



**HAL**  
open science

## **ProxMetrics: modular proxemic similarity toolkit to generate domain-adaptable indicators from social media**

Maxime Masson, Philippe Roose, Christian Sallaberry, Marie-Noelle Bessagnet, Annig Le Parc-Lacayrelle, Rodrigo Agerri

### ► To cite this version:

Maxime Masson, Philippe Roose, Christian Sallaberry, Marie-Noelle Bessagnet, Annig Le Parc-Lacayrelle, et al.. ProxMetrics: modular proxemic similarity toolkit to generate domain-adaptable indicators from social media. *Social Network Analysis and Mining*, 2024, 14 (1), pp.124. 10.1007/s13278-024-01282-1 . hal-04629357

**HAL Id: hal-04629357**

**<https://univ-pau.hal.science/hal-04629357>**

Submitted on 29 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# ProxMetrics: modular proxemic similarity toolkit to generate domain-adaptable indicators from social media

Maxime Masson<sup>1,2</sup> · Philippe Roose<sup>1</sup> · Christian Sallaberry<sup>1</sup> · Marie-Noelle Bessagnet<sup>1</sup> · Annig Le Parc Lacayrelle<sup>1</sup> · Rodrigo Agerri<sup>2</sup>

Received: 15 December 2023 / Revised: 18 May 2024 / Accepted: 12 June 2024  
© The Author(s) 2024

## Abstract

In this paper, we introduce *ProxMetrics*, a novel toolkit designed to evaluate similarity among social media entities through proxemic dimensions. Proxemics is the science that studies the organization of space and the effects of distances on behavior and interactions. It encompasses 5 core dimensions: Distance, Identity, Location, Movement, and Orientation. Adapting the principles of traditional physical proxemics to the digital world of social media, we present a method and a modular similarity function to determine proxemic similarity scores across heterogeneous social media entities (*users, groups, places, themes and times*) based on these dimensions. The approach used is intended to be modular and generic, ensuring adaptability across various application domains and requirements. The calculated scores act as indicators and offer valuable insights for stakeholders, aligning with distinct domain requirements. Empirical testing in the tourism domain highlights the toolkit's extensive applicability across a variety of requirements.

**Keywords** Proxemics · Similarity · Social Media · Computational Social Science · Multidimensional · Tourism

---

Philippe Roose, Christian Sallaberry, Marie-Noelle Bessagnet, Annig Le Parc Lacayrelle, Rodrigo Agerri have contributed equally to this work.

✉ Maxime Masson  
maxime.masson@univ-pau.fr

Philippe Roose  
philippe.roose@univ-pau.fr

Christian Sallaberry  
christian.sallaberry@univ-pau.fr

Marie-Noelle Bessagnet  
marie-noelle.bessagnet@univ-pau.fr

Annig Le Parc Lacayrelle  
annig.lacayrelle@univ-pau.fr

Rodrigo Agerri  
rodrigo.agerri@ehu.eus

<sup>1</sup> LIUPPA, E2S, University of Pau and Pays Adour, Pau, France

<sup>2</sup> HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia-San Sebastian, Spain

## 1 Introduction

In recent years, social media have transformed from simple digital platforms for informal conversations to massive networks that shape modern society (Akram and Kumar 2017). They now encapsulate a broad spectrum of daily life, ranging from communication and commerce to politics and entertainment, serving as a valuable data source providing insights into how people perceive and engage with the world.

The versatility of social media data makes it valuable for analyzing and gaining insights into a wide array of vastly different domains, including but not limited to tourism, public policy, or healthcare. For example, in the tourism sector, stakeholders are increasingly leveraging social media data for diverse purposes (Hvass and Munar 2012). These purposes range from analyzing frequently practiced leisure activities and their correlations with climatic or temporal factors, to understanding typical tourist routes and gauging levels of satisfaction. This data can assist in tasks such as the creation of tailored tour packages or the identification of tourist areas needing refinement. In the domain of public policy, social media data helps to contextualize public sentiment, enabling government agencies to make informed decisions (Charalabidis and Loukis 2012) to improve public

services, enhance land use planning or address societal issues. Lastly, in healthcare (Smailhodzic et al. 2016), social media provide a platform for patients to share their experiences and for healthcare providers to disseminate information and monitor health trends, ultimately contributing to a better understanding of medical interventions and patient outcomes.

Analysis of social media data often involves the calculation of indicators (Neiger et al. 2012), which are metrics that aim to measure or evaluate the state or level of a particular aspect of interest (e.g., the affluence associated with a given place, the level of friendship between users, etc.). Typically, indicators are defined based on the domain of application. Indicators are usually related to heterogeneous social media entities. These entities may be directly associated with the social media platform, such as individual users, groups of users, posts, etc., but they can also be informational entities extracted from the content or metadata of social media posts, including cities, themes, events, persons, organizations, etc. A prevalent category of indicators are *similarity measures*. These are quantitative indicators used to assess the degree of *resemblance* or *closeness* between several entities. For instance, in the context of social media, a similarity measure might compare user profiles based on shared features (e.g., similar age, same gender, nearby home location, same language) to suggest potential connections with other users or content recommendations (Mazhari et al. 2015).

Similarity measures are crucial in various domains for many applications, including content recommendation (Jiang and Yang 2017), targeted advertising (Zhang et al. 2019), understanding social dynamics, and event detection (Becker et al. 2010). In the context of content recommendation, similarity measures are essential because they allow the recommender system to identify and suggest content that is most relevant and appealing to the user. We will give two examples. (1) In collaborative filtering on a social media platform, user-based collaborative filtering calculates the similarity between users based on their engagement history, such as likes and shares. If *User A* and *User B* have liked many of the same posts, the recommender system infers that they could have similar tastes and recommends to *User A* the posts that *User B* has liked but *User A* has not yet seen. (2) Conversely, in item-based collaborative filtering, the similarity between content items is assessed, recommending posts to a user based on their engagement with similar posts (Zangerle and Bauer 2022). This helps in personalizing the content feed, keeping the user engaged by presenting them with content that aligns closely with their interests.

When interpreted correctly, they can provide insights into user connectivity and preferences, the nature of interactions, and common patterns found online.

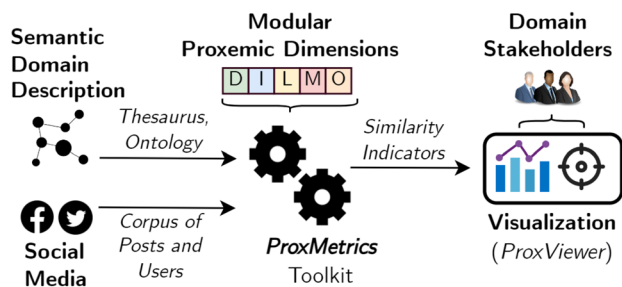
However, determining what *similarity* means in the context of social media can be challenging. Different platforms, users, and objectives lead to multifaceted interpretations of similarity (Anderson et al. 2012). Consequently, there is a pressing need for universal similarity measures capable of serving as meaningful indicators in an extensive array of domains, using data from social media. These measures should not be domain-specific but instead, offer versatility and modularity to adequately satisfy the diverse requirements of end users.

In response to these challenges, this paper introduces *ProxMetrics*, a modular and generic toolkit designed to assess the similarity between multidimensional social media entities (*that encompass users, groups, places, times, and themes*). Proxemics is the study of human use of space and the effects of distances on behavior, communication, and social interaction (Hall 1966). Traditionally applied to physical environments, we hypothesize that adapting the 5 dimensions of proxemics (DILMO), namely Distance (D), Identity (I), Location (L), Movement (M) and Orientation (O) (Greenberg et al. 2011) to the digital world of social media could provide modular and scalable dimensions to build personalized indicators to evaluate similarity between multidimensional social media entities in a domain-independent and versatile manner. We define *proxemic similarity* on social media as the perceived relational closeness or association between entities on social media (e.g., users, groups, places, periods, themes), based on the nature, frequency, and depth of their interactions or mentions within social media platforms. Various factors can be considered when evaluating proxemic similarity, including analysis of individual posts, user feeds, engagement levels, sentiment analysis, or even profile information. The *ProxMetrics* toolkit has been designed around the 5 dimensions of proxemics with both modularity and genericity in mind (see Fig. 1). It can be used with any social media platforms (*source genericity*) along any domain that can be described semantically (*thematic genericity*) without requiring extensive modifications or ad hoc solutions. Additionally, a user-friendly dashboard (*ProxViewer*) is provided to visualize the results; it is aimed at non-computer scientists, such as stakeholders in various domains.

This paper is structured as follows: We begin by reviewing existing work related to proxemics and similarity measures (Sect. 2). Subsequently, we introduce *ProxMetrics*, a modular toolkit designed for universal similarity measurement in social media, grounded in the theory of Proxemics and its foundational data model (Sects. 3 and 4). We then proceed to experiment it and evaluate its performance based on requirements in the tourism domain (Sect. 5). We conclude by discussing future prospects (Sect. 6).

**Table 1** The 5 Dimensions of Proxemics, as defined by Greenberg et al. (2011)

Dimension	Definition	Example
Distance (D)	The measure of separation between several entities (e.g., <i>persons, objects</i> )	Physical distance in meters ( <i>numerical</i> ). Whether two entities are in the same room or not ( <i>categorical</i> )
Identity (I)	A set of characteristics describing the individuality and the role of an entity	Age, height ( <i>numerical</i> ). Name, gender ( <i>categorical</i> )
Location (L)	A qualitative description of the space. Position of static (e.g., <i>furniture</i> ) and dynamic (e.g., <i>persons</i> ) entities	Euclidean coordinates: x, y, z ( <i>numerical</i> ). Room in which an entity is ( <i>categorical</i> )
Movement (M)	The change of location and orientation over time	Spatio-temporal sequence ( <i>numerical</i> ). Descriptors of speed such as fast, slow, rapid, or steady ( <i>categorical</i> ).
Orientation (O)	The direction in which an entity is facing	Bearing ( <i>numerical</i> ). Facing toward or away from something ( <i>categorical</i> )



**Fig. 1** Overview of the *ProxMetrics* toolkit

## 2 Related work

To design the *ProxMetric* toolkit, we conducted an evaluation of existing work in 3 key areas: (1) the use of proxemics in various contexts (Sect. 2.1), (2) the commonly used measures for assessing similarity (Sect. 2.2), and (3) the current approaches to combining measures in order to obtain a composite one (Sect. 2.3).

### 2.1 The proxemics theory

Before going further, it is important to define the concept of proxemics, as it is a core element of our project. Proxemics was introduced in the seminal work of the American anthropologist Edward T. Hall (Hall 1966). He defines proxemics as “the science that studies the organization of space and the effect of distance on interpersonal relations”. Hall studied physical distance and the way it affects and regulates interactions between people. He then went further and linked the concept of distance to *proxemic zones* (Hall 1966). There are 4 core proxemic zones: (1) the intimate zone (0 to 0.45m) which is mainly used for close physical contact, (2) the personal zone (0.45 to 1.2m) for interactions with very close people such as family or

friends, (3) the social zone (1.2 to 3.6m) for regular conversations with strangers, and finally (4) the public zone (more than 3.6m) which is used when speaking to an audience or gathering. It is crucial to note that cultural, social, and physical factors can affect the definition of proxemic zones. In 2006, Greenberg et al. extended Hall’s definition of proxemics to introduce the notion of proxemic dimensions (Greenberg et al. 2011) (also referred to as DILMO dimensions). They identified 5 dimensions that can be used to express proxemics (Distance, Identity, Location, Movement and Orientation, DILMO). These dimensions are presented in Table 1.

Proxemics can be studied at several levels: (1) the individual level (*how and why does an individual express specific traits and cognitive or affective states through their proxemic behavior?*) and (2) the group level (*how does the behavior of individuals affect the group?*) (McCall 2015). A crucial concept in proxemics is known as *centrality*. Proxemics requires the selection of a reference, central entity (e.g., a specific individual, an object, etc.) that will serve as the focal point for behavioral observations and analyses. For example, in an urban planning context, a central landmark building can be chosen as the reference entity in a city. Urban planners might study the behavior and movement patterns of individuals in and around this central entity to gain insight into how people use the space, how they interact with each other, and how other entities (like nearby businesses) relate to it.

When it comes to existing usage, (refer to Table 2) proxemics is used primarily to analyze physical interactions using tangible, physical metrics, such as for robot navigation (Rios-Martinez et al. 2015) and classroom behavior analysis (Castañer et al. 2013). The concept has also been extended to other applications like picture annotation (Yang et al. 2012) and assessing the impact of social distancing during COVID-19 (Mehta 2020). Recent works have tried adapting proxemics to virtual spaces but typically still relied on physical metrics, such as in video games or Virtual Reality (VR) worlds (Llobera et al. 2010). Recently, the concept of *digital*

**Table 2** Overview of Existing Research Works Using *Proxemics* for Practical Applications

Reference	Space	Metrics	Level	Appl. Domain	Use of <i>proxemics</i>
Llobera et al. (2010)	Cyber	Physical	Individual	Virtual Reality	<i>Detect people reaction in a VR world.</i>
Cristani et al. (2011)	Physical	Physical	Group	Psychology	<i>Social relation inference.</i>
Yang et al. (2012)	Physical	Physical	Individual	Psychology	<i>Picture annotation.</i>
Castañer et al. (2013)	Physical	Physical	Both	Education	<i>Study teachers' behaviors.</i>
Mueller et al. (2014)	Physical	Physical	Individual	Video Games	<i>Proxemic strategies for videogames.</i>
Hans and Hans (2015)	Physical	Physical	Individual	Psychology	<i>Analysis of non-verbal communication</i>
Rios-Martinez et al. (2015)	Physical	Physical	Individual	Robotics	<i>Socially-aware robot navigation</i>
Yeh et al. (2017)	Physical	Physical	Individual	Robotics	<i>Human-drone interactions.</i>
Pérez et al. (2021)	Physical	Physical	Both	Engineering	<i>Development of proxemic mobile apps.</i>
Luxey (2019)	Cyber	Other	Individual	Engineering	<i>Middleware configuration.</i>
Mehta (2020)	Physical	Physical	Both	Health	<i>Social distancing on human behavior.</i>
Williamson et al. (2021)	Cyber	Physical	Group	Education	<i>Group behavior in a virtual workshop.</i>
Medeiros et al. (2021)	Physical	Physical	Individual	Virtual Reality	<i>Safety when using a VR headset.</i>
Gunawan et al. (2021)	Cyber	Other	Individual	Digital Forensics	<i>Cybercrimes analysis.</i>

*proxemics* has emerged, focusing on non-physical interactions in virtual spaces, such as in the analysis of cybercrimes (Gunawan et al. 2021) or for middleware reconfiguration (Luxey 2019).

We hypothesize that by adapting proxemics to social media and redefining the 5 DILMO dimensions, we can establish a foundation for a generic and modular approach to measuring similarity in social media. This approach could then be used by domain stakeholders to easily build custom indicators to analyze behaviors on social media around their domain. Several factors motivate this choice:

- **Flexibility** (versatility) proxemics is versatile and can be adapted to various requirements. Its five dimensions are broad and can be used to model many use cases. From Table 2, it can be observed that proxemics has been extensively used to model user-oriented similarity measures for a wide range of objectives, including physical metrics like human-drone interactions, behavior analysis in classrooms, picture annotation, social distancing during the COVID-19 pandemic, and other metrics like middleware reconfiguration and cybercrime analysis. These diverse applications highlight the versatility of the proxemics theory in addressing various requirements.
- **Domain-Agnostic** it has no strong correlation to a specific domain; it is a very domain-unaware theory. Table 2 shows that existing work leveraging proxemics spans vastly different domains of application: Virtual Reality, Health, Video Games, Robotics, Education, Software Engineering, Digital Forensics, and more. This demonstrates the domain-agnostic nature of proxemics.
- **Fitness for Social Media** As shown previously, proxemics can be applied to spaces of different natures (e.g., physical space, VR space, social media space) with interacting enti-

ties (e.g., real people, video game characters, social media users). As social media platforms can be conceptualized as spaces where various entities interact and maintain distances between each other, it is possible to hypothesize that proxemics could be applied to social media and that many aspects of proxemics could be naturally linked with social media concepts. For example, *distance* can refer to the distance between social media users, posts, or entities contained in posts such as hashtags, place mentions, etc. Additionally, *location* can refer to the community in which a social media user is positioned in terms of their interests. *Orientation* might signify a sentimental orientation toward certain topics, and *identity* can characterize a social media user by attributes like age, language, etc.

- **Tangible Dimensions** Proxemic dimensions were originally designed around the physical world. They are practical and tangible, making them easier to understand and manipulate, even for non-computer scientist users. These dimensions can therefore serve as abstractions for more complex concepts. For example, the distance between individuals can be used to represent their level of interaction or social engagement. By leveraging these tangible dimensions, we aim to create similarity measures that simplify the understanding of more abstract or complex phenomena.

Let's now examine existing similarity measures that could be used to assess similarity between entities' proxemic dimensions.

## 2.2 Existing similarity measures

Social media platforms have become a focal point for the research and application of various algorithms (Anderson

**Table 3** Comparison of the Applicability of Existing Similarity Measures for Social Media Data

	Profiles' attributes		Posts' attributes					
	Numerical	Categorical	Coordinate	Toponym	Date	Themes	Sentiment	Engagement
<b>Example</b>	25 yo.	Male	43.47,-1.41	Paris	2023-10-09	Natural::Beach	Positive	128 likes
<b>Traditional</b>								
Euclidean (Johansson et al. 2013)	✓		*		*	*	✓	✓
Manhattan (Wang et al. 2016)	✓		*		*	*	✓	✓
Minkowski (Groenen et al. 1995)	✓		*		*	*	✓	✓
Haversine (Nguyen et al. 2017)			✓					
Mahalanobis (Leys et al. 2018)	✓				*	*	*	*
Jaccard (Zangerle et al. 2013)		✓ (set)		✓ (set)	* (set)	✓ (set)		
Pearson (Sponcil and Gitimu 2013)	✓				✓ (seq)			
Dice (Duarte et al. 1999)		* (set)			* (set)	✓ (set)	*	
Cosine (Lahitani et al. 2016)	✓				* (seq)			
Hamming (Bookstein et al. 2002)		*			* (seq)		✓ (seq)	
Levenshtein (Navarro 2001)		✓		✓ (seq)	✓ (seq)	✓ (seq)		
Chebyshev (Coghetto 2016)	✓	*			* (seq)	* (seq)		
Earth Mover (Rubner et al. 2000)	✓		* (seq)		* (seq)			*
Wu-Palmer (Wu and Palmer 1994)				*		✓		
<b>Series-based</b>								
DTW (Müller 2007)			* (seq)		✓ (seq)			
TraFoS (Varlamis et al. 2021)			✓ (seq)					
Hausdorff (Huttenlocher et al. 1993)				✓ (set)				
Frechet (Alt and Godau 1995)			✓ (seq)					
TRACCLUS (Jiashun 2012)			✓ (seq)					
LCSS (Bergroth et al. 2000)			✓ (seq)		✓ (seq)	✓ (seq)		
CED (Moreau et al. 2020)						✓ (seq)		
<b>Graph-based</b>								
Node Similarity (Tang et al. 2016)	✓	✓	*	*	*	*		
Random Walk (Xia et al. 2019)	*	*	*	*	*	*		
<b>Deep Learning</b>								
Word Embeddings (Liu et al. 2015)				✓		✓	*	
User Embeddings (Amir et al. 2016)	✓	✓						
Pretrained Models (Devlin et al. 2018)	*	*	*	*	*	*	*	*

et al. 2012), especially with regard to trend analysis (Bhor et al. 2018), content recommendation (Jiang and Yang 2017), event detection (Huang et al. 2021), and ad targeting (Knoll 2016). One of the key components for these tasks is the ability to assess similarity between entities, whether these items are words, posts, users, or media. We have chosen to divide

existing similarity measures into 4 core families (*traditional, series-based, graph-based and deep learning*).

Table 3 provides a side-by-side comparison of existing similarity measures aligned with the various types of social media data contained in the APs Trajectory Model (refer to Fig. 2 for details).

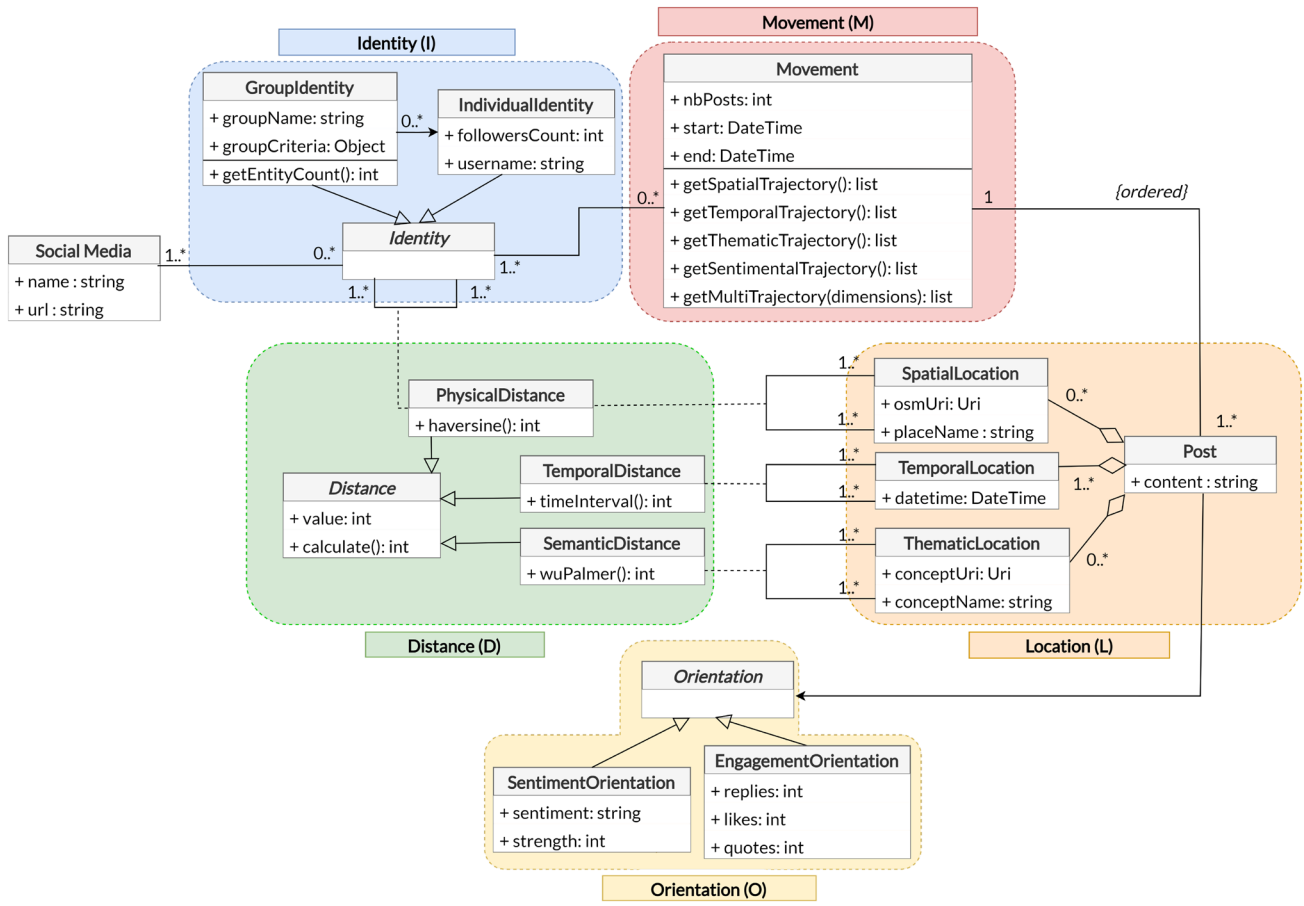


Fig. 2 APs Trajectory Model - Designed around the 5 dimensions of proxemics (DILMO)

We consider that a user’s attributes can be either categorical (e.g., gender, occupation, etc.) or numerical (e.g., age, height, etc.).

We have specified with the *sets* keyword when the similarity is calculated between unordered sets of entities (rather than individual ones), while the *seq* keyword indicates ordered sequences. The ✓ symbol indicates that the measure is fully compatible and fits the given data type, requiring little to no modification. The \* symbol indicates that the measure could be adapted but would need substantial redesign or modification. The absence of a symbol suggests that the measure is not fit or that adapting the measure would require extensive work.

The underlined measures are those we selected to adapt in the *ProxMetrics* toolkit. We hypothesized that we could rely on these measures as a foundation for the *ProxMetrics* toolkit and adapt them to assess similarity based on proxemic dimensions. These similarity measures are widely used, cover most of the data types we are dealing with, are less complex than graph or series-based ones, and do not necessitate time consuming or costly training dataset.

These include the Haversine (Nguyen et al. 2017; Baucom et al. 2013) and Euclidean (Johansson et al. 2013) distances to measure the straight-line distance between two coordinates, the first taking into account the curvature of the Earth, which is essential for dealing with spatial coordinates such as geocoded toponyms or post geotags. The Jaccard Index assesses the similarity of two sets of data, and is applicable for evaluating social media trends like hashtag usage (Zangerle et al. 2013). Lastly, the Wu-Palmer similarity determines the closeness of semantic concepts in a taxonomy, it calculates similarity by considering the elements’ depths in the taxonomy and the depth of their Least Common Subsumer (LCS), making it particularly useful for semantic similarity measures (Wu and Palmer 1994).

We will now examine how similarity metrics can be combined to create composite ones, which is crucial in our use case, given that we consider five distinct proxemic dimensions (DILMO).

### 2.3 Main techniques for criteria combination

The diversity of social media data often requires a nuanced approach to measuring similarity. A single similarity measure may not adequately address the multifaceted nature of the data. Therefore, the combination of various similarity measures can provide a more robust and comprehensive understanding. This section outlines strategies for combining multiple measures to improve indicators' accuracy in various domains.

In healthcare, criteria such as patient history, symptom severity, and laboratory test results are often combined. While weighted means are common, other methods like decision trees or Bayesian networks might be used, depending on the complexity and nature of the data. For example, Khan et al. (2008) used fuzzy decision trees to combine various biological indicators for disease prediction.

Financial institutions might use weighted means, logistic regression, or machine learning models to combine criteria like credit history, current debts, and income levels for risk assessment. The approach chosen often depends on the need for interpretability compared to predictive accuracy. For example, Bolton et al. (2010) explored the application of logistic regression in credit scoring models.

In education, combining criteria such as student engagement, performance metrics, and feedback can involve weighted means, cluster analysis, or even neural networks. The choice depends on the educational context and the specific goals of the analysis. For instance, Ng et al. (2016) employed cluster analysis to categorize student motivation and learning behaviors based on multiple metrics.

Lastly, environmental scientists often rely on spatial analysis techniques to combine criteria like pollutant levels, biodiversity indices, and land use patterns. The method chosen often reflects the scale and complexity of the environmental data. A study by Lu et al. (2015) leveraged spatial analysis to assess the impacts of various environmental factors on ecosystem health.

In the *ProxMetrics* toolkit, we plan to first experiment the weighted mean. This decision is driven by three main factors. Firstly, as demonstrated above, various domains employ distinctly different methods for combining indicators. Therefore, to ensure domain independence, it is crucial for us to adopt a combination technique that can be flexibly adapted to each domain's specific requirements and that is not too specialized. Secondly, we aim to integrate the expertise of domain stakeholders into the measure. Stakeholders should have the ability to adjust the impact of each dimension to tailor the indicators to their unique requirements. Thirdly, we require a combination technique that is easily interpretable by users who are not computer scientists. A weighted mean, in contrast to more complex combination methods such as machine learning models,

provides a clear and understandable rationale for the assigned weights of different measures.

We will now explain: (1) how we adapted traditional physical proxemics to social media data into a generic data model, and (2) how this model and the similarity measures we have selected are used in the *ProxMetrics* toolkit.

## 3 From physical to social media proxemics

The *ProxMetrics* toolkit rely on a common data model. This model adheres to the five dimensions of proxemics: *Distance*, *Identity*, *Location*, *Movement*, and *Orientation* (DILMO dimensions) (Greenberg et al. 2011). However, we have adapted these dimensions to model entities and interactions in social media rather than in physical spaces. The process behind this adaptation is detailed in Masson et al. (2023b).

The model is designed to be compatible with any type of social media data as long as it conforms to this model, thus ensuring its applicability across various domains and social media platforms. Figure 2 presents the UML class diagram of the model, which is instantiated step-by-step throughout the framework's pipeline.

The IDENTITY (I) dimension allows for the modeling of the studied population: individual users (along with their profile attributes) or demographics (user groups) featuring common characteristics or traits.

The MOVEMENT (M) dimension provides the ordered sequence of posts belonging to a given user. It gives a comprehensive view of a user's activities on the chosen social media and allows linking posts together to create multidimensional trajectories. Additionally, it can be broken down into several sub-trajectories (*spatial*, *thematic*, *spatio-thematic*, *tempo-sentimental*, etc.).

The LOCATION (L) dimension models the posts themselves along with their associated *locations*. We move away from the solely physical definition of *location* and consider 3 types of locations. These can be (1) spatial (*places based on toponyms extracted from posts or geotag metadata*), (2) temporal (*temporal entities or timestamps*) or (3) thematic (*domain concepts aligned with a semantic resource*). Thematic locations are defined according to the studied domain's description (domain specific ontology, thesaurus or dictionary). These semantic resources provide additional hierarchy information. When it comes to spatial locations, they are associated with a unique identifier linked to a spatial database. This allows for featuring relationships (e.g., *a city is within a region, itself within a country*). A given *post* can be in several *locations* at the same time, making the model *ubiquitous*.

The ORIENTATION (O) dimension contains contextual and enrichment data, such as sentiment of the associated post



(*positive, negative or neutral*) and the engagement associated with it (*based on the number of replies, likes and quotes*). The classes for both locations and orientations are designed to be extensible, thus providing the flexibility to incorporate new classes as desired (e.g., when studying politics, one could imagine a *political orientation* dimension).

Lastly, the DISTANCE (D) dimension helps in modeling and storing static distances between entities of the same type, specifically between (1) identities (*user or group*) and (2) locations (*spatial, temporal or thematic*). These static distances will be used to calculate proxemic similarity, we will detail this aspect later (see Sect. 4.6).

Let's now explore the *ProxMetrics* toolkit, which relies on this model and novel redefinition of proxemics, allowing for the calculation of similarity indicators between various types of social media entities.

## 4 ProxMetrics: modular toolkit to evaluate proxemic similarity in social media

We introduce *ProxMetrics*, a modular toolkit designed to assess the proxemic similarity between multidimensional entities on social media based on the 5 dimensions of proxemics (DILMO). This toolkit is designed to be fully modular and generic, adaptable to any social media platform and flexible to accommodate a wide variety of user requirements.

### 4.1 Entity definitions

We define *proxemic similarity* on social media as the perceived relational closeness or association between entities on social media, based on the nature, frequency, and depth of their interactions or mentions within social media platforms. Various factors can be considered when evaluating proxemic similarity, including analysis of individual posts, user feeds, engagement levels, sentiment analysis, or even profile information. In this work, we consider the following categories of social media entities. Firstly, we have dynamic entities: *users* and *groups*. They actively *interact* and *move* within the landscape of social media, analogous to people in physical proxemics.

- *Users* ( $u \equiv \text{👤}$ ). Individual social media users, whether physical (e.g., *a person*) or corporate, institutional ones (e.g., *an institution, a company*).
- *Groups* or *demographic s* ( $g \equiv \text{👥} = \{i_1, i_2, \dots, i_n\}$ ). Groups of social media users defined according to the domain of study. They can be based on shared traits (e.g., *French users, influencers, foreign tourists, etc.*).

Secondly, there are static entities: *places, dates* (or periods), and *themes*. These are extracted from users' posts and, unlike dynamic entities, do not interact on their own. Instead, they *appear* in the user posts. These static entities are analogous to objects in physical proxemics.

- *Places* or *spatial entities* ( $s \equiv \text{📍}$ ). Places mentioned on social media. Different levels of granularity are possible, such as points of interest, districts, cities, regions, countries, etc. Extracted from posts' metadata (*geotags*) or from the content of the posts.
- *Themes* or *thematic entities* ( $th \equiv \text{📖}$ ). Domain-specific thematic concepts mentioned on social media are assigned to a semantic resource (e.g., *dictionary, thesaurus, ontology*). Different levels of granularity, like levels within a thesaurus or ontology, are possible. Extracted from posts' content.
- *Dates/Periods* or *temporal entities* ( $t \equiv \text{📅}$ ). Dates or time periods associated with social media posts. Different levels of granularity are possible, including hour, day, week, month, year, season, day of the week, etc. Extracted from posts' timestamps (metadata) or content.

### 4.2 Proxemic similarity design

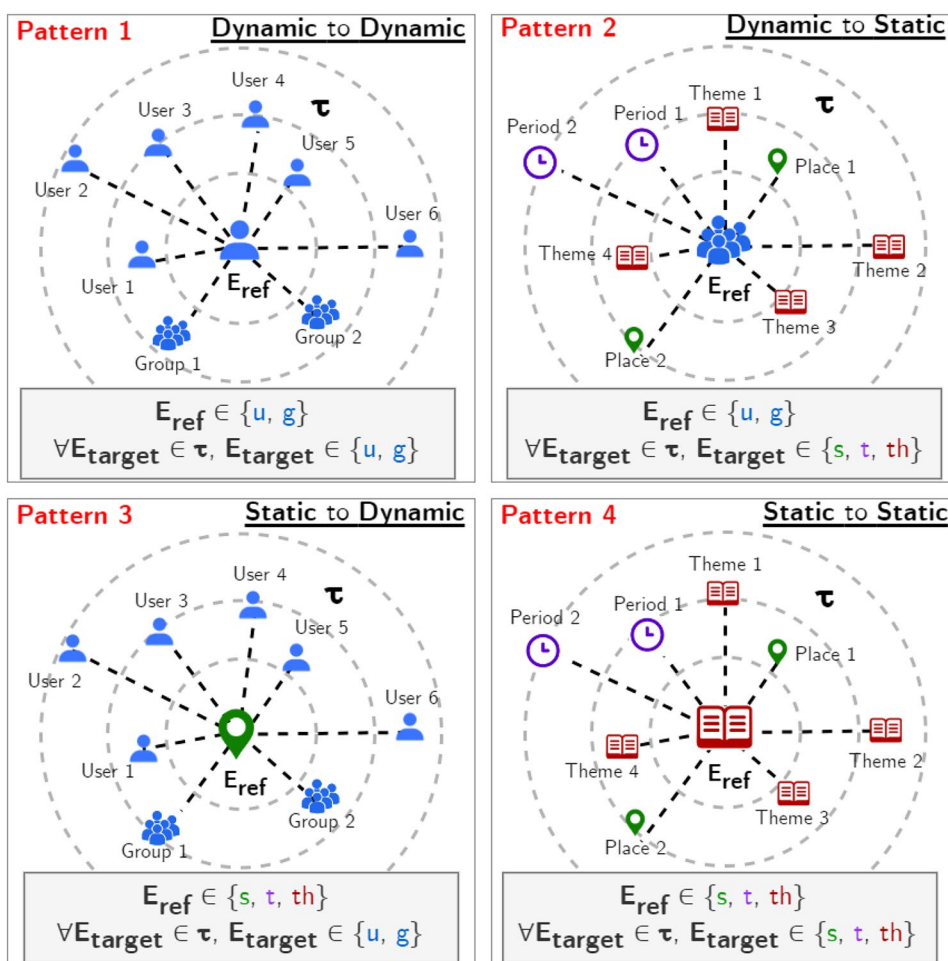
In proxemics, the selection of a **reference entity**, or the *center entity* ( $E_{ref}$ ), is essential. While the traditional physical context of proxemics mostly uses individuals or sensors as references (whose interactions are under observation), the landscape of social media provides a more diverse range of entities as potential references. This could be, for example, a specific user ( $u$ ), a group ( $g$ ), a place ( $s$ ), a theme ( $th$ ), or a date ( $t$ ).

$$E_{ref} \in \{u, g, s, th, t\} \quad (1)$$

Proxemic similarity ( $P_s$ ) is measured between a chosen reference entity  $E_{ref}$  and a set of target entities denoted as  $\tau$ . This relationship gives rise to a wide array of potential entity pairings (25 combinations), including *user to users, user to groups, place to users, place to places*, among others. To systematically categorize these pairings, we have identified 4 primary proxemic similarity patterns, as illustrated in Fig. 3.

In the center of the proxemic reticle is the reference entity  $E_{ref}$ . Surrounding it is a set of target entities  $\tau$ . We denote entities in this set as  $E_{target}$  (with  $E_{target} \in \tau$ ). The visual distance between these and the reference entity represents their relative proxemic similarity. This distinction in pattern is necessary because, unlike in physical proxemics, determining the proxemic similarity between dynamic entities (users or groups issuing posts), static entities (informational entities found in posts), or a combination of both cannot be done in the same manner.

**Fig. 3** The 4 proxemic similarity patterns in the *ProxMetrics* toolkit

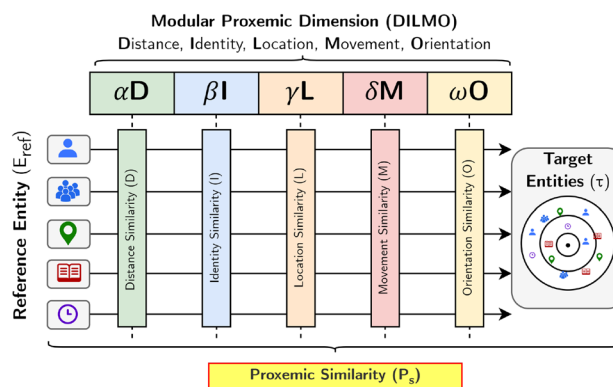


Depending on their level of proxemic similarity in regard to the reference entity  $E_{ref}$ , target entities  $\tau$  might be categorized into different *proxemic zones*. These zones are demarcated by the light gray dashed lines in Fig. 3. The number and range of proxemic zones are determined based on the study domain and by the end users. For example, in the tourism domain, they might represent the degree of attraction a tourist feels towards nearby POIs (Points of Interest). Unlike in traditional physical proxemics (with intimate, personal, social, and public zones), there are no universal definitions of zones when it comes to social media.

### 4.3 Proxemic similarity definition

The formula for proxemic similarity ( $P_s$ ) is grounded in the 5 proxemic dimensions (Distance, Identity, Location, Movement, Orientation), as detailed in Fig. 4. It is therefore composite and modular; the user can modulate (amplify or reduce) the impact of certain dimensions depending on the requirements he wishes to address. This is done through 5 coefficients (here  $\alpha, \beta, \gamma, \delta, \omega$ ).

Given a reference entity  $E_{ref}$  and a target entity  $E_{target}$ , we introduce a composite measure  $P_s(E_{ref}, E_{target})$ . This measure is an aggregation of 5 sub-measures, each corresponding to one of the proxemic dimensions (DILMO). These sub-measures quantify the similarity or dissimilarity between the



**Fig. 4** Overview of the *ProxMetrics* toolkit

2 entities across each dimension, providing a comprehensive assessment of their proxemic relationship.

Proxemic similarity ( $P_s$ ) is expressed as a numerical value ranging from 0 to 1. A value of 1 indicates a strong proxemic similarity, while a value of 0 points to a lack of similarity.

$$\begin{aligned}
 P_s(E_{ref}, E_{target}) &= \alpha D(E_{ref}, E_{target}) \\
 &+ \beta I(E_{ref}, E_{target}) + \gamma L(E_{ref}, E_{target}) \\
 &+ \delta M(E_{ref}, E_{target}) + \omega O(E_{ref}, E_{target}) \quad (2)
 \end{aligned}$$

with  $\alpha + \beta + \gamma + \delta + \omega = 1$   
and  $0 \leq P_s, D, I, L, M, O \leq 1$

We hypothesize that by blending and modulating these five dimensional measures, we can create domain-specific indicators. This approach aims to address business requirements in a manner that is *generic*, functioning across various social media platforms, *domain-independent*, applicable to different domains of application, and *versatile*, capable of accommodating a wide range of end user requirements. We will now explore how each component is calculated.

#### 4.4 Distance similarity (D)

As previously mentioned, we opted to align with physical proxemics for the Distance  $D(E_{ref}, E_{target})$ . For simplicity, we currently allow the distance dimension to be applied only when calculating proxemic similarity between two entities of the same kind (e.g., two users, two cities, two themes, two dates or time periods, etc.). Our approach allows us to leverage established and efficient similarity measures for this dimension.

For users and places, we leverage the Haversine distance (Nguyen et al. 2017; Baucom et al. 2013) (denoted as  $D_{physical}$ ), which measures the straight-line distance between two points while accounting for the Earth's curvature. This decision is grounded in the utility of physical metrics in social media contexts. For example, it helps determine whether two social media users are in close proximity or if two cities are in the same region. This physical metric could be useful in recommendation scenarios, where we want to recommend POIs (*Points of Interest*) to users that are physically close to them. In cases where real-time positioning of a user is not supported by the social media platform, we rely on their most recently recorded location.

For user groups, we calculate the centroid position of all members, providing an indication of the group's predominant location.

For themes, we evaluate their semantic similarity (denoted as  $D_{semantic}$ ) using the Wu-Palmer methodology (Wu and Palmer 1994), which determines similarity based on their least common subsumer (LCS). This method helps

ascertain whether two themes are semantically *close* (e.g., *Beach* with *Sea*) or *far* (e.g., *Beach* with *Museum*).

For dates, we evaluate the interval ( $D_{interval}$ ) in hours between them, and for periods, we reference the median date. This enables us to determine the temporal proximity of dates or periods, determining whether they occurred close or not.

Lastly, it is important to note that the Haversine distances and time intervals undergo normalization between 0 and 1 using min-max normalization. Normalization parameters need to be tweaked depending on the spatial area and time range covered by the social media corpus in use.

#### 4.5 Identity similarity (I)

The Identity dimension  $I(E_{ref}, E_{target})$  uses profile features to calculate the similarity. Social media users possess various profile features such as age, gender, occupation, and more. These features are crucial for understanding behavior by emphasizing the unique characteristics of individual profiles. The primary goal of this dimension is to bridge demographic differences and to detect similar groups or users. If we refer back to the patterns illustrated in Fig. 3, here are examples of how this dimension is applied:

- PATTERN 1  $E_{ref} = u, \forall E_{target} \in \tau, E_{target} = u$  Detection of users ( $u$ ) with similar profiles based on their features (e.g., user connection based on their age, language, gender, etc.).
- PATTERN 2  $E_{ref} = u, \forall E_{target} \in \tau, E_{target} = th$  Recommendation of places ( $p$ ) or themes ( $th$ ) to users ( $u$ ) based on places visited or themes mentioned by similar users (based on their profiles features).
- PATTERN 3  $E_{ref} = s, \forall E_{target} \in \tau, E_{target} = g$  Detection of which demographics ( $g$ ) primarily visit a given city ( $p$ ) or are active during a certain period of the day or year ( $t$ ).
- PATTERN 4  $E_{ref} = s, \forall E_{target} \in \tau, E_{target} = s$  Comparison of the demographics associated with themes ( $th$ ), places ( $p$ ), or time periods ( $t$ ). For example, two places might want to identify how similar their visitors are.

Let's begin by defining  $I$  between 2 individual users. Given 2 users  $E_{ref}$  and  $E_{target}$ , the component  $I_{individual}(E_{ref}, E_{target})$  is defined as:

$$\begin{aligned}
 I_{individual}(E_{ref}, E_{target}) &= \frac{\sum_{i=1}^n w_i \times s_{feature}(E_{ref}^i, E_{target}^i)}{\sum_{i=1}^n w_i} \quad (3) \\
 &\text{with } E_{ref} \in u \text{ and } E_{target} \in u
 \end{aligned}$$

- $n$  represents the number of attributes considered.

- $w_i$  represents the weight of the  $i^{th}$  attribute.
- $s_{feature}(E_{ref}^i, E_{target}^i)$  denotes the similarity between the  $i^{th}$  attributes of users  $E_{ref}$  and  $E_{target}$  respectively.

For numerical attributes (e.g., age, height), we use the normalized Manhattan distance (Wang et al. 2016):

$$s_{feature}(E_{ref}^i, E_{target}^i) = \frac{|E_{ref}^i - E_{target}^i|}{\max_{attr} - \min_{attr}} \tag{4}$$

- $\max_{attr}$  and  $\min_{attr}$  are the maximum and minimum possible values for the attribute respectively.

For categorical attributes (e.g., gender):

$$s_{feature}(E_{ref}^i, E_{target}^i) = \begin{cases} 0 & \text{if } E_{ref}^i = E_{target}^i \\ 1 & \text{if } E_{ref}^i \neq E_{target}^i \end{cases} \tag{5}$$

We can now extend this to groups of users, given two groups of users  $E_{ref}$  and  $E_{target}$ , the average similarity  $I_{group}(E_{ref}, E_{target})$  is:

$$I_{group}(E_{ref}, E_{target}) = \frac{1}{m \times p} \sum_{i=1}^m \sum_{j=1}^p I_{individual}(E_{ref}^i, E_{target}^j) \tag{6}$$

with  $E_{ref} \in g$  and  $E_{target} \in g$

- $m$  is the number of profiles in group  $E_{ref}$ .
- $p$  is the number of profiles in group  $E_{target}$ .
- $I_{individual}(E_{ref}^i, E_{target}^j)$  is the similarity between the  $i^{th}$  profile in  $E_{ref}$  and the  $j^{th}$  profile in  $E_{target}$  computed using the individual formula described before.

When one of the entities considered ( $E_{ref}$  or  $E_{target}$ ) is static (e.g. themes, places, or dates / periods), we assimilate it to the subset of users who have referenced it in their posts. This enables us to assess whether these entities are associated with similar demographics.

### 4.6 Location similarity (L)

The Location dimension  $L(E_{ref}, E_{target})$  operates differently based on the pattern of proxemic similarity in use (see Fig. 3). Specifically, it has three variations.

For PATTERN 1, which covers 4 proxemic similarity entity pairing (user to users, user to groups, group to users, and group to groups), we adapt the Jaccard Index (Zangerle et al. 2013) used to compute similarity between sets of

spatial, temporal, and thematic locations. We evaluate the cooccurrences of locations between  $E_{ref}$  and  $E_{target}$ . The more locations (spatial, temporal and thematic) two users or groups share, the more similar they are considered to be. For example, if two users frequently mention visiting the same cities or attending the same events, they are considered to have a high location similarity. Additionally, we want locations mentioned in recent posts to weight more (we consider that, as they are more recent, they are more relevant to domain stakeholders). Therefore, we introduce a time decay factor  $w_{freshness}$ , which is based on the freshness of posts (e.g., a post issued today will have a weight of 1, while a post issued  $x$  days ago will weight less).

$$w_{freshness}(p) = e^{-\lambda \times (currentDate - p.date)} \tag{7}$$

- $(currentDate - p.date)$  is the difference in days between the current date and the post's issuance date.
- $\lambda$  is a constant that controls the decay rate (how fast older locations are weighted less). We use  $\lambda = 0.01$  for a balanced effect.

The time-weighted locations similarity score between users or groups could be defined as:

$$L_{individual}(E_{ref}, E_{target}) = \sigma \times S_{spatial} + \theta \times S_{temporal} + \zeta \times S_{thematic} \quad \text{with } \sigma + \theta + \zeta = 1 \text{ and } E_{ref} \in \{u, g\} \text{ and } E_{target} \in \{u, g\} \tag{8}$$

- $\sigma, \theta, \zeta$  are coefficients used to modulate the strength of each type of locations (spatial, temporal, thematic).
- $S_{spatial}, S_{temporal}, S_{thematic}$  are individual locations similarity scores for the three types of locations, as defined below.

$$S_{type} = \frac{\sum_{l \in E_{ref}^{type} \cap E_{target}^{type}} w_{freshness}^{ref}(l,p) + w_{freshness}^{target}(l,p)}{\sum_{l \in E_{ref}^{type} \cup E_{target}^{type}} w_{freshness}^{ref}(l,p) + w_{freshness}^{target}(l,p)} \tag{9}$$

with  $type \in \{spatial, temporal, thematic\}$

- $type$  is type of locations considered (spatial, temporal or thematic).
- $E_{ref}^{type}$  is the set of locations mentioned by  $E_{ref}$  of type  $type$ . The same applies for  $E_{target}$ .
- $w_{freshness}^{ref}(l,p)$  represents the time decay factor  $w_{freshness}$  applied to the most recent post in  $E_{ref}^{type}$  where location  $l$  is mentioned. Similarly,  $w_{freshness}^{target}(l,p)$  applies to  $E_{target}$ .

This formula allows us to assess whether users or groups are similar based on: (1) where they interacted (spatial, where),

(2) when they were active (temporal, *when*), and (3) what they interacted about, their interests (thematic, *what*).

When applying PATTERN 2 and PATTERN 3 (e.g., user to themes, places, etc.), we use the ratio of the users’ posts containing the theme to his other posts. This allows us to detect the affinity of users or groups to specific themes, places, and time periods. We also weight occurrences with the time decay factor to give more weight to recent ones. We define  $L_{occurrences}(E_{ref}, E_{target})$  as:

$$L_{occurrences}(E_{ref}, E_{target}) = \frac{\sum_{p \in P_{E_{ref}|E_{target} \in p.locations}} w_{freshness}(p)}{\sum_{p \in P_{E_{ref}}} w_{freshness}(p)} \tag{10}$$

with  $E_{ref} \in \{u, g\}$  and  $E_{target} \in \{s, t, th\}$

- $P_{E_{ref}|E_{target} \in p.locations}$  represents the set of all posts from user or group  $E_{ref}$  that contain  $E_{target}$ .
- $\sum_{p \in P_{E_{ref}|E_{target} \in p.locations}} w_{freshness}(p)$  is the sum of the weights of all posts from user or group  $E_{ref}$  that contain  $E_{target}$ .
- $\sum_{p \in P_{E_{ref}}} w_{freshness}(p)$  is the sum of the weights of all posts from user or group  $E_{ref}$ .

Lastly, when dealing with PATTERN 4, which consists in determining the proxemic similarity between themes, places and dates or periods (e.g., place to themes, theme to periods, etc.), we consider the co-occurrences of spatial, temporal, and thematic locations found within posts or user feeds. For instance, consider the post: “We went swimming at the beach in Paris”. In this example, “swimming” (*thematic*), “beach” (*thematic*), and “Paris” (*spatial*) are co-occurring elements. For this, we can use a variant of the time-weighted locations similarity score described above (the time decay factor  $w_{freshness}$  is the same as defined before).

$$L_{cooccurrences}(E_{ref}, E_{target}) = \frac{\sum_{p \in E_{ref}^{posts} \cap E_{target}^{posts}} w_{freshness}(p)}{\sum_{p \in E_{ref}^{posts} \cup E_{target}^{posts}} w_{freshness}(p)} \tag{11}$$

with  $E_{ref} \in \{s, t, th\}$  and  $E_{target} \in \{s, t, th\}$

- $E_{ref}^{posts} \cap E_{target}^{posts}$  is the set of posts containing both  $E_{ref}$  and  $E_{target}$ .
- $E_{ref}^{posts} \cup E_{target}^{posts}$  is the set of posts that contain either  $E_{ref}$  or  $E_{target}$ .

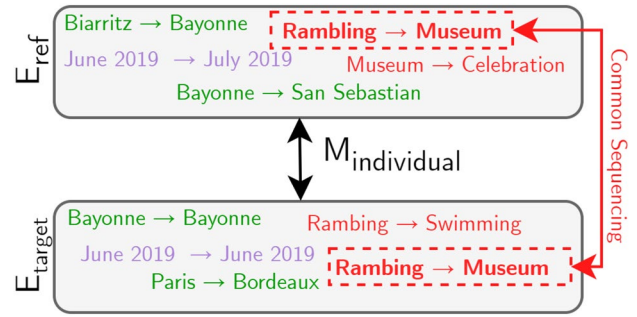


Fig. 5 Two sets of spatial, temporal, and thematic sequencings

### 4.7 Movement similarity (M)

The Movement dimension  $M(E_{ref}, E_{target})$  focuses on sequential relationships within multidimensional trajectories. A sequence is formed when two locations are mentioned consecutively on different posts. Using an example, if a post reads “We are *visiting* the museum in Paris”, following another that mentions “swimming”, sequences like *swimming*→*visiting*, *swimming*→*museum*, and *swimming*→*Paris* emerge. Here, for simplicity reasons, we limit ourselves to sequences of two contiguous locations, representing transitions from one place, theme, or time period to another.

For PATTERN 1 (such as user-to-users or group-to-users relations), we employ the same formula as location similarity ( $L_{individual}$ ). However, instead of considering occurrences of spatial, temporal, or thematic locations, we compare the sequencings of these elements across posts.

This approach ( $M_{individual}$ ) helps identify common travel patterns among tourists, for example. An illustrative example of the sequencings sets compared is presented in Fig. 5.

For PATTERN 2 and PATTERN 3, the goal is to determine if users consistently stay (or *focus*) in a particular location (e.g., a place, a theme, or a time period) in their posts. We achieve this by calculating the entropy associated with that location. In the following example,  $E_{ref}$  represents a user, and  $E_{target}$  a theme.

1. We start by counting the number of posts containing occurrences of the location  $E_{target}$  (denoted as  $C_{target}$ ) and all other locations in  $E_{ref}$ ’s sequence of posts (denoted as  $C_{others}$ ):
2. We calculate the total number of posts ( $C_{total}$ ):
3. We compute the entropy ( $M_{entropy}$ ) using the following formula (the lower this value is, the less predictable the user’s sequence of posts is in regard to the reference entity):

**Table 4** Implementation of the *ProxMetrics* toolkit. Each line represents a Proxemic Similarity Pattern from Fig. 3

	$E_{ref}$	$E_{target}$	D	I	L	M	O
1			$D_{physical}$	$I_{individual}$ $I_{group}$	$L_{individual}$	$M_{individual}$	$O_{individual}$
2			n/a	$I_{group}$	$L_{occurrences}$	$M_{entropy}$	$O_{occurrences}$
3			n/a	$I_{group}$	$L_{occurrences}$	$M_{entropy}$	$O_{occurrences}$
4			$D_{physical}$ $D_{semantic}$ $D_{interval}$	$I_{group}$	$L_{cooccurrences}$	$M_{sequencing}$	$O_{cooccurrences}$

$$= 1 - \sum_{T \in \{target, others\}} \left( \frac{C_T}{C_{total}} \right) \log_2 \left( \frac{C_T}{C_{total}} \right)$$

with

$$C_{target} = |\{p \in E_{ref} \mid E_{target} \in p.locations\}|$$

$$C_{others} = |\{p \in E_{ref} \mid E_{target} \notin p.locations\}|$$

$$C_{total} = C_{target} + C_{others}$$

$$E_{ref} \in \{u, g\} \text{ and } E_{target} \in \{s, t, th\}$$
(12)

In PATTERN 4, where we evaluate the relationships among themes, locations, or time periods (or any combination thereof), we use the conditional probability to determine how frequently one entity consistently appears after another. Given a sequencing set (see Fig. 5), the probability that an entity  $E_{target}$  follows  $E_{ref}$  is:

$$M_{sequencing}(E_{ref}, E_{target}) = \frac{N(E_{ref} \cap E_{target})}{N(E_{ref})}$$

with  $E_{ref} \in \{s, t, th\}$  and  $E_{target} \in \{s, t, th\}$

(13)

- $M_{sequencing}$  is the probability that  $E_{target}$  occurs after  $E_{ref}$ .
- $N(E_{ref} \cap E_{target})$  is the count of times both  $E_{ref}$  and  $E_{target}$  appear sequentially.
- $N(E_{ref})$  counts occurrences of  $E_{ref}$ .

This measure can help in determining social media users' subsequent destinations after visiting a particular city or what people tend to do after practicing a given activity (*theme*). It can also easily be reversed to determine what people do *before*.

### 4.8 Orientation similarity (O)

In the orientation, both contextual sentiment and engagement data are taken into account, and they are linked on a per-post basis. In our model, sentiment is categorized with labels: *positive*, *negative*, or *neutral*. Engagement is quantified by aggregating metrics such as the number of likes, reposts, comments, etc.

For every proxemic similarity pattern, our formula mirrors that of the location dimensions with one key modification: the weighting. Rather than applying the time-decay factor  $w_{freshness}$ , we introduce a new orientation factor,  $w_{orientation}$ , that adjusts post weights based on sentiment and engagement levels associated with it.

$$w_{orientation}(p) = \mu \times p.eng \times \nu \times p.sent$$

with  $e \in \mathbb{N}$  and  $s \in \{0, 1, 2\}$

(14)

$p.sent$  represents the sentiment value of the post  $p$ , which we map to the following values:  $\{0, 1, 2\}$ . Here, 0 corresponds to negative posts (*reducing the weight*), 1 to neutral posts (*no change in weight*), and 2 to positive posts (*increasing the weight*). This scaling, for instance, helps in identifying whether multiple users or groups favor the same content, or in emphasizing positive posts when establishing connections between themes or places (as positive posts will weight more than neutral ones and negative posts will not be considered).

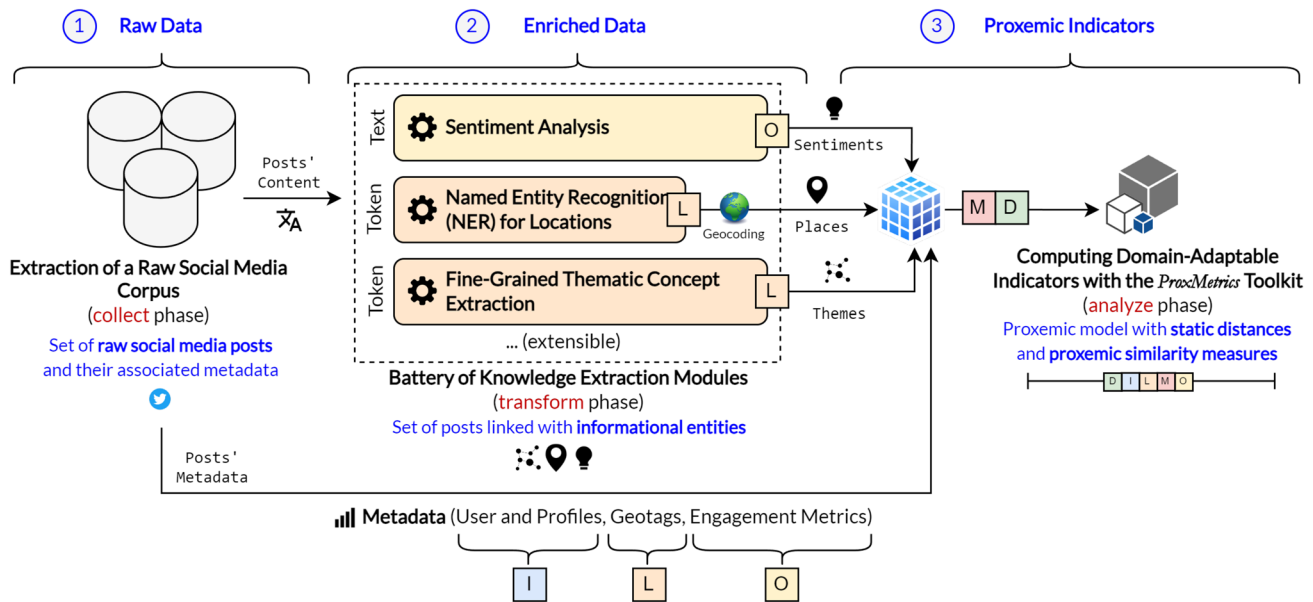
$p.eng$  stands for the engagement value of the post  $p$ . A post with higher engagement will carry greater weight. For example, this can help to discern whether various users gain popularity around similar themes.

Lastly,  $\mu$  and  $\nu$  are coefficients employed to prioritize either sentiment or engagement in the weighting calculation.

Formulas with  $w_{orientation}$  weighting are called  $O_{individual}$ ,  $O_{occurrences}$  and  $O_{cooccurrences}$ . It is important to note that the *ProxMetrics* toolkit is extensible. While we have chosen to implement it using these formulas, users are free to select others, provided they adhere to the proxemic patterns and the APs Trajectory Model. Table 4 offers a summary of the appropriate formulas for each similarity pattern. We will now experiment with the toolkit and manipulate the dimensions to address domain-specific requirements.

## 5 Experimentation: application to the tourism domain

Firstly (Sect. 5.1), we present the input data that we use for this experiment. This includes (1) the dataset (its collection and preprocessing for use with the *ProxMetrics* toolkit),



**Fig. 6** Overview of the Instantiation Process of the APs Trajectory Model

as well as (2) the tourism knowledge base, and (3) the requirements from stakeholders in the domain of tourism (specifically, tourism offices) that we gathered.

Secondly (Sect. 5.2), we demonstrate how these requirements can be expressed as proxemic similarity indicators through the *ProxMetrics* toolkit (e.g., choice of proxemic pattern, relevant dimensions). We then elaborate on a chosen request from tourism offices.

Thirdly (Sect. 5.3), we generalize from this and qualitatively evaluate the four defined proxemic patterns with various dimensions, as well as the methodology used to combine them, to assess the validity of our working hypotheses and the relevance of the indicators produced by the toolkit. Additionally, we conclude the experiment by comparing the toolkit with existing platforms dedicated to analyzing touristic data to highlight its complementarity in this domain.

## 5.1 Input data and enrichment

Let's begin by describing the input data and the enrichment process used in our experiments with the *ProxMetrics* toolkit. Fig. 6 illustrates the pipeline used to instantiate the APs Trajectory Model (note that the toolkit is applied to this model). We have also specified the dimensions of the model and the corresponding steps at which they are instantiated. The data collection and enrichment process was organized as follows:

1. We collected data from the popular social media platform *Twitter* (now X). Specifically, we gathered a corpus of 3,154 multilingual tweets in French, English, and Spanish that originated from the French Basque Coast during the summer of 2019 (see Fig. 6, ①). This area is known as one of the top tourist destinations in France. These tweets were posted by 655 unique users. Details on the data collection process can be found in Masson et al. (2022).
2. The data model was instantiated using two types of data: metadata and tweet content (see Fig. 6, ②). Metadata-based instantiation was a straightforward process, encompassing elements like engagement metrics, profile features, timestamps, and geotags, among others. The textual content of the tweets was processed using three deep learning-based NLP modules to generate automatic annotations at the token level for (1) locations and (2) thematic concepts, and at the text level for (3) sentiment. We divided the dataset into three parts with a split of 60% for training, 20% for validation, and 20% for testing, ensuring a uniform distribution of languages and users across these splits. More precisely:
  - **For Tourism-related Concepts:** we used the *The-saurus on Tourism & Leisure Activities* provided by the World Tourism Organization (2002) to define them. This extensive resource covers approximately 1,300 touristic concepts. These concepts were linked to tweets through prompt-based few-shot learning

with EntLM (Ma et al. 2022) paired with a multilingual BERT (Devlin et al. 2018) language model. The hyperparameters used are those recommended in the EntLM repository,<sup>1</sup> namely a batch size of 4, a learning rate of 1e-4, and a weight decay of 0. We found that prompt-based few-shot learning with EntLM was significantly more effective than fine-tuning approaches, achieving an F1-score of 0.840, compared to 0.241 for fine-tuning for this task.

- **For Spatial Locations:** we extracted them by geotagging named entities from toponyms aligned with *OpenStreetMap* and from the geotags of posts. Here, fine-tuning multilingual BERT (Devlin et al. 2018) demonstrated superior performance with an F1-score of 0.848, compared to 0.788 with the few-shot technique for sequence labeling implemented by EntLM (Ma et al. 2022). We used the OSM Nominatim API (Clemens 2015) to geocode extracted toponyms.
- **For Sentiments:** they were derived using the XLM-RoBERTa language model (Conneau et al. 2019). More precisely, we used the version which was fine-tuned for sentiment analysis (Barbieri et al. 2022). It achieved a very high accuracy of 0.939.
- **Users** in the dataset were manually categorized into three groups (local tourists, external tourists, and photographers). This technique is not efficient, but the objective is to use this annotated dataset as a backbone to train deep learning-based classifiers to automate the task, as there were no existing annotated resources for this purpose publicly available at the time of our study.

More details on these NLP-related aspects are given in Masson et al. (2023a). In the end, this experimentation dataset of 3,154 tweets covers around 315 unique concepts (which were instantiated 7,022 times) and 7,128 locations. 1,357 tweets were positive, 366 negative and 1,238 neutral. The dataset has then been converted to our proxemic model (from Fig. 2) and loaded into the *ProxMetrics* toolkit (see Fig. 6, ©).

Below is a set of simplified requirements for indicators in the tourism domain, compiled through collaborative discussions with local tourism authorities and stakeholders on the French Basque Coast. Stakeholders are looking for indicators on:

1. *Leisure Activities Practiced Together* (which activities do tourists often practice together)

2. *Cities Visited in Sequence* (after visiting a given city, where do tourists usually go)
3. *City-Specific Demographics* (types of tourists per city)
4. *Weather-Based City Preferences* (choice of city by tourists based on the weather)
5. *Seasonal Activity Preferences* (choice of leisure activity by tourists based on the season)
6. *Identification of Popular Events* (which events are popular in the region)
7. *Lacking Tourism Infrastructure* (places where infrastructure is found disappointing)
8. *Tourist Satisfaction about POIs* (what POIs do tourists primarily enjoy)
9. *Trends in Cross-Border Tourism* (who, where, when, and what)
10. *Connection of Similar Tourists* (for a tourists connection system)

These requirements vary in scope, with some being broad and others more specific. Tourism stakeholders need indicators to better understand these diverse aspects of tourism in their region. To address this, we will take advantage of the *ProxMetrics* toolkit to address these requirements and calculate relevant proxemic similarity indicators, allowing for a deeper understanding of the various aspects of tourism in the region.

## 5.2 Toolkit experiment on tourism

Table 5 presents the requirements from tourism offices as shown above. Each line number corresponds to a requirement from the list in Sect. 5.1. For each requirement, we have proposed a manner to express it as a proxemic similarity indicator using the *ProxMetrics* toolkit. This includes: (1) the proxemic environment that could be used to model the requirement, identifying both the reference entity ( $E_{ref}$ ) and the target entities ( $\tau$ ), whose proxemic similarities to the reference element will be assessed; and (2) the relevant proxemic dimensions of our model that are essential for addressing it. While a variety of dimension combinations and proxemic environments (e.g., reference and target entities) may be applicable for most requirements, we present only a selected one here due to space constraints.

### 5.2.1 Designing tourism indicators with *ProxMetrics*

Let's examine the requirement "City-Specific Demographics" (requirement 3 in Table 5). In this context, the goal is to calculate indicators allowing tourism stakeholders to identify the types of tourists most commonly associated with various cities. Using the *ProxMetrics* toolkit, this requirement can be modeled by selecting a specific user group (e.g., *short-stay*

<sup>1</sup> <https://github.com/rtmaww/EntLM>.



**Table 5** Example of end users requirements for social media analysis: the case of tourism

Requirement	Proxemic Environment		Dimensions				
	Reference ( $E_{ref}$ )	Targets ( $\tau$ )	D	I	L	M	O
1	Leisure Activity	Leisure Activities			•		•
2	City <i>or</i> POI	Cities, POIs	•			•	
3	User Group	Cities		•	•		•
4	Weather	Cities			•		
5	Leisure Activity	Seasons	•		•		
6	Season <i>or</i> City	Events	•		•		•
7	City	Infrastructures			•		
8	City	Points of Interest			•		•
9	FrontierArea	Groups,  Cities,  Periods		•	•	•	•
10	User	Users	•	•	•	•	•

tourists) as the proxemic reference and then determining the proxemic similarity of different cities to this group. Therefore, it falls into the proxemic PATTERN 2 (refer to Table 4), which links a dynamic reference entity (in this case, a user group) to a static entity (in this case, a place).

To calculate this similarity, it may be useful to take into account multiple proxemic dimensions to achieve a more comprehensive analysis, such as the characteristics of the reference user group and of users visiting cities (*Identity dimension*), the frequency with which members of the reference group mention the city in their posts (*Location dimension*), and the sentiments they generally express towards it (*Orientation dimension*). For example, a given group of tourists and city could be perceived *similar* if positive sentiments are frequently expressed by the group about the city. This requirement could be articulated in various alternative ways (e.g., by choosing a specific city as the reference point instead of a user group). As the toolkit is modular, the end user is able to give more importance to certain dimensions by adjusting the dimensional weighting. In summary, Table 5 shows that the *ProxMetrics* toolkit is versatile and capable of modeling various domain-specific requirements.

### 5.2.2 Indicator case study: City-specific demographics

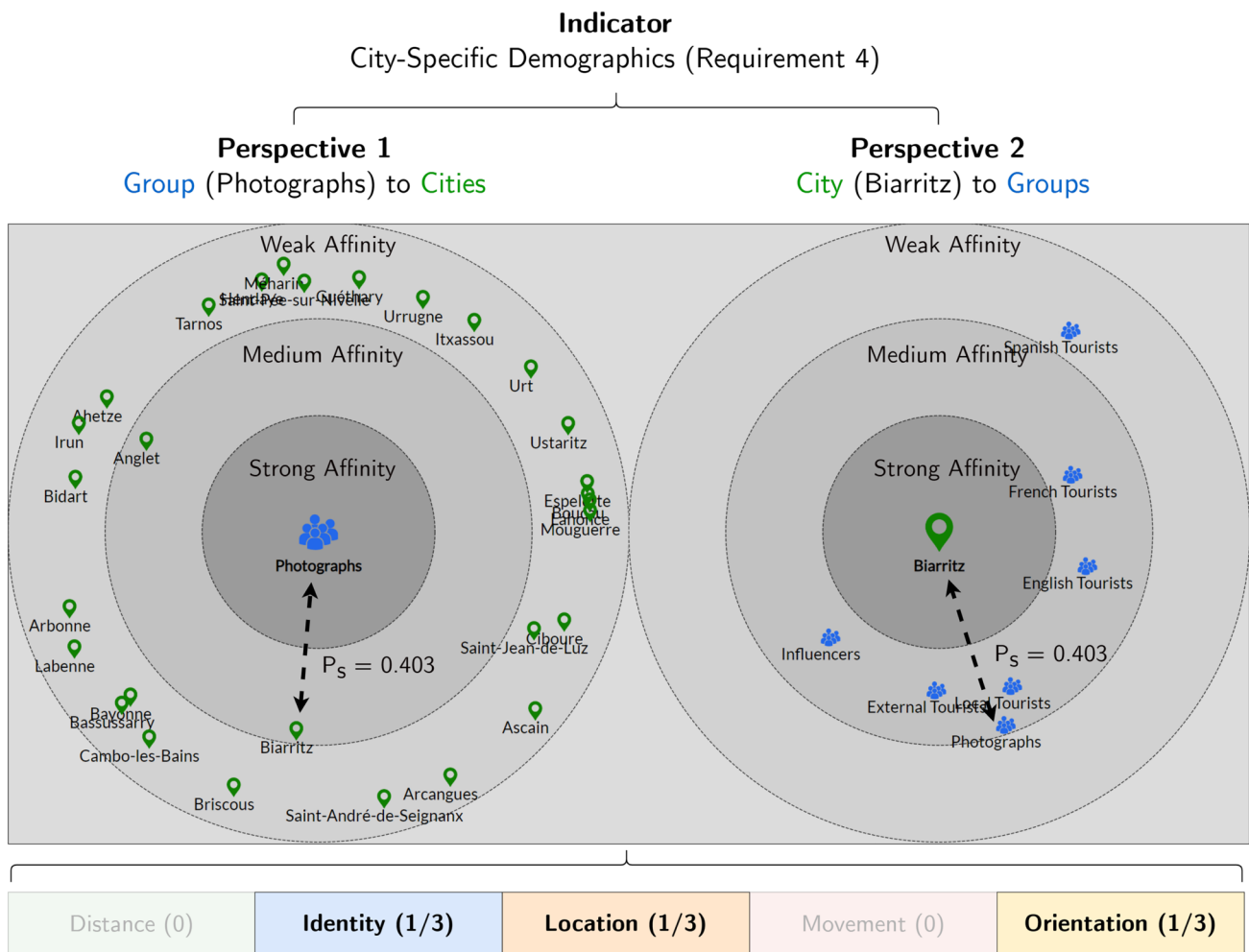
We will now go into more details with the example requirement presented above. Figure 7 focuses on the indicator “City-Specific Demographics”. Visuals were created using the *ProxViewer* dashboard, a web interface powered by the *ProxMetrics* toolkit, able to visualize results through proxemic reticles. This interface also helps end users setup the toolkit (dimensions, entities, etc.). Due to space constraints, we are unable to thoroughly present this dashboard. However, we have prepared a demonstration video. It is available here.<sup>2</sup>

In this example, the “City-Specific Demographics” indicator is expressed using two distinct proxemic environments (*Perspective 1* and *2* in Fig. 7). One where the demographic (user group entity) is the reference (here, *Photographers*), and similar cities are positioned relative to it, and another perspective where a chosen city of interest (here, the touristic city of *Biarritz*) is at the center, and we want to identify groups similar to this city. These 2 perspectives correspond to the proxemic PATTERN 2 and PATTERN 3 defined earlier (see Fig. 3). By default, we set proxemic dimensions (ILO) to equal weighting ( $\frac{1}{3} \approx 0.33$ ), but this default weighting can be dynamically changed by end users if they wish to give more weight to profile features or city mentions.

- The IDENTITY (I) dimension influences the result based on whether the typical user profiles of a city are similar to the group based on profile features (e.g., are the tourists of the city of Biarritz usually of the same age compared to photograph tourists, etc.).
- The LOCATION (L) dimension is based on how often members of the user groups mention the given city in their posts, which may indicate a specific affinity for the city.
- The ORIENTATION (O) dimension is considered and weights city mentions by sentiment and engagement values, which means that influential or positive users will have more impact on the similarity results.

Let’s illustrate *Perspective 1* from Fig. 7 using the pairing  $E_{ref} = Photographers$  and  $E_{target} = Biarritz$ . We can refer back to Table 4 to determine the formula for this proxemic environment pattern, specifically PATTERN 2. We calculate the proxemic similarity between Biarritz and Photographers as follows, using the  $I_{group}$ ,  $L_{occurrences}$ , and  $O_{occurrences}$  formula corresponding to the Identity, Location, and Orientation (ILO) dimensions considered in this example. We obtain a proxemic similarity of

<sup>2</sup> <https://youtu.be/x714RKvo9Cg>.



**Fig. 7** Visualization of the indicator “City-Specific Demographics” in proxemic reticles, using proxemic patterns 2 and 3

around 0.403 by aggregating the I, L and O dimensions with equal weighting. It appears that users classified as *Photographers* have a moderate interest in this city when considering these 3 dimensions.

$$\begin{aligned}
 &P_s(\text{Photographers}, \text{Biarritz}) \\
 = &\frac{1}{3} \times \underbrace{I_{\text{group}}}_{0.52} + \frac{1}{3} \times \underbrace{L_{\text{occurrences}}}_{0.34} + \frac{1}{3} \times \underbrace{O_{\text{occurrences}}}_{0.36} \\
 &\approx 0.4026
 \end{aligned}$$

In Fig. 7, results are displayed through a proxemic reticle with the reference entity in the center and the target entities around it, scattered in different proxemic zones. Here, we have 3 zones: strong affinity, medium affinity, and weak affinity. However, other zones could be defined according to the domain and the requirement of interest. Any number

of zones is possible. As we can observe, these visualizations are useful and flexible for domain stakeholders as they allow immediate identification of similar or dissimilar entities.

To further elaborate and demonstrate the *ProxMetrics* toolkit’s capability to calculate a wide array of indicators, we will provide an overview of four additional selected indicators.

### 5.2.3 Overview of additional indicators

Figure 8 presents example results for four additional indicators from those introduced in Sect. 5.1.

These four indicators were chosen to demonstrate the capabilities of the *ProxMetrics* toolkit, as they leverage various types of entity combinations and proxemic

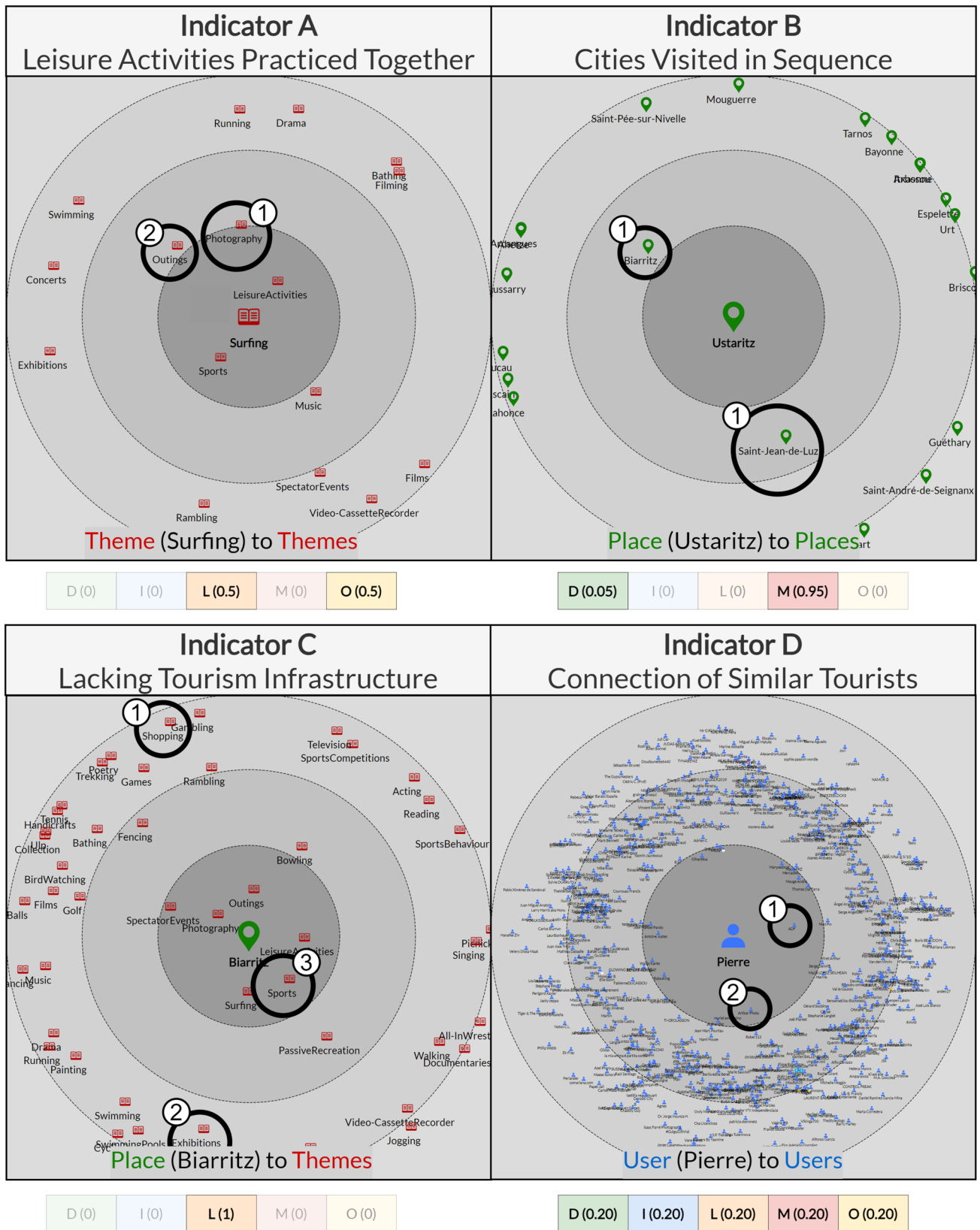


Fig. 8 Results of Four Selected Proxemic Similarity Indicators from the Tourism Office Requirements

dimensions. The proxemic similarity indicators displayed are as follows:

- *Leisure Activities Practiced Together* (Fig. 8, Indicator A): for this indicator, we are examining similar touristic activity-related themes compared to a reference one, in this example, *Surfing*. We use two equally weighted dimensions: *Location* (L), to consider the co-occurrences of themes in posts, and *Orientation* (O), so that positive co-occurrences weigh more heavily, as we consider those to be more significant. The results show that *Surfing* is often paired with *Photography* (see ①) and *Outings* (see ②). Perhaps, it appears that photographing surfers is particularly popular in the region.
- *Cities Visited in Sequence* (Fig. 8, Indicator B): here, we are interested in the spatial movements of tourists, namely, where they tend to go after visiting a given place. In this case, we are focusing on the city of *Ustaritz*, positioned as the reference entity, and it will be compared to other cities in the region. We use the *Movement* (M) dimension with stronger weighting because we are particularly interested in the places that are often sequenced in tourists' trajectories. Minor weighting is given to the *Distance* (D) dimension to slightly boost places that are physically close, as they are more likely to attract tourists. The results show that the cities of *Biarritz* (see ①) and *Saint-Jean-De-Luz* (see ②) are more likely to attract tourists after having visited *Ustaritz*. An extension of this indicator could be to calculate it for specific user categories, for example, *Photographers*. This would allow a recommender system to determine where to recommend another photographer to go based on the behavior of other photographers (via a collaborative filtering approach (Liu et al. 2014)).
- *Lacking Tourism Infrastructure* (Fig. 8, Indicator C): for this requirement, the objective is to identify themes related to infrastructure that are broadly considered lacking in a given city. We use the city of *Biarritz* as a reference and observe similar themes based on the *Orientation* dimension. This dimension will bring positive themes closer to the reference city and push negative ones further. The results show that *Shopping* (see ①) and *Exhibitions* (see ②) are quite dissimilar, indicating that these aspects are severely lacking in this city. On the other hand, *Sports* (see ③) is very similar, suggesting that the city is viewed very positively in this aspect by tourists.
- *Connection of Similar Tourists* (Fig. 8, Indicator D): Finally, the last requirement is to detect similar tourists to build a system that connects them. If we select a reference tourist, here *Pierre*, we can observe close matches. This indicator is based on all dimensions because we want to get the overall similarity based on various criteria. This

can be tweaked depending on the use case, as the toolkit is modular. ① and ② show examples of other tourists that are closer to the reference one and could therefore be recommended because they have similar behavior and likely share interests with the reference user. This indicator could also be used for a recommender system to suggest themes or places based on what similar users liked.

We have demonstrated that *ProxMetrics* effectively models a wide range of diverse indicators in the tourism domain by combining various entities and modular proxemic dimensions according to the requirements of domain stakeholders. However, it is now essential to evaluate the relevance and significance of the produced proxemic similarity indicators for the stakeholders.

### 5.3 Qualitative evaluation of proxemic similarity patterns

To determine whether the *ProxMetrics* toolkit produces indicators that are accurately representative of real-world phenomena or behaviors on social media, we conducted a qualitative evaluation of the results and compared the results obtained by *ProxMetrics* with the assessments of domain experts (colleagues specialized on cultural heritage and tourism practices), as depicted in Table 6. For each identified proxemic similarity pattern (from Fig. 3), we selected an indicator within this pattern as a case study. As mentioned earlier, the proxemic similarity is typically calculated between a reference entity and multiple target entities. However, for this evaluation, we focused on calculating the similarity between the reference entity and a single target entity to facilitate the work of the experts.

We implemented the following protocol. Firstly, for each of the four proxemic patterns, we asked five domain experts to assess the similarity (on a scale from 1 to 10, with 1 being extremely dissimilar and 10 being extremely similar) of the reference and the selected target entity for each dimension individually. To do this, experts were provided with only excerpts of the data model corresponding to the given dimension. For example, in the case of PATTERN 1, when evaluating the proxemic similarity between two users within the identity dimension, two sets of user profiles along with their characteristics (age, number of followers, etc.) were given but no information on the tweets they posted, places visited, etc. When dealing with the location dimension, they were only given two sets of tweets, but not the sequence in which they were issued or the users to whom they belong, etc. This allows us to determine whether our individual dimension formulas are relevant and meaningful to domain stakeholders. Then, we asked the domain experts to assess

**Table 6** Qualitative Evaluation of the *ProxMetrics* toolkit on the 4 proxemic similarity patterns (see Fig. 3) with selected indicators

Pattern 1 - Dynamic to Dynamic										Pattern 2 - Dynamic to Static									
Indicator	Connection of Similar Tourists									Indicator	City-Specific Demographics								
Prox. Environment	User (Dominique) to User (Luco)									Prox. Environment	Group (Photographers) to Place (St-Jean-De-Luz)								
	Evaluators				$\sigma$	$\bar{x}$	<i>ProxMetrics</i>		$\Delta$		Evaluators				$\sigma$	$\bar{x}$	<i>ProxMetrics</i>		$\Delta$
Distance	5	8	8	5	8	1,47	6,80	8,60	1,80		N/A								
Identity	3	3	3	2	4	0,63	3,00	6,50	3,50		6	6	6	6	5	0,40	5,80	4,80	1,00
Location	2	5	2	4	2	1,26	3,00	1,70	1,30		3	2	3	4	3	0,63	3,00	2,60	0,40
Movement	2	5	3	2	2	1,17	2,80	1,70	1,10		3	3	2	3	2	0,49	2,60	1,50	1,10
Orientation	1	5	2	2	3	1,36	2,60	2,80	0,20		3	6	5	8	3	1,90	5,00	1,40	3,60
Combination	DILMO									Combination	ILO								
	3	5	2	2	4	1,17	3,20	4,26	1,06		3	4	3	6	3	1,17	3,80	2,93	0,87
Pattern 3 - Static to Dynamic										Pattern 4 - Static to Static									
Indicator	Trends in Cross-Border Tourism									Indicator	Leisure Activities Practiced Together								
Prox. Environment	Theme (FrontierArea) to User (Daniel)									Prox. Environment	Theme (Surfing) to Theme (Photography)								
	Evaluators				$\sigma$	$\bar{x}$	<i>ProxMetrics</i>		$\Delta$		Evaluators				$\sigma$	$\bar{x}$	<i>ProxMetrics</i>		$\Delta$
Distance	N/A									Distance	1	3	3	3	1	0,98	2,20	1,00	1,20
Identity	6	8	5	4	4	1,50	5,40	3,50	1,90		7	7	7	2	5	1,96	5,60	3,90	1,70
Location	5	5	5	7	5	0,80	5,40	5,00	0,40		2	3	2	3	2	0,49	2,40	1,10	1,30
Movement	5	6	5	8	5	1,17	5,80	5,00	0,80		2	3	2	2	1	0,63	2,00	0,20	1,80
Orientation	2	7	5	6	7	1,85	5,40	7,80	2,40		3	3	2	3	2	0,49	2,60	0,80	1,80
Combination	ILMO									Combination	LO								
	4	6	5	2	5	1,36	4,40	5,33	0,93		3	4	3	7	2	1,72	3,80	0,95	2,85

the similarity of all relevant dimensions combined in regard to the example indicator chosen. This helps us to determine whether our method of combining dimensions (weighted mean) is relevant. For this evaluation, we deliberately selected entities that are not overly represented in the entire dataset, aiming to facilitate the experts' evaluation process.

As seen in Table 6, for each dimension and pattern, we calculated the standard deviation (depicted as  $\sigma$ ) between the five experts' measures (1, 2, 3, 4, and 5) in order to assess their degree of agreement. Then, we averaged the measures provided by the experts (depicted as  $\bar{x}$ ) and compared them against the results obtained by *ProxMetrics*. This allowed us to calculate the difference (depicted as  $\Delta$ ) between the experts' results and those of *ProxMetrics*, thus determining whether they are in concordance or not.

As we can observe in Table 6, for the individual dimensional evaluation, out of the 18 evaluation cases, there are only 3 cases (□ in Table 6) where the *ProxMetrics* assessment significantly differs from that of the evaluators (e.g., the  $\Delta$  is greater than 2). These are PATTERN 1 with the I dimension ( $\Delta = 3.50$ ), PATTERN 2 with the O dimension ( $\Delta = 3.60$ ), and PATTERN 3 with the O dimension ( $\Delta = 2.40$ ). In cases of poor assessment for the O dimension, we notice they are also correlated with very low agreement between evaluators ( $\sigma = 1.90$  for PATTERN 2 and  $\sigma = 1.85$  for PATTERN 3), indicating there are various ways to interpret the similarity for this dimension. Therefore, other formulas might be more appropriate depending on the

domain requirements. For the other dimensional measures, *ProxMetrics* provides assessments that are quite similar to those of the experts, demonstrating the relevance of our formula in the domain of tourism.

In the case of multidimensional combinations (DILMO, ILO, ILMO, and LO in the examples chosen), for most of them (3 out of 4, □ in Table 6), the aggregation with default parameters (equal weighting) of *ProxMetrics* closely matches that of the evaluators, with  $\Delta = 1.06$ ,  $\Delta = 0.87$ , and  $\Delta = 0.93$ . However, for PATTERN 4, the result diverges with  $\Delta = 2.85$  (□ in Table 6), and the agreement among evaluators is also significantly weaker ( $\sigma = 1.72$ ) compared to the other patterns. In all cases, it would have been possible to improve the accuracy of the results by altering the weighting of each dimension. Let's consider PATTERN 1 as an example: if we had doubled the impact of the L and O dimensions and reduced that of D by half, we would have obtained a result of 3.14, similar to that of the experts. This highlights the importance of choosing appropriate weights during the process to get accurate results. These weight values are highly dependent on the domain and specific requirements; therefore, it is necessary for experts in each domain to tweak them.

Let's conclude this experiment by comparing the *ProxMetrics* toolkit applied to the domain of tourism with indicators produced by other platforms in this domain. *ProxMetrics* is highly complementary for several reasons:

- *Dynamic, User-Parameterized Indicators*: It is a dynamic and modular tool that allows users to build their own indicators in real time using proxemic dimensions (DILMO). This contrasts with static dashboards such as Pilat Tourisme (2022), Isère Attractivité (2023), or Atout France (2023), where users have no say about what is presented to them.
- *Distinctive Analytical Insights*: ProxMetrics introduces a type of indicator (*proxemic similarity*) rarely seen in existing solutions. This allows tourism stakeholders to gain new analytical perspectives by assessing the similarity of domain-specific entities (*tourist activities, points of interest, types of tourist, etc.*). Additionally, usual tourism dashboards like the one from Visit Paris Region (2023) usually allow combined filtering but are limited in terms of blending multidimensional indicators together.
- *New Visualization (Proxemic Reticles)*: This offers a mode of visualization that contrasts with the classic visualizations used in tourism dashboards such as UNWTO (2023), INSEE (2023), or OECD (2023), which contain spatial maps, timelines, tables, and numerical charts but do not highlight the similarity of entities relative to each other.

## 6 Conclusion and future perspectives

In this paper, we introduced *ProxMetrics*, a modular toolkit designed to assess the similarity of social media entities (namely users, groups, themes, places, and time periods). This toolkit is grounded in the theory of proxemics, a traditionally physical theory that we adapted for social media. The adaptation of this theory allows the toolkit to be highly versatile, enabling domain stakeholders to build insightful indicators for various requirements by manipulating the five dimensions of proxemics (DILMO): Distance, Identity, Location, Movement, and Orientation.

Our experimentation within the tourism domain demonstrated not only the practical applicability of *ProxMetrics* but also its potential to contribute meaningfully to domain stakeholders. The proxemic similarity scores that it generates facilitate a deeper understanding of the relationships and behavioral patterns within social media. Through the proxemic viewer dashboard (*ProxViewer*), we have also addressed the accessibility of the toolkit, ensuring that it can be used by non-computer scientist users.

We are currently seeking to enhance this work in various ways. Firstly, we plan to conduct a more extensive evaluation of the toolkit on larger datasets (to determine whether it scales effectively to massive volumes of data, on the order of millions of posts) and across other domains (such as fashion and public policy). This evaluation will be

conducted in collaboration with management researchers and will cover a broader range of end users through semi-structured interviews. Secondly, although our current choice of foundational measures to evaluate proxemic similarity has proven effective in the tourism experiment, we want to explore other measures (such as series-based or graph-based ones) to determine whether they would significantly improve the quality of results for each proxemic dimension. Thirdly, we aim to improve the explainability of how these dimensions influence the calculated proxemic similarity scores for non-specialist users by adding reactive visual aids to the *ProxViewer* dashboard. Lastly, we envisage using the *ProxMetrics* toolkit for the detection of bots and avatars on social media, addressing one of the world's most pressing issues (Ferrara 2023). The *ProxMetrics* toolkit can calculate proxemic similarities between users, which may be indicative of automated behavior. For effective bot detection, it will be necessary to incorporate additional dimensions, such as temporal activity patterns (Chavoshi et al. 2017), interaction diversity (Kosmajac and Keselj 2019), linguistic consistency (Cardaioli et al. 2021), and network centrality (Shinan et al. 2023), to enhance the accuracy and reliability of the detection process.

The toolkit and associated dashboard will be made open source and released to the public in the coming months.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Akram W, Kumar R (2017) A study on positive and negative effects of social media on society. *Int J Comput Sci Eng* 5(10):351–354
- Alt H, Godau M (1995) Computing the fréchet distance between two polygonal curves. *Int J Comput Geom Appl* 5(01n02):75–91
- Amir S, Wallace BC, Lyu H, et al (2016) Modelling context with user embeddings for sarcasm detection in social media. arXiv preprint [arXiv:1607.00976](https://arxiv.org/abs/1607.00976)
- Anderson A, Huttenlocher D, Kleinberg J, et al (2012) Effects of user similarity in social media. In: Proceedings of the fifth ACM international conference on Web search and data mining, pp 703–712
- Atout F (2023) Synthèse et sources de données. [https://www.atout-france.fr/sites/default/files/imce/synthese\\_et\\_sources\\_de\\_donnees\\_-\\_atout\\_france\\_18102023\\_vd\\_0.pdf](https://www.atout-france.fr/sites/default/files/imce/synthese_et_sources_de_donnees_-_atout_france_18102023_vd_0.pdf), accessed: 2023-11-20

- Barbieri F, Espinosa Anke L, Camacho-Collados J (2022) XLM-T: multilingual language models in Twitter for sentiment analysis and beyond. In: Proceedings of the thirteenth language resources and evaluation conference. European language resources association, Marseille, France, pp 258–266. <https://aclanthology.org/2022.lrec-1.27>
- Baucom E, Sanjari A, Liu X, et al (2013) Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In: Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing, pp 61–68
- Becker H, Naaman M, Gravano L (2010) Learning similarity metrics for event identification in social media. In: Proceedings of the third ACM international conference on Web search and data mining, pp 291–300
- Bergroth L, Hakonen H, Raita T (2000) A survey of longest common subsequence algorithms. In: Proceedings Seventh International symposium on string processing and information retrieval. SPIRE 2000, IEEE, pp 39–48
- Bhor HN, Koul T, Malviya R, et al (2018) Digital media marketing using trend analysis on social media. In: 2018 2nd International conference on inventive systems and control (ICISC), IEEE, pp 1398–1400
- Bolton C, et al (2010) Logistic regression and its application in credit scoring. Phd thesis, University of Pretoria
- Bookstein A, Kulyukin VA, Raita T (2002) Generalized hamming distance. *Inf Retr* 5:353–375
- Cardaioli M, Conti M, Di Sorbo A, et al (2021) It's a matter of style: detecting social bots through writing style consistency. In: 2021 International conference on computer communications and networks (ICCCN), IEEE, pp 1–9
- Castañer M, Camerino O, Anguera MT et al (2013) Kinesics and proxemics communication of expert and novice pe teachers. *Qual Quant* 47(4):1813–1829
- Charalabidis Y, Loukis E (2012) Participative public policy making through multiple social media platforms utilization. *Int J Electron Gov Res (IJEGR)* 8(3):78–97
- Chavoshi N, Hamooni H, Mueen A (2017) Temporal patterns in bot activities. In: Proceedings of the 26th international conference on world wide web companion, pp 1601–1606
- Clemens K (2015) Geocoding with openstreetmap data. *GEOProcessing* 2015:10
- Coghetto R (2016) Chebyshev distance. *Formaliz Math* 24(2):121–141
- Conneau A, Khandelwal K, Goyal N, et al (2019) Unsupervised cross-lingual representation learning at scale. *CoRR* abs/1911.02116. [arXiv:1911.02116](https://arxiv.org/abs/1911.02116)
- Cristani M, Paggetti G, Vinciarelli A, et al (2011) Towards computational proxemics: inferring social relations from interpersonal distances. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, IEEE, pp 290–297
- Devlin J, Chang MW, Lee K, et al (2018) Bert: pre-training of deep bidirectional transformers for language understanding. [arXiv preprint arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
- Duarte JM, Santos JBD, Melo LC (1999) Comparison of similarity coefficients based on rapid markers in the common bean. *Genet Mol Biol* 22:427–432
- Ferrara E (2023) Social bot detection in the age of chatgpt: challenges and opportunities. *First Monday*
- Greenberg S, Marquardt N, Ballendat T et al (2011) Proxemic interactions: the new ubicomp? *Interactions* 18(1):42–50
- Groenen PJ, Mathar R, Heiser WJ (1995) The majorization approach to multidimensional scaling for minkowski distances. *J Classif* 12:3–19
- Gunawan AB, Pratama B, Sarwono R (2021) Digital proxemics approach in cyber space analysis—a systematic literature review. *ICIC Express Lett* 15(2):201–208
- Hall ET (1966) The hidden dimension, vol 609. Anchor
- Hans A, Hans E (2015) Kinesics, haptics and proxemics: aspects of non-verbal communication. *IOSR J Humanit Soc Sci (IOSR-JHSS)* 20(2): 47–52
- Huang L, Liu G, Chen T et al (2021) Similarity-based emergency event detection in social media. *J Saf Sci Resil* 2(1):11–19
- Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993) Comparing images using the hausdorff distance. *IEEE Trans Pattern Anal Mach Intell* 15(9):850–863
- Hvass KA, Munar AM (2012) The takeoff of social media in tourism. *J Vacat Mark* 18(2):93–103
- INSEE (2023) Insee - statistiques locales. <https://statistiques-locales.insee.fr/>, accessed: 2023-11-20
- Isère Attractivité (2023) Carnet observatoires. [https://isere-attractivite.com/sites/default/files/minisite/pro/rat\\_carnet\\_observatoires\\_maquette\\_mise\\_en\\_ligne.pdf](https://isere-attractivite.com/sites/default/files/minisite/pro/rat_carnet_observatoires_maquette_mise_en_ligne.pdf), accessed: 2023-11-20
- Jiang L, Yang CC (2017) User recommendation in healthcare social media by assessing user similarity in heterogeneous network. *Artif Intell Med* 81:63–77
- Jiashun C (2012) A new trajectory clustering algorithm based on trachus. In: Proceedings of 2012 2nd international conference on computer science and network technology, IEEE, pp 783–787
- Johansson F, Kaati L, Shrestha A (2013) Detecting multiple aliases in social media. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining, pp 1004–1011
- Khan MU, Choi JP, Shin H, et al (2008) Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In: 2008 30th annual international conference of the IEEE engineering in medicine and biology society, IEEE, pp 5148–5151
- Knoll J (2016) Advertising in social media: a review of empirical evidence. *Int J Advert* 35(2):266–300
- Kosmajac D, Keselj V (2019) Twitter bot detection using diversity measures. In: Proceedings of the 3rd international conference on natural language and speech processing, pp 1–8
- Lahitani AR, Permanasari AE, Setiawan NA (2016) Cosine similarity to determine similarity measure: Study case in online essay assessment. In: 2016 4th International conference on cyber and IT service management, IEEE, pp 1–6
- Leys C, Klein O, Dominicy Y et al (2018) Detecting multivariate outliers: use a robust variant of the mahalanobis distance. *J Exp Soc Psychol* 74:150–156
- Liu H, Hu Z, Mian A et al (2014) A new user similarity model to improve the accuracy of collaborative filtering. *Knowl-Based Syst* 56:156–166
- Liu Y, Liu Z, Chua TS, et al (2015) Topical word embeddings. In: Proceedings of the AAAI conference on artificial intelligence
- Llobera J, Spanlang B, Ruffini G, et al (2010) Proxemics with multiple dynamic characters in an immersive virtual environment. *ACM Trans Appl Percept* 8(1)
- Lu Y, Wang R, Zhang Y et al (2015) Ecosystem health towards sustainability. *Ecos Health Sustain* 1(1):1–15
- Luxey A (2019) E-squads: a novel paradigm to build privacy-preserving ubiquitous applications. Phd thesis, Université Rennes 1
- Ma R, Zhou X, Gui T, et al (2022) Template-free prompt tuning for few-shot NER. In: Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies. Association for computational linguistics, ACL, pp 5721–5732. <https://doi.org/10.18653/v1/2022.naacl-main.420>
- Masson M, Sallaberry C, Agerri R, et al (2022) A domain-independent method for thematic dataset building from social media: the case of tourism on twitter. In: International conference on web information systems engineering, Springer, pp 11–20

- Masson M, Agerri R, Sallaberry C, et al (2023a) Optimal strategies to perform multilingual analysis of social content for a novel dataset in the tourism domain. arXiv preprint [arXiv:2311.14727](https://arxiv.org/abs/2311.14727)
- Masson M, Roose P, Sallaberry C, et al (2023b) Aps: A proxemic framework for social media interactions modeling and analysis. In: International symposium on intelligent data analysis, Springer, pp 287–299
- Mazhari S, Fakhrahmad SM, Sadeghbeygi H (2015) A user-profile-based friendship recommendation solution in social networks. *J Inf Sci* 41(3):284–295
- McCall C (2015) Mapping social interactions: the science of proxemics. *Social behavior from rodents to humans* pp 295–308
- Medeiros D, Dos Anjos R, Pantidi N, et al (2021) Promoting reality awareness in virtual reality through proxemics. In: 2021 IEEE virtual reality and 3D user interfaces (VR), IEEE, pp 21–30
- Mehta V (2020) The new proxemics: Covid-19, social distancing, and sociable space. *J Urban Des* 25(6):669–674
- Moreau C, Devoegele T, Peralta V, et al (2020) A contextual edit distance for semantic trajectories. In: Proceedings of the 35th annual ACM symposium on applied computing, pp 635–637
- Mueller F, Stellmach S, Greenberg S, et al (2014) Proxemics play: Understanding proxemics for designing digital play experiences. In: Proceedings of the 2014 conference on designing interactive systems. Association for computing machinery, New York, NY, USA, DIS '14, p 533–542
- Müller M (2007) Dynamic time warping. *Information retrieval for music and motion* pp 69–84
- Navarro G (2001) A guided tour to approximate string matching. *ACM Comput Surv (CSUR)* 33(1):31–88
- Neiger BL, Thackeray R, Van Wagenen SA et al (2012) Use of social media in health promotion: purposes, key performance indicators, and evaluation metrics. *Health Promot Pract* 13(2):159–164
- Ng BL, Liu W, Wang JC (2016) Student motivation and learning in mathematics and science: a cluster analysis. *Int J Sci Math Educ* 14:1359–1376
- Nguyen TT, Camacho D, Jung JE (2017) Identifying and ranking cultural heritage resources on geotagged social media for smart cultural tourism services. *Pers Ubiquit Comput* 21:267–279
- OECD (2023) Indicateurs clés du tourisme | statistiques de l'ocde sur le tourisme. [https://www.oecd-ilibrary.org/fr/economics/data/statistiques-de-l-ocde-sur-le-tourisme/indicateurs-cles-du-tourisme\\_a5ace089-fr](https://www.oecd-ilibrary.org/fr/economics/data/statistiques-de-l-ocde-sur-le-tourisme/indicateurs-cles-du-tourisme_a5ace089-fr), accessed: 2023-11-20
- Pérez P, Roose P, Cardinale Y et al (2021) An approach to develop mobile proxemic applications. *J Data Intell* 2(2):166–189
- Pilat Tourisme (2022) Tableau de bord 2022. [https://www.pilat-tourisme.fr/sites/pilat-tourisme/files/content/files/tableau\\_de\\_bord\\_2022.pdf](https://www.pilat-tourisme.fr/sites/pilat-tourisme/files/content/files/tableau_de_bord_2022.pdf), accessed: 2023-11-20
- Rios-Martinez J, Spalanzani A, Laugier C (2015) From proxemics theory to socially-aware navigation: a survey. *Int J Soc Robot* 7(2):137–153
- Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover's distance as a metric for image retrieval. *Int J Comput Vision* 40:99–121
- Shinan K, Alsubhi K, Ashraf MU (2023) Botsward: centrality measures for graph-based bot detection using machine learning. *Comput Mater Continua* 75(1)
- Smailhodzic E, Hooijsma W, Boonstra A et al (2016) Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Serv Res* 16(1):1–14
- Sponcil M, Gitimu P (2013) Use of social media by college students: relationship to communication and self-concept. *J Technol Res* 4(1):37–49
- Tang J, Chang Y, Aggarwal C et al (2016) A survey of signed network mining in social media. *ACM Comput Surv (CSUR)* 49(3):1–37
- UNWTO (2023) Tableau de bord de l'omt de données sur le tourisme. <https://www.unwto.org/fr/unwto-tourism-dashboard>, accessed: 2023-11-20
- Varlamis I, Sardanios C, Bogorny V, et al (2021) A novel similarity measure for multiple aspect trajectory clustering. In: Proceedings of the 36th annual ACM symposium on applied computing, pp 551–558
- Visit Paris Region (2023) Tableau de bord. <https://pro.visitparisregion.com/chiffres-du-tourisme/conjoncture/tableau-de-bord>, accessed: 2023-11-20
- Wang W, Zhang G, Lu J (2016) Member contribution-based group recommender system. *Decis Support Syst* 87:80–93
- Williamson J, Li J, Vinayagamoorthy V, et al (2021) Proxemics and social interactions in an instrumented virtual reality workshop. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp 1–13
- World Tourism Organization (2002) Thesaurus on tourism and leisure activities. World Tourism Organization
- Wu Z, Palmer M (1994) Verb semantics and lexical selection. arXiv preprint [cmp-19/9406033](https://arxiv.org/abs/1904.06033)
- Xia F, Liu J, Nie H et al (2019) Random walks: a review of algorithms and applications. *IEEE Transact Emerg Top Comput Intell* 4(2):95–107
- Yang Y, Baker S, Kannan A, et al (2012) Recognizing proxemics in personal photos. In: 2012 IEEE conference on computer vision and pattern recognition, pp 3522–3529
- Yeh A, Ratsamee P, Kiyokawa K, et al (2017) Exploring proxemics for human-drone interaction. In: Proceedings of the 5th International Conference on Human Agent Interaction. Association for computing machinery, New York, NY, USA, HAI '17, p 81–88
- Zangerle E, Bauer C (2022) Evaluating recommender systems: survey and framework. *ACM Comput Surv* 55(8):1–38
- Zangerle E, Gassler W, Specht G (2013) On the impact of text similarity functions on hashtag recommendations in microblogging environments. *Soc Netw Anal Min* 3:889–898
- Zhang Y, Wang X, Sakai Y, et al (2019) Measuring similarity between brands using followers' post in social media. *Proceedings of the ACM Multimedia Asia* pp 1–6

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.