Cartography and Geographic Information Science

ISSN: 1523-0406 (Print) 1545-0465 (Online) Journal homepage: http://www.tandfonline.com/loi/tcag20

Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment

Bernd Resch , Florian Usländer & Clemens Havas

CARTOGRAPHY AND GEOGRAPHIC INFORMATION SCIENCE, 2017
https://doi.org/10.1080/15230406.2017.1356242

Taylor & Francis
Taylor & Francis Group

ARTICLE

Check for updates

Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment

Bernd Resch [a,b], Florian Usländer [a] and Clemens Havas [a]

[a]Department of Geoinformatics – Z_GIS, University of Salzburg, Austria; [b]Center for Geographic Analysis, Harvard University, Cambridge, MA, USA

ABSTRACT

Current disaster management procedures to cope with human and economic losses and to manage a disaster's aftermath suffer from a number of shortcomings like high temporal lags or limited temporal and spatial resolution. This paper presents an approach to analyze social media posts to assess the footprint of and the damage caused by natural disasters through combining machine-learning techniques (Latent Dirichlet Allocation) for semantic information extraction with spatial and temporal analysis (local spatial autocorrelation) for hot spot detection. Our results demonstrate that earthquake footprints can be reliably and accurately identified in our use case. More, a number of relevant semantic topics can be automatically identified without a priori knowledge, revealing clearly differing temporal and spatial signatures. Furthermore, we are able to generate a damage map that indicates where significant losses have occurred. The validation of our results using statistical measures, complemented by the official earthquake footprint by US Geological Survey and the results of the HAZUS loss model, shows that our approach produces valid and reliable outputs. Thus, our approach may improve current disaster management procedures through generating a new and unseen information layer in near real time.

ARTICLE HISTORY
Received 18 January 2017
Accepted 13 July 2017

KEYWORDS
Social media; disaster management; machine-learning; semantic topic analysis; spatiotemporal analysis

Introduction

Natural disasters like earthquakes, floods, tsunamis, or hurricanes can cause severe property damage and harm personal life. Current procedures in disaster management, particularly in the immediate response phase, which focuses on dealing with human and economic losses and to mitigate a disaster's aftermath, are characterized by a variety of deficiencies. Among others, remote-sensing-based methods face the central disadvantage (aside from all well-known advantages) of temporal lags of approximately 48–72 h before information layers can be produced, which are relevant to disaster management. Furthermore, remotely sensed data are limited in terms of their spatial, spectral, and temporal resolution; their usability (e.g. through cloud cover); and their cost-efficient availability. More, lacking pre-disaster imagery oftentimes prevents successful and accurate change detection (Panagiota, Jocelyn, & Erwan, 2011).

A number of recent research efforts have shown that nontraditional data sources like social media networks and other crowdsourcing platforms can significantly improve disaster management (Boulos et al., 2011; Roche, Propeck-Zimmermann, & Mericskay, 2013). The advantages of these new approaches are rooted in the real-time nature of the data (inputs are available without significant temporal delays), and their *in-situ* character (information can be gained about the local situation like accessibility of streets and obstructed routes, the location of injured persons, the degree of damage of buildings, log jams in rivers, etc.). Like this, a large amount of messages from social media platforms or dedicated smartphone apps is created that can be regarded as real-time *in-situ* sensor data (Resch, 2013), which supports the assessment of a post-disaster situation through the provision of an additional, up-to-date information layer that can be produced in considerably less time compared to remote-sensing-based approaches (approx. 1–3 h). Naturally, these approaches require a functioning wireless network (mobile phone networks, WiFi networks, etc.), which is usually given even after major disasters, partly with some restrictions (Bengtsson, Xin, Thorson, Garfield, & Johan, 2011).

This paper proposes a new approach to analyze social media posts to assess the footprint of and the damage caused by natural disasters through combining semantic machine-learning techniques with spatial and temporal analysis. This goes clearly beyond previous

CONTACT Bernd Resch bernd.resch@sbg.ac.at

research that mostly pursued keyword-based approaches (Herfort, Schelhorn, de Albuquerque, & Zipf, 2014), neglected the spatial dimension (Cresci, Cimino, Dell'Orletta, & Tesconi, 2015), or purely focused on geographic analysis of manually classified social media posts (de Albuquerque, Herfort, Brenning, & Zipf, 2015). In our specific case study, we use Tweets to analyze the characteristics of the earthquake in Napa (CA, USA) in 2014. Yet, the approach is transferable to other data sources (Flickr, Instagram, blogs, short messaging services, etc.) and other geographic areas, even though cultural differences in social media usage need to be considered. It shall be noted that our research focuses on post-disaster analysis (social media posts are analyzed after a natural disaster has occurred), but it can also be applied in the fields of event detection or early warning, depending on the characteristics of the disaster.

Our approach is divided into two main parts. First, the textual content of social media posts is analyzed using an unsupervised, self-learning topic modeling approach, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). Like this, latent topics are extracted from the text corpus without the necessity for a priori knowledge about the disaster event, which is required with most previous approaches, particularly keyword-based ones. Thus, our approach aims to find the relevant "set of words" to be used for the analysis. The extracted topics can then be interpreted with respect to their relevance to a natural disaster. A cascading procedure identifies Tweets which are related to a disaster and then extracts a damage subtopic. After this semantic analysis, we identify spatial hot spots and temporal anomalies in the dataset. In other words, our method discovers similarities in space, time and semantics in a combined fashion, which goes beyond traditional approaches of social media analysis as explained above (see the "Introduction" and "Related Work" sections). From this information, we can infer disaster footprints and assess the damage caused by the disaster. The results of our analysis are validated through comparison with official earthquake and damage maps, following technical guide provided by the Federal Emergency Management Agency (FEMA) (Kircher, Whitman, & Holmes, 2006).

Related work

The central challenge in detecting events in online systems in near real time is that no prior knowledge about an event and no previously collected data are analyzed, but data that arrive as continuous streams. The general goal typically is to detect new events with low latency [the event needs to be identified shortly after it happened (Middleton, Middleton, & Modafferi, 2014)]. Robinson, Power, and Cameron (2013) developed a burst detector for earthquakes in Australia and New Zealand. An assumption of the authors is that if a new event happens, the frequency about event-related topics increases and can therefore be detected. Thus, they analyzed the real-time Twitter stream and used the keywords "earthquake" and "#eqnz" to identify user-related messages to a new earthquake. They limited their research on New Zealand and Australia and on earthquakes and used a keyword-based approach which has clear limits in expanding to other events or languages or collecting all available information. Spielhofer, Greenlaw, Markham, and Hahne (2016) present another method for analyzing social media streams. While their noise reduction algorithm is highly valuable, it is only the basis for further (geospatial) analysis, which we aim to cover in our research.

Kongthon, Haruechaiyasak, Pailai, and Kongyoung (2012) examine how people react to a crisis event like the Thailand floods in 2011 on social media by analyzing the textual content of Twitter posts. The aim of their work is to identify user-generated messages that are related to the flood event and to categorize these messages into five classes: situational announcements and alerts, support announcements, requests for assistance, requests for information, and other. For the text analysis, all Tweets with the hashtag #thaiflood are preselected and afterward categorized by a rule-based classification approach based on keywords. Imran, Castillo, Lucas, Meier, and Vieweg (2014) present a more sophisticated approach called Artificial Intelligence for Disaster Response, which uses a learning system to classify tweets. Yet, no geospatial analysis is performed, but the system purely focuses on text analysis.

Herfort et al. (2014) compare the spatial and temporal distribution of Twitter data related to flood phenomena with the footprint of the Elbe flood event (Germany) in June 2013. Like de Albuquerque et al. (2015), they use keyword-based filtering and portray the results on a map. Similarly, Guan and Chen (2014) categorize Twitter messages into "disaster-related" and "not disaster-related" through selecting predefined keywords and hashtags. Terpstra, de Vries, Stronkman, and Paradies (2012) also apply a keyword-based approach to detect damage and casualties reports from Twitter data during a storm at the Pukkelpop Festival in the year 2011.

One example, where LDA is applied to Tweets in the context of a natural disaster, is a study by

Kireyev, Palen, and Anderson (2009) where the algorithm is leveraged for the extraction of disaster-related Tweets. Their focus is on improving the LDA algorithm itself by using a term weighting function, therefore addressing the reduced information content as a consequence of the shortness of the single Tweets. Topics covering single events like an earthquake or a tsunami are successfully extracted, but no geospatial analysis is applied. Similarly, Sakaki, Okazaki, and Matsuo (2010) were able to detect crisis-related Twitter messages using a support vector machine and Kalman filtering. Yet, the geospatial nature of the results is critical to supporting an operational picture (Starr Roxanne Hiltz, Kushma, & Plotnick, 2014).

The approaches discussed in this section show two central shortcomings. First, most of them use keyword-based approaches (some authors actually suggest the use of machine-learning techniques for text analysis to improve the classification results). Second, the spatial information contained in the Tweets is not always considered, which is an essential drawback for first responders (Spence, Lachlan, & Rainear, 2016). In fact, many approaches only investigate one of the three dimensions (spatial, temporal, and semantic).

Method: spatial, temporal, and semantic analysis

The approach presented in this paper is illustrated in Figure 1. After preprocessing the raw Tweets, a topic modeling technique (LDA) is applied in a cascading fashion, i.e. we apply LDA to the results of the first iteration in a second iteration, to extract "earthquake"- and "damage"-related Tweets. Then, the generated topics are validated through statistical accuracy assessment, before spatial hot spot analysis is applied to investigate local spatial autocorrelation. This results in footprint maps for the earthquake itself and the damage it caused, which are finally validated using official earthquake information. The single steps of this workflow are described in more detail in the following subsections.

Data and study area

Our study uses Twitter data around the Napa (CA, USA) earthquake on 24 August 2014, geographically covering the larger Bay Area. Table 1 summarizes the characteristics of the dataset. The study area comprises not only several metropolitan areas (San Francisco, Oakland, San Jose, amongst others), but also a wide variety of comparatively rural, sparsely populated areas around the epicenter of the earthquake.

Figure 2 shows the spatial distribution (determined through Kernel Density Estimation with a bandwidth of 5 km and an output cell size of 100 m) of Tweets on two different days: left, 1 week before the earthquake, and right, on the day of the earthquake. It can be clearly observed that the number of Tweets heavily increased on the day of the disaster. However, this spatial density clustering method, which has been widely used in previous approaches, shows that Tweets strongly accumulate in urban areas, making it impossible to draw conclusions on the location of the earthquake and on potential damage.

Figures 3 and 4 show the number of Tweets posted per hour and per minute, respectively. It can be seen that a distinct peak in the overall Tweets per hour can be observed on the day of the earthquake, whereas the

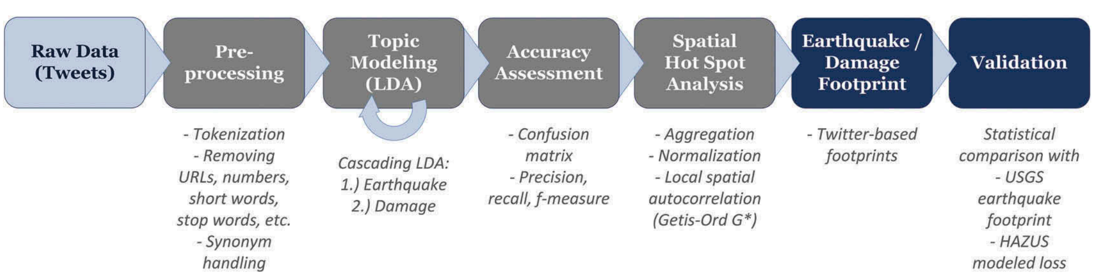

Figure 1. Overall workflow.

Table 1. Characteristics of the dataset.

Data description summary	
Geographic bounding box (WGS84)	−123.05°, 37.19°, −121.04°, 38.99°
Time period (UTC)	16 August 2014–31 August 2014
Number of georeferenced Tweets	1,012,650
Date of the earthquake event	24 August 2014
Number of Tweets on the event day	94,485

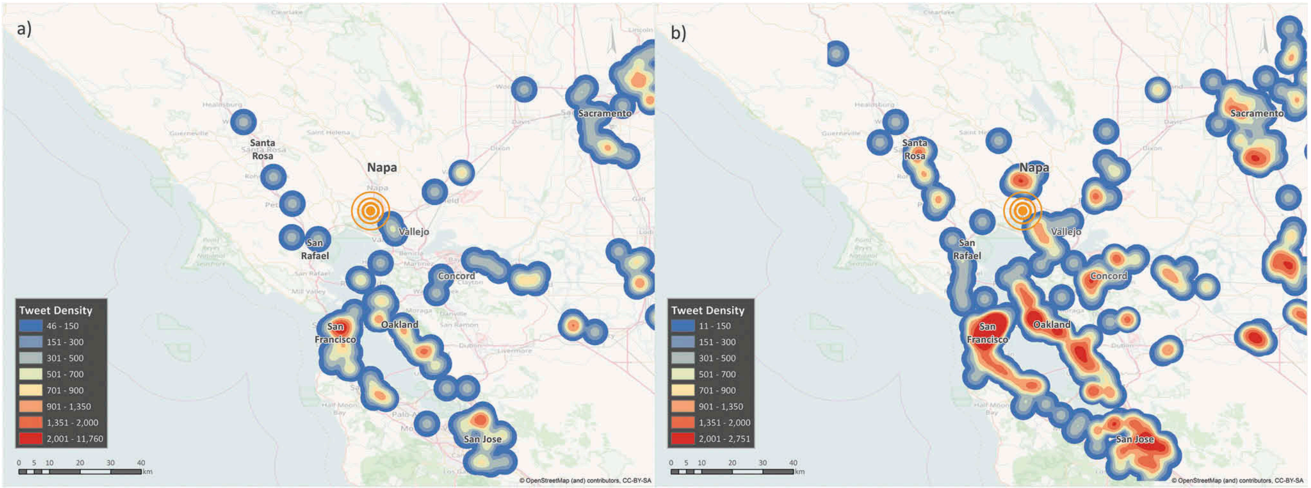

Figure 2. Spatial distribution of Tweets 7 days before the earthquake (left) and on the day of the earthquake (right). © OpenStreetMap (and) contributors; available under CC-BY-SA Licence.

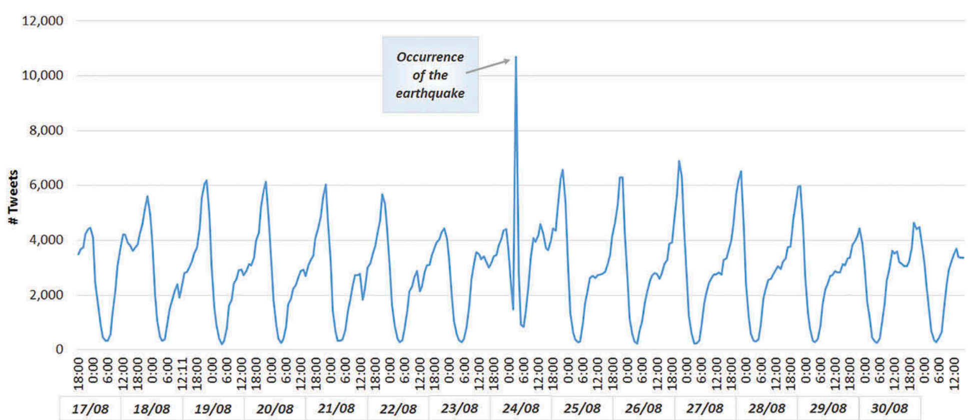

Figure 3. Number of Tweets per hour around the day of the earthquake (PDT).

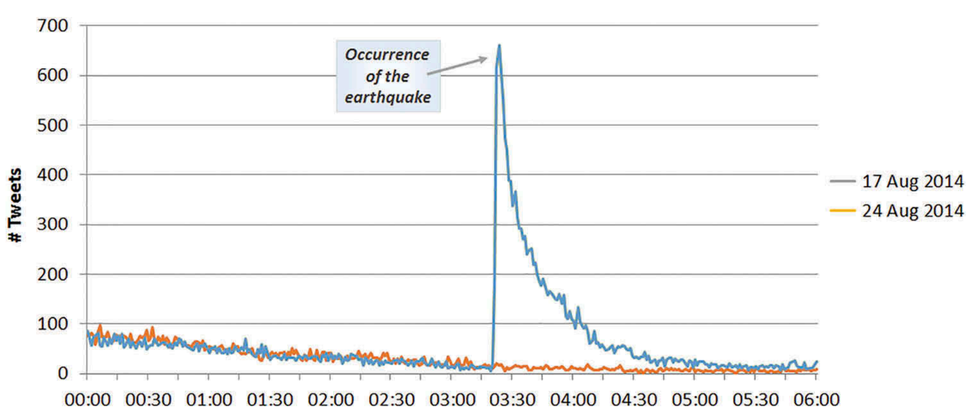

Figure 4. Number of Tweets per minute during the night of the earthquake (PDT).

minute-wise illustration shows an even stronger anomaly, particularly because the earthquake happened at night (3:20 a.m.), where the number of posted Tweets is generally low.

Data preprocessing

Before performing the actual semantic analysis on the Tweets, the dataset needs to be preprocessed to reduce potential statistical "noise," which is inherently present in social media posts (Steiger, Resch, & Zipf, 2015). The following paragraphs summarize the preprocessing steps. It shall be noted that the sequence of the preprocessing steps is obviously critical to achieve maximum reliability of the results.

Tokenization

This step splits Tweet text at each blank character to create a list of single tokens (stand-alone words, numbers, signs, or a concatenated string like a URL). This processing step allows for grasping the single tokens in further processing steps to filter, for instance, single words or symbols.

Tokens to Lowercase

This step is required to make the tokens more similar and to reduce sensitivity to typos, preventing that words with identical spelling, but differing cases would be treated as two semantically separate tokens (Ortigosa, Martiin, & Carro, 2014). One problem that arises from converting all characters to lowercase is that words, which differ in their semantic meaning, cannot be differentiated any more. However, this is not a central issue in the English language except for specific expressions like geographical location names, personal, company and brand names, or academic titles.

Removing URLs

URLs are not be considered in our topic modeling approach because they contain unspecific and hardly interpretable semantic information (Pak & Paroubek, 2010). Thus, they are regarded as noise in the data and are removed from the Tweet text.

Removing Numbers

Numbers are not considered in the text corpus because they do generally not contain semantically viable information for our purposes, but they would bias the outcomes of the topic modeling process. One exception may be the magnitude description for the earthquake. Yet, the word "magnitude" will indicate such a description and is represented in the topic model.

Removing Special Characters

Special characters are removed to preserve the focus on words in the dataset. The underlying assumption of this decision is that emoticons are frequently used in Tweets to express emotions in messages (Resch, Summa, Zeile, & Strube, 2016), but they are not standardized and thus hard to interpret. Yet, extracting emotion information from Tweets may increase interpretability of the results (Resch, Summa, Sagl, Zeile, & Exner, 2015; Zhao, Dong, Junjie, & Ke, 2012), but this is out of the scope of this research.

Synonym Handling

In the context of the Napa earthquake, several different expressions have been used to describe or talk about the earthquake. Thus, synonyms need to be handled to achieve realistic weighting (frequency) for the term "earthquake." Otherwise, different words, which are frequently co-occurring not only with the term "earthquake" but also with different synonyms like e.g. "quake," will create new, separate topics that are not combinable after the analysis. In our research, we manually created a list of synonyms (e.g. earthquake, quake, eq, shake). Synonym handling is critical to successful topic modeling and is thus separately discussed in the "Discussion" section.

Removing Short Words

Short words with three characters or less are removed from the text corpus because common agreements in the literature state that such words contain little semantic meaning (Pak & Paroubek, 2010).

Removing Stop Words

"Stop words" are commonly used words which do not carry distinct semantical meaning (e.g. auxiliary verbs, conjunctions and articles). These terms appear in almost every document and would form a topic on their own because they co-occur frequently with all other words (Ikonomakis, Kotsiantis, & Tampakas, 2005). Thus, we remove stop words using the predefined list from the NLTK Toolkit. Additionally, we manually identified additional stop words, which are specific to unedited text like Tweets, including "ain't," "gonna," "wanna," among others.

Removing Unique Words

Words, which only appear once in the text corpus, are removed because they will not contribute significantly to a topic but in turn drastically decrease the algorithm's performance.

Stemming

We used a Porter stemmer that reduces single words to their root, resulting in a condensation of the text corpus by combining different forms of a word, which in turn increases significance of the topic–word associations.

Vectorization and Market Matrix

The preprocessed data have to be transformed into a vector format for further calculation steps because

LDA expects a document-word-count matrix and a word dictionary. From these two representations, a corpus is built in a "bag-of-words" format, i.e. a collection of words without information on word order or grammar.

Machine-learning for Extracting Semantic Information from Tweets: Topic Modeling with Cascading LDA

To extract topics from our Twitter dataset, we use the LDA model (Blei et al., 2003), as shown in Figure 5. LDA assumes that each document d of a set of documents D contains one or more topics z, which is again defined by a probability distribution of single words w, the only observed variable in the model. Therefore, the latent variable ϕ represents a multinomial distribution of words within a topic. The other latent variable θ constitutes a multinomial distribution of topics in a document. α and β are two concentration parameters – α represents prior knowledge about the distribution of topics in a document, whereas β contains prior knowledge about the distribution of words in a topic. A higher value of α leads to a more smoothed distribution of topics over document, whereas a lower value, especially lower than zero, leads to a higher concentration of topics. ϕ, θ, and z are latent and therefore unobserved variables, which are generated when the process is running (Griffiths & Steyvers, 2004).

To achieve most reliable results, we test and compare different parameter combinations (number of passes, alpha, number of topics, beta, etc.), and the topics are then manually interpreted with regard to their semantic meaning (see the "Discussion" section for a critical review of this procedure). Steyvers and Griffiths (2006) propose an α value of $50/T$, but the best working value in our case seems to be an α value of 0.0001. This is reasonable because of the nature of our Twitter dataset, in which each Tweet expectedly only comprises one topic due to the shortness of the text (with up to 140 characters). Therefore, the probability distribution is more concentrated when $\alpha \ll 1$.

In our study, a 1000 *passes* deliver the best results – a higher number does not improve the learning effect any more but decreases performance. The *number of topics* strongly influences the topic–word distributions in terms of the granularity of the topics. Therefore, a lower number of topics leads to larger topics (containing more words), whereas a higher number of topics leads to smaller topics (containing fewer words). Effectively, this is a trade-off between minimal information loss through high granularity, and the generation of distinct and meaningful topics. In our case, the optimal number of topics turned out to be 25, as stated in the "Results and Validation" section.

LDA generates a document–topic matrix that contains the topics with corresponding probabilities for each document. This means that one Tweet comprises one or more topics. In a next step, the latent topics have to be semantically interpreted to identify disaster-related or damage-related Tweets.

One issue in the course of our research was the extraction of damage-related Tweets because no such topic was generated in the first LDA run. Thus, we applied cascaded LDA approach, i.e. we ran LDA a second time only on the earthquake-related Tweets. The results of these two runs are presented in the section on the "Results and Validation" section.

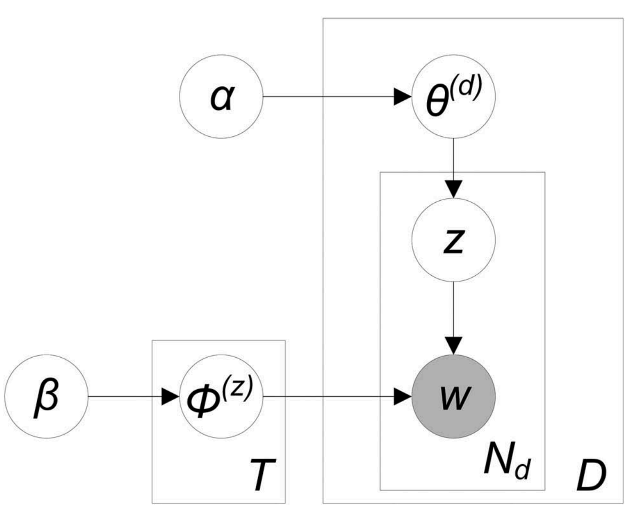

Figure 5. LDA plate notation (Based on Griffiths & Steyvers, 2004).

Accuracy assessment of the text classification

For assessing the accuracy of the Tweet classification, we create a confusion matrix, in which positives represent all Tweets, which are categorized as earthquake related, whereas negatives represent all Tweets which are categorized as non-earthquake related. From the confusion matrix, we compute the statistical measures *accuracy*, *error rate*, *precision*, *recall*, and *F-measure*, which are commonly applied in text mining (Feldman & Sanger, 2007). The computation of these evaluation statistics requires a gold standard for validation, which we created through manual labeling of a subset of 1509 Tweets by 3 independent annotators. For the annotation procedure, the three annotators were asked to label each Tweet with "earthquake related" or "not earthquake related." The annotators were instructed not to use context information (information about the earthquake, potential emotional state of the person sending a Tweet, etc.) when labeling the Tweets. The inter-annotator agreement (Fleiss Kappa) showed a comparatively high value of 0.886, as explained in section the "Results and Validation" section).

Spatial hot spot analysis

In the next step, we investigate spatial hot spots in the categorized earthquake-related Tweets. Therefore, we create a regular grid (cell size 1 km × 1 km) and summarize all Tweets within a grid cell. We chose this grid size by considering the number of points located in the study area and its size. Concretely, we used a technique from the area of point pattern analysis (Wong & Lee, 2005), which takes the size of the study area and the number of points into account to determine the optimal cell width:

$$l = \sqrt{2\frac{A}{n}},$$

where l is the side length of a grid cell, A is the size of study area, and n is the number of points in the study area.

Then, the earthquake-related Tweets are normalized over the population – we used the LandScan population layer at a resolution of 800 m (Oak Ridge National Laboratory, 2017) – and the overall number of Tweets per cell to prevent the undesirable effect that densely populated areas dominate the picture. Finally, we perform the actual hot spot analysis by applying the well-known *Getis-Ord G** method (Getis & Keith Ord, 1992; Ord & Getis, 1995) that detects local attribute clusters in our data.

Validation: earthquake and damage footprint modeling

For validation purposes, we compare the earthquake and damage footprints generated by our methodology to the official footprints, provided by public authorities. For the earthquake footprint, we use the pre-computed map by the US Geological Survey (USGS) that can be obtained from their website (US Geological Survey, 2016).

For the damage footprint, we use the loss model for earthquakes provided by the FEMA. We apply the basic equation $Loss = Hazard \times Vulnerability \times Exposure$ for generating a modeled damage map for the earthquake event. The *Hazard* is given by the earthquake footprint provided by USGS, representing the measured peak ground acceleration (PGA) that the earthquake has triggered, i.e. how hard the earth has been shaking at different geographic locations. For the *Exposure* and the *Vulnerability* factors, we used the HAZUS building grid, which contains information about the aggregated building type and building cost. Using this information, a simplified damage footprint can be modeled.

Results and validation

This section presents the results that we obtained from applying the methodology described in the previous section with respect to topic modeling, according accuracy assessment, and spatial hot spot analysis.

Topic modeling

Table 2 shows the topic–word distribution (in fact, the stems of the original words in the Tweets) of the first LDA iteration for the "earthquake" topic, in which the word "earthquake" has a probability of 43.30%, which is extraordinarily high. Furthermore, the words "california" and "damage" appear in this topic.

The results presented in Table 2 are highly appropriate to generate the earthquake footprint. However, in this first iteration, no stand-alone "damage" topic was generated. Thus, we applied a cascading LDA approach to examine the earthquake topic in more detail. Table 3 shows the according topic–word distribution for four subtopics.

As expected, the word "earthquake" is by far the word with the highest probability in all subtopics. The other words provide an indication about some granular differences within the earthquake topic. Like this, it can be clearly distinguished between different reporting processes: ad-hoc earthquake reports during the night, *post-hoc* earthquake reports in the morning after the event, damage reports, and the wine bucket challenge.

The identification of a distinct *damage* subtopic shows that this information is hidden in the overall earthquake topic from the first LDA iteration. The last subtopic (*bucket challenge*) represents an event called the "wine bucket challenge," a commonly used

Table 2. Word (stems) distribution for the "earthquake" topic.

Word	Probability (%)
Earthquak	43.30
California	5.90
Damag	2.80
Slept	2.50
Colleg	1.80
White	1.70
Report	1.60
Challeng	1.60
Hope	1.20
Stand	1.20

Table 3. Subtopics of the "earthquake" topic.

Subtopic interpretation	Word probability distribution
Ad-hoc reports during the night	0.326*earthquak; 0.045*california; 0.032*wake; 0.024*feel; 0.019*report; 0.018*usg; 0.015*shit; 0.015*holi; 0.013*area; 0.012*sanfrancisco
Post-hoc reports in the morning after	0.238*earthquak; 0.071*feel; 0.060*sleep; 0.026*wake; 0.020*last; 0.019*night; 0.012*morn; 0.012*right; 0.009*damn; 0.009*good
Damage reports	0.110*earthquak; 0.068*napa; 0.057*damag; 0.050*california; 0.020*northern; 0.016*colleg; 0.012*love; 0.010*magnitud; 0.009*north; 0.009*report
Bucket challenge	0.071*earthquak; 0.046*challeng; 0.037*bucket; 0.030*napa; 0.028*near; 0.025*hit; 0.020*nomin; 0.020*white; 0.017*worri; 0.012*stand

campaign in social media to raise awareness and raise funds for people who were affected by the Napa earthquake. Thus, the bucket challenge subtopic can be considered earthquake related, even though it has no direct value for disaster management.

To further investigate the validity of our conclusions, we performed some explorative analysis steps. Overall, 7160 Tweets were categorized as earthquake related in the entire dataset of 94,458 Tweets that were posted on the day of the earthquake. The distribution of Tweets across the four subtopics presented in Table 3 within the overall set of earthquake-related Tweets is as follows: 42% are ad-hoc earthquake reports, 32% are *post-hoc* reports, 15% are damage-related posts, and 11% refer to the wine bucket challenge.

Figure 6 illustrates the different temporal signatures of the four subtopics. These signatures confirm the interpretation of the subtopics as described above. The ad-hoc report subtopic shows a peak between 03:00 a.m. and 06:00 a.m., the *post-hoc* report subtopic between 06:00 a.m. and 12:00 noon, the damage reports keep steadily increasing over the day, and the bucket challenge is strongly present in the afternoon and the evening.

The spatial distribution of the points within the subtopics does not vary considerably. The hot spot analysis revealed concentrations around the epicenter for the damage and the ad-hoc report topics, whereas the *post-hoc* report and bucket challenge topics show no significant hot spots.

Accuracy assessment of the text classification

Table 4 summarizes the results of our statistical validation for a varying number of topics that are generated by the LDA process. These figures are based on three independent annotators whose Fleiss Kappa value (denoting the inter-annotator agreement) of 0.886 is particularly high. The numbers in Table 4 show that the precision (true positives over true positives plus false positives) is extraordinarily high, ranging from 94.97% to 97.29%, depending on the number of topics. The same applies to the accuracy (all positives over all Tweets), ranging from 81.05% to 86.41%. Yet, there is a clear trade-off between precision versus accuracy and recall (true positives over true positives plus false negatives), which are inversely proportional when changing the number of topics.

For our study, a high precision value means that Tweets, which have been recognized as earthquake related, are to a great degree actually earthquake related, and few Tweets have been incorrectly categorized as earthquake related. On the other side, the recall value can also be seen as the recognition rate. A high recall value means that a large proportion of actually earthquake-related Tweets have been recognized and few Tweets have been incorrectly

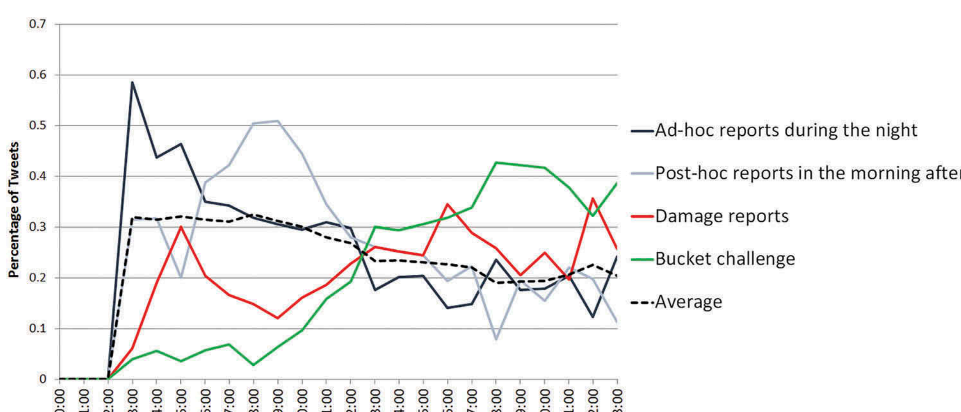

Figure 6. Temporal signatures of the four subtopics.

Table 4. Statistical validation of LDA outputs.

No. of topics	Accuracy (%)	Error rate (%)	Precision (%)	Recall (%)	*F*-measure (%)	#Tweets eq. topic	Topic–word distribution
15	86.41	13.59	94.97	73.28	82.73	9167	0.328*earthquak + 0.038*california + 0.023*class + 0.021*ariana + 0.015*grand + 0.015*second + 0.014*minut + 0.012*earth + 0.012*kill + 0.011*queen
20	85.95	14.05	95.26	71.94	81.97	8284	0.376*earthquak + 0.051*california + 0.035*morn + 0.034*Twitter + 0.020*damag + 0.015*tire + 0.015*earli + 0.015*magnitud + 0.013*parent + 0.009*fall
25	85.35	14.65	96.48	69.55	80.83	7160	0.433*earthquak + 0.059*california + 0.028*damag + 0.025*slept + 0.018*colleg + 0.017*white + 0.016*report + 0.016*challeng + 0.012*hope + 0.012*stand
40	81.05	18.95	97.29	58.96	73.42	4906	0.566*earthquak + 0.077*california + 0.016*forev + 0.010*giant + 0.010*broke + 0.010*northern + 0.009*epicent + 0.008*finna + 0.007*window + 0.007*behind

categorized as non-earthquake related. To combine these two parameters, the *F*-measure (harmonic mean between precision and recall) can be used, which seems to be the most meaningful parameter for our evaluation because it accounts for both, incorrectly categorized earthquake-related and non-earthquake-related Tweets. In addition, the probabilities in the topic–word distributions are considered, which change with a growing number of topics. In our particular setting, 25 topics, an α-value of 0.0001, and 1000 passes (learning steps) deliver the most promising results with respect to identifying damage.

Spatial hot spot analysis

Figure 7 illustrates the results of our spatial analysis of earthquake-related Tweets. The grid cells (colored in red–yellow–blue) represent the significant hot spots of earthquake-related Tweets, while the polygons in the background (colored in shades of green) outline the official USGS earthquake footprint in different PGA intensities.

For statistical validation, we overlaid the USGS footprint with the hot spots generated by our method. In a first step, we created a simple confusion matrix of grid cells matching the USGS and the Twitter footprints. The numbers in Table 5 result in the following statistical results: accuracy 86.45%, error rate 13.55%, precision 41.12%, recall 99.52%, and *F*-measure 58.19%.

In the next step, we compared the affected population as determined by the USGS and Twitter footprints. Therefore, we mapped the eight USGS classes for PGA to the four confidence interval levels of the geospatial hot spot analysis (see Figure 7). Figure 8 shows that the mapping generally expresses a good match between the two footprints. Yet, for high PGA values, the Twitter footprint seems to overestimate the affected population, which results from the imprecise mapping of confidence interval levels the USGS earthquake classes.

Figure 7 and our validation demonstrate that the Twitter hot spots match up remarkably well with USGS's earthquake footprint. Naturally, the hot spots correspond to populated areas to some degree, for instance, along highway 101 (going North, starting from San Francisco) where the cities of San Rafael, Petaluma and Santa Rosa are located. These cities are relatively close to the earthquake's epicenter and within the official footprint, so the corresponding hot spots can be considered as correctly identified.

On the contrary, other large cities like San Francisco, Oakland and San Jose are located outside USGS's footprint and are correctly identified as cold spots by our method. This is noteworthy because previous research mostly identified hot spots in large cities due to the vast amount of Tweets posted in metropolitan areas (cp. Figure 2).

Figure 9 illustrates the results of our spatial analysis of damage-related Tweets. The grid cells (colored in red–yellow–blue) represent the significant hot spots of damage-related Tweets, while the polygons in the background (colored in shades of green) outline the damage footprint as produced by FEMA's HAZUS loss model. Similarly to the earthquake footprints, the damage footprints generated by Tweets and the official model match quite precisely. Although the general patterns correlate well and the larger cities are again identified as cold spots, the intensity of Tweet hot spots and monetary losses show some slight differences. This particularly concerns the identification of the area around Vallejo as a hot spot with slightly lower confidence, even though heavy damage was caused in the area.

Discussion

Although the results demonstrate that our approach works reliably and accurately, there are still a number of aspects that deserve thorough discussion. These

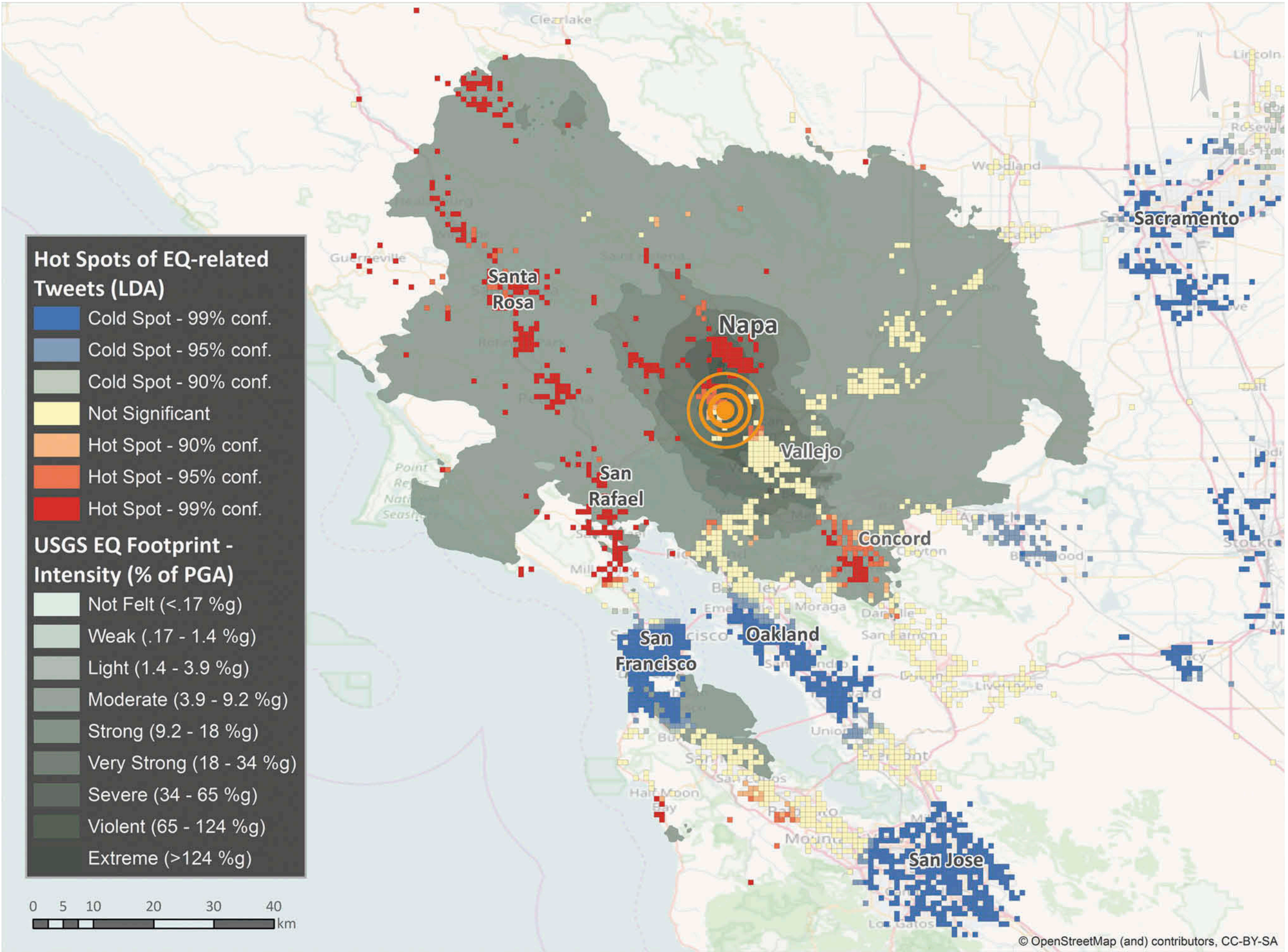

Figure 7. Twitter earthquake hot spots and USGS earthquake footprint. © OpenStreetMap (and) contributors; available under CC-BY-SA Licence.

Table 5. Statistical validation of spatial analysis.

	Twitter: EQ-related	Twitter: Not EQ-related
Within USGS footprint	824	4
Outside of USGS footprint	1180	310

particularly concern not only the topic modeling approach but also the interpretation of the semantic topics, and the spatial hot spot analysis.

Machine-learning topic modeling

First of all, typical topic models have been designed for edited text like newspaper articles or blog entries. Social media posts usually contain a large portion of *noise and irregularities* such as slang words, abbreviations, emoticons, irregular punctuation, "yoof speak," or other words that cannot be found in standard dictionaries, with which most previous approaches work (Eisenstein, 2013; Kireyev et al., 2009). These challenges can greatly be overcome through extensive preprocessing algorithms, which may, however, lead to information loss through removal of important words. Yet, a pertaining central research challenge is how to integrate information derived from user-generated data with measurements from technical sensors and contextual information (Sagl, Resch, & Blaschke, 2015). Furthermore, the number of topics covered by a Tweet is typically low (due to the shortness of the text) because each Tweet is treated as a single document. This bag-of-words model is characterized by low lexical redundancy and may thus lead to incoherent words within one topic–word distribution (Mehrotra, Sanner, Buntine, & Xie, 2013).

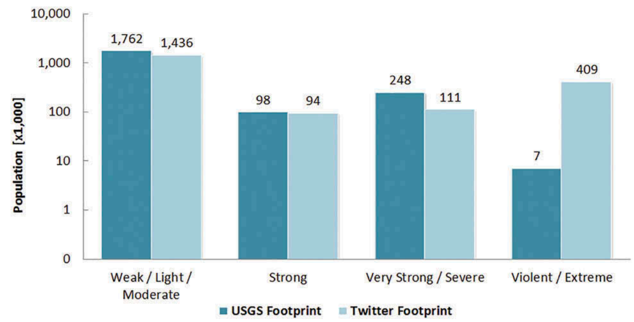

Figure 8. Population affected by USGS and Twitter footprints.

Another critical aspect using a topic modeling approach is the *semantic interpretation of topics*. Currently, there is no standardized method for assigning semantic meaning to a topic, leading to the uncertainty whether the interpretation of the topic is actually appropriate (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). This aspect becomes even more critical because the multinomial distribution of topics over documents leads to potential shortcomings in giving priority to the topic with the highest probability in that a topic with similar probability may be a better fit for a specific document. Also, this problem does not only apply to the interpretation step but also to the manual labelling procedure, which is prone to subjectivity. Moreover, the bag-of-words approach that LDA uses may lead to over-representation (e.g. a pop star's heavily re-tweeted post containing "a flood of tears") or underrepresentation (e.g. the use of several different hashtags for the same event) of words in topics. These aspects currently impede the development of an automated method for interpreting the topics. In a next short-term step, we aim to address this shortcoming through a keyword-based interpretation to infuse the generated topics with semantic meaning. Furthermore, a more formal integration of the temporal dimension into the process of interpreting semantic topics needs to be developed in the future. Currently, the temporal variations in the topic can be regarded as a strong indication of anomalies in the temporal domain, but our approach does not use a

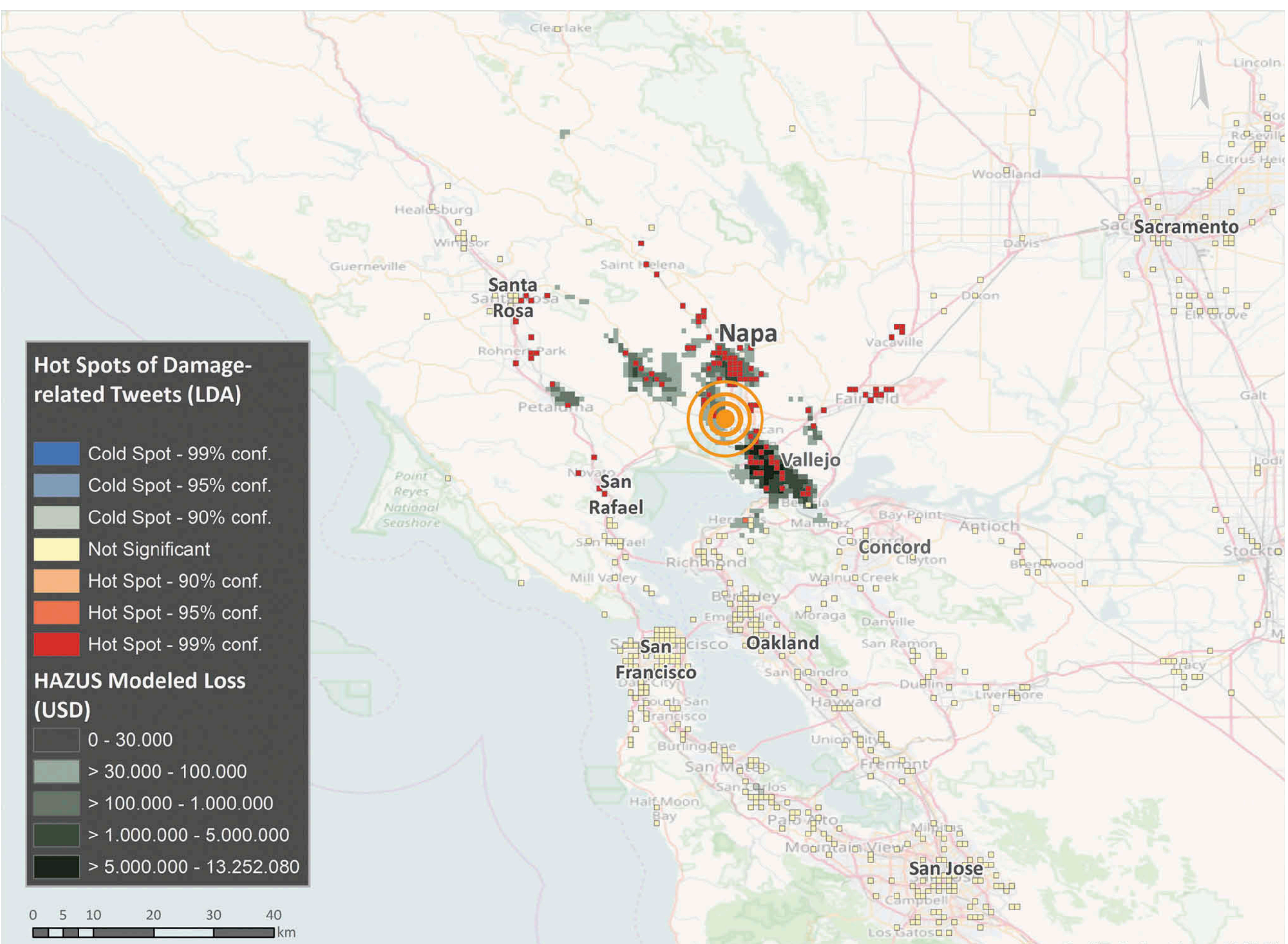

Figure 9. Twitter damage hot spots and FEMA's HAZUS loss modeling result. © OpenStreetMap (and) contributors; available under CC-BY-SA Licence.

structured method of integrated analysis of semantic and temporal yet.

The *cascaded approach of LDA* allows for identifying topics in higher granularity, as the extraction of the "damage" subtopic shows in our case: We were able to distinguish identify earthquake-related and damage-related Tweets. Moreover, we could observe a distinct difference in the temporal signatures of the subtopics. Naturally, one shortcoming of this approach is that the identification of such subtopics is limited by the recognition rate (recall) of a topic in the first iteration. Consequently, Tweets, which are not contained in a topic in the first step, cannot be used in the cascading approach.

A central advantage of LDA is its *transferability to other text corpora, and languages* because of the rapid adaptation of the unsupervised learning approach to another text corpus (Kireyev et al., 2009). Applying this approach to other scenarios and datasets only requires slight modifications in the preprocessing procedures and the interpretation of the generated topics. A central remaining challenge is the establishment of a meaningful, representative and non-redundant synonym list.

Finally, a pertinent challenge is the *choice of the model parameters*. Particularly for unstructured and small text documents as social media posts, there is no structured and formal approach to balance the priors, the number of topics and the number of passes. Thus, these parameters have to be defined in a trial-and-error approach. This step is important because these parameter choices are crucial for producing reliable outputs and for ensuring performance using larger datasets. We found that a lower number of topics in the first LDA iteration lead to a more comprehensive earthquake-related topic, which can then be decomposed into subtopics in a second iteration. Generally speaking, we can state that the optimal parameter settings presented in this paper are not representative for different analysis contexts. The settings may vary considerably depending on the type of disaster, the size of the study area, language, population distribution, tweet density, or the absolute number of tweets.

The procedure of parameter selection was one of the major issues in our research presented in this paper. As our aim was to identify earthquake *and* damage footprints, there was *no uniform parameter selection* for both topics. Therefore, we had to take a compromise between correctly identifying relevant Tweets and narrowing down the identification procedure to minimize false positives. This possibility for parametrization distinguishes our approach from keyword-based methods, which check whether a certain word is present in a dataset or not. Figure 10 shows the resulting maps for a keyword-based Tweet selection (earthquake-related Tweets on the left, damage-related ones on the right). Two aspects are striking: First, only hot spots with a confidence level of <99% and nonsignificant values are present in the earthquake footprint, but neither are hot spots with confidence levels between 90% and 99% nor are cold spots. Second, the damage footprint only contains nonsignificant values (no hot spots or cold spots). This fact is subject to further investigation in future research. Although there are some open research questions with respect to our approach, we can state that the advantages of our approach are the combination of words into topics (co-occurrence of words rather than single words), the identification of non-anticipated words (social media users oftentimes use their own terminology), and the correspondence of word

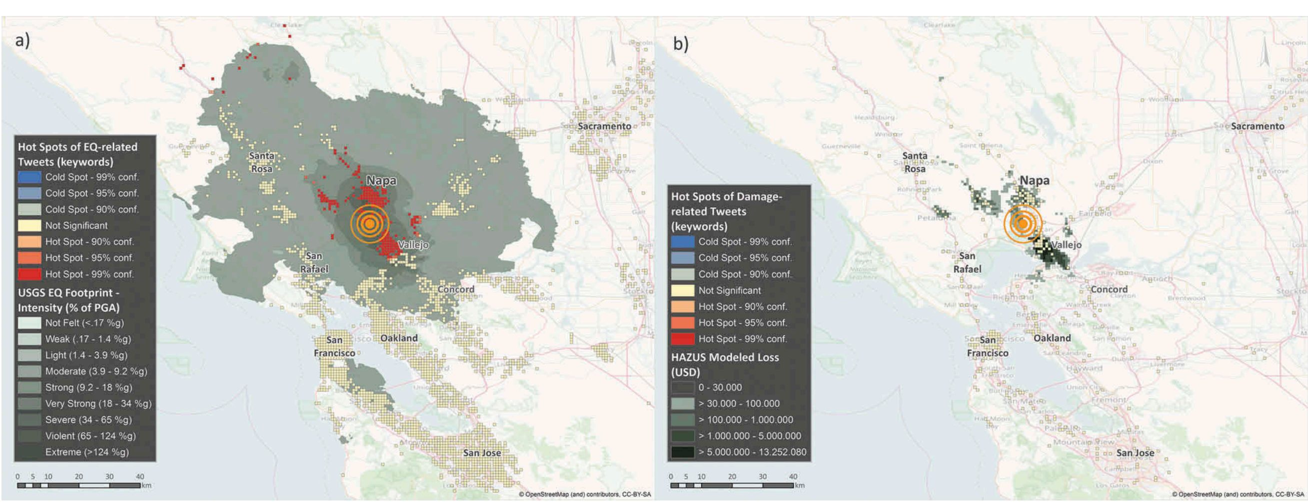

Figure 10. Keyword-based Tweet hot spots (left: earthquake; right: damage). © OpenStreetMap (and) contributors; available under CC-BY-SA Licence.

combinations and topics (which is important because of the particular und oftentimes non-standardized text style in social media).

Spatial hot spot analysis

We assess local spatial autocorrelation using the *Getis-Ord G** statistics that delivers statistically significant *hot spots and cold spots*. One issue with using this method with Tweets is that single outliers (e.g. one single highly active Twitter user) may produce a local hot spot. Yet, this effect can be reduced by normalizing the earthquake-related Tweets by the overall number of Tweets and the population per cell. In addition, a more fundamental question is whether traditional analysis methods, which greatly rely on Waldo Tobler's "First Law of Geography" (Tobler, 1970), can actually be applied to social media datasets because they may not conform to steadily decaying distance relationships.

Another important aspect is that *spatial scale* strongly determines the granularity of the topics. For instance, the use of social media messages, which have been posted in the vicinity of a disaster event, results in clearly distinguishable and crisp topic definitions, whereas data from larger areas may dilute topic generation. Another aspect of scale concerns the environment that is investigated. While peaks in disaster-related posts will mostly be distinct and stand out as an anomaly, they may only cause a smaller peak in densely populated urban areas with a high number of regular (non-disaster-related) posts. This issue can be addressed through a rigorous definition of an evaluation method, balancing precision, recall, and F-measure.

In addition, the *characteristics of the social media dataset* itself lead to constraints in their analysis. These shortcomings have been thoroughly discussed in previous literature (Steiger, Westerholt, Resch, & Zipf, 2015; Sui & Goodchild, 2011) and are therefore not elaborated on in detail here. First, social media users are not uniformly distributed over all age groups and education levels and are thus not representative for the entire population. Furthermore, the geolocation of Tweets is not necessarily the actual location of the observation of a real-world phenomenon even though Tweets are oftentimes considered *in-situ* reports. The same applies to temporal uncertainty. Finally, only a few percent of all Tweets contain an explicit geolocation, further biasing the dataset. The above-mentioned shortcomings may be greatly mitigated by using a large enough dataset and consciousness of the implications of interpreting the results.

Moreover, the *different characteristics of diverse disaster types* may be reflected in social media. For instance, a flood event, which is usually preceded by a significantly long early warning phase, may cause a steady, but slow increase in the number of Tweets sent. On the contrary, an earthquake event typically causes a distinct spike with steep slopes in the number of posted Tweets because there is literally no early warning phase. This also imposes implications on the analysis method, requiring different method sets and parameter settings, depending on the type of disaster to be assessed.

Conclusion

This paper presented an approach to analyze social media posts to assess the footprint of and the damage caused by natural disasters through combining semantic machine-learning techniques (LDA) with spatial and temporal analysis (local spatial autocorrelation for hot spot detection). This significantly advances the state of the art because previous research mostly pursued keyword-based approaches, neglected the spatial dimension, or purely focused on geographic analysis without taking semantics into account. Furthermore, our machine-learning approach does not require far-reaching a priori knowledge about the disaster event (e.g. keywords to search for), which is an important factor in social media-based analysis because social networks often develop their own terminology (hashtags, expressions, etc.). Thus, our approach allows for detecting or quantifying events that may not be characterized through well-known keywords. In other words, our approach aims to find the relevant "set of words" to be used for the analysis.

Our results demonstrate that earthquake footprints can be reliably and accurately identified in our case study, and significant hot spots in local spatial autocorrelation provide valuable insights into the nature and the specific spatial distribution of areas affected by an earthquake. In addition, a number of relevant semantic topics could be automatically identified, which clearly show differing temporal and spatial signatures. Furthermore, we were able to generate a damage map that indicates where significant losses have occurred. The validation of our results using statistical measures, complemented by the official earthquake footprint by USGS and the results of the HAZUS loss model, shows that our approach produces valid and reliable outputs.

Concluding, it can be stated that the research presented in this paper may improve current disaster management procedures through generating a new and unseen information layer. Furthermore, result maps can be produced in near real time because of the small temporal lag in social media posts (as opposed to remote-sensing-based approaches), and their *in-situ* character.

Yet, there are a number of open challenges, which will be addressed in future research, including the optimization of the preprocessing routines, automated interpretation of the generated semantic topics, the development of a standardized method for determining the best parameters for the semantic analysis, dealing with the lacking representativeness (population-wise and spatially) of social media posts, the creation of an automated validation procedure, and the development of a real-time monitoring system.

Acknowledgments

We would like to thank Dr. Wendy Guan from Harvard University's Center for Geographic Analysis for her support through providing us with the Twitter data for our study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

We would like to express our gratitude to the European Commission's Horizon 2020 program for supporting the project "E2mC – Evolution of Emergency Copernicus services," reference number 730082. Furthermore, we would like to thank the Doctoral College GIScience [DK W 1237-N23] at the Department of Geoinformatics – Z_GIS, University of Salzburg, Austria, funded by the Austrian Science Fund (FWF) for their support.

ORCID

Bernd Resch http://orcid.org/0000-0002-2233-6926
Florian Usländer http://orcid.org/0000-0002-8484-7393
Clemens Havas http://orcid.org/0000-0003-0390-5094

References

Bengtsson, L., Xin, L., Thorson, A., Garfield, R., & Johan, V. S. (2011). improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine*, 8(8), e1001083. doi:10.1371/journal.pmed.1001083

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022. http://www.jmlr.org/papers/v3/blei03a.html

Boulos, M. N. K., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., … Chuang, K.-Y. S. (2011). Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, OGC standards and application examples. *International Journal of Health Geographics*, 10(1), 1–29. doi:10.1186/1476-072X-10-67.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., & Culotta, A. (Eds.), *Advances in neural information processing systems* (pp. 288–296). Red Hook, NY: Curran Associates.

Cresci, S., Cimino, A., Dell'Orletta, F., & Tesconi, M. (2015). Crisis mapping during natural disasters via text analysis of social media messages. In Wang, J., Cellary, W., Wang, D., Wang, H., Chen, S.-C., Li, T., & Zhang, Y. (Eds.), *Web Information Systems Engineering – WISE 2015* (pp. 250–258). Cham, Switzerland: Springer International. doi:10.1007/978-3-319-26187-4_21

de Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, 29(4), 667–689. doi:10.1080/13658816.2014.996567

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 359–369). Atlanta, GA: Association for Computational Linguistics.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Getis, A., & Keith Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5228–5235. doi:10.1073/pnas.0307752101

Guan, X., & Chen, C. (2014). Using social media data to understand and assess disasters. *Natural Hazards*, 74(2), 837–850. doi:10.1007/s11069-014-1217-1

Herfort, B., Schelhorn, S.-J., de Albuquerque, J. P., & Zipf, A. (2014). Does the spatiotemporal distribution of tweets match the spatiotemporal distribution of flood phenomena? A study about the River Elbe flood in June 2013. In Hiltz, S. R., Pfaff, M. S., Plotnick, L., & Shih, P. C. (Eds.), *Proceedings of the 11th International ISCRAM Conference* (pp. 747–751). University Park, PA: Pennsylvania State University. Available from http://www.iscram.org/legacy/ISCRAM2014/papers/p101.pdf

Hiltz, S. R., Kushma, J., & Plotnick, L. (2014). Use of social media by US public sector emergency managers: Barriers and wish lists. In S. R. Hiltz, M. S. Pfaff, L. Plotnick, & P. C. Shih (Eds.), *Proceedings of the 11th International ISCRAM Conference* (Vol. 279: pp. 602–611). University Park, PA: Pennsylvania State University. Available from http://www.iscram.org/legacy/ISCRAM2014/papers/p11.pdf

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8), 966–974.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). AIDR: artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 159–162). New York: ACM. doi:10.1145/2567948.2577034

Kircher, C. A., Whitman, R. V., & Holmes, W. T. (2006). HAZUS earthquake loss estimation methods. *Natural*

Hazards Review, 7(2), 45–59. doi:10.1061/(ASCE)1527-6988(2006)7:2(45)

Kireyev, K., Palen, L., & Anderson, K. (2009, December). *Applications of topics models to analysis of disaster-related Twitter data*. NIPS Workshop on Applications for Topic Models: Text and Beyond, Whistler, BC. Retrieved from http://www.umiacs.umd.edu/~jbg/nips_tm_workshop/15.pdf

Kongthon, A., Haruechaiyasak, C., Pailai, J., & Kongyoung, S. (2012). The role of Twitter during a natural disaster: Case study of 2011 Thai flood. In Kocaoglu, D.F., Anderson, T. R., Daim, T. U., Jetter, A., Weber, C. M. (Eds.), *2012 proceedings of PICMET'12: Technology management for emerging technologies* (pp. 2227–2232). Piscataway, NJ: IEEE.

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 889–892). New York: ACM. doi:10.1145/2484028.2484166

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2), 9–17. doi:10.1109/MIS.2013.126

Oak Ridge National Laboratory. (2017). LandScan Global Population Database. Retrieved from: http://web.ornl.gov/sci/landscan/.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4), 286–306. doi:10.1111/j.1538-4632.1995.tb00912.x

Ortigosa, A., Martiin, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior*, 31, 527–541. doi:10.1016/j.chb.2013.05.024

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, … D. Tapias (Eds.), *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC'10)* (1320–1326). Valletta, Malta: European Language Resources Association (ELRA).

Panagiota, M., Jocelyn, C., & Erwan, P. (2011). *State of the art on remote sensing for vulnerability and damage assessment on urban context*. Grenoble, France: URBASIS Consortium.

Resch, B. (2013). People as sensors and collective sensing - Contextual observations complementing geo-sensor network measurements. In J. M. Krisp (Ed), *Progress in location-based services* (pp. 391–406). Berlin: Springer.

Resch, B., Summa, A., Sagl, G., Zeile, P., & Exner, J.-P. (2015). Urban emotions — Geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data. In G. Gartner & H. Huang (Eds.), *Progress in location-based services 2014* (pp. 199–212). Cham, Switzerland: Springer International.

Resch, B., Summa, A., Zeile, P., & Strube, M. (2016). Citizen-centric urban planning through extracting emotion information from Twitter in an interdisciplinary space-time-linguistics algorithm. *Urban Planning*, 1(2), 114–127. doi:10.17645/up.v1i2.617

Robinson, B., Power, R., & Cameron, M. (2013). A sensitive Twitter earthquake detector. In *Proceedings of the 22nd International Conference on World Wide Web*, (999–1002). New York: ACM. doi:10.1145/2487788.2488101.

Roche, S., Propeck-Zimmermann, E., & Mericskay, B. (2013). GeoWeb and crisis management: Issues and perspectives of volunteered geographic information. *GeoJournal*, 78(1), 21–40. doi:10.1007/s10708-011-9423-9

Sagl, G., Resch, B., & Blaschke, T. (2015). Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities. *Sensors*, 15 (7), 17013–17035. doi:10.3390/s150717013

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web* pp. (851–860). New York: ACM.

Spence, P. R., Lachlan, K. A., & Rainear, A. M. (2016). Social media and crisis research: data collection and directions. *Computers in Human Behavior*, 54, 667–672. doi:10.1016/j.chb.2015.08.045

Spielhofer, T., Greenlaw, R., Markham, D., & Hahne, A. (2016). Data mining Twitter during the UK floods: Investigating the potential use of social media in emergency management." In *3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)* (pp. 1–6). doi:10.1109/ICT-DM.2016.7857213.

Steiger, E., Resch, B., & Zipf, A. (2015). Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*. doi:10.1080/13658816.2015.1099658

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, 54, 255–265. doi:10.1016/j.compenvurbsys.2015.09.007

Steyvers, M., & Griffiths, T. (2006). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Hillsdale, NJ: Lawrence Erlbaum.

Sui, D., & Goodchild, M. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*, 25(11), 1737–1748. doi:10.1080/13658816.2011.604636

Terpstra, T., de Vries, A., Stronkman, R., & Paradies, G. L. (2012, Apri). "Towards a realtime Twitter analysis during crises for operational crisis management. In Proceedings of the 9th International ISCRAM Conference, Vancouver, BC.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240. doi:10.2307/143141

US Geological Survey. (2016). Earthquakes. Retrieved from http://earthquake.usgs.gov/earthquakes.

Wong, D. W. S., & Lee, J. (2005). *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. Hoboken, NJ: Wiley.

Zhao, J., Dong, L., Junjie, W., & Ke, X. (2012). MoodLens: An emoticon-based sentiment analysis system for Chinese Tweets. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (pp. 1528–1531). New York: ACM. doi:10.1145/2339530.2339772