Accepted for publication at IEEE RCIS 2017, Brighton, UK

IMEXT: A method and system to extract geolocated images from Tweets — Analysis of a case study

Chiara Francalanci, Paolo Guglielmino, Matteo Montalcini, Gabriele Scalia, Barbara Pernici
Politecnico di Milano - DEIB
Milan, Italy
{chiara.francalanci, barbara.pernici}@polimi.it
{paolo.guglielmino, matteo.montalcini, gabriele.scalia}@mail.polimi.it

***Abstract*—Extracting useful information from social networks raises several challenges that still represent open research issues. In this paper we focus on the problem of extracting geolocated images from Tweets to support emergency response. A Tweet analysis process is discussed, focusing on the selection of posts, their geolocation based on their text content, and the subsequent analysis of the images linked by geolocated tweets. A prototype system has been built and tested on a case study based on the Tweets posted in the two days after the earthquake that occurred in Central Italy in August 2016. Results indicate that focusing on images linked by geolocated tweets represents a good criterion to identify useful information that can aid emergency response.**

***Keywords*—Emergency Management Systems (EMS), Tweet selection, geolocation, media extraction**

I. Introduction

When a natural disaster occurs, emergency management is a crucial aspect to consider. Nowadays, with the support of information technology, the aim is to improve this important process to mitigate its effects, being prepared and organizing rescue and aid efforts as soon as possible. An important aspect to achieve these goals is information exchange and understanding, processing relevant information and coordinating actions in the shortest possible time.

In order to develop an effective management plan we need to collect enough relevant information in some way. Latest technologies can help us a lot to obtain real-time and almost real-time information, for example images from satellites, GPS, IoT devices, and the Internet, in particular social media. During an emergency, communication can be difficult for various reasons, therefore these sources of information can be all helpful, but they need to be integrated each other to achieve better performance and provide real time information that can be crucial during extreme events.

An emergency management process is widely represented as composed by four phases: preparedness, response, recovery, and mitigation. In this paper we focus on supporting the emergency response phase, i.e., the phase that begins when an emergency occurs or, if the event is predicted, immediately before the event.

Nowadays there is not a standard system to manage emergency, however many different project exist. We take into account the Copernicus Emergency Management Service, which is an EU programme aimed at developing European information services based on satellite Earth Observation and in situ (non space) data [2].

Since February 2015, Copernicus EMS Rapid Mapping (RM)[1], part of the Copernicus Emergency Management Service, an EU programme aimed at developing European information services based on satellite Earth Observation and in situ (non space) data [2], provides information about the extent and impact of disasters (natural or man-made) to European Civil Protection Authorities and to International Humanitarian Aid organizations. The information has a high quality, is reliable and timely. It covers in particular: floods, earthquakes, landslides, severe storms, fires, technological disasters, volcanic eruptions, humanitarian crises and tsunamis. The system is currently facing specific challenges and operational gaps. Starting from the Mapping Component we can say that the main source of data used by EMS are satellite images acquired on purpose following a map request. Timeliness is one of the main user requirements of the EMS, although it is not yet fully achieved. In fact it is not unusual to experience delays up to 72 hours, while users have expressed the requirement of receiving first crisis information within the first 24 hours after the disaster. Many factors can cause this delay, for example satellite orbital constraints and bad weather conditions preventing the collection of optical images. Moreover in case of large-scale disasters map production throughput could be insufficient to analyze the affected area in a short period of time. Lastly, satellite-based emergency maps have also limitations in quality.

As an additional information source, social media have been considered in the literature, as further discussed in Section II. Social media are Internet services that allow the management of social relations and communications. One of the main characteristic of social media is that they allow users to generate and share content (text, multimedia, etc...). Twitter is one of the most widespread social media sites worldwide. It is an online free social networking and microblogging service, it allows users to publish text messages of up to 140 characters, which are called *tweets*. Twitter is entirely based on an Open Source architecture [16]. In the literature, tweets have been mainly used for early warning and geolocation of events,

[1]Rapid Mapping is the process of enrichment of a geographical map with some details (e.g., picture of unusable road, damaged areas) within the Copernicus Emergency Management Service (EMS) [2]

978-1-5090-5476-3/17/$31.00 2017 IEEE

as discussed in the following. However, the extent to which the information extracted from tweets is actually useful can be an open issue, especially when operating activities focus on multimedia data. However, an analysis of recent tweets in emergency situations, has made it clear that interesting information, useful in particular for rapid mapping in cases when satellite or aerial images are not available, is not only contained in tweets text, but also in associated images and links.

Several types of messages can be conveyed through social media. A reference classification is presented in a governmental report [10], discussing the different types of messages. In particular, messages can help establishing situational awareness and also assisting in damage estimation projects on the basis of shared media (e.g., pictures).

The objectives of this paper are to focus on information that can be derived from tweets, and in particular, to analyze the characteristics of tweets in an emergency event with an earthquake as a case study, to study how to support rapid mapping with images extracted from social media, geolocated by analyzing the text of associated tweets. Moreover, in the paper we describe the IMEXT (Image Extraction from tweets) environment, illustrating its main characteristics and the results obtained from a large case study.

In the present paper, we focus on two main goals: to extract relevant tweets and geolocalize them and to extract information, and in particular media contents, that can support the rapid mapping tasks in the emergency management response phase.

In the following sections, first we discuss the state of the art focusing on social media support for emergency response in Section II. In Section IV, we illustrate an extraction process for tweets, focusing on how to extract geolocated images. In Section V we illustrate the developed analysis environment, illustrating empirical results and discussing their validity in Section VI.

II. Related work

The work described in this paper is part of a new H2020 European project E^{2}mC Evolution of Emergency Copernicus services, which has started in November 2016, to support early warning and rapid mapping with information extracted from social media. In the initial phases of the project, a critical review of crowdsourcing and social media use in emergency management has been performed [4].

The exploitation of social media and related information in emergency management represents a subject of a vast literature, facing both technical and emergency-specific challenges. This literature is not limited to the past decade, when large-scale social media such as Facebook and Twitter have become prevalent, but dates back to the late '90s, when social interactions were already supported with ad hoc newsgroups, email clients or Web sites. Clearly, the amount of information, especially multimedia information, that is currently available is by far larger than the amount of information that was available in the late '90s. As a matter of fact the focus of the early studies was on how to make information available in the first place. However, these studies have played an important role in demonstrating the importance of collecting information from the people involved in a crisis situation and using technology to support crisis management before, during and after the response phase. This and the more recent literature is extensively surveyed in [8]. This survey paper provides the most important and most widely cited reference framework to understand the collective body of research on social media and emergency management until 2014. The paper supports the following key statements that represent our starting point:

- Both on a global and European scale, in most cases Twitter can function as the main source of social media information. Although other sources of social media information are considered in literature, they tend to be country-specific (for example, the less intensive use of Twitter in Germany) or emergency-specific.
- Different users (that is, different official response agencies and companies) have different expectations and requirements and there is no one-fits-all solution. Particularly, the information that different stakeholders find useful and look for depends on their objectives.
- While multimedia information is often acknowledged as important, it is seldom the focus of previous literature, which is far more concerned with text information and related NLP (Natural Language Processing) techniques. The issue of exploiting multimedia information gathered from social media for emergency mapping purposes is rarely addressed and largely unexplored.

The literature published after 2014 (and, thus, excluded from [8]) provides interesting insights from the perspective of emergency management stakeholders. In [8], authors note that although emergency management stakeholders acknowledge the potential value of social media, they are also skeptical towards the dependability of social media information (see, also, the survey dedicated to the issue of "trust" published in [14]). They express concerns for the information overload that they experience when looking at social media during a mass disaster, as well as various organizational issues related to roles, responsibilities and liabilities. A recent research stream deals with some of these issues, with an underlying focus on technology usage. For example, the research published in [12] studies the issue of information overload in the use of social media by emergency managers. The paper discusses the results of a survey of 477 U.S. county-level emergency managers that examines the relationship of the perception of information overload as a barrier to social media use for gathering information, to the intention to use social media. Results show how "managers' perception of information overload as a barrier to use is negatively related to their intention to use it, while perceptions of the usefulness of NLP technologies are positively related to intention to use." In [15], authors provide "a content analysis of Nepal Police Tweets from the aftermath of the 2015 Nepal earthquake." The paper "uses

the typology of convergence behaviors in emergency response as an attempt to categorize the public interaction with social media platforms." In [11], authors discuss the Social Media Intelligence Analyst, defined as "a new operational role within a State Control Centre in Victoria, Australia dedicated to obtaining situational awareness from social media to support decision making for emergency management." The paper outlines "where this role fits within the structure of a command and control organization, describes the requirements for such a position and detail the operational activities expected during an emergency event." Even research on the software tools for social media and emergencies is taking a stakeholders' perspective. For example, in [7], the authors criticize existing tools as they do not entirely meet the needs of PIOs (Public Information Officers) involved in emergency management. The paper describes a new prototype tool "that supports the work practice of emergency public information officers and their need to gather, monitor, sort, and report social media activity." A recent survey [9] concludes that "there is lack of knowledge towards investigating empirically the intricate and dynamic nature of crisis management operations and to determine how knowledge coordination can assist managers in enhancing emergency management task performance."

In the context of the E^{2}mC project, we take the perspective of the rapid mapping specialists. Our goal is to take a first step towards helping our stakeholder in exploiting multimedia information gathered from social media for emergency mapping purposes, focusing in particular on natural disasters. In this paper, we discuss the results that we have obtained in the case study described in the next section.

III. Case study description

A. Description

Following the recent happening of seismic events in Italy, we decided to concentrate our work on a recent event in central Italy. In particular, we collected Twitter data about the earthquake measuring 6.0 on the moment magnitude scale, that hit Central Italy on 24 August 2016 at 03:36:32 CEST (01:36 UTC). Its epicentre was close to Accumoli, approximately 75 km southeast of Perugia and 45 km north of L'Aquila, in an area near the borders of Umbria, Lazio, Abruzzo and Marche regions. Several damages were indicated in the town of Amatrice, near the epicentre, in Accumoli and Pescara del Tronto.

B. Tweet analysis

1) Detection: As illustrated in Section IV, we collected tweets relevant to the event with a keyword-based approach. As mentioned in previous work, e.g. [13], the impact on the number of tweets of each tremor is evident, as shown in Fig. 1.

It has be noted that several tremors and aftershocks occurred within a brief time period. Only the main ones, according to the classification of National INGV Institute[2], are evident in the tweets analysis (magnitude 6.0 and 5.4, respectively).

[2]http://cnt.rm.ingv.it

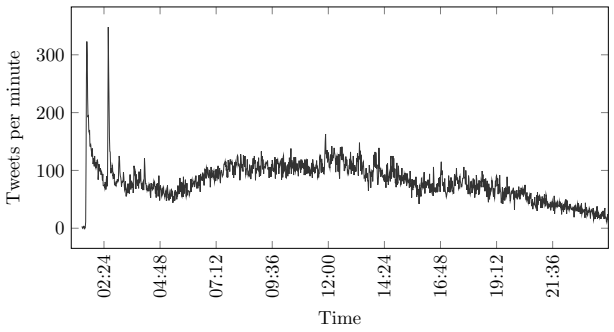

Fig. 1. Numbers of tweets in Italian talking about the earthquake that occurred in central Italy on August 24, 2016.

A total of 152,062 tweets were collected in a two-day time frame after the first high-magnitude tremor on August 24. A 40 MB database was obtained, with 323,051 records, containing information about the posts and their authors, attached images, links to other sites (including links to YouTube videos and other social media).

2) General Tweet analysis: Here we present some general characteristics of the collected tweets.

- language: 86.6% are in Italian.
- Geographical coordinates: only 533 tweets have geographical coordinates, approximately 0.35%, in line with what reported in [6].
- Links: 51.23% of tweets have a link. Links are mainly to Facebook (20,328 links) and news sites (15,989 links).
- Media: 26,914 (17.7%) tweets contain images. However images loose all metadata, including their geographical coordinates, when stored by Twitter inside the repository that is used to respond to data retrieval queries from the official Twitter APIs.
- Geolocated tweets with links: 53.54% out of the 533 tweets with geographical coordinates have a link.
- Instagram: 846 Instagram pictures were linked by a tweet, 68.16% of them are geotagged, 23.67% of them are not geotagged.

IV. Extraction methodology

As presented in the previous section, social media can provide relevant information for emergency management. We focus on the improvement of emergency maps. The key idea is to leverage social media popularity and diffusion in order to collect valuable information directly from citizens, which can play the role of a distributed sensor network. Our goal in this paper is to thoroughly analyze our case study to understand whether and to what extent social media information can be useful to support existing mapping systems and activities. As a general observation, given that our ultimate goal is rapid mapping, multimedia information represents the most useful type of information. For example, a picture of a building taken after the event can clearly show the extent of the damage of the event on that building and, thus, support mapping specialists

in deciding whether the building should be marked in green (no damage), yellow (some damage), or red (destroyed) in the map that they are preparing. Similarly, a picture of a road could help them mark it as unusable on the map. In order for them to do this, they have to 1) identify the picture within a mass of pictures gathered from Twitter (around 26,000 in our case study), 2) geolocate the image, 3) possibly group the image with other related images (same building taken from a different angle, same street including the building and adjacent buildings, etc.) and 3) look at the image(s) and make their decision on the extent of the damage. A full support to this set of activities raises a number of research questions. The first ones are whether pictures useful for mapping purposes exist, how many they are in a typical event that requires satellite mapping, how they can be extracted and geotagged automatically. To make initial field experiments with our stakeholders, a repository of useful and geotagged images would be an excellent starting point, as they have declared that the manual inspection of potentially useful and geotagged images would be feasible and, in fact, could aid their daily work.

In this section, we illustrate the main steps of the proposed extraction methodology, shown in Fig. 2. In Section V we illustrate a system developed to perform automatically all the steps of the methodology, except the last one of image analysis and fine grained localization.

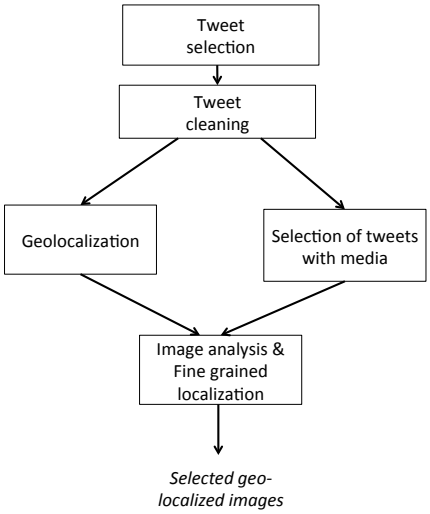

Fig. 2. Image extraction methodology from tweets.

A. Tweet selection

In order to identify a potential emergency situation, tweets are retrieved using a keyword-based approach. A multilingual set of keywords has been defined for major natural emergency events (the keywords for earthquakes are shown in the appendix). The languages being used are English, Italian, French, and German.

Such an approach enables the identification of tweets in a multilingual environment in general, as it is possible to automatically translate the seed terms in most languages with a good accuracy for the purpose of tweet selection. The relevant information on tweets are collected, including the tweet time, text, image, coordinates (latitude and longitude) if available, and links.

B. Tweet cleaning

A first cleaning phase is performed to exclude tweets that are not carrying new information. In the current case study, we excluded all retweets. Given the location of the event, we also excluded all the posts that were not written in Italian, since, as it was mentioned in the previous section, in this way we still retain the vast majority of posts, and exclude tweets from users that are not likely to be present in the location of the event.

C. Geolocalization

As mentioned previously, geolocalization of tweets from shared locations can be extracted only in a tiny number of tweets (0.35% in our case study). As a consequence, in this step we use the the text of tweets to extract location information. As mentioned in the related work section, some authors propose to use the geolocation available from user profiles, but it has been observed that only a few users provide geolocation information in their profiles and, when available, this information does not provide the exact coordinates of the position where the tweet (and related picture) was sent from.

The proposed approach is to perform geotagging based on the text of the tweet. Text is analyzed by searching for specific keywords that could indicate a precise geographical position, such as “road,” “street,” “church,” etc. The complete keyword list is provided in Appendix B. Some further cleaning operation might be needed to eliminate language-specific ambiguities. For instance, in Italian “street” is translated into “via”, which has multiple meanings, one of them being the same as in the English expressions “via msg”, “via internet”, etc. Since these expressions are very common, they have been blacklisted and are used to further filter the tweet collected with the keywords reported in Appendix B.

We performed entity recognition by using a text analytics tool capable of doing named entity recognition in different languages including Italian. Within the recognized entities we found locations. If the text contains more than one location we concatenate them in a single string, because for example a tweet may contain both the city name and the address, giving us a more accurate location. A geolocation service is used in order to retrieve, starting from the previous mentioned string, the coordinates of the location, and in this way a latitude and a longitude are attributed to the tweet.

After this process, in our case study we increased the number of tweets that have a coordinate attribute from 533 (0.35%) to 3,357 (2.20%). These tweets provide a link to 973 pictures.

D. Selection of tweets with media

The textual information contained in tweets resulted in general not very significant to support the rapid mapping

activity. This supports the idea of focusing on extracting tweets with images or links to images.

In addition to images, tweets link YouTube videos (1299 in our case study when all tweets are considered). We have manually inspected the videos linked by tweets in our case study. This analysis has shown that videos do not provide information useful for the mapping activities, as they mainly focus on details or interiors. A notable exception are videos filmed by drones, which are associated with precise geographic information as they were provided mainly by Public Authorities working on the emergency. However, these videos were available only at the end of our two-day observation period, so their usefulness for the rapid mapping activity is limited and they have not been considered further in the process.

E. Image analysis and fine-grained localization

In this phase, the images which are associated with geolocalized tweets are analyzed. A further filtering is performed on the result searching for tweets accompanied by images using words that limit the images to the ones containing text referring to damages, using a set of rapid mapping keywords (listed in Italian in the appendix).

We found that of the previously 973 filtered images, we were able to associate a location with 39 of them.

Examples of the images obtained with this last filtering are shown in the following. An example tweet from the news is presented in Fig. 3. The picture clearly shows that a road is blocked and the text of the tweet associated with this image provides the name of the road. In this case, the geolocalization is rather approximate as it is limited to the name of a very long main road. However, knowing that a main road is blocked is potentially very important, especially in the early phases of the emergency (the photo was posted within a short time from the event).

A more accurate geolocation is shown in Fig. 4. In this case, the street name is recognized, but not the exact location of the church (resulting in a 180 m. distance from the exact location).

Manual work, or possibly an image processing tool, could be very useful to remove duplicate images or images that are not significant for rapid mapping. Furthermore, if the image is significant, users can trace back to the tweet, read the text and contact the author of the tweet in order to ask for more information.

We have performed this work manually on the 973 images that were associated with the tweets filtered with location- and damage-related keywords. In addition to eliminating duplicates, we have also filtered out images that were clearly useless (e.g., screenshots of tweets). After this manual inspection, we were left with 541 selected images, 20 of which were among the 39 images filtered by means of damage-related keywords.

Overall, the semi-automated work described so far has allowed us to automatically retrieve 973 geolocated images with geographical coordinates, out of which 541 were potentially useful for rapid mapping purposes. Attempts to further filter these images with damage-related keywords have not proved effective as they eliminate the majority of potentially useful images and do not guarantee that remaining few images are actually useful. In the next sections, we discuss a more sophisticated approach supporting the identification of pictures that are providing information that is useful for damage assessment.

Fig. 3. Example of approximate geolocation: lat:42.79167460 lon:13.09473350

Fig. 4. Example of more accurate geolocation: lat:43.29958590 lon:13.44990280

V. Extraction tool environment

In order to develop a system which is able to perform the functionalities pointed out in the previous section, we set up the IMEXT (Image Extraction from Tweets) environment (Fig.

5), that leverages on integrating existing tools and developing some components to support the above mentioned analysis. The development was done while studying the specific case study, but it can be easily generalized to analyze other types of natural emergencies and extended to other languages in addition to Italian.

The IMEXT environment allows the user to perform automatically most of the steps illustrated above, from tweet selection to fine-grained geolocalization and the selection of tweets with media. The main components of IMEXT are illustrated in Fig. 5. As mentioned before, existing tools have been integrated to support the methodology illustrated in the previous section.

The integration modules were implemented in Java, storing data using MySQL and using NetBeans as an IDE to import libraries using Maven. As external services, we used only free services that provide well documented APIs for developers, accessing these services using the HTTP protocol.

Twitter provides official REST APIs for developers to allow the use of its services in third-party applications [16]. Twitter provides a set of operators[3] that can be combined with keywords to perform complex queries to Twitter's database, in this way we can have a first-level filter to select interesting tweets.

Regarding the lexical database, while in papers like [6] lexical data are based on WordNet[4], in this paper we adopted an event-specific glossary tailored to emergency events. The generic glossary for different event categories is under development, based on the Torcia glossary [5]; the specific glossary adopted in this paper is illustrated in the appendix. Such glossary is multilingual and the goal is to have the possibility to translate it automatically into other languages, to be able to support with this method events occurring in any place of the world, as the rapid mapping environment within the Copernicus Emergency Management Service has a satellite coverage worldwide [2].

The simplest raw implementation of the Twitter crawler leverages Twitter's streaming APIs[5]. In the specific case study, we downloaded using Twitter queries all the tweets identified as relevant to the event with the above mentioned queries, using the Twitter search API.

To support geolocalization, we used a text analysis tool to identify the components in the text and Geocoding to transform a (postal) address description to a location on the Earth's surface (spatial representation in numerical coordinates). Rosette text analytics is a robust toolkit for processing language, documents and names[6]. It is a service that can perform various tasks, transforming unstructured text into valuable information. In fact, Rosette is able to process text strings and extract proper nouns, geographic locations and other features. In the context of our analysis we used Rosette to extract geographical entities from tweets' text. When several geographical entities were identified in a post, they were concatenated to get a string with all localization information (for instance, name of a town and street address), to be processed in the following steps. Relying on Rosette entails some drawbacks, first of all the number of requests we can perform on Rosette's server is limited and can not be enough for a wide scale application. Then there is the latency, in fact every request and response takes a little bit of time (milliseconds) which added to many requests become seconds. Finally we noticed that Rosette's service is not able to recognize ambiguities in some languages, for example the name "San Benedetto del Tronto" which is an Italian city is cataloged as a person instead of geographic location (a similar problem can be seen in Fig. 4, where Rosette identifies correctly the name of the town and the name of the street as geographic names, but does not recognized the name of the church "Chiesa San Giovanni" as a location, but rather as a person name, thus losing in precision).

As one of the goals of our project is to deduce geographical coordinates from tweets, we rely on Google Maps[7] for the second step, that is extracting coordinates from geographic names provided by Rosette, and in particular the Geocoding service, that consists in extracting coordinates (latitude and longitude) from city names and addresses. Also this service has some limitations of use imposed by Google, but they are much less stringent than those of Rosette. However, it could be possible to move this process locally instead of relying on the web, for example configuring Geonames[8] database on our platform; this would increase the performance of the system, but we would not be able to cover all the world's geographic location because of the excessive size of the database needed.

To represent our data on geographical maps we used the Google Maps service, which allows producing custom maps that can display our data of interest.

The geolocation process for all collected non-geolocalized tweets required about half an hour and as a result we obtained the geographical coordinates of about 3,000 tweets.

The components we described above interact with each other in different ways in order to achieve complex tasks. There is a main component we called "Core" which is responsible of the coordination of the others, including the crawlers, the storage interface and the HTTP interface. The Core decides when to start/stop the crawlers and connects to the database to store the collected data, after that queries the storage in order to obtain data to analyze with the help of external services provided by the HTTP interface. The details of this architecture and the flow of information are shown in Fig. 6.

VI. Validation

In this section, we analyze the validity of the obtained results with respect to the selection of useful images using the IMEXT process and tools.

[3]Source: https://dev.twitter.com/rest/public/search

[4]https://wordnet.princeton.edu/

[5]https://dev.twitter.com/streaming/overview

[6]Source: https://www.rosette.com/

[7]https://developers.google.com/maps

[8]http://www.geonames.org

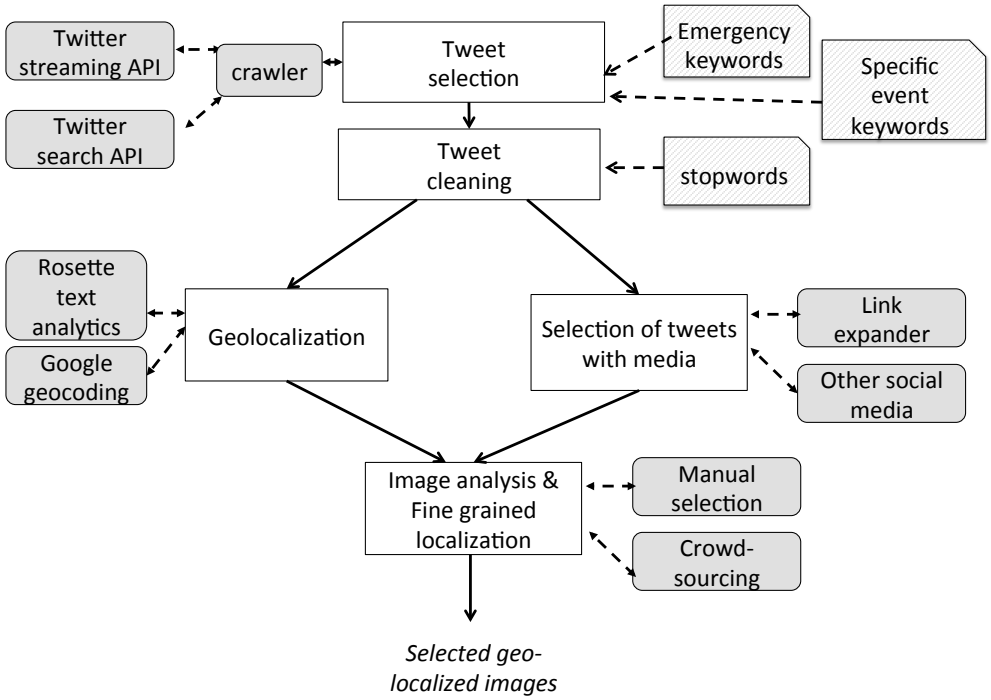

Fig. 5. The IMEXT tool environment for the extraction methodology

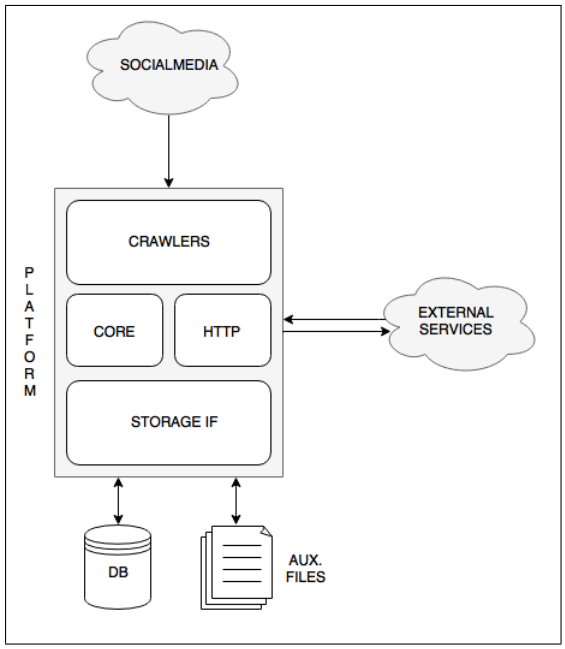

Fig. 6. IMEXT system architecture

As mentioned above, the initial image set contains 973 images filtered from the tweets potentially related to the event and geotagged by means of the text analytics techniques provided by the IMEXT tool.

For validating the approach and evaluating the obtained results, we manually analyzed this set of initial images to derive how many could be relevant in the considered context and to compare the extracted number with the number derived from the manual analysis. In particular, we will focus on the evaluation of precision and recall.

Initially, we removed duplicate images. These should include not only identical images but also different versions of the same original photos in terms of dimensions, resolution and colors. For this purpose, the perceptual hashing algorithm pHash[9] has been employed. We preferred a conservative approach which minimizes the risk of false positives, thus leaving some duplicate images in the result. When more images were available, the highest resolution one has been chosen. This step has reduced the 973 images initially considered to 783 images, with the corresponding tweets which constitute the *main* set.

We manually annotated the images in the *main* set to identify the useful ones. Since the idea is to find images useful for the emergency management task, the following criteria has been adopted:

- are considered useful the images which show damaged or blocked streets and roads, damage to buildings and population or aerial images;
- are excluded the images which focus on the rescue operations, minor damages, taken inside, focusing only on a small detail, with a resolution too low to be understandable or photos of other photos.

The image classification task has taken into consideration only the images and not the text or any metadata of the tweets.

[9]https://github.com/JohannesBuchner/imagehash

	Damage	*No Damage*	*Not Relevant*
Main set	**500 (63.86%)**	175 (22.35%)	108 (13.79%)

TABLE I
INFORMATIVENESS FOR THE *main set*

	Useful images	*Not useful images*
Main set	**14.18%**	85.82%
Damage set	**19.40%**	80.60%

TABLE II
USEFULNESS OF THE IMAGES IN THE MAIN SET AND IN THE SUBSET OF TWEETS WITH THE "DAMAGE" CLASS

At the end, 111 images were marked as useful, that is 14.18% of the *main* set.

To study the possible correlations between texts and images, we developed a text classifier similar to those in [1], [13]. We used linear Support Vector Machines (SVM) and a preprocessing phase to normalize and abstract the text both from a syntactic and a semantic point of view. To train it with a totally disjoint dataset but with the same domain and language, we chose the dataset presented in [3], in particular the historical earthquake of *L'Aquila* in 2009. Accordingly to this dataset, the considered classes are: "damage", i.e. tweets related to the disaster and carrying information about damages to the infrastructures or on the population, "no damage", i.e. tweets related to the disaster but not carrying relevant information for the assessment of damages, "not relevant", i.e. tweets not related to the disaster.

Testing the tweets in the *main* set, 500 tweets have been labeled as "damage" (they constitute the *damage* set). This is shown in table I.

Considering the *damage* set, it includes 97 images marked as useful, that is 19.40%. Therefore there is an increment in the useful images with respect to considering all the tweets in the *main* set. This is shown in table II.

These results highlight a correlation between a text-only feature (*damage*-relatedness) and an image-only feature (*usefulness*), with the first one obtained automatically and without any training data from the same dataset. This analysis is promising in showing that on Twitter the results for an image analysis task can be enhanced by a text-analysis task.

Comparing to the results obtained through the process and the tools illustrated in the paper, we have a precision of approximately 20% of the useful images with a damage-related text. This can be considered a good result as the images obtained automatically have also a location information, which is also derived from the text. Concerning the granularity of the locations, this is variable, since the text in some cases only allows localization at locality level rather than more accurate locations in terms of exact addresses. Further research work is still needed to provide ways of extracting geographical information also indirectly through tweet relationships as discussed for instance in [13]. In the project we are also going to investigate how further information can be gathered directly from twitterers on place or through crowdsourcing.

VII. CONCLUDING REMARKS

In this paper we presented a methodology and illustrated a prototype environment to support it. The methodology and the tool were tested in a case study to prove that Twitter can provide a significant number of geolocated images that can be useful for rapid mapping purposes. The development was done while studying the specific case study, but it can be easily generalized and improved to achieve better performance. Several aspects need further investigation. In particular, precision versus recall have to be analyzed in greater detail. In the present paper we chose to keep only images where a location was found, with the aim of increasing precision to provide only a relative small selection of images for further analysis. However, if a more extensive human-supported analysis is available, e.g., from crowdsourcing, this selection could favor recall instead, at the price of having to analyze a larger set of images. In the paper, tweets have been analyzed individually. In future work, we could leverage on clustering tweets, for instance merging tweets posted by the same author within a time interval as was done for user location identification in [6] or clustering tweets in a thematic way when they refer to the same location, or area, with the goal of increasing the precision of geolocation. Further analyses could involve a more sophisticated use of image recognition tools. While most open source/free image recognition tools are still in the initial phases, their fast development could provide further support in identifying locations and also associate further information starting from the contents of the tweets, thus making rapid mapping increasingly easier.

VIII. ACKNOWLEDGMENTS

This work has been partially funded by the European Commission H2020 project E^{2}mC "Evolution of Emergency Copernicus services". This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work.

REFERENCES

[1] M. Avvenuti, S. Cresci, M. N. La Polla, A. Marchetti, and M. Tesconi. Earthquake emergency management by social sensing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 587–592. IEEE, 2014.

[2] Copernicus. Copernicus, a European system for monitoring the earth. http://www.copernicus.eu/, 2016.

[3] S. Cresci, M. Tesconi, A. Cimino, and F. Dell'Orletta. A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1195–1200. ACM, 2015.

[4] E2mC Consortium. Critical review of crowdsourcing and social media use associated with copernicus ems service evolution challenges, e2mc deliverable d1.1, jan 2017.

[5] C. Francalanci and P. Giacomazzi. Torcia - a decision-support collaborative platform for emergency management. In *Date Conference*, 2015.

[6] R. Gonzalez, G. Figueroa, and Y. Chen. Tweolocator: a non-intrusive geographical locator system for twitter. In G. Ghinita, J. Neville, and S. D. Newsam, editors, *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN 2012, Redondo Beach, California, USA, November 6, 2012*, pages 24–31. ACM, 2012.

[7] A. L. Hughes and R. Shah. Designing an application for social media needs in emergency public information work. In *Proceedings of the 19th ACM International Conference on Supporting Group Work*, pages 399–408, 2017.
[8] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4):67:1–67:38, 2015.
[9] N. A. lateef Saeed, N. H. Zakaria, and M. N. Ahmad. The use of social media in knowledge integration for improving disaster emergency management task performance: Review of flood disasters. *Indian Journal of Science and Technology*, 9(34):1–12, 2016.
[10] B. Lindsay. Social media and disasters: Current uses, future options, and policy considerations. Congressional Research Service 7-5700, http: www.crs.gov, R41987, 2011.
[11] R. Power and J. Kibell. The social media intelligence analyst for emergency management. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 313–322, 2017.
[12] R. Rao, L. Plotnick, and R. Hiltz. Supporting the use of social media by emergency managers: Software tools to overcome information overload. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 304–312, 2017.
[13] T. Sakaki, M. Okazaki, and Y. Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Trans. Knowl. Data Eng.*, 25(4):919–931, 2013.
[14] W. Sherchan, S. Nepal, and C. Paris. A survey of trust in social networks. *ACM Comput. Surv.*, 45(4):47:1–47:33, 2013.
[15] R. Subba and T. Bui. Online convergence behavior, social media communications and crisis response: An empirical study of the 2015 nepal earthquake police twitter project. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, pages 284–293, 2017.
[16] Twitter. Open Source at Twitter. (English). https://engineering.twitter.com/opensource.

APPENDIX

A. Earthquake glossary

The earthquake glossary is multilingual. In the table the terms are shown in English and Italian, which are the prevalent languages in the case study. It contains a set of generic terms used to refer to earthquakes.

ENGLISH	**ITALIAN**
earthquake	terremoto, sisma
fault	faglia
quake	tremito, tremare
seismic	sismico
shake	scuotere
shock	scossa
tremor	tremore, scossa sismica

B. Places keywords

The following keywords have been used to recognize places and streets, to give a precise geolocation of tweets, beyond the locality.

“struttura” (“building”), “convento” (“convent”), “chiesa” (“church”), “stazione” (“station”), “piazza” (“square”), “via” (“street”), “zona” (“area”), “quartiere” (“district”), “monumento” (“monument”), “metro” (“subway”), “tram”, “ospedale” (“hospital”), “museo” (“museum”), “biblioteca” (“library”), “hotel”, “palazzo” (“palace”), “palestra” (“gym”), “campo” (“field”), “campus”, “castello” (“castle”), “torre” (“tower”), “campanile” (“bell tower”), “tribunale” (“court”), “municipio” (“town hall”), “corso” (“avenue”), “centrale” (“headquarter”), “tunnel”, “galleria” (“gallery”), “ponte” (“bridge”), “viadotto” (“overpass”), “universitá” (“university”), “fontana” (“fountain”), “teatro” (“theater”), “porto” (“harbor”), “strada” (“road”), “cinema”, “parco” (“park”), “duomo” (“cathedral”).

C. Rapid Mapping keywords

The keywords listed here are used to recognize tweets referring to damages.

“% croll%” (building/infrastructure collapse), “% maceri%” (debris), “% dann%” (damage), “% ferit%” (injured).