Journal of Geovisualization and Spatial Analysis (2018) 2: 2
https://doi.org/10.1007/s41651-017-0010-6

TAGGS: Grouping Tweets to Improve Global Geoparsing for Disaster Response

Jens A. de Bruijn[1] · Hans de Moel[1] · Brenden Jongman[1,2] · Jurjen Wagemaker[3] · Jeroen C. J. H. Aerts[1]

Published online: 26 December 2017

Abstract
Timely and accurate information about ongoing events are crucial for relief organizations seeking to effectively respond to disasters. Recently, social media platforms, especially Twitter, have gained traction as a novel source of information on disaster events. Unfortunately, geographical information is rarely attached to tweets, which hinders the use of Twitter for geographical applications. As a solution, geoparsing algorithms extract and can locate geographical locations referenced in a tweet's text. This paper describes TAGGS, a new algorithm that enhances location disambiguation by employing both metadata and the contextual spatial information of groups of tweets referencing the same location regarding a specific disaster type. Validation demonstrated that TAGGS approximately attains a recall of 0.82 and precision of 0.91. Without lowering precision, this roughly doubles the number of correctly found administrative subdivisions and cities, towns, and villages as compared to individual geoparsing. We applied TAGGS to 55.1 million flood-related tweets in 12 languages, collected over 3 years. We found 19.2 million tweets mentioning one or more flood locations, which can be towns (11.2 million), administrative subdivisions (5.1 million), or countries (4.6 million). In the future, TAGGS could form the basis for a global event detection system.

Keywords Geoparsing · Geocoding · Geotagging · Floods · Twitter · Geolocation · Disaster response

Introduction

Each year, natural disasters affect roughly one million people, causing thousands of deaths and tens of billions of US dollars in damages (UNISDR 2015). The availability of timely and accurate information about the impacts of an ongoing event can assist relief organizations in enhancing their disaster response activities, and thus mitigate the consequences of disasters (de Perez et al. 2014; Messner and Meyer 2006; Pennington-rowsell et al. 2005). Information about an ongoing event is, however, often difficult to obtain. Such data is generally collected using measurement instruments such as remote sensors (e.g., Sun et al. 2000), from local relief and response professionals, and analyses of media reports (Jongman et al. 2015; Kordopatis-Zilos et al. 2015). Recently, social media, and in particular Twitter, has gained traction as a novel source of information on disaster events. The Twitter posts ("tweets") that are sent out by millions of users around the globe hold great potential in disaster management (Carley et al. 2016; Jongman et al. 2015; Sakaki et al. 2010). When correctly analyzed, they can improve the detection of disasters (Ghahremanlou et al. 2014) and provide valuable information about the societal impacts of ongoing disaster events (Fohringer et al. 2015; Gao et al. 2011; Jongman et al. 2015). In computer science, social media has been studied extensively. Researchers have also developed several applications for applied geographic research. Examples of such applications include detection of flood events (Jongman et al. 2015) and earthquake disasters (Crooks et al. 2013; Sakaki et al. 2010).

One of the key issues in using Twitter information to assess the impacts of natural disasters entails accurately localizing individual tweets. Twitter allows users to automatically attach their current GPS location to a tweet, specifying their position at the moment a tweet is posted (Sakaki et al. 2010). However, because this feature is turned off by default, only 0.9% of the tweets have

✉ Jens A. de Bruijn
j.a.debruijn@outlook.com

[1] Institute for Environmental Studies, VU University, Amsterdam 1081, HV, The Netherlands

[2] Global Facility for Disaster Reduction and Recovery, World Bank Group, D.C, Washington 20433, USA

[3] FloodTags, The Hague 2511, BE, The Netherlands

coordinate information attached (Lee et al. 2013). Several approaches exist for the localization of social media posts. Language models typically use a collection of training posts with corresponding geotagged images (i.e., the location where the image was taken is known) to determine the most likely location of a new post (Kordopatis-Zilos et al. 2015). However, a very large training corpus is required to apply language models to temporally volatile events, such as floods. Otherwise, new posts are unlikely to be geotagged to a location where no event occurred in the training data (Kordopatis-Zilos et al. 2015). Other approaches employ text and/or metadata matching to a gazetteer to detect a user's residence, the location from which the tweet was sent, or the location to which a tweet refers (e.g., Schulz et al. 2013). Geoparsing (also referred to as geotagging or geolocalization) algorithms extract and locate these referenced geographical locations (also known as *toponyms*) from a text. Research has demonstrated that geoparsing algorithms can dramatically increase the number of geoparsed posts (e.g., Gelernter and Balaji 2013; Karimzadeh et al. 2013; Paradesi 2011).

Geoparsing has been discussed in numerous studies (Amitay et al. 2004; Ghahremanlou et al. 2014; Lieberman et al. 2010). This literature domain has identified two distinct steps: (1) *toponym recognition*, which entails identifying geographical names. and (2) *toponym resolution* which entails disambiguation of a toponym to assign it to a specific location (Leidner 2007; Lieberman et al. 2010).

For *toponym recognition*, the simplest approach is to extract single and consecutive words from a text and then match them to a comprehensive set of toponyms (i.e., geographical locations; Schulz et al. 2013). Such a pre-existing list of toponyms is known as a "gazetteer." This approach yields a list of candidate locations independent of the language used in the tweet. The use of a comprehensive gazetteer makes it likely that the algorithm will find locations mentioned in a tweet. Unfortunately, since many location names also have other meanings in normal language usage (e.g., "Darwin" is both a place name and a family name), the results also include many erroneous matches. In contrast, named-entity recognition (NER) analyzes (through natural language processing) the structure and grammar of the tweet's language (Al-Rfou et al. 2015; Van Erp et al. 2013). Employing NER can help to distinguish, for example, among similarly named places and persons (Amitay et al. 2004). These tools have mostly been developed and trained using more formal texts, such as newspapers (Sultanik and Fink 2012). Nonetheless, researchers have developed several NER approaches for Twitter (Dittrich 2016; Li et al. 2012; Van Erp et al. 2013), most of which are designed for English-language tweets. However, the short, error-prone, multi-lingual nature of tweets, along with that medium's frequent use of slang and abbreviations, has limited the applicability of NER (Li et al. 2012). Middleton et al. (2014) show that named entity matching (NEM) performs better than NER on tweets. This approach tokenizes tweets and matches these tokens first to places, then streets, and finally regions, while discarding matched tokens to avoid double matches.

The *toponym resolution* step is required, because many place names have multiple occurrences worldwide (e.g., Leidner 2007). Most studies have restricted their gazetteers to only include unambiguous place names with a relatively high population or assigned tweets to the candidate location with the highest population (Amitay et al. 2004). Unfortunately, both approaches introduce errors when an event occurs in a town with both a low population and a name shared with another location. These errors arise because either the town is not included in the limited gazetteer or a city with larger population takes precedence. For Twitter in particular, challenges persist regarding the automated geoparsing analysis of text and other metadata. For example, users rarely post an unambiguous name or a combination of a place and country name, mainly because of the limited length of a tweet (Sultanik and Fink 2012). Several studies have addressed this issue by using the tweet's metadata as additional spatial information, with examples including the user's hometown (Hecht et al. 2011), the relationships between users (Takhteyev et al. 2012), and mentions of super regions or nearby locations (Middleton and Krivcovs 2016). Unfortunately, in many cases, these additional spatial indicators are unavailable or unreliable. Therefore, Schulz et al. (2013) and Zhang and Gelernter (2014) analyzed several spatial indicators, such as the time zone, the user location field, and other textual clues, to obtain a more reliable estimate of a particular tweet's location. Their results revealed that tweet geoparsing outcomes can be improved using these methods, but only for those tweets with available spatial indicators. As additional spatial information is not always available, this approach cannot be easily applied to all tweets. Moreover, even when this data is available, it does not always match the location mentioned by the user.

The aim of this study is to develop a global geoparsing algorithm for tweets without assuming a priori knowledge about an event so that the algorithm can be employed for event detection. To geoparse tweets on a global scale without a local focus, additional spatial information from the tweets is required for disambiguation. Therefore, we develop a new toponym disambiguation system which builds upon the approach by Schulz et al. (2013). This new toponym-based algorithm for grouped geoparsing of social media (TAGGS) uses grouped geoparsing to reliably find a much larger percentage of locations compared to the standard approach of individual geoparsing. The TAGGS approach permits spatial information from related tweets to be incorporated in the analysis, allowing users to geoparse tweets with few or no spatial indicators of their own. TAGGS could form the basis for a global flood detection algorithm in future research.

In the remainder of this paper, we outline the development of the TAGGS and show its applications using approximately

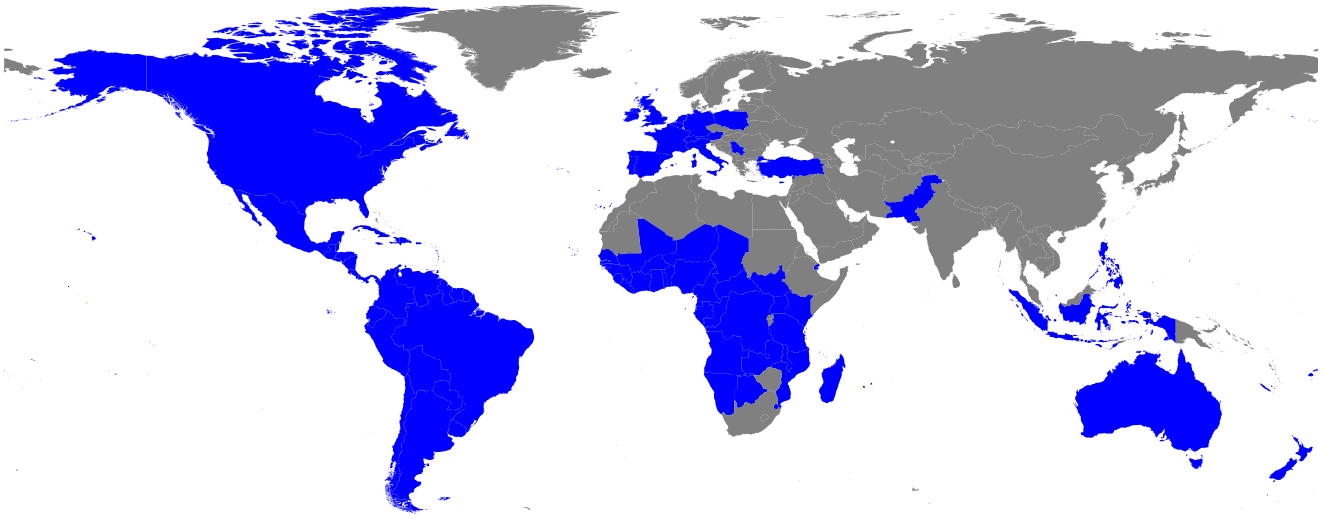

Fig. 1 Countries of which we included keywords in the first official language. University of Groningen Open Data. Retrieved Nov, 2017, from http://opendata.rug.nl/

3 years of globally sourced tweets with flood-related keywords, collected between July 29, 2014 and July 18, 2017, and validate the algorithm using a set of manually labeled tweets.

Methodology

The TAGGS algorithm uses geoparsing to match a tweet's location references to one or more geographic locations at a country, administrative subdivision or city, town and village-level. To that end, a database containing known geo-locations (a gazetteer) was used to match a tweet's text to one or more candidate locations (*toponym recognition*). Thereafter, additional spatial information obtained from both the tweet itself and related tweets was employed to determine the actual location(s) that the user mentioned in the tweet (*toponym resolution)*. The collection of the input dataset is described in the "Input Data" section. Afterwards, the process of geoparsing via toponym recognition and resolution is outlined in the "Geoparsing" section.

Input Data

The TAGGS algorithm uses three types of input data: a gazetteer, tweets collected using the Twitter API, and additional GIS-based geographical information. To build our gazetteer, we used the GeoNames database,[1] a geographical database containing over 4 million cities, towns, villages, and administrative divisions. The main dataset in GeoNames contains towns and villages, including their administrative parent area, geographical location, and population. Another dataset lists alternative names, like translations, slang terms, and abbreviations (e.g., for "New York," it includes, for example, *New York*, *The Big Apple*, *NY*, *Nueva York*), and the language of each alternative name. For an analysis of the accuracy of the GeoNames database, we refer to Ahlers (2013).

The tweets and their associated metadata (e.g., the user's hometown, the user's time zone, and the GPS coordinates of the device from which the tweet was sent) were collected in real time via the Twitter streaming API using a series of keywords in 12 major languages, covering a considerable part of the globe (see Fig. 1 and Table 1). We note that the languages

Table 1 Keywords related to floods and percentage of tweets per language over the period 2014–2017

Language	Keywords	Number of tweets per language (%)
English	Flood, floods, flooding, flooded, inundation, inundations, inundated	63.31
Indonesian	Banjir, banjirjkt, bantubanjir	24.71
Filipino	Baha, bumabaha, pagbaha	1.76
French	Inonder, inondation	0.53
German	Flut, hochwasser, Überflutung	0.28
Italian	Inondazione, inondacioni,alluvione	0.18
Dutch	Overstroming	0.03
Polish	Powódź, powodzie	0.01
Serbian	Poplava, poplave, поплава, поплаве	0.03
Portuguese	Inundação, inundacão, inundaçao, inundacaoinundações	0.45
Spanish	Inundación, inundacion, inundarinundaciones	8.65
Turkish	Su taşkın, su baskını, sel bastı, sel suyu, sel yüzünden, taşkın oldu, sel suyunun	0.06

[1] GeoNames. Retrieved August 1, 2017, from http://www.geonames.org

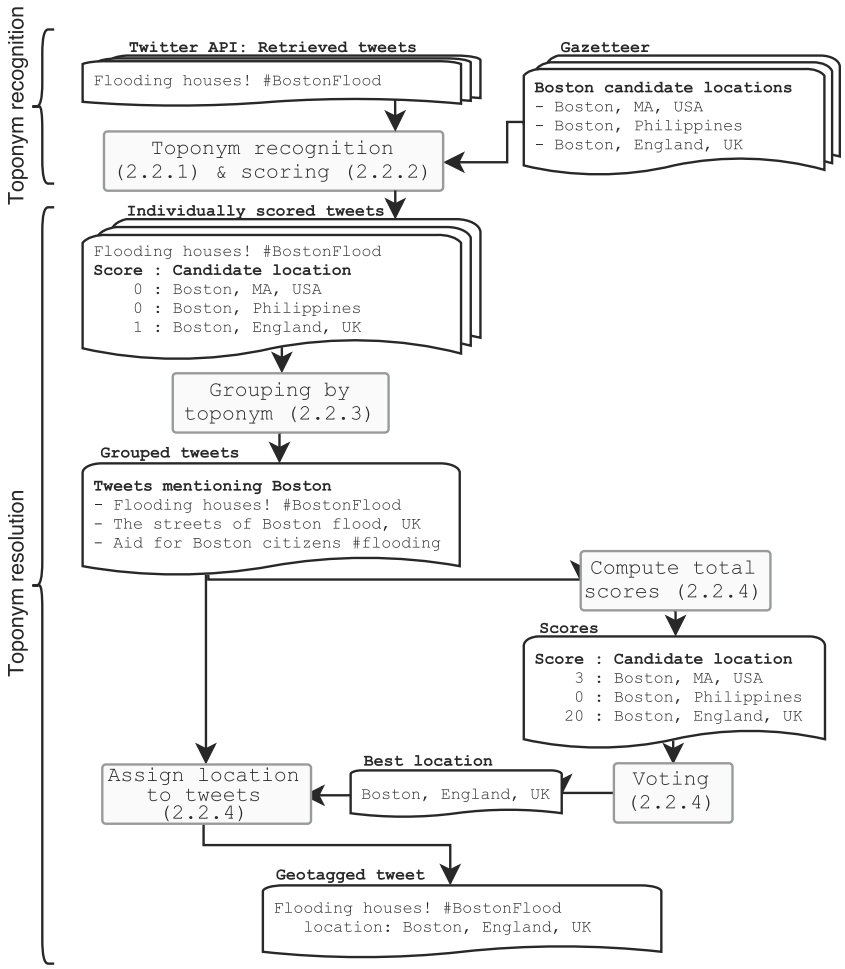

Fig. 2 Overview of the TAGGS geoparsing process

used are space-separated, because in a later step, we perform tokenization of the sentence utilizing the spaces in the sentences. To apply the model for non-space-separated languages, such as Mandarin and Japanese, other types of tokenization should be employed. We collected 55.1 million tweets, posted between July 29, 2014 and July 18, 2017. We used GIS shapefiles of the global time zones to match locations and country and administrative boundaries to time zones.[2] Finally, we analyzed a large corpus of Wikipedia articles to obtain lists of the 1000 most commonly used words per language. To avoid discarding commonly used toponyms, such as New York, from these lists, we omitted words that are used to refer to a location with a population greater than 100,000.

Geoparsing

Figure 2 describes the procedure followed by the new TAGGS algorithm. First, we collected tweets over a 24-h period. Each tweet from this timeframe was analyzed on an individual basis (the "Toponym Recognition" section) by matching its text to our gazetteer (toponym recognition). Next, each of the tweets' candidate locations was assigned a score indicating how well it matched the tweet's additional spatial information (the "Scoring" section). While previous approaches have relied on the spatial information of the individual tweet in question, we grouped all tweets according to the toponym (the "Grouping" section) identified during the toponym recognition step. Then, we computed the total score for each candidate location by summing the scores of the individual tweets and using a voting process to assign the best location (toponym resolution) to all tweets in the group (the "Voting and Assigning Locations" section; Fig. 3). In addition, a toponym resolution table was made to store the toponyms and their resolved geographic locations of the last time step. This table is later used to geoparse tweets in real time. Once locations had been assigned to the tweets, the same procedure was applied to a later scanning window (the "Iteration" section; Fig. 4), which included new incoming tweets. At that stage, tweets that are outside the scanning window were no longer considered. Meanwhile, new incoming tweets were immediately geoparsed using the toponym resolution table.

[2] Natural Earth. Retrieved March 1, 2017, from http://www.naturalearthdata.com/

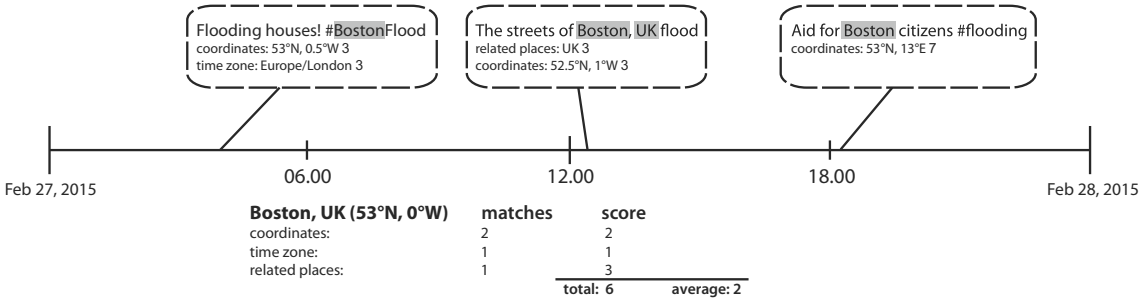

Fig. 3 In this example, three tweets mention Boston within 24 h. The metadata of the first and second tweet match Boston, UK, while the third one did not have any matching spatial indicators. Using the spatial information for the first two tweets makes it possible to correctly assign the third tweet to Boston, UK. In this example, the total score for Boston, UK, is 6 and the average score is 2

Toponym Recognition

To identify candidate locations for a tweet, a tweet's text was matched to the gazetteer. Tweets are often written in informal language and contain content unnecessary for geoparsing. Therefore, we applied Dittrich's (2016) approach to delete URLs and punctuation, split words with medial capitals (camelCase) or underscores, and convert all text to lowercase. Then, all contiguous sequences of one and two words were extracted from the text ("uni- and bi-grams"). Subsequently, similar to Schulz et al. (2013), we looked up all uni- and bi-grams in our gazetteer, and the result was a list of potential candidate locations for each toponym mentioned in the text. We then further filtered the results to obtain the candidate locations, using the following approach:

1. All uni- and bi-grams among the 1000 most common words in a tweet's language are discarded.
2. Locations are frequently referenced to using alternative names in other languages (e.g., New York is Nueva York in Spanish). If a tweet referenced a location via an alternative name, we only considered that tweet if it uses the same language as that alternative name (e.g., an alternative name for "Cameroon" is "Cameron," meaning that all mentions of Cameron, the UK's former prime minister, would otherwise be located in the "Cameroon").
3. We consider small towns, with a population of at least 1, if the town mentioned in the text is written with a capital letter. For tweets written in German, we discard all locations with population lower than 5000 because, in this language, all nouns are capitalized.
4. Next, all locations with names that were part of another location's name are discarded. For example, a tweet containing New York matches both York and New York, although the user was clearly referring to New York. Therefore, we excluded York.
5. If a tweet mentioned a location with a name simultaneously used for a province or country and an identically named town within that area (e.g., the city of New York is within the state of New York), the location with the most translations provided in the gazetteer—a criterion that we used as a proxy for importance—took precedence (i.e., the city of New York took precedence over the state).

Scoring

For each tweet for which we found one or more candidate locations, as described in the "Toponym Recognition" section, its additional spatial indicators were matched to each candidate location. We use these indicators as contextual clues to provide additional information for toponym disambiguation.

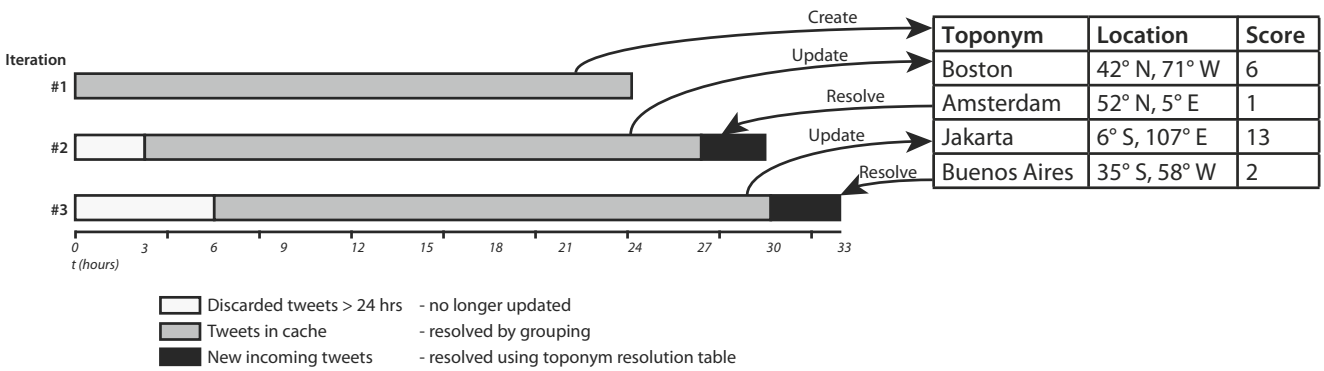

Fig. 4 Schematic overview of the geoparsing process. As indicated, 24 h worth of cached tweets were geoparsed by grouping and analyzing their additional spatial information. Moreover, this step entailed creating or updating the toponym resolution table used to geoparse subsequent incoming tweets. Tweets that fall outside of the scanning window were excluded from this process

A higher score indicates a higher confidence that a specific candidate location is correct. The bulleted list below describes the matching process for each of the spatial indicators (metadata and textual).

- UTC offset (metadata): Twitter's time zone field, available on the user's profile page, signifies an area with a uniform time standard. Twitter initially sets users' time zones, but users can manually adjust this setting. If this field was set, the Coordinated Universal Time (UTC) offset was available for each of the user's tweets and was converted to a list of time zones matching the UTC offset. Our gazetteer contained a list of time zones for each location, used to match these to the found time zones.
- Coordinate-based indicators (metadata): We extracted geographical coordinates for two spatial indicators (see below). We considered the coordinates extracted from these indicators a match for a candidate location if they were located within 200 km of each other or, for administrative areas, if the coordinates were within the same country as the candidate location.

  – User hometown: Users can specify their hometowns in their user profile. In doing so, users receive assistance from a dynamic menu of location options that appears when they start typing in the Twitter text field. Although the box can be ignored, most users do make use of it. This means that in most cases, the location field is either (1) a town and country name separated by a comma or (2) a country name. However, many variations are possible, including fantasy places (Schulz et al. 2013), multiple locations, and incomplete data entries (e.g., a user who lives in Washington, D.C. might simply enter "Washington" in the location field). We searched for both the town and country in the gazetteer to create a list of candidate towns within the specified country. If no comma was present, we looked up the entire field in the gazetteer.
  – Location: When a tweet is sent from a GPS-enabled device, and when the user's privacy settings or manual adjustments assign a location to the tweet, that user's location at the time of posting is attached to the tweet. Additionally, the user can attach a geographic entity to a tweet, by manually selecting it from a dynamic list.

- Mentions of related places (textual): We matched user mentions of other locations higher (geographical parent) or lower (geographical child) in the hierarchy (e.g., "Los Angeles" is the geographical child of "California," which is the geographical child of "the USA"), other towns within 200 km of a candidate place name, and other administrative areas within the same geographical parent (e.g., Serdyukov et al. 2009; Amitay et al. 2004).

Next, we used a scoring system to indicate the likelihood of a match between the location referenced in the tweet and each candidate location. An overview of the scores for each of the five spatial indicators is provided in Table 2. These scores were summed to obtain the total score (maximum of 7), which indicated the likelihood of a match.

Grouping

We assumed that multiple tweets that mentioned the same toponym within a given timeframe referred to the same location. For example, if a flood occurred in Boston, UK, we expected that all users mentioning "*flood*" and "*Boston*" were referencing Boston, UK, rather than Boston, Massachusetts, USA. All tweets mentioning the same toponym were then grouped together. Thus, the greater the number of tweets mentioning a location, the larger was the associated group—and therefore, the higher the probability of metadata being available for that group. Since tweets could contain multiple toponyms, individual tweets could belong to more than one group.

Voting and Assigning Locations

In this step, for each group, the total score of each candidate location was computed by averaging the scores for the candidate locations of the individual tweets (Fig. 3). Hence, if only few tweets (a small group) mention a location, the score is more likely to fluctuate compared to a larger group of tweets. If multiple tweets originated from the same user and thus had the same metadata, only the most recent tweet was considered. In addition, because the mentions of related places rely on textual clues, and since users frequently copy each other's tweets, we only considered the oldest tweet for clusters of similar tweets. To that end, we created word vectors for the tweets within a group and then compared those vectors. If the vectors were similar, we eliminated the newest tweet. For further details on this approach, refer to Hürriyetoğlu et al. 2016.

Finally, we assigned the location with the highest score to all tweets in the group if that tweet's referenced toponym was the official name of the location or if the tweet's language matched the toponym language. If multiple locations had an

Table 2 Score for each of the spatial indicators assigned to individual tweets

Indicator	Score
UTC offset	0.5
User home town	1
Coordinates	2
Mentions of related places	3

The scores are in the order of magnitude found by Schulz et al. (2013)

equally high score, we assumed that the correct location was the candidate with the highest population.

Moreover, it was also possible to discard potential locations for which the average score was below a certain threshold. We do this only for countries, because a country name is often used by people outside a country and thus spatial indicators of the post are less likely to match the mentioned country. In contrast, local information (i.e., administrative subdivisions, cities, towns, and villages) was more likely to be provided by locals and thus the post's spatial indicators are more likely to match the geographical entity. When no minimum score (i.e., a minimum score of 0) was set, a large number of tweets was assigned to incorrect locations, due to a lack of matching metadata (e.g., numerous tweets were assigned to the city of "Mobile" in Alabama, USA). By increasing the threshold to, for example, 0.2, groups with little to no metadata matching any of the candidate locations were discarded. This meant that the recall decreased (i.e., fewer tweets were assigned locations), while the precision of the algorithm increased. Introducing a higher threshold, such as 1.0, could improve precision even further, but would have also meant discarding a much higher percentage of tweets. Therefore, we decided to initially set a 0.2 threshold and to perform a sensitivity analysis (the "Validation of TAGGS" section).

In addition, the toponyms and their respective resolved locations were saved in a *toponym resolution* table. That table indicated the location with the highest score per toponym, and therefore, the location most likely for a future tweet to reference. This toponym resolution table was continuously updated and used to geoparse new incoming tweets.

Iteration

To continuously geoparse tweets, we used an iterative process (Fig. 4). After the geoparsing of all tweets within the scanning window was finished, the window was shifted by 6 h. All new tweets were retrieved from the tweet database and separately analyzed for toponyms and respective spatial indicators (the "Toponym Recognition" and "Scoring" sections), while tweets that fall outside of the scanning window were discarded. The locations mentioned in the tweets within the scanning window were, again, grouped (the "Grouping" section) and resolved (the "Voting and Assigning Locations" section) and used to update the *toponym resolution table*. As the tweet geoparsing process included information from other tweets (including locations referred to in future tweets), it was possible for a tweet's location and respective score to change. In such cases, we updated the database accordingly. Such alterations only occurred when in a subsequent iteration we found a higher score for a specific location or identified another location with a higher score (but the same toponym).

In addition, when the first iteration was completed, another process analyzing incoming tweets in real time was initiated. Using the procedure described in the "Input Data" section, the text of the tweets was processed, and its uni- and bi-grams are matched with the *toponym resolution table*. This resulted in an initial guess regarding the locations mentioned in each tweet.

Results

Application of TAGGS

First, we applied TAGGS on the 55.1 million tweets in a historical dataset, applying the algorithm as if the data were available in real time, shifting the scanning window by 6 h in each step. We first discuss the results obtained using baseline settings. For this, a 24-h scanning window and a threshold of 0.2 were used, which causes all locations found in tweets that scored below the threshold (the "Voting and Assigning Locations" section) to be discarded. The results for the baseline settings are summarized in Table 3. Next, we discuss the results of a sensitivity analysis for the threshold value and the size of the scanning window (the "Sensitivity Analysis" section).

Of the 55.1 million tweets, we found that 19.2 million mentioned at least one location, and 3.4 million tweets referenced multiple locations. In addition, when distinguishing between administrative levels, roughly half of the locations mentioned refer to a city, town, or village, while country and the lower administrative level locations each account for a quarter of the mentions.

To gain insight into the geoparsed tweets, those countries covered by the algorithm (Fig. 1) with a population of at least 10 million people were grouped according to economic development. For that purpose, we employed the income groups defined by the World Bank.[3] For each group, the number of geoparsed tweets between August 2014 and December 2016 was plotted (Fig. 5) against the total flood losses over this period, as described in the Munich Re's NatCatSERVICE on a purchasing power parity (PPP) basis.[4] This gives an impression of how Twitter reporting relates to flood impacts. The data made clear that in high-income (green) countries, there were about one to two orders of magnitude more tweets than in low-income (red) countries. The number of tweets in middle-income (blue and orange) countries fell between the other two groups, with a particularly large spread in the lower-middle-income (orange) countries. Notably, these numbers likely reflect a size effect, as Indonesia (IDN) and Pakistan (PAK), which had the highest number of tweets within the lower-middle-income group, also have large populations.

[3] World Bank Knowledge base. Retrieved May 1, 2017, from https://datahelpdesk.worldbank.org/knowledgebase

[4] World Bank Open Data. Retrieved May 1, 2017, from http://data.worldbank.org/

Table 3 Results of the automated geoparsing of 58.9 million tweets (using the base settings; see the "Application of TAGGS" section)

	Number of tweets
Total	55.1 million
One or more location	19.2 million
Multiple locations	3.4 million
Country level	4.6 million
Lower administrative level	5.1 million
City, town, village, etc.	11.2 million

However, the results underscored that relatively small countries, such as the Philippines (PHL) and Venezuela (VEN), generated a significant number of (geoparsed) flood tweets within their respective groups. These findings suggest that flood events, and not just the size of the population or the Twitter user base, are responsible for the high number of tweets during the investigated time period.

The plots also illustrate that in general, more flood tweets seemed to be linked to higher levels of flood damage over the study period, as the points roughly go from the bottom left-hand corner to the top right-hand corner of the diagrams. This relation is influenced by many other factors, including (but not limited to) variations in the extent of Twitter usage per country, language use per country, and keyword selection, and is therefore by no means strong enough to have any predictive power after regression analysis. That said, the existence of this relationship was in line with expectations. Namely, in countries that suffered from disastrous flood events that caused significant damage, a substantial number of tweets about flooding were generated. This illustrates that the algorithm seemed to be successful in capturing flood events around the globe.

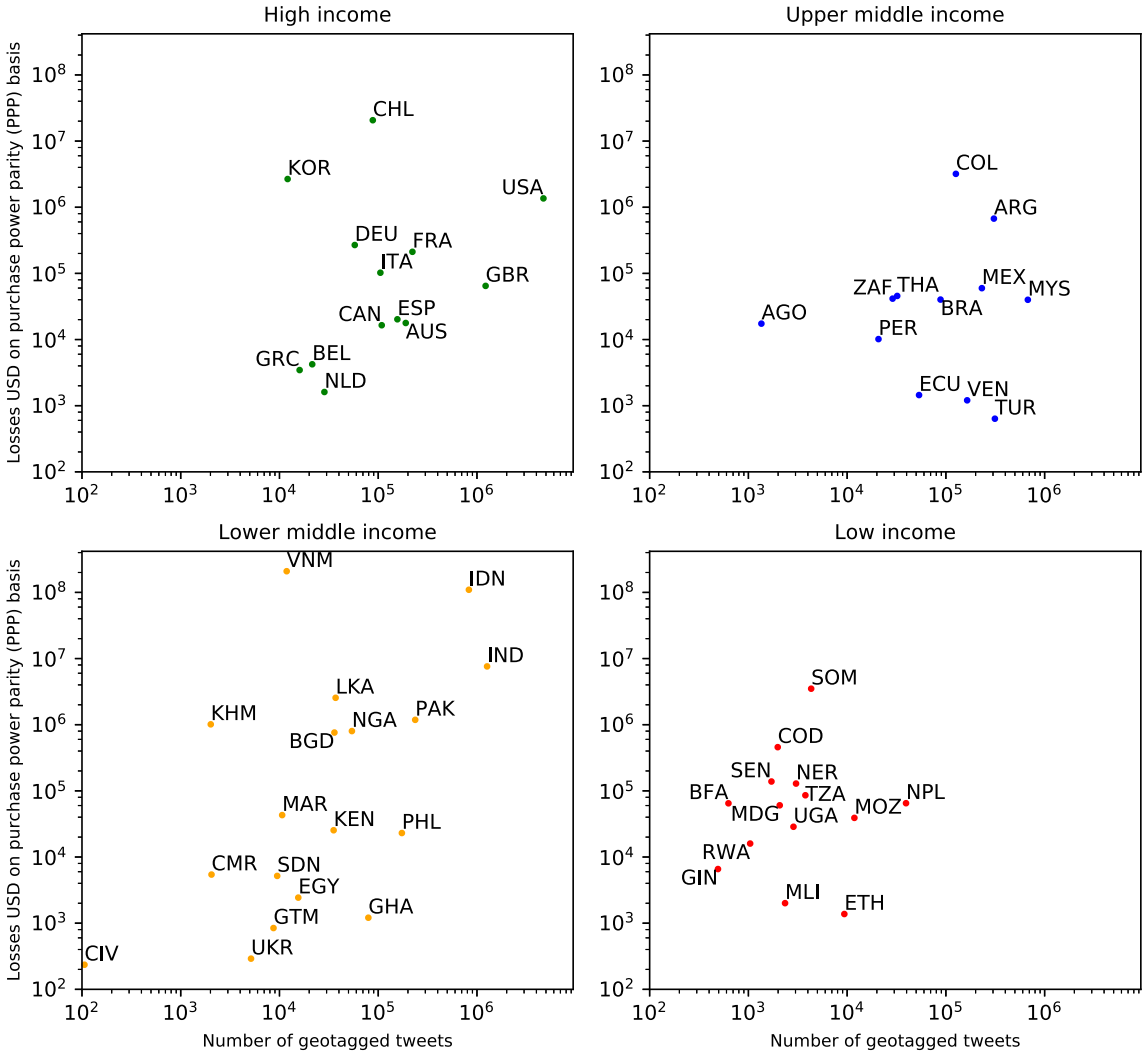

Fig. 5 The number of geoparsed tweets relative to losses due to flood events between July 29, 2014 and December 31, 2016 for four country income groups

Validation of TAGGS

To properly validate TAGGS, we defined a golden standard with *manually* tagged tweets. To the best of our knowledge, no other study provides a global dataset focusing on a specific event type. Therefore, we compile a random dataset using 2785 flood-related tweets from two separate days and manually assign locations to the tweets.

- Dec 12, 2015: To check if our model properly for small flood events in multiple languages, we selected a day during which multiple such events occurred across the globe, including in Indonesia, India, Kenya, Congo, Norway, the UK, Canada, and Paraguay (1282 tweets).
- Dec 27, 2015: When the number of tweets that mentioned a specific location is higher, the probability of sufficient metadata being available is also higher. Therefore, we validated our algorithm on a date with multiple large events. On the date in question, several major floods received global news coverage, including floods in the USA, the UK, and Argentina (1503 tweets).

Each tweet can be labeled with one, multiple, or no locations at all. We recognized all mentions of locations on the different administrative levels that we apply the algorithm to (i.e., country, administrative subdivisions and cities, towns, and villages), including abbreviations, shorter versions, and slang, but excluded possessive pronouns (e.g., the Irish weather) and mentions of geographical features within towns and other geographical features, such as valleys and rivers. We do include location mentions when they are combined with other words (e.g., #leedsfloods) but exclude any information in the Twitter handles (e.g., @PakistanToday) because these locations are not necessarily related to the location of a possible event.

Using the manual approach, of the 2785 total tweets in our validation set, we found 2079 references to countries, administrative subdivisions and cities, towns, and villages in 1497 tweets. Then, we compared the manually labeled tweets to both the automated individual and automated grouped geoparsing (TAGGS) approaches. For individual geoparsing, we use the location metadata but did not consider other tweets mentioning the same geographical entities, similar to Schulz et al. (2013).

Trade-Off between Recall and Precision

With geoparsing algorithms, there is a trade-off between the number of tweets that are parsed (recall) and the number of correctly parsed tweets (precision; Leidner 2007). Precision measures the number of correctly geoparsed tweets relative to the total number of geoparsed tweets. Hence, precision markers do not provide an indication of the total number of tweets within a location. Recall measures reflect the number of correctly geoparsed tweets relative to the total number of tweets with a spatial reference. In essence, the greater the level of precision (i.e., the smaller the number of incorrect tags), the smaller the total number of geoparsed tweets. Inversely, if one wants to geoparse more tweets (higher recall), the number of errors within the geoparsed tweets (in terms of incorrect location assignments) will also increase (lower precision).

Sensitivity Analysis

In the following sensitivity analysis, we show two series of plots (Figs. 6 and 7) delineating both individual (red) and grouped (blue) geoparsing for various model settings, namely, a varying threshold and a varying size of the scanning window. In these figures, we show three plots: (1) a plot that shows recall and precision measures for all locations that the model accounts for (i.e., countries, administrative subdivision and cities, towns, and villages), using all 2785 tweets; (2) a plot that shows these measures for administrative subdivisions, using only those tweets that mention such a location according to our validation set; and (3) a plot that shows precision and recall measures for all cities, towns, and villages, using only those tweets that mention such a location.

Figure 6 shows the recall and precision scores for individual and grouped geoparsing with a varying threshold. The trade-off between precision and recall is visible in the first window: When a higher threshold is chosen, more location matches are discarded, while the likelihood of a correct match is higher for the residual locations. For individual geoparsing, as only the spatial indicators of the post itself are considered, the scores behave discreet. In contrast, for grouped geoparsing, the scores are averaged between tweets within the same group, and therefore the decrease is more gradual. At very high thresholds, the precision for grouped geoparsing starts to drop (for administrative subdivisions and cities/town/villages). This is likely because the scores assigned to tweets in small groups fluctuate more than for large groups (the "Voting and Assigning Locations" section) and hence there is more uncertainty in the location being assigned correctly. Therefore, when the threshold increases, small groups have a larger share in the response set (as large groups will always have averaged medium scores) which causes the precision to drop. Approximately between a threshold of 0.1 and 0.25, precision and recall measures for grouped geoparsing are optimal and higher than using any other threshold for individual geoparsing.

Figure 7 shows the recall and precision measures for a varying scanning window size, ranging between 6 min and 48 h. In theory, when using an infinitesimally small scanning window for grouped geoparsing, the results would be identical to the individual geoparsing. It is clearly visible that, in general, both precision and recall increase when the size of the scanning window is larger. This is expected, because a larger number of tweets are grouped, and therefore, the likelihood that spatial

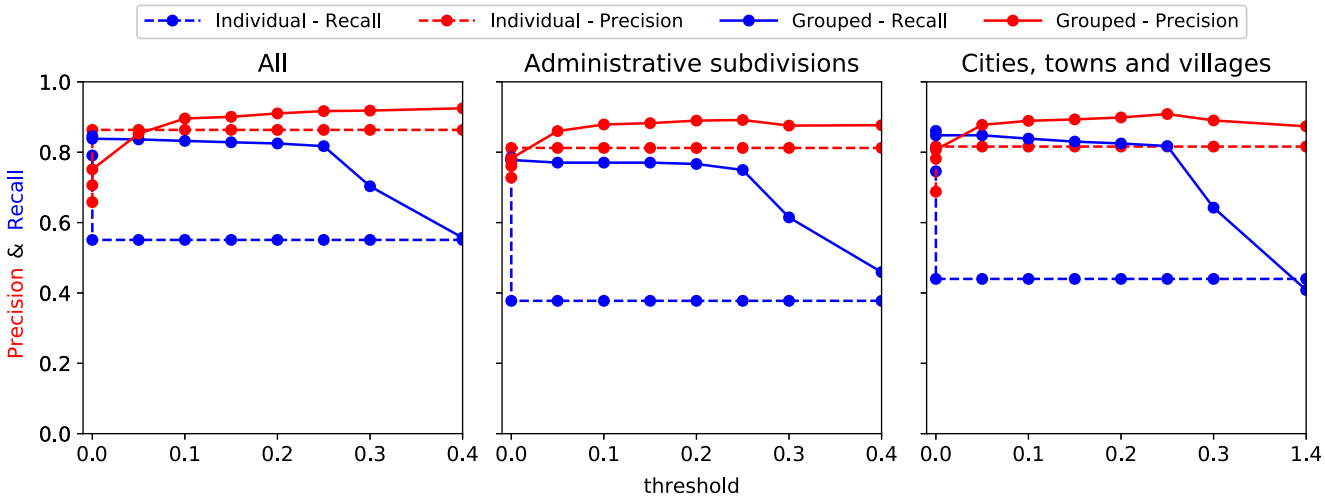

Fig. 6 Recall and precision scores for individual and grouped geoparsing with a varying threshold

information is available increases. Although an increase of recall and precision is still visible for a larger scanning window, the increase is not substantial, which indicates that spatial information is available for most toponyms. When new floods occur, it is not feasible to take location mentions of previous floods into account. Therefore, we hypothesize that when the scanning window becomes too large, the performance of the model will be lower. Unfortunately, because of memory (RAM) constraints in our current setup, we cannot test this. Ideally, the size of the scanning window depends on the volatility of the event type, where events with a longer average duration (people will likely refer to the same event over a longer time span), such as droughts, could benefit from a larger scanning window and vice versa for shorter events.

Effect of the Event Size

Figure 8 highlights differences in performance due to different flooding circumstances using a varying threshold. On December 12, 2015, there were various smaller flood events, while on December 27, 2015, a couple of very large flood events took place (the "Validation of TAGGS" section). These two cases make clear that using optimal model settings, TAGGS was slightly more accurate for larger-scale flood events than smaller-scale flood events. Such a finding is to be expected, because during the large flood events in the USA and UK, a larger percentage of tweets mentioned the same toponym, due to a high level of Twitter usage in both countries. The grouping approach meant that most of these tweets were scored, even though not all of them had spatial information available. In contrast, when a location is mentioned in a single or small group of tweets without location metadata, this tweet was not geoparsed. This latter situation is more common when a higher number of smaller events occur, as was the case on Dec 12, 2015. Similar to the large groups of tweets' drop-in precision at a lower threshold compared to small groups' drop-in precision (the

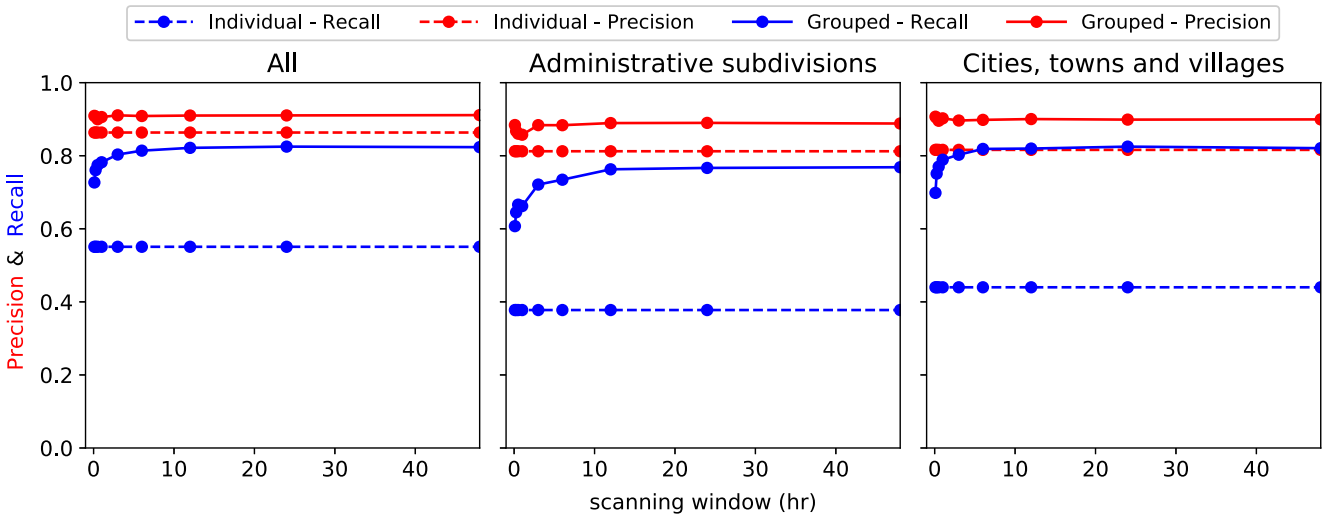

Fig. 7 Recall and precision scores for individual and grouped geoparsing with a varying scanning window size

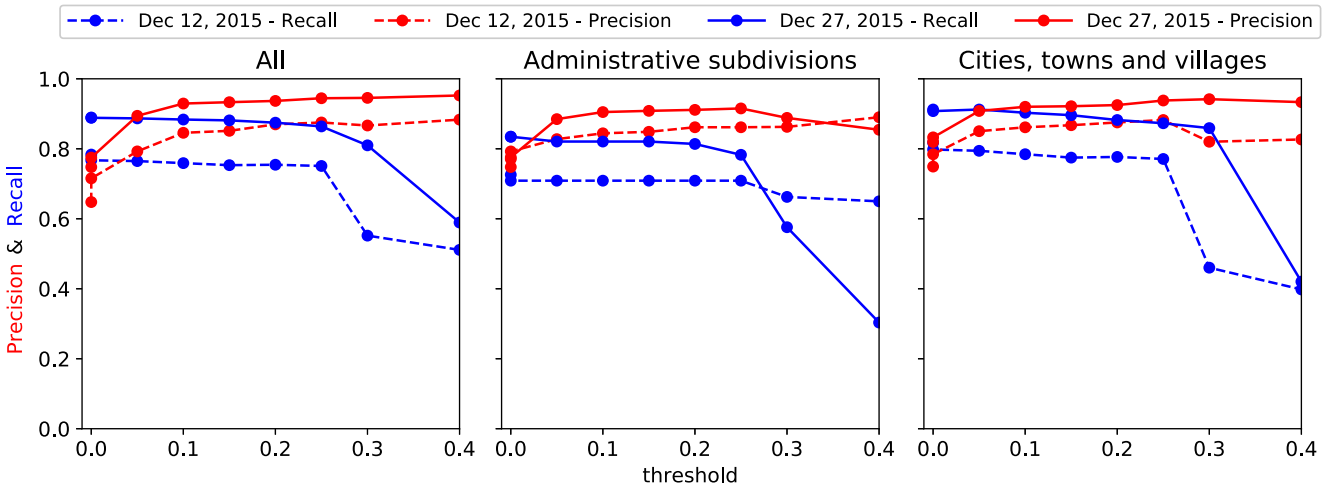

Fig. 8 Recall and precision scores for grouped geoparsing on Dec. 12, 2015 and Dec. 27, 2015

"Sensitivity Analysis" section), the precision of Dec 27, 2015 also declines at a lower threshold compared to the precision of Dec 12, 2015. We argue that the more drastic drop in precision for the tweets posted on Dec 27, 2015 is because most groups of tweets are larger and therefore all tweets have a relatively low score, which are then discarded at a higher threshold. Nevertheless, the grouped algorithm still correctly geotagged about two thirds of the tweets with a location, even on days with predominantly smaller flood events.

	Number of locations	
	Total	*Correct (%)*
Manual		100
Coordinates	33	100
Individual geoparsing	1148	84
Grouped geoparsing	1718	90

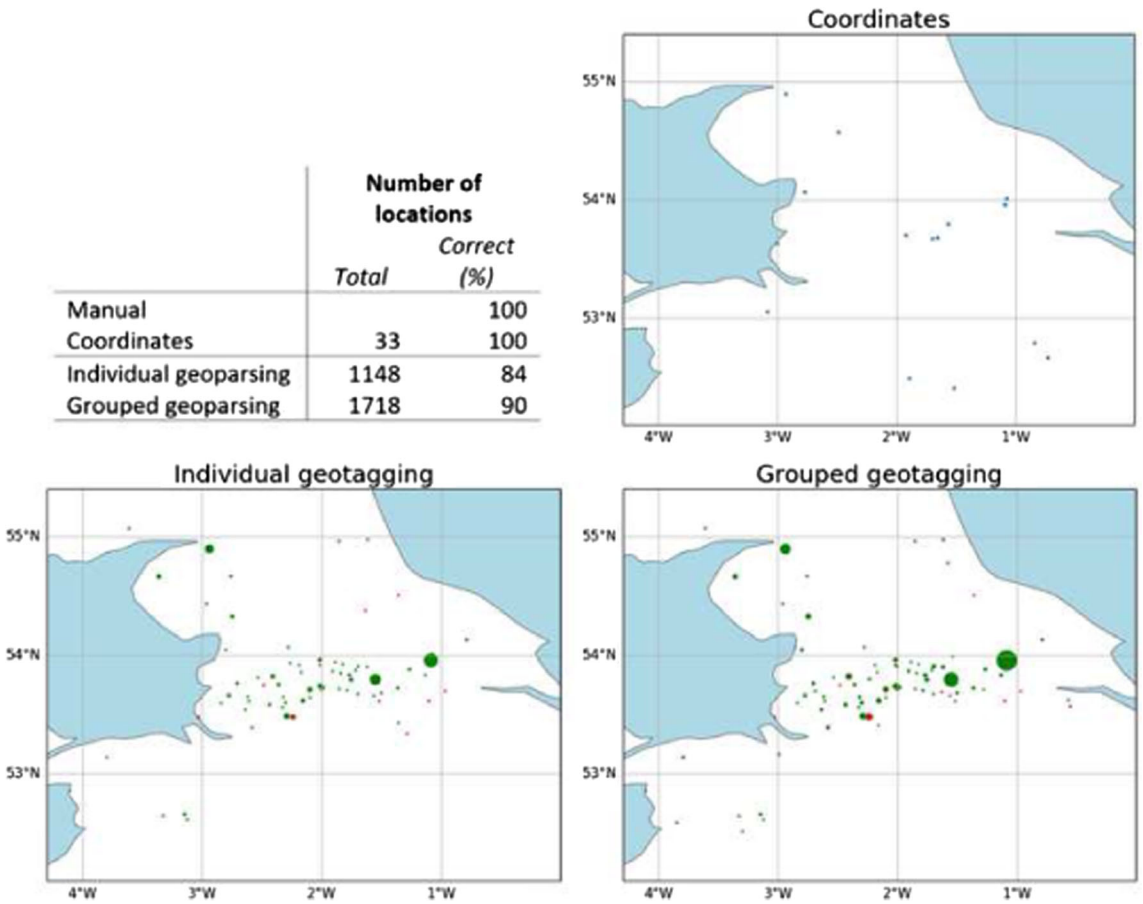

Fig. 9 Comparison of the number of geo-located tweets in the validation set in the middle of the UK for various geo-location methods. The green dots represent correctly identified locations, and the red dots represent incorrectly identified locations

Table 4 Comparison of precision and recall scores for the TAGGS and Carnegie Mellon geolocator 2 algorithms

	TAGGS (precision/recall)	Carnegie Mellon geolocator 2 (precision/recall)
All	92/84	42/47
First-order administrative subdivision	92/86	51/48
Towns	92/83	44/41

Comparison to Other Spatial Indicators

Figure 9 illustrates the number of locations identified using the different approaches and the number of erroneous matches for the base settings. Using individual geoparsing, we found approximately 55% of these locations—of which roughly 86% were correct. The grouped geoparsing technique, developed for this research, increased the number of found locations to approximately 82%—of which about 91% are correct. In contrast, of the 2785 tweets, only 33 (~1.2%) have coordinate information attached. This suggests that the TAGGS approach makes significantly more spatial information available than does a strategy relying on either individual geoparsing or coordinates alone.

Comparison to Other Work

Several other studies have addressed similar problems as this paper. For example, Middleton et al. (2014) and Gelernter and Balaji (2013) investigated geoparsing for crisis mapping in a local setting, assuming a priori knowledge about an event. This allowed the authors to collect detailed information from the focus area of the event, which is unfortunately not possible for our approach. Zhang and Gelernter (2014) developed the Carnegie Mellon geolocator 2 algorithm. We analyzed the performance of these algorithms using the English tweets in our validation set. As shown in Table 4, TAGGS performs considerably better for both precision and recall.

Concluding Remarks and Outlook

In this paper, we presented TAGGS, a multi-lingual algorithm that groups topologically related tweets based on their referenced toponyms and then geoparses those tweets using the mutual spatial information of the entire group. In addition, the algorithm successfully differentiates between various administrative levels.

Studies on event detection often work with geo-located tweets by using the coordinates attached to them. In our validation set, however, only 2% of all tweets had coordinates attached. By geoparsing tweets using the tweet content, this study roughly doubles the number of correctly geoparsed administrative subdivisions and cities, towns, and villages (using 0.2 threshold) compared to individual geoparsing. Moreover, using the grouping approach developed in TAGGS also boosted the precision level without lowering the number of geoparsed tweets (i.e., lowering recall) to an unacceptable degree. As a result, applying optimal model settings, recall is approximately 0.82 and precision 0.91 (F1-score 0.865), which means that approximately 74.6% of mentioned locations are both found and correct, while only ~10% of locations are incorrect. We note that these scores could vary for different event types, especially depending on the total number of location mentions relative to the total number of tweets.

Unfortunately, our algorithm also introduced several minor problems: (1) Using the individual geoparsing approach, a tweet is only parsed in a location if the metadata matches that location. When a tweet mentions a location with a toponym that is also frequently used in normal speech, all tweets mentioning this word can be localized to that location, rather than only those tweets that used the word as a toponym. An example of this is "turkey," a term that can refer to both the country of Turkey and the bird of the same name. (2) In rare cases, when a flood occurs in two different places with identical place names, all tweets are put into one group and hence tagged in only one of these locations. (3) Tweets often mentioned areas (e.g., the East Coast), rivers, and airports. Although the algorithm can resolve such locations using metadata, many such areas have not been included in this study's gazetteer. Including these entities in the gazetteer could improve the recall of the algorithm.

The TAGGS algorithm can form the basis for a flood detection algorithm to detect sudden changes in the number of flood tweets. Moreover, while this paper focused on geoparsing tweets, the approach outlined within this paper can be further developed and, for example, be combined with other types of mass data, such as newspapers and other social media platforms, to yield even more geoparsed information.

In future work, we aim to continue improving our algorithm. Currently, using the approach described in this paper, we only parse each tweet using the spatial information from that tweet itself and from other tweets mentioning the same toponym. In future research, we plan to expand on this approach by detecting sudden changes in the number of mentioned locations in an area. This technique would allow us to improve the geoparsing algorithm by considering sudden increases in mentions of nearby locations, using such a peak as an additional spatial indicator. Other improvements could be made by taking into account additional context, such as entity co-occurrence (Hu et al. 2014; Ju et al. 2016) or the geography of Twitter networks (Takhteyev et al. 2012).

Funding Information The research leading to these results has received funding from Netherlands Organization for Scientific Research (NWO) VICI (grant number 453-14-006) and the European Community's Seventh Framework Programme (FP7) ENHANCE (grant number 308438).

Data Availability The datasets analyzed and generated during the current study are not publicly available due to Twitter's privacy policy but are available from the corresponding author upon a reasonable request in line with the policy.

Code Availability The algorithm uses Python 3 scripts and connects to an Elasticsearch and PostgreSQL (with PostGIS) database. All code is publicly available on GitHub (https://github.com/jensdebruijn/TAGGS) as well as Zenodo (https://doi.org/10.5281/zenodo.1069678).

Compliance with Ethical Standards This study is in full compliance with all applicable ethical standards.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval Obtaining any specific ethics approval for this research was not required under VU University Amsterdam's policies and its ethics review board. Hence, no ethics approval was sought for this study.

Informed Consent This study does not involve any subject participation and therefore informed consent is not applicable.

References

Ahlers D (2013) Assessment of the accuracy of GeoNames gazetteer data. Proceedings of the 7th workshop on geographic information retrieval - GIR '13, pp 74–81. https://doi.org/10.1145/2533888.2533938

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 (pp. 586–594). https://doi.org/10.1137/1.9781611974010.66

Amitay E, Har'El N, Sivan R, & Soffer A (2004) Web-a-where. Proceedings of the 27th annual international conference on research and development in information retrieval—SIGIR '04, 273–280. https://doi.org/10.1145/1008992.1009040

Carley KM, Malik M, Landwehr PM, Pfeffer J, Kowalchuck M (2016) Crowd sourcing disaster management: the complex nature of Twitter usage in Padang Indonesia. Saf Sci 90:48–61. https://doi.org/10.1016/j.ssci.2016.04.002

Crooks A, Croitoru A, Stefanidis A, Radzikowski J (2013) #Earthquake: Twitter as a distributed sensor system. Trans GIS 17(1):124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

de Perez EC, Monasso F, van Aalst M, Suarez P (2014) Science to prevent disasters. Nat Geosci 7(2):78–79. https://doi.org/10.1038/ngeo2081

Dittrich A (2016) Real-time event analysis and spatial information extraction from text using social media data. In: Doctoral dissertation, Dissertation, Karlsruhe, Karlsruher Institut für Technologie (KIT). Photogrammetrie und Fernerkundung, Institut für. https://doi.org/10.5445/IR/1000057058

Fohringer J, Dransch D, Kreibich H, Schröter K (2015) Social media as an information source for rapid flood inundation mapping. Nat Hazards Earth Syst Sci 15(12):2725–2738. https://doi.org/10.5194/nhess-15-2725-2015

Gao H, Barbier G, Goolsby R (2011) Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell Syst 26(3):10–14. https://doi.org/10.1109/MIS.2011.52

Gelernter J, Balaji S (2013) An algorithm for local geoparsing of microtext. GeoInformatica 17(4):635–667. https://doi.org/10.1007/s10707-012-0173-8

Ghahremanlou L, Sherchan W, Thom JA (2014) Geotagging twitter messages in crisis management. Comput J 58(9):1937–1954. https://doi.org/10.1093/comjnl/bxu034

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, 237–246. https://doi.org/10.1145/1978942.1978976

Hu Y, Janowicz K, & Prasad S (2014) Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia. Proceedings of the 8th Workshop on Geographic Information Retrieval, 8:1–8:8. https://doi.org/10.1145/2675354.2675356

Hürriyetoğlu A, Gudehus C, Oostdijk N, & van den Bosch A (2016) Relevancer: finding and labeling relevant information in tweet collections. In international conference on social informatics (pp. 210–224). Springer, Cham. https://doi.org/10.1007/978-3-319-47874-6_15

Jongman B, Wagemaker J, Romero B, de Perez E (2015) Early flood detection for rapid humanitarian response: harnessing near real-time satellite and twitter signals. ISPRS International Journal of Geo-Information 4(4):2246–2266. https://doi.org/10.3390/ijgi4042246

Ju Y, Adams B, Janowicz K, Hu Y, Yan B, & McKenzie G (2016) Things and strings: improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19–23, 2016, proceedings 20 (pp. 353–367)

Karimzadeh M, Huang W, Banerjee S, Wallgrün JO, Hardisty F, Pezanowski S, … MacEachren AM (2013) GeoTxt: a web API to leverage place references in text. In Proceedings of the 7th Workshop on Geographic Information Retrieval—GIR '13 (pp. 72–73). ACM https://doi.org/10.1145/2533888.2533942

Kordopatis-Zilos G, Papadopoulos S, Kompatsiaris Y (2015) Geotagging social media content with a refined language modelling approach. https://doi.org/10.1007/978-3-319-18455-5

Lee K, Ganti R, Srivatsa M, & Mohapatra P (2013) Spatio-temporal provenance: identifying location information from unstructured text. In 2013 I.E. International Conference on Pervasive Computing and Communications Workshops (PERCOM workshops) (pp. 499–504). IEEE. https://doi.org/10.1109/PerComW.2013.6529548

Leidner JL (2007) Toponym resolution in text. ACM. SIGIR Forum 41(2):124. https://doi.org/10.1145/1328964.1328989

Li C, Weng J, He Q, Yao Y, Datta A, Sun A, & Lee B.-S. (2012) TwiNER: named entity recognition in targeted twitter stream. In Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval—SIGIR '12 (pp. 721–730). ACM. https://doi.org/10.1145/2348283.2348380

Lieberman MD, Samet H, & Sankaranarayanan J (2010) Geotagging with local lexicons to build indexes for textually-specified spatial data. Proceedings—international conference on data engineering, (May 2009), 201–212. https://doi.org/10.1109/ICDE.2010.5447903

Messner F, Meyer V (2006) Flood damage, vulnerability and risk perception—challenges for flood damage research. Flood risk management: hazards, vulnerability and mitigation measures. NATO Science Series 67:149–167. https://doi.org/10.1007/978-1-4020-4598-1_13

Middleton SE, Krivcovs V (2016) Geoparsing and Geosemantics for social media: spatiotemporal grounding of content propagating rumors to support trust and veracity analysis during breaking news. ACM Trans Inf Syst 34(3):16:1–16:26. https://doi.org/10.1145/2842604

Middleton SE, Middleton L, Modafferi S (2014) Real-time crisis mapping of natural disasters using social media. IEEE Intell Syst 29(2):9–17. https://doi.org/10.1109/MIS.2013.126

Paradesi SM (2011) Geotagging tweets using their content. Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 355–356. Retrieved from https://aaai.org/ocs/index.php/FLAIRS/FLAIRS11/paper/view/2617

Penning-rowsell E, Johnson C, Tunstall S, Tapsell S, Morris J, Chatterton J, Green C (2005) The benefits of flood and coastal risk management: a manual of assessment techniques. Flood Hazard Research Centre Flood Hazard Research Centre. https://doi.org/10.1596/978-0-8213-8050-5

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users. In: Proceedings of the 19th international conference on World Wide Web, (January 2010), pp 851–860. https://doi.org/10.1145/1772690.1772777

Schulz A, Hadjakos A, Paulheim H, Nachtwey J, & Mühlhäuser M (2013) A multi-indicator approach for geolocalization of tweets. Seventh International AAAI Conference on Weblogs and Social Media, 573–582. http://doi.org/papers3://publication/uuid/62449928-74D1-4674-A1A7-24D5F6813F85

Serdyukov P, Murdock V, van Zwol R (2009) Placing flickr photos on a map. Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval—SIGIR '09, (May):484. https://doi.org/10.1145/1571941.1572025

Sultanik EA, Fink C (2012) Rapid geotagging and disambiguation of social media text via an indexed gazetteer. Proceedings of ISCRAM 12:1–10

Sun X, Mein RG, Keenan TD, Elliott JF (2000) Flood estimation using radar and raingauge data. J Hydrol 239(1–4):4–18. https://doi.org/10.1016/S0022-1694(00)00350-4

Takhteyev Y, Gruzd A, Wellman B (2012) Geography of twitter networks. Soc Networks 34(1):73–81. https://doi.org/10.1016/j.socnet.2011.05.006

UNISDR (2015) Global assessment report on disaster risk reduction 2015. Retrieved from https://www.unisdr.org/we/inform/publications/42809

Van Erp M, Rizzo G, & Troncy R (2013) Learning with the web: spotting named entities on the intersection of NERD and machine learning. CEUR Workshop Proceedings, 1019, 27–30

Zhang W, Gelernter J (2014) Geocoding location expressions in twitter messages: a preference learning method. J Spat Inf Sci 9(9):37–70. https://doi.org/10.5311/JOSIS.2014.9.170