Geoinformatica (2022) 26:125–157
https://doi.org/10.1007/s10707-021-00446-x

CIME: Context-aware geolocation of emergency-related posts

Gabriele Scalia[1] · **Chiara Francalanci**[1] · **Barbara Pernici**[1]

Received: 11 March 2020 / Revised: 19 May 2021 / Accepted: 8 July 2021 /
Published online: 28 July 2021

Abstract
Information extracted from social media has proven to be very useful in the domain of emergency management. An important task in emergency management is rapid crisis mapping, which aims to produce timely and reliable maps of affected areas. During an emergency, the volume of emergency-related posts is typically large, but only a small fraction is relevant and help rapid mapping effectively. Furthermore, posts are not useful for mapping purposes unless they are correctly geolocated and, on average, less than 2% of posts are natively georeferenced. This paper presents an algorithm, called CIME, that aims to identify and geolocate emergency-related posts that are relevant for mapping purposes. While native geocoordinates are most often missing, many posts contain geographical references in their metadata, such as texts or links that can be used by CIME to filter and geolocate information. In addition, social media creates a social network and each post can be enhanced with indirect information from the post's network of relationships with other posts (for example, a retweet can be associated with other geographical references which are useful to geolocate the original tweet). To exploit all this information, CIME uses the concept of context, defined as the information characterizing a post both directly (the post's metadata) and indirectly (the post's network of relationships). The algorithm was evaluated on a recent major emergency event demonstrating better performance with respect to the state of the art in terms of total number of geolocated posts, geolocation accuracy and relevance for rapid mapping.

Keywords Geolocation · Information extraction · Graph analysis · Social media analysis · Emergency mapping

✉ Gabriele Scalia
gabriele.scalia@polimi.it

Chiara Francalanci
chiara.francalanci@polimi.it

Barbara Pernici
barbara.pernici@polimi.it

[1] Politecnico di Milano - DEIB, piazza Leonardo da Vinci, 32, Milano, Italy

1 Introduction

Social media play an important role in emergency management. Indeed, information extracted from social media has proven to be very informative in crisis situations [1, 2]. An important activity in this context is *rapid crisis mapping*. Rapid mapping consists of producing timely and reliable maps of the areas affected by the emergency situation — typically a natural disaster — with the goal to support rescue teams and operators. For instance, the Copernicus Emergency Rapid Mapping Service[1] is a rapid crisis mapping service based on Copernicus satellite data.

The production of maps needs to be rapid, that is, fast enough to support emergency management as soon as the emergency begins, during the so-called *response phase*. In this phase, issues such as the delineation of the perimeter of the emergency or the severity of the event are explored in a rush of turbulent aid-organization processes. Rapidly available maps represent an invaluable input for these life-saving activities.

Rapid mapping has very specific and challenging requirements, including:

- Multimedia content, especially pictures, is far more useful than text.
- Content is not useful unless geolocated. A map is a representation of selected information of a geographical area. To be represented on a map, information needs to be (automatically or manually) geolocated.
- The geolocation precision needs to be high enough to support the tasks at hand. For rapid crisis mapping, geolocation information needs to be associated with the extension of the target object: a bridge, a house, a road, a flooded area, etc.
- Geolocated information needs to be redundant to reinforce mapping decisions. Redundancy increases trust in information. For example, if all pictures of adjacent roads show a flooded situation, an entire zone can be mapped as flooded with high confidence.

During an emergency, the volume of emergency-related posts is large, but only a small fraction of posts are relevant and provide objective information and media that can effectively help rapid mapping activities [1]. Furthermore, posts are not useful for mapping purposes unless they are correctly geolocated and, on average, less than 2% of posts are natively georeferenced. In theory, rapid mapping professionals can find it beneficial to be supported by on-site visual information created and shared by social media users. However, in practice, without adequate tools to filter and automatically extract information, leveraging social media posts during emergencies is unsustainable. This creates a demand for integrated platforms to extract information from social media in a timely fashion [1], which need to be supported by algorithms and methods tailored to the needs and challenges posed by this domain.

This research was conducted as part of the E2mC European Project [3][2] with the aim of designing and testing a platform called *Witness* to support rapid mapping professionals with information extracted from social media. In this context, we designed a geolocation algorithm, called CIME, to identify and geolocate emergency-related posts which can be relevant for mapping purposes.

While native geocoordinates are most often missing, many posts contain geographical references in their metadata, such as texts or links, that can be used by CIME to filter and geolocate information. In addition, social media posts are not isolated entities, but

[1] http://emergency.copernicus.eu/mapping
[2] https://cordis.europa.eu/project/id/730082

exist within a social network. CIME exploits this phenomenon to improve the geolocation process: each post is enhanced with indirect information from the post's network of relationships with other posts (for example, a retweet can have geographical references associated with it, which are useful to geolocate the original tweet). To exploit all this information, CIME is designed to build a *context* for each candidate location extracted from the post. The context is defined as the information characterizing a post both directly (the post's contents and metadata) and indirectly (the post's network of relationships).

As discussed in the following section, other network-based geolocation techniques have been proposed previously. These aim to geolocate the authors of the posts, based on the analysis of static social relationships, such as friendship. In contrast, CIME aims to geolocate content (individual posts) on the basis of *behavioral* and *dynamic* social networks that are created by communication patterns among authors, who may or may not have a static social relationship. The rationale for this approach is that crisis situations create new social relationships that did not exist before the crisis, but are needed to manage the crisis [1]. Geolocating posts, rather than authors, can prove to be particularly useful for rapid mapping activities.

CIME exploits the behavioral social networks naturally arising during emergencies [1]. In these situations, social media posts are not isolated entities, but often belong to crisis-related social contexts. For example, a post can belong to a conversation or a crisis-related trending topic (hashtag). It can be observed how emergencies create multiple behavioral social networks, usually including a high number of messages about the event posted in a relatively short timeframe. CIME is capable of identifying and exploiting these behavioural networks to enrich the context which supports the geolocation task.

In principle, the proposed methodology is independent with respect to the social media at hand. In the E2mC project, *Twitter*, *Flickr*, *Facebook/Instagram* and *YouTube* have been identified as the most important sources of information for rapid mapping purposes. The motivations and the concepts behind these aspects have been explored in [3–7]. In particular, Twitter offers data access publicly and represents the most widely studied source of information for emergency management [1]. This paper presents CIME as a general algorithm that can be used on any social network. However, the examples and the experimental evaluation will refer, in particular, to Twitter, to facilitate the data collection/analysis and the comparison with previous approaches mentioned in the literature.

The main contributions of this work are as follows:

- We face the problem of geolocating social media posts, focusing on the requirements and challenges arising from supporting rapid mapping activities. These include: (1) the focus on posts, rather than users, (2) not relying on prior knowledge about the event or the area, since emergency events are, in general, unpredictable, (3) the focus on media contents associated with posts (in particular, images).
- To this end, we develop a context-based geolocation algorithm, named CIME, to disambiguate locations mentioned in posts. The context is both "local" (that is, built based on elements from the same post), and "global" (that is, built based on elements from other connected posts in a dynamic behavioral social network).
- We develop a complete pipeline to execute CIME. CIME is tested on two case studies.
- We evaluate CIME from the point of view of supporting the needs arising from leveraging social media analysis for rapid mapping activities. In particular, we focus on images associated with posts geolocated through CIME, and we compare their volume, precision and relevance to native (GPS-based) georeferences provided by Twitter for a selected subset of posts.

The structure of the paper is as follows: Section 2 discusses the state of the art. Section 3 discusses different types of interpretations that can be associated with locations associated with social media objects, to better define the scope of the proposed method and the experimental evaluation. Section 4 illustrates the proposed geolocation algorithm and Section 5 analyzes the outcomes of the proposed methodology with reference to a rapid mapping case study. Section 6 discusses the implementation details.

Finally, Section 7 discusses open issues and future work.

2 State of the art

2.1 Geolocation in social media

Geolocation means assigning a location to an entity, which, in the context of social media, is usually a user, a post or another content associated with users/posts. On social media, users can have a profile that statically associates a location to them. The location can be precise (a home address) or broad (a city or a region). Posts can be dynamically linked to a location indicating either the position of the user when the post was shared or the location of the post itself (for example, the location of a picture that is part of the post's content).

Therefore, different types of locations can be extracted from social media: locations related to users, including the home residence [8–11], locations related to the posting position [12], as well as locations related to the posts themselves [13–16]. Recently, [17] surveyed the task of location prediction on Twitter. In this work, the authors distinguished three types of target locations: *home location*, *tweet location* and *mentioned location*, discussing different methods and evaluation metrics for each one. With respect to these categories, the present work focuses on mentioned locations. See Section 3 for an in-depth analysis of location types.

In some cases, a single piece of geolocation information is available and it can be difficult to associate it with a posting location, a location associated with the post topic or with the post multimedia content unequivocally. Often, a combination of different types of locations have been used in practical applications, as in [18]. In general, the different location types are independent of each other and they are all independent with respect to the features used to infer them. For example, a location mentioned in a post does not necessarily imply that the user or the message content are linked to that geolocation. Different features have been used to geolocate tweets [17, 19]: location mentions in the text of the tweet, friend's network, metadata of the tweet, website IP addresses, geotags, URL links and time zones.

An important difference should be made between *location disambiguation* and *location inference*. While the former, also called "toponym resolution", aims to geocode an expression — that is to find the location in terms of its coordinates — given a more or less qualified name, the latter aims to discover the location based on features such as the content, the metadata, the relationships with other documents, etc., even if no location is explicitly mentioned.

The proposed algorithm targets location disambiguation, surveyed in Section 2.2. Recent works about location inference with a similar goal — post-level geolocation — are mentioned in Section 2.3.

2.2 Location disambiguation

Location disambiguation is typical of text-based methods. Indeed, toponyms mentioned in unstructured text are, in general, ambiguous and unreliable. First of all, it is necessary to extract them from text, an activity called *named entity recognition* (NER) in general, when focused on names of locations, people, companies, etc., or also "geoparsing" when focused exclusively on names of locations [16]. The NER, which is typically performed through supervised techniques such as conditional random field (CRF), allows the extraction of a set of *surface forms* from the text. Each surface form is an *n*-gram potentially referring to a named entity [20] — that is, to a location when only location names are considered. Then, each surface form must be disambiguated, i.e., linked to a precise named entity in a knowledge base[3]. Generally speaking, this activity is called *named entity linking* (NEL), also named "geocoding" or "location disambiguation" when focused on location names. This phase is challenging because the relationship between surface forms and entities is, in general, many-to-many: on the one hand, the same entity can have many names and variations, on the other hand, multiple entities can share the same name or part of it. Moreover, being downstream, the performance of the NEL phase is necessarily affected by the NER phase: wrongly recognized surface forms negatively affect the linking/disambiguation phase. Disambiguation of location names is made especially difficult by the existence of *geo/non-geo* ambiguities, such as the Italian city called None which coincides with a very common word in the English language, and *geo/geo* ambiguities such as in London, UK; London, ON, Canada; London, OH, USA; London, TX, USA; London, CA, USA [13].

These tasks have been mainly addressed through supervised learning techniques, as in [15, 16, 22–24]. Supervised models are used not only for the NER task, but also for the disambiguation itself. The main limitations of this kind of approach include the need for labeled data and the limited generalization ability (in terms of language, area and event). In some cases, supervised models are supported by heuristic disambiguation rules, as in [13]. In other cases, supervised techniques are avoided and only gazetteer insights are exploited, as in [25]. However, this requires knowledge of the target area in advance and the pre-loading of the related location names in the memory. In [26], the work presented in [25] is extended, demonstrating that a gazetteer-based approach outperforms other methods, including third-party services such as Geo-Names and Google Geocoder API. In some cases, disambiguation could target location names mentioned in an informal way (abbreviations and out-of-vocabulary words), which introduces a further degree of ambiguity. This is, for example, the goal of the two-stage supervised system presented in [24].

CIME uses pre-trained models only for the NER task. One of the goals of the proposed algorithm is to not rely on trained components for the disambiguation task, thus allowing no prior knowledge of the target area and avoiding any pre-loading, indexing or training phases.

2.3 Location inference

Supervised techniques have also been recently proposed for fine-grained geolocation inference at the venue level [27, 28]. In this case, rather than recognizing and disambiguating locations mentioned in tweets, the goal is to infer the tweet's posting location in terms of

[3]Examples of knowledge bases include Wikipedia, DBpedia, Freebase etc. Focusing on location names, knowledge bases correspond to gazetteers such as GeoNames (http://www.geonames.org) or OpenStreetMap [21]

specific venue starting from the tweet's content. Such approaches allow geolocating fine-grained POIs, but require labeled data at the venue level, which can be difficult to acquire in general. For example, [27] proposes using data crawled from the location app Foursquare, where users associate their posts with specific venues. The paper [19] presents an overview of location inference techniques on Twitter, including those requiring prior knowledge of the target location or having an updated model of the target area. This kind of location inference technique is not appropriate, in general, for emergency events, when the target area is not known in advance and previous posts about it could be scarce. In contrast to this kind of technique, the methodology proposed in this work focuses on *disambiguating* location names in posts, rather than associating locations based on the post's content. Moreover, the proposed methodology does not require any prior knowledge or model about the target event or area.

2.4 Network-based geolocation

Network-based geolocation methods exploit the social networks characterizing social media to *propagate* information. Network-based geolocation is mainly based on the property according to which users on social media tend to interact primarily with the same people with whom they interact in their everyday life. Therefore, this approach has been mainly used to identify users' location on the basis of their friends' locations. It has been noted how, "in many cases, a person's social network is sufficient to reveal their location" [11]. For example, [8] uses an extension of label propagation through the social network to assign a location to nearly all users, starting from few labeled users, [29] proposes an algorithm that allows the selection of the right interpretation of the self-reported toponym of users starting from the (possibly ambiguous) self-reported toponyms of their friends. The geolocation performance depends on the users' number of connections: users with too few or too many friends cannot be geolocated easily [30]. Supervised techniques are also employed: in [31], a decision tree is trained to show that some features of relationships (for example users who mention each other) are correlated with physical proximity, [11] formulates geolocation as a classification task where the class is the inferred location and the features include the locations of users' friends.

Network-based and text-based methods are not mutually exclusive: [10] shows that a hybrid method (both text and network based) outperforms two baselines that consider only the text or the network. A special case of social network is the one containing all the posts from the same user, i.e., the list of posts from the same person in a given timeframe. Even if the size of this network is drastically smaller than a network considering multiple users, posts from the same user can have a higher probability of being (spatially) related. This concept is exploited in [28], where tweets are linked to fine-grained locations through a supervised approach that also leverages same-user posts.

All these network-based techniques target user home locations [8, 10, 11, 29, 31]. This is due to the fact that they focus on *explicit (articulated)* networks, such as friendship, as opposed to *implicit (behavioral)* networks, which are inferred from dynamic communication patterns. Explicit networks model "static" relationships among *users*, not capturing the relationships among actual *posts*. Whereas "implicit networks are particularly interesting in the crisis scenario because many exchanges happen among people who were not connected before the crisis" [1]. In [32], the authors discuss how different information diffusion patterns appear in Twitter in different contexts, and how the patterns typical of natural emergency situations are different, for instance, from the ones of videogame users.

Similarities among individual posts have also been exploited for post geolocation, for example in [33]. In this work, authors estimated the location of a post by leveraging similarities in the content of the target and a set of other geotagged posts, as well as their time-evolution features, using a supervised model. However, there is no real concept of behavioral social network (in terms of graph) in their work. Their approach is particularly effective for the geolocation of tweets related to time-focused events, especially those with a relatively short time-span (e.g., concerts). However, it has the same limitations of other supervised techniques in terms of required training and limited domain adaptation.

Most of the network-based techniques described in the literature focus on the location inference task, targeting the location disambiguation task more rarely [29]. Behavioral social networks have been successfully exploited for other tasks, such as topic identification [34], however, to the best of our knowledge, they have never been used for post/media geolocation on Twitter.

Different to most of the disambiguation algorithms previously mentioned in Section 2.2, CIME does not only rely on textual features, but also on contextual features, and in particular on posts' *social networks*. An initial description of this approach was introduced in [35] and is presented in detail in this paper. Different with respect to the mentioned network-based methods, the proposed algorithm i) relies on social networks for post-level geolocation instead of user-level geolocation, thus focusing on behavioral social networks instead of static networks and ii) leverages social networks for location disambiguation instead of inference.

2.5 Context-aware social media analysis

The role of context in the proposed algorithm is central. In CIME, the context is built based on post metadata and social network features. Recently, different kinds of contextual information have been exploited for social media analysis. The concept of geographical context has been investigated and analyzed in [36], in particular with respect to hierarchical spatial relationships characterizing the user's context.

It has been noted that the combination of social media and contextual authoritative sources of data can be helpful to identify useful information in disaster management [37, 38]. The joint analysis of GPS data, authoritative data and news report data (but not social media sources) has been recently described in [39], where a supervised approach for modeling human behavior following natural disasters has been proposed, demonstrating the advantages of integrating heterogeneous sources. Another way to exploit context is by considering multiple social media providing complementary information [40, 41].

2.6 Gazetteers

Different gazetteers have been employed for the geolocation task on social media [17, 19]. In a previous work [4], the authors used GeoNames. This initial case study helped us to define the requirements for the work presented in this paper, which instead adopted OpenStreetMap as gazetteer, as it provides a richer and larger knowledge base but also introduces new challenges.

Regarding gazetteer misses, they can be tackled only partially by geolocation algorithms.

Gazetteer misses can be considered unavoidable since "the output of any geocoding algorithm is only as exact as the knowledge base that underlies it" [16].

3 Understanding the meaning of Tweet geolocation

As mentioned in Section 2, different types of locations can be associated with a social media post.

In [17], a distinction was made among different types of predictions: author's home geolocation, tweet geolocation and mentioned geolocation. For this work we focus on geolocating locations mentioned in a post, with the assumption that media (in particular, images) associated with a post show the mentioned locations. This focus is due to the goal of our work, which is aimed at helping rapid mapping activities, providing evidence through images with an associated location information. As noted previously, visual content is of primary importance in these activities, as it can provide tangible evidence of the conditions of an area.

This section clarifies the approach adopted in this paper: social media posts were analyzed to extract locations on visual information associated to the posts. In the following, we illustrate which are the available elements in the post data and metadata, and which are the geolocations that can be inferred from their analysis. Figure 1 shows a model for locations in Twitter. The model can be adapted to other social media, taking into consideration their specific meta-data.

The figure shows the main entities that can be identified in a tweet: the content of the post itself (*post*), the user posting the tweet (*user*), and the images associated to a tweet (*post image*). For each of these entities, the available data and metadata are shown in the figure as attributes. In addition, in the following of the paper we will also consider the links that can be found in a post, either to posts in the same social network, i.e. through *reply*, *quote*, and *retweet* links, or as links to posts in other social networks, indicated as *post link*.

With regards to tweets, there are several possible sources of geographical information. The *meta-data* of the post and of the user provide the following information:

- `location`: The location defined by users for their account profile. This is free text inserted by users. In addition, the users in their profile can set a *geo_enabled* flag that enables the possibility of geotagging tweets based on their current position when posting (see below).
- `coordinates`: Tweets can be associated with an automatically derived geographical location, usually extracted from the GPS-enabled device from which the tweet is sent. This type of geographical location is referred to as *georeference* (or sometimes *geotag*), indicated as *georef* in the figure. Georefs are reportedly unusual in tweets, ranging from 0.3% [4] to 3%, and only in rare cases reaching 11% [37]. Georefs are longitude-latitude pairs.
- `place`: Recently, Twitter has added the possibility for users to specify a *place*, which is displayed at the bottom of the post when specified, next to the *posting time*. In general, places are given at city or country level and are not used to specify a precise address.

Notice that all these fields are nullable. The `location` can be unparsable and may not be a real location, and both `location` and `place` do not necessarily have an actual relationship with the users' message content or actual position. Considering the global tweets randomly sampled, [18] reports that 41% of users agreed to share their location at least once, 35% of users have specified a location, only 2.5% of the tweets come with a non-null `place` field (which is at city-level in 89% of the cases) and only 2% of the tweets are geotagged with precise location coordinates.

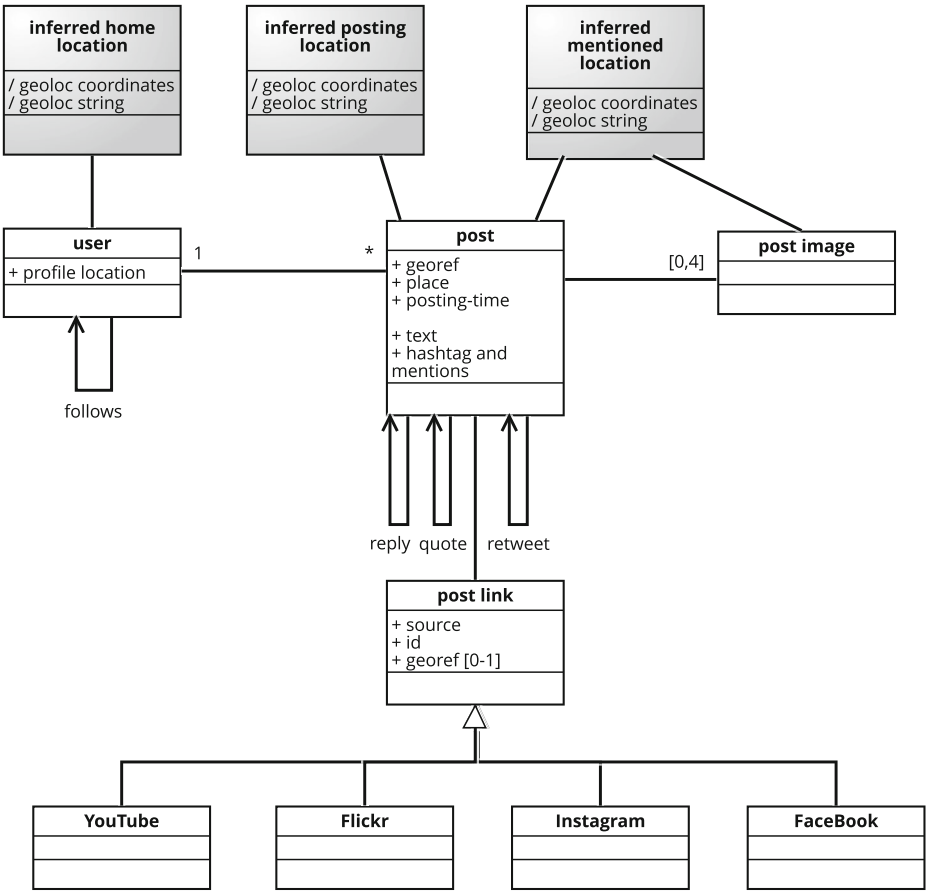

Fig. 1 Modeling locations associated with Tweets

Posts may be associated with media contents. In particular, *images* can be attached to tweets. In Twitter, images are stripped of any metadata when uploaded, so no geolocation information can be found in their metadata. As mentioned above, posts can also include links to other social media posts. A linked post may or may not be associated with a georeference.

In general, all these geographical locations (related to the user, the posting location, the content, the media, etc.) may or may not be consistent. For instance, users might be visiting a location far from their home location, and might post their tweets at a different time and in a different place from where the images or linked media were taken. These locations can also be given or derived at different levels of granularity.

Since georeferences are usually very scarce, research addressed the problem of automatically geolocating tweets (see Section 2) as a way to increase the number of posts linked to a location. Recent work has exploited tweets' locations to analyze events. However, as discussed in Section 2, the concept of *location* needs to be clearly defined. In fact, inferred locations can be associated with different objects: i) inferred home locations; ii)

inferred posting locations; iii) inferred mentioned locations. Figure 1 shows possible different inferred locations as gray classes. The three inferred locations can be different, i.e., the user can have a home location which is different from the posting location, which in turn can be either the same as the one of the image (for instance, when a tweet is posting a just captured image) or different, in case images are posted later of derived from other sources.

Inferred locations are associated with two possible representations: *geolocation coordinates* ("geoloc coordinates") - that is a point or an area - when a geographical representation is given in terms of coordinates; *geolocation string* ("geoloc string"), when a textual representation is provided. The location precision depends on the gazetteer being used.

As this paper focuses on posts related to emergencies, and on the need of extracting first hand information about a disaster, the proposed methodology focuses on locations mentioned in posts (inferred mentioned locations). In the following section, we illustrate our approach to disambiguate mentioned locations from the information available in posts (including text and metadata) and other related posts in the social network. Inferred locations are associated also to images contained in the post.

4 CIME geolocation algorithm

In this section we describe CIME, a novel algorithm for social-media content geolocation. With respect to the distinction discussed in Chapter 2, CIME can be considered a text-based geolocation algorithm, since it aims to disambiguate location references contained in posts. However, it can also be considered a *network-based* geolocation algorithm, since it exploits the social network to build a context exploited for geolocation. The main structure of the algorithm is social-media independent, even though the implementation and the evaluation of this work are specific to *Twitter*.

The algorithm is based on the idea of building a *context* for each ambiguous surface form extracted in a preceding NER step and using it as *reinforcement* to infer the most likely candidate location as the one with the highest "reinforcement level", provided that the reinforcement level is higher than a *confidence threshold*. The context is built in two steps: firstly, using *local* features (that is, features associated with the individual post), and secondly, it is then extended based on *global* features (that is, features associated with other posts that are connected to the target post in the social network).

This approach aims to handle the following situations:

- If a surface form is a false positive outcome of the NER phase,[4] it should not find a reinforcement in the context (local or global), and therefore the surface form should not be disambiguated at all.[5] In this respect, the algorithm acts as a *filter* for the NER, addressing potential NER errors.
- If a surface form is contained in a post with a strong enough local context, the algorithm should be able to disambiguate it "locally," without analyzing the global context of the post.
- If a surface form can not be disambiguated through the local context, it can be either a NER false positive or a disambiguation false negative. NER false positives are mainly due to the noisy and unstructured nature of social media texts, while disambiguation

[4] A NER false positive corresponds to the situation where the NER has recognized an *n*-gram as location while actually it is not a location, for example for a geo/non-geo ambiguity.

[5] Or, equivalently, it should be disambiguated to the *null* location.

false negatives are mainly due to the short and de-contextualized nature of social media posts. The global context can be used as support information to differentiate between NER false positives and disambiguation false negatives. Indeed, a NER false positive should not find a reinforcement in the global context since it is not a real location. Otherwise, the global context should allow the lacking local context to be overcome and find a reinforcement to disambiguate the surface form to a location with enough confidence.

The algorithm is described starting with the *local* phase in Section 4.1 and then with the *global* phase in Section 4.2. For the sake of readability, the main sets and functions defined in the following sections are summarized in Table 1.

4.1 Local geolocation based on a single-tweet analysis

The local phase of the algorithm is based on the analysis of a single post, in particular its text and metadata. The main steps are shown in Fig. 2. The input is a post p_i. The surface forms, the `place` metadata and the hashtags are extracted from p_i. Then, the candidate locations are retrieved from the gazetteer for each of these elements and a *local context* is built for each surface form. The local context-based ranking step allows the candidate locations for each surface form to be ranked given their local context. Finally, there are two possible outputs. If the confidence threshold has been reached for some candidate locations, the candidate location(s) with the highest score are selected as post location(s). Otherwise, the post is not geolocated, the surface forms are considered ambiguous and their candidate locations are kept for the subsequent global geolocation phase. We detail each of these steps in the following paragraphs.

Table 1 Notation used in the description of the geolocation algorithm

Notation	Description	Notes
P	Set of posts	Input
SF	Set of surface forms	
LOC	Set of locations	Locations in the gazetteer
$sf : P \rightarrow \mathcal{P}(SF)$	Surface forms of each post	NER
$C : SF \rightarrow \mathcal{P}(LOC)$	Candidate locations of each surface form	Gazetteer search
$\mathcal{C} : P \rightarrow \mathcal{P}(LOC)$	Candidate locations of each surface form of each post	NER + gazetteer search
$hashtag(p)$	Hashtags associated to each post	
$place(p)$	Place associated with each post	Post metadata
$lcxt(sf, p)$	Local context of the surface form sf in the post p	$\subseteq LOC$
$score(c, l)$	Score of a candidate location c by the context element l	$c, l \in LOC$
$rk : C \rightarrow \mathbb{R}^+$	Rank of each candidate location of a surface form	
$rk^* : \mathcal{C} \rightarrow \mathbb{R}^+$	Rank of each candidate location of a post	
$loc : P \rightarrow \mathcal{P}(LOC)$	Location(s) associated with a post p	Post geolocation
$G = (P, E)$	Social network of posts	
$globalcontext(p)$	*Global context* seen by the post p	
$gcxt(sf, p)$	Global context of the surface form sf in the post p	$\subseteq LOC$

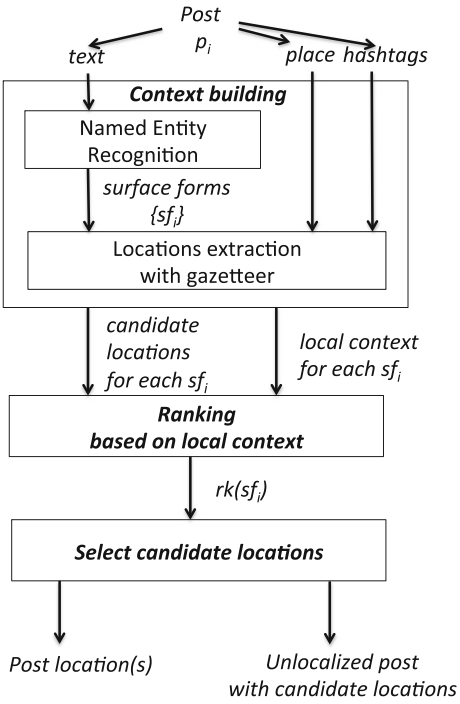

Fig. 2 Local CIME algorithm

4.1.1 Candidate locations and local context building

For each post p, the text is analyzed to extract a set of surface forms. Each surface form is an n-gram potentially referring to a location [20]. This is formalized by the function $sf : P \rightarrow \mathcal{P}(SF)$, which links to each post $p \in P$ a set of surface forms: $sf(p) = \{sf_1, sf_2, \ldots, sf_N\}$, with $sf_1, sf_2, \ldots, sf_N \in SF$. This function corresponds to a NER. In the current implementation, the multilingual Stanford CoreNLP library [42, 43] has been used, however, the methodology is independent of the chosen NER library.

In general, since the focus of the algorithm is on the *disambiguation* of the surface forms, including the recognition of false positives,[6] it can be convenient to employ a NER algorithm which privileges recall instead of precision. Indeed, if a surface form is a false positive, CIME should not be able to link it to anything, thus avoiding false positive locations, while a false negative in the NER phase cannot be fixed by CIME in the disambiguation phase and irremediably leads to a missed location. In this respect, CIME acts as a *filter* for false positives from the NER phase. Moreover, as discussed in the introduction, the management emergency domain often requires the recall to be maximized for operative reasons. To account for this, the CoreNLP NER library has been used, with a traditional model and with a caseless model which allowed handling "all lowercase, all uppercase, or badly and

[6]The disambiguation phase should link a surface form to the "*null*" entity if the referred entity, for any reason, does not exists in the knowledge base (gazetteer), therefore detecting the false positive.

inconsistently capitalized texts,"[7] keeping the union of the results. Moreover, the results of the dependency parsing analysis were used to increase the coverage, adding as surface forms those tokens that have a strong relationship with the detected surface forms (e.g., a *compound* relationship[8]). Finally, to further increase the NER recall, the surface forms were enhanced with a rule-based regex NER to include typical location names of streets, roads, rivers and so on[9]. This is based on the *TokensRegexNERAnnotator*[10] provided in the CoreNLP library.

The surface forms extracted from a tweet are, in general, *ambiguous*, due to the many-to-many relationships existing between toponyms and locations, geo/geo and geo/non-geo ambiguities (see Section 2). Therefore, it is necessary to retrieve all the *candidate locations* for each surface form and *disambiguate* the correct one. We used OpenStreetMap [21] as gazetteer to retrieve all possible candidate locations for each surface form. OpenStreetMap is accessed through the Nominatim[11] search engine in a local instance. Nominatim links to each location entity, among many other tags, a set of *names* referring to the location (including alternative names), a *hierarchical system* which allows the definition of relationships among locations and the *polygon(s)* corresponding to the area of the locations.[12] Even if a location is typically a polygon, a *center* is also defined, simplifying the representation as a pair of coordinates.

Through the gazetteer, each surface form sf is assigned to a set (0 or more) of candidate locations fully or partially matching sf. This is formalized by the function $C : SF \rightarrow \mathcal{P}(LOC)$ which links to each surface form $sf \in SF$ a set of locations: $C(sf) = \{c_1, c_2, \ldots, c_M\}$ with $c_1, c_2, \ldots, c_M \in LOC$. In turn, it is possible to associate *all* candidate locations related to *all* surface forms of a post through the function $\mathcal{C} : P \rightarrow \mathcal{P}(LOC)$, considering the union of the individual sets $\mathcal{C}(p) = C(sf_1) \cup C(sf_2) \cup \ldots \cup C(sf_N)$.

The local step of the proposed methodology aims to disambiguate the surface forms extracted from the post's text using other information available in the post itself as context. This is referred to as the *local context*. In particular, the information exploited to build a surface form's local context is:

- Other surface forms, through the associated candidate locations. This reflects the fact that the locations mentioned in a post have a high likelihood of being somewhat related to each other. For example, a user can mention a street and the related city or two streets belonging to the same city.
- Contextual information in the post. These are social-media dependent; focusing on Twitter, the contextual information that can be exploited are:
  - The `hashtags` in the tweet. Hashtags are used to assign a topic to a post. Topics can correspond to localities, especially when the post includes content related to a (damaged) area.

[7] https://stanfordnlp.github.io/CoreNLP/caseless.html

[8] http://universaldependencies.org/u/dep/

[9] The regex NER step detects capitalized words followed by a word in pre-defined set such as: "street", "road", "river", etc.

[10] https://stanfordnlp.github.io/CoreNLP/regexner.html

[11] http://nominatim.openstreetmap.org/

[12] Different formats are provided; GeoJSON was chosen for its interoperability.

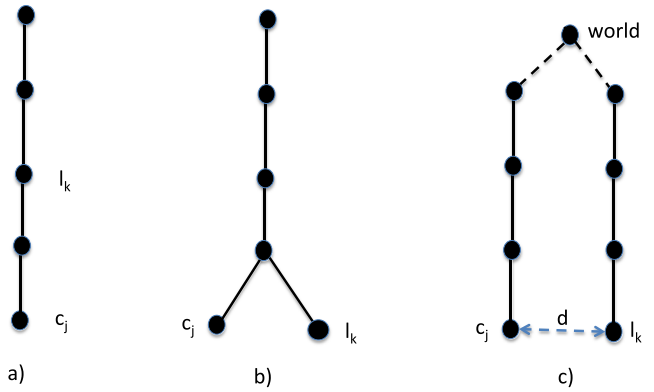

Fig. 3 Relationships between locations

– The `place` field of the post, which, "when present, indicates that the tweet is associated with (but not necessarily originates from) a place".[13] The `place` field often corresponds to a coarse-grained location, but it can help to contextualize more fine-grained locations mentioned in the text of the tweet or enhance their confidence.

Therefore, the local context for the surface form $sf_i \in sf(p)$, defined as $lcxt(sf_i, p)$, is given by:

$$lcxt(sf_i, p) := \left(\bigcup_{sf_j \in sf(p), j \neq i} C(sf_j) \right) \cup \left(\bigcup_{h \in hashtags(p)} C(h) \right) \cup C(place(p))$$

4.1.2 Local context-based ranking

The candidate locations of each surface form get a *score* for each element in the context, thus defining a *ranking function* $rk(sf) : C(sf) \rightarrow \mathbb{R}^+$ that associates a positive number to each candidate location. In particular, the ranking is given by the sum of the individual scores:

$$\forall c_j \in C(sf_i),\ rk(c_j) = \sum_{l_k \in lcxt(sf_i, p)} score(c_j, l_k)$$

In this formula, $score(c_j, l_k)$ — the score given to the candidate location c_j by the contextual element l_k — is computed as follows (see Fig. 3):

(a) Based on the *hierarchical relationship*[14] between c_j and l_k. In particular, a score is given if c_j is part of l_k. For example, if c_j is a street belonging to the town l_k. A fixed score value is assigned to c_j, since it is more fine-grained (Fig. 3a).

[13] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

[14] Information on hierarchical relationships among locations are extracted from the OpenStretMap gazetteer.

(b) Based on the *common hierarchy* between c_j and l_k. For example, if both c_j and l_k are streets which exist in the same town, a fixed score value is assigned to both c_j and l_k (Fig. 3b).

(c) Based on the distance between c_j and l_k, considering their centers, if lower than a fixed threshold (20 *km* in the current implementation). c_j and l_k can be any type of location. A fixed score is assigned if the *distance* is 0 meters and decreases as the distance increases up to the threshold (Fig. 3c).
Specifically, $score = \left(1 - \frac{distance}{distance\,threshold}\right) \cdot max\_score$

These scores allow the algorithm to privilege, respectively, the situations in which a post describes a location as a specific part of another location (e.g., a street in a city), more locations in the same area (e.g., streets in the same city or cities in the same region) and places close to each other. In the current implementation, these fixed values were set at 3, 2 and 1, respectively.

4.1.3 Select post location

The ranking can be extended to all the candidate locations of a post p through the function $rk^* : \mathcal{C}(p) \rightarrow \mathbb{R}^+$. rk^* is the global ranking for a post, based on all the individual rankings $rk_i(sf_i)$ for all the surface forms in the post. The global ranking allows the identification of *the most representative locations*[15] of the post, if any, defined as $loc(p)$.

In particular, the location(s) with the highest global ranking are selected if $rk^*(p)$ is greater than a threshold. Otherwise, no location is selected (that is, the post's location remains ambiguous). The *threshold* is a confidence or reinforcement level that the candidate location(s) need to satisfy to be confidently disambiguated. This is summarized in the following equation:

$$loc(p) = \begin{cases} \emptyset & \nexists c \in \mathcal{C}(p) \mid rk^*(c) > threshold \\ \arg\max(rk^*) & \text{otherwise} \end{cases} \quad (1)$$

4.2 Global geolocation based on diffusion network

As previously discussed, in the CIME local geolocation step, a confidence threshold is essential to prevent false positives.

Social media posts are inherently short[16], and they often need to be contextualized in terms of event, time and/or conversation flow to be meaningful.

Therefore, in some cases, even though all the surface forms have been correctly extracted from the text, it may still be impossible to disambiguate them with enough confidence because the local context is missing or not strong enough (see first case in Eq. 1).

These cases are tackled by CIME's *global geolocation step*.

As discussed in Section 2, each post belongs to a *social network*[17]. CIME relies on *behavioral* (or *implicit*) social networks rather than the articulated (or explicit) social networks. While the latter are based on codified and static relationships among users, such as

[15] Note that, in general, more than one location can be assigned to the same post. Indeed, the same post can have more than one topic location (see Fig. 1).

[16] E.g., standard Twitter posts, which used to be limited to 140 characters, are currently limited to 280 characters.

[17] In this context, a social network indicates a network built over one or more static or dynamic social interactions.

"friendship," behavioral social networks are *dynamically inferred* from communication patterns and connect posts to other posts related to the same event, time or conversation, even if their authors have no explicit connections, which is a common case in emergency situations [1].

The proposed algorithm incrementally builds a behavioral social network based on the interactions among posts. This social network is encoded as a graph, called *global graph*. The global graph is then used to provide an additional context to posts that could not be geolocated in the local phase.

Moreover, the algorithm supports information propagation among posts recursively, in order to exploit already disambiguated posts in the global disambiguation step of other posts. Each of these steps is discussed in detail in the following paragraphs.

4.2.1 Behavioral social network

The behavioral social network is modeled through a graph $G = (P, E)$, called *global graph*. The nodes represent the posts, while the edges represent social interactions among posts. The graph is built incrementally as new posts are crawled. Because a behavioral social network is meant to capture transient relationships among posts, a *sliding window* is used to remove posts once they exceed a maximum timeframe (72 hours in the current implementation).

Posts are connected based on the following criteria:

- When they belong to the same *conversation*, for example through *replies* or *quotes*.
- When their authors interact with each other within a given time-frame (24 hours in the current implementation[18]). For example, when a user *mentions* or *retweets* another user, a temporary connection is created and the posts of the two users in that time-frame are connected.
- When they share the same *image* (or very similar images). The assumption is that posts sharing the same image have a high probability of sharing the same topic. Perceptual hashing[19] is used to compare images robustly.
- When the same *hashtag* is used in a given time-frame (four hours in the current implementation). A hashtag is a user-defined topic for a post. Therefore, posts using the same hashtags in the same time-frame are likely to be related to the same topic/event.

4.2.2 Global context building and global context-based ranking

The global graph allows each post p to "see" a global context $globalcontext(p)$ built based on the union of the neighbors' locations (that is, the other posts connected) in the graph which were previously disambiguated:

$$globalcontext(p) = \bigcup_{p_j | (p,p_j) \in E \wedge loc(p_j) \neq \emptyset} loc(p_j)$$

[18]This time-frame has been established based on preliminary analysis, varying the window in the 8-72 hours range. We have observed that results are robust (i.e., they change only marginally) with respect to the choice of the time-frame. Primarily, a fixed time-frame allows the resulting social graph to be limited, reducing the average number of neighbors for each node (keeping, in particular, the most relevant ones) and therefore improving the performance of the global disambiguation step. This also allows the reduction of the computational and memory requirements.

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

The *cardinality*, that is the cumulative number of times each location appears in any of the disambiguated neighbors, is stored for each location in $globalcontext(p)$:

$$\forall l_k \in globalcontext(p),\ cardinality(l_k, p) = \left|\left\{p_j \mid (p, p_j) \in E \wedge l_k \in loc(p_j)\right\}\right|$$

The global context is combined with the local context of each ambiguous surface form of the post p, defining $gcxt(sf_i, p)$:

$$gcxt(sf_i, p) \coloneqq \langle lcxt(sf_i, p),\ globalcontext(p),\ cardinality(l_k, p)\rangle$$

$gcxt(sf_i, p)$ is the global context for the surface form sf_i of the post p, and it provides a new chance to select a candidate location with enough confidence when the local context $lcxt(sf_i, p)$ is not enough.

A *global ranking* is defined through the same rules and score functions illustrated for the local context-based ranking in Section 4.1. However, the ranking function has been re-defined to take into account the cardinality of each location in the global context:

$$\forall c_j \in C(sf_i),\ rk(c_j) = \sum_{l_k \in lcxt(sf_i, p)} score(c_j, l_k) \ + \sum_{l_k \in globalcontext(p)} score(c_j, l_k) \cdot cardinality(p, l_k)$$

Doing this, disambiguated locations that appear in more than one neighbors contribute with their scores multiple times. This models a *majority voting* mechanism that accounts for the fact that neighbors can contribute with contradictory information. The assumption is that a post is more frequently connected to more similar (that is, coherent) posts. Finally, the ranking function for the post $rk^*(p)$ is defined as in the local phase based on the individual rankings rk_i for all the surface forms $sf_i \in p$.

Given the differences between the global context $gcxt$ and the local context $lcxt$, we define a reinforcement confidence $threshold_{global} \neq threshold$ to accept a candidate as location in the global step. Preliminary experiments have shown that locations in the global context have less chance of being related to the target post than locations in the local context. Therefore, locations in the global context must provide, collectively, a greater reinforcement. It follows that $threshold_{global}$ should be greater than $threshold$. In the current implementation $threshold = 3$ and $threshold_{global} = 5$.

4.2.3 Propagation and disambiguation over the social network

The algorithm also uses the global graph G to *propagate* the information from disambiguated nodes to ambiguous nodes. A global disambiguation attempt is performed in the following two cases:

- On a post p, in case the local phase was not able to geolocate it.
- After the (local or global) disambiguation of a post p, on the set S_D of all its neighbors p_i which are still ambiguous: $S_D = \{p_i \mid (p, p_i) \in E \wedge loc(p_i) \neq \emptyset\}$.

These rules are applied *recursively*, allowing *chains* of disambiguations. Once a location has been disambiguated in a post, this information can propagate over the social network allowing the disambiguation of other posts, recursively. At each step, a global disambiguation attempt is performed on a node only if its $globalcontext$ has changed with respect to the previous attempt.

There exist surface forms which cannot be disambiguated. These include NER false positives (i.e., surface forms incorrectly detected by the NER) and gazetteer misses (i.e., locations not available in the gazetteer). To avoid an indefinite number of global disambiguation attempts on the same node, which would prevent the application of the algorithm in real-time scenarios, the number of global disambiguation attempts is bounded by a `max_attempts` parameter. After `max_attempts`[20] global disambiguation attempts, the surface forms of a node are considered to be *unable to disambiguate* and the node is removed from the graph G.

5 Experimental evaluation

This section presents the experimental results of the proposed methodology. In the following, for ease of discussion, a location provided by Twitter (geotag) is indicated as a "georeference" and a location provided by the CIME algorithm is indicated as a "geolocation." Moreover, the "local" step of the CIME algorithm that builds a context based on elements from the same post (Section 4.1) is referred to as *CIME local*, while the "global" step of the CIME algorithm that extends the context with elements from other connected posts in the behavioral social network (Section 4.2) is referred to as *CIME global*.

As discussed in the introduction, the CIME algorithm was developed with the goal of increasing the amount of geolocated images extracted from social media during an emergency to ultimately support rapid mapping activities. Therefore, we evaluated the proposed algorithm in the context of emergency events, evaluating the amount, the quality and the relevance of the geolocated posts and images, and discussing the impact of the geolocated images on rapid mapping activities.

Section 5.1 describes the experimental setting in terms of processing pipeline and the analysis method. Section 5.2 describes the emergency events selected as case studies. Section 5.3 summarizes the results and, finally, Section 5.4 discusses them.

5.1 Analysis method and processing pipeline

The post/image extraction process includes the following phases:

- *Crawling*. Posts from Twitter were crawled through the *Twitter search API* using a list of event-specific keywords as search parameters [3, 4].
- *Geolocation of posts*. Posts were geolocated as described in Section 4. Georeferences natively provided by Twitter were also stored for comparison and as a baseline.
- *Geolocation of images*. The geolocation assigned to a post by the algorithm was associated with all the images included in the post.

The results were analyzed considering the following criteria:

- *Recall*: defined as the percentage of items correctly retrieved with respect to a given target set. Since the focus has been placed on extracting and geolocating images, 100% corresponds to geolocating all images.
- *Precision*: defined as the percentage of relevant items among those retrieved. We can distinguish between:

[20] This parameter has been set to 3 in the current implementation.

- The precision of the geolocations.
- The relevance of the geolocated images to rapid mapping.

The geolocation precision is assessed by comparing CIME's geolocations against Twitter's native georeferences as well as manually annotated ones.

While evaluating recall and geolocation precision aims to demonstrate the validity of CIME by itself, evaluating the relevance of the geolocated images aims to demonstrate the usefulness of CIME for rapid mapping activities.

These evaluation criteria have been translated in the analysis summarized as follows:

- *Overall number of media geolocated by CIME.* Even if this value does not take into account the total number of media that can be potentially geolocated, it can be compared to the number of natively georeferenced tweets, quantifying the increase given by CIME.
- *Comparison to natively georeferenced locations.* In order to evaluate the precision of the media geolocated through CIME, we selected the ones which are *also* natively georeferenced, comparing the two locations. Georeferences are not always precise, but they are commonly used in applications and generally provide a reasonable quality [1]. To assess whether Twitter georeferences or CIME geolocations are more precise, we selected posts which include a media and we manually verified them using *Google Street View* if the locations were coherent with the media content.
- *Geolocation precision.* Considering the entire dataset, CIME geolocations were compared to manual annotations, quantifying the agreement.
- *Geolocation recall*. Considering the entire dataset, the percentage of tweets geolocated by CIME among those manually geolocated is calculated.
- *Geolocation relevance to the emergency management task*. We calculated the percentage of media geolocated by CIME which were manually labeled as "useful". Images are useful if they can contribute to rapid mapping activities (e.g., a river image, a flooded road, a blocked road, groups of people gathered in a street and so on). This value was compared to the percentage of useful natively georeferenced media.

Note that not all these analyses were conducted on all case studies introduced in Section 5.2. The reason for this is that manual annotations are heavily time consuming and not practically feasible over certain volumes. For example, the *geolocation precision* analysis requires us to label all the tweets in the dataset, to recognize the mentioned locations for each tweet, and to find them in a gazetteer so as to annotate the associated coordinates. Therefore, this analysis was performed on one case study, namely the Hurricane Sandy case study. The dataset for the Hurricane Sandy case study was manually annotated starting from a dataset included in [25, 26], which was already annotated to identify location references in the text. The goal of our further annotation was to manually disambiguate location references using OpenStreetMap.

5.2 Case studies

5.2.1 Floods in Southern England, 2014

The first case study considered in this work is based on tweets related to the floods which occurred in Southern England in 2014.

Twitter was crawled with the keywords: "England" and "flood England" from 10th to 15th of February 2014, obtaining a dataset of 108,757 tweets.

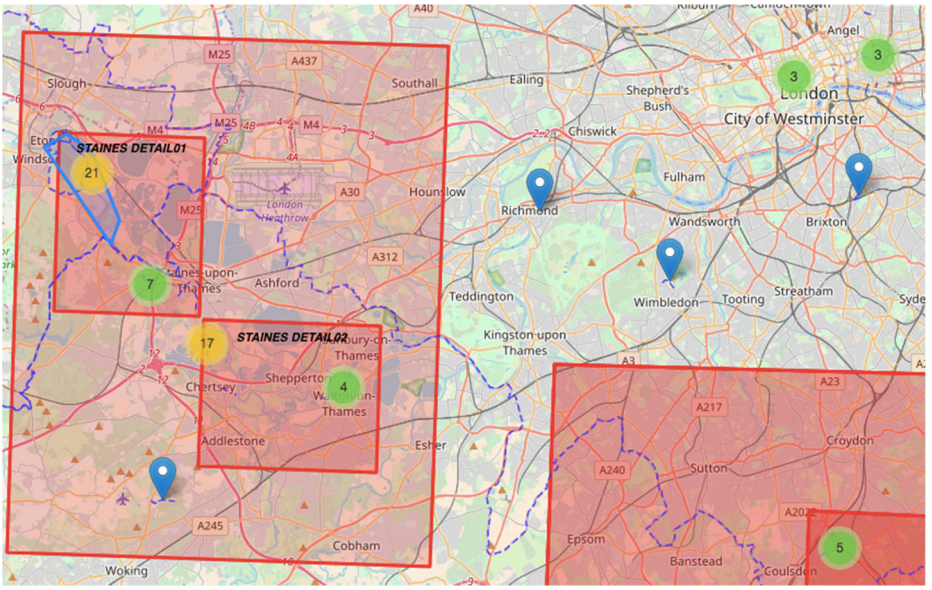

Fig. 4 A portion of a mapped area in the UK. The areas in red are the areas for which Copernicus EMS069 produced maps

In this time frame, the Copernicus EMS rapid mapping service was activated (activation EMSR069), producing rapid maps for the affected areas. In total 22 delineation maps were produced for the areas of Bridgwater, Hambledon, Kenley, Maidenhead, Staines and Worcester. Maps produced from Copernicus EMS are based on the analysis of high-resolution SAR (Synthetic Aperture Radar) images from the Copernicus satellite system. It has to be noted that, as stated in the map description, "the thematic accuracy might be lower in urban and forested areas due to known limitations of the analysis technique"[21].

Figure 4 shows a portion of a mapped area in the UK. The areas in red are the areas for which Copernicus EMS produced maps. Dots with numbers and markers refer to geolocated tweets with images in the area of interest. The produced maps can be interactively browsed and the geolocated tweets/images individually inspected. Figure 5 is a detail of the delineation map produced by Copernicus EMS for Staines, with the areas in blue representing flooded areas.

Figure 6 shows an example of a tweet geolocated in Queen's Road, Datchet, by the CIME algorithm (local). The tweet is not georeferenced, but the position was derived from the tweet's text through CIME. On the right, the extracted information is shown. The text in bold is the textual description of the location, while the coordinates correspond to the center of the location, where the tweet has been placed on the map.

Figure 7 shows other cases where CIME identified the locality rather than the street. In these cases, the posts (and the attached media) were all linked to the center of the locality. Locations at this level of precision can be useful since media can carry additional

[21] http://emergency.copernicus.eu/mapping/list-of-components/EMSR069

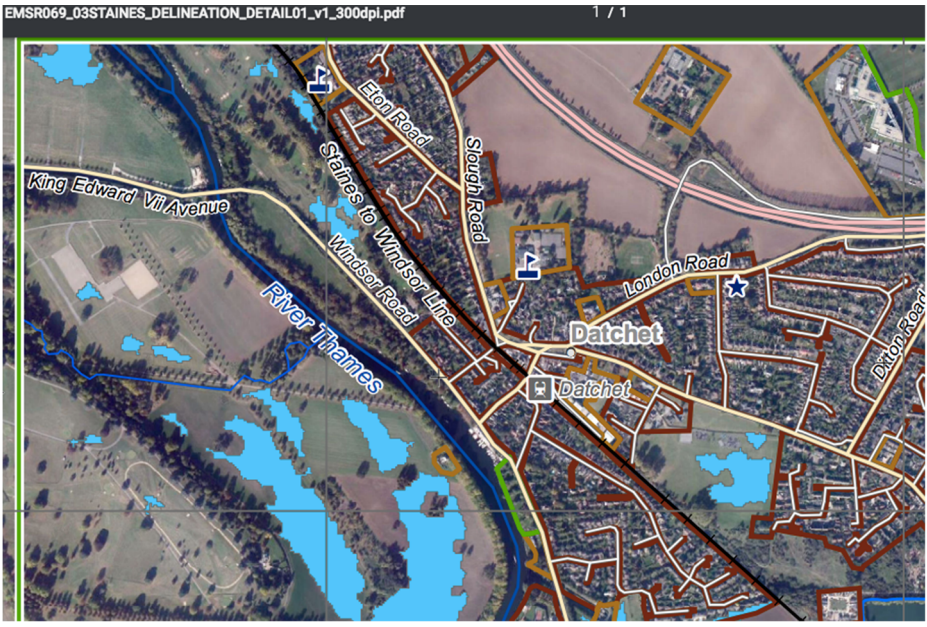

Fig. 5 Example of delineation map (detail of a part of the map Staines Detail01, Copernicus EMSR069)

information which can help human operators creating the maps. For instance, the image on the right contains the name of a restaurant, which is clearly visible, and the image on the left shows rail tracks that can be easily identified in the area inferred by CIME.

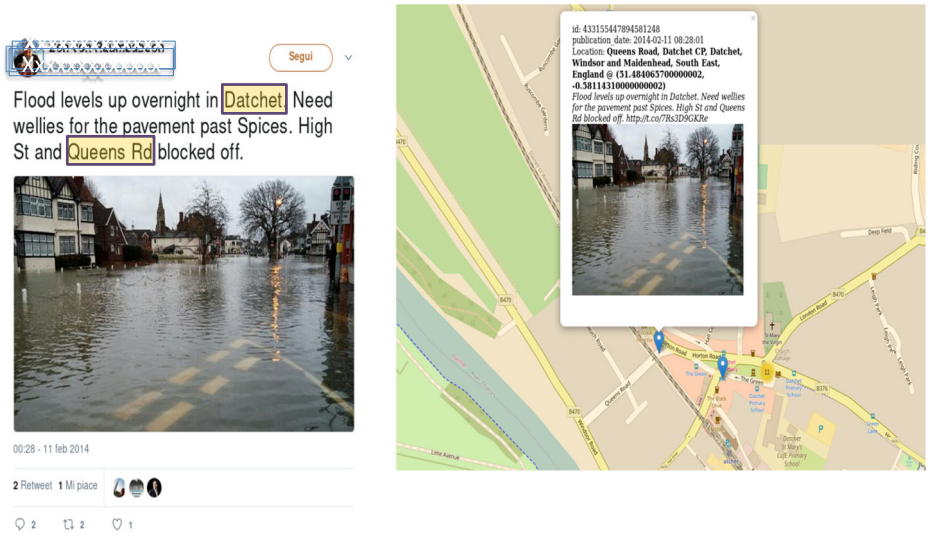

Fig. 6 Geolocated tweet with image (CIME local)

Fig. 7 Geolocated coarse-grained images (CIME)

5.2.2 Hurricane Sandy in New York, 2012

Hurricane Sandy was the most devastating Atlantic storm in 2012, causing severe damage in New Jersey and New York. An annotated dataset of tweets mainly related to different locations of New York City and its surroundings was provided by the University of Southampton IT Innovation Centre [25]. The original dataset contains 1,996 tweets with manual annotations of the locations mentioned in the tweets. The dataset has been used as a basis for the evaluation of the algorithm, with an additional annotation performed by two persons for each post for location disambiguation. A total of 280 different locations that can be manually disambiguated has been obtained.

5.3 Results

5.3.1 Overall number of media geolocated by CIME

The first part of Tables 2 and 3 describes the datasets in terms of overall number of tweets, tweets with images and natively georeferenced tweets. We can observe that only the 0.28% and the 0.15% of tweets with images are natively georeferenced in the two case studies.

The second part of Tables 2 and 3 summarizes the results of CIME in terms of the total number of geolocated tweets with images, also distinguishing between images geolocated through the local and the global phase. In the Southern England case, 0.64% of tweets

Table 2 Overall number of tweets with images, natively georeferenced and geolocated by CIME for the Southern England case

Post type	Number	Percentage
Tweets	108,757	100%
Tweets with images	6,256	5.7%
Georef tweets	3,333	3.06%
Georef tweets with images	310	0.28%
Geolocated tweets with images		
CIME local	378	0.35%
CIME local and global	695	**0.64**%

with images have been geolocated, which is more than twice the number of natively georeferenced images. In the Hurricane Sandy case, 1.10% of tweets with images have been geolocated, which is more than seven times the number of natively georeferenced images.

To evaluate the *usefulness* of the geolocated images in terms of *new information* provided to operators, we assessed whether the sets of geolocated and georeferenced images overlap. In the Southern England case, there are 45 overlapping images, which is only 6.5% of the CIME geolocated images, while in the Hurricane Sandy case, there are 3 overlapping images, which is only 12% of the geolocated images. This means that the vast majority of images geolocated through CIME would not be available by considering only georeferenced tweets.

Note that the percentages in Tables 2 and 3 are computed with respect to the total number of tweets in the datasets. The reason for such low outcome percentages for CIME is that only the tweets with images were considered in this analysis, which are only 3.7% and 5.7% of the tweets in the two datasets. If we focus solely on tweets with images, CIME was able to geolocate 11.1% and 33.78% of the posts in the two cases.

5.3.2 Comparison with natively georeferenced locations.

To evaluate the accuracy of the images geolocated through CIME, we focused on those which are *also* natively georeferenced for comparison.

This analysis was performed only on the Southern England case, because the other case study did not provide enough georeferenced and geolocated tweets (see Table 3).

Table 3 Overall number of tweets with images, natively georeferenced and geolocated by CIME for the Hurricane Sandy case

Post type	Number	Percentage
Tweets	1996	100%
Tweets with images	74	3.7%
Georef tweets	51	2.55%
Georef tweets with images	3	0.15%
Geolocated tweets with images		
CIME local	22	1.10%
CIME local and global	25	**1.25**%

The georeferences and the locations provided by CIME for each tweet were validated by checking them on Google Maps and also by using Street View when needed, to assess whether they locate the content shown in the images correctly. The correct place for an image was considered as the place to be mapped (e.g., if a picture of a church was taken from a distance, the true location is the position of the church rather than the position where the picture was taken).

The locations were evaluated as follows: *precise* if on sight (max 1,000 mt.), *same area* if between 1 and 5 km, *imprecise* if $>$ 5 km.

Figure 8 shows the results of the comparison of georeferenced and geolocated images. On the left, all images were considered for the comparison, while on the right, images which were manually tagged as unuseful (i.e., they do not show places, and therefore cannot be evaluated) were excluded.

To summarize the results, the following cases were identified: *precise correspondence* when both locations are precise, *same area* when both locations are in the same area or one of the two is in the same area and the other one is precise, *geoloc correct, geotag incorrect* when the geolocation is precise/same area while the other one is unprecise and vice-versa for *geoloc correct, geotag incorrect*. *Imprecise* denotes both imprecise locations.

We can observe that in 53% of cases, both CIME geolocations and georeferences referred precisely to the same place (precise correspondence or same area between 1 and 5 km). In several cases, only one of the two locations was accurate, and, in particular, CIME geolocations appeared to be more accurate than georeferences (23% vs 8%). It has also to be noted that about half of the images analyzed did not show useful content.

5.3.3 Geolocation precision

To evaluate the precision of CIME in terms of correctly disambiguated locations, all tweets were manually annotated to identify mentioned locations. The manual annotation involved both *location recognition* and *location disambiguation* (see Section 2 for details about these phases). While location recognition was purely a manual task, location disambiguation was performed manually with the aid of OpenStreetMap gazetteer accessed through the Nominatim web interface, thus using the same knowledge base of CIME. The gazetteer allows the human annotator to select a specific location, rather than just a name, so that a location mention can be associated with a pair of coordinates. This kind of manual annotation is heavily time-consuming and not practically feasible over certain volumes. Therefore, this

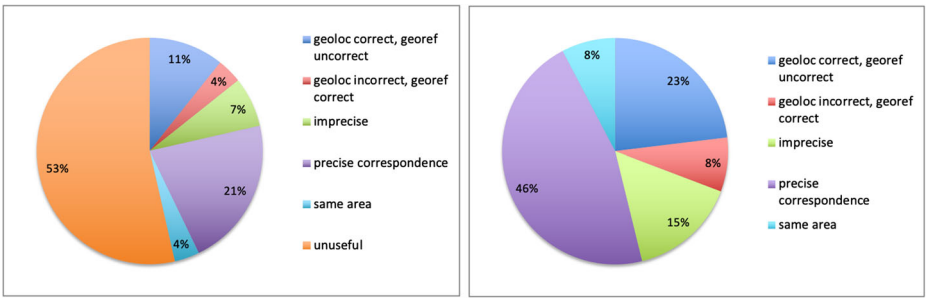

Fig. 8 Comparison of native georeferences and CIME geolocations. *Left*: all images. *Right*: only useful images

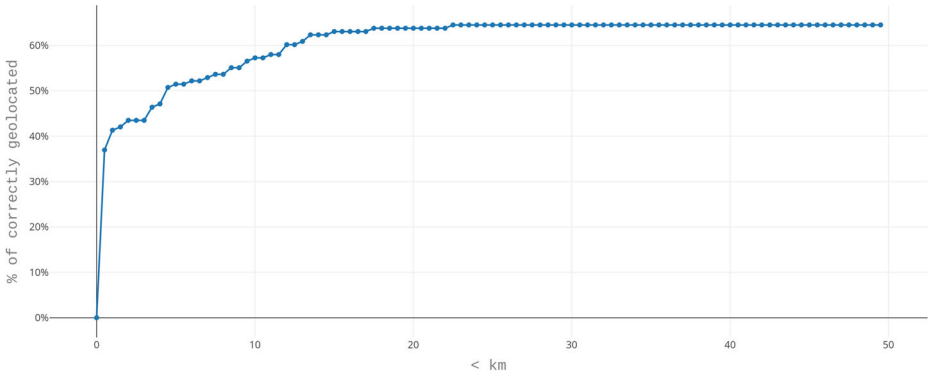

Fig. 9 Geolocation precision of CIME with respect to manual annotation varying the allowed distance in the range 0–50 km, 0.5 km steps

analysis was performed on the Hurricane Sandy case study, which included approximately 2,000 tweets, but not on the Southern England case study, which included over 100,000 tweets. For this analysis, each location was interpreted as a point by considering its center. A CIME location is considered "precise" if the distance to the manually annotated location is below a certain threshold. We varied the distance threshold in the range of 0—50 km with a 0.5 km step. The distance was computed using the *geodesic* distance with the WGS-84 ellipsoid.

Results are given in Fig. 9. Precision at 1 km is equal to 41%, at 10 km it is 57%, and it settles at 64% at 18 km. Precision does not increase significantly beyond this threshold up to 50 km. Indeed, we can observe that the largest increase in precision was obtained in the first few kilometers. Figure 10 focuses on the range of 0–5 km with a 0.1 km step and shows that the precision at 0.1 km is already 37%. Therefore, the majority of locations correctly disambiguated by CIME were identified with a very high geographic accuracy.

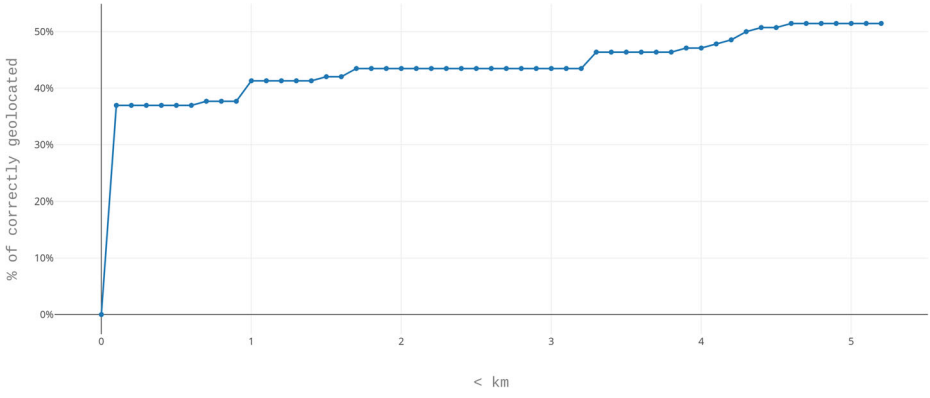

Fig. 10 Geolocation precision of CIME with respect to manual annotation varying the allowed distance in the range of 0—5km, 0.1 km steps

5.3.4 Geolocation recall

The second part of Tables 1 and 2 summarizes the total number of images geolocated by CIME. However, to evaluate the recall of the algorithm, it is necessary to take into consideration *all* the locations mentioned in tweets and to calculate the ratio of locations correctly identified. This analysis was based on the same annotations used to evaluate the precision on the Hurricane Sandy case study (Section 5.3.3).

Overall, CIME disambiguated 21% of the mentioned locations. The recall of CIME appears to be relatively lower than the precision. This is mainly due to the fact that CIME always seeks a *context* to disambiguate a location reference. Therefore, all the tweets mentioning just one isolated reference are excluded, even when that location is intuitively easy to disambiguate. This discussion is further developed in Section 5.4.

5.3.5 Relevance for the task to be performed

Geolocated and georeferenced tweets with images were annotated with respect to their relevance (that is, their potential usefulness) for the rapid mapping activities. A prerequisite for being useful is being in the area of interest (mapped area), therefore only those images were considered. An image was considered relevant if it shows roads, streets, areas, blocks or gatherings of people, with or without water. We were interested in comparing the relevance of georeferenced and geolocated images. This analysis was performed only on the Southern England case, because the other case study did not provide enough geolocated and georeferenced images (see Section 5.3.1).

The results of this analysis are reported in Table 4, where the relevance for images geolocated through the local and the global phase was evaluated separately. As shown, the majority of images with any kind of attached location is relevant, but CIME geolocations (local or global) are more relevant than natively georeferenced images. Possible reasons for this phenomenon are discussed in Section 5.4.

5.4 Discussion

Results demonstrate the validity of locations inferred by CIME, both in terms of *volume* and in terms of *accuracy*, with respect to natively georeferenced ones. Since natively georeferenced tweets are typically used in applications where located tweets are needed [1], the results show that CIME geolocated tweets can be reliably used in the same applications. Moreover, most of the CIME geolocated tweets are not natively georeferenced, thus providing an additional and complementary source.

Not only CIME geolocated tweets (local and global) double the number of geolocated media with respect to those natively georeferenced, but those media were also evaluated as more relevant than natively georeferenced ones (66% vs 79/83%). This phenomenon can be explained by the fact that, often, the text linked to relevant images is more descriptive

Table 4 Image relevance of natively georeferenced and geolocated tweets

Tweet type	Percentage of relevant images
Georeferenced	66%
Geolocated (CIME local)	79%
Geolocated (CIME global)	83%

and contains more references to locations, thus being a better target for CIME. Moreover, relevant media often lead to many interactions in terms of replies, retweets, etc., thus giving CIME global disambiguation an advantage.

The analysis reported in Fig. 8 for the Southern England case has shown that CIME geolocations are, on average, more precise than geotags. Additionally, this plot highlights the limitations of natively georeferenced tweets, which are often taken as ground truth in the analysis of existing algorithms (see for instance [17], which reports many methods using geotags as ground truth in Table 2). Indeed, among all the georeferenced tweets with a place-related image, 38% do not show a precise correspondence between the place depicted in the image and the tweet's geotag.

One of the main limitations of CIME is related to its recall, comparatively lower than the precision on the same dataset. This is mainly the result of an algorithm's design choice, which sees a tweet's context as a *necessary* (but not sufficient) element to disambiguate location mentions. Therefore, all the isolated location references (i.e., location references without a context) are inherently unable to be disambiguated. CIME has been designed to privilege the precise disambiguation of context-dependent location references, which usually correspond to streets, roads and small villages, rather than locations which can be reliably disambiguated from the references themselves (e.g., "New York" in the Hurricane Sandy case study). Indeed, there are at least three important goals of CIME which are not reflected in a precision/recall analysis and which stem from the rapid mapping task. First of all, not all the locations have the same importance for rapid mapping, and often information about less known and comparatively small locations can add more value than information about large cities already covered by other media. Secondly, we can define a precision only by setting a certain accuracy threshold (see Fig. 9), which is application-dependent and needs to be low for rapid mapping. Finally, to aid rapid mapping activities, it is necessary to ensure both the relevance of the geolocated media and the precise correspondence of the locations depicted in the images to the geolocations. The analysis and the evaluation criteria presented in this paper try to capture these aspects. Nonetheless, nothing prevents the use of CIME as part of larger systems in conjunction with other high-recall geolocation methods to disambiguate isolated location references.

Comparing geolocation algorithms tailored to specific domains and applications (such as rapid mapping) is not an easy task. Indeed, as discussed, taking into consideration only precision/recall is often reductive and several other features — the relevance of which is largely domain-dependent — must also be taken into account. First of all, as described earlier, the precision is a function of the threshold to consider a location as "accurate". Therefore, two different geolocation methodologies with different goals could perform better for different thresholds. CIME, in particular, privileges accuracy rather than volume, given a target domain that requires highly accurate geolocated items for rapid mapping purposes. This can be seen in Fig. 10, which highlights the precision at 1km (about 40%) and at 5km (about 50%). Moreover, other features and constraints of a geolocation algorithm have a major impact on its applicability in specific domains. For example, language-specific features, the necessity of prior knowledge about the target event/area, or application-specific preprocessing phases. All these different features hinder comparisons between different geolocation algorithms, because stronger constraints could bring about better performance but limit the applicability in specific domains and contexts. Finally, different geolocation algorithms could target different steps of the geolocation process (see Section 2) or different aspects of social media locations (such as home or post location, see Section 3), further hindering comparisons.

On this basis, the results of CIME on the Hurricane Sandy case study have been compared to those presented in [26] for the same dataset (originally described in [25]). A major difference between the two methodologies is the need for a preprocessing phase for the algorithm described in [26], with indices built limited to a specific target geographical area being considered. In contrast, CIME is designed to analyze tweets without any preparation or training phase, aimed at disambiguating locations by reacting to events in real-time.

The precision @1km obtained by CIME is comparable and slightly higher than the precision obtained in the tested datasets for geographical distances mentioned in [26] (40% vs 18-36%). The recall appears to be lower for CIME (21% vs 76-81%), and the motivations for a lower recall have been previously discussed.

6 Implementation

The algorithm has been developed in Python. The NER module is based on the Stanford CoreNLP library [42, 43] accessed through a Python wrapper[22], with a caseless model to account for the frequently miscapitalized social media text [1]. OpenStreetMap is the gazetteer through which candidate locations are retrieved. It is accessed via a local instance of the Nominatim search engine, both through its API[23] and through direct calls to its PostgreSQL database to extract information not exposed by the API, such as complete hierarchical relationships. The Python libraries used by CIME include `geopy` to compute geographical distances, `numpy` and `pandas` for data processing and `networkx` to build and manage both the *global* network of tweets and the *local* networks of tweets' locations. Note that both the NER module and the gazetteer can be replaced by other functionally equivalent libraries without affecting the core algorithm.

CIME can be accessed as a service through dedicated REST APIs based on `django`. The APIs allow the user to submit a tweet and receive back the geolocation output in JSON. For each disambiguated location, the algorithm returns its geographic structure as a GeoJSON (point, line, polygon or multi-polygon), its centre, its fully qualified name (as returned by the gazetteer) and other gazeteeer-dependent information such as the "class" and the "type" provided by Nominatim.

CIME has been designed to use several caches, at the gazetteer, at the NER and at the tweet level to minimize executions times. On average, it is able to geolocate a tweet in less than 1 second. The main computational requirements are those related to the local execution of Nominatim and CoreNLP.

7 Concluding remarks and future work

A fundamental objective of this work was to understand whether it is possible to automatically retrieve and geolocate multimedia content from social media in real-time and with high precision for rapid mapping purposes. The evidence gathered from the analyzed case studies suggests that relevant, geolocated content can be retrieved by complementing the extraction of natively georeferenced data with a context-based geolocation algorithm that exploits

[22] https://github.com/Lynten/stanford-corenlp
[23] https://wiki.openstreetmap.org/wiki/Nominatim

many available modalities (text, social network, and social interactions) for geolocation purposes.

Given that social media content is timely, as it is available in the first few hours after an emergency situation, it can represent an enabler for the faster delivery of maps of the affected areas in the first 24 hours after the event, when data from satellite and other official sources are not yet available. Clearly, increasing the volume of relevant images increases the benefits.

In this regard, the experimental evaluation carried out in this work has shown that CIME can double the number of geolocated images extracted from the stream of after-event posts with respect to using only natively georeferenced tweets. Moreover, we have shown that the accuracy of the locations inferred by CIME with respect to the locations shown in the images is higher or similar than natively georeferenced tweets, with also a higher percentage of locations linked to images relevant to rapid mapping activities. This analysis shows that images geolocated through CIME can effectively be used in applications, as natively georeferenced tweets are usually considered the standard for the task [17]. Furthermore, we have evaluated the precision of the inferred locations, showing how CIME is particularly effective in geolocating posts at very low distances, with most of the posts correctly geolocated at less than 5 km (41% precision at 1km, 57% precision at 10km).

Possible future directions include the extension of the algorithm to other information sources, including other social media, news sources and crowdsourcing. The flexible context-based approach of the proposed methodology can be easily extended to account for other contextual information that can improve the disambiguation process.

Future work also includes more extensive validations. The algorithm was analyzed in detail only in the case of emergency situations, and in particular floods and storms. The general validity of the results obtained in other types of events is still an open research topic. With regard to the evaluation of distances, this research applied a simple formula measuring distances between centers, independently from the content shown in images. A more precise evaluation should also take into account the type of picture (aerial, landscape, close detail). However, this appears difficult to automate without using image recognition techniques, which can, in turn, introduce other sources of errors. In addition, there is a need to evaluate the maximum level of precision that can be achieved given an image and its associated text and metadata. Even through crowdsourcing and manual annotations, a precise location cannot be associated to images in many cases, thus limiting the maximum theoretical precision achievable by an algorithm.

CIME relies on a number of external components (including the crawler, the NER and the gazetteer) that inevitably affect the performance of the resulting system. Even though CIME is, in principle, independent with respect to these components, the need to develop a system that can be deployed and tested on real events has inevitably led to specific choices with respect to these components. Taking into consideration multiple alternatives for each of these components would lead to an exponential number of possible configurations that would be impossible to evaluate experimentally. Nonetheless, further validation in this respect should be the subject of future work.

Another threat is related to the parameters of the algorithm presented in Section 4. The parameters are currently set based on assumptions related to their meaning and the experience acquired from the analyzed case studies. However, more sophisticated techniques can be employed to optimize the parameters for a single event or even a single phase in an event, further improving the geolocation performance. For example, parameters could be learned through machine learning techniques. Given the fact that CIME does not rely on any prior

information about the event, automatically learning event-specific parameters appears to be challenging.

Current work in the literature is going in the direction of using machine learning techniques to also filter out irrelevant images and classify useful ones, such as, for instance, in [44], where a social media pipeline combining human and machine evaluation is proposed to filter images and support damage assessment. The combination of text-based approaches, such as the one presented in this paper, with image-based techniques, should be the target of future work.

The results presented in this paper consider Twitter as the main information source, as many other papers in the literature do, given its relatively accessible API. However, the proposed methodology can be easily extended to other social media. Future research may explore the benefits of crawling other sources. This could provide additional content and help meet the information richness and reliability requirement needed for rapid mapping tasks by reinforcing decisions with possibly redundant information. Exploiting multiple social media has recently proved to be useful for enhancing the extracted information in terms of relevant keywords [41], also using CIME as a tool for geolocating YouTube posts. However, following the evolution of the content published on different social media during an emergency represents a rather unexplored research area that is very relevant for the rapid mapping context.

Acknowledgements This work was funded by the European Commission H2020 projects E2mC "Evolution of Emergency Copernicus services" under project No. 730082 and project Crowd4SDG "Citizen Science for Monitoring Climate Impacts and Achieving Climate Resilience" under project no. 872944. 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 of the information contained in this work. The authors thank Paolo Ravanelli for his support in developing the crawlers for E2mC and Crowd4SDG and the E2mC post-management platform, Andrea Autelitano and Nicole Gervasoni for some of the tweet annotations and Bernard Allenbach from the University of Strasbourg for his comments on a previous draft of this paper. Finally, the authors would like to thank Stuart E. Middleton from the University of Southampton for providing access to the labeled Sandy dataset.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

References

1. Castillo C (2016) Big crisis data: Social media in disasters and time-critical situations. Cambridge University Press
2. Avvenuti M, Cresci S, Del Vigna F, Fagni T, Tesconi M (2018) Crismap: a big data crisis mapping system based on damage detection and geoparsing. Inf Syst Front 20(5):993–1011
3. Havas C, Resch B, Francalanci C, Pernici B, Scalia G, Fernandez-Marquez JL, Achte TV, Zeug G, Mondardini MRR, Grandoni D, Kirsch B, Kalas M, Lorini V, Rüping S (2017) E2mc: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. Sensors 17(12):2766. https://doi.org/10.3390/s17122766

4. Francalanci C, Guglielmino P, Montalcini M, Scalia G, Pernici B (2017) IMEXT: A method and system to extract geolocated images from tweets - analysis of a case study. In: 11th International Conference on Research Challenges in Information Science, RCIS 2017, Brighton, pp 382–390
5. Scalia G (2017) Network-based content geolocation on social media for emergency management, Master Thesis, Politecnico di Milano
6. Fernandez-Marquez JL, Francalanci C, Mohanty S, Mondardini R, Pernici B, Scalia G (2019) E2mC: Improving rapid mapping with social network information. In: Organizing for the Digital World. Springer, pp 63–74
7. Francalanci C, Pernici B, Scalia G (2018) Exploratory spatio-temporal queries in evolving information. In: Doulkeridis C, Vouros GA, Qu Q, Wang S (eds) Mobility Analytics for Spatio-Temporal and Social Data. Springer International Publishing, Cham, pp 138–156
8. Jurgens D (2013) That's what friends are for: Inferring location in online social media platforms based on social relationships. In: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013. AAAI, Cambridge, pp 273–282
9. Li R, Wang S, Deng H, Wang R, Chang KC-C (2012) Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 1023–1031
10. Rahimi A, Vu D, Cohn T, Baldwin T (2015) Exploiting text and network context for geolocation of social media users. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Denver, pp 1362–1367
11. Rout D, Bontcheva K, Preoţiuc-Pietro D, Cohn T (2013) Where's@ wally?: a classification approach to geolocating users based on their social ties. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. ACM, pp 11–20
12. Cheng Z, Caverlee J, Lee K (2010) 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, pp 759–768
13. Inkpen D, Liu J, Farzindar A, Kazemi F, Ghazi D (2015) Detecting and disambiguating locations mentioned in Twitter messages. In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, pp 321–332
14. Kinsella S, Murdock V, O'Hare N (2011) I'm eating a sandwich in Glasgow: modeling locations with tweets. In: Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM, pp 61–68
15. Ji Z, Sun A, Cong G, Han J (2016) Joint recognition and linking of fine-grained locations from tweets. In: Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp 1271–1281
16. Zhang W, Gelernter J (2014) Geocoding location expressions in twitter messages: A preference learning method. J Spatial Inf Sci 2014(9):37–70
17. Zheng X, Han J, Sun A (2018) A survey of location prediction on Twitter. IEEE Trans Knowl Data Eng 30(9):1652–1671. https://doi.org/10.1109/TKDE.2018.2807840
18. Laylavi F, Rajabifard A, Kalantari M (2016) A multi-element approach to location inference of twitter: A case for emergency response. ISPRS Int J Geo-Inf 5(5):56
19. Ajao O, Hong J, Liu W (2015) A survey of location inference techniques on twitter. J Inf Sci 41(6):855–864
20. Shen W, Wang J, Han J (2015) Entity linking with a knowledge base: Issues, techniques, and solutions. IEEE Trans Knowl Data Eng 27(2):443–460
21. Haklay MM, Weber P (2008) OpenStreetMap: User-generated street maps. IEEE Pervasive Comput 7(4):12–18
22. Liu F, Vasardani M, Baldwin T (2014) Automatic identification of locative expressions from social media text: A comparative analysis. In: Proceedings of the 4th International Workshop on Location and the Web. ACM, pp 9–16
23. Mahmud J, Nichols J, Drews C (2012) Where is this tweet from? inferring home locations of twitter users. ICWSM 12:511–514
24. Li C, Sun A (2017) Extracting fine-grained location with temporal awareness in tweets: A two-stage approach. J Assoc Inf Sci Technol 68(7):1652–1670
25. Middleton SE, Middleton L, Modafferi S (2014) Real-time crisis mapping of natural disasters using social media. IEEE Intell Syst 29(2):9–17
26. Middleton SE, Kordopatis-Zilos G, Papadopoulos S, Kompatsiaris Y (2018) Location extraction from social media: Geoparsing, location disambiguation, and geotagging. ACM Trans Inf Syst 36(4):40:1–40:27. https://doi.org/10.1145/3202662

27. Chong W-H, Lim E-P (2018) Exploiting user and venue characteristics for fine-grained tweet geolocation. ACM Trans Inf Syst (TOIS) 36(3):26
28. Chong W-H, Lim E-P (2019) Fine-grained geolocation of tweets in temporal proximity. ACM Trans Inf Syst (TOIS) 37(2):17
29. Ghufran M, Quercini G, Bennacer N (2015) Toponym disambiguation in online social network profiles. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, p 6
30. Davis Jr CA, Pappa GL, de Oliveira DRR, de L Arcanjo F (2011) Inferring the location of Twitter messages based on user relationships. Trans GIS 15(6):735–751
31. McGee J, Caverlee J, Cheng Z (2013) Location prediction in social media based on tie strength. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, pp 459–468
32. Sakaki T, Okazaki M, Matsuo Y (2013) Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans Knowl Data Eng 25(4):919–931
33. Paraskevopoulos P, Palpanas T (2016) Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. Soc Netw Anal Min 6(1):89:1–89:16
34. Nugroho R, Yang J, Zhao W, Paris C, Nepal S (2017) What and with whom? identifying topics in Twitter through both interactions and text. IEEE Transactions on Services Computing, Early access
35. Francalanci C, Pernici B, Scalia G, Zeug G (2018) Talking about places: considering context in geolocation of images extracted from tweets. GI_Forum 2018 1:243–250
36. Thomason A, Griffiths N, Sanchez V (2016) Context trees: Augmenting geospatial trajectories with context. ACM Trans Inf Syst (TOIS) 35(2):14
37. de Albuquerque JP, Herfort B, Brenning A, Zipf A (2015) A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. Int J Geogr Inf Sci 29(4):667–689
38. Brangbour E, Bruneau P, Marchand-Maillet S (2018) Extracting flood maps from social media for assimilation. In: 2018 IEEE 14th International Conference on e-Science (e-Science). IEEE, pp 272–273
39. Song X, Shibasaki R, Yuan NJ, Xie X, Li T, Adachi R (2017) Deepmob: learning deep knowledge of human emergency behavior and mobility from big and heterogeneous data. ACM Trans Inf Syst (TOIS) 35(4):41
40. Panteras G, Wise S, Lu X, Croitoru A, Crooks A, Stefanidis A (2015) Triangulating social multimedia content for event localization using flickr and twitter. Trans GIS 19(5):694–715
41. Autelitano A, Pernici B, Scalia G (2019) Spatio-temporal mining of keywords for social media cross-social crawling of emergency events. GeoInformatica 23(3):425–447. https://doi.org/10.1007/s10707-019-00354-1
42. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014. System Demonstrations, Baltimore, pp 55–60
43. Al-Rfou R, Kulkarni V, Perozzi B, Skiena S (2015) POLYGLOT-NER: Massive multilingual Named Entity Recognition. In: Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, pp 586–594
44. Alam F, Ofli F, Imran M (2018) Processing social media images by combining human and machine computing during crises. Int J Hum Comput Interact 34(4):311–327

Gabriele Scalia completed his Ph.D. in Computer Science at Politecnico di Milano, with an interdisciplinary thesis titled "Machine Learning-driven Integration, Knowledge Extraction and Uncertainty Management for Scientific Data". During his Ph.D. he has mainly focused on machine learning and deep learning techniques with applications to chemical and biological systems, spending one year at MIT and an additional period at the Broad Institute of MIT and Harvard. His research interests also include information extraction from complex and highdimensional data, with applications to biochemical, geospatial data, and citizen science data. Currently, he is an Associate Scientist (ML/AI) at Roche, doing research on methods for the analysis of graph-based and sequence-based data in the biomedical domain to facilitate target discovery.

Chiara Francalanci is associate professor of information systems at Politecnico di Milano. She has a Ph.D. in Computer Science from Politecnico di Milano and a Master in Management Engineering from the Business School of Politecnico di Milano. As part of her post doctoral studies, she has worked for two years at the Harvard Business School as a Visiting Researcher where she has graduated in September 1995 in Management of the Information Systems Resource. Her research focuses on information system engineering and, in particular, on feasibility analyses. Her main research interests in this area are information design, architectural design of information systems, and cost-benefit analyses, as fundamental components of a feasibility study.

Barbara Pernici is full professor in Computer Engineering at the Politecnico di Milano. Her research interests include adaptive information systems, data quality, IS energy efficiency, and social media analysis. She has published more than 70 papers in international journals and about 350 papers at international level. She has lead the information systems group of Politecnico di Milano in many European and national projects. She is currently responsible of the Politecnico unit of the Crowd4SDG European H2020 project. She was an elected chair of TC8 Information Systems of the International Federation for Information Processing (IFIP), of IFIP WG on Information Systems Design, vice-chair of the IFIP WG on Services-Oriented Systems, and chair of the Steering Committee of the international Conference on Advanced Information Systems Engineering (CAiSE). She has chaired or cochaired main conferences, as general chair or program chair, including CAiSE, ER, BPM, ICSOC, Coopis, tracks in ICSE and ICIS.