AGILE Association of Geographic Information Laboratories in Europe

Corrigendum to "A taxonomy for classifying user groups in location-based social media" published in AGILE: GIScience Series, 1, 5, 2020

Thomas Gründemann and Dirk Burghardt

Institute of Cartography, Technische Universität Dresden, Dresden, Germany

Published: 2 September 2020

The following paper has to be considered as a replacement of the originally published version. The paper was provided by mistake in an incorrect version for publication. Corrected error: Figures were not displayed correctly in the pdf version created from the original file format.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://agile-giss.copernicus.org/articles/1/5/2020/
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

A taxonomy for classifying user groups in location-based social media

Thomas Gründemann[0000-0002-0596-7451] and Dirk Burghardt[0000-0003-2949-4887]

Institute of Cartography, Technische Universität Dresden, Dresden, Germany
e-mail: thomas.gruendemann@tu-dresden.de

Abstract. Location-based social media provide great opportunities to monitor and map social, natural or health-related events. Due to the vast amount of data, it is appropriate for many researchers to use a judiciously selected sample of data. However, many of the datasets from social media sources do not consist of representative samples of the overall population because they do not take into account the users who generate the social media content. The consequences can be a bias of particular user groups and a misinterpretation of the analysis results. To overcome these shortcomings, this paper develops a taxonomy of user groups in social media based on a thorough literature analysis. The different approaches can be summarized to the five dimensions: character, connectivity, communication, content and coordinates. The expected use of the taxonomy is to support the selection of social media datasets by choosing only those user groups that provide relevant information and to improve the analysis by identifying significant groups. Both application areas are illustrated by using a dataset that includes the members of the German parliament who registered on Twitter.

Keywords: Taxonomy; User groups; Classification; Social media; Population bias

1 Introduction

The mining and analysis of location-based social media (LBSM) has become an important task for the better understanding of social events and the functioning of the human society by evaluating information about public opinion on a topic or event. Such research includes, among others, the investigation of natural disasters [1], [2], protests [3], disease surveillance [4], and marketing campaigns [5].

While the amount of data generated on social media platforms is huge, the typical data collection tools provided by the publicly available application programming interfaces (APIs) are often insufficient to capture all the generated data. The authors in [6] describe this phenomenon as the data acquisition bottleneck. For example, about 350,000 messages per minute are created on Twitter[1]. At the same time, due to the rate limits, only 180,000 tweets per hour can be queried[2]. This challenge necessi-

[1] https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html
[2] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

tates the need for collecting a sample of the social media data that is suitable for answering the research questions.

However, the data selection and analysis are limited in many studies to the user-generated content, without taking into account the additional information available on the creators and their network properties. Such an approach ignores the fact that the content results from the presence of certain user groups, e.g. few actors with high number of contributions or stakeholders to a certain topic. The same applies to the different emphases of the social media platforms used by different groups, such as professionals on LinkedIn or microbloggers on Twitter. These methodological shortcomings and structural biases result in over-representation of particular user groups and can lead to misinterpretation if left unconsidered [7–9].

For both factors - the collection and the analysis of social media data - there is a need to take greater account on the users. Only through the knowledge of the user characteristics we can select suitable datasets and investigate complex social processes. This paper aims at developing a taxonomy for describing and classifying user groups in LBSM, thereby benefiting three areas:

1. To support the selection of appropriate and valid LBSM data by choosing only those user groups that provide relevant information to answer the respective research questions (e.g. users discussing a particular topic or who are active in a particular region);
2. To improve the analysis of LBSM data and the interpretation of the results by identifying significant user groups and recognizing their over- or under-representation;
3. To provide a clear terminology for distinguishing user groups on various social media platforms by merging different classification approaches.

At this point, it should be noted that the paper classifies generic user groups based on attributes that appear in a variety of social media. Methods that describe how attributes can be derived if they are not explicitly given are only mentioned briefly, since they would go beyond the scope of this work.

In the following, we first review the literature on classifying user groups in social media. Based on this review, we then explain the method for developing the taxonomy before introducing the proposed taxonomy and describing its categories in detail. Finally, we demonstrate the usefulness of the taxonomy by applying it to two typical use cases. To illustrate various examples throughout the paper and to assess the applicability of the taxonomy, we use two different datasets:

- A dataset retrieved from Twitter (hereinafter referred as ‘Twitter dataset’), that includes the user profiles, the follower network, and the timelines of the years 2017 and 2018 of the members of the German parliament (MPs) who signed up for Twitter. This dataset consists of 504 users with 5,190,044 follower connections and a total of 354,299 contributions (tweets).
- In addition, we used the MPs biographical information which is subject to the mandatory disclosure and published on the website of the German parliament

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

(hereinafter referred as 'master dataset')[3], to obtain socio-demographic characteristics of the MPs that could not be extracted from the Twitter dataset. The master dataset also serves as a comparison for the over- and under-representation of the groups of the Twitter dataset.

2 Classification of users in social media: a state of the art

With the emergence of online social networks in the mid-2000s, research in different fields, such as social science, computer science, GIScience etc. has shown interest in characterizing and classifying users in social media. Based on a thorough literature study five general approaches have been identified, that can be summarized as the '5 Cs of user classification' (Fig. 1):

- *Character:* classifies users based on their personal identity, in particular by demographic attributes;
- *Connectivity:* describes the collective identity of the users and the affiliation to social groups and social positions;
- *Communication:* groups users on the basis of their communication role;
- *Content:* classifies users with regard to the topics in their contributions;
- *Coordinates:* divides users in terms of their spatial and temporal characteristics.

In this section we give an overview of important related work concerning the classification of user groups in social media. Based on this research, we then describe our own synthesized taxonomy.

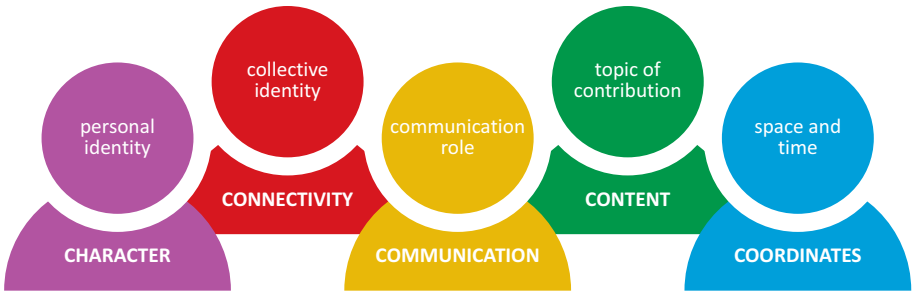

Fig. 1. The different approaches for classifying user groups in social media can be summarized as the '5 Cs of user classification'.

2.1 Classification based on character

One common feature of LBSM is the ability for the users to create a profile. This profile can include disclosing information such as name, age, gender, profession,

[3] https://www.bundestag.de/resource/blob/472878/e207ab4b38c93187c6580fc186a95f38/MdB-Stammdaten-data.zip

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

location, and also information that describe the user characteristics and preferences [10]. Various related works have used this information to classify users based on their character.

A simple but very concise classification is provided by [11]. The authors distinguish between real person, institution (abstract entity, e.g. company or organization) and fictional entity (the person or organization is not real) based on the user profile type. Work presented in [12] classifies users to detect humans, bots and cyborgs. The authors in [13] divide the users into two groups, each containing three classes: real users (personal users, professional and business users) and digital actors (spam users, feed/news and viral/marketing services). The two latter papers classify Twitter users, based on their personal attributes that they mention in profiles as well as on their communication behavior.

Other related works divide users on the basis of demographic attributes derived from the profile description. In [14], the authors categorize users in different classes for each of the personal attributes sex, age, political orientation, religious affiliation, ethnicity and sexual orientation. The class affiliation is derived from the name of the user and by a pattern-matching of the users' biographies. The classification was then used to label unknown users in the social network. Work presented in [15] follow a similar approach. They distinguish users according to ethnicity, place of origin, gender, language, and race, using only the attributes first name, last name and user-provided location from the profile. Such classifications take advantage of the fact that users who communicate with each other often have similar characteristics. The authors in [16] design and evaluate two tools for the automated classification of the age group, occupational group and the social class.

2.2 Classification based on connectivity

By communicating with each other, sharing information, or simply listing someone as a contact or as a friend or fan users connect to each other and thus form a social network. Consequently, how users of a social media platform are connected to each other determines to which group the users belong. Related work dealing with network-based user classification pursue two different approaches. The first examines the structure of the network and uses clustering methods to group users that are similar according to some definition. The second approach is based on the relative position of the users within that network and computes a numerical value for each user in the network [17].

The tendency of users to establish connections with those sharing similar characteristics - also known as homophily - can be used to group users with similar interests. Clustering algorithms are utilized to detect these groups (also called online communities) in the network. These algorithms harness the structural property of the network that users of a group are more densely connected to each other than to users outside the group [18]. A typology of online communities is given in [19]. The proposed classification scheme consists of two levels. The first level distinguishes according to the establishment and includes the categories 'Member-initiated' and 'Organization-sponsored'. At the second level, online communities are categorized based on their

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

relationship orientation. Member-initiated communities are formed either around social or professional member relationships. Organization-sponsored communities foster relationships from commercial, nonprofit and government members. In [20] the authors provide an example of communities based on shared interest. For their study the authors examined Instagram users in Amsterdam and Copenhagen. By online activities such as liking and commenting posts, the users constitute social ties, which form a network. A combination of network analysis and manual analysis of the profiles and image content of the ten most central users of each cluster was applied to characterize the different groups. This results in interest groups rooted in shared professions, lifestyles, hangouts, etc. such as 'Visual Professionals', 'Designers', 'College Students', Lifestyle Vanguards', 'Coffee Aficionados', or 'Party Buffs'.

LBSM users are not just the creators and recipients of the information, they also have certain positions within the network. To classify the users with regard to their function, a number of commonly accepted centrality measures have been proposed for the second approach of the network-based user classification. For each user in the network, a value is calculated according to the measure used. The authors of [21] were one of the first researchers who used the links in the social network to categorize users on Twitter. The authors not only divide the users in different groups based on the shared interest described above, but also used the link analysis algorithm HITS (Hyperlink-Induced Topic Search) to find the hubs and authorities in the network [22]. The computation of the values results in a rough categorization of the users into information source, friends and information seeker. Three prominent functions were identified by [23]: the hub (a person with links to many others), the broker or gatekeeper (a person who is the only connection between groups) and the bridge or pulsetaker (a user who links several groups and can see opportunities for exchange between them). In [17] the author classifies users based on the centrality measures 'degree', 'closeness', 'betweenness' and 'eigenvector'. The key users found are divided into the following groups: popular users (users with a high degree value have a high number of direct link to other users), amplifiers (users with a high closeness centrality are close to all other users in the network), disseminators (users with a high betweenness centrality are on a direct path to other groups and act as link between different communities) and influentials (users with a high eigenvector centrality are close to the most important people). The approach in [24] uses the node's in-degree, which refers to the number of users following the node and categorize users into mass media (in-degree > 100,000), evangelists (influentials, opinion leaders, hubs, or connectors with an indegree between 200 and 100,000) and grassroots (common users with an in-degree < 200). The authors in [25] provide a similar classification with a different approach based on network position and message activity. They divide Twitter users into 'influentials', 'hidden influentials', 'broadcasters' and 'common users' depending on their ratio of following/followers and messages received/sent.

2.3 Classification based on communication

Social media are primarily designed to facilitate communication among the users and groups. The way of communication, or rather how the members use the social media,

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

leads to different classification approaches in the literature. The authors of [26] use the communication tools provided by the Twitter platform and divide the users on a continuum between the categories 'personal-interactive' (higher usage of @ and RT compared to http://) and 'topical-informative' (higher usage of http:// compared to @ and RT). Further related works study the communication process in two different ways. The approaches described in [27, 28] focus on the type of user participation and differentiate between active and passive users. Whereas the authors in [29] examine the position of the users in the information propagation and categorize the users in idea starter, amplifier, curator, commentator and viewer.

2.4 Classification based on content

One of the key features of social media is user-generated content. The user's contributions consist of personal information, news, images, videos, status reports or links to content such as articles. Furthermore, the content often includes references to events and provides many different and inclusive views of the events. The approaches to classify users in terms of their content in their contributions differ mainly in regard to the granularity of the topics.

Dann [30] provides a classification at a low granularity level. The content classification framework is based on 16 existing studies and offers 6 main categories (Conversational, Pass-along, News, Status, Phatic and Spam) and 23 subcategories (e.g. Headlines, Sport, Event and Weather in the category News). In a similar vein, the authors in [31] examine the content of Instagram photos and identified five types of users (e.g. Selfie-lovers, Captioned photo users, outdoor and indoor activity users). A finer granularity with regard to the topics is pursued by approaches which focus on keywords and tags or generally words with a high frequency of mentions. Examples of user groups identified from single tags are given in [32]. In their work, the authors examined Twitter users who have posted messages with a specific hashtag. It is argued that by writing or forwarding topic-related posts, the users become engaged in the issue and hence, a member of the topic-related group. In the paper, the groups are further subdivided according to their attitude. The structure of the groups differs, depending on the subject and the users that are driving the conversation. An example in the paper is the polarized crowd with a divided structure which is common observation in political discussions. The users are focused on the same topic, but the discussion is highly divisive, resulting in two big and dense groups (e.g. proponents and opponents). The authors of the study observed five other different structures: unified, fragmented, clustered, and inward and outward hub and spoke structures.

2.5 Classification based on coordinates

Spatial and temporal analysis are two key parts for LBSM analysis [33] and are summarized here under the term coordinates. Research in GIScience mainly focus on the contributions, which represent a spatiotemporal signal (geolocation and timestamp) with a semantic information layer (content) as described in [34]. The three components place, time and topic are additionally specified by [35] as 'interdependent and

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

human-centered, which means they are originally defined or created by human beings (messengers)'. Based on these social media characteristics, there is a large body of research work that classifies users in terms of space and time.

For determining users as locals or visitors in the greater Seattle area, [36] used the time stamps of the contributions. Users being within the study area for more than nine days and outside for less than nine days were classified as locals. Users who did not meet this requirement were grouped as tourists. In their work on the credibility of social media information, [37] uses the three components place, time and topic to characterize micro-bloggers according to whether they perceive an event directly or indirectly. The authors distinguish between witness, potential witness, relay on witness and not witness, impact or relay. This classification depends on whether the user's contribution is spatially on-the-ground or unknown, temporally at present or delayed and thematically direct observed, direct impacted, relayed or not. A comprehensive analysis based on various facets is described in [38]. The authors demonstrate how the content of the user contributions varies according to the place, time and user. For example, the authors correlate the topics of the contributions with the demographic and socio-economic characteristics of the users and observe behavioral differences in different user groups. In the work of [39] various socio-demographic, spatial and temporal variables are used to classify Twitter users in London. Examples of such groups are 'Residents', 'Commuting Professionals', 'Spectators', 'Visitors'.

3 Research approach

Most of the work surveyed in the previous chapter classifies users on the basis of specific features that are appropriate for the intended use. Our proposed taxonomy for classifying user groups in LBSM focus on a broader set of categories based on the synthesis of the reviewed research papers. Due to the extensive availability of literature in the field of user groups classification, the empirical-to-conceptual approach, as described in [40], was chosen. The method depicted in Fig. 2 illustrates the different steps of the taxonomy development.

The approach used begins with the specification of a meta-characteristic as a starting point for the development of the taxonomy. The meta-characteristic forms the basis for the choice of the characteristics in the taxonomy and should be based on the purpose and the expected use of the taxonomy. Each characteristic should be a logical consequence of the meta-characteristic. Our purpose is to distinguish user groups in social media with the resulting expected use to support the selection of social media datasets by choosing only those user groups that provide relevant information and to improve the analysis by identifying significant groups. Therefore, we choose the following meta-characteristic:

Meta-characteristic: Attributes that are included in social media data and that characterize the social media user.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

Since the approach is an iterative method, we next determine the following objective and subjective conditions that end the development process:

- Objective ending conditions:
  - no new dimensions are added in the last iteration;
  - no additional user groups need to be examined.
- Subjective ending conditions:
  - the taxonomy is determined to be concise, robust, comprehensive, extendible, and explanatory.

The selection of a (new) subset of users (step 3 in Fig. 2) has already been performed by the authors of the research papers reviewed in chapter 2. Here, the datasets used in this literature form the basis. In step 4 we then identified common characteristics of these users and formed the respective groups. All user groups that represent values of a specific attribute are then grouped into a category in step 5. In addition, we have combined categories into dimensions that classify the groups from a certain angle. Steps 3 through 5 have been repeated until all objective and subjective conditions were met. The resulting taxonomy is presented in the following section.

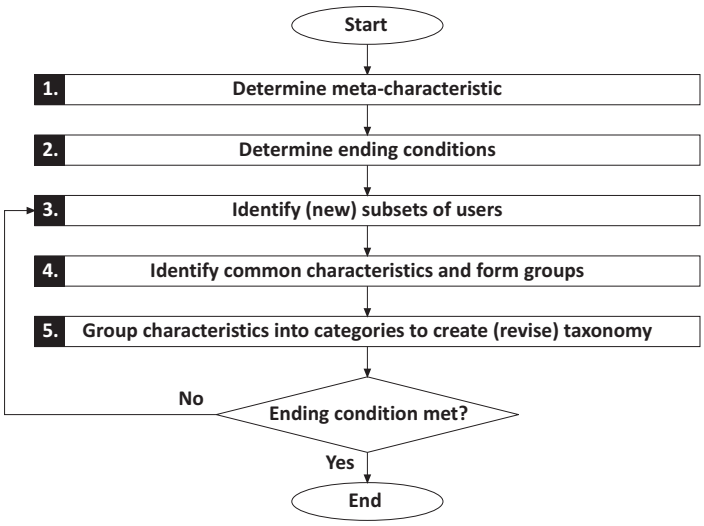

Fig. 2. The empirical-to-conceptual approach to develop the taxonomy (adapted from [40]).

4 Proposed Taxonomy

The proposed taxonomy (see Fig. 3) is derived from the previously described research model. Based on the stepwise development process, we have identified five main dimensions that have already been used to structure the different research approaches. They form the main directions to describe the user groups in a sufficiently compre-

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

hensive manner. The classification of the user groups in terms of social characteristics takes place in the dimensions 'character' and 'connectivity'. The first focuses on the personal identity, the second on the collective identity of the users. In the dimension 'communication' the user groups are classified according to modal characteristics, in the dimension 'content' according to thematic characteristics, and in the dimension 'coordinates' according to spatial and temporal characteristics. For each of these dimensions, we found one or more types of user attributes, producing 11 categories in our taxonomy. Some of the categories have a limited set of possible groups, others are naturally infinite. For example, the category 'Type of social actor' can be considered as complete with the three groups 'Individuals', 'Organizations' and 'Virtual entities'. In contrast to that, the category 'Topic of contribution' is a clear example of an unlimited number of possible user groups. For categories which do not explicitly define the number of possible user groups, it is left to the analyst to choose them in such a way that they describe the object in sufficient detail. For example, the category 'Age' may include the user groups 'young', 'middle age' and 'old' or the age ranges 10-19, 20-29, 30-39, ..., depending on the desired level of granularity and the available data.

Data availability may be the biggest hurdle for user classification. In this context, we distinguish between explicit and implicit data. Explicit data is provided directly by an API or is directly available in a database and can be used immediately for classification. This includes, among others, voluntary information (e.g. date of birth and origin) provided in the user profile or the data and metadata of published contributions and photos. However, datasets are often incomplete, or the user attributes are not explicitly available. In this case, the information has to be gained with additional processing, which is why we call it implicit data. In general, implicit data are derived from explicit data or from user behavior. If, for example, the IP address is explicitly available, it can be used to infer the actual location of the user. Missing user attributes can also be derived from the combination of two (or more) explicit information. For example, if a significant number of contacts live in a city, one can conclude that the user of the social network might live there as well. By analyzing the activity patterns (e.g. timestamps and contents of the contributions), additional information about the communication role or the type of social actor can be derived that is generally not explicitly provided. If missing attributes cannot be derived from the dataset, other data sources must be used where the user characteristics are available in an explicit or implicit manner. Furthermore, we do not specify that the groups in a category must be mutually exclusive or may overlap. In each category, however, the total quantity of users (i.e. all users included in a dataset) always forms the basis for the classification into the different groups.

The taxonomy is not an exhaustive list of user groups. Rather, it is intended to reflect the many facets and possibilities of the user classification that are useful for a variety of analytical purposes. It is up to the analyst to extend it to cover specific use cases. For example, a new category 'Language' in the dimension 'Character' may be helpful for certain analytical purposes to group users in different linguistic areas. In the following, the different categories and user groups are briefly introduced, by describing the concepts, explaining the terms and presenting data and classification methods that can be used to identify the user groups.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

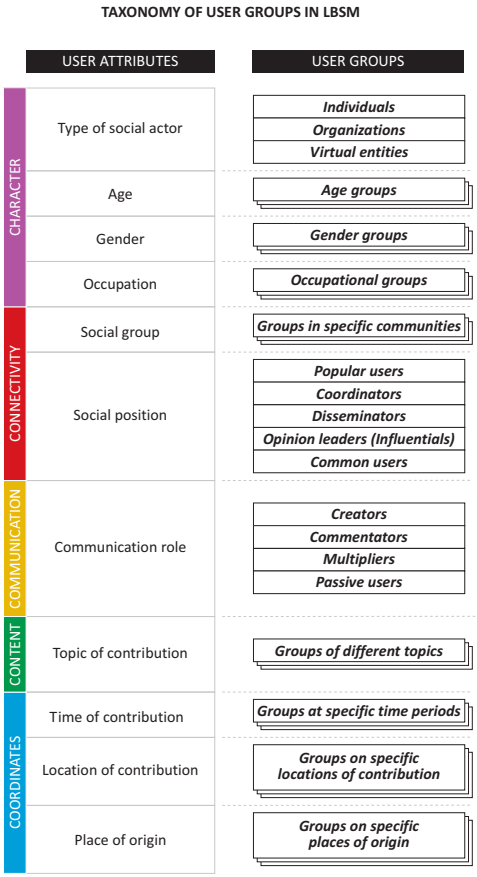

Fig. 3. Proposed taxonomy for classifying user groups in location-based social media.

4.1 User groups based on character

The dimension 'character' describes the personal identity of the users and is characterized by the category 'social actor' and different categories which can be subsumed under the term 'socio-demographic features".

Type of social actor. In the reviewed literature, the concept of the social actor refers mainly to individuals, organizations and virtual entities. The social actor is only the digital representation of the type of user. The social actor of the type *individual* is a real person. In LBSM, a social actor can also be an *organization* (institution, company, association). Responsible for the online presence of an organization is often the public relations department or the social media manager of the organization. A *virtual entity* is an actor that does not directly represent a physical person or organization. Examples for this are fictional characters and social bots. While fictional characters

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

are virtual figures or avatars behind which real people can hide, social bots are computer algorithms that automatically produce content.

Specific consideration should be given to the intentions or purposes affecting the actions of the different social actors. Individuals often want to promote themselves, to get the latest news, or to stay in touch with friends. Similarly, organizations use social network services for target-oriented advertising, marketing campaigns, and to communicate with customers, vendors, and the public at large [10]. Virtual entities of the type 'fictional character' want to obfuscate their identity or distribute certain content under a certain name (for example, the parody Twitter account @FakeScience). Social bots intent to inform other users in the case they provide content from automated sources (e.g. sensors, news feeds) or to alter the behavior of other individuals or organizations by exhibiting human-like behavior [12].

To detect different types of social actors a variety of methods for pattern detection and natural language processing are used. Important features include the first name in the username and the user description to distinguish individuals from organizations. Virtual entities can be identified by analyzing the connections to other users, content and sentiment features and the temporal patterns of activity [41].

Socio-demographic features. Socio-demographic features are used in the reviewed literature to form groups of users that primarily characterize the social actor of the type 'individual'. Common groups that describe the user structure are, for example, *age groups*, *gender groups* or *occupational groups*. The extent and type of socio-demographic features vary in LBSM. This depends primarily on the input fields provided by the social media platforms and the willingness of the users to disclose the information.

The same applies to the derivation of user groups in terms of their socio-demographic features. Not all attributes are explicitly provided by the users and can be extracted directly from the profile entries. The attributes can also be derived from one or more user-provided information (e.g. using the user's name to determine the age and gender) or it is possible to derive attributes from the combination of information. For example, it can be concluded that a user could have the same nationality as the majority of its contacts.

4.2 User groups based on connectivity

In the research literature examined, the dimension 'connectivity' describes the collective identity of the users and is characterized by the categories 'social group' and 'social position'.

Social group. An essential characteristic of LBSM is the formation of relationships to other users that lead to social groups. In the context of social media, social groups are typically referred to as communities. The social group consists of users who are in regular contact with each other, feel that they belong together and pursue common objectives and interests [42]. These characteristics distinguish social groups from

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

groups of different topics. Given this description, it is clear that social groups can be formed around an infinite number of common objectives and interests. However, three main classes of social groups can be identified from the literature examined:

- private communities: evolve around leisure activities, hobbies or other non-professional interests
- professional communities: formed around shared professional interests
- commercial communities: formed around products or companies

In the network structure, the social group is characterized by the fact that users of the group are more densely connected internally than with the rest of the network. Several methods have been developed which take advantage of this feature to detect social groups in LBSM, such as vertex clustering and community quality optimization methods [43].

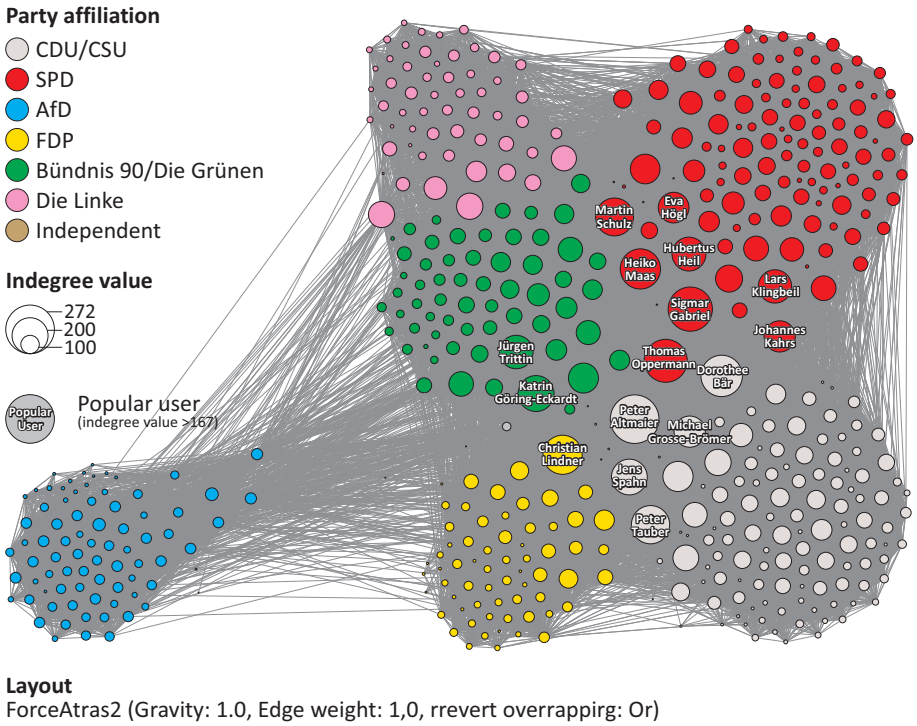

Fig. 4. The relationship network of social media users can be used to determine the affiliation to social groups and the social position. This is illustrated here using the example of the members of the German parliament, who are signed in for Twitter.

A good example of professional communities is the party affiliation of the members of parliament (MPs). The party affiliation can be observed after applying the Louvain method for community detection to the network of the previously described Twitter dataset. Fig. 4 shows the network structure in a force-directed layout [44] as a graph

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

consisting of nodes that represent the MPs and edges that represent the follower relationships among them. The nodes are colored according to their party affiliation, and it is obvious that MPs belonging to the same party are more closely connected than MPs of another party.

Social position. In analogy to relative spatial arrangements, the social position is the place of a user in a network of social relationships. In social media, the social position is linked to and defined by certain tasks and functions of the users and can be associated with different immaterial resources, such as power, influence and prestige. Link analysis algorithms are often used to determine specific social positions. From the reviewed research papers four different groups can be derived.

- *Popular users:* quickly spread information to a many directly connected users in a localized area. They can be determined by a high degree centrality value.
- *Coordinators:* gather information from users and share it with other users. The closeness centrality is an indicator for this group, since it measures the average length of paths from a user to all other users in the network. Users with small length path to all users are considered more likely to be coordinators, since they get information faster than those with high length path.
- *Disseminators:* act as an important link between different users and social groups. They can be determined by the betweenness centrality that measures the extent to which a user lies on a path between other users.
- *Opinion leaders (Influentials):* are users who initiate most activities, and with whom other users tend to interact most. The basis for the classification of this group is the eigenvector centrality. It assigns relative scores to all users in the network based on the concept that connections to users with high scores contribute more to the score of the user in question than equal connections to users with low scores.
- Users who do not have a specific function and do not meet the above-mentioned criteria can be grouped into *Common users*.

It should be pointed out that the identification of the social position based on centrality measures is more a point of reference than a factual statement. Furthermore, users can take multiple social positions within the network, depending on which threshold is used for each link analysis algorithm. The division into a specific group can then be based on additional statistical methods or ranking procedures.

As an example, we divided the users of the Twitter dataset into popular and common users. The classification into these groups is based on the indegree centrality (indicated by the node size in Fig. 4), since Twitter supports directed relationships. The indegree value of a member of parliament indicates how many follower connections exist to this user. For this example, we define a popular user as a member of parliament, who is followed by at least one third of all other MPs in the network (indegree value >167). The setting of the threshold arises from the condition that a popular user has followers from at least two different parties. Since already around 25% of all users belong to the party CDU/CSU, it has been determined that at least a

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

quarter of the followers of a popular user should be a member of another party. A total of 16 users were classified as popular users. These are marked with their name inside the node. Due to their strong networking, including connections to other party members, they have a central position within the network.

4.3 User groups based on communication

The research papers that classifies users in terms of communication follows three different approaches, (1) by the tools of communication, (2) by the type of user participation, and (3) by the position in information propagation. We have discovered that the tools of communication of the investigated social media platforms are very different and work in this regard is mainly restricted to the Twitter platform. Therefore, we do not consider a general user classification based on the tools of communication to be practical. We combine the two approaches (2) and (3) into a 3x3 matrix, dividing the type of user participation into action, reaction and inaction, and the position in the information propagation in creation, sharing and consumption. The classifications of the examined works can be divided in this matrix into four different user groups, which reflect the communication role (Fig. 5).

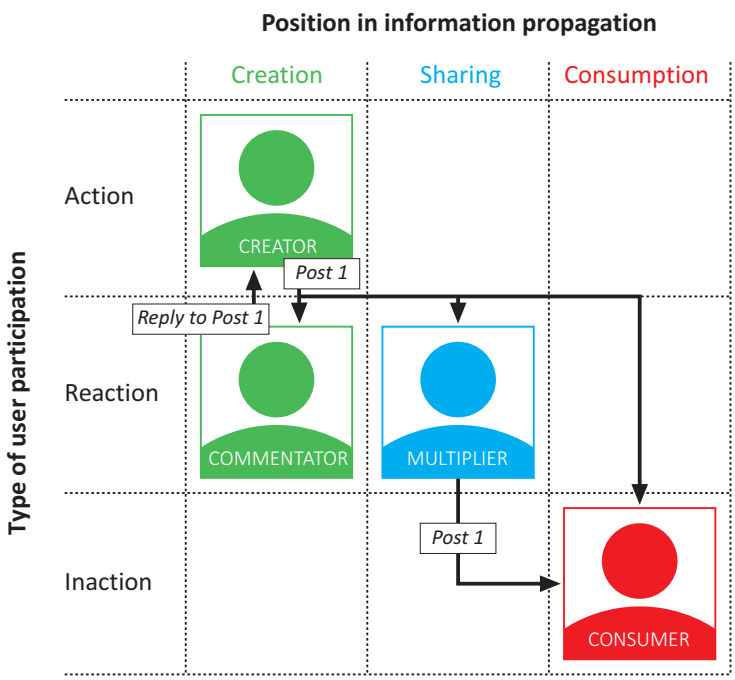

Fig. 5. User groups that are classified according to their communication role are formed by a relationship between the type of user participation and the position in the information propagation.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

- *Creators:* Actively create new content. They are at the beginning of the flow of information;
- *Commentators:* React to the content of the creators. Through their answers, they also create new content;
- *Multipliers:* Get content from the creators and share it with their own relationship network;
- *Consumers:* Consume the content they are interested in from creators and multipliers. They often do not appear actively.

Which communication role the users play often depends on other dimensions, for example the thematic dimension. A user may be an expert in one topic and thus a source of information or creator, or he may be a consumer in another topic. Therefore, the classification of users should be based on the role they play in the majority of their communications.

4.4 User groups based on content

In the literature reviewed, the classification of user groups based on content mainly uses the topics covered and the granularity of the content. Due to the wide variety of themes, the users are divided into *groups of different topics*. Depending on the content, a user's contribution can be assigned to one or more topics. Thus, users with contributions to the same or similar topics can be aggregated to one thematic group. In contrast to the social group, which is characterized by the relationships among each other, a thematic group includes all users who refer to the same topic, regardless of whether they are linked to each other or not.

The formation of subgroups is appropriate when classifying users who create contributions to subfields of a topic. For example, the user group with contributions to the topic 'Computer' can be further subdivided into the subgroups 'Games' or 'Programming'. Users with different opinions, attitudes, viewpoints towards a topic, can also be divided into subgroups. Especially political discussions form different camps, such as liberals and conservatives; proponents and opponents. The users are focused on the same topic, but their views are often opposed [32]. Possible further subgroups are based on the different sentiments, feelings or emotions (e.g., positive, negative, neutral; joy, surprise, sadness, anger, fear) [45].

Specific topics are extracted from keywords, tags, hashtags, n-grams by using natural language processing techniques [46]. Tags and Hashtags are user-selected terms to emphasize content, to refer to subjects or events, to facilitate data access, and to enable community networking [47]. Therefore, such terms can be used for topic grouping. With regard to the thematic analysis of user-generated data, it should be noted that users are not professional authors. User generated documents often contain very diverse vocabulary, abbreviations and typos. The classification of groups of different topics dependents to a large degree on the context of the analysis. The analyst must define user groups that are appropriate for his purpose.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

4.5 User groups based on coordinates

In the examined works, user groups based on coordinates are mainly classified according to various dimensions, for example, by means of content or socio-demographic features. However, the basis is always the spatial and temporal behavior of the users. In terms of time, the users are classified exclusively according to the timestamp of their contribution. The spatial classification of users is based on two different location information. These are the place of origin sometimes provided in the profile information and the geotagged location of the contribution. Therefore, we propose the classification of users into the three categories 'Time of contribution', 'Place of origin' and 'Location of contribution'.

Time of contribution. Regarding the temporal perspective, the users are grouped according to specific periods of time. The aggregation can be based on specific time frames (hours, days, weeks), recurring periods (Sundays, Weekdays) or seasonal time intervals (holidays, summer). The period in which the users are grouped depends on the subject of investigation. A grouping of users may also be possible in a temporal context with a particular event that has occurred over a specific period of time (reference time). Thus, statements can be made as to whether the users made contributions before, during or after this reference time. Users contributing before the reference time make statements about an expected event, e.g. the purchase of tickets (statement) for a concert (event) next week (reference time). Users contributing during the reference time create contributions as the event takes place. Due to the immediate temporal coincidence between the event and the creation of the contribution, these contributions can contain comments directly from the event reflecting, for example, current feelings and emotions. The group of users contributing after the reference time reports after an event has occurred. This may be the case, for example, when users tell about their visit at the concert when they return home. The choice of an appropriate time stamp is a decisive factor, as different temporal information may be available. For example, Flickr data contain three different types of time attributes: 'taken' - the timestamp the picture was created by the user's camera, 'posted' - the timestamp the picture was uploaded to the Flickr platform and 'lastupdate' - the timestamp when the description of the picture has been modified.

Place of origin. Users often disclose the home location in their profile, which describes the place of residence of individuals or the place of business of organizations. Consequently, this information can be considered as an indicator of the cultural background [48]. Depending on the intended purpose, users can be aggregated in different ways (e.g. city level, state level, country level or inside and outside of a specific area). This results in groups on specific places of origin. The aggregation is also dependent on the availability and granularity of the data. The location field in the user's profile is often an optional text field. Thus, the entry can remain empty, contain a wrong or even a fictional place name. In terms of their granularity, the information in the location field covers the entire range from exact coordinates to continents [49].

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

Location of contribution. In contrast to the place of origin, the location of contribution shows the current place where the user was when he created or published the contribution. Thus, the location of contribution may reflect the place where the immediate situational aspects influenced and triggered the reaction (impression, attitude, emotion). Groups on specific locations of contribution can be classified according to meaningful regions relevant for the analysis (e.g. the same neighborhood, city or country). Furthermore, users frequently comment on events happening at or affecting their location or refer to locations representing momentary social hotspots (e.g. by referring to the area hit by a natural disaster, or to the location of a protest) [50]. Therefore, knowing whether a user is inside or outside the affected location or area during an event is an important factor in determining which user is likely to publish relevant information for the event. For example, in an earthquake, contributions coming from a place affected by the earthquake are more relevant than contributors outside the affected area. The definition of whether a user is inside or outside a reference place or affected area is the task of the analyst and depends on the spatial scope of the event and the analysis purpose.

As in the case of the place of origin, the location of contribution is often an optional information. If users decide to make their current location available, the position is determined with an accuracy of a few meters (using GPS satellites or WiFi networks), or at city, state and country level (using the IP address).

5 Usage of the taxonomy and application examples

Dealing with social media often means dealing with 'big data', signifying the collection and analysis of datasets about users and their activities [51]. The proposed taxonomy contributes to both LBSM data collection and analysis. In the following, we illustrate the use of the proposed taxonomy by applying it to the Twitter dataset. The examples should serve to examine the usefulness for the intended users and purpose. The target audience of the taxonomy are researchers and analysts working with data from LBSM. A typical area of application is to collect or select appropriate datasets to answer their research questions and to analyze this data with regard to involved user groups

5.1 The taxonomy as a tool to filter LBSM users

It is often not necessary to look at the full LBSM data to draw certain conclusions about the subject of investigation. This is the case, for example, when only those users are to be considered who comment on a particular topic or if only the users in a particular region are to be examined. With regard to the collection of LBSM data, the taxonomy allows the selection of user groups that are suitable for the context of the analysis. According to the various dimensions and categories, an analyst can reduce the required data based on its relevance for the analysis. The data restriction based on the selected criteria and the resulting exclusion and extraction of information, improves the relevancy of the data.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

A prerequisite for the data selection is that the analyst defines his research subject (determining which question should be answered or which problem should be solved) and selects suitable LBSM data sources. Using the taxonomy, the analyst can then restrict the data required for his purpose according to the various dimensions and categories. In doing so, he compares the attributes of the LBSM data source with the attributes of the subject of investigation to define criteria for the data collection - e.g. user groups that provide contributions on a particular topic, which actively participate in a certain period of time and in specific locations, or which belong to a particular social group. These criteria allow the analyst to query suitable datasets from the LBSM data source, thereby improving the relevance of the data for the purpose of the analysis. In this regard, however, the analyst should target a balance between increasing the relevance based on the filtering techniques used and a reduced significance and validity of the data due to sampling effects [6]. It is also important to note that the filtering of user groups according to specific attributes also affects the composition of the resulting population in the filtered dataset. The following example is intended to clarify the procedure just mentioned.

The majority of research, using Twitter as a data source, focus on event detection and the related investigation of unusual spatial, temporal, and semantic activity patterns [34]. For this purpose, semantic information such as hashtags are predominantly used. We adopt this approach for our analysis to answer the following question:

- Analysis task: What spatial, temporal and party-political patterns are formed by the members of the German Bundestag (MdB), who made contributions (tweets) on 'climate protection' in 2017?

Two restrictions for the selection of relevant data from our Twitter dataset can be derived from the question. Using the taxonomy, we only select (1) from the category 'Time of contribution' the 'Group at time period 2017' and (2) from the category 'Topic of contribution' the 'Group with topic "Klimaschutz" (climate protection)'. Although the term climate protection generally may include a variety of keywords that can be assigned to this topic, we only choose users who made contributions containing the hashtag "Klimaschutz" for the sake of simplicity and to illustrate the example. With the use of the above-mentioned constraints, the Twitter dataset returns a total number of 716 contributions from 86 users. By choosing the two criteria, we now have an appropriate selection to answer the question of our analysis task.

5.2 The taxonomy as a tool to describe and analyze LBSM users

As in our example, the analysis of social media data often only refers to a specific platform or service. As a result, the sampling frame also applies only to the users who have decided to join and use the service. The characteristics, behavior, and perspectives of a user who does not use the service are excluded from the analysis. In this regard, data from social media is often biased to the extent that certain user groups are over-represented and thus not representative of the whole population. In order to interpret the results, it is therefore crucial to know which user groups are active in the social media. The taxonomy helps to classify user groups with relevant and especially

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

demographically important characteristics in order to make statements about the representativeness of social media data.

The goal of our analysis is to identify spatial, temporal and party-political patterns of user groups in the selected dataset. For this purpose, we have aggregated the MPs by month of their contributions and according to the federal state of their electoral districts. The Party affiliation has already been assigned to each member in advance. By doing so, we created an interpretable number of user groups, shown in Fig. 6.

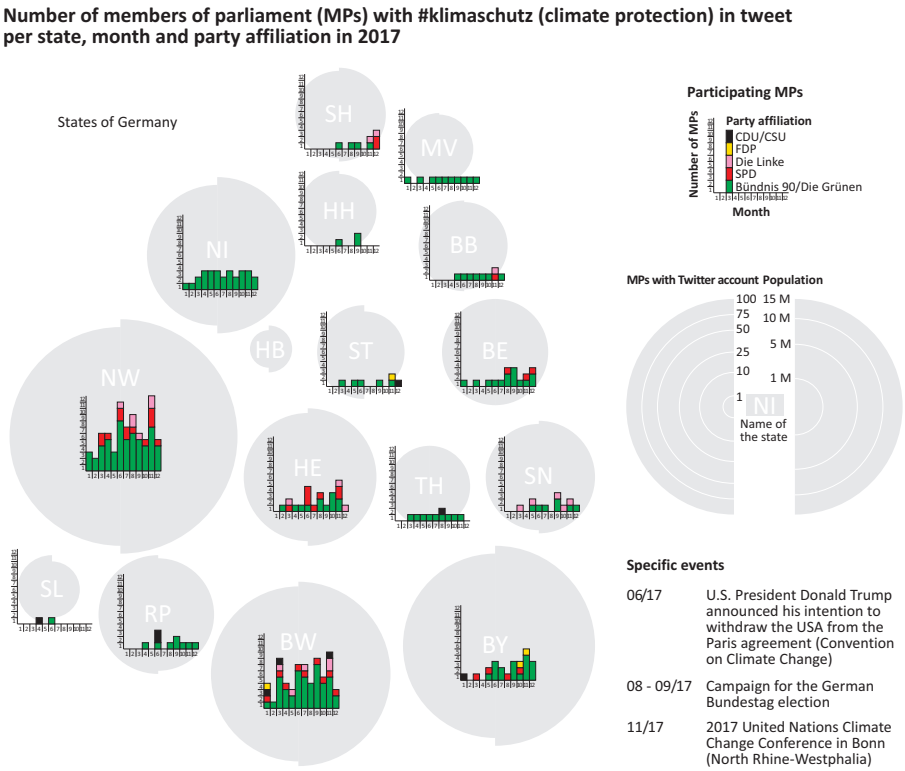

Fig. 6. Map showing the number of members of the German parliament (MPs) with #klimaschutz (climate protection) in tweet per state, month and party affiliation in 2017.

For the visual analysis we have chosen the Dorling cartogram, which on the one hand encodes the population of the federal states and on the other hand the number of MPs with a Twitter account by the size of the semicircles. In this way, the number of the MPs involved in the topic 'climate protection' can be put in relation to these two values and not in relation to the geographical area. Although the topology is not preserved, the approximate spatial position relative to the other states is sufficient for this analysis. The number of the participating MPs in the different months is represented by stacked bar charts, where the color represents the party affiliation. Fig. 7 shows in addition to the map, the proportions of the user groups in the selected dataset and their over- and under-representation in relation to the proportions of the groups in the mas-

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

ter dataset. For the analysis of the user groups, the categories gender, age, social group, place of origin and time of contribution were selected.

Some patterns are confirmed by the visual analysis of the map, shown in Fig. 6 and the charts, shown in Fig. 7. With regard to the social group the MPs of the green political party Bündnis 90/Die Grünen dominate the discussion and form the largest group with 58%, although they are only account for 9% in the parliament. This results in an over-representation of the party members on the topic of 'climate protection' of 49 percentage points (pp). The center-left party SPD and the far-left party DIE LINKE with a total of 25 users form groups with a proportion of 16% and 13% respectively. While the party DIE LINKE is over-represented with 3 pp, the party SPD is under-represented with 5 pp. The group of the center-right parties CDU and FDP with a total of 11 MPs are hardly involved in the discussion. They only appear in a few time periods and in six states. No contributions are made by the MPs of the far-right party AfD. The reasons are probably the political orientation of the party together with the low interest in environmental issues, as well as the fact that the members have only been present in the German parliament since 27 September 2017.

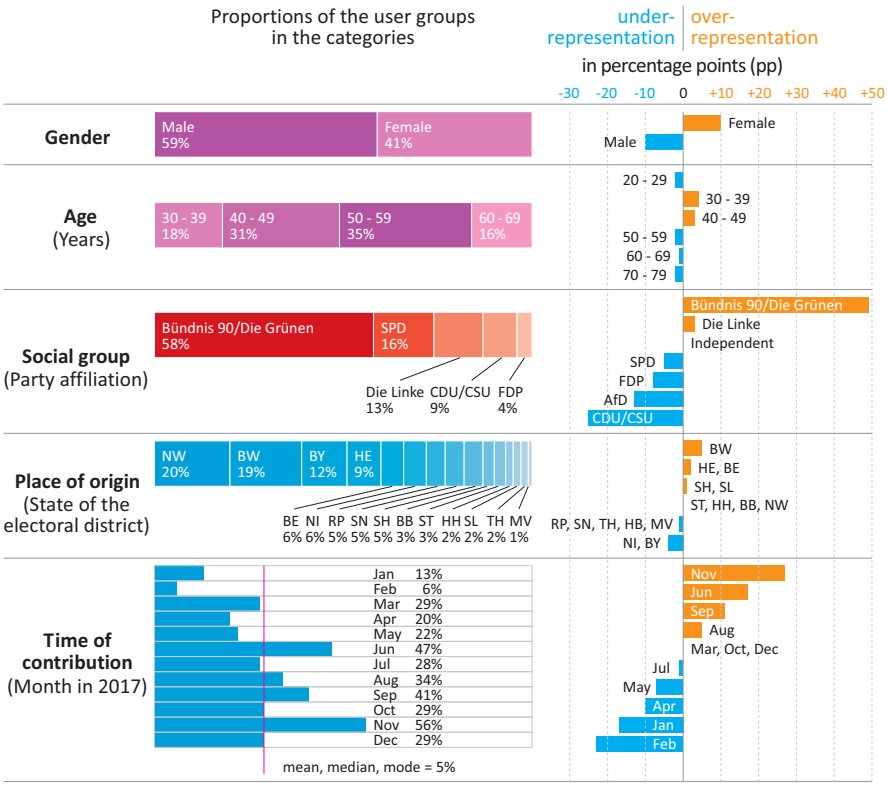

Fig. 7. Proportions of the user groups in the selected dataset and their over- and under-representation in relation to the proportions of the groups in the master dataset.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

With respect to the spatial distribution, the participating MPs in the three most populous southwestern states form the largest groups (North Rhine-Westphalia (NW) 20%, Baden-Württemberg (BW) 19%, and Bavaria (BY) 12%). While Baden-Württemberg is over-represented with 5 pp, Bavaria is under-represented with 4 pp. The reasons are the above-average proportion of MPs of party Bündnis 90/Die Grünen in the state parliament in BW and the lower participation of the MPs in the topic ‘climate protection’ in BY. No contribution with the hashtag ‘klimaschutz’ could be localized in Bremen. This may be due to the fact that only two MPs from this state have registered on Twitter.

The temporal perspective indicates that larger groups formed in November (48 of the 86 participating MPs - 59%), in June (47%), in September (41%), and in August (34%) especially in the southwestern states. A semantic analysis of the contributions reveals that the groups are mainly composed of MPs that respond to certain events that have occurred during the respective time periods. These are

- 06/2017: Announcement of U.S. President Donald Trump to withdraw from the Paris Agreement (Convention on Climate Change);
- 08 - 09/2017: Campaigns for the German Bundestag election;
- 11/2017: United Nations Climate Change Conference in Bonn (North Rhine-Westphalia).

Since no temporal comparison of the selected dataset with the master dataset is possible, the arithmetic average, median and mode were calculated to describe the over- and under-representation. All three parameters have the value 25 MPs (29%). Therefore, this value forms the basis of the representativeness. The group of the MPs using the hashtag 'climate protection' in November 2017 is consequently over-represented by 27 pp. The term climate protection used for the selection of the contributions and MPs is directly related to the three events, which explains the large participation of green-political MPs of the party Bündnis 90/Die Grünen during these periods.

A further analysis with regard to the socio-demographic categories shows the following results: the proportion of male users is 59% and of female users 41%. This results in a 10% over-representation of female MPs and consequently an under-representation of male MPs by the same amount. Six groups were formed for the category "age", broken down as follows: 20 - 29 years 0%, 30 - 39 years 18%, 40 - 49 years 31%, 50 - 59 years 35%, 60 - 69 years 16% and 70 - 79 years 0%. This shows that the age groups between 30 and 49 years are over-represented by about 3 percentage points and the age groups under 30 and from 50 are under-represented by about 2 percentage points.

This example illustrates that the selection of relevant data plays a crucial role in the analysis and interpretation of the results. In order to avoid misinterpretations, the dataset should not be restricted too much. Furthermore, additional data from other LBSM sources may be used to verify the results.

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

6 Conclusions

Location based social media has received significant research interest from many different disciplines in the recent years, as a means for understanding real-world phenomena. However, most existing work simply applies algorithms to analyze the LBSM data without knowing the population and user characteristics, even though the users form the center of social media. Any connection between them, every contribution is created by the users. Without a detailed knowledge of the social media users, this can lead to misinterpretations of the results of the analysis – especially for self-selected datasets provided by publicly available APIs.

To address these shortcomings, we developed a taxonomy for classifying user groups. Based on an empirical-to-conceptual approach, first the body of literature was analyzed to identify possible user characteristics. Subsequently, a structured taxonomy was derived by studying fundamental user groups and stepwise classifying them into non-redundant categories. This procedure revealed that social media users can be divided into five main dimensions - the 5 Cs of user classification: character (personal identity), connectivity (collective identity), communication role, content of the contribution and coordinates (space and time). The proposed taxonomy represents on the one hand the starting point for the selection of appropriate social media data by choosing only the relevant user groups for answering the research questions. On the other hand, it forms the basis for the classification and characterization of user groups represented in social media, which are crucial for the interpretation of a large number of research results. The use of the taxonomy for the two fields of application is illustrated by a dataset consisting of the user profiles, the follower connections and the timelines of the 504 members of the German parliament registered on Twitter. The extraction of attributes to characterize the users proved to be difficult. Not all attributes are available in LBSM datasets. Therefore, we have described some ways how specific attributes can be derived from the data. Furthermore, it can prove to be reasonable to use additional datasets from other sources that can provide missing attributes, as we have done for our example.

In the example described, we have only classified user groups according to individual characteristics such as gender, occupation, place of origin. However, significant user groups can often be described by several combined characteristics (academically educated women, young IT experts from Asia). In future work, we want to develop methods which, starting from the "basic groups" of the taxonomy, automatically recognize particularly important groups which can be characterized by multiple attributes.

7 References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. Proc. 19th Int. Conf. World Wide Web. 851–860 (2010). https://doi.org/10.1145/1772690.1772777
2. Middleton, S.E., Middleton, L., Modafferi, S.: Real-time crisis mapping of natural

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

disasters using social media. IEEE Intell. Syst. 29, 9–17 (2014). https://doi.org/10.1109/MIS.2013.126

3. He, J., Hong, L., Frias-Martinez, V., Torrens, P.: Uncovering Social Media Reaction Pattern to Protest Events: A Spatiotemporal Dynamics Perspective of Ferguson Unrest. In: Liu, T.-Y., Scollon, C.N., and Zhu, W. (eds.) Social Informatics. pp. 67–81. Springer International Publishing, Cham (2015)
4. Charles-Smith, L.E., Reynolds, T.L., Cameron, M.A., Conway, M., Lau, E.H.Y., Olsen, J.M., Pavlin, J.A., Shigematsu, M., Streichert, L.C., Suda, K.J., Corley, C.D.: Using social media for actionable disease surveillance and outbreak management: A systematic literature review. PLoS One. 10, 1–20 (2015). https://doi.org/10.1371/journal.pone.0139701
5. Zhou, L., Wang, T.: Social media: A new vehicle for city marketing in China. Cities. 37, 27–32 (2014). https://doi.org/10.1016/j.cities.2013.11.006
6. Choudhury, M. De, Lin, Y.-R., Sundaram, H., Candan, K.S., Xie, L., Kelliher, A.: How does the data sampling strategy impact the discovery of information diffusion in social media? Proc. 4th Int. AAAI Conf. Weblogs Soc. Media. 34–41 (2010). https://doi.org/papers3://publication/uuid/8A1DA629-F17A-4F21-9FBD-4CAA6C0D32A7
7. Boyd, D., Crawford, K.: Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15, 662–679 (2012). https://doi.org/10.1080/1369118X.2012.678878
8. Tufekci, Z.: Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media. 505–514 (2014)
9. Malik, M.M., Lamba, H., Nakos, C., Pfeffer, J.: Population Bias in Geotagged Tweets. ICWSM-15 Work. Stand. Pract. Large-Scale Soc. Media Res. 18–27 (2015)
10. Kietzmann, J.H., Hermkens, K., McCarthy, I.P., Silvestre, B.S.: Social media? Get serious! Understanding the functional building blocks of social media. Bus. Horiz. 54, 241–251 (2011). https://doi.org/10.1016/j.bushor.2011.01.005
11. Osteso, J.M., Claes, F., Deltell, L.: Teoría de la urdimbre comunicativa. Política, activismo y formación de líderes de opinión por medio de Twitter en España. In: XXVI Seminario internacional AISOC (2013)
12. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is Tweeting on Twitter: Human, Bot, or Cyborg? Acsac 2010. 21 (2010). https://doi.org/10.1145/1920261.1920265
13. Uddin, M.M., Imran, M., Sajjad, H.: Understanding Types of Users on Twitter. arXiv Prepr. 6 (2014)
14. Gayo-Avello, D.: All liaisons are dangerous when all your friends are known to us. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia. pp. 171--180 (2011)
15. Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., Yarowsky, D.: Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. Hlt-Naacl. 1010–1019 (2013). https://doi.org/10.1007/s00256-005-0933-8
16. Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-

https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

data. PLoS One. 10, 1–20 (2015). https://doi.org/10.1371/journal.pone.0115545
17. Kaati, L.: Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities. In: Alhajj, R. and Rokne, J. (eds.) Encyclopedia of Social Network Analysis and Mining. pp. 19–25. Springer New York, New York, NY (2014)
18. Lehmann, S.: Community Detection, Current and Future Research Trends. In: Alhajj, R. and Rokne, J. (eds.) Encyclopedia of Social Network Analysis and Mining. pp. 214–220. Springer New York, New York, NY (2014)
19. Porter, C.E.: A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. J. Comput. Commun. 10, (2004). https://doi.org/10.1111/j.1083-6101.2004.tb00228.x
20. Boy, J.D., Uitermark, J.: How to study the city on instagram. PLoS One. 11, 1–16 (2016). https://doi.org/10.1371/journal.pone.0158161
21. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: An Analysis of a Microblogging Community. In: Zhang, H., Spiliopoulou, M., Mobasher, B., Giles, C.L., McCallum, A., Nasraoui, O., Srivastava, J., and Yen, J. (eds.) Advances in Web Mining and Web Usage Analysis: 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007, and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007, San Jose, CA, USA, August 12-15, 2007. Revised Papers. pp. 118–138. Springer Berlin Heidelberg, Berlin, Heidelberg (2009)
22. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. J. ACM. 46, 604–632 (1999). https://doi.org/10.1145/324133.324140
23. Denning, P.J.: Network laws. Commun. ACM. 47, 15–20 (2004). https://doi.org/10.1145/1029496.1029510
24. Cha, M., Benevenuto, F., Haddadi, H., Gummadi, K.: The World of Connections and Information Flow in Twitter. IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans. 42, 991–998 (2012). https://doi.org/10.1109/TSMCA.2012.2183359
25. González-Bailón, S., Borge-Holthoefer, J., Moreno, Y.: Broadcasters and Hidden Influentials in Online Protest Diffusion. Am. Behav. Sci. 57, 943–965 (2013). https://doi.org/10.1177/0002764213479371
26. Thimm, C., Frame, A., Einspänner-Pflock, J., Leclercq, E., Anastasiadis, M.: The EU-Election on Twitter: Comparison of German and French Candidates' Tweeting styles. In: Holtz-Bacha, C. (ed.) Europawahlkampf 2014 : Internationale Studien zur Rolle der Medien. pp. 175–204. Springer Fachmedien Wiesbaden, Wiesbaden (2016)
27. Pagani, M., Hofacker, C.F., Goldsmith, R.E.: The influence of personality on active and passive use of social networking sites. Psychol. Mark. 28, 441–456 (2011). https://doi.org/10.1002/mar.20395
28. Lin, J.H.: Differential gains in SNSs: effects of active vs. passive Facebook political participation on offline political participation and voting behavior among first-time and experienced voters. Asian J. Commun. 26, 278–297 (2016). https://doi.org/10.1080/01292986.2016.1148184
29. Tinati, R., Carr, L., Hall, W., Bentwood, J.: Identifying Communicator Roles in Twitter. Proc. 21st Int. Conf. World Wide Web. 1161–1168 (2012). https://doi.org/10.1145/2187980.2188256
30. Dann, S.: Twitter content classification. First Monday. 15, (2010). https://doi.org/10.5210/fm.v15i12.2745

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

31. Hu, Y., Manikonda, L., Kambhampati, S.: What We Instagram: A First Analysis of Instagram Photo Content and User Types. Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media. (2014)
32. Smith, M.A., Rainie, L., Himelboim, I., Shneiderman, B.: Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters, http://www.pewinternet.org/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/
33. Gao, H., Tang, J., Hu, X., Liu, H.: Exploring temporal effects for location recommendation on location-based social networks. Proc. 7th ACM Conf. Recomm. Syst. RecSys '13, 93–100 (2013). https://doi.org/10.1145/2507157.2507182
34. Steiger, E., de Albuquerque, J.P., Zipf, A.: An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data. Trans. GIS. 19, 809–834 (2015). https://doi.org/10.1111/tgis.12132
35. Tsou, M.H., Leitner, M.: Visualization of social media: Seeing a mirage or a message? Cartogr. Geogr. Inf. Sci. 40, 55–60 (2013). https://doi.org/10.1080/15230406.2013.776754
36. Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., Thom, D.: Thematic patterns in georeferenced tweets through space-time visual analytics. Comput. Sci. Eng. 15, 72–82 (2013). https://doi.org/10.1109/MCSE.2013.70
37. Truelove, M., Vasardani, M., Winter, S.: Towards credibility of micro-blogs: characterising witness accounts. GeoJournal. 80, 339–359 (2015). https://doi.org/10.1007/s10708-014-9556-8
38. Lansley, G., Longley, P.A.: The geography of Twitter topics in London. Comput. Environ. Urban Syst. 58, 85–96 (2016). https://doi.org/10.1016/j.compenvurbsys.2016.04.002
39. Longley, P.A., Adnan, M.: Geo-temporal Twitter demographics. Int. J. Geogr. Inf. Sci. 30, 369–389 (2016). https://doi.org/10.1080/13658816.2015.1089441
40. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. Eur. J. Inf. Syst. 22, 336–359 (2013). https://doi.org/10.1057/ejis.2012.26
41. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The Rise of Social Bots. Commun. ACM. 59, 96–104 (2016). https://doi.org/10.1145/2818717
42. Schäfers, B. ed: Einführung in die Gruppensoziologie. Geschichte, Theorien, Analysen. 3., korr. Aufl. (1999)
43. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media performance and application considerations. Data Min. Knowl. Discov. 24, 515–554 (2012). https://doi.org/10.1007/s10618-011-0224-z
44. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One. 9, 1–12 (2014). https://doi.org/10.1371/journal.pone.0098679
45. Hauthal, E., Burghardt, D.: Extraction of Location-Based Emotions from Photo Platforms. Prog. Locat. Serv. 49, 3–28 (2013). https://doi.org/10.1007/978-3-642-34203-5
46. Hu, B., Song, Z., Ester, M.: Topic Modeling in Online Social Media, User Features, and Social Networks for. In: Alhajj, R. and Rokne, J. (eds.) Encyclopedia of Social

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.

Network Analysis and Mining. pp. 2178–2191. Springer New York, New York, NY (2014)

47. Croitoru, A., Crooks, A., Radzikowski, J., Stefanidis, A.: Geosocial gauge: a system prototype for knowledge discovery from social media. Int. J. Geogr. Inf. Sci. 27, 2483–2508 (2013). https://doi.org/10.1080/13658816.2013.825724
48. Dunkel, A.: Visualizing the perceived environment using crowdsourced photo geodata. Landsc. Urban Plan. 142, 173–186 (2015). https://doi.org/10.1016/j.landurbplan.2015.02.022
49. Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., Almeida, V.: Beware of what you share: Inferring home location in social networks. Proc. - 12th IEEE Int. Conf. Data Min. Work. ICDMW 2012. 571–578 (2012). https://doi.org/10.1109/ICDMW.2012.106
50. Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J.: #Earthquake: Twitter as a Distributed Sensor System. Trans. GIS. 17, 124–147 (2013). https://doi.org/10.1111/j.1467-9671.2012.01359.x
51. Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W., Zook, M.: Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartogr. Geogr. Inf. Sci. 40, 130–139 (2013). https://doi.org/10.1080/15230406.2013.777137

Corrigendum to AGILE: GIScience Series, 1, 5, 2020. https://doi.org/10.5194/agile-giss-1-5-2020
Full paper Proceedings of the 23rd AGILE Conference on Geographic Information Science, 2020.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras
This contribution underwent peer review based on a full paper submission.
https://doi.org/10.5194/agile-giss-1-5-2020-corrigendum | © Authors 2020. CC BY 4.0 License.