs (top and bottom).

related to education in general. While we identified university campuses in both distributions, they pointed towards different areas in the city resulting in significantly low overlap. In this case, one can argue that content-location correlation is low as the spatial similarity is close to zero. However, the geo-tagged content from Twitter collectively refers to the location of the University of Lisbon. The intricate relationship between thematic and spatial similarities between geo-tagged activity and the corresponding place-based content is also exemplified in the last topic pair. Despite being linked to different neighborhood and landmarks, evidenced by a weak cosine similarity, their spatial overlap is mostly located in a historic and touristic region. Both Belém and Alfama are historic districts in Lisbon, enclosing important landmarks and attractions.

6. Discussion

The information we harvested from space-based and place-based sources of unstructured text were collectively analyzed in the form of topics. Both sources yielded thematic profiles that described locations, activities and functions of the urban landscape. The steps of our framework are able to quantitatively compare topics derived from geo-tagged social media activity with the most similar topics emerged from place-based sources (Google Places and OSM). However, elucidating content-location relationships based on thematic and spatial correlations depend on careful interpretations of the results.

Although we applied the embedding-based BERTopic without others models for comparison, the topic clusters showed that the algorithm was able to output many coherent and interpretable topics, including geo-indicative topics of interest that are related to specific activities, functions and affordances of different regions within the city. The algorithm is freely available to the public and does not require substantial text preprocessing. In addition, the algorithm yields an optimal number of topics according to cluster parameters and hence does not force instances to belong to topics, which is a better option for oftentimes noisy or incomplete data. Therefore, studies that source data from geo-tagged online activity should not only take advantage of the advances in embedding-based models, but also compare with other traditional and novel topic models

On the other hand, topics with unclear thematic profiles (such as Topic 5 and 6 from place-based sources) frequently scored high values of cosine similarity with topics retrieved from Twitter. Textual data sourced from user-generated content is noisy, unstructured and messy by nature. When adding the spatial dimension, a new layer of complexity is included and researchers must be aware of the limitations of the data themselves prior to the analysis. By developing a straightforward reproducible framework using an embedding-based topic model, researchers can test thematic content-location relationships by changing model parameters, confidence levels, thresholds, preprocessing steps as well as the resolution of the spatial unit of analysis.

We observed that the degree to which geo-tagged content from social media is connected to the corresponding place-based characteristics of the city will vary depending on thematic profiles. Similar insights were found in related literature, but differences were portrayed by place types (Hahmann *et al.* 2014, McKenzie and Adams 2017). Here, we represent both space-based and place-based geo-text dimensions as collective topics to be objectively evaluated against each other. Although previous works have developed methods to geo-locating social media activity, we developed an approach to extend the discussion on how discursive information in intentionally geo-tagged text might be associated with urban settings and activities (Adams and Janowicz 2021).

Football, a topic that potentially has a high disconnect between content and location, was characterized by one of the highest interpretable correlation between sources. The relation suggests that in Lisbon, geo-tagged content linked to football is connected to locations that afford football related activities. We observed the same relationship in the airport topic, indicating that geo-tagged content thematically associated with airports is mostly generated near the airport location. However, uncertain thematic profiles and different types of categories (activities, neighborhoods and places) show that choosing topic modeling to explore content also reveals that correlations between content and spatial context is intricate and open to discussion.

We were able to identify similar topic pairs coming from space-based and place-based sources using the cosine similarity metric. The following spatial similarity analysis disclosed distinct relationships from which interpretations are not necessarily straightforward. Our results can be translated through two somewhat contrasting viewpoints. One hand, dissimilarities reinforce the limitation of using geo-tagged UGC, as it only connects spatial footprints with textual data (Papadakis *et al.* 2020). On the other hand, similarities between sources strengthen the justifications of using UGC to infer the interaction between people and places within the urban environment (Lansley and Longley 2016, Heikinheimo *et al.* 2020).

Furthermore, our study supports the inquisitive discussions on the reliability and accuracy of geospatial information collected or inferred from online sources, problems that are not only a product of well-known biases (Twaroch *et al.* 2019), but also of the theoretical and methodological approaches behind these practices. Although our analysis was bounded to the same limitations and biases, we hope to incite other researchers to extend analytical and conceptual frameworks aimed at validating the use of geo-tagged UGC to unravel human-centered urban dimensions.

Some limitations should be pointed out. First, both sources of user-generated content are biased regarding their users' demographic profiles and do not fully cover the whole extent of the city, which in turn affects representativeness (Zhang *et al.* 2018, Gao *et al.* 2021). In addition, aggregating geo-tagged textual data into cells can result in biases that stem from the MAUP problem (Openshaw effect), whereby thematic and spatial relationships might differ according to cell size or scale (Goodchild 2022). Lastly, results also show that interpretation of thematic and spatial relationships are often constrained to prior familiarity with and knowledge about the city in regard to specific places, activities and neighborhoods. To improve interpretability as well as insights about content-location relationships, future work should consider applying spatially explicit topic models, gathering additional data from online sources as well as implementing alternative metrics and spatial and temporal units of analysis.

7. Conclusions

Geo-tagged social network data has become an extremely popular data source in urban studies as information is used to map, explore and infer the several dimensions of human-environment interactions, including human mobility, urban perception, sentiment analysis among many other activities and opinions. However, the content-location relationships in social media activity are intricate and not always clear. In this article, we introduced a methodological framework to explore the vector-space and geographic-space similarities between thematic profiles emerged from space-based (Twitter) and place-based (Google Places and OSM) sources of geographic user-generated content.

The stages included applying a transformer-based topic modeling, retrieving cosine similarity measurements between topics, running Getis-Ord G_i^* hotspot analysis to extract representative topic cells as well as computing Jaccard indices to calculate spatial similarities. The results showed that content-location relationship between the surrounding urban settings and the thematic content of in-situ online activity are heavily dependent on the thematic signatures. Nonetheless, the framework can easily be implemented and extended in other cities in order to explore novel insights and support discussions on the use of geo-tagged UGC in GIScience.

Author contributions

Vicente Tang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review and Editing, Visualization.

Marco Painho: Conceptualization, Resources, Writing – Original Draft, Writing – Review and Editing, Supervision, Project administration, Funding acquisition

Disclosure statement

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

Funding

The authors acknowledge the funding from the Portuguese national funding agency for science, research and technology (Fundação para a Ciência e a Tecnologia – FCT) through the CityMe project (EXPL/GES-URB/1429/2021; https://cityme.novaims.unl.pt/) and the project UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Notes on contributors

Vicente Tang is a PhD candidate in Geoinformatics at the NOVA Information Management School of the NOVA University Lisbon and also a research member of the CityMe project (https://cityme.novaims.unl.pt/). His research interests include the use of user-generated content and participatory approaches within GIScience to explore places and regions in the city. Twitter: @CityMe5 @vicetang_

Marco Painho is a full professor of Geographic Information Science at the NOVA Information Management School of the NoOVA University Lisbon and the scientific coordinator of the Geoinformatics and Analytics Laboratory. He holds a Master in Regional Planning from the University of Massachusetts, Amherst and a PhD in Geography from the University of California Santa Barbara. Twitter: @painho

ORCID

Vicente Tang http://orcid.org/0000-0002-5591-9108
Marco Painho http://orcid.org/0000-0003-1136-3387

Data and codes availability statement

The data and code that support the findings of this study are available at: https://doi.org/10.6084/m9.figshare.19307936

References

Adams, B., and Janowicz, K., 2021. On the geo-indicativeness of non-georeferenced text. *Proceedings of the International AAAI Conference on Web and Social Media*, 6 (1), 375–378.

Alhaj, F., *et al.*, 2022. Improving arabic cognitive distortion classification in twitter using BERTopic. *International Journal of Advanced Computer Science and Applications*, 13 (1), 854–860.

Andrade, R., Alves, A., and Bento, C., 2020. POI mining for land use classification: a case study. *ISPRS International Journal of Geo-Information*, 9 (9), 493.

Belcastro, L., Marozzo, F., and Perrella, E., 2021. Automatic detection of user trajectories from social media posts. *Expert Systems with Applications*, 186, 115733.

Berragan, C., *et al.*, 2023. Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, 37 (4), 747–766.

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet Allocation. *The Jorunal of Machine Learning Research*, 3, 993–1022.

Bo, H., and Martin, E., 2013. Spatial topic modeling in online social media for location recommendation. *RecSys '13: Proceedings of the 7th ACM conference on Recommender systems*, Hong Kong, 25–32.
Borges, F., 2019. Soccer clubs as media organizations: a case study of Benfica TV and PSG TV. *International Journal of Sport Communication*, 12 (2), 275–294.
Chen, M., Arribas-Bel, D., and Singleton, A., 2019. Understanding the dynamics of urban areas of interest through volunteered geographic information. *Journal of Geographical Systems*, 21 (1), 89–109.
Chu, D., *et al.*, 2022. A machine learning approach to extracting spatial information from geological texts in Chinese. *International Journal of Geographical Information Science*, 36 (11), 2169–2193.
Cocola-Gant, A., and Gago, A., 2021. Airbnb, buy-to-let investment and tourism-driven displacement: a case study in Lisbon. *Environment and Planning A*, 53 (7), 1671–1688.
Comber, S., and Arribas-Bel, D., 2019. Machine learning innovations in address matching: a practical comparison of word2vec and CRFs. *Transactions in GIS*, 23 (2), 334–348.
de Oliveira, T.H.M., and Painho, M., 2021. Open geospatial data contribution towards sentiment analysis within the human dimension of smart cities. *In*: A. Mobasheri, ed. *Open Source Geospatial Science for Urban Studies*. Cham: Springer International Publishing, 75–95.
Devlin, J., *et al.*, 2019. BERT: pre-training of deep bidirectional transformers for language understanding. *In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies*, Vol. 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.
Dunkel, A., 2015. Visualizing the perceived environment using crowdsourced photo geodata. *Landscape and Urban Planning*, 142, 173–186.
Egger, R., 2022. Text representations and word embeddings. *In*: R. Egger, ed. *Applied data science in tourism: interdisciplinary approaches, methodologies, and applications*. Cham: Springer International Publishing, 335–361.
Egger, R., and Yu, J., 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology*, 7, 2297–7775.
Fu, C., *et al.*, 2018. Identifying spatiotemporal urban activities through linguistic signatures. *Computers, Environment and Urban Systems*, 72, 25–37.
Gao, S., *et al.*, 2017. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *International Journal of Geographical Information Science*, 31 (6), 1–27.
Gao, S., *et al.*, 2021. User-generated content: a promising data source for urban informatics. *In*: W. Shi, *et al.*, eds. *Urban Informatics*. Singapore: Springer, 503–522.
Gao, S., Janowicz, K., and Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in Gis*, 21 (3), 446–467.
Getis, A., and Ord, K., 2010. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24 (3), 189–206.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.
Goodchild, M.F., 2022. The openshaw effect. *International Journal of Geographical Information Science*, 36 (9), 1697–1698.
Grootendorst, M., 2022. BERTopic: neural topic modeling with a class-based TF-IDF procedure. *arXiv*.
Hahmann, S., Purves, R., and Burghardt, D., 2014. Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes. *Journal of Spatial Information Science*, 2014 (9), 1–36.
Heikinheimo, V., *et al.*, 2020. Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning*, 201, 103845.
Herfort, B., *et al.*, 2014. Exploring the geographical relations between social media and flood phenomena to improve situational awareness. *In*: J. Huerta, S. Schade, C. Granell, eds.

Connecting a digital alati through location and place. Heidelberg: Springer International Publishing, 55–71.
Hobel, H., *et al.*, 2015. A Semantic region growing algorithm: extraction of urban settings. *In*: F. Bação, F. U. Andrew, M. Y. Santos, M. Painho, eds. *AGILE 2015: geographic information science as an enabler of smarter cities and communities*. Cham: Springer International Publishing, 19–33.
Hobel, H., Fogliaroni, P., and Andrew, F.U., 2016. Deriving the geographic footprint of cognitive regions. *In*: S. Tapani, M. S. Yasmina, S. L. Tiina, eds. *Geospatial data in a changing world*. Cham: Springer International Publishing, 67–84.
Hu, Y., *et al.*, 2021. A framework to detect and understand thematic places of a city using geospatial data. *Cities*, 109, 103012.
Hu, Y., 2018a. 1.07 – geospatial semantics. *In*: B. Huang, ed. *Comprehensive Geographic Information Systems*. Oxford: Elsevier, 80–94.
Hu, Y., 2018b. Geo-text data and data-driven geospatial semantics. *Geography Compass*, 12 (11), e12404.
Huang, Q., and Xiao, Y., 2015. Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, 4 (3), 1549–1568.
Iranmanesh, A., Cömert, N.Z., and Hoşkara, ŞÖ., 2022. Reading urban land use through spatio-temporal and content analysis of geo-tagged Twitter data. *GeoJournal*, 87 (4), 2593–2610.
Jenkins, A., *et al.*, 2016. Crowdsourcing a collective sense of place. *PLOS One*, 11 (4), e0152932.
Kumar, S., and Parida, B.R., 2021. Hydroponic farming hotspot analysis using the Getis–Ord Gi* statistic and high-resolution satellite data of Majuli Island, India. *Remote Sensing Letters*, 12 (4), 408–418.
Lamurias, A., Couto, F.M., *et al.*, 2019. Text mining for bioinformatics using biomedical literature. *In*: S. Ranganathan, eds. Oxford: Academic Press, 602–611.
Lansley, G., and Longley, P.A., 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85–96.
Liu, Z., *et al.*, 2019. Recommending attractive thematic regions by semantic community detection with multi-sourced VGI data. *International Journal of Geographical Information Science*, 33 (8), 1520–1544.
Martí, P., Serrano-Estrada, L., and Nolasco-Cirugeda, A., 2019. Social media data: challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, 74, 161–174.
McKenzie, G., and Adams, B., 2017. Juxtaposing thematic regions derived from spatial and alatial user-generated content. *Proceedings of the 13th International Conference on Spatial Information Theory*.
Melo, F., and Martins, B., 2017. Automated geocoding of textual documents: a survey of current approaches. *Transactions in GIS*, 21 (1), 3–38.
Naseem, U., *et al.*, 2021. A comprehensive survey on word representation models: from classical to state-of-the-art word representation language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20 (5), 1–35.
Ng, Q.X., *et al.*, 2022. Public sentiment on the global outbreak of monkeypox: an unsupervised machine learning analysis of 352,182 twitter posts. *Public Health*, 213, 1–4.
Niu, H., and Silva, E.A., 2020. Crowdsourced data mining for urban activity: review of data sources, applications, and methods. *Journal of Urban Planning and Development*, 146, 2.
Jay, A. Number of twitter users 2022/2023: demographics, breakdowns & predictions [online]. FinancesOnline. 2022. Available from: https://financesonline.com/number-of-Twitter-users/ [Accessed 28 January 2022].
Papadakis, E., Resch, B., and Blaschke, T., 2020. Composition of place: towards a compositional view of functional space. *Cartography and Geographic Information Science*, 47 (1), 28–45.
Psyllidis, A., *et al.*, 2022. Points of interest (POI): a commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2 (1), 20.

Qiu, Q., *et al.*, 2022. ChineseTR: a weakly supervised toponym recognition architecture based on automatic training data generator and deep neural network. *Transactions in GIS*, 26 (3), 1256–1279.
Rossi, F., and Becker, G., 2019. Creating forest management units with Hot Spot Analysis (Getis-Ord Gi*) over a forest affected by mixed-severity fires. *Australian Forestry*, 82 (4), 166–175.
Saidi, F., Trabelsi, Z., and Thangaraj, E., 2022. A novel framework for semantic classification of cyber terrorist communities on Twitter. *Engineering Applications of Artificial Intelligence*, 115, 105271.
Shang, S., *et al.*, 2016. Finding regions of interest using location based social media. *Neurocomputing*, 173, 118–123.
Suwaileh, R., *et al.*, 2022. When a disaster happens, we are ready: location mention recognition from crisis tweets. *International Journal of Disaster Risk Reduction*, 78, 103107.
Twaroch, F.A., *et al.*, 2019. Investigating behavioural and computational approaches for defining imprecise regions. *Spatial Cognition & Computation*, 19 (2), 146–171.
Wang, B., *et al.*, 2020. Understanding the spatial dimension of natural language by measuring the spatial semantic similarity of words through a scalable geospatial context window. *PLOS One*, 15 (7), e0236347.
Yan, B., *et al.*, 2017. From ITDL to Place2Vec: reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 10.
Yang, J., Jang, J., and Yu, K., 2023. Analyzing geographic questions using embedding-based topic modeling. *ISPRS International Journal of Geo-Information*, 12 (2), 52.
Yao, Y., *et al.*, 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, 31 (4), 825–848.
Yu, Z., Xiao, Z., and Liu, X., 2022. A data-driven perspective for sensing urban functional images: place-based evidence in Hong Kong. *Habitat International*, 130, 102707.
Zhang, C., *et al.*, 2022. W-TextCNN: a TextCNN model with weighted word embeddings for Chinese address pattern classification. *Computers, Environment and Urban Systems*, 95, 101819.
Zhang, G., Zhu, A., and Zhu, 2018. The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS*, 24 (3), 151–162.
Zohar, M., 2021. Geolocating tweets via spatial inspection of information inferred from tweet meta-fields. *International Journal of Applied Earth Observation and Geoinformation*, 105, 102593.

Appendix

Table A1. List of identified topics from Twitter including the ones mentioned throughout the paper (cont.): words, translations, probabilities as well as the count of documents and instances.

Topic 0 Hexagons: 179/tweets: 110757			Topic 1 Hexagons: 105/tweets: 53074			Football (Topic 2) Hexagons: 49/tweets: 41079			Downtown (Topic 3) Hexagons: 43/tweets: 33237			Alfama (Topic 4) Hexagons: 43/tweets: 26754			Topic 5 Hexagons: 179/tweets: 25049		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Vou	(I) go	0.0242	Padrão	Touristic landmark	0.0341	Estádio	Stadium	0.1462	Comércio	Business	0.0754	Duetos	Duet	0.0698	Vou	(I) go	0.0268
Vasco	Vasco da Gama mall	0.0144	Descobrimentos	Touristic landmark	0.0338	Benfica	Benfica football team	0.1183	Terreiro	Terreiro do Paço plaza	0.0611	Bar	Bar	0.0472	Escola	School	0.0213
Gama	Vasco da Gama mall	0.0143	Alfama	Alfama neighborhood	0.0249	Sport	Benfica football team	0.0995	Praça	Plaza	0.0610	Amp	Organization in Alfama	0.0434	Tou	(I) am	0.0208
Casa	House/home	0.0117	Arena	–	0.0209	Alvalade	José Alvalade stadium	0.0604	paço	Terreiro do Paço plaza	0.0596	Alfama	Alfama neighborhood	0.0420	Chiado	Chiado neighborhood	0.0167
Tou	(I) am	0.0114	Parque	Eduardo VII park	0.0204	José	José Alvalade stadium	0.0568	fil	FIL exhibition center	0.0173	Gastronomia	Gastronomy	0.0400	Brasileira	Brasileira café	0.0160
Parque	Park	0.0111	Eduardo	Eduardo VII park	0.0193	slbenfica	Benfica football team	0.0470	Museu	Musesum	0.0171	Restaurante	Restaurant	0.0375	Accessories	–	0.0139
Dormir	To sleep	0.0108	vii	Eduardo VII park	0.0178	Xxi	–	0.0414	cinema	–	0.0167	Café	Café	0.0341	Jewelry	–	0.0139
Escola	School	0.0107	Meo	Telecommunication company	0.0163	Sporting	Sporting football team	0.0374	lxfactory	LXFactory open mall	0.0164	Praça	Plaza/square	0.0299	Jewellery lover	–	0.0139
Mim	Me	0.0103	Nações	Parque das Nações parish	0.0160	Luz	Luz stadium	0.0301	Nacional	National	0.0160	Mercado	Market	0.0297	Jewellery design	–	0.0139
Acho	(I) think	0.0099	Centro	Center	0.0123	Carregabenfica	Benfica football team	0.0267	Feira	Market/fair	0.0155	Ribeira	Area in Lisbon	0.0277	Jewellery designer	–	0.0138
Técnico	IST university	0.0099	Telheiras	Telheiras neighborhood	0.0115	Campo	Field	0.0266	Internacional	International	0.0144	Fado	Fado music	0.0269	Jewellery addict	–	0.0138
Amanhã	Tomorrow	0.0098	Museu	Museum	0.0106	Bairro	Neighborhood	0.0235	Cinemateca	Cinemateca film archive	0.0133	Sobremesa	Dessert	0.0268	Paiva	Portuguese surname	0.0137
Vida	Life	0.0095	Vou	(I) go	0.0102	Alto	Tall/high	0.0232	Posted	–	0.0132	Comércio	Business	0.0260	Silver	–	0.0137
IST	IST university	0.0094	Saldanha	Saldanha neighborhood	0.0095	Slb	Benfica football team	0.0214	Livraria	Bookshop	0.0122	Música	Music	0.0252	Jewelry gram	–	0.0132
Centro	Center	0.0094	Melhor	Best/better	0.0090	Sportingcp	Sporting football Team	0.0204	Main	–	0.0111	Concerto	Concert	0.0248	Carmo	Carmo convent	0.0132

(*continued*)

Topic 6			Topic 7			Topic 8			Topic 9			Topic 10			Topic 11		
Hexagons: 37/tweets: 14304			Hexagons: 32/tweets: 16763			Hexagons: 32/tweets: 10100			Hexagons: 27/tweets: 14514			Hexagons: 26/tweets: 12839			Hexagons: 24/tweets: 12316		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Ajuda	Ajuda neighborhood	0.0322	Pombal	Marquês de Pombal (square)	0.1006	Campo	Field	0.0257	Parque	Parque das Nações parish	0.1621	Conhecimento	Pavilhão do Conhecimento (museum)	0.0361	Vou	(I) go	0.0358
Alcântara	Alcântara neighborhood	0.0179	Marquês	Marquês de Pombal (square)	0.0971	Ilha	Island	0.0237	Oceanário	Oceanarium	0.1404	Pavilhão	Pavilhão do Conhecimento (museum)	0.0279	Livro	Book	0.0193
Europe	–	0.0137	Amaz	–	0.0369	Vou	(I) go	0.0230	Nações	Parque das Nações parish	0.1376	RTP	Portuguese public broadcasting	0.0270	Mim	Me	0.0180
República	Republic	0.0132	Metro	Subway/metro	0.0258	Amores	Love (plural)	0.0224	expo	Lisbon Expo '98	0.0599	Vou	(I) go	0.0223	Casa	House/home	0.0164
Ponte	Bridge	0.0120	Chafariz	Fountain	0.0224	Jardim	Garden	0.0168	Naçõesparque	Parque das Nações parish	0.0505	Televisão	Television	0.0195	Feira	Market	0.0161
Hotelaria	Hospitality	0.0110	Enoteca	Wine cellar	0.0217	Peninsular	–	0.0166	Tweetgram	–	0.0324	Rádio	Radio	0.0192	Chilihotgirl	–	0.0159
Turismo	Tourism	0.0109	Vou	(I) go	0.0211	Rua	Street	0.0164	Tweegram	–	0.0322	Ciência	Science	0.0179	Jogos	Games	0.0153
Abril	April (bridge)	0.0104	Edinho	Portuguese football player	0.0184	Heróis	Heros	0.0158	Webstagram	–	0.0320	Casa	House/home	0.0163	Comigo	With me	0.0132
Marquês	Marquês de Pombal (square)	0.0102	Vinho	Wine	0.0172	Campos	Fields	0.0154	Wow	–	0.0320	Pessoa	Fernando Pessoa (portuguese writer)	0.0149	Falar	(To) speak/talk	0.0128
Avenida	Avenue	0.0102	Praça	Plaza/square	0.0151	Tou	(I) am	0.0131	Congressos	Congresses	0.0275	Fernando	Fernando Pessoa (portuguese writer)	0.0143	Tou	(I) am	0.0121
Rato	Rato neighborhood	0.0098	Blog	–	0.0145	Guerra	War	0.0131	Okportugal	–	0.0246	Twitter	–	0.0136	Merda	Swear word	0.0120
Pombal	Marquês de Pombal (square)	0.0090	Prof	–	0.0140	Metro	Subway/metro	0.0122	Amarlisboa	(to) love lisbon	0.0213	Viva	Alive/cheer	0.0130	Amanha	Tomorrow	0.0115
Espaco exibicionista	Local art gallery	0.0087	Vídeo	–	0.0140	República	Republic	0.0121	Posted	–	0.0200	Damaralines	–	0.0123	CRL	Swear word	0.0114
Nações	Parque das Nações Parish	0.0086	Acabou	(it) ended/ran out/just happenEd	0.0125	Casa	House/home	0.0113	Monsanto	Monsanto park	0.0198	Tou	(I) am	0.0113	Melhor	Best/better	0.0113
Exhibition	–	0.0084	Hotel	–	0.0105	Campolide	Campolide neighborhood	0.0113	tejo	Tagus river	0.0192	Campo	Field	0.0108	Dormir	(to) sleep	0.0111

(*continued*)

Topic 12 Hexagons: 23/tweets: 9592			Topic 13 Hexagons: 23/tweets: 16909			Topic 14 Hexagons: 23/tweets: 11105			Topic 15 Hexagons: 23/tweets: 15015			Belém and tourist attractions (Topic 16) Hexagons: 23/tweets: 20123			Topic 17 Hexagons: 20/tweets: 7132		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Amoreiras	Amoreiras mall	0.0695	Chiado	Chiado neighborhood	0.1190	Mega	–	0.0822	Vou	(I) go	0.0294	Torre	Tower (Belém tower)	0.1179	Lusófona	Lusófona university	0.0354
Hospital	–	0.0689	Estação	Station	0.0942	Craque	ACE/expert	0.0766	Dágua	Linha d'Água restaurant	0.0230	Belém	Belém neighborhood	0.1036	ULHT	Lusófona university	0.0291
Humanas	Humanities	0.0603	Oriente	Oriente train station	0.0850	Clube	Club	0.0597	Linha	Linha d'Água restaurant	0.0197	Castelo	Saint George castle	0.0872	Laranjeiras	Laranjeiras neighborhood	0.0273
Martim	Martim Moniz square	0.0581	Armazéns	Warehouses/ Armazéns do Chiado mall	0.0849	Rios	Rivers	0.0397	Mim	Me	0.0130	Jorge	Saint George castle	0.0776	Carlotatavares	–	0.0272
Moniz	Martim Moniz square	0.0573	Gare	Oriente train Station	0.0754	Rua	Street	0.0332	Vida	Life	0.0121	Belem	Belém Neighborhood	0.0560	Humanidades	Humanities	0.0255
Ciências	Sciences	0.0521	Ferroviária	Railway	0.0654	Catedral	Cathedral	0.0315	Acho	(I) think	0.0118	Miradouro	St. Peter of Alcântara viewpoint	0.0482	Cidadão	Citizen	0.0236
Sociais	Social	0.0520	Rossio	Rossio square	0.0588	Mundial	Worldwide	0.0285	Dormir	(to) sleep	0.0118	Alcântara	St. Peter of Alcântara viewpoint	0.0469	Universidade	University	0.0225
Fcshunl	Social sciences university	0.0484	Baixachiado	Chiado neighborhood	0.0469	Esperança	Hope	0.0264	Falar	(to) speak/talk	0.0116	pedro	St. Peter of Alcântara viewpoint	0.0438	Tecnologias	Technologies	0.0208
Faculdade	faculty/university	0.0453	Baixa	Lisbon's downtown	0.0379	Crossfit	–	0.0264	Tou	(I) am	0.0115	Tower	–	0.0432	Morgas	–	0.0195
Center	–	0.0411	Starbucks	–	0.0345	Augusta	Augusta street	0.0250	Mãe	Mother	0.0112	Cultural	–	0.0402	Vou	(I) go	0.0189
Shopping	–	0.0385	AZVD	–	0.0334	Estação	Station	0.0238	Casa	House/home	0.0111	Saldanha	Saldanha neighborhood	0.0293	Loja	Store/shop	0.0168
SAMS	private hospital	0.0287	Metro	Subway/metro	0.0333	Ferroviária	Railway	0.0238	Pessoa	Person	0.0111	CCB	Belém culture center	0.0274	Amigadop	–	0.0155
FCSH	Social sciences university	0.0284	Entrecampos	Entrecampos metro station	0.0328	Susanagateira	Local store	0.0237	Amanhã	Tomorrow	0.0102	Centro	Center	0.0269	Modernização	Modernization	0.0147
Clínico	Clinic	0.0277	Others	–	0.0225	Megacraque	(Mega) ACE/expert	0.0211	Ihugof	–	0.0099	Berardo	Berardo museum	0.0231	Administrativa	Administrative	0.0144
Universidade	University	0.0233	Pontinha	Pontinha neighborhood	0.0224	Vitória	Victory	0.0200	Ficar	(to) stay	0.0097	Mosteiro	Monastery (in Belém)	0.0214	Academica	Academic	0.0141

Table A2. List of identified topics from Twitter including the ones mentioned throughout the paper: words, translations, probabilities as well as the count of documents and instances.

Topic 18 Hexagons: 20/tweets: 12159			Picoas and Saldanha neighborhoods (Topic 19) Hexagons: 19/tweets: 6312			Topic 20 Hexagons: 17/tweets: 30270			Topic 21 Hexagons: 16/tweets: 6189			Airport (Topic 22) Hexagons: 16/tweets: 14991			Topic 23 Hexagons: 15/tweets: 4778		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Camões	Luis de Camões square	0.0694	Picoas	Picoas neighborhood	0.0332	Senhora	Madam/woman	0.0199	chapitô	Chapitô circus	0.0539	Aeroporto	Airport	0.3283	Iscal	ISCAL university	0.0527
Chiado	Chiado neighborhood	0.0693	Petiscar	To nibble (food)	0.0303	Acessórios	Accessories	0.0165	Adrianamars	–	0.0418	Lis	Lisbon airport	0.1976	Contabilidade	Accounting	0.0311
Luís	Luis de Camões square	0.0612	Villaret	Villaret heatre Picoas	0.0293	Isabel	Isabel Ribeiro wedding clothes	0.0153	Marcação	Appointment	0.0392	Others	–	0.0559	Administração	Administration	0.0268
Gulbenkian	Calouste Gulbenkian garden	0.0592	Gulbenkian	Gulbenkian park	0.0269	Ribeiro	Isabel Ribeiro wedding clothes	0.0146	Gel	–	0.0376	Delgado	Lisbon airport	0.0467	Noivos	Grooms (wedding)	0.0255
Jardim	Garden	0.0585	Jardim	Garden	0.0266	Homem	Man	0.0136	Blog	–	0.0373	Humberto	Lisbon airport	0.0466	Noivas	Brides	0.0232
Doces	Sweets	0.0565	Fontes	Fontes P.M avenue	0.0261	Cerimónia	Cerimony	0.0133	Strength	–	0.0330	Chegadas	Arrivals	0.0446	Acessórios	Accessories	0.0229
Calouste	Calouste gulbEnkian garden	0.0553	Entrenospetisca	To nibble (Food)	0.0258	Corroios	Corroios pariSh	0.0128	Imersaoempo	Manicure	0.0326	Airport	–	0.0432	Vestimos	(we) dress	0.0227
Sonhos	Dreams	0.0520	Saúde	Health	0.0258	Glamour	–	0.0126	Verniz	Nail polish	0.0233	Arrivals	–	0.0431	Instituto	Institute	0.0217
Fundação	Foundation	0.0491	Tecnologia	Technology	0.0250	Criança	Child	0.0126	Thestrengthclinic	Local gym	0.0232	Terminal	–	0.0423	Rapaz	Young man	0.0214
Zoológico	Zoo	0.0483	Melo	Fontes P.M avenue	0.0244	Noivos	Groom (wedding)	0.0122	Health	–	0.0191	Departures	–	0.0367	Superior	–	0.0211
Bairroaltohotel	Hotel	0.0410	Estesl	Estesl university	0.0235	Noiva	Bride	0.0121	Vou	(i) go	0.0171	Partidas	Departures	0.0367	Altis	Altis hotels	0.0208
Praça	Square	0.0338	Conversar	To chat	0.0231	Qualidade	Quality	0.0121	Vitoriapti	Pti election winning	0.0165	Comunidades	Communities	0.0332	Portela	Portela parish	0.0196
Largo	Square	0.0307	Pereira	Fontes P.M avenue	0.0229	Requinte	Refinement	0.0118	Longevity	–	0.0161	Lisboalis	Lisbon airport	0.0321	Senhora	Madam/woman	0.0192
Taberna	Tavern	0.0296	Superior	Higher	0.0220	Vida	Life	0.0118	Escola	School	0.0159	Internacional	International	0.0318	Avenidas	Avenidas Novas parish	0.0181
Botânico	Botanic	0.0262	Escola	School	0.0216	Lojas	Stores	0.0117	Areeiro	Areeiro neighborhood	0.0149	Portuguesas	Portuguese	0.0312	Especializados	Specialized	0.0180

Topic 24 Hexagons: 15/tweets: 31609			Topic 25 Hexagons: 14/tweets: 22732			Topic 26 Hexagons: 13/tweets: 7744			Topic 27 Hexagons: 12/tweets: 10069			University (Topic 28) Hexagons: 11/tweets: 3580			Topic 29 Hexagons: 10/tweets: 2386		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
District	–	0.0459	Graça	Graça neighborhood	0.0247	Monte	Nossa Senhora do Monte viewpoint	0.1055	Mude	(you) change	0.0470	Faculdade	Faculty/university	0.1310	Estufa	Estufa Fria garden	0.0851
Cais	Cais do Sodré neighborhood	0.0229	Urban	Urban Beach nightclub	0.0227	Senhora	Nossa Senhora do Monte viewpoint	0.0858	Augusta	Augusta street Arch	0.0404	Ciências	Sciences	0.1254	Fria	Estufa Fria garden	0.0759
Sodré	Cais do Sodré neighborhood	0.0219	Beach	Urban Beach nightclub	0.0206	Miradouro	Nossa Senhora do Monte viewpoint	0.0720	Arco	Augusta street Arch	0.0370	Universidade	University	0.1218	Eduardo	Eduardo VII park	0.0748

(*continued*)

Pensão	Pensão Amor bar	0.0212	Miradouro	Viewpoint	0.0178	Lust	Lust in Rio nightclub	0.0690	Design	–	0.0364	Colombo	Colombo mall	0.0557	vii	Eduardo VII park	0.0595
Amor	Pensão Amor bar	0.0192	Européia	European	0.0156	Micasaenlisboa	–	0.0611	Moda	Fashion	0.0318	Cinemas	–	0.0552	Eleven	Venue in Eduardo VII park	0.0564
Iscteiul	ISCTE university	0.0171	Cidade	City	0.0154	Rio	Lust in Rio nightclub	0.0605	Comércio	Commerce/comércio square	0.0301	Campus	–	0.0530	Depósitos	Portuguese bank	0.0423
Vídeo	Video	0.0156	Marina	Sea park	0.0153	Lav	–	0.0385	Marcações	appointments	0.0290	Justiça	Justice	0.0505	Caixa	Portuguese bank	0.0393
Universitário	University	0.0143	União	Union	0.0150	Bomdia	Good morning	0.0383	Rua	Street	0.0279	FCUL	University of Lisbon	0.0381	Sede	headquarters	0.0392
Acabou	(it) ended/ran out/just happened	0.0130	Vida	Life	0.0120	Mstattoos	Tattoo place	0.0335	Whatsapp	–	0.0255	Medicina	Medicine	0.0355	Park	–	0.0365
ISCTE	ISCTE university	0.0126	Photography	–	0.0107	Vivo	(I) live	0.0239	Disponíveis	Availables	0.0253	Dentária	Dental	0.0343	Parque	Parque	0.0287
Justa	Santa Justa neighborhood and llift	0.0120	Photooftheday	–	0.0107	Tattoolisboa	–	0.0237	Agende	(you) schedule	0.0240	Holmes	Local gym chain (Holme's place)	0.0265	Roma	Roma avenue	0.0282
Instituto	Institute	0.0115	Celso	Celso barbershop	0.0101	Tattooportugal	–	0.0236	Praça	Square/plaza	0.0236	Campo	Field	0.0264	Avenida	Roma avenue	0.0175
Msg	Message	0.0110	Barbeiro	Celso barbershop	0.0098	Tattoo	–	0.0199	Museu	Musesum	0.0231	Place	Local gym chain (Holme's place)	0.0208	Claudiaccamposh	–	0.0148
Vamuver	YouTube channel	0.0102	Melhor	best/better	0.0097	Tattoos	–	0.0199	Horário	Time/timeslot	0.0219	Filme	Film	0.0168	Frutalmeidas	Café at Roma avenue	0.0135
Carmo	Carmo convent/Carmo square	0.0099	Igers	Online photography community	0.0097	Meninos	Boys	0.0189	Triunfal	Augusta street Arch	0.0198	IMAX	IMAX cinema	0.0160	Lar	Home	0.0134

Topic 30
Hexagons: 10/tweets: 7112

Word	Translation	Prob.
Coliseu	Coliseu dos Recreios auditorium	0.1654
Recreios	Coliseu dos Recreios auditorium	0.1286
Marriott	–	0.1185
Hotel	–	0.0963
Restauradores	Restauradores square	0.0639
Hard	Hard Rock Café	
Rock	Hard Rock Café	0.0477
bbc	–	0.0440
Café	–	0.0225
Praça	Square/Plaza	0.0221
Bar	–	0.0216
Coliseudelisboa	Coliseu dos Recreios auditorium	0.0154
Belém	Belém neighborhood	0.0148
MAAT	MAAT museum	0.0143

Table A3. List of identified topics from OSM and Google Places including the ones mentioned throughout the paper (cont.): words, translations, probabilities as well as the count of documents and instances.

Topic 0			Topic 1			Topic 2			Topic 3			Health (Topic 4)			Services (Topic 5)		
Hexagons: 93/OSM: 1555/Google POI: 515			Hexagons: 88/OSM: 341/Google POI: 417			Hexagons: 53/OSM: 432/Google POI: 198			Hexagons: 52/OSM: 546/Google POI: 263			Hexagons: 41/OSM: 198/Google POI: 232			Hexagons: 40/OSM: 196/Google POI: 209		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Hotel	–	0.0262	Profissionais	Professionals	0.0175	Igreja	Church	0.0461	Food	–	0.0232	Good	–	0.0145	Recomendo	(I) recommend	0.0219
Hostel	–	0.0211	Excelente	Excellent	0.0174	Church	–	0.0294	Oriente	Oriente train station	0.0193	Atendimento	Service/treatment	0.0144	Serviço	Service	0.0182
Location	–	0.0208	Serviço	Service	0.0171	Santo	Saint	0.0180	Great	–	0.0182	Excelente	Excellent	0.0140	Atendimento	Customer service	0.0180
Place	–	0.0175	Atendimento	Service/treatment	0.0165	Senhora	Madam/woman	0.0157	Nice	–	0.0174	Farmácia	Pharmacy	0.0133	Ajuda	Ajuda neighborhood	0.0168
Good	–	0.0169	Lda	Company	0.0156	Capela	Chapel	0.0156	Staff	–	0.0169	Service	–	0.0129	Excelente	Excellent	0.0152
Clean	–	0.0168	Empresa	Company	0.0148	Convento	Covent	0.0153	Place	–	0.0166	Saúde	Health	0.0127	Restelo	Restelo neighborhood	0.0152
Great	–	0.0154	Melhor	Best/better	0.0147	Place	–	0.0143	Good	–	0.0160	Clínica	Clinic	0.0124	Rua	Street	0.0152
Stay	–	0.0151	Recomendo	(I) recommend	0.0143	Santa	Saint	0.0142	Amazing	–	0.0135	Centro	Center	0.0114	Adorei	(I) loved (it)	0.0138
Room	–	0.0141	Rua	Street	0.0142	Nice	–	0.0142	Friendly	–	0.0134	Hospital	–	0.0112	Melhor	Best/better	0.0124
Nice	–	0.0132	Preços	Prices	0.0127	Beautiful	–	0.0141	Restaurant	–	0.0129	Simpatia	Sympathy	0.0103	Profissional	Professional	0.0121
Apartment	–	0.0132	Profissional	Professional	0.0127	Cruz	Cross	0.0112	Best	–	0.0119	Café	–	0.0103	Avenida	Avenue	0.0116
Staff	–	0.0131	Excelentes	Excellent	0.0118	Colégio	Private school	0.0101	Service	–	0.0116	Great	–	0.0103	Lda	Company	0.0116
Rooms	–	0.0120	Qualidade	Quality	0.0107	Clara	Claire/bright	0.0099	One	–	0.0110	Lda	Company	0.0102	Profissionais	Professionals	0.0116
Friendly	–	0.0116	Caixa	Box/cashier	0.0100	Memória	Memory	0.0097	Super	–	0.0107	Appointment	–	0.0092	Bairro	Neighborhood	0.0114
Breakfast	–	0.0108	Top	–	0.0095	Good	–	0.0091	Estação	Station	0.0107	Rua	Street	0.0091	Simpatia	Sympathy	0.0109
Topic 6			Education (Topic 7)			Topic 8			Topic 9			Topic 10			Topic 11		
Hexagons: 35/OSM: 225/Google POI: 241			Hexagons: 35/OSM: 140/Google POI: 96			Hexagons: 34/OSM: 170/Google POI: 251			Hexagons: 33/OSM: 244/Google POI: 212			Hexagons: 32/OSM: 366/Google POI: 204			Hexagons: 29/OSM: 126/Google POI: 157		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Escola	School	0.0155	Escola	School	0.0800	Lda	Company	0.0169	Great	–	0.0216	Hotel	–	0.0154	Service	–	0.0188
Ajuda	Ajuda neighborhood	0.0148	School	–	0.0652	Excelente	Excellent	0.0167	Good	–	0.0194	Rato	Rato neighborhood	0.0130	Shop	–	0.0149
Lda		0.0127	University	–	0.0312	Atendimento	Customer serviec	0.0165	Service	–	0.0174	embaixada	embassy	0.0128	Atendimento	Customer service	0.0142
Atendimento	Customer service	0.0124	Faculdade	faculty/university	0.0218	Recomendo	(i) recommend	0.0139	Food	–	0.0156	Place	–	0.0111	Great	–	0.0140
Service	–	0.0121	Universidade	university	0.0216	Serviço	Service	0.0134	Atendimento	Customer service	0.0154	Lda	–	0.0106	Good	–	0.0136
Loja	Store	0.0115	Azulejos	Portuguese tiles	0.0197	Simpatia	Sympathy	0.0122	Nice	–	0.0136	Staff	–	0.0103	Santogal	–	0.0136
Benfica	Benfica neighborhood	0.0113	Registo	Registration	0.0181	Avenida	Avenue	0.0112	Staff	–	0.0132	Hostel	–	0.0102	Place	–	0.0131
Good	–	0.0113	Teachers	–	0.0169	Top	–	0.0109	Marquês	Marquis	0.0124	Great	–	0.0095	Oficina	Workshop	0.0126
Excelente	Excellent	0.0112	José	José Fontana square	0.0166	Super	–	0.0109	Belém	Belém neighborhood	0.0121	Chile	Chile square	0.0095	Staff	–	0.0125
Marquês	Marquês de P. square	0.0108	Ensino	Teaching/education	0.0164	Profissionalismo	Professionalism	0.0106	Place	–	0.0117	Room	–	0.0094	Super	–	0.0123
Pombal	Marquês de P. square	0.0107	Professores	Professors	0.0154	Profissionais	Professionals	0.0105	Excelente	Excellent	0.0107	Rua	Street	0.0093	Lda	Company	0.0115
Espaço	Space	0.0107	Fontana	José Fontana square	0.0152	Great	–	0.0102	Serviço	Service	0.0105	Excelente	Excellent	0.0093	Nice	–	0.0113
Nice	–	0.0104	Faculty	–	0.0142	Empresa	Company	0.0101	Friendly	–	0.0099	Good	–	0.0092	Kia	–	0.0108
Centro	Center	0.0102	Superior	Higher (education)	0.0137	Campo	Field	0.0101	Qualidade	Quality	0.0094	Atendimento	Customer service	0.0091	Friendly	–	0.0104
Rua	Street	0.0100	Campus	–	0.0131	Rua	Street	0.0097	Espaço	Space	0.0089	Avenida	Avenue	0.0091	Helpful	–	0.0101

(*continued*)

Topic 12 Hexagons: 26/OSM: 224/Google POI: 150			Topic 13 Hexagons: 26/OSM: 96/Google POI: 44			Topic 14 Hexagons: 24/OSM: 272/Google POI: 67		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Good	–	0.0230	Monsanto	Monsanto park	0.1403	Benfica	Benfica neighborhod	0.0349
Pingo	Pingo Doce supermarket	0.0173	Parque	Monsanto park	0.0644	Sapadores	Firefighters	0.0290
Doce	Pingo Doce supermarket	0.0164	Interpretação	Interpretation	0.0611	Avenida	Avenue	0.0255
Sebastião	Portuguese name	0.0147	Infantil	Infantile	0.0325	Comida	Food	0.0226
Store	–	0.0130	Acesso	Access	0.0320	Farmácia	Pharmacy	0.0216
Place	–	0.0124	Alvito	Alvito neighborhood	0.0276	Simpatia	Sympathy	0.0179
Expo	Lisbon Expo '98	0.0123	Centro	Center	0.0269	Grão	Grain	0.0178
Nice	–	0.0121	João	Portuguese name	0.0268	Atendimento	Customer service	0.0178
Service	–	0.0112	Água	Water	0.0251	Laranjeiras	Laranjeiras neighborhood	0.0172
Fonte	fountain	0.0109	Tribunal	Court	0.0244	Prazeres	Prazeres cemitery	0.0171
Well	–	0.0107	Oliveiras	Olive trees	0.0244	Good	–	0.0170
Sushi	–	0.0106	Campolide	Campolide neighborhood	0.0238	Produtos	Products	0.0157
One	–	0.0104	Estacionamento	Car park	0.0237	Tiffosi	Portuguese clothing store	0.0155
Etnologia	Ethnology	0.0102	Gil	Portuguese name	0.0231	Qualidade	Quality	0.0154
Food	–	0.0102	Running	–	0.0228	Restaurante	Restaurant	0.0152

Airport (Topic 15) Hexagons: 24/OSM: 91/Google POI: 135			Topic 16 Hexagons: 23/OSM: 112/Google POI: 114			Topic 17 Hexagons: 23/OSM: 142/Google POI: 111		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Aeroporto	Airport	0.0424	Work	–	0.0204	Food	–	0.0265
Airport	–	0.0403	Nice	–	0.0171	Caselas	Caselas neighborhood	0.0222
Flight	–	0.0168	Top	–	0.0155	Ice	–	0.0221
Terminal	–	0.0145	Great	–	0.0144	Cream	–	0.0207
Taxi	–	0.0138	Bar	–	0.0136	Portuguese	–	0.0150
Euros	–	0.0136	Company	–	0.0135	Great	–	0.0150
Worst	–	0.0135	Edifício	Building	0.0128	Campolide	Campolide neighborhood	0.0147
Olivais	Olivais neighborhood	0.0131	Staff	–	0.0128	Typical	–	0.0146
Ever	–	0.0131	Rua	Street	0.0119	Lda	Company	0.0143
Queiroz	Portuguese surname	0.0129	Studio	–	0.0112	Comida	Food	0.0126
Waiting	–	0.0126	People	–	0.0112	Excelente	Excellent	0.0126
Car	–	0.0124	Job	–	0.0112	Amazing	–	0.0126
Excelente	Excellent	0.0122	Gestão	Management	0.0111	Restelo	Restelo neighborhood	0.0125
DPD	DPD parcel delivery	0.0121	Place	–	0.0111	Good	–	0.0125
Minutes	–	0.0120	Figura	Figure	0.0110	Rua	Street	0.0124

Table A4. List of identified topics from OSM and Google Places including the ones mentioned throughout the paper: words, translations, probabilities as well as the count of documents and instances.

Shopping mall (Topic 18) Hexagons: 20/OSM: 107/Google POI: 156			Football (Topic 19) Hexagons: 19/OSM: 55/Google POI: 77			Alfama (Topic 20) Hexagons: 18/OSM: 202/Google POI: 125			Topic 21 Hexagons: 18/OSM: 44/Google POI: 34			Topic 22 Hexagons: 18/OSM: 70/Google POI: 116			Topic 23 Hexagons: 16/OSM: 78/Google POI: 77		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Atendimento	Customer service	0.0215	Stadium	–	0.0523	Alfama	Alfama neighborhood	0.0393	miradouro	Viewpoint	0.0925	Boat	–	0.0448	Penha	Penha de França neighborhood	0.0369
Colombo	Colombo mall	0.0161	Club	–	0.0501	Fado	Fado music	0.0279	Claros	Montes Claros garden	0.0829	Tours	–	0.0314	França	Penha de França neighborhood	0.0363
Serviço	–	0.0147	Gym	–	0.0308	Casa	House	0.0177	Montes	Montes Claros garden	0.0829	Sailing	–	0.0314	Car	–	0.0216
Good	–	0.0144	Football	–	0.0293	Amazing	–	0.0157	View	–	0.0709	Experience	–	0.0288	Ctt	Portuguese postal service	0.0213
Lda	Company	0.0139	Sport	Benfica football team	0.0284	Palace	–	0.0153	Moinho	Moinho park	0.0510	River	–	0.0287	Police	–	0.0205
Loja	Store	0.0131	Benfica	Benfica football team	0.0279	Apartment	–	0.0147	Nice	–	0.0485	Amazing	–	0.0286	Portuguesa	Portuguese	0.0176
Empresa	Company	0.0130	Estádio	Stadium	0.0265	Place	–	0.0145	Place	–	0.0480	Tour	–	0.0245	Rotunda	Roundabout	0.0171
Centro	Center	0.0127	Great	–	0.0235	Food	–	0.0139	Parque	Park	0.0471	Crew	–	0.0236	Matos	Júlio de Matos hospital	0.0155
Excelente	Excellent	0.0119	Best	–	0.0207	Like	–	0.0136	Park	–	0.0435	Great	–	0.0227	Psp	Portuguese civil police	0.0152
Really	–	0.0111	Good	–	0.0192	Great	–	0.0130	Vista	View	0.0430	View	–	0.0224	Pay	–	0.0140
Service	–	0.0099	Staff	–	0.0176	Location	–	0.0124	Nature	–	0.0406	Nice	–	0.0221	Charneca	Charneca neighborhood	0.0139
Profissionalismo	Professionalism	0.0096	Luz	Luz stadium	0.0175	Beautiful	–	0.0123	Kids	–	0.0377	Tejo	Tagus river	0.0209	Rent	–	0.0134
Preço	Price	0.0095	Place	–	0.0169	Chafariz	Fountain	0.0123	Picnic	–	0.0372	Trip	–	0.0204	Hospital	Júlio de Matos hospital	0.0129
Telheiras	Telheiras neighborhood	0.0094	Olivais	Olivais neighborhood	0.0162	Clean	–	0.0122	Olhão	Portuguese town	0.0367	Recommend	–	0.0176	Júlio	Júlio de Matos hospital	0.0128
Equipa	Team/staff	0.0093	Courts	–	0.0161	Everything	–	0.0121	Mocho	Moinho park	0.0355	Docas	Docks	0.0162	Lda	Company	0.0126
Topic 24 Hexagons: 15/OSM: 117/Google POI: 55			Topic 25 Hexagons: 15/OSM: 185/Google POI: 91			Topic 26 Hexagons: 14/OSM: 50/Google POI: 84			Topic 27 Hexagons: 13/OSM: 114/Google POI: 112			Topic 28 Hexagons: 12/OSM: 42/Google POI: 63			Topic 29 Hexagons: 12/OSM: 47/Google POI: 76		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Museu	museum	0.0348	Campo	Field	0.0406	Radio	–	0.0284	Lumiar		0.0217	Cuf	CUF hospital	0.0284	Portas	Portas de Benfica building	0.0289
Carnide	Carnide neighborhood	0.0312	Good	–	0.0228	Renascença	Portuguese radio station	0.0173	Alameda	Alameda square	0.0210	Carnide	Carnide neighborhood	0.0253	Atendimento	customer service	0.0286
Teatro	Theater	0.0286	mouraria	Mouraria neighborhood	0.0226	Great	–	0.0172	Microsoft	–	0.0209	Excelente	Excellent	0.0199	Spacio	Spacio mall	0.0279
Arte	Art	0.0246	Great	–	0.0207	Soluções	Solutions	0.0168	Clinic	–	0.0189	Lumiar	Lumiar neighborhood	0.0196	Boavista	Boavista neighborhood	0.0275
Nacional	National	0.0229	Place	–	0.0176	Best	–	0.0163	Clínica	Clinic	0.0184	Hospital	–	0.0183	Benfica	Benfica neighborhood	0.0231
Theater	–	0.0202	Nice	–	0.0171	Lda	Company	0.0161	Atendimento	Customer service	0.0148	School	–	0.0173	Shopping	–	0.0226

(*continued*)

Cultural	–	0.0188	Best	–	0.0169	Alvalade	Alvalade neighborhood	0.0159	Nice	–	0.0140	Pavilhão	Pavilion	0.0167	Produtos	Products	0.0210
Roque	Saint Roch church	0.0185	Belém	Belém neighborood	0.0164	Empresa	Company	0.0152	Emel	Lisbon's mobility	0.0126	Desportivo	Sports	0.0167	Qualidade	Quality	0.0186
Museum	–	0.0185	Moniz	Martim Moniz square	0.0157	Encarnação	Encarnação neighborhood	0.0148	Excelente	Excellent	0.0125	Pay	–	0.0167	Roupa	Clothing	0.0184
Piano	–	0.0181	Loureiro	Loureiro neighborhood	0.0151	Rádio	Radio	0.0144	Farmácia	Pharmacy	0.0119	Lispolis	LISPOLIS technology hub	0.0165	Costa	Coast	0.0182
Venue	–	0.0176	Chão	Ground/floor	0.0146	Equipa	Team/staff	0.0142	Henriques	Portuguese surname	0.0118	Dresses	–	0.0161	Excelente	Excellent	0.0176
Antiga	Old/ancient	0.0169	Food	–	0.0144	Professional	–	0.0140	Medicina	Medicine	0.0116	Macau	–	0.0157	Loja	Store	0.0173
Estufa	Greenhouse	0.0168	Mamede	São Memede neighborhood	0.0138	Instruments	–	0.0135	Centro	Center	0.0111	Escola	–	0.0155	Serviço	Service	0.0155
Lourenço	Portuguese name	0.0156	Martim	Martim Moniz square	0.0128	Rfm	Portuguese radio station	0.0135	Best	–	0.0109	Staff	–	0.0152	José	Portuguese name	0.0153
Locanda	–	0.0153	Hostel	–	0.0125	Mascarenhas	Portuguese surname	0.0135	Clínicas	Clinics	0.0109	Portuguese	–	0.0150	Good	–	0.0151

Sports (Topic 30) Hexagons: 12/OSM: 23/Google POI: 26			Topic 31 Hexagons: 11/OSM: 27/Google POI: 15			Topic 32 Hexagons: 11/OSM: 142/Google POI: 92			Topic 33 Hexagons: 10/OSM: 32/Google POI: 64			Topic 34 Hexagons: 10/OSM: 77/Google POI: 60		
Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.	Word	Translation	Prob.
Futebol	Football	0.0554	Parque	Park	0.1208	Pizza	–	0.0409	Nice	–	0.0168	Bike	–	0.0254
Campo	Field	0.0516	Keil	Keil do Amaral park	0.1008	Good	–	0.0186	Staff	–	0.0163	Climbing	–	0.0243
Musgueira	Musgueira sports complex	0.0451	Amaral	Keil do Amaral park	0.0968	Chiado	Chiado neighborhood	0.0176	Service	–	0.0161	Tours	–	0.0237
Bandeiras	Flags	0.0451	Park	–	0.0909	Food	–	0.0171	Diamond	–	0.0159	Cais	Cais do Sodré neighborhood	0.0235
Desportivo	Sports	0.0449	Dog	–	0.0779	Staff	–	0.0162	Friendly	–	0.0159	Tour	–	0.0220
Ténis	Tennis	0.0407	Canino	Canine	0.0707	Browns	–	0.0155	Good	–	0.0156	Great	–	0.0210
Park	–	0.0387	Alameda	Keil do Amaral park	0.0517	Restelo	Restelo neighborhood	0.0154	Aranha	Spider	0.0153	Beer	–	0.0199
Universitário	University	0.0368	Picnic	–	0.0510	Place	–	0.0154	Patudos	Bigeye fish	0.0153	Excellent	–	0.0195
Tennis	–	0.0346	Dogs	–	0.0504	Best	–	0.0148	Happy	–	0.0153	Bar	–	0.0193
Clube	Club	0.0327	Pedra	Rock/stone	0.0454	Dominos	–	0.0146	Helpful	–	0.0147	Good	–	0.0192
Ferreira	Portuguese surname	0.0311	Infantil	Infantile	0.0403	Great	–	0.0141	Place	–	0.0144	Place	–	0.0173
Condições	Conditions	0.0311	Beautiful	–	0.0373	Hotel	–	0.0129	Apply	–	0.0141	Service	–	0.0167
Desportiva	Sports	0.0306	Garden	–	0.0367	Loja	Store	0.0124	Work	–	0.0139	Vespa	Wasp	0.0165
Court	–	0.0302	Animais	Animals	0.0363	Tour	–	0.0120	Leitão	Pork	0.0138	Sodré	Cais do Sodré neighborhood	0.0160
Amaral	Portuguese surname	0.0283	Size	–	0.0348	Nice	–	0.0118	Crossfit	–	0.0138	Pesca	Fishing	0.0154