Analyzing spatial data from twitter during a disaster

Analyzing spatial data from Twitter during a disaster

Luca Venturini, Evelina Di Corso
Dipartimento di Automatica e Informatica
Politecnico di Torino
Torino, Italy
Email: {luca.venturini, evelina.dicorso}@polito.it

***Abstract*—Social media can be an invaluable help in a mass emergency, but the information handling can be challenging. One major concern is identifying posts related to the area, or pinning them on a map. This exploratory study analyzes the spatial data coming with tweets during two natural disasters, an earthquake and a hurricane. Geo-tagged tweets confirm to be a small fraction of all tweets and disasters within a limited region appear to be a niche topic in the whole stream. The results can help researchers and practitioners in the design of tools to identify these messages.**

***Keywords*-spatial data analysis; Twitter; disasters and mass emergencies**

I. Introduction

In the last years, social media have met an unprecedented success and become a widespread, fast, and economical tool to access and share information. As such, they are an invaluable help in a mass emergency situation like that of a natural disaster, and are already actively used to communicate to the population involved in the preparation or in the aftermath of a disaster.

However, exploiting social media to lead decisions in a mass emergency presents multiple challenges, including parsing information, handling the information overload, and prioritizing different types of information, as discussed in [1]. One of these challenges is handling geographical information, which translates in identifying the content produced or related to a specific area, and placing this content on a map. Coping with these two issues would mean, in a scenario hit by a natural disaster, being able to have instantaneous and immediate feedback from the population in the area, possibly with reports of the damages, requests for support or availability of food, shelters, or help.

This work is an exploratory study on the quantity and the quality of the geographical data officially provided by a social media service like Twitter, in contexts of mass emergencies and natural disasters. The aim is to assess how many tweets contain geographical information and of what kind, and whether these tweets contain useful information for disaster relief and management.

II. Related work

In recent years, social media emerged as a potential resource to improve the management of crisis situations (e.g., earthquakes, tsunamis, floods). Authors of [1] have extensively investigated the subject, surveying the methods available in literature and their shortcomings, among which important issues of privacy, reliability, and accuracy of information. We refer the reader to this survey for a detailed review of the literature on the topic.

The spatiality of social messages has been addressed in previous works and raised concerns. The authors of [2] identify general spatial patterns in the occurrence of tweets through statistical analysis. The results show that messages near (up to 10 km) to severe crisis areas have a much higher probability of being related to the crisis. Although, in a review of the use of SMS and social media in the Haiti earthquake [3], the authors note that the value of such information at a detailed level was mainly useless on the field, while the aggregate information from various sources proved very helpful to focus work to areas where relief was most needed. Moreover, tweet datasets depict a specific period in time, typically defined by the use of particular hashtags. Thus, the analysis of social media during and after a disaster can resemble traditional media coverage, which has been often accused of paying attention to only the most sensational stories in a truncated timeframe [4].

Several works outside the scope of mass emergencies have already showed that social media contain very limited spatial information. In [5], for example, only 2% of the tweets in the study contained GPS location. The authors of [6] reported that only 0.42% of all tweets in their study had GPS-provided coordinates, and thus proposed a system to infer city-level location from the content of the tweet. In some contexts, these percentages do not impede a thorough spatial analysis. In the large dataset of tweets related to 2014 FIFA World Cup, for example, authors of [7] found more than 300 thousand out of 23,5 million tweets to be geo-located, which allowed a very large-scale analysis of the event.

III. Data collection and preprocessing

Twitter APIs provide access, with some restrictions, to the 140-character texts and the rich source of metadata associated with it. Among the optional fields in the metadata of a tweet we find geographical coordinates and a place id. The user, or rather the application posting on his or her

behalf, can choose to add the precise location given by the GPS sensors, or instead associate the tweet to a point of interest in the nearby, which translates in a place id. A place is defined as an area with predefined geographical bounds, and can range from a venue to an entire region or country.

The Twitter developers' documentation states that roughly 1% of all tweets are geo-located. In addition to these, the documentation hints that natural language processing is used to enrich the results of a geo-spatial search. Thus, a search of tweets around the coordinates of Rome would return tweets with coordinates in the area and possibly tweets mentioning Rome in the text, or tweets by users who set Rome as location in their profile. No filter by country code or state is possible with the public APIs.

The two datasets used in this study were collected as follows.

Ischia The dataset scope is to represent all the tweets in Italy on 21/08/2017[1]. On that date, an earthquake with magnitude 4.2 hit the island of Ischia, causing 42 injuries and 2 deaths and extended damage to the buildings [8]. The tweets were searched with two different queries. The first query searches for tweets in a radius of 600 km from the city of Rome, which covers approximately all the country. The second query searches for all tweets in Italian, which is a good proxy of all tweets in Italy, as Italian is spoken by the majority in Italy and little spoken abroad. All tweets belong to the day of 21/08/2017. The tweets coming from the first query were labeled as geo-referenced, with the broad meaning of having either geographical coordinates or NLP-enriched geographical references.

Texas The dataset aims to represent tweets in an area largely affected by hurricane Harvey, an Atlantic hurricane formed on 17/08/2017 that has caused the death of 78 people and the evacuation of more than 30,000 [9]. The tweets were downloaded within a circle of 300 km centered in Rockport, Texas, the city where the hurricane made the first landfall. The radius was set as to cover all the Texas coastline. The date of the tweets is 27/08/2017, the day after the initial landfall, when hurricane Harvey reclassified to storm and the heavy raining caused widespread floods in the whole state.

IV. Discussion

The Ischia dataset contains 409392 tweets, of which 3566 (0.87%) are geo-referenced. This percentage is similar to the one stated by Twitter. The earthquake in Ischia hit the island at 20:57. The results for a search on a given date go until 2 in the morning, which means the dataset contains 19 hours of tweets written before the earthquake, and 5 after. 67813 tweets in the dataset contain one or more hashtags, which thus offer a good sample of the topics discussed on the platform. The word clouds in Figure 1 show the most frequent hashtags, where the size of the word is proportional to its frequency. Figure 1a depicts the most frequent hashtags in non-geo-referenced tweets before the earthquake. TV programs and football teams take most of the social interest, with a little attention to exceptional events like the solar eclipse, happening at 20:26 local time and not visible from this timezone. The first tweet on the earthquake appears at 20:59. Figure 1b shows the frequent hashtags from then to 1:59. Ischia is indeed the most frequent hashtag, and we can spot other related terms like *terremoto* (earthquake) and Casamicciola, the location of the epicenter. Most of the word cloud, though, is yet crowded with mentions of TV shows broadcasted in that evening. The number of tweets containing the words Ischia, Casamicciola, or synonyms of earthquake is 9668, 5.2% of the tweets after 20:57. Care must be taken in taking this number as a measure of the interest to the event, as most tweets are made of stopwords and might follow up a conversation, and are therefore less likely to contain keywords or refer directly to the fact.

Figures 1c and 1d show frequent hashtags for geo-referenced tweets, respectively before and after the earthquake. These two clouds appear to speak of the same topic, and the abundance of English terms and the vocabulary used suggests that these are mainly promotional tweets that sponsor touristic areas, in which the earthquake is mostly ignored. Indeed, the percentage of geo-referenced tweets referring to the earthquake after 20:57 is only 2.8%.

Tables I and II show, respectively, the top 10 domain names of links in geo-referenced and non-geo-referenced tweets in the Ischia dataset. Links can be a hint on the kind of content posted, e.g. a video in the case of a link to Youtube, on the app used to post the tweet, e.g. Swarm in the case of *www.swarmapp.com*, or on an interaction on the Twitter platform, e.g. a reply to a tweet contains a link to the original tweet. The top 10 websites linked in non-geo-referenced tweets (Table II) depict the behavior we would expect from a user of Twitter: an high interaction with other users (*twitter.com* is the most linked domain), videos (Youtube is the third most linked domain) and links to other popular social platform like Facebook and Instagram, that are respectively the second most and the fifth most linked domains. The top domains in geo-referenced tweets (Table I) are very different, and in the top 10 list we do not see any of the top 3 websites linked in non-geo-referenced tweets, i.e. Twitter itself (with retweets), Youtube and Facebook. The top two domain names in geo-referenced tweets come from apps that post tweets from third parties, like Instagram or Foursquare (Swarm). These differences are a second clue of the different nature of geo-referenced tweets, and suggest that they are not a representative sample of the whole stream of tweets.

Texas dataset is made only of geo-referenced tweets, but with different granularities of geographical information. Its

[1]note that Twitter developers' guide states that some tweets and users may be missing from search results

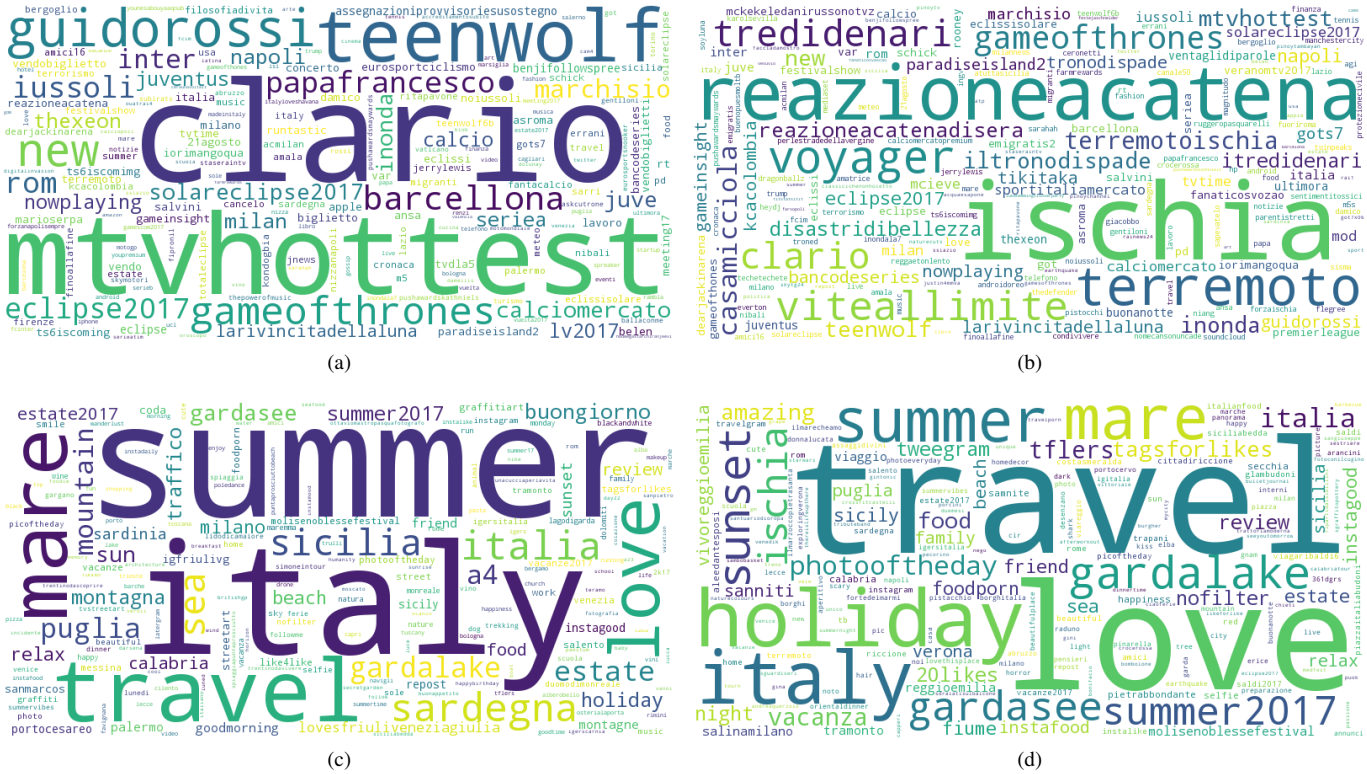


Figure 1. Frequent hashtags in the Ischia dataset in non-geo-referenced tweets, (a) before the earthquake and (b) after the earthquake, and in geo-referenced tweets, (c) before the earthquake and (d) after the earthquake.

Table I
TOP 10 OF DOMAIN NAMES LINKED IN GEO-REFERENCED TWEETS IN ISCHIA DATASET

urls	#
https://www.instagram.com/	845
https://www.swarmapp.com/	51
https://goo.gl/	25
https://www.trendsmap.com/	16
http://dlvr.it/	8
http://www.olevanometeo.it/	5
http://n.mynews.ly/	5
https://www.gpone.com/	4
http://www.montedarena.com/	3
http://crwd.fr/	3

Table II
TOP 10 OF DOMAIN NAMES LINKED IN NON-GEO-REFERENCED TWEETS IN ISCHIA DATASET

urls	#
https://twitter.com/	30890
http://fb.me/	21139
http://youtu.be/	7011
http://ift.tt/	5484
https://www.instagram.com/	4848
https://goo.gl/	4516
http://bit.ly/	3785
http://dlvr.it/	3024
http://l.ask.fm/	2271
https://curiouscat.me/	1890

Table III
TOP 10 OF DOMAIN NAMES LINKED IN TWEETS IN TEXAS DATASET

urls	#
https://www.instagram.com/	4000
http://waterdata.usgs.gov/	871
http://bit.ly/	360
https://www.swarmapp.com/	240
http://bubly.us/	165
https://mesonet.agron.iastate.edu/	106
http://untp.beer/	94
https://twitter.com/	70
http://www.allaboutbirds.org/	42
http://tour.circlepix.com/	13

study can lead to a better assessment of the quality of the geographical information inside a tweet corpus. Of 6938 tweets resulting from the geographical query, 1275 have geographical coordinates, of which 1240 have also a place id. The remainder of the tweets is supposedly assigned to a location through named entity recognition, though no metadata gives details on this. Places divide further in three types (i.e. city, admin, and neighborhood), of which city is the most numerous, with 1039 records. Figure 2 lists the names of top 10 locations with their frequency. Not surprisingly, 9 out of 10 are cities, although Texas is the second most frequent place. Already the fifth most frequent city, Conroe, has less than 25 geo-tagged tweets, and the tenth, Mission,

Figure 2. Frequency of top 10 locations for Texas dataset

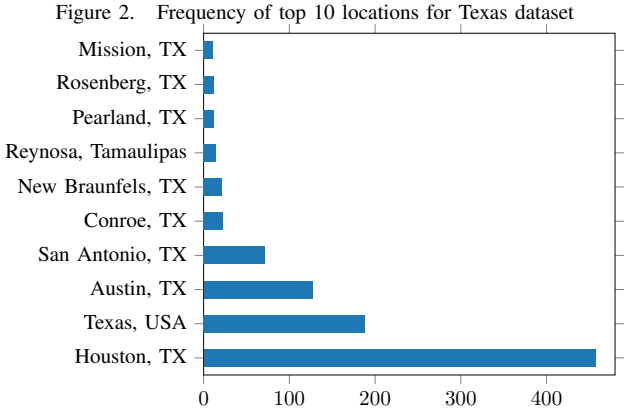
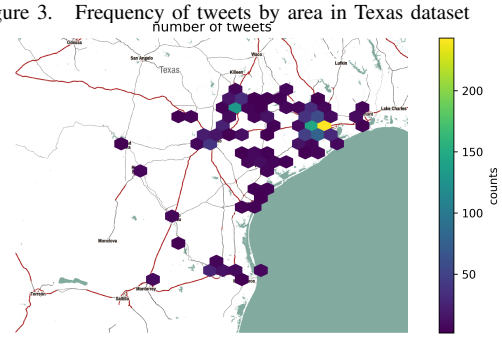


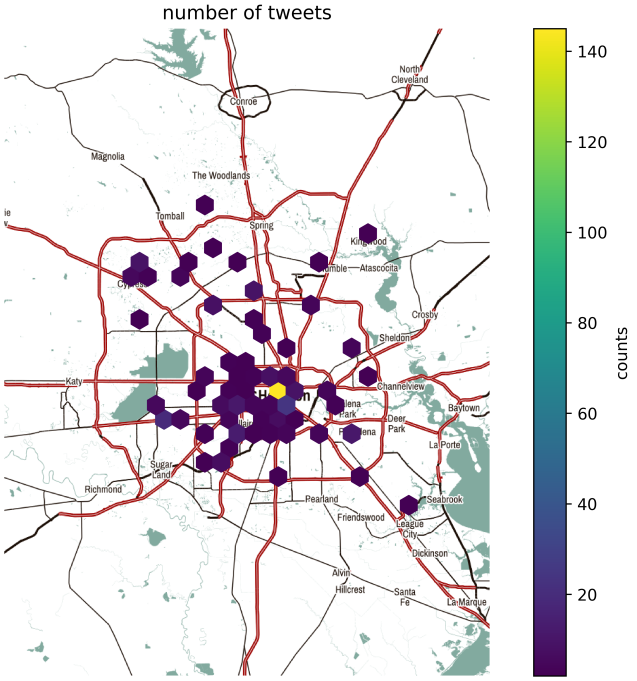
Figure 3. Frequency of tweets by area in Texas dataset



has only 11 of them. Figure 3 shows a frequency map of tweets with coordinates, binned in hexagonal cells of equal size. A large part of the map does not see any tweet, and most of the cells have less than 50 tweets. The most of them are located in the bigger cities, probably following the distribution of the population, and only the cell over Houston sees more than 200 tweets. In the city, we see again a similar behavior, with more than 120 tweets located in the city center and the rest scattered around (Figure 4). If, as it is likely, people tweeted from the blank spots of the map, their tweet was ignored by this geographical search and there might be no information to link it to this area of the world.

Table III lists the top 10 domain names in links in this dataset. Similarly to Table I, many of them link to Instagram and Swarm, and Facebook and Youtube are totally absent. Among them we can also notice domains of public services, that post warnings and tweets of public interest. Figure 5 shows the most frequent hashtags. The vocabulary used, differently from Ischia, seems to be largely related to the event. This can be due to the predictability of the hurricane, to the fact it is weather-related, or to the large extent of the area and population affected. The hashtags seem to belong mostly to weather warnings and automatic reports, with tags that refer directly to the city or area involved or in several cases are specific to the USGS service [10]. This, in line with the findings in Ischia dataset, shows how geo-referenced tweets belong to a special subset of users, and are not apt to describe the whole population, at least in a special situation like that of a natural disaster.

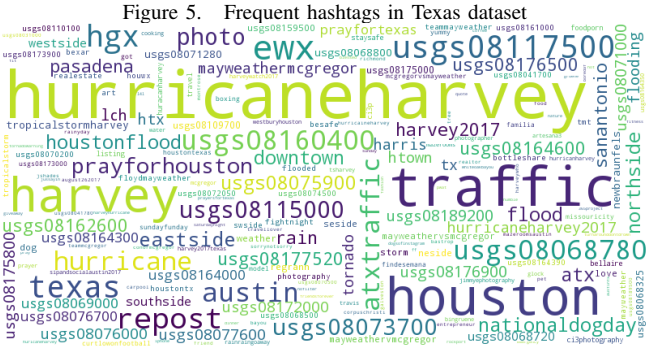
Figure 4. Frequency of tweets by area in Texas dataset (place_id=Houston, TX)



V. Conclusion

In this exploratory study, we shed light on a number of features of the social medium during a mass emergency. First, we found that only a small fraction, under 1%, of tweets is georeferenced, which is in line with the numbers in [5], [6]. This implies that a spatially bounded search, like the one that has produced our Texas dataset, excludes from the results the most of messages and is therefore not recommendable in the handling of an emergency. Furthermore,

Figure 5. Frequent hashtags in Texas dataset



less than a fifth of these tweets had precise GPS coordinates associated with them. Moreover, we have shown as the kind of results that return from a geographical search belong to special categories of users or services, e.g. tweets from third party apps like Instagram or official weather warnings. In the case of the Ischia earthquake, these tweets did not resemble, in their contents and in the vocabulary used, the entirety of the community. Researchers and practitioners should therefore be aware of the bias introduced in making a search of this sort.

Future research should aim at searching for tweets related to the emergency in ways that do not rely on spatial information, like for example [11]. As evidenced in the Ischia scenario, these methods should be able to identify a stream of tweets that may not surge in the trends or make exclusively use of hashtags, as already suggested in [4].

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700256 ("I-REACT" project).

REFERENCES

[1] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.

[2] J. P. De Albuquerque, B. Herfort, A. Brenning, and A. Zipf, "A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management," *International Journal of Geographical Information Science*, vol. 29, no. 4, pp. 667–689, 2015.

[3] J. Dugdale, B. Van de Walle, and C. Koeppinghoff, "Social media and sms in the haiti earthquake," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12 Companion, 2012, pp. 713–714.

[4] K. Crawford and M. Finn, "The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters," *GeoJournal*, vol. 80, no. 4, pp. 491–502, 2015.

[5] S. H. Burton, K. W. Tanner, C. G. Giraud-Carrier, J. H. West, and M. D. Barnes, ""right time, right place" health communication on twitter: value and accuracy of location information," *Journal of medical Internet research*, vol. 14, no. 6, 2012.

[6] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[7] X. Xiao, A. Attanasio, S. Chiusano, and T. Cerquitelli, "Twitter data laid almost bare: An insightful exploratory analyser," *Expert Systems with Applications*, vol. 90, pp. 501–517, 2017.

[8] E. Povoledo. (2017, 8) Deadly earthquake hits italian island of ischia. [Online]. Available: https://www.nytimes.com/2017/08/22/world/europe/italy-ischia-earthquake.html

[9] E. Moravec. (2017, 10) Storm deaths: Harvey claims lives of more than 75 in texas. [Online]. Available: http://www.chron.com/news/houston-weather/hurricaneharvey/article/Harvey-Aftermath-Houston-police-officer-dies-19-12159139.php

[10] USGS. U.s. geological survey. last access: October 2017. [Online]. Available: http://www.usgs.gov

[11] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises," in *International AAAI Conference on Web and Social Media*, 2014.