ISSN 1870-4069

When Twitter Becomes a Data Source for Geospatial Analysis

Magdalena Saldana-Perez[1], Camille Cavalière[2], Miguel Torres-Ruiz[1], Marco Moreno-Ibarra[1]

[1] Instituto Politécnico Nacional, Centro de Investigación en Computación, Laboratorio de Procesamiento Inteligente de Información Geoespacial, Mexico
[2] University of Grenoble Alpes, Laboratoire d'Informatique de Grenoble, France

{amagdasaldana, mtorres, marcomoreno}@cic.ipn.mx,
Camille.Cavaliere@imag.fr

Abstract. Social media has become a useful data source for processes and researches interested in improving people's life. Publications done by social media users provide details about the people perceptions of their environment, and updated observations about what happens in real world. In this approach the relevance of Twitter as a data source for scientific purposes is analyzed, as well as its use in geospatial researches. Tweets have two main characteristics, a text where user describes its ideas, and metadata, where features such as the coordinates of the place where the tweet was posted are stored. Different computing procedures are applied over tweets in order to make them useful for different tasks; commonly, text mining, classification, and regression algorithms are used to process tweets. The coordinates of tweets make possible to link events described in tweets with the geographical area where they occur. The analysis of tweets and coordinates provides updated data, useful in natural disasters control, decision taking processes and urban studies. The present approach studies what motivates users to tweet, and analyzing the messages produced by Twitter users, classifies the tweets into three groups according to the information expressed in their text: *self-centered, social information* and *collective information*. Additionally, some methods to extract information from tweets are studied, common problems presented when working with tweets, and some researches that use them in a geospatial domain are presented.

Keywords: twitter, social media, data treatment, classification, text mining, geospatial analysis, data problems.

1 Introduction

Twitter is commonly used as data source at scientific and commercial data analysis; the microblogging features are analyzed by researches interested in emergency information diffusion, on the dissemination of news, and on the way in which authorities and civil societies help citizens with problems or crisis situations [45].

The microblogs posted on Twitter are called tweets, they have two main elements: metadata, and a text limited to 280 characters; their metadata is composed of a timestamp (the date and time of the tweet's publication), the coordinates of the place

where the tweet was posted, and the id of the user who wrote the tweet [71]. In Twitter, users are able to see some others tweets, acting as followers. Users can share their tweets with all the media members (public profile), or with specific users (private profile) [3].

Considered as a new data source, Twitter provides people's points of view from many different topics such as health, environment, politics, economy, sports, and natural disasters, among others. The social media lets researchers know the people's feelings and ideas about factors that disturb their daily activities; governments and institutions interested in improving people's lives and in solving specific problems can use its updated information [57]. Twitter facilitates the gathering of crowd-sourced information [46]. Twitter provides information in less time than the methods used to collect data before the social media explosion. In the past, data related to people's feelings and opinions about their environment were collected applying surveys and census [34].

The data processing of the tweets text and metadata components becomes a challenge, sometimes users do not allow the social media to know their location, tweets do not have coordinates. Sometimes people just tweet comments about a topic, this represents a problem for scientific researches since most of the times personal comments do not provide relevant information [42]. For example, when terrorism acts occur special hashtags are created in order to group all the tweets related to the act; many tweets use hashtags, some of them provide information about the event, and some others describe people's feelings (these last generate noisy information). Noisy tweets affect the quality of the tweets sample when it is processed, personal comments must be filtered in order to assure that the collected tweets contain valuable information. When tweets regarding to a specific topic are tweeted far from the study area, is important to consider the variations of space and time between them. In developed countries, citizens volunteering and the information from social media or collective systems without an official agency affiliation are undervalued (Whittaker et al., 2015). The use of extra data sources when working with Twitter helps validate the tweets data and ameliorate the researches accuracy [67].

According to Haworth (2016), the citizens observing, collecting, sharing and analyzing data have led the development of many scientific researches that would not have been possible otherwise. Similarly to the collaborative information and social media data, the quality of citizen's science has been questioned, but it has been shown that applying the adequate preprocessing procedures, the citizens' data can meet the same quality as data collected from official sources [10]. People's participation has lead the citizen science, which refers to engage public citizens in scientific research projects [7].

2 Objective

The present approach analyses the use of Twitter in scientific researches related to the geospatial domain. The objective of the approach is show the importance and utility of data obtained from Twitter when are used in the geographic domain, and propose a classification of tweets based on the information they provide for different data analysis. The approach presents the relevance of data extracted from tweets, the

methods used to manage them, and some problems presented when working with the most famous microblog service. In addition, research works that merge tweets with geographic information systems (GIS) are described.

The paper is organized as follows, in section two are presented the material and methods of the investigation, in section three the theory of the approach and the motivation of people to tweet are analyzed, in section four results and discussions are presented; finally, the conclusions of the approach are given.

3 Material and Methods

Volunteered geographic information (VGI) refers to the wide creation of geographic information by citizens using web platforms, free mapping tools, mobile applications and social media. VGI has changed the creation and propagation of information generated by people [34].

In disaster management context VGI can also be labelled as digital volunteering (McLennan et al., 2015), or digital humanities [13]. Digital volunteering has become a source of asserted information used to complement geographic information from governmental agencies and private organizations [37].

Many researching works that use Twitter as a data source meet in one point: the data provided by tweets is an updated vision about people's perceptions of their environment. Landwehr et al. (2016) see in Twitter the opportunity to get information out of public, which increases the rate of dissemination. According to Haworth and Bruce (2015), VGI academic researches and emergency management studies have focused on citizens' science. For example, the 2010 Haití earthquake where volunteers from all over the world worked together to map the affected area [51]; the volunteered mapping of the Nepal earthquake crisis [34] and the crowdmaps responding to the cyclone and floods in Queensland [49]. Another valuable example of emergency management using citizen's data is the OSM tasking manager [54], a mapping tool designed for the Humanitarian OSM Team to give solutions when critical situations occur, such as the Hurricane Maria (20/09/2017) and the Mexican Earthquake (19/09/2017).

Some studies emphasize the benefits of VGI and social media data meanwhile some others argued that web platforms marginalize people without internet and technology access. Benefits of VGI are the timely information exchange and the connectedness (Taylor et al., 2012), the provisional information for disaster mapping [49], and the most important, the availability of data in near-real time [34].

There are researching works that study the human behavior in social media [1]; some others study people's feelings (Agarwal et al., 2011). There exists researches that model urban factors analyzing tweets at specific places and developing forecasting processes [77]. In some projects, tweets are considered as social and political indicators [40]; even, there are researches focused on study what motivates people to tweet [58].

On Han et al. (2017) the authors propose a model to study the factors that influence and motivate people to tweet. The researchers stablished four possible kinds of gratification people get from their participation in Twitter: content, technology, process, and social gratification. Their research shows that users became satisfied

after tweeting relevant information, since they consider themselves useful for other people (social gratification). Burnap and Williams (2016) describe the importance of considering big data and social media as data sources at policy decisions processes, implementing a machine learning classifier to detect cyber hate speech in tweets. Many researches classify human emotions from tweets considering the Ekman and Friesen model, which identifies six basic emotions: joy, anger, fear, sadness, surprise and disgust [25]. Resch et al. (2016) consider such model to identify human emotions implementing a semi-supervised machine-learning algorithm.

Two sciences deeply interested in Twitter data analysis are Geography and Computing. Geography has become indispensable when working with tweets to georeferenced them; computing helps to apply complex processes over tweets to produce data models and statistical studies.

When forecasting and modeling from tweets there are three challenges to consider: i) the spatio-temporal relations between tweets, ii) the temporal evolution of spatially distributed tweets related to specific events; iii) the use of prior geographical knowledge since different geolocations could derive in different events.

In Albuquerque et al. (2015), social media and authoritative data are used to identify information for disaster control. Tweets that provide relevant information about the Elba's river floods in 2013 are analyzed, such data are merged with hydrologic sensors and digital models of the affected areas with a view to create flooding models. Some researches [24, 36, 74] use geographic information systems to georeferenced tweets. This kind of researches are relevant since few tweets have the coordinates of the place where they were posted, as studied in Section 3.1.

One of the main advantages of VGI and social media data are the community ties generated when people share local knowledge with disaster management systems, increasing the survival opportunity for more people [34]. In Landwehr et al. (2016), a social media response system for tsunami warnings is presented, the project is a web application designed for the community of Padang Indonesia, to warn people of tsunamis. The application collects and analyzes tweets from the region to provide immediate feedback when tsunamis happen, supporting an early warning system.

4 Theory. What is the whole point of tweeting? Caution, Users Tweeting

In this approach, we propose a classification for tweets: *the self-centered tweets, the social information tweets* and *the collective information tweets (*represented in Figure 1). The individual information or *self-centered tweets*, share personal information of the users answering to the questions "*What am I doing right now?"* or *"What am I seeing right now?"* These tweets are posted quickly and their temporal importance is short depends of the time the user spends doing the action.

Social information tweets express the user's thoughts about social events (political events, strikes, terrorism), which brings about mass mobilization. Opinion mining and sentiment analysis studies are often based on this kind of tweets. Researches as the Resch et al. (2015) make use of this kind of tweets to get information from the people's environment treating Twitter users as sensors.

Collective information tweets, report information about an unusual event occurrence that disturbs the daily life, those tweets are based on individual's observations but are not self-centered. Figure 2 presents some examples of the proposed classification of tweets.

Tweets provide two valuable data: the user's perspective of an event, and the coordinates of the place where it is probably happening.

The present analysis is focused on the collective information and the social information tweets. These two kinds of tweets provide information about problems that affect the normal activities of people, or modify their environment, such as natural phenomena or socio-political changes.

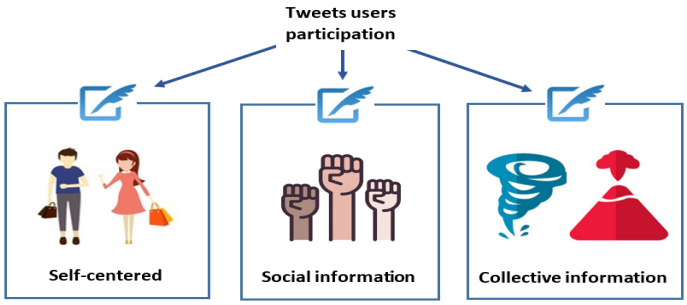

Fig. 1. The tweets can be classified as: 1) self-centered, 2) social information related or 3) collective information related.

4.1 Why are tweets scientifically relevant?

The collective tweets can indeed spread life-saving information. This kind of tweets are frequently collected as VGI for emergency warning and situational awareness, their relevance lies in their facility to spread important information in a short time [65]. When an unusual event occurs, some users near to the place report in Twitter what they see using hash tags and specific words, in some minutes the information is available for millions of users near or not to the event.

Two characteristics of tweets that make them relevant for scientific purposes are their coordinates and timestamp; unfortunately, not all the posted tweets have coordinates, and not all the cities present the same citizen's participation in the social network.

Geographic science has an important role managing the coordinates of the tweets; furthermore, tweets commonly refer to geographic places on their texts [31]. There exist two possible cases when analyzing tweets coordinates, their presence or absence. Some users let their smartphones add their geographic coordinates when tweeting, in such cases, the Twitter app will automatically enable the device's embedded GPS to add the latitude and longitude coordinates of the user to the tweets metadata [[17].

When the user tweets from a device without embedded GPS, its geographic coordinates can be manually added; Twitter provides a list of cities based on the users IP address [48]. Those georeferenced tweets however represent a small part (1%-2%) of the whole Twitter stream [8]. The coordinates of the tweet can be represented as point-shaped data using Geographic Information Systems (GIS).

When the user does not allow Twitter to add his coordinates, the tweet can be georeferenced if needed. There have been implemented different kinds of georeferencing models; some of them use specialized gazetteers to estimate the coordinates of the tweet based on its text [38]. The general process consists of three steps: 1.-identification of geographic elements such as roads, avenues, monuments, and buildings names in the text of the tweet; 2.-search the coordinates of the geographic elements in a gazetteer; 3.-compute the coordinates of the tweet, applying geospatial functions over the coordinates found in step 2.

A useful element in tweets is the hashtag (#) used to cluster tweets with similar information [11]. Names and words ambiguity must be considered when using any georeferencing method, to differentiate geographic entities with the same name (as roads, avenues and bridges that share their name), and to prevent the data lack of precision. These issues are addressed by special techniques, which identify road segments and key words from a particular geographic context [5]. One of such techniques is the Named Entity Recognition (NER), which determines the geographic place referenced in a text, considering the disambiguation of names and the most appropriate geographic context (Daly et al., 2013).

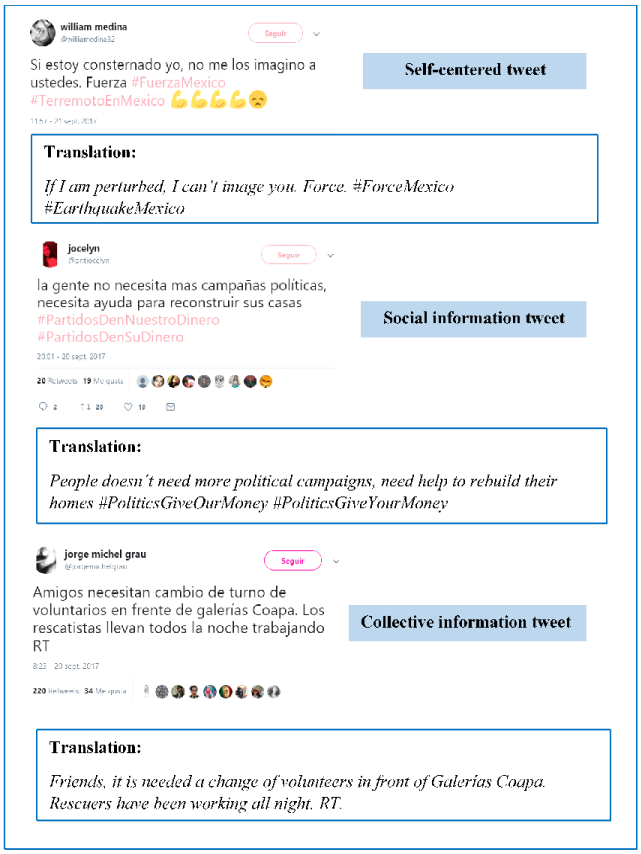

Fig. 2. Examples of the three different proposed classes of tweets. The tweets were collected on 21st September 2017, they refer to the earthquake occurred in Mexico on 19th September 2017.

In Roller et al. (2012), NER and Wikipedia are merged to analyze and find geographic terms in tweets. Graham et al. (2014) blend NER and specialized gazetteers to identify geographic sites. Lee et al. (2015) propose a machine learning approach to extract named entities from tweets text and determine their location considering disambiguation.

In Escamilla et al. (2016), is proposed a methodology to geocoding traffic events reported in tweets, the approach identifies the traffic events, geographic features, and their possible spatial relations, using natural language processing (NLP), an ontology, and classification algorithms. In its geocoding stage the proposed methodology extracts geographic entities using NER, finds the spatial relations between them using a syntactic dependency tree, and defines spatial operations to apply over the geographic elements.

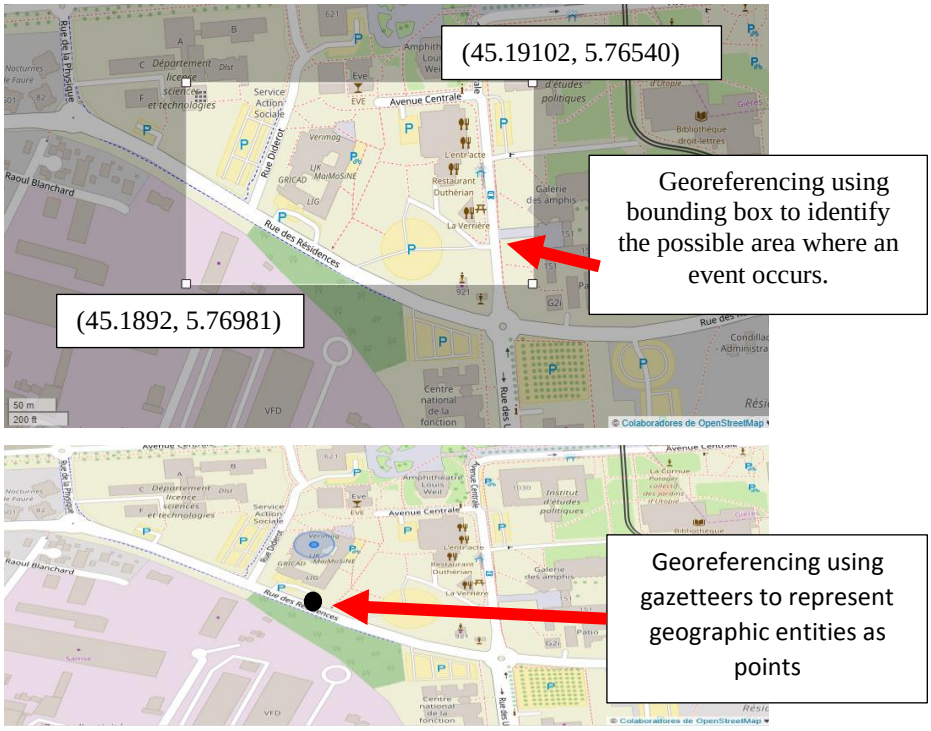

Fig. 3. The two described georeferencing methods when working with tweets. The image on the top shows a bounding box used to search the possible location of a tweet. The bottom image shows a point shape location assigned to a tweet. The cartography belongs to ©OpenStreetMap Contributors.

Another common georeferencing method consist in creating a bounding box; a bounding box is a rectangle created to represent the possible area where a tweet was posted; the box corners are coordinates of places mentioned in the tweets text, or places inferred from other user's tweets [59]. Tweets georeferenced by bounding box are represented as rectangles in GIS.

The geocoding procedures are applied over texts in order to identify geographic entities such as points of interest, events of interest, geographic relations between entities (intersections or their position with respect to others), or geographic features of certain events. Figure 3 shows examples of the mentioned georeferencing methods.

5 Results and Discussions: Geospatial Analysis Based on Tweets

There are researches interested in analyzing tweets to obtain information for problems solving and to develop decision-making processes. Most of the times, the text of the tweets is the main studied feature. Until 2015, a study of 92 revised researching articles from different disciplines working with Twitter showed that only 33% of them made use of the spatio-temporal and semantic characteristics of tweets [67]. Most of the researches interested in Twitter are focused on their text treatment to get information about specific topics, as users emotions [75, 38, 23], or activities. Such researches have propitiate the generation of text mining and natural language processing techniques, and more accurate geo-location procedures.

In this section, some methods to collect and pre-process tweets are explained; also, some common problems present when managing tweets are exposed. Finally, examples of researches that use tweets as data sources are described.

5.1 Preprocessing Tweets

Tweets need to be pre-processed before be used at specific tasks. There is possible to apply over them common data treatment processes, such as text mining, natural language processes, classification, identification of patterns and elements, among others [68].

Twitter lets programmers and developers access to tweets for educational or scientific purposes, there have been developed specialized API's for different programming languages such as Python, R, Java, and others, that allow programmers to extract tweets and its metadata in real time [60].

When the data feature of interest is the text of the tweet, data mining and the NLP are two relevant tools. Commonly, texts are tokenized into words or n-grams to analyze each textual element in order to identify relevant terms and stop words [66]. There exist a variety of algorithms and libraries to implement NLP and text mining procedures over text [78].

When using NLP and textual elements from tweets is important to consider the language used in the region where the tweets were posted [29], there are particularities in the language used by citizens in a region and people who lives in another area, even if the two places share the same official language. For example, in England people use the term motorway to refer to a highway. In addition, it is important to consider the regional terms and their meanings, in Mexico the term *ahorita*, according to the Oxford dictionary [55] means a moment close to the present, immediately after the moment when term is being used. For Mexicans *ahorita* means in a moment or never, according to the context where the word is used.

When working with tweets is important to consider the language in which they are written, in some cases the geographic place where they were tweeted is also relevant. Twitter trending topics change from one geographical area to another; what is more,

people use words that sometimes have a specific local meaning. The miss consideration of the tweets language generates a noisy analysis (Figure 4).

The recurrent task done over the coordinates of tweets is their georeferencing using GIS functions. The spatio-temporal researches that use Twitter, consider three main features of the tweets: their texts, their coordinates and their timestamps [69].

5.2 What are the Main Problems when Working with Data from Tweets?

From the analysis of methodologies that use Twitter as a data source, it is possible to identify two main problems when working with tweets, the data quality and the data quantity they provide.

5.3 Data Quality

Although some tweets are deliberately created to spread useful information within a large community, most of the tweets gathered by scientists for scientific purposes are merely individual information, users are usually not aware their data can be analyzed for scientific purposes. Harvey (2013) proposes the term *contributed data*, to describe any data generated by users' activities on social media platforms. Tweets are part of soft data, which defines all information created and shared on social media platforms; contrary to hard data (traditional datasets managed by governmental agencies), tweets come from heterogeneous sources that cannot be all confirmed.

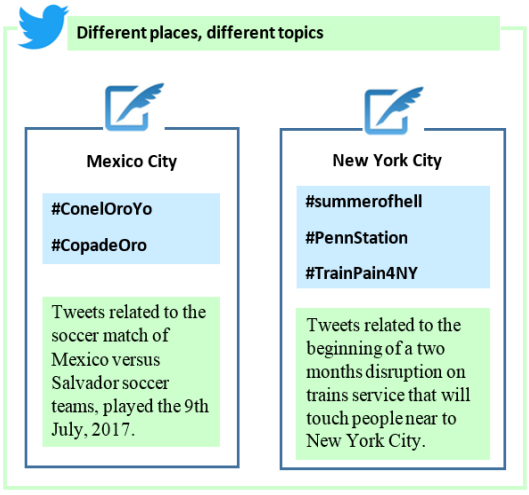

Fig. 4. Trending topics on Twitter at Mexico City and New York City on 10th July 2017. In the first city, the hashtags are associated to a soccer match; in the second case, the hashtags communicate train disruptions. The user´s interest change from one geographic area to another.

Moreover, the process of the acquisition is undocumented and there is no data quality assertion [26]. Two recurrent questions are therefore: *Are tweets trustworthy? What is the degree of trust of tweets?*

Tweets quality is therefore heterogeneous and highly variable, while some tweets report a mere observation with no further information (e.g. "*It's raining*" without any

detail about the rainfall intensity), others may include accurate information (e.g. "Hail cracked the windshield" may provide clues about hail size and intensity). Also, the geo-references given in the text of tweets sometimes are vague or their description is general (e.g. "*Earthquake in Mexico*" does not points the exact area where the event occurred, meanwhile "*Earthquake in Chiapas southwest of Mexico*" provides more information to locate the event).

The quality in VGI data is variable for many reasons such as the volunteers experience producing information, their activities and educational level, the multiple technologies to generate information, the purpose of the transmitter, and the precision of the message. Senaratne et al. (2017) review different VGI quality measures and indicators with a view to show the importance of evaluate the VGI information before use it.

In Dashti et al., (2014) the scientists deal with natural hazards using georeferenced tweets and real-time information from a recovery system of disaster information. Crosschecking tweets location with official datasets (flood hazards cartography and Landsat remote imagery), the authors show that georeferenced tweets provide relevant and valuable information about flood events.

Tweets are reports of what users perceive in their local environment; such reports vary from one individual to another, according to their personal life experiences. Observations shared in tweets may be subjective and some bias may be introduced.

Do tweets show people true thoughts? Every time a political event of global interest occurs, *traditional surveys* (based on representative population samples) and *big data analytics algorithms* (opinion mining based on machine learning algorithms that analyse people's publications on Twitter and Facebook) compete to predict the most accurate results.

Which of them got the most relevant results? Opinion mining algorithms based on social media content analysis performed by private companies turned out to be closer to the results than the surveys. The algorithms predicted Trump's election [9] and the Brexit decision [15], but they finally failed in predicting the results of the first French presidential election round [41]. Why? People using social media to express their opinion are not likely to give their true thoughts, or they are ironic, which makes no sense for the machine learning algorithms causing their failure [52]. Finally, users may discuss distant events; users can produce information about a particular event, even if they did not attend it; also, they can keep spreading information after the end of the event. Miller (2017) calls this phenomenon *asynchronous telepresence*.

5.3.1 Data Quality

Tweets are not representative samples of a population, the number of Twitter users is a portion of the total habitants in a specific geographic area (see Figure 5); most of the times people interested in Twitter are part of groups with a specific age and educational level. Tweets can be considered as digital footprints.

Li et al. (2013) tried to draw social profiles of Twitter users in California crossing socio-economic data, demographic data, and geolocated tweets; these results cannot be generalized to the tweeters in the whole world. There is an issue dealing with results generalization, Goodchild (2013) pointed out that results of experiments using random representative samples can be generalized because they apply to the whole

population in a specific area, but the results of a study based on tweets cannot be generalized to a complex geographic and social world. Tweeting indeed involves a minority of the population; tweets show the users concerns but, How can be known if all the people living at the same place share these concerns? Whatever could be concluded from a study case might not show the same results when analyzing the same case in another space or in another period.

Tweets spatial distribution is quite irregular. Studies show that a strong correlation between tweets and populations densities results in high concentrations of tweets in highly populated cities, whereas in rural areas less tweets are produced [48]. That is a reason why the specific geographic areas of study must be delimited when working with Twitter.

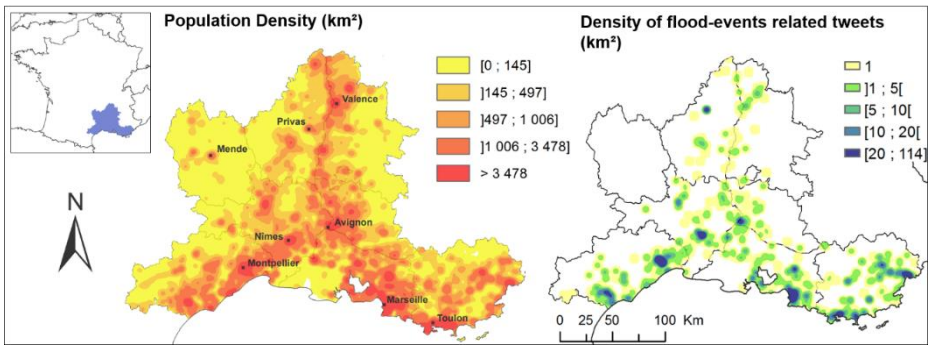

Fig. 5. Analysis of population density and tweets related to floods in the southeast of France. These maps confirm that the number of tweets is not proportional to the population density of an area. Not all the habitants are interested in participating in Twitter.

If events occur in sparsely populated places, just a few users may be therefore involved in creating sparse information. At a local scale, population densities vary depending on time: if an event is occurring within a period of time when a particular place is unoccupied (e.g. a residential neighborhood during daylight), it may be missed and undocumented.

On one hand, events with different features do not have the same impact in the media. Major events, are widely spread in the media and bring about high streams of tweets; on the other hand, local events do not have such a media propagation, so the spatiotemporal extent of the event does not spread over the edges of the physical event. When working with tweets related to events at a local scale, researchers must deal with small datasets.

5.4 Where Twitter Meets GIS

There exist researches that merge GIS and Twitter with a view to develop decision-taking and urban computing procedures. Nowadays some GIS systems let people manage geographic information, even if they do not have a specialized formation [39].

GIS can manage tweets coordinates to represent them as geographic points, or to apply over them geospatial functions; there are open source GIS such as QGIS, Grass, gvSIG, SAGA, and open source databases to manage geographic data such as

PostGIS, SpatialLite and MySQLSpatial [28]. Some GIS-Twitter researches are focused on analyzing people's reports of weather, rain, snow storms, floods [4, 6]; some others model the activities of people in urban environments, such as traffic, pollution, public demonstrations, taxi routes [18, 42, 44]. In addition, there exists some researches interested in studying and managing critical situations [56, 53, 16].

One of the main purposes of Twitter interested researches is to develop analytical tools to study events that represent a risk for people, in order to prevent them from possible damages. In Cavalière et al. (2016), the researches analyze georeferenced tweets posted by users and authorities from the southeast France that provide information about extreme rainfalls and floods. Their proposed methodology processes the tweets using the abductive reasoning. The approach is proved considering tweets posted in the study area during the 2014-fall season.

The Peta Jakarta Project has been implemented to gather crowdsourced data and tweets posted when a flood event occurs. These data are crosschecked with other data sources (topographic, demographic and socio-economic data) and shown in a map, in order to visualize the affected areas and to organize emergency responses [51].

Events forecasting from social media streams is a novelty-researching trend. In Zhao et al. (2016), the authors propose a new batch and online approaches for spatio-temporal events forecasting from social media data; their approach characterizes the evolutionary pattern of both spatial burstiness and structural context. The authors remark the Twitter fundamental characteristics: the timeless of messages, the ubiquity of social sensors, and the geo-information availability. The researchers conclude that the forecasting of spatiotemporal events requires the consideration of spatial features and their correlations, in addition to the temporal dimension.

Machine learning (ML) algorithms are used to classify and compute regressions over the tweets; for example, in He et al. (2017), the authors propose an optimization framework to extract traffic related information from tweets using a transformation matrix, also make long-term traffic predictions using linear regressions; their study case is San Francisco Bay, California.

An example where GIS and ML are used over tweets is the traffic analysis based on short text from social media [64]. Traffic related tweets are pre-processed by text mining and NLP procedures, georeferenced, and classified by ML algorithms into different traffic events; finally, the classified tweets are geovisualized using an open source GIS, with a view to analyze traffic in Mexico City.

The relevance of Twitter usage in geographic studies relapses in the tweets coordinates and the information that can be inferred from their texts [34]. Tweets are a source of updated reports about what is happening in real world. The analysis of tweets regarding natural hazard lets geography has a perception of the territory changes, to generate updated maps, and to design evacuation routes [46].

When geographic sciences and urban computing use Twitter, they have the opportunity to obtain citizen's information about what happens in their environment, in order to ameliorate their life's modifying urban aspects such as traffic, pollution, and dangerous places. Twitter opens a new opportunity to sense the real world problems, and to implement efficient and fast solutions with the purpose of solve them.

6 Conclusions

Twitter has demonstrated to be a useful tool for data analytics at different science disciplines. Tweets are a new way to obtain information about what happens in real world. People are getting more interested in participating in the social media; they think their information helps others.

Many factors motivate people to tweet; the factors are related to the kind of gratification people gets after do it. Tweets have important features such as metadata and text that can be used to georeference the tweet. Technological advances in mobile devices such as smartphones with embedded GPS let users keep in touch with the social media all the time, and add their coordinates to the tweet metadata. Unfortunately, for scientific analysis based on tweets, few of them have coordinates

In the present approach, tweets can be classified according to the kind of information they provide into three types: self-centered, social, and collective information. The self-centered tweets provide information about the mood and particular activities of the user. The social information tweets are interested in spreading information about particular events that disturbs people's life such as socio-political movements. The collective information tweets main purpose is to provide data about natural hazards and urban circumstances.

Many social researches study the self-centered and social information tweets since they provide information about people's feelings, activities and opinions. Most of times social and collective information tweets are used on researches interested in forecasting, or modelling urban factors and natural phenomena, considering the geographic features of the tweets and their timestamps.

Computing and geographic sciences work together in order to generate data managing procedures to extract tweets, and to analyze the microblogs features for scientific purposes. There have been developed different API's, text mining and NLP procedures to manage the tweets textual component; also, there have been designed and implemented different methodologies to georeference tweets without coordinates in its metadata.

Data obtained from Twitter can be crosschecked with authoritative data and some other information sources to increase its accuracy, and to provide more opportunities of data analysis. Twitter has open an option to monitor the real world dynamics through the user's perspectives; it could be say that Twitter provides first-hand information.

References

1. Aramo-Immonen, H., Kärkkäinen, H., Jussila, J. J., Joel-Edgar, S., Huhtamäki, J.: Visualizing informal learning behavior from conference participants' Twitter data with the Ostinato Model. Computers in Human Behavior, 55, 584–595 (2016)
2. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proceedings of the workshop on languages in social media, pp. 30–38. Association for Computational Linguistics (2011, June)
3. Balazs, J. A., Velásquez, J. D.: Opinion mining and information fusion: a survey. In: Information Fusion 27: 95–110 (2016)

4. Basnyat, B., Anam, A., Singh, N., Gangopadhyay, A., Roy, N.: Analyzing Social Media Texts and Images to Assess the Impact of Flash Floods in Cities. In: Smart Computing (SMARTCOMP), 2017 IEEE International Conference on, pp. 1–6 (2017, May)
5. Blanco, R., Ottaviano, G., Meij, E.: Fast and space-efficient entity linking for queries. In: ACM, Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 179–188 (2015, February)
6. Bobb, J. F., Ho, K. K., Yeh, R. W., Harrington, L., Zai, A., Liao, K. P., Dominici, F.: Time-course of cause-specific hospital admissions during snowstorms: an analysis of electronic medical records from major hospitals in Boston, Massachusetts. American journal of epidemiology 185(4), 283–294 (2017)
7. Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., Parrish, J. K.: Next steps for citizen science. Science 343(6178), 1436–1437 (2014)
8. Bouillot, F., Poncelet, P., Roche, M.: How and why exploit tweet's location information? In: AGILE'2012: 15th International Conference on Geographic Information Science, pp. N–A (2012, August)
9. Breur, T.: US elections: How could predictions be so wrong? (2016)
10. Bruce, E., Albright, L., Sheehan, S., Blewitt, M.: Distribution patterns of migrating humpback whales (Megaptera novaeangliae) in Jervis Bay, Australia: A spatial analysis using geographical citizen science data. Applied Geography, 54, 83–95 (2014)
11. Bruns, Axel, Jean E. Burgess.: The use of Twitter hashtags in the formation of ad hoc publics. In: Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011 (2011)
12. Burnap, P., Williams, M.L.: Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Science, 5(1), 11. Earth observation sources: a workflow for volunteered geographic information sensing, Doctoral dissertation, UCL (2016)
13. Burns, R.: Moments of closure in the knowledge politics of digital humanitarianism. Geoforum, 53, 51–62 (2014)
14. Cavalière, C., Davoine, P. A., Lutoff, C., Ruin, I.: Analyser des tweets géolocalisés pour explorer les réponses sociales face aux phénomènes météorologiques extrêmes. In: SAGEO'2016 (2016, December)
15. Celli, F., Stepanov, E., Poesio, M., Riccardi, G.: Predicting Brexit: Classifying agreement is better than sentiment and pollsters. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES), pp. 110–118 (2016)
16. Chatfield, A.T., Brajawidagda, U.: Twitter early tsunami warning system: A case study in Indonesia's natural disaster management. IEEE, System sciences (HICSS), 46th Hawaii international conference on, pp. 2050–2060 (2013, January)
17. Cheng, S., Fang, J., Hristidis, V., Madhyastha, H.V., Mithun, N.C., Perkins, D., Tsotras, V.J.: OSNI: Searching for Needles in a Haystack of Social Network Data. In: EDBT, pp. 616–619 (2016)
18. Cottrill, C., Gault, P., Yeboah, G., Nelson, J. D., Anable, J., Budd, T.: Tweeting Transit: An examination of social media strategies for transport information management during a large event. Transportation Research Part C: Emerging Technologies, 77, 421–432 (2017)
19. Daly, E.M., Lecue, F., Bicer, V.: Westland row why so slow? fusing social media and linked data sources for understanding real-time traffic conditions. In: ACM, Proceedings of the 2013 international conference on Intelligent user interfaces, pp. 203–212 (2013, March)
20. Danielsen, F., Jensen, P.M., Burgess, N.D., Altamirano, R., Alviola, P.A., Andrianandrasana, H., Enghoff, M.: A multicountry assessment of tropical resource monitoring by local communities. BioScience 64(3), 236–251 (2014)
21. Dashti S., Palen L., Heris M., Anderson K, Anderson S., Anderson T.: Supporting disaster reconnaissance with social media data: a design-oriented case study of the 2013 Colorado

floods. In: Proceedings of the 11th international conference on information systems for crisis response and management. ISCRAM, pp. 630–639 (2014)

22. De Albuquerque, J.P., Herfort, B., Brenning, A., Zipf, A.: A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. International Journal of Geographical Information Science 29(4), 667–689 (2015)
23. Do, H.J., Lim, C.G., Kim, Y.J., Choi, H.J.: Analyzing emotions in twitter during a crisis: A case study of the 2015 middle east respiratory syndrome outbreak in Korea. In: Big Data and Smart Computing (BigComp), 2016 International Conference on, pp. 415–418. IEEE (2016, January)
24. Dredze, M., Paul, M.J., Bergsma, S., Tran, H. Carmen: A twitter geolocation system with applications to public health. In: AAAI workshop on expanding the boundaries of health informatics using AI (HIAI), pp. 20–24 (2013, June)
25. Elfenbein, H.A., Ambady, N.: On the universality and cultural specificity of emotion recognition: a meta-analysis. Psychological bulletin 128(2), 203 (2002)
26. Elwood, S., Goodchild, M.F., Sui, D.: Prospects for VGI research and the emerging fourth paradigm. Crowdsourcing geographic knowledge, pp. 361–375) Springer Netherlands (2013)
27. Escamilla, I., Torres-Ruiz, M., Moreno-Ibarra, M., Quintero, R., Guzmán, G., Luna-Soto, V.: Geocoding tweets approach based on conceptual representations in the context of the knowledge society. International Journal on Semantic Web and Information Systems (IJSWIS) 12(1), 44–61 (2016)
28. Farkas, G.: Applicability of open-source web mapping libraries for building massive Web GIS clients. Journal of Geographical Systems 3(19), 273–295 (2017)
29. Gómez-Adorno, H., Markov, I., Baptista, J., Sidorov, G., Pinto, D.: Discriminating between similar languages using a combination of typed and untyped character n-grams and words. VarDial 2017, 137 (2017)
30. Graham, M., Hale, S.A., Gaffney, D.: Where in the world are you? Geolocation and language identification in Twitter. The Professional Geographer 66(4), 568–578 (2014)
31. Gutierrez, C., Figuerias, P., Oliveira, P., Costa, R., Jardim-Goncalves, R.: Twitter mining for traffic events detection. In: Science and Information Conference (SAI), pp. 371–378. IEEE (2015, July)
32. Han, Y., Hong, B., Lee, H., Kim, K.: How do we Tweet? The Comparative Analysis of Twitter Usage by Message Types, Devices, and Sources. The Journal of Social Media in Society 6(1), 189–219 (2017)
33. Harvey, F.: To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. Crowdsourcing Geographic Knowledge, pp. 31–42. Springer Netherlands (2013)
34. Haworth, B.: Emergency management perspectives on volunteered geographic information: Opportunities, challenges and change. Computers, Environment and Urban Systems, 57, 189–198 (2016)
35. He, J., Shen, W., Divakaruni, P., Wynter, L., Lawrence, R.: Improving Traffic Prediction with Tweet Semantics. In: IJCAI, pp. 1387–1393 (2013, August)
36. Huang, B., Carley, K.M.: On Predicting Geolocation of Tweets using Convolutional Neural Networks. arXiv preprint arXiv:1704.05146 (2017)
37. Jiang, B., Thill, J.C.: Volunteered Geographic Information: Towards the establishment of a new paradigm (2015)
38. Jusupova, Â., Batista, F., Ribeiro, R.: Characterizing the Personality of Twitter Users based on their Timeline Information. In: ATAS Conferência APSI, Vol. 16, No. 16, pp. 292–299 (2017, February)

39. Johnson, P.A., Sieber, R.E.: The Geoweb for community-based organizations: Tool development, implementation, and sustainability in an era of Google Maps. The Journal of Community Informatics 13(1) (2017)
40. Kavanaugh, A.L., Sheetz, S.D., Sandoval-Almazan, R., Tedesco, J.C., Fox, E.A.: Media use during conflicts: Information seeking and political efficacy during the 2012 Mexican elections. Government Information Quarterly 33(3), 595–602 (2016)
41. Kennedy, R., Wojcik, S., Lazer, D.: Improving election prediction internationally. Science, 355(6324), 515–520 (2017)
42. Kumar, K.E., Ahmed, H.A.: Estimation of traffic with accuracy through Twitter stream analysis. International journal of innovative technologies 4(8), 1317–1324 (2016)
43. Kumar, S., Morstatter, F., Liu, H.: Twitter data analytics. In: Springer Science and Business Media (2013)
44. Krueger, R., Sun, G., Beck, F., Liang, R., Ertl, T.: TravelDiff: Visual comparison analytics for massive movement patterns derived from twitter. In: Pacific Visualization Symposium (PacificVis), 2016 IEEE, pp. 176–183 (2016, April)
45. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., Cebrian, M.: Rapid assessment of disaster damage using social media activity. Science advances, 2(3), e1500779 (2016)
46. Landwehr, P.M., Wei, W., Kowalchuck, M., Carley, K.M.: Using tweets to support disaster planning, warning and response. Safety science, 90, 33–47 (2016)
47. Lee, S., Farag, M., Kanan, T., Fox, E.: A Read between the lines: A Machine Learning Approach for Disambiguating the Geo-location of Tweets. In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 273–274. ACM (2015, June)
48. Li, L., Goodchild, M.F., Xu, B.: Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartography and geographic information science 40(2), 61–77 (2013)
49. McDougall, K.: Using volunteered information to map the Queensland floods. In: Proceedings of the 2011 Surveying and Spatial Sciences Conference: Innovation in Action: Working Smarter (SSSC 2011), pp. 13–23. Surveying and Spatial Sciences Institute (2011)
50. Miller, H.J.: Time geography and space–time prism. The International Encyclopedia of Geography (2017)
51. Meier, P. Digital humanitarians: how big data is changing the face of humanitarian response. Crc Press (2015)
52. Mostafa, M.M.: More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications 40(10), 4241–4251 (2013)
53. Ngo, M.Q., Haghighi, P.D., Burstein, F.: A Crowd Monitoring Framework using Emotion Analysis of Social Media for Emergency Management in Mass Gatherings. arXiv preprint arXiv:1606.00751 (2016)
54. OSM Tasking Manager. Retrieved from http://tasks.hotosm.org/about. 16th September (2017)
55. Oxforddictionaries. Spanish Oxford living dictionaries. Retrieved from https://es.oxforddictionaries.com/translate/spanish-english/ahorita?locale=en. 16th September (2017)
56. Panagiotopoulos, P., Barnett, J., Bigdeli, A.Z., Sams, S.: Social media in emergency management: Twitter as a tool for communicating risks to the public. Technological Forecasting and Social Change, 111, 86–96 (2016)
57. Park, H., Reber, B.H., Chon, M.G.: Tweeting as health communication: health organizations' use of twitter for health promotion and public engagement. Journal of health communication 21(2), 188–198 (2016)
58. Pentina, I., Basmanova, O., Zhang, L.: A cross-national study of Twitter users' motivations and continuance intentions. Journal of Marketing Communications 22(1), 36–55 (2016)

59. Pereira, J., Pasquali, A., Saleiro, P., Rossetti, R.: Transportation in Social Media: an automatic classifier for travel-related tweets. arXiv preprint arXiv:1706.05090 (2017)
60. Perera, R.D., Anand, S., Subbalakshmi, K.P., Chandramouli, R.: Twitter analytics: Architecture, tools and analysis. In: Military Communications Conference, 2010-MILCOM 2010, pp. 2186–2191. IEEE. (2010, October)
61. Resch, B., Summa, A., Zeile, P., Strube, M.: Citizen-Centric Urban Plannin through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm. Urban Planning 1(2), 114–127 (2016)
62. Ribeiro Jr, S.S., Davis Jr, C.A., Oliveira, D.R.R., Meira Jr, W., Gonçalves, T.S.: Pappa, G.L.: Traffic observatory: a system to detect and locate traffic events and conditions using Twitter. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, pp. 5–11. ACM (2012, November)
63. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1500–1510. Association for Computational Linguistics (2012, July)
64. Saldana-Perez, A.M.M., Moreno-Ibarra, M.: Traffic analysis based on short texts from social media. International Journal of Knowledge Society Research (IJKSR) 7(1), 63–79 (2016)
65. Sheppard, S.A.: wq: A modular framework for collecting, storing, and utilizing experiential VGI. In: Proceedings of the 1st acm sigspatial international workshop on crowdsourced and volunteered geographic information, pp. 62–69. ACM (2012, November)
66. Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L.: Syntactic n-grams as machine learning features for natural language processing. In: Expert Systems with Applications 41(3), 853–860 (2014)
67. Steiger, E., Albuquerque, J.P., Zipf, A.: An advanced systematic literature review on spatiotemporal analyses of Twitter data. Transactions in GIS 19(6), 809–834 (2015)
68. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Social computing (socialcom), 2010 IEEE second international conference on, pp. 177–184, IEEE (2010, August)
69. Summa, A., Resch, B., GIS, G.Z., Strube, M.: Microblog emotion classification by computing similarity in text, time, and space. In: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 153–162 (2016, December)
70. Taylor, M., Wells, G., Howell, G.: Raphael, B.: The role of social media as psychological first aid as a support to community resilience building. Australian Journal of Emergency Management 27(1), 20 (2012)
71. Van Dijck, J.: The culture of connectivity: A critical history of social media. Oxford University Press (2013)
72. Wang, Q., Bhandal, J., Huang, S., Luo, B.: Classification of Private Tweets Using Tweet Content. In: Semantic Computing (ICSC), In: 2017 IEEE 11th International Conference on, pp. 65–68. IEEE (2017, January)
73. Whittaker, J., McLennan, B., Handmer, J.: A review of informal volunteerism in emergencies and disasters: Definition, opportunities and challenges. International journal of disaster risk reduction, 13, 358–368 (2015)
74. Williams, E., Gray, J., Dixon, B.: Improving geolocation of social media posts. Pervasive and Mobile Computing, 36, 68–79 (2017)
75. Wu, X., Bartram, L., Shaw, C.: Plexus: An Interactive Visualization Tool for Analyzing Public Emotions from Twitter Data. arXiv preprint arXiv:1701.06270 (2017)

76. Zhao, L., Chen, F., Lu, C.T., Ramakrishnan, N.: Dynamic theme tracking in Twitter. In: Big Data (Big Data), 2015 IEEE International Conference on, pp. 561–570. IEEE (2015, October)
77. Zhao, L., Chen, F., Lu, C.T.: Ramakrishnan, N.: Online Spatial Event Forecasting in Microblogs. ACM Transactions on Spatial Algorithms and Systems (TSAS) 2(4), 15 (2016)
78. Zhou, Y., Zhang, Z., Lan, M.: ECNU at SemEval-2016 Task 4: An Empirical Investigation of Traditional NLP Features and Word Embedding Features for Sentence-level and Topic-level Sentiment Analysis in Twitter. In SemEval@ NAACL-HLT, pp. 256–261 (2016)