Imagined territorial communities in social media: A language-based approach on Twitter

Sara Colella, Andreas Jungherr

11th October 2019

1 Introduction

Investigating the level of attention paid to each topic by a particular population has been made recently possible by the interactivity of social media. In fact, any online activity (e.g. sharing information, comments and opinions) leaves a digital trace that partially reveals the audience interests. Therefore, social media platforms are a big opportunity for researchers to study people's attention dynamics that would otherwise remain hidden. Nevertheless, to perform these analyses, we often need to select users who share the same cultural proximity, a territorial community, keeping in mind that the Internet revolution faded the impact of geographic closeness on the sense of belonging.

In this paper we will focus on territorial communities and how to identify their members in social media. We will identify members of a territorial community focusing on what "glues" them together creating ties of cohesion (for example language), following the imaginary communities approach (Anderson, 1991). Selecting community members is an open problem mainly because it is based on the subjective sense of belonging of the members. Furthermore, the unifying factors can be several and not of easy identification. In fact, even if all the current sociological literature agrees that the unifying factor of a territorial community is not necessarily interaction or spatial proximity, there is not complete agreement on its definition. The cohesive factors may be common values, traditions, interests, language or religion regardless of the territorial dispersion of the community members (Newby and Bell, 1974).

According to Elias (2008), a place stops to be characterized as a community when the mutual interdependence among the members is so feeble that they stop to be involved in local gossip remaining indifferent to any form of community control. This means that members of the same community do not need to interact directly, but they need to have a common area of discussion since they are part of the same cultural space which is mainly formed by "Little-c" culture (i.e. what people talk about at the hairdresser or on the bus). As the paradigms from choral movels in (Anderson, 1991) show, shared knowledge and habits may be the basis for the sense of belonging. In other words, considering common cultural space as cohesive criteria, territorial communities are composed by people with any interest about the selected geographical area.

Since people belonging to the same cultural community often have shared interests due to their common cultural background, cultural proximity has a deep impact also on social media due to the high range of topics shared online. Therefore, for example in attention dynamics research, it is often necessary to select an audience belonging to a common cultural background to study what are the motives behind the audience shifts of focus.

Since community membership has a deep impact on people's attitude toward themselves and others, its investigation presents several applications. First of all, belonging to different territorial communities might have a consequence on attention dynamics, testified also by the presence of the concept of proximity among the criteria that determine news value for the audience (Shoemaker and Reese, 1996).

Belonging to a territorial community has also an impact on the choice of cultural products. Its effect was firstly noticed by Straubhaar (1991) about television genres. As soon as national programs started to be available, South American people "seem to prefer nationally or locally produced material that is closer to and more reinforcing of traditional identities, based in regional, ethnic, dialect/language, religious, and other elements" (Straubhaar, 1991, p.51).

Even if the existing techniques claim to select territorial communities on social media, they map only sections of communities that satisfy also additional requirements such as spatial proximity (e.g. (Schulz et al., 2013; Bruns et al., 2014)). Nevertheless, the existence of categories of territorial community members beyond spatial proximity is well known in sociology. For example children of immigrants often have a lower national identification with the country of residence than the natives (Leszczensky et al., 2019). This phenomenon is partially due to a certain sense of belonging to their parents' country of origin, making them part of such territorial community, even if their connection is limited to only certain aspects such as religion and language (Fleischmann and Maykel, 2016; Maykel and Martinovic, 2012).

Instead, other existing literature focuses on identifying communities through the level of interaction among the members (van Meeteren et al., 2010). However, not in all communities members do interact directly. For example, in cultural or territorial communities most members do not contact (or even know) each other, but they still share a sense of belonging and they react to the same stimuli (e.g. religious communities). Therefore, only focusing on direct interaction ties to identify territorial communities would lead to considerably underestimate their size.

Hence, due to the multifaceted nature of the actual sociological concept of community, its identification presents several challenges. In this paper we propose a technique to select members of the same territorial community on social media; groups with geographic interests regardless of their territorial dispersion (Newby and Bell, 1974).

Therefore, due to the limitations of the existing literature, we propose a novel methodology to identify territorial communities from a sociological point of view. We select, for the first time, social media users not considering only their location, but also their territorial interests through a multi-signal approach. Users whose individual cultural space intersects the one of the geographical area of interest. They are selected through the connection of any selection criteria with the selected territory. This prospective leads to a broader concept of community taking into account users that would have not been selected according to the existing literature procedures, but that belong to the local cultural space. Such as tourists, emigrants, children of immigrants, fans of the culture and even supporters of the local football teams.

Furthermore, these users are not only members of a group with shared interests, but part of a community since they exhibit their connection with the geographical area of interest on the platform through several signals, such as the language of their posts, showing their own public identity (Hogan, 2010).

We will start presenting the theory on which our approach is based (Sec. 2), then we will compare it with the existing methods (Sec. 3). After having implemented our methodology on Twitter for a geographical area chosen as study case and tested its validity with some sanity checks (Sec. 4), we will conclude by underlining its strengths and limitations sketching avenues for further research and applications (Sec. 5).

2 Theory

The concept of community is open to wide interpretation and often used as reference point in multifaceted discourses about identity in several fields, from psychology to anthropology (Cohen, 1985). In fact social identities are constituted by people's overlapping, competing and conflicting memberships in several communities and community membership has a deep impact on people's attitude toward themselves and the others.

In sociology the concept of community has a long history characterized by many variations in its conceptualization also because of the political and social fragmentation that occurred in time. According to the recent literature (Anderson, 1991), communities are considered imaginary, specific and not exclusive. Based on "a relational idea: the opposition of one community to others or to other social entities" (Cohen, 1985, p.12), communities are defined through symbolic boundaries. Community are based on what "glue" the members together differentiating them from the others, the people outside the community.

Even if early sociological debates stressed community territorial base (e.g. groups whose members share a well defined location (Hawley, 1950)), the later concept of community is rather an imaginary state of mind. A boundary-expressing symbol as postulated by Cohen (1985) defined by what

differentiate the community members from the others. The recent definition of territorial community goes beyond simply spatial proximity, but there is not consensus on the nature of the unifying factor among the members that leads to their sense of belonging. Furthermore, even if the chance of perspective on the definition of territorial communities started before the digital revolution, the developments in technology further reduced the impact of the geographical component on the sense of belonging. The cyberspace has undetermined whatever geographical basis communities have ever had, blurring its geographical demarcation lines.

According to Elias (2008), the members of the same community do not need to interact directly, but they need to have a common area of discussion, to be part of the same cultural space. As Anderson (1991)'s paradigms from choral novels show, shared knowledge and habits may be the basis for the sense of belonging. Discussing a dinner party by hundreds of unnamed inhabitants of Manila conjures up the image of a community.

However, the concept of common knowledge may change according to the geographical area under examination; only people in Manila know about that specific dinner party.

For a more formal expression of this concept of community, let us consider the cultural space of each category of people as a set. Following Elias (2008)'s approach, we considered members of a territorial community anyone whose cultural space intersects the cultural space regarding the geographical area of interest. For example, also the foreigners living in Manila may know about the dinner party having therefore a cultural space which intersects the Philippine one. Nevertheless their cultural space is not a subset since it intersects also the one of the foreigners' country of origin. In turn, the citizens of Manila abroad may have a cultural space that intersects also the one of their place of emigration and their level of intersection would depend on their level of integration (Berry, 2011). Considering intersecting communities (not disjoint sets) with the cultural space of the geographical area of interest has the advantage to include in the selection also who has personal interests that range beyond their territory (for example USA citizen who like Filipino pop music). Finally, let us also consider that the intersection of the cultural spaces does not have to include all the aspects of the local cultural area. For example fans of Manila football team may be interested not in all the Philippine news, but only in the ones about football.

Following (Elias, 2008)'s approach, we defined a territorial community as composed by anyone who has personal interests that partially intersect any news about the geographical area of interest. Therefore, studying people's interests is essential to establish territorial community membership. Social media are a very suitable tool to perform this analysis due to the available digital traces left by the user behaviour online.

As stated in the introduction, the Internet fated the impact of geographic closeness. Hence, our approach to focus on the cohesive factors of territorial factors can be applied to several social media platform. We decided to implement it for a specific platform: Twitter.

Twitter defines itself as a tool to describe *"what's happening in the world and what people are talking about right now"*[1]. Therefore its focus is on dynamics that occur outside the platform and especially on people's reactions towards them. In particular Twitter is a suitable tool to observe the users' reaction to news since, according to a survey from the Pew Research Center[2], it is on percentage the platform most used by the American population to get informed. Hence, Twitter is a suitable tool to implement our approach to identify members of a territorial community according to their interest in news about the geographical area of interest.

Since we want to select any Twitter user whose cultural space intersects the one of the geographical area of interest and the territorial community does not have to include all the aspects of the local cultural area, but only any part of it, we propose to select any user with at least one selection criteria connected with the geographical area of interest. Similarly to the cultural spaces (Sec. 1), we can consider all the Twitter accounts that satisfy each selection criteria as a set. The presence of discordant indicators does not exclude the users from the sample as we are not considering the intersection of multiple signals, but rather their union. In this way we are considering also, but not just, the residents, a category selected by the existing techniques (Sec. 3), and also subgroups of Twitter users that would have not been selected otherwise (e.g descendants of emigrants or immigrants).

[1] https://about.twitter.com/en_us.html
[2] www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017

3 Existing methods: State of the art of Twitter user selection according to their geolocation

Several authors have developed techniques to identify Twitter users according to their residence. The high level of attention given to this matter is mainly due to the importance of users' sense of closeness to a certain territory, as we argued before, to determine personal interests. This sense of closeness can be objective and be location based (as Facebook's chairman stated:*"A squirrel dying in front of your house may be more relevant to your interests right now than people dying in Africa"*), or subjective and depending on a personal feeling of connection to a territory through cultural ties. If the news is personally relevant to an individual, it will be more likely to discuss it with others, as asserted by Basil and Brown (1994) studying the spread of the news of "Magic" Johnson's positive HIV test. The fans of the American basketball player Earvin "Magic" Johnson were more likely to spread the news of this infection with the HIV virus rather than the non-fans.

The existing approaches in literature to perform community detection on social media are based on both geolocation and frequent interactions. These are interesting, but only optional aspects that are not always present among the members of a territorial community (Sec. 2). To the best of our knowledge, there are no existing techniques to identify territorial communities disregarding members' interactions and their spatial proximity on Twitter. In this section we are going to investigate the state of art of the main procedures of the Twitter user selection according to their location and compare them to our alternative approach.

Beside the conceptual limitations, identifying social network users' localization presents some challenges, especially for platforms that do not require such information to open an account, as Twitter. Users can provide the geolocalization of their account on a voluntary basis, and, according to Hecht et al. (2011) only 66% of the users provide any kind of geographic information (counting also very general indications such as "Asia"). These issues aside, processing this information presents some challenges because of the presence of ambiguous locations or fake information (such as "flying" or "anywhere but here"). An interpretation of this field, its coordinates, is provided by Twitter Firehose, the paid version of the Twitter Streaming API data collection. Nevertheless the obtained geolocation, since it is not obtained on additional information provided by the user, may not be accurate for ambiguous names (beside Paris in France there are more than 20 other Parises in the USA). Therefore, due to the challenges of correctly geolocating the Twitter user profile, the several methodologies have been developed to determine Twitter users' residence.

3.1 Snowball approach

To the best of our knowledge, the first attempt to select the entire Twitter population related to a certain nation was implemented by Bruns et al. (2014), who developed a methodology to collect the Australian Twittersphere. This user selection approach is a snowball selection that chose as seeds Twitter accounts that set an Australian time zone for their Twitter profile and that employed hashtags related to particular Australian topics (e.g. $\#ausvotes$ for the 2010 Australian election or $\#qldfloods$ for the 2011 floods in Queensland). Therefore, the selected users were most likely in Australia when they registered their Twitter account (till 25/5/18 Twitter suggested the time zone to insert according to the IP addess) and interested in territorial topics.

Bruns et al. (2014) selected also the followers and followees of the seeds that satisfied the time zone criteria on the assumption that *"most Australian Twitter users are more likely to connect with other Australians than with other international accounts"*(Bruns et al., 2014, p.5). Several works support this hypothesis showing that the frequency of people's contacts on social media, versus the distance between the contacts' location and their own, decays following a power law distribution. This had been shown to be true for several platforms, such as for Facebook friends (Backstrom et al., 2010) and partially for Twitter followers and mentions (McGee et al., 2011). The spatial proximity of friends is not limited to social media platforms, but it had a significant impact also before the digital revolution (Mok and Wellman, 2007). Nevertheless, even if geographical distance has an impact on users' level of interaction making plausible the initial assumption, it may be independent from membership to territorial communities.

Beside the non-generalizability of the initial assumption to identify territorial communities, the user selection approach developed by Bruns et al. (2014) has several downsides. First of all the thematic focus of the selected hashtags; in case of hashtags with a political focus (e.g. $\#ausvotes$, $\#auspol$,...) the snowballing process will be likely to find accounts with the same focus rather than with an interest in any kind of territorial topic. Even if the attention paid to the discussed topics is a very interesting aspect, overcoming this problem is not trivial since it would be necessary to find hashtags relating to any Australian theme.
Finally not every tweet contains hashtags (from a maximum of 18% of German tweets to a minimum of 5% of Japanese ones (Hong et al., 2011)) creating problems in the seed selection.

3.2 Social network-based geo-inference

To overcome the problems regarding the small number of Twitter posts associated with a geolocation (0.85% of the total Twitter posts (Hecht et al., 2011)), recent work has focused on geo-inference using social networks to predict the location of posts. This class of methods consists in the recent practice and it is based on the assumption that the locations of users' relationships is evidence of the location of the user and of its posts. This hypothesis has been introduced in Sec. 3.1, but there are essential differences between the snowball approach and the social network-based geo-inference. The first methodology assumes that the majority of your contacts has your same location, the second one reverses the concept: the nationality of your contacts can be used to detect your location.

However, being part of a territorial communities do not imply interaction. In other words, the geographical location of the people known by a person is not a sufficient condition for its belonging to a territorial community. Knowing a Manila citizen is not enough for a Hungarian to be part of the cultural space of Manila; this would be true only if they would discuss topics regarding Manila.

Furthermore, Twitter social relationships have several roles beyond signifying friendships. Therefore several geolocation inference algorithms have different definitions of what they consider a relationship (e.g. followers, mentions or reciprocal followers) to create social networks. To identify territorial communities it may make sense not to consider all the contacts of the Twitter users, but only the ones regarding news agencies, politicians or any topic regarding the geographical area under interest. In fact, this may show a territorial interest of the users. Nevertheless, it is not trivial to detect accounts considering local topics about any possible theme (a similar problem to the one discussed in Sec. 3.1).

In addition, social network-based geo-inference use some ground truth data as inference. The problem is how the choice of this data sample is conducted. Usually the ground truth data is considered to be users' self reported location or GPS-based location of the tweets, that gives better performances (Jurgens et al., 2015)). Therefore, these selection criteria choose a subpopulation of users not representative of the complete Twitter user sample (Sloan and Morgan, 2015). Finally, there is also the temporal impact to consider in training these methods. Geo-inference performance uniformly declines in time and its half life is roughly four months. This suggests the need for testing periodically the data to update the effectiveness of the ground truth data.

3.3 Multi-Signal Intersection

Han et al. (2014) showed that tweets' language has a strong influence on geolocation prediction, especially for geographically-focused languages, however integrating several approaches can significantly improve the performance prediction. Following this suggestion, Schulz et al. (2013) employed a multi-signal method to determine both the tweet location and the user's residence.

To detect user residence, Schulz et al. (2013) considered the area determined by different signals: "Location" field, language of the tweet (German since they were selecting residents in German cities), tweets' geolocation and the website's country code as well as the geolocated IP adresses for each shared URL. Each indicator defines an area from which the post can be sent from. Through the intersection of these areas, Schulz et al. (2013) determined the geolocation of the posts. This approach was used to determine both the location of each post and the user's residence, defined as the predicted location of the last post sent by the user.

One of the downsides of this technique is that the user location is defined by the geolocation of only one tweet of the user ignoring its other messages. In addition, in case of discordant signals the user is excluded from the selection not considering that in this way false negative accounts may be excluded as well. It may happen that the geolocated tweets refer to a particular occasion, such as a trip abroad, not being therefore a valid indicator for users' residence.

Finally a user may relocate over the course of its social history making possible the presence of several home locations keeping territorial interests towards each of them. Hence, we reversed this approach (that selects only the locals) and considered any user that had a connection with the geographical area of interest considering the union of multiple signals.

4 Our Approach: Multi-Signal Union for cultural connection

Beside the several difficulties in identifying users' nationality on Twitter, we are not only interested in identify users with a certain home location, but more broadly to their connection with the cultural space of a specific geographical area. Since we consider the cultural identity of the users, we developed an alternative methodology to select users that may belong to a geographic community. To the best of our knowledge, this technique is the first one to select users according to their cultural connection with a certain territory that may be motivated by several reasons, such as residence, family's origins or business reasons.

To implement our novel methodology to identify territorial communities, we tested it on the Italian study case. This choice was driven by several factors, mainly because of the geographically-focused language, the large emigration and the strong connection ties. The importance of the first factor is driven by the selection criteria employed by our methodology, whose majority is language oriented, therefore this approach has the limitation to work only for geographically-focused languages such as German and Italian. The importance of a common language in the process of creation of the concept of community belonging has been observed by Anderson (1991), even highlighting that this condition is not necessary. In case of the Italian language, it is "a vehicle of transmission of an original culture" (Fortier, 1998, p.208). Quite all the community of Italian speakers is in Italy, that experienced a serious voluntary emigration that involved over 5 millions of registered Italians abroad[3]. Furthermore, Fortier (1998) estimated that the majority of people with "Italian origins" (65 million) lives outside the county of origin because of the long history of mass emigration. An additional factor is that the descendants of Italian emigrants, even several generation apart, tend to keep strong connection ties with their home country (e.g. through traditions, culinary culture and religion) (Fortier, 1998), but not a particular high level of nationalism due their complicated relationship to national identity (Mose and Shive, 2011). These citizens (as well as supporters of Italian teams or tourists enthusiastic of ancient Rome history), even keeping on being part of the Italian cultural space, would have not been selected through the existing methods in literature. Therefore, due to their high number, the selection of the Italian community presents an optimal study case to implement our technique that can, nevertheless, be implemented for any other territorial communities with geographically-focused languages.

[3] https://www.esteri.it/mae/en/servizi/italiani-all-estero

Even if affiliation and identification may be unclear even to the actors themselves that leave several identifying signals unintentionally, their membership to communities shape their own individual social identities and therefore their behaviour. Therefore, we propose a multi-signal methodology, to which we will refer as *multi-signal union*, considering several indicators:

- *language of the user interface:* the choice of the language of the Twitter interface is compulsory and always present.
- *time zone of the user:* the time zone the user declares himself in setting its Twitter account. This information is not available anymore because of changes to the developer platform occurred on 25/05/18.
- *language of the tweet:* the machine-detected language identified in the text by Twitter[4][5]. In case of retweets the detected language regards union of the text of both the added words of comment and of the original message.
- *location associated to the tweet:* if present, it indicates that the tweet is associated to (but not necessarily originated from) a place by the user.

As highlighted before, half of these criteria are language oriented, thus this approach has the limitation of not working properly for geographically-diverse languages such as English and Spanish. It can instead be employed for geographically-focused languages such as Dutch and Italian (the study case). Furthermore, considering Italian as a geographically-focused language may be reductive. Even if the largest community of Italian speakers is in Italy (over 65 millions of inhabitants), Italian is also an official language of other countries: San Marino (32,448 of inhabitants), Vatican City (829 of inhabitants), Switzerland (over 600,000 of Italian speakers, 8.3% of the population) and Western Istria (around 2200 Italian speakers). Nevertheless, beside Slovenia, these minorities have access to Italian mass media (e.g. Italian TV stations), therefore we considered possible for them to be part of the Italian cultural space.

4.1 Implementation

The *multi-signal union* consists of 5 steps:

1. **Baseline: Internet Archive**

   Users are selected from a downloadable collection of tweets made available from the Internet Archive[6]. It consists of the "spritzer" stream, around 1% of all the worldwide Twitter posts, and the related metadata[7] from September 2011 till nowadays. The Internet Archive is a nonprofit digital library that collects and makes available a digital library of Internet sites and other cultural artifacts in digital form as historical sources. The importance of Internet data, Twitter in particular, as a historical sorce for our "collective memory" was well reconceptualized by Bruns and Weller (2016). Also Sequiera and Lin (2017) recognized the importance of the Twitter data collection provided by the Internet Archive, but rather from a methodological point of view since it provides the research community with a downloadable test collection of tweets.

   The main advantage of using the Twitter data collection of the Internet Archive as baseline for the user selection is to avoid to choose an arbitrary time period to perform streaming data collection and then to filter the users according to the selection parameters. Almost all the existing methods in literature (Sec. 3) selected their sample of users from arbitrary periods of data collection through the Twitter Streaming API. The only exception is Bruns et al. (2017) that considered all the existing Twitter users.

   Even if the Twitter data collection of the Internet Archive is very comprehensive, the data regarding some time intervals is missing (especially in 2014 and 2015). Another limitation is that, since the Twitter data collection of the Internet Archive provides only around 1% of the published posts, our user selection is pushed to the most active users. The more active the users are, the higher is the probability that their posts will be collected by the "spritzer" stream.

[4] https://tools.ietf.org/html/bcp47
[5] https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/premium-operators
[6] https://archive.org/
[7] https://developer.twitter.com/en/docs/basics/things-every-developer-should-know

Since we started the data collection on 1/01/18, we employed all the data of the Internet Archive downloadable at that moment: from September 2011 to June 2017 due the delay between the data collection and its publication.

2. **First selection: User selection from Internet Archive**

   Among all the users whose tweets are contained in the Internet Archive's collection, we selected only the users that satisfy at least one of the selection parameters for the geographical area under consideration.

   Some of the selection parameters have been employed also by some of the existing techniques to select Twitter users according to their residence (Sec. 3). The time zone of the Twitter account has been employed by both the snowball approach (Sec. 3.1) and the multi-signal intersection (Sec. 3.3). Instead the location associated with the tweet by the user is usually employed as ground truth, it is the one that gives the best prediction of users' location, in the social network-based approaches (Sec. 3.2) and in the multi-signal intersection (Sec. 3.3). Finally, the language oriented criteria (language of the user interface and of the tweet) are employed only by Han et al. (2014) and partially by the multi-signal intersection (Sec. 3.3). Mainly because the other techniques focus on geographically-diverse languages (e.g. the Australian Twittersphere in Sec. 3.1) or on the location of users' residence at a city level.

   Other parameters that could have been useful to complete the user profile are the user Profile Geo and the self-reported user location. The Profile Geo[8] is a field available only in the enriched version of the tweets (contained in Twitter Firehose[9]), therefore not collected by the Internet Archive. If activated, it provides an interpretation for the geographic place described in the profile location string. The self-reported user location is the user-defined location for the account and its interpretation presents several challenges (Han et al., 2014). The performance of the prediction of this indicator varies broadly according to the gazetteer (i.e. geographical dictionary) in use due to the not standardization of the field. In addition, since different language communities use Twitter in different ways, the accuracy of self-reported user location varies also according to their cultural differences (Hong et al., 2011). It seems reasonable to assume that different levels of patriotism may influence user self-reported location and description. Due to Italian low nationalism and identification with the state (characterized by a high level of parochialism) (Mose and Shive, 2011), we considered unlikely that the users would explicitly write Italy in the field under examination. Finally, beside the interpretation problems of the "Location" field, it refers only to the user-defined location, not to its community identity. However, we employed this field in the sanity checks (Sec. 4.3).

   For the study case, the selection of the Italian Twitter community, we selected roughly 10.8 million of users that satisfied at least one parameter. Due to the Twitter Developer Agreement[10], the Internet Archive should update its archives every time that an account or a post is deleted (Sequiera and Lin, 2017) and therefore should not contain any deleted account. Nevertheless, for safety, in the next step we filtered the users according their existence, activity and consistency of the selection parameters.

[8] https://developer.twitter.com/en/docs/tweets/enrichments/overview/profile-geo
[9] http://support.gnip.com/apis/firehose/overview.html
[10] https://developer.twitter.com/en/developer-terms/agreement-and-policy

3. **Existence and activity of the accounts**

After having selected the users from the dataset of the Internet Archive, we checked if the accounts still exist and if the selection criteria are still satisfied. We verified as well if the users are still active: if their last post occurred less than a month before the data collection. One month has been chosen as threshold mainly because, as we can observe in the last bar of Figure 1, the vast majority of the users that did not tweet in the last month did not tweet in the last year (65%). In addition, having tweeted at least once in the last month is one of the criteria used by Twitter to detect users that have an "active usage" of the platform[11].

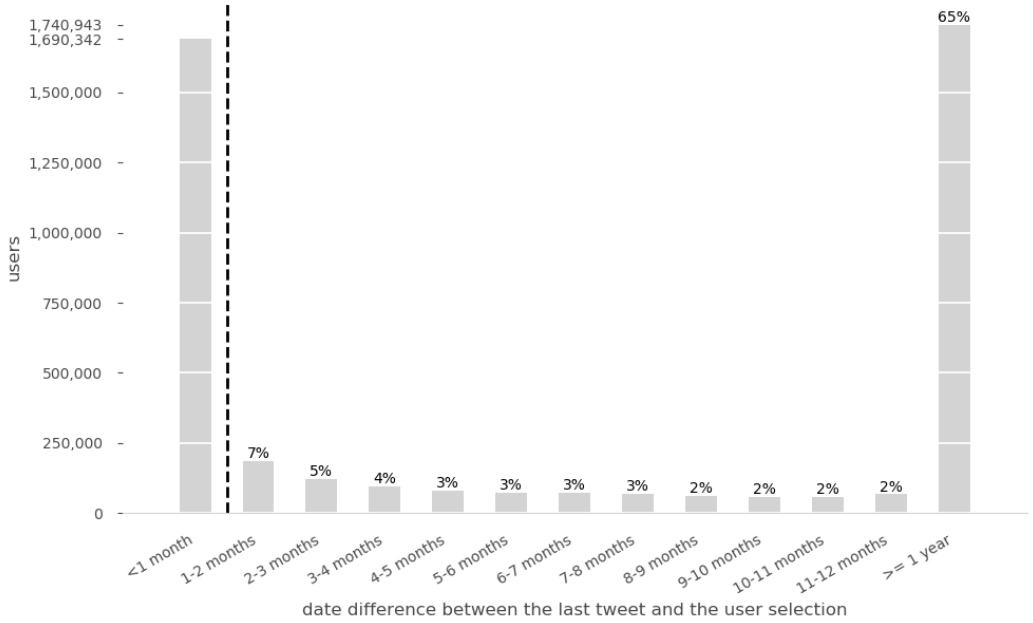

Figure 1: Frequency distribution of the users that satisfy the selection criteria according to the difference between the date of the last tweet and of the user selection. The bars to the right of the dashed line show the distribution of the excluded Twitter users with, for each month, their percentage over the excluded accounts.

To check these criteria we considered only the first page of the user archive (the last 200 tweets) to streamline the analysis.

For our study case the number of active accounts that satisfied the criteria was 1,698,460. To consider the impact of each month on the user selection, we plotted in Figure 2 how many new active users are added to the selected Twitter population going backwards in time. Figure 2 shows that going backwards in time allows to add less and less new users to the selection. This is due both to the abandoned accounts and to the presence of the posts of the same active users in several months (also the recent ones) of the dataset from the Internet Archive.

Since we performed the user selection in December 2018 with the data from the dataset of the Internet Archive till June 2017, we can observe that the highest impact (the highest bar) is given by the last month available in the archive. Till March 2018 we updated the user sample adding also the users obtained from the dataset of the Internet Archive till November 2017 (following the first 3 steps of the user selection). Because of the risk of not being able to obtain the complete collection of posts of the new users (2% in March according to the estimated frequency of the users' tweeting behavior) would have further increased going forward in time, we stopped updating our user sample.

[11] https://www.cbsnews.com/news/many-twitter-users-dont-tweet-finds-report/

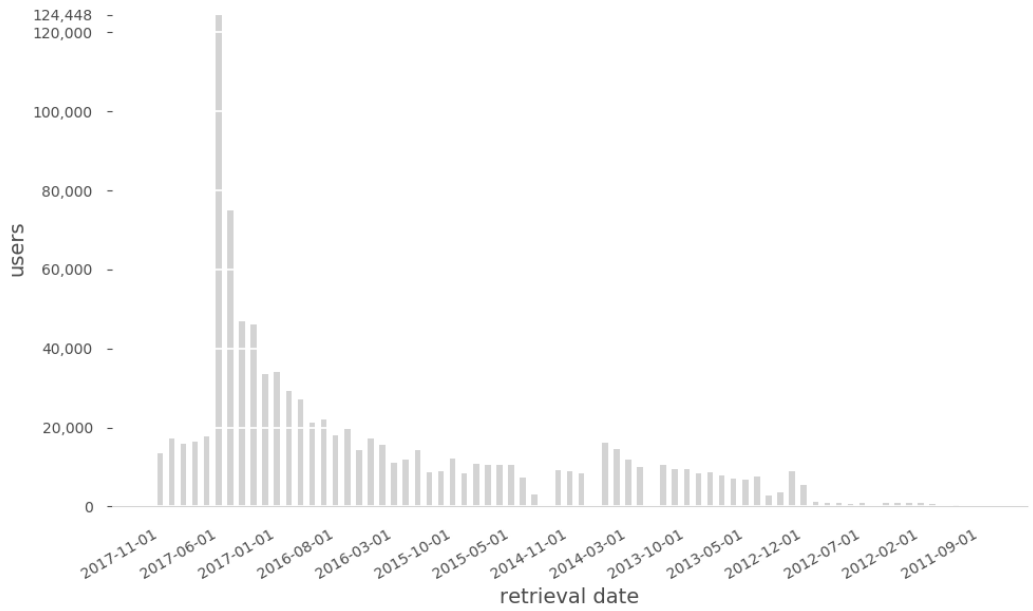

Figure 2: Number of the users added to the user selection for each month of the Internet archive, going backwards in time.

Let us now consider the impact of each selection parameter and their combination. Table 1 shows that more than half of the users was selected only because of the language of the published posts which may be a very noisy selection criteria. In fact, more than half of them (777,632) had a very low level of consistency in the use of the language: only 1 in 200 tweets was detected as Italian. This may be due to the ambiguous Twitter language recognition (e.g. Spanish and Portuguese are often labeled as Italian) and to the very large diffusion of Italian words (e.g. Ciao). To address this issue, we investigated the reliability of the detected tweet language.

Table 1: Impact of the selection criteria and their combination in absolute values and as percentage of the total selected users.

indicator	*users*	*percentage*
interface	13,707	8.07%
time zone	7,273	0.43%
tweet language	1,167,301	69.73%
associated location	6,220	0.37%
interface & tweet language	291,604	17.17%
interface & time zone	1,677	0.10%
interface & associated location	220	0.01%
tweet language & time zone	14,019	0,82%
tweet language & associated location	15,671	0,92%
time zone & associated location	391	0,02%
interface & tweet language & time zone	75,559	4.45%
tweet language & time zone & associated location	5,371	0.32%
interface & associated location & time zone	79	0.05‰
interface & associated location & tweet language	66,874	3.94%
interface & associated location & time zone & tweet language	32,584	1.92%

4. **Filter: consistence in the use of the language**

After having selected any user that tweeted in Italian at least once over its last 200 tweets, we investigated the ambiguity of the Twitter language recognition. Therefore, we performed a sanity check on the users selected only through the language criteria. We manually examined a random sample of Twitter accounts, for several values of consistency of the language in the tweets, to determine if they had any connection with the Italian cultural space. To do so we employed the criteria described in the following decision tree (Figure 5) and the uncertainty over the percentages has been computed considering maximum variability and a confidence level of 95%[12]. Figure 3 shows both the total number of users selected only through the language criteria for several values of consistency in the use of the language (in gray) and the percentage of users that, according to the sanity check criteria (Sec. 4.3), has been detected as part of the Italian community (in green).

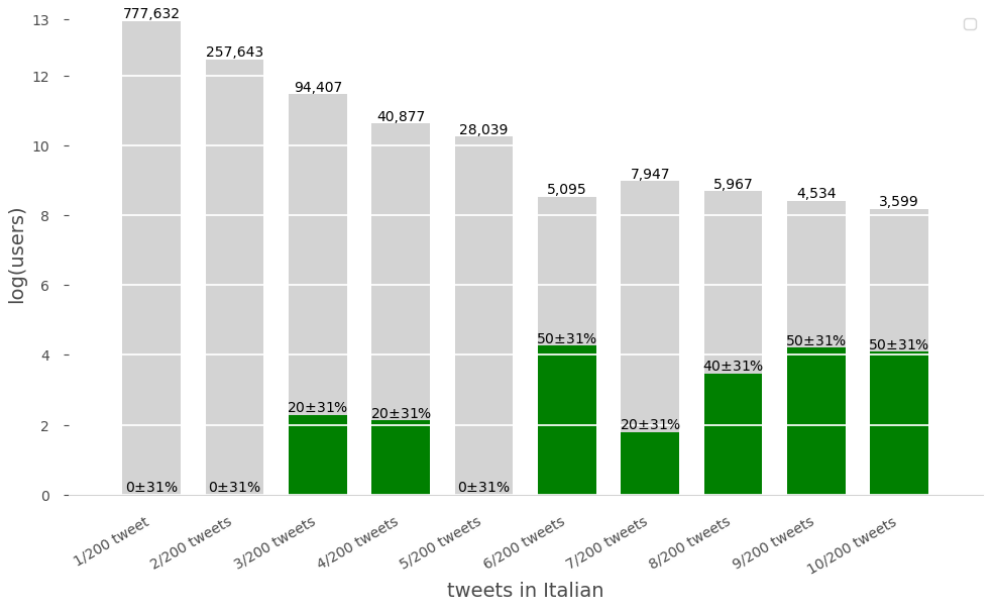

Figure 3: Number of users selected only through the language criteria for several values of consistency in the use of the language (in gray) and their percentage that has been detected as part of the Italian community (in green). The uncertainty over the percentages has been computed considering maximum variability and a confidence level of 95%[12].

As conclusion of this analysis, we decided to exclude from our selection the users that were not consistent in the use of the language of their posts below a certain threshold (3 tweets over 200) if that is the only satisfied criterion. In fact, that is the first level of language consistency that shows the presence of accounts intersecting the Italian community, even if lower than the confidence interval. The first level of language consistency that shows unmistakably the presence of users belonging to the Italian community is 3%. According to the analyzed sample cutting these users (905,009) meant not to exclude any user, with the uncertainty level shown in the plot (Figure 3), that is connected with the Italian cultural space. Therefore, the estimated maximum number of false negative accounts that was cut is 154,600.

This cut led to the selection of 944,996 users and, even if much less effective, the parameter with the highest impact was still the tweet language.

[12]The level of precision over the percentages has been computed employing Cochran's formula $e = \sqrt{\frac{Z^2 p(1-p)}{n_0}}$ (Israel, 1992) with n_0 as sample size, in this case only 10 individuals were tested due the cost of this sanity check (Sec. 4.3). Z is the abscissa of the normal curve and it depends on the desired confidence level (1.96 for 95%) and p is the estimated proportion of the attribute that we want to measure in the population. Since we do not have previous knowledge of p we assume massimum variability ($p = 0.5$).

5. **Filter: consistency of the selection criteria**

   Since we intended to perform a data collection one year long, we wanted to check if some of the users selected so far would have modified their connection with the Italian cultural space or if they would have stopped to be active on Twitter in that period. Therefore in 8/18 we run a robustness test checking if the selection criteria illustrated in the third and fourth steps (both for the activity level of the users on the platform and their connection with the territory) were still satisfied. This was not true anymore for 464,565 Twitter accounts showing a high variation in time of some users' use of Twitter and personal cultural space.

   Since, by the time of the sanity check, the time zone of the accounts was not provided anymore by Twitter, the robustness check could not be performed on 12,562 users that were selected only by this parameter. This is a very small number compared to the total size of the selected users (2.61%). Therefore the strength and replicability of this technique remains unvaried.

   Table 2 shows how, from the composition of the excluded accounts, this last cut affected mainly the users that were selected only through the language of their tweets (more than half of the excluded users). This seams reasonable, especially considering that the other criteria are mainly set during the profile setting. Nevertheless, we can observe that also a consistent number of users that were selected by the combination of several criteria were affected by the cut. This is mainly due to the accounts that are not active anymore on Twitter rather than their chance in their connection with the Italian cultural space.

Table 2: Impact of the selection criteria and their combination in absolute values and as percentage of the selected users who do not satisfy any more the selection criteria: both regarding their tweet frequency and the indicators regarding their connection to Italy.

indicator	*users*	*percentage*
interface	6,823	1.47%
tweet language	271,701	58.48%
associated location	4,066	0.87%
interface & tweet language	113,574	24.45%
interface & time zone	543	0.12%
interface & associated location	86	0.02%
tweet language & time zone	2,774	0.60%
tweet language & associated location	6,989	1.50%
time zone & associated location	139	0.03%
interface & tweet language & time zone	21,579	4.64%
tweet language & time zone & associated location	1,613	0.38%
interface & associated location & time zone	17	0.04‰
interface & associated location & tweet language	24,816	5.34%
interface & associated location & time zone & tweet language	9,845	2.12%

   As we can observe more in detail in Figure 4, the cut occurred mainly among users that did not tweet in Italian on a regular basis. In fact more than half of the cut users had less than 5% of their posts in Italian and, furthermore, this was the only satisfied selection criteria.

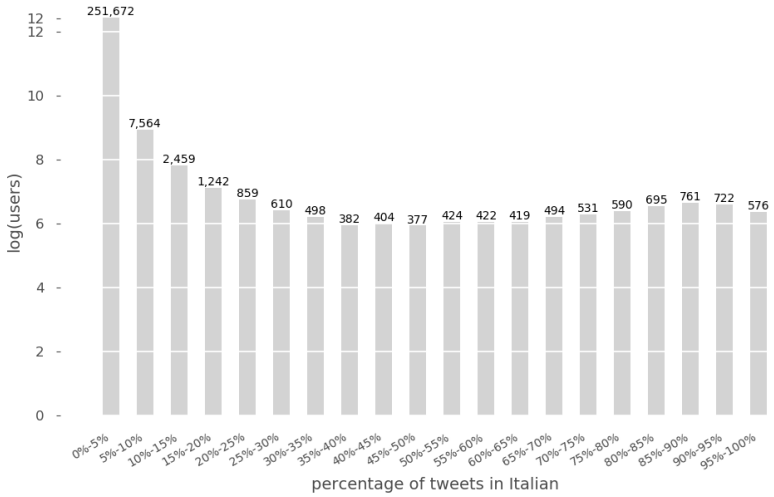

Figure 4: Users that do not satisfy any more the selection criteria with the language of their tweets as the only selection criteria for several percentages of consistency in the use of the Italian language in their tweets.

The actual population is formed by 480,431 active users that show their connection to the Italian cultural space satisfying the selection criteria consistently in time.

4.2 Results: population analysis

After having selected the population sample, Table 3 shows that the combination of the language of the tweets and of the account interface was the selection criteria with the highest impact. Compared with the selection steps (Sec. 4.1), where the most untactful selection criteria was always the language of the tweets, the final selected accounts have a stronger cultural connection with the territory in consideration. This is due to the fact that our approach led to a population in which the majority satisfies two selection criteria rather than only one.

Table 3: Impact of the single selection criteria and their combination in absolute values and as percentage of the 480,431 selected users.

indicator	*users*	*percentage*
interface	9,391	1.95%
time zone	7,758	1.61%
tweet language	101,173	21,06%
associated location	2,943	0.61%
interface & tweet language	206,740	43.03%
interface & time zone	1,231	0.25%
interface & associated location	168	0.03%
tweet language & time zone	11,790	2.45%
tweet language & associated location	10,108	2.10%
time zone & associated location	176	0.04%
interface & tweet language & time zone	56,222	11.70%
tweet language & time zone & associated location	3,850	0.80%
interface & associated location & time zone	65	0.01%
interface & associated location & tweet language	45,505	9.47%
interface & associated location & time zone & tweet language	23,311	4.85%

Table 4 shows that nearly all the users that have an Italian interface tweet in Italian (97%), but the other way around occurs in a less prominent way (72%). This is partially due to the international companies that have an English user interface and market their products in Italian, but also to the foreigners whose cultural space intersects the Italian one (such as fans of Italian teams who retweet news from the official account, thus in Italian) without knowing the language well enough to choose it as interface.

Table 4: Overlap of the selection criteria in absolute values and as percentage of the total users. In the second table the percentages show the rate of the accounts identified by each signal (in the rows) that satisfy also other criteria (in the column).

480,431	**interface**	**time zone**	**tweet language**	**associated location**
interface	342,633	80,829	331,778	69,049
time zone		104,403	95,173	27,402
tweet language			458,699	82,774
associated location				86,126

	interface	**time zone**	**tweet language**	**associated location**
interface	**71.32%**	23.59%	96.83%	20.15%
time zone	77.42%	**21.73%**	91.16%	26.25%
tweet language	72.33%	20.75%	**95.48%**	18.04%
associated location	80.17%	31.82%	96.11%	**17.93%**

Likewise, almost all the users that have selected an Italian time zone tweet in Italian and their majority (77%) has an Italian interface, but not the other way around. It seams reasonable to assume that the majority of the users who set their Twitter account in Italy are locals proficient in the language (that tweet in Italian with an Italian Interface). Therefore, since nearly all the users selected through the time zone parameter satisfy also other indicators (only 1.61% of the total population was selected only through this signal) and that this information is not available anymore since 25/05/18, the replicability of the methodology remains unvaried. Repeating the *multi-signal union* approach nowadays for the Italian territorial community would lead to almost the same population size, due to the changes to the developer platform only less than 3% of the users would not be able to be selected.

Finally nearly all the users that associated Italy to their tweets tweet in Italian (90%) and have an Italian interface (80%). Therefore, the overlap of the selection criteria suggests that the majority of users that tweet about Italy (to associate tweets to a location usually indicates that also the topic of the tweets is related) has a strong connection to the country having two other matching selection criteria. In conclusion, most of the users that tweet about Italy are robustly linked with it, but not not necessarily the other way around.

4.3 Sanity checks

To verify if among the members of the selected population there are not only residents in Italy, but also Twitter users who have a territorial interest to the country regardless of their territorial dispersion, we performed an incomplete evaluation of the composition of the selected population. Therefore, we performed a manual analysis following the steps described in the following decision tree (Figure 5) on random samples of users selected among the ones that satisfy each of the selection criteria and their combination. These steps were performed analyzing features of the Twitter accounts that where not considered as selection criteria, but that were taken into consideration by other methods presented in the literature (Sec. 3).

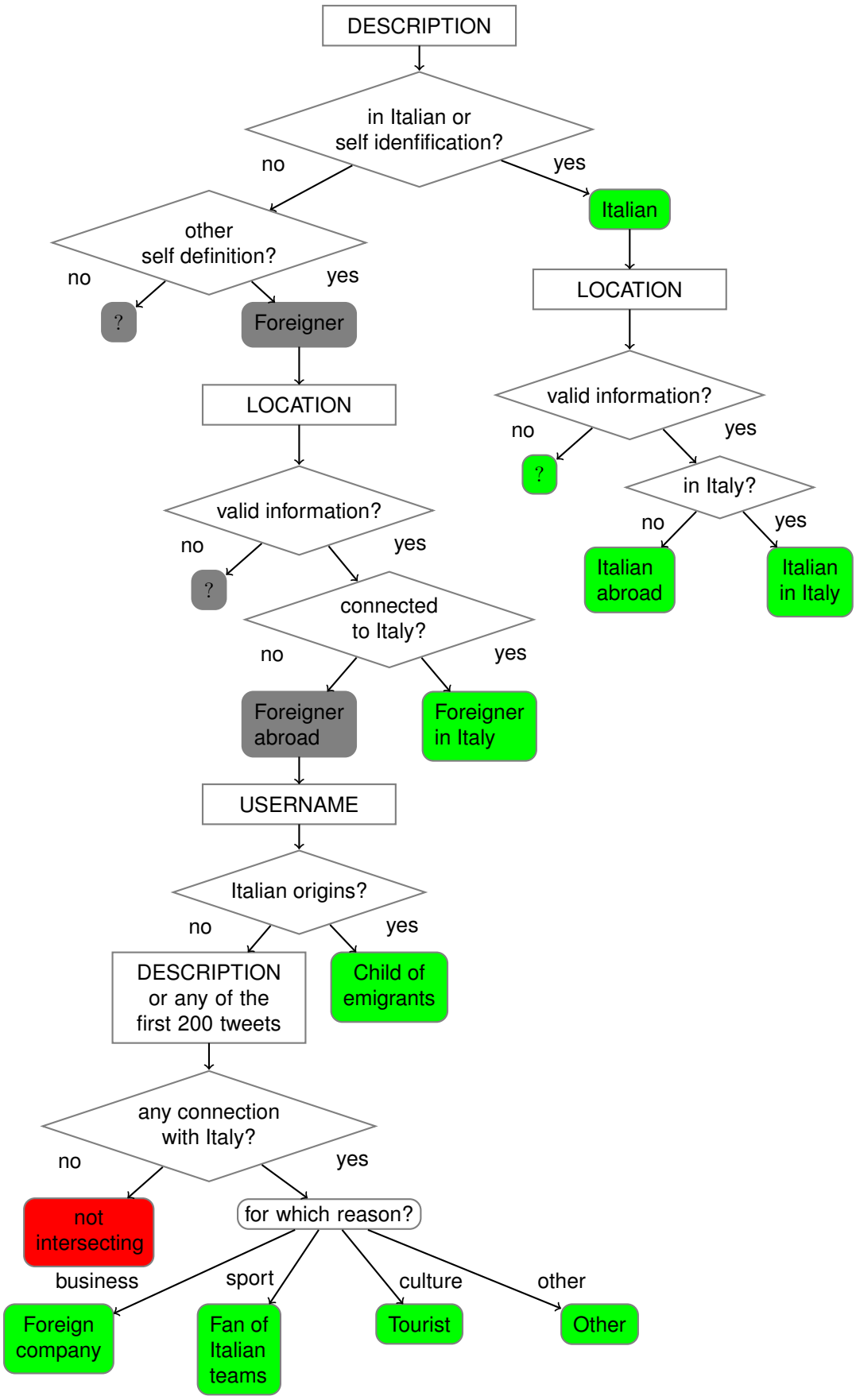

Figure 5: Decision tree to determine if the Twitter account under examination is intersecting the Italian cultural space (green outcome), exclude it (red) or give an inconclusive result (grey).

Due to the highly time consuming manual evaluation of the composition of the selected population, we could analyze only a small sample of Twitter accounts. Nevertheless the task of this analysis was to only determine the presence of Twitter users who have a territorial interest to the Italian culture, without being part of the residents. In fact, if this class of users had not been present, there would not have been any reason to have developed an alternative approach to the ones present in the literature.

Finally, this incomplete evaluation of the composition of the selected population does not only lead to the conclusion if the user belongs to the territorial community, but also to which intersecting category it belongs. Each category has a different level of intersection with the Italian cultural space and, according to the different levels of concordance with the selection criteria (Tab. 5), we can observe seven different categories of users.

Table 5: The percentage of the accounts whose cultural space intersects the Italian one over the total sample size is displayed in the first column. Its composition is illustrated in the remaining columns and the uncertainty over the percentages has been computed considering maximum variability and a confidence level of 95%[13]. The composition of the not intersecting accounts has been omitted from the table.

indicator	*intersecting accounts*	*Italians in Italy*	*Italians abroad*	*Children of emigrants*	*Foreign companies*	*Immigrants*	*Others*
interface (int)	70%	40%	10%	10%	10%		
time zone (tz)	30%	10%			10%	10%	
tweet language (tw)	50%	20%	20%				10%[a]
associated location (loc)	50%		10%	10%		10%	20%[b]
int & tw	90%	90%					
int & tz	70%	70%					
int & loc	60%	20%	20%			20%	
tw & tz	90%	90%					
tw & loc	80%	30%	20%		30%		
tz & loc	8%	30%	10%		10%	30%	
int & tw & tz	100%	100%					
tw & tz & loc	100%	90%	10%				
int & loc & tz	70%	60%				10%	
int & loc & tw	90%	90%					
int&loc&tz&tw	100%	100%					

[a] Fans of Italian teams
[b] Tourists

Table 5 displays how, as expected, if the number of the satisfied selection criteria increases also the percentage of Twitter accounts whose cultural space intersects the Italian one raises. This occurs especially if the language of the tweets is among the criteria showing its high impact as selection criteria.

Furthermore, according to the different level of concordance of the indicators, several profiles can be observed in Table 5. Unsurprisingly, for high level of agreement of the selection criteria (in particular for the combination of the language of the tweets and of the interface) we obtain Italians who live in Italy replicating the results of the multi-signal intersection technique (Sec. 3.3). Instead, loosening the selection criteria we can observe the presence of other profiles, such as Italians who live abroad, foreigners in Italy, foreign companies, children of emigrants, tourists and even supporters of Italian teams (mainly football). Therefore we were able to determine not only the presence of the residents, but also of other categories of users who are part of the same geographic community without being resident in the territory.

[13] The level of precision of the percentages has been computed employing Cochran's formula (see footnote 12). This led to a confidence level of $\pm 31\%$ for all the criteria beside than for the combination of location and language of the interface and tweets ($\pm 30\%$) because of the so selected small number of Twitter accounts.

As shown in Table 5, each signal has a specific meaning and therefore it identifies a certain class of users. Since the time zone was deeply influenced by the location of the user during the set of the account, because of the suggestions that Twitter provided through the IP address, it is reasonable that, if it is not associated with any parameter language related, it identifies accounts of foreigners in Italy.

Finally we wish to underline how the multi-signal union approach was able to select classes of users that have a cultural connection with Italy beside from the residents, would have not been selected by the existing techniques (Sec. 3). Mainly because their focus was the residence, not the cultural connection of the user population.

4.4 Triangulation

To perform a credibility check on the size of the selected sample, due to the lack of directly comparable data, we performed a triangulation employing different sources.

Audiweb[14] is an impartial monitoring organization about internet usage in Italy. It collects data from a panel of people whose devices have a monitoring software. According to *Audiweb*, in 2017 around 7,100,000 Italians had a Twitter account whose high majority accessed through mobile. They are between 35 and 54 years old; much older than the users of other social media or than the USA's audience. Instead, in agreement with the American Twitter users, most Italian users are men (57%) and, among all the social media, Twitter has the highest number of people with a degree.

According to the analysis performed by Twopchartsin 2013[15], only 44% of the users that has a Twitter account has tweeted at least once. Furthermore only 23% of them has tweeted in the last month. Considering these percentages valid also for the Italian users in 2018, then around 718,528 Italian Twitter users should have twitted in the last month. We selected 65.33% of this number; a reasonable amount considering that the percentages are 5 years old and that they may have varied in the meantime. Finally, we selected not only Italian users, but more broadly users who have a cultural connection with Italy, therefore we can not directly compare these estimates.

5 Conclusion

To account for mobility and the current sociological research, we introduced the concept of territorial community. As shown in the introduction (Sec. 1), the most recent conceptualisation of community is an imaginary state of mind (Anderson, 1991). Groups of people whose personal interests intersect, even if only partially, the cultural space of the selected geographical area regardless of the territorial dispersion (Elias, 2008). The existing identification approaches (Sec. 3) do not consider the current sociological definition of territorial community leading to the selection of only subsets of the territorial communities, such as residents and locals. Therefore, we proposed a technique to select members of geographic communities on social media disregarding members' interaction and residence (Sec. 2). Hence any user whose cultural space intersects the one of the geographical area of interest. This was obtained considering any account that satisfies any of the selection criteria. In fact, the presence of discordant indicators does not exclude the users from the so selected community, but it only indicates that their personal interests may range beyond the territorial ones.

Our study implemented the proposed approach to identify territorial communities on Twitter, because of its focus on news. However, this methodology can be generalized also for other social media platforms; not only the platforms that do not require users' nationality (e.g. Twitter). In fact, nationality and belonging to a certain territorial cultural space are different concepts.

To test the efficiency of the developed technique (called *multi-signal union* since it considers the union of the users selected through each signal), we implemented it for a specific geographical area: Italy. Nevertheless this methodology can be generalized employing it for any territory characterized by a geographically-focused language.

Even if the language of the tweets is the parameter with the highest coverage among the selected population, the other signals (language interface, time zone and associated location) identify mostly users with a very strong cultural-linguistic identity (higher than the ones identified only by the language of the posts) since they express their belonging to the Italian cultural space through multiple

[14] http://www.audiweb.it

[15] https://www.cbsnews.com/news/many-twitter-users-dont-tweet-finds-report/

signals. Furthermore, we can consider these users as part of a community, not only a group with shared interests, because they express their belonging to the territory through the selection criteria.

Trough *multi-signal union* we were able to select not only the residents, that could have been identified also by the existing methods in the literature, but also other categories of users intersecting the territory by a cultural connection. In fact, for the study case (Sec. 4.2), we were able to select not only Italians living in Italy, but also Italians abroad, immigrants, foreign companies, children of emigrants and even supporters of Italian teams. Categories that could not have been selected employing the existing geo-inference techniques, due to their focus on spatial proximity.

An additional advantage of the use of multiple signals is that this practice makes this technique robust to Twitter policy variations. In the future Twitter may not make any more some Tweet information available for the data collection. For example since 25/05/18 Twitter no longer offers information regarding users' time zone, but the replicabiltiy of the technique remains intact. We would not be able anymore to select only 2.61% of the actual population size without this information.

The main limitation of this approach is that half of the selection criteria are language oriented, thus this approach has the limitation of not working properly for geographically-diverse languages. This restriction has no easy fix; it may be possible to analyze the text in the tweets trying to identify dialectal words employed in a limited geographical area. This is slightly similar to the approach employed by Han et al. (2014) in identifying indicative words that may give some indication on the location of the users (e.g. Piccadilly Circus for Londoners).

Even if this approach can be implemented for any territory with a geographically-focused language, some of the decision that we took during the implementation steps (Sec. 4.1) may have to be modified during the implementation of this methodology for other languages. In particular, cutting the users selected only through the language of the tweets below a certain level of consistency in the use of the language (step 4) may not be necessary for a more accurate language recognition or for languages that are less unambiguous than Italian (e.g. Hungarian or Turkish).

Another step that may not be necessary in the implementation of other geographical languages is the robustness check employed to check the consistence of the selection criteria (step 5). We employed it since we wanted to perform a one year long data collection, therefore we wanted to select only the users who were still part of the territorial community after 8 months. Nevertheless, if the analysis does not require a long data collection, this step may not be necessary.

This methodology presents several applications in multiple fields. For example detecting territorial communities, that depend on people's interests, can give us insights on attention dynamics. Such as replicating on Twitter Basil and Brown (1994)'s experiment (showing that if news items are personally relevant to an individual, they will be more likely to share it with others) for the several categories composing the cultural space.

Another possible use of this technique is to test if the different categories so selected present differences in the use of the platform. As Hong et al. (2011) pointed out, language communities differ significantly in the use of: URLs, Hashtags, Mentions, Replies and Retweets. For example in 2010 German users had a strong propensity for content-related behavior (high use of URLs and hashtags) and instead Indonesians and Malays preferred social behaviors (high use of retweets and mentions). Since some differences might be attributed to cultural differences or to how long Twitter had been used in that community, we may observe these variations also among the categories previously selected.

Finally, our paper shows the value of theory-driven operationalization. As we decide on which signals to include in our community detection approach based on concepts developed in sociological theory, our paper illustrates the potential and challenges of doing data science following well-established sociological theory developed in contexts diverging radically from the environment of digital communication (Sec. 3). To the best of our knowledge, the developed approach is the first to perform user selection going beyond spatial proximity selecting user categories that would have not being identified otherwise, but whose existence is well known in sociology (e.g. children of emigrants).

References

Anderson, B. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso, London, New York.

Backstrom, L., Sun, E., and Marlow, C. (2010). Find me if you can: improving geographical prediction with social and spatial proximitys. In *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, pages 61–70.

Basil, M. D. and Brown, W. J. (1994). Interpersonal communication in news diffusion: A study of "magic" Johnson's Announcement. *IEEE International Conference on Data Mining*, 71(2):305–320.

Berry, J. W. (2011). Integration and multiculturalism: Ways towards social solidarity. *Papers on Social Representations*, 20(1):1–21.

Bruns, A., Burgess, J., and Highfield, T. (2014). A "big data" approach to mapping the Australian Twittersphere. In *Advancing Digital Humanities*, pages 113–129. Springer.

Bruns, A., Moon, B., Münch, F. V., and Sadkowsky, T. (2017). The Australian Twittersphere in 2016: Mapping the follower/followee network. *Social Media+Society*, 3(4):1–15.

Bruns, A. and Weller, K. (2016). Twitter as first draft of the present and the challenges of preserving it for the future. In *Proceedings of the 8th ACM Conference on Web Science*, pages 183–189. ACM.

Cohen, A. P. (1985). *The Symbolic Construction of Community*. Psychology Press.

Elias, N. (2008). *Essays II: On Civilising Processes, State Formation and National Identity*. University college Dublin press.

Fleischmann, F. and Maykel, V. (2016). *Dual identity among immigrants: Comparing different conceptualizations, their measurements, and implications*, volume 2. Educational Publishing Foundation.

Fortier, A.-M. (1998). The politics of "italians abroad": Nation, diaspora, and new geographies of identit. *Diaspora: A Journal of Transnational Studies*, 7(2):197–224.

Han, B., Cook, P., and Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Hawley, A. H. (1950). *Human ecology; a theory of community structure*. Ronald Press.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 11 of *CHI 2011*, pages 237–246. ACM.

Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386.

Hong, L., Conventino, G., and Chi, E. H. (2011). Language Matters in Twitter: A Large Scale Study. In *ICWSM*, pages 518–521.

Israel, G. D. (1992). Determining sample size. In *University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS, Florida*, pages 1–5.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in twitter using social networks: A critical analysis and review of current practices. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 188–197.

Leszczensky, L., Maxwell, R., and Bleich, E. (2019). What factors best explain national identification among Muslim adolescents? Evidence from four European countries. *Journal of Ethnic and Migration Studies*, pages 1–17.

Maykel, V. and Martinovic, B. (2012). Immigrants' national identification: Meanings, determinants, and consequences. *Social Issues and Policy Review*, 6(1):82–112.

McGee, J., Cattuto, C., and Cheng, Z. (2011). A geographic study of tie strength in social media. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2333–2336. ACM.

Mok, D. and Wellman, B. (2007). Did distance matter before the internet?: Interpersonal contact and support in the 1970s. *Social Networks*, 29(3):430–461.

Mose, A. and Shive, S. (2011). Patriotism in your portfolio. *Journal of Financial Markets*, 14(2):411–440.

Newby, H. and Bell, C. (1974). *The Sociology of Community: A Collection of Readings*. Frank Cass and Company, London, UK.

Schulz, A., Hadjakos, A., Paulheim, A., Nachmtwey, J., and Mühlhäuser, M. (2013). A multi-indicator approach for geolocalization of tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM '13)*, pages 573–582.

Sequiera, R. and Lin, J. (2017). Finally, a downloadable test collection of tweets. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1225–1228. ACM.

Shoemaker, P. J. and Reese, S. D., editors (1996). *Mediating The Message: Theories of Influences on Mass Media Content*, volume 52. Longman, second edition.

Sloan, L. and Morgan, J. (2015). Who tweets with their location? Understanding the relationship between Demographic Characteristics and the Use of the Geoservices and Geotagging on Twitter. *PLoS ONE*, 10(11):1–15.

Straubhaar, J. D. (1991). Beyond media imperialism: Assymetrical interdependence and cultural proximity. *Critical Studies in media communication*, 8(1):39–59.

van Meeteren, M., Ate, P., and Elenna, D. (2010). Mapping communities in large virtual social networks: Using twitter data to find the Indie Mac community. In *2010 IEEE International Workshop on: Business Applications of Social Network Analysis (BASNA)*, pages 1–8. IEEE.