ORIGINAL ARTICLE

What happens where during disasters? A Workflow for the multifaceted characterization of crisis events based on Twitter data

Jens Kersten (iD) | Friederike Klan (iD)

Institute of Data Science, German Aerospace Center (DLR), Jena, Germany

Correspondence
Jens Kersten, Institute of Data Science, German Aerospace Center (DLR), Maelzerstr. 3, Jena 07745, Germany.
Email: jens.kersten@dlr.de

Abstract

Twitter data are a valuable source of information for rescue and helping activities in case of natural disasters and technical accidents. Several methods for disaster- and event-related tweet filtering and classification are available to analyse social media streams. Rather than processing single tweets, taking into account space and time is likely to reveal even more insights regarding local event dynamics and impacts on population and environment. This study focuses on the design and evaluation of a generic workflow for Twitter data analysis that leverages that additional information to characterize crisis events more comprehensively. The workflow covers data acquisition, analysis and visualization, and aims at the provision of a multifaceted and detailed picture of events that happen in affected areas. This is approached by utilizing agile and flexible analysis methods providing different and complementary views on the data. Utilizing state-of-the-art deep learning and clustering methods, we are interested in the question, whether our workflow is suitable to reconstruct and picture the course of events during major natural disasters from Twitter data. Experimental results obtained with a data set acquired during hurricane Florence in September 2018 demonstrate the effectiveness of the applied methods but also indicate further interesting research questions and directions.

KEYWORDS
event characterization, natural disasters, spatio-temporal analysis, Twitter

1 | INTRODUCTION

Twitter enables users to post tweets containing text, images and videos with their current locations shared. With an average rate of 0.85%–3% tweets being geo-tagged, around 7,000,000 geo-tagged tweets are posted per day (Huang, Li, & Shan, 2018). A 1% fraction of this data stream is freely available and can be accessed via the Twitter Streaming API (Twitter, 2019).

In several studies and use cases, the great value and importance of analysing social media streams to extract information for supporting rescue and helping activities in case of natural disasters and technical accidents are demonstrated (Imran, Castillo, Diaz, & Vieweg, 2018; Reuter, Hughes, & Kaufhold, 2018). Even though social media can be a useful information source for all phases of natural disaster management, this work focuses on disaster response activities during and immediately after an event has occurred. Especially in large-scale

events, where a huge number of citizens are affected, geo-located information about the current situation is required in order to manage rescue and helping activities. We are therefore interested in applying data analysis methods to social media data that are able to reveal important information about the impacts. As an example, the identification of citizens in danger of life, in temporarily inaccessible and isolated settlement areas or affected by power outages may be extremely helpful for relief forces. From an information consumer's perspective, the obtained insights from Twitter data have to be considered as one of many different information sources, such as satellite imagery and news sites, feeds and databases, that have to be fused in order to obtain a wide-scale representation of the real situation.

According to recent reviews in the field of social media analytics in general (Wang & Ye, 2018) and in the context of crisis management (Imran et al., 2018), spatio-temporal analysis of Twitter data (Steiger, de Albuquerque, & Zipf, 2015) and social media visual analytics (Ngamassi, Malik, Zhang, & Edbert, 2017), the following needs for further research can be pointed out:

- Two main types of information about crisis situations are usually desired: understanding "the big picture" (for humanitarian and governmental emergency management organizations) versus finding "implicit and explicit requests related to emergency needs that should be fulfilled or serviced as soon as possible" ("actionable insights" for local police forces and firefighters).
- Methods that can automatically identify actionable messages from a live data stream during emergency events, including their urgency and information category, are essential for crisis responders to launch rapid relief efforts.
- One research objective should be to develop novel methods and approaches towards the spatio-temporal analysis and exploration of social media data, also by leveraging existing geographic knowledge.
- The spatial dimension in social media data has been given particular attention, while other dimensions have not been fully exploited in data fusion.
- Visual analytics tools should be developed to help determining the progress of an emergency and to help identifying the next possible solutions that could be used to alleviate damages.
- The fusion of social media data with census data and remote-sensing imagery is currently mainly about simple overlaying and aggregation and could therefore be further investigated.

Motivated by these findings, a flexible exploration of Twitter data streams based on spatio-temporal analyses and taking into account multiple dimensions of tweets, that is space, time and message content, is focused in this work. Our proposed workflow is intended to provide both a "big picture" of the course of events within an affected area and the identification of actionable tweets. The big picture describes when, where and what real and virtual event-related topics are discussed and how they change over time and space, whereas actionable information is conveyed by single messages indicating that help is urgently required.

The workflow covers data acquisition, analysis and visualization, and is intended to provide a multifaceted and detailed picture of events that happened in the affected area. The involved methods for short text analysis and clustering are identified according to the criteria flexibility, low computational costs, applicability to short messages and the successful application in other related works. Furthermore, each of the methods is intended to provide a different complementary view on the data. For instance, local hotspot analysis is applied to analyse the spatial distribution of crisis-related tweets within affected areas, whereas spatio-temporal tweet clustering and classification is used to analyse the discussed topics within clusters. In order to investigate the capabilities of the workflow to reconstruct and picture the course of natural disaster events, it is exemplarily applied to a Twitter data set acquired during hurricane Florence in September 2018.

The article is organized as follows. In the next section, related work in the context of spatial and spatio-temporal data analysis is discussed. From a methodological point of view, this mainly involves clustering techniques, whereas from an application-oriented perspective, event detection and spatio-temporal process analysis are focused. Based on the review, our workflow for the multifaceted analysis of Twitter data related to natural disasters is proposed in section 3. The components of the workflow are investigated in detail in section 4. In section 4.1, the Twitter data set used for evaluation is described. After a discussion of the obtained results in section 5, conclusions are drawn and open research questions are formulated.

2 | RELATED WORK

Exploring and reconstructing the course of (sub-) events potentially involves many different tasks, such as crawling, filtering, localizing and ranking of information, the detection, tracking and summarization of events, text classification, semantic enrichment and topic modelling. Systems for event-related social media data analysis usually focus on a specific challenge and therefore comprise a fixed sequence of tasks for which a fixed set of methods is utilized. Abel, Hauff, Houben, Stronkman, and Tao (2012) introduced Twitcident, a framework and Web-based system for filtering and searching information about real-world incidents or crises. Information filtering is adapted to the current temporal context by event profiling and semantic enrichment of tweets. Faceted search and analytic tools allow to retrieve particular information from the enriched data. With a focus on automated classification of crisis-related tweets, a platform using Artificial Intelligence for Disaster Response (AIDR) is proposed by Imran, Castillo, Lucas, Meier, and Vieweg (2014). The Emergency Analysis Identification and Management System (EAIMS) proposed by McCreadie, Macdonald, and Ounis (2016) focuses on real-time detection of emergency events, related information finding and credibility analysis tools for use over social media during emergencies. Event Tracker (Thomas, McCreadie, & Ounis, 2019) aims to provide a unified view of an event, integrating information

from emergency response officers, the public (via social media) and also volunteers from around the world. Components are real-time identification of critical information, automatic grouping of content by the information needs of response officers, and real-time volunteer management and communication.

All mentioned systems significantly contribute to deduce an overview of the situation in areas affected by natural or man-made disasters. However, we can observe that the spatial and the spatio-temporal information of the data is not fully exploited. Despite the grouping of content in Thomas et al. (2019), rather tweet-wise analyses are conducted. A further result of our review is that spatio-temporal analyses are usually done in the context of a specific application, for example earthquake damage assessment (Resch, Uslaender, & Havas, 2018), or small-scale events (Huang et al., 2018), and therefore are often not transferable to other event types. This motivates our generic approach and the exploitation of message content, space and time.

In order to be application independent, we chose a rather method-oriented structure of the subsequent review. Since clustering is one of the key techniques in the field of spatio-temporal data mining (Shi et al., 2016), common clustering techniques are reviewed in the following. Clustering is often utilized for event detection and for spatio-temporal process analysis. Therefore, related work in this context is also discussed.

2.1 | Clustering approaches

A statistical approach for hotspot analysis is the *Getis-Ord G**-method (Ord & Getis, 1995). This local spatial autocorrelation approach is applied for disaster footprint estimation and damage assessment based on social media data in (Resch et al., 2018).

Kernel density estimation (KDE) methods estimate a density surface from a set of point-based locations by adding functions, for example Gaussians, centred at each data point. In Lee, Gong, and Li (2017), spatial KDE is extended to incorporate temporal attributes of point data. Tackling the main problem of different units in space and time, the authors propose different extensions of KDE (space–time KDE, STKDE). Experimental results show that STKDE methods can identify hot and cold clusters better than KDE.

Space Time Scan Statistics (STSS) (Cheng & Wicks, 2014) is an approach to analyse data points (incidents) within a space–time cube. A cylindrical window, of varying radius (space) and height (time), is moved across all possible space–time locations. Based on the number of observed incidents compared to the number of expected incidents, clusters of interest are identified.

In density-based clustering algorithm for applications with noise (DBSCAN) (Ester, Kriegel, Sander, & Xu, 1996), the number of clusters is identified based on the quantity of densely connected components. Method parameters are the radius and the minimum number of neighbours within this radius. From these parameters, clusters with different shapes and similar densities are found. ST-DBSCAN (Birant & Kut, 2007) is an extension of DBSCAN discovering clusters according to non-spatial, spatial and temporal features.

Gaussian mixture modelling is a parametric approach for clustering data. Advantages of such model-based approaches are that they can handle differently scaled variables, for example resulting from different units of measure. Furthermore, since mixtures of different distributions are allowed, the joint analysis of continuous and categorical data is enabled. Thorough surveys of state-of-the-art mixed data clustering algorithms are provided in Ahmad and Khan (2019) and Hunt and Jorgensen (2011).

With special emphasis on the joint analysis of numerical and categorical data, a fast density-based clustering algorithm (FDCA) is proposed in Jinyin, Huihao, Chen, Shanqing, and Zhaoxia (2017). In a one-time scan through all data points, the cluster centres are determined automatically. A data similarity metric is designed to enable the joint clustering of numerical, categorical and ranking attributes. However, the choice of the radius defining the local neighbourhood around each data point is not trivial and small changes can significantly affect the shapes and sizes of the clusters.

In contrast to the clustering approaches reviewed above, topic modelling techniques model tweets or documents as a mixture of topics. Latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) is a widely used probabilistic topic model, where each topic has a probability distribution over the terms contained in the documents. Capturing representative topics from short texts with limited context is known to be challenging. Additionally, topic modelling approaches usually incur high computational costs while being not quite effective in handling parallel events.

A suitable technique for dimensionality reduction and data analysis is non-negative matrix approximation (NNMA) (Dhillon & Sra, 2005), also known as non-negative matrix factorization (NMF) (Cichocki & Phan, 2009). As pointed out in Casalino, Castiello, Buono, and Mencar (2018), NMF is known to provide meaningful interpretations of mined information and therefore is successfully applied in various domains including the analysis of tweets.

2.2 | Event detection

Supervised, unsupervised and semi-supervised methods for the detection of physically occurring events as well as for emerging or popular topics based on Twitter data are discussed in Ramachandran and Ramasubramanian (2018) and Hasan, Orgun, and Schwitter (2018). The authors conclude that further work is required to propose effective measures to filter out spam and trivial events. Furthermore, geo-tagging, the additional incorporation of videos and images as well as the detection of spambots, rumours and false reports is yet to be explored in more detail. Detecting events based on bursty keywords alone as a measure is not encouraged, as it may also return noisy data.

In Cheng and Wicks (2014), STSS is applied for event detection using Twitter data. STSS looks for clusters within a data set across both space and time, regardless of tweet content. It is expected that clusters of tweets will emerge during spatio-temporally relevant events, as people will tweet more than expected in order to describe the event and spread information. The method successfully detects

events such as football matches, train and flight delays, and a helicopter crash from Twitter data.

With a special emphasis on real-world events, an approach for geo-spatial Twitter event detection is proposed in Walther and Kaisser (2013). Target events are often expected on a rather small-scale, meaning that they happen at a specific place in a given time period, and are often covered by only few tweets. After a pre-selection of tweets based on their geographical and temporal proximity, clusters (event candidates) are constituted based on simple distance measures in space and time. 41 features that address various aspects of the event candidates are then used to rank the tweets and to make a binary decision as to whether a tweets cluster constitutes a real-world event or not.

In Berlingerio et al. (2013), event identification is done by utilizing a spatio-temporal clustering approach based on the Louvain method (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). The original method, intended to analyse graph networks for community detection, was extended to also take into account the textual content of a tweet. The method successfully identified major events of the occupy Wall Street movement in 2011.

2.3 | Spatio-temporal process analysis

The detection of small-scale spatio-temporal events from geo-tagged tweets based on ST-DBSCAN is focused in Huang et al. (2018). After summarization of the word frequencies for each cluster, potential topics are modelled by latent Dirichlet allocation (LDA). ST-DBSCAN has shown to be a good method for the detection of several different small-scale event types, for example gunshot events. However, a careful selection of the clustering parameters is required.

An approach based on machine learning and geo-visualization to identify events in cities and trace the development of these events in real-time based on Twitter data is presented in Zhou and Xu (2017). The stream data are pre-processed and grouped into one-hour intervals. In order to find tokens related to events, bursty word detection techniques are applied. DBSCAN is utilized for spatial clustering and random forests are trained to classify tokens as event-related or unrelated. Due to the low proportion of geo-tagged tweets within the processed stream, only major local events could be detected.

A statistical approach for studying the spatio-temporal distribution of geo-located tweets in urban environments is proposed in Santa, Henriques, Torres-Sospedra, and Pebesma (2019). Instead of detecting and characterizing unique events, repeated patterns are identified by utilizing negative binomial regression analysis, PCA and hierarchical clustering. This approach may help to characterize the usual Twitter usage behaviour as a starting point for anomaly or event detection.

In Liang, Lin, and Peng (2018), a method for discovering the spatio-temporal process of a typhoon using Sina Weibo microblog data is presented. Support vector machines are utilized for classifying the text messages into four crisis-related classes. Further spatio-temporal analyses are based on searching keywords. The authors point out, that more research is required to determine the most appropriate scale for spatial analyses.

In Resch et al. (2018), machine learning topic models (LDA) and spatio-temporal analysis (local spatial autocorrelation) of social media data are combined for hotspot detection, disaster footprint estimation and damage assessment. LDA is applied to a Twitter data stream in a cascading fashion in order to extract earthquake- and damage-related tweets. After statistical topic validation, hotspot analysis is conducted based on local spatial autocorrelation (*Getis-Ord G**).

In Martin, Li, and Cutter (2017), the spatio-temporal variability in social media response is examined and an approach to leverage geo-tagged tweets to assess the evacuation responses of residents is developed. The approach involves (keyword- and location-based) retrieval of tweets, creation and filtering of data sets, and statistical and spatial processing to extract and map the results. Spatial and temporal analyses, which are done separately in this study, are mainly based on counting related tweets (per time and/or per county). Evacuation-related analyses, that is how many Twitter users were evacuated, their evacuation destination and return date, involve the identification and tracking of active local users during different event periods.

In contrast of directly using Twitter data, a framework for discovering evolving domain-related spatio-temporal patterns is proposed in Shi et al. (2016). Given a target domain, a dynamic query expansion is employed to extract related tweets which are then used to form spatio-temporal Twitter events. Spatial clustering is based on the use of multi-level constrained Delaunay triangulation to capture the spatial distribution patterns of Twitter events. An additional spatio-temporal clustering process is then performed to reveal spatio-temporal clusters and outliers that are evolving into spatial distribution patterns.

A method for spatio-temporal anomaly detection through visual analysis of geo-located Twitter messages is proposed in Thom, Bosch, Koch, Woerner, and Ertl (2012). The approach enables the interactive analysis of location-based microblog messages in real-time by means of scalable aggregation and geo-located text visualization. For this purpose, a novel cluster analysis approach is used to distinguish between local event reports and global media reaction to detect spatio-temporal anomalies automatically. The main tool for data analysis and visualization is the so-called ScatterBlogs system. Its purpose is to enable analysts to work on quantitative and qualitative findings by not only automatically identifying anomalies, but also summarizing and labelling event candidates, and providing interaction mechanisms to examine them.

In Hu, Wang, Guin, and Zhu (2018), an STKDE-based framework for predictive crime hotspot mapping and evaluation is proposed. A statistical significance test is designed to filter out false positives in the density estimates. Furthermore, a new metric is proposed to evaluate predictive hotspots at multiple scales.

3 | PROPOSED WORKFLOW

The main goal is to reconstruct and picture the course of events during natural disasters based on Twitter data. The workflow and the involved methods therefore should have the following properties:

- The workflow should be application- and event-independent
- The analysis of geo-located tweets posted within the affected area is focused
- A complete tool chain ranging from data acquisition to data visualization is desired
- Both, the processing of full crisis data sets and the incremental processing of incoming streamed data, are focused
- The utilized methods shall provide a multifaceted, summarizing and also detailed picture of (sub-) events that happen in the affected area
- Involved methods should be flexible, computationally tractable, applicable to short messages and successfully applied in other related works

Our resulting Twitter data stream analysis workflow is depicted in Figure 1. Due to the sequential structure, the process can either be applied to a full data set or in a repeated manner, for example to hourly analyse the latest incoming data. Data acquisition is done by using the Twitter Streaming API. Capturing reactions from affected individuals is ensured by restricting crawling to a specified area of interest. The first step of data analysis is the filtering of crisis-related tweets. The tweets identified as crisis-related can directly be used for spatio-temporal and topic analyses. In case of spatio-temporal analysis, clustering techniques are applied in order to group posted messages with respect to time and space. Topic modelling is applied to automatically identify the discussed semantic topics in a set of messages, for example posted during one day or within a spatio-temporal cluster. Each identified topic is characterized by a set of describing keywords that serve as a summarization. Each message is assigned to multiple topics based on probabilities.

In parallel, the tweets can further be classified into crisis-related information classes, such as *Donations and Volunteering*, *Infrastructure and Utilities* and *Affected Individuals*. The latter class is an example for potentially actionable tweets that might directly trigger an alert mechanism that should be confirmed and further processed by a human interpreter.

Spatio-temporal analysis is conducted by applying methods for local hotspot detection, density estimation and clustering. The results can be visualized on interactive maps, map sequences and 3D plots. Furthermore, motivated by Alam, Ofli, Imran, and Aupetit (2018) and Thom et al. (2012), the class distribution within clusters and word clouds are useful tools to characterize and visualize tweet contents and the changes over time. The results from topic modelling methods applied to all crisis-related tweets (in contrast to cluster-wise analysis) is intended to provide a complementary view on the data.

Tweets not related to natural disasters are discarded in this work. Following the approach in Santa et al. (2019), these tweets might be useful to characterize the usual local twitter behaviour when no event takes place and therefore may be utilized to support event and anomaly detection in future works.

In the following, the choice of the workflow design and the identified methods for data analysis are described in detail.
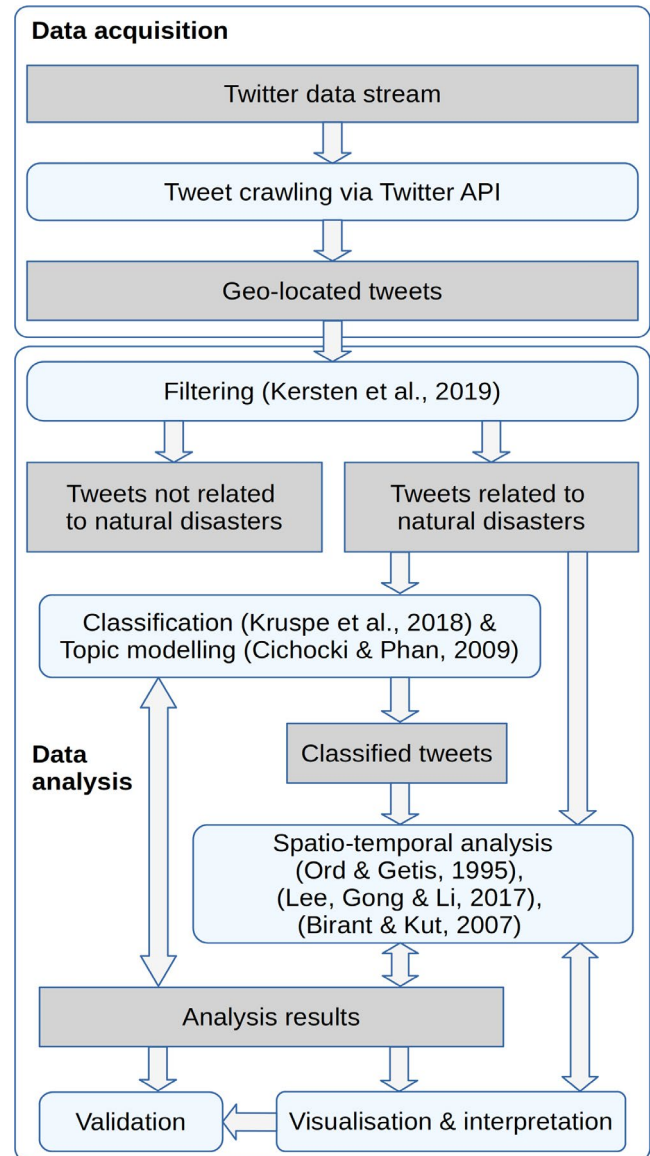

FIGURE 1 Proposed workflow [Colour figure can be viewed at wileyonlinelibrary.com]

3.1 | Tweet filtering

According to Shi et al. (2016), we can distinguish between the analysis of unfiltered and of domain-related Twitter data. By focusing on the domain of natural disasters, that is, by filtering crisis-related tweets, the amount of data to analyse in the subsequent steps is likely to be reduced significantly. We therefore favour a domain-related approach in this work. In Zhao et al. (2014), dynamic query expansion is utilized for the extraction of domain-related tweets. In Resch et al. (2018), LDA is applied in a cascading fashion in order to identify messages related to earthquake events and damages. The use of recently proposed deep learning methods for extracting either disaster-related tweets (Burel & Alani, 2018; Kersten, Kruspe, Wiegmann, & Klan, 2019), or more specifically, event-related Tweets (Kruspe, Kersten, & Klan, 2019) have shown to be a promising alternative. A convolutional neural network (CNN) model (Kersten et al., 2019)

is therefore utilized in this work. With this approach, where filtering is understood as a binary classification problem, superior results compared to other methods, such as Naïve Bayes, support vector machines (Burel & Alani, 2018), random forests and logistic regression were obtained (Nguyen et al., 2017). Implementation was done in Python 3.6 by utilizing tensorflow[1] and keras.[2]

Each message is pre-processed with a set of standard operations, that is lowercasing, multiple space replacing, normalizing URLs, usernames and hashtags, removing special characters, normalizing digits and tokenization. Each word is then transformed into a real-valued vector of 300 elements by using a pre-trained word embedding specifically trained on 52 million crisis-related Twitter messages (Imran, Mitra, & Castillo, 2016). Hence, each tweet is represented as a matrix of size $n \times k$, where n is the maximum number of words per tweet occurring in the training data and $k = 300$ is the embedding vector length. In case the number of words is lower than n, zero padding is applied. Then, a convolutional layer applies m kernels of different widths w in parallel to the input matrix. We use standard values $m = 128$ and $w = (3, 4, 5)$, because parameter variation did not affect the results significantly. Global max pooling is performed after each of these convolutional layers, and the results are concatenated to a vector. This new embedding is then fed into a fully connected layer to determine the final class.

3.2 | Tweet classification and topic analysis

The application of state-of-the-art methods to classify crisis-related tweets into information classes (Alam, Imran, & Ofli, 2019; Alam et al., 2018; Burel, Saif, Fernandez, & Alani, 2017; McCreadie & Soboroff, 2018; Nguyen et al., 2017) provides additional valuable information. In particular, this is a reasonable way to identify actionable messages that should immediately be forwarded to involved local disaster relief organizations. The distribution of classes may also provide insights regarding the discussed topics within spatio-temporal clusters. A CNN model proposed in Kruspe, Kersten, Wiegmann, Stein, and Klan (2018) is therefore applied here. Despite the fact that the number of nodes of the final output layer corresponds to the number of thematic classes, the model is identical to the one used for filtering. The model was trained using two different class definitions: CrisisLexT26 (Olteanu, Vieweg, & Castillo, 2015) and TREC-IS 2018 (McCreadie & Soboroff, 2018), comprising 7 and 25 information classes, respectively.

Topic models, such as latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF), are used to discover the underlying semantic structure of a corpus of documents by identifying and quantifying the importance of representative topics throughout the documents (MacMillan & Wilson, 2017). As recently demonstrated in Resch et al. (2018) and Casalino et al. (2018), these are useful approaches to reliably identify specific topics discussed on Twitter by avoiding explicit classification. We decided to utilize an own Python 3.6 implementation of NMF using the sklearn[3] library in this work, since the NMF results have shown to offer a high

degree of understandability of the extracted topics in several application areas such as bioinformatics, pattern recognition, image analysis, educational data mining and document clustering (Casalino et al., 2018). In contrast to use topic modelling for filtering a specific subset of messages (e.g. messages related to specific event types), we use it to reveal insights on the discussed topics and their spatio-temporal distribution within the fraction of crisis-related tweets.

As a starting point, the corpus (in our case the crisis-related tweets) is encoded in a term-document (or term-tweet) matrix, whose rows correspond to n terms in a vocabulary and whose columns are related to m tweets. NMF (Cichocki & Phan, 2009) basically is a dimensionality reduction technique which can be applied to decompose this matrix into two low-rank factor matrices constrained to have only non-negative elements and such that the product of these matrices approximates the term-tweet matrix. The rank k defines the number of hidden topics and has to be defined manually. Compared to other traditional dimensionality reduction methods, NMF is able to uncover latent low-dimensional structures intrinsic in high-dimensional data and provides a non-negative, part-based, representation of data enhancing meaningful interpretations of mined information (Casalino et al., 2018).

3.3 | Spatio-temporal analysis

The filtered and classified tweets can then be further analysed by taking into account space and time. Since local hotspot analysis and STKDE and (ST-) DBSCAN clustering have shown to provide reasonable results in different applications related to tweet analysis (Hu et al., 2018; Huang et al., 2018; Lee et al., 2017; Resch et al., 2018; Zhou & Xu, 2017), they were identified as suitable methods for this work. Message content is not taken into account with these approaches. However, the tweet labels obtained by the CNN classifier can be used to visualize the content of the clusters. Furthermore, a visualization of the discussed topics in terms of word clouds (see for example (Thom et al., 2012) and (Casalino et al., 2018)) is applied. In the following, the selected methods are described in detail.

3.3.1 | Local hotspot detection

The goal of applying the *Getis-Ord G**-statistics is to detect local coldspots and hotspots of crisis-related Twitter activity. This is accomplished by clustering high and low tweet occurrences and measuring the concentration of occurrences in specific areas. We used the tweets posted within congressional districts for our experiments. A JavaScript implementation[4] and the open-source Geographic Information System Software QGIS[5] for pre- and post-processing and visualization was utilized in this work.

The *Getis-Ord G**-statistic measures the degree of association that results from the concentration of weighted points (or area represented by a weighted point) and all other weighted points included within a radius of distance d from the original weighted point (Ord

& Getis, 1995). Hence, it is a tool for hotspot and cold-spot analysis, in which high and low values are clustered and the concentration of these values in a specific area is measured. A simple definition of the G^*-statistic of a specific point i is given by

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j}{\sum_{j=1}^n x_j} \tag{1}$$

where x_j is the attribute value (e.g. number of tweets) for feature j, $w_{i,j}$ is the spatial weight between entity i and j, and n is the number of entities (e.g. points or areas). G_i^* is usually standardized by the sample mean X and the sample standard deviation S:

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - X \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{\left[n \sum_{j=1}^n w_{i,j}^2 - \sum_{j=1}^n w_{i,j}^2 \right]}{n-1}}} \tag{2}$$

with

$$X = \frac{\sum_{j=1}^n x_j}{n} \tag{3}$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - X^2} \tag{4}$$

An entity with a high value might be interesting, but may not be a statistically significant hotspot in terms of the G^*-statistic. To be a statistically significant hotspot, an entity will have a high value and be surrounded by other features with high values as well. The local sum for an entity and its neighbours is compared proportionally to the sum of all entities. When the local sum is much different than the expected local sum, and that difference is too large to be the result of random chance, a statistically significant G_i^* value (Z-score) results. G_i^* values around zero indicate a random distribution.

3.3.2 | STKDE density estimation

In contrast to the G^*-statistic, space–time kernel density estimation (STKDE) might be directly applied to data points in the spatiotemporal domain. Given a spatial distribution of geographic events or points in a region, KDE methods compute a density surface that describes the intensity of the geographic events' distribution. KDE can be understood as a simple adding of basic density functions, for example Gaussians, located at each event location. Intensities of spatial distributions of point data are estimated in the following way:

$$f(x,y) = \frac{1}{nh_s^2} \sum_{i=1}^n k_s \left(\frac{x-x_i}{h_s}, \frac{y-y_i}{h_s} \right) \tag{5}$$

where n is the number of data points, x_i and y_i define the location of point i, h_s is the bandwidth of the kernel density function, and k_s represents the kernel density function.

Space–time kernel density estimation (STKDE) is an extension of the traditional KDE, used for identifying spatio-temporal patterns of underlying data sets through visualizing the resulting densities in the space–time cube. The output of STKDE can be expressed as a 3D raster volume, where for each voxel a density estimate based on the surrounding point data is assigned. The corresponding density values can be computed with the following equation:

$$f(x,y,t) = \frac{1}{nh_s^2 h_t} \sum_{i=1}^n I\left(d_i < h_s, t_i < h_t \right) k_s \left(\frac{x-x_i}{h_s}, \frac{y-y_i}{h_s} \right) k_t \left(\frac{t-t_i}{h_t} \right) \tag{6}$$

where the density f is computed based on all points (x_i, y_i, t_i) for which the spatial distance d_i and the temporal distance t_i to the current point is lower than the thresholds h_s and h_t, respectively. Hence, the indicator function I returns 0, if one of these thresholds is exceeded and 1 otherwise. For the kernel functions k_s and k_t, the Epanechnikov kernel (Epanechnikov, 1969) is applied, where each data point is weighted according to its distance in time (t_i) and space (d_i) to the current voxel (the closer the data point, the higher the weight).

Since information regarding the population is not taken into account with this approach, the resulting density estimates picture absolute occurrences of crisis-related tweets. This is a useful information that can support disaster (sub-) event detection. In case a normalization by population density is desired, the dual KDE approach proposed in Wang, Ye, and Tsou (2016) might be applied instead.

3.3.3 | ST-DBSCAN clustering

ST-DBSCAN (Birant & Kut, 2007) is a clustering approach based on DBSCAN (density-based spatial clustering of applications with noise) (Ester et al., 1996). In DBSCAN, the density associated with a point is obtained by counting the number of points in a region of specified radius around the point. Points with a density above a chosen threshold are constituted to clusters. Compared to so-called parametric approaches, such as Gaussian mixture models, no assumption about the cluster shapes is introduced. Hence, DBSCAN has the ability to discover clusters with arbitrary shape such as linear, concave and oval. Furthermore, the number of clusters is estimated. ST-DBSCAN can cluster spatio-temporal data according to non-spatial, spatial and temporal attributes. Besides a distance measure (e.g. Euclidean distance) between data points, a second similarity measure is introduced in order to describe non-spatial similarity between attributes, such as temperature or time. To speed up computations, similarity measures are only computed for temporal neighbours. If a point does not contribute to any of the clusters, that is the number of neighbours is lower than a threshold, it is considered to be a noise point. We used an open-source Python implementation[6] of ST-DBSCAN in this work.

4 | EXPERIMENTAL RESULTS

In this section, the proposed workflow is applied and evaluated in detail. Besides tweet filtering and classification, special emphasis is put on spatio-temporal analysis. For a comparison of the obtained insights regarding local and general impacts of hurricane Florence on the population located in the defined area, reference information from the freely accessible news databases GDELT (GDELT Project, 2019), Europe Media Monitor (EMM) (Europe Media Monitor, 2019) and Wikipedia (Wikipedia, 2019), and situation reports published by Humanity Road (Humanity Road, 2019) is used. In the following, the used data, conducted experiments and results of the three main processing steps of the proposed workflow, that is filtering, classification and topic modelling as well as spatio-temporal analysis are described.

4.1 | Data set

Hurricane Florence, the wettest tropical cyclone on record, made landfall as a category 1 hurricane on early 14 September, 2018. Due to weather forecasts predicting the trajectory of the hurricane to pass the area of North and South Carolina, a corresponding area (see Figure 2) was defined for Twitter data acquisition. From September 12 to 19, around 600,000 geo-located tweets from this area were acquired.

Since only for a small fraction of Twitter data a geo-location is available, the location-based crawling introduces bias by neglecting the significantly larger fraction of information without geo-location. On the other hand, filtering by keywords might be problematic as well, since unrelated messages containing the keywords will be

retrieved, while important messages not containing any of these will be discarded. However, location-based filtering is required in order to provide tweets from directly affected individuals, whereas tweets from users who are not directly involved, but contribute to discussions, are discarded.

4.2 | Tweet filtering

The CNN model was trained with a balanced set of around 120,000 crisis-related and unrelated tweets covering various types of natural and man-made disasters. The resulting temporal distribution of related and unrelated tweets binned into intervals of two hours is shown in Figure 3. A daily tweet activity pattern with maximum values of around $10k – 12k$ until around midnight can be observed. 14.9% (around 30,000) of all tweets (around 600,000) are classified as related to a disaster event. Interestingly, the daily peaks of disaster-related tweets are shifted and can be observed around and after noon.

It was pointed out in Kersten et al. (2019) that an average F1 score of 0.83 ± 0.13 in case of unseen events can be expected. Furthermore, an average precision and recall of 0.80 ± 0.16 and 0.88 ± 0.08 for related tweets were obtained in their experiments. For an independent evaluation, a subset of 2,000 tweets classified as related by the CNN was manually labelled. According to these labels, the CNN achieved a precision of 0.7 indicating that 30% of the actually unrelated tweets are misclassified as related. Even though this demonstrates that pre-trained models should further be enhanced in terms of robustness, this behaviour is favoured over a lower recall, where related and therefore potentially important tweets might be omitted.
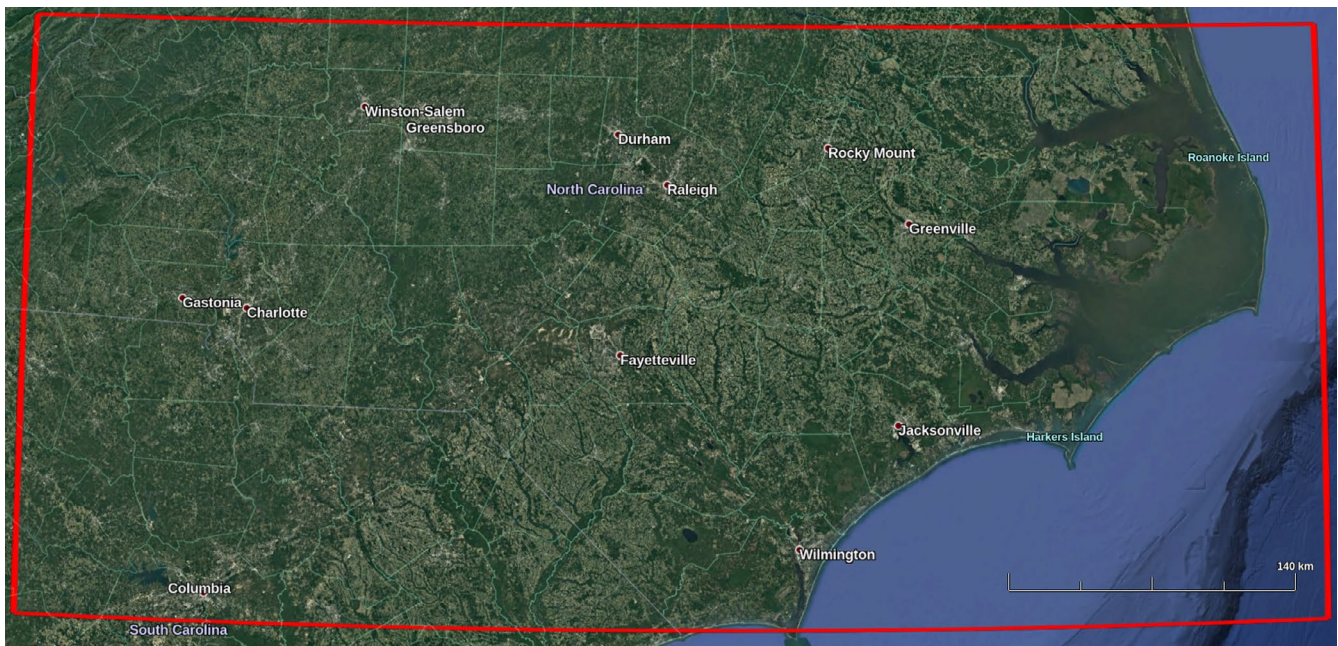

FIGURE 2 Area of interest for Twitter data acquisition during hurricane Florence (Google Maps, 2019) [Colour figure can be viewed at wileyonlinelibrary.com]
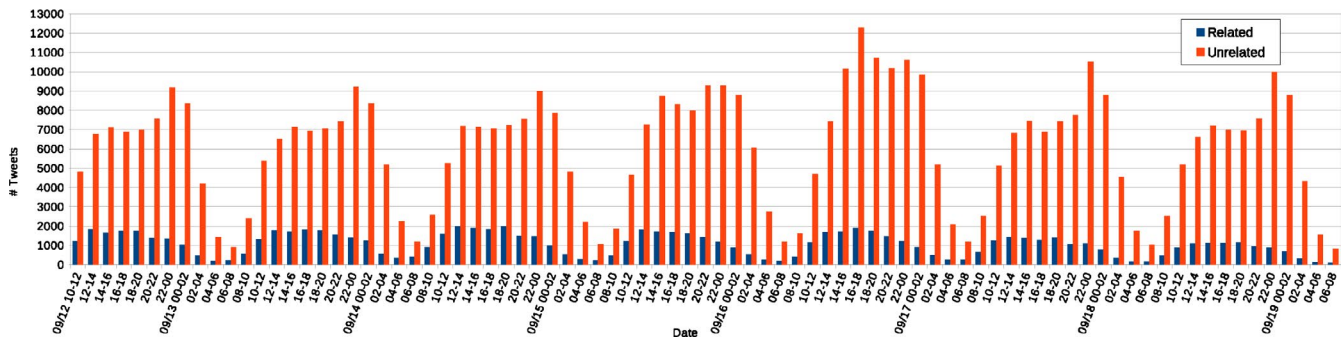
FIGURE 3 Filtering results during hurricane Florence, 12–19 September 2018 [Colour figure can be viewed at wileyonlinelibrary.com]

As a further simple test of the model, we applied it to Twitter data from a non-disaster period. 8.3% of all 526,000 tweets acquired within the same area of interest in 8 January – 15 January 2019 were classified as related, which is a quite satisfactory result. A manual inspection of Tweets classified with a high confidence (>0.9) revealed that general crisis-related aspects and the heavy 2018 California wildfires were discussed. However, most of these tweets turned out to be false positives.

The filtering approach was recently enhanced and further investigated in terms of cross-event and cross-type applications (Wiegmann, Kersten, Klan, Potthast, & Stein, 2020). Here, an average misclassification rate of 3,8% for 5 million unrelated tweets was achieved.

4.3 | Tweet classification

Tweet classification is also accomplished by utilizing CNN models (Kruspe et al., 2018). Two different information class definitions were used. CrisisLexT26 consists of seven rather general classes, such as *Infrastructure and Utilities*, *Caution and Advice*, and *Sympathy and Support*, whereas TREC-IS 2018 comprises 25 more fine-grained classes, such as *SearchAndRescue*, *Volunteer* and *EmergingThreats*. For both class definitions, the validation accuracy after training with 10-fold cross validation was around 0.65. Regarding the TREC classes, F1 scores of around 0.55 were reported from independent evaluations conducted by the TREC organizers. Even though this seems not to be a good result at a first glance, one of the proposed models provided the best results in terms of F1 measure for actionable classes among all submissions to the TREC 2018 Incident Streams competition (McCreadie, 2019).

4.4 | Spatio-temporal analysis

In this section, the results obtained with local hotspot detection, STKDE density estimation, ST-DBSCAN clustering and NMF topic modelling are discussed.

4.4.1 | Local hotspot detection

As pointed out in Resch et al. (2018), useful spatial information cannot be derived without taking into account the population density. We therefore normalized the number of tweets in each district with the number of inhabitants per square kilometre available in (ESRI, 2018). To be a statistically significant hotspot, an area will have a high normalized tweet occurrence and be surrounded by other areas with high values as well. The local sum for an area and its neighbours is compared proportionally to the sum of all areas. When the local sum is significantly different than the expected local sum, a statistically significant Z-score results.

In our experiments, we determined the G^{*}-statistics with a spatial bandwidth of 50 and 100 km. The daily results of this analysis are depicted in Figure 4. Interestingly, the area covering Raleigh, Durham, Cary and Burlington are identified as hotspots every day, indicating a comparable high crisis-related Twitter activity in this area. The area around Charlotte has the highest population density and was identified as hotspot on 16 September. According to official news, Florence "slogged" through Charlotte over the weekend (15 and 16 September) as a tropical depression with wind speeds of around 35 *mph*. As an effect of heavy and continuing rainfalls, greenways, backyards and low-lying streets across the city were flooded. Similar to this, a correlation between the hurricane landfall on September 14 and increased Twitter activities in the coastal areas around Wilmington and Jacksonville can be observed.

The two maps on the lower right side in Figure 4 show the overall results by analysing the data of the whole week. A significant impact of the chosen maximum analysis radius can be observed here. For a radius of 50 km the hotspots tend to highlight the districts that actually show up a high Twitter activity. With a radius of 100 km, these areas appear to be much larger and tend to overrule or influence neighbouring districts.

Since this method is applied to all crisis-related messages, only general trends regarding the relative occurrences of these messages can be observed here. By observing the development over time, we could demonstrate that this is a useful approach to localize changes in crisis-related twitter activities. Applying the method to messages related to more specific topics, such as damages or donations and
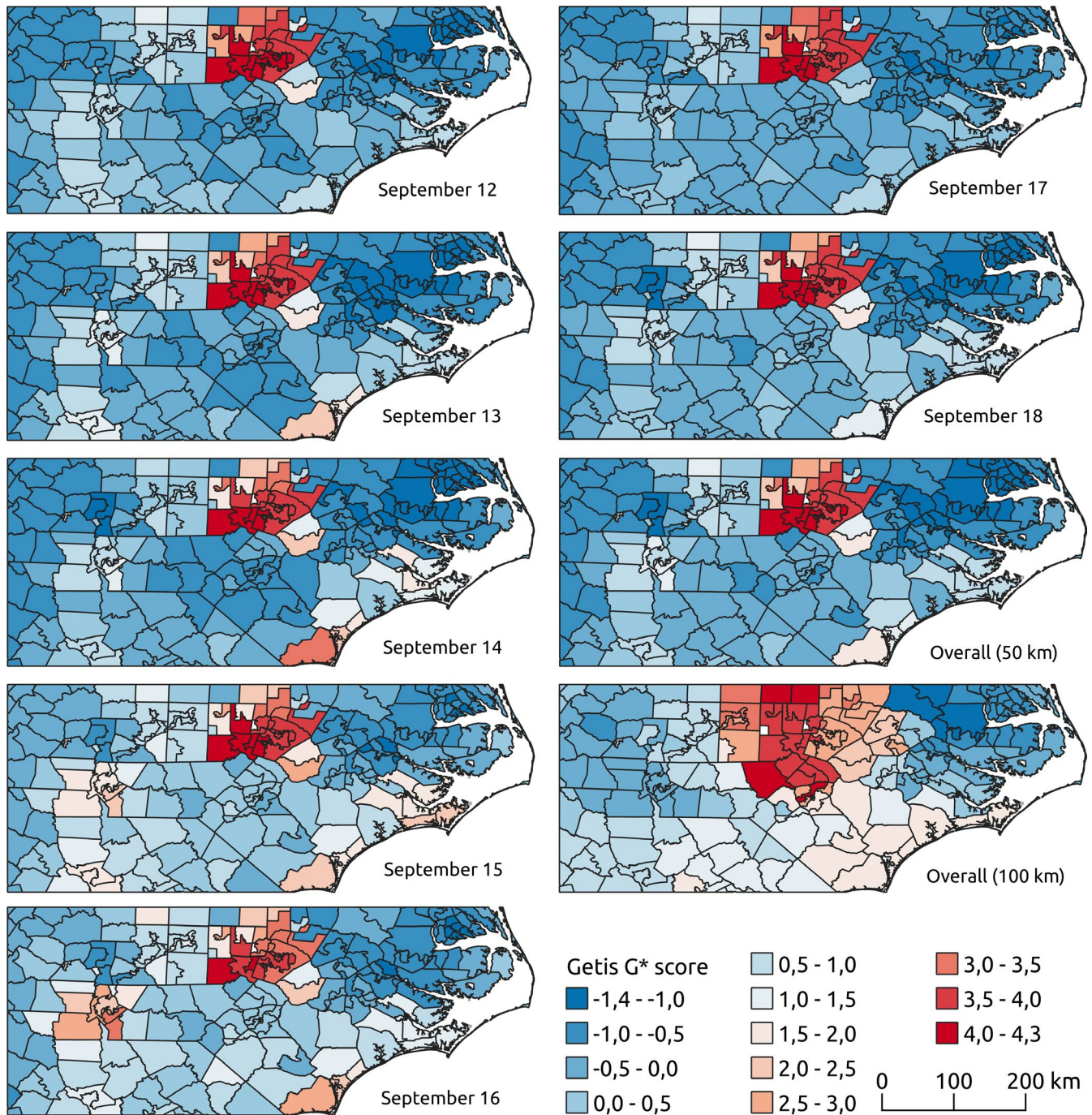
FIGURE 4 Daily and overall hotspot analysis results obtained with a maximum distance of $max_d = 50$ km as well as an overall result with $max_d = 100$ km (lower right) [Colour figure can be viewed at wileyonlinelibrary.com]

volunteering, might be quite useful to gain more specific information (see for example Resch et al., 2018).

4.4.2 | STKDE density estimation

In STKDE, densities are estimates for each point in a regular grid with a chosen spatial resolution of 5 km and a temporal resolution of 0.1 d. Optimal values of $h_s = 7.5$ km and $h_t = 0.3$ d for the spatial and temporal analysis bandwidth were empirically identified.

A 2D view of the results is shown in Figure 5. According to the estimated density, each point is colour-coded and adjusted in size. As expected, the areas around Charlotte and Raleigh are identified as the densest crisis-related tweet activity areas. Less dense areas correspond to less densely populated areas, for example Durham, Fayetteville, Columbia and Wilmington. It is worth noting that the 2D view in Figure 5 does not represent the result of a 2D analysis in which all disaster-related tweets from different points in time are projected into the 2D plane. A 3D view on the same STKDE result in the spatio-temporal domain is depicted in Figure 6. In this
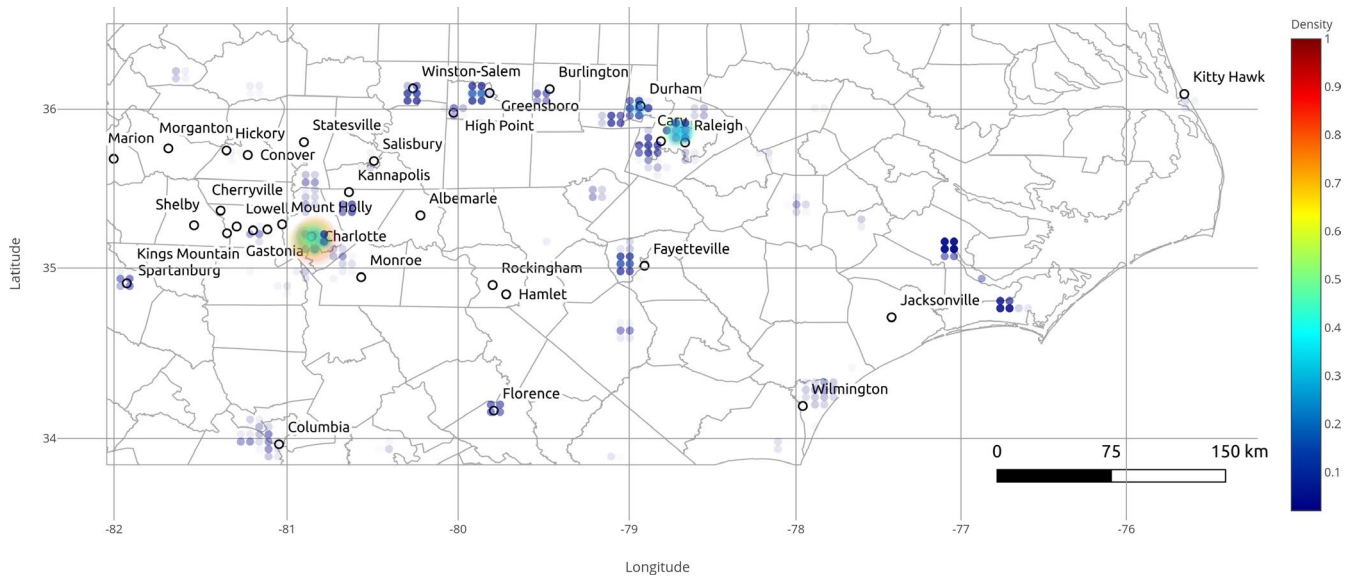
FIGURE 5 2D view of STKDE results derived from crisis-related Twitter activities during the hurricane (spatial bandwidth: 7.5 km, temporal bandwidth: 0.3 d, spatial resolution: 5 km and temporal resolution: 0.1 d). The map is based on OpenStreetMap data (OpenStreetMap, 2019) [Colour figure can be viewed at wileyonlinelibrary.com]
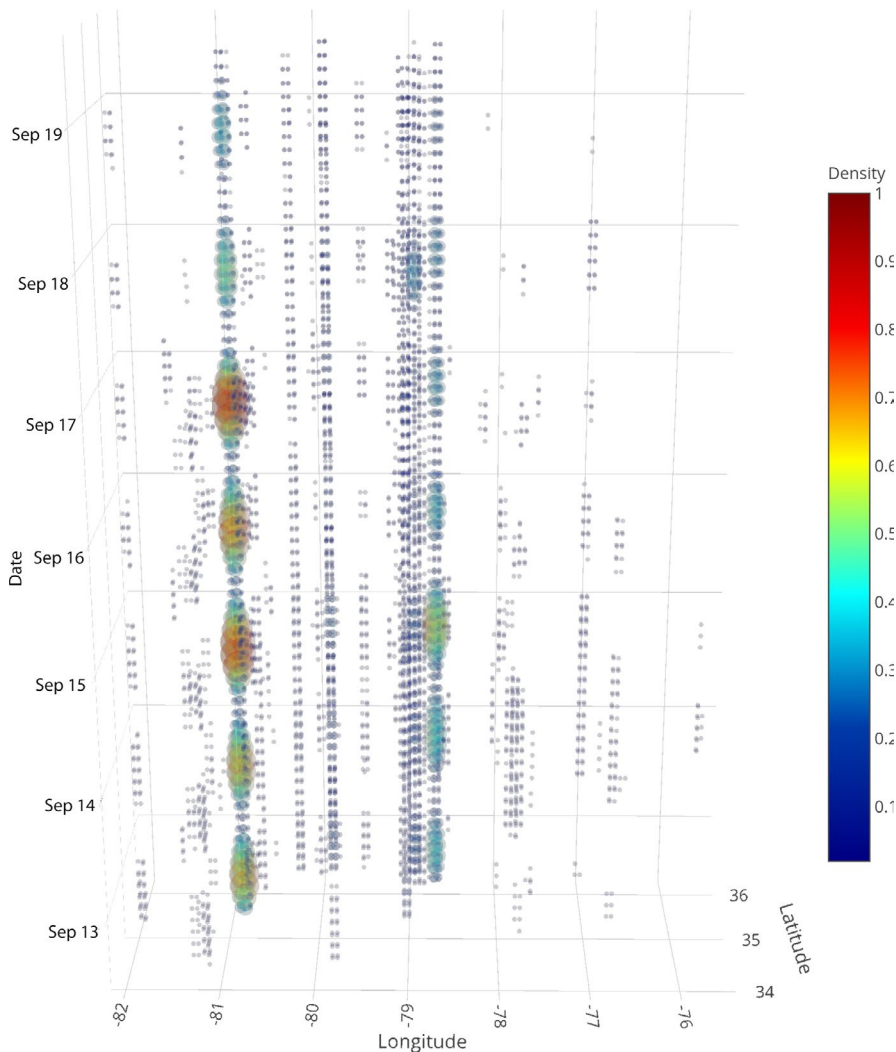

FIGURE 6 3D view of STKDE results derived from crisis-related Twitter activities during the hurricane (spatial bandwidth: 7.5 km, temporal bandwidth: 0.3 d, spatial resolution: 5 km and temporal resolution: 0.1 d) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 7 3D view of ST-DBSCAN results derived from crisis-related tweets (spatial threshold $t_s = 5$ km, temporal threshold $t_t = 0.5$ hr and minimum number of tweets $min_t = 5$). Each coloured dot represents a cluster and is scaled according to the number of tweets contributing to this cluster [Colour figure can be viewed at wileyonlinelibrary.com]
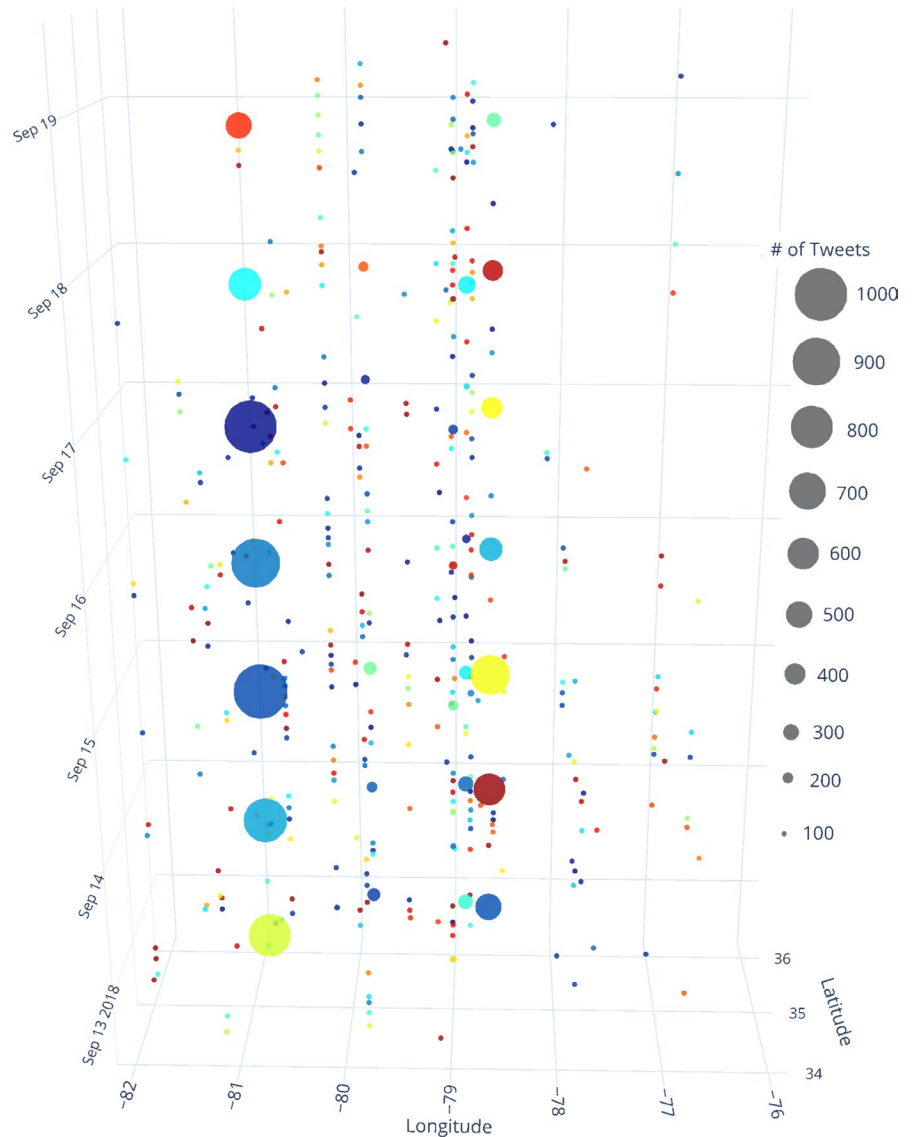
plot, spatio-temporal clusters corresponding to specific regions as well as daily periods can be observed. The hurricane's landfall occurred in the morning of September 14. During the two days before the landfall, approximately constant daily tweet activity patterns can be observed. Around the landfall time, a significant increase of crisis-related tweets can be observed. From this point on, the amount of tweets slowly decreases towards the end of the recording period. A second peak of tweet activity can be observed on 16 September, in Charlotte confirming the results obtained with hotspot analysis. A further increased tweet activity is visible in the Durham area on 17 September. This corresponds to reported flooding in this area (e.g. Durham-Chapel Hill Blvd and Northgate Park) as well as official tornado warnings for this region. The results demonstrate that STKDE is a quite useful approach for visualizing and analysing spatio-temporal distributions of crisis-related Twitter activities. However, similar to the G^*-statistic, rather general trends can be observed, if crisis-related tweets are taken into account. Applying STKDE to tweets representing more specific topics is a useful option but might introduce challenges due to data sparsity.

4.4.3 | ST-DBSCAN clustering

So far, rather high-level and broad event developments in space and over time could be observed. According to Walther and Kaisser (2013), specific real-world events (as part or sub-event during major events) can often be expected on a rather small-scale, meaning that they happen at a specific place in a given time period, and are often covered by only few tweets. In order to capture these types of (sub-) events, ST-DBSCAN is applied on our set of crisis-related tweets.

A colour-coded scatter plot of cluster centres obtained with a spatial threshold of $t_s = 5$ km, a temporal threshold of $t_t = 0.5$ hr and a minimum number of tweets $min_t = 5$ is depicted in Figure 7. Each of the 385 identified clusters is represented by a circle with a randomly chosen colour, where the circle size depends on the number of tweets contributing to a cluster. At first sight, the ST-DBSCAN clusters correspond to the densities of the STKDE results in Figure 6. Dense STKDE areas, in which a lot of tweets were posted within a specific time range, also exhibit large ST-DBSCAN clusters. The crucial difference between STKDE and ST-DBSCAN is that the latter

assigns each tweet to a cluster, whereas STKDE provides a density value for each point in space and time.

High Twitter activities around Charlotte and Raleigh lead to a single large ST-DBSCAN cluster for each day and region. For all other regions, smaller clusters with significantly fewer tweets, shorter time spans and areas are identified. This implies that the choice of ST-DBSCAN parameters is difficult in case of irregular point distributions. Clusters with a few tweets might be related one specific topic whereas it can be expected that a large cluster, representing all tweets posted during one day within a city, covers several different topics. This motivates further analyses of the clusters and messages, which is described in the next section.

4.4.4 | Cluster analysis and interpretation

Two different approaches are chosen here: the visualization of class distributions and the visualization of word clouds for each cluster. ST-DBSCAN identified 385 spatio-temporal clusters. Around 300 clusters contain only few tweets (around 10) and 100 clusters are rather large (approximately 100–1,000 tweets). In Figure 8, the CrisisLex information class distributions for the ten largest clusters, comprising around 500–4,000 tweets, are depicted. All tweets identified as noise by ST-DBSCAN (around 13,500) are summarized in the cluster Noise, whereas the average category distribution of the ten valid clusters is denoted as Average. An interesting observation is that the noise and average distributions are quite similar, indicating that potentially important information is not taken into account when the Noise cluster is neglected. Also, the distribution of the largest clusters (4 and 8) is quite similar to the average distribution and therefore does not provide detailed insights. In contrast, the per-cluster distributions for smaller clusters turn out to be different.

For instance, the number of messages expressing sympathy and support (see cluster 61 with 556 tweets in Figure 8) has a peak on 14 September, the day of the landfall. After the landfall, more messages related to donations and volunteering (cluster 14 and 28) as well as infrastructure and utilities and affected individuals (cluster 11) can be observed. More significant changes, for example the absence of one or even more of the information classes cannot be observed here.

Similar to class distribution histograms, also word clouds visualizing the most frequent terms tend to be quite unspecific in case of large clusters. However, especially for clusters containing only a few messages, class distribution histograms are not appropriate. Instead, word clouds turn out to be quite useful and informative in this case. Instead of reading messages, the word clouds immediately provide an overview about the frequently used terms. Six selected example word clouds are shown in Figure 9. Interestingly, we found that the small clusters can often be connected to real happenings, such as flood and tornado warnings, road and beach closures, power outages, damages, fallen trees and power lines as well as the need or availability of donations in a specific area. As shown in Figure 9 (f), also virtual events, such as a critical discussion regarding the handling of outdated emergency posts and news prioritizing emergency news sources, are found.

4.5 | NMF topic modelling

A disadvantage of ST-DBSCAN is that, depending on their spatio-temporal distribution, a lot of messages may not be assigned to a cluster. In our case, 13,576 noise points (approximately 1/3 of all crisis-related tweets) potentially containing important crisis-related information are not taken into account. We therefore additionally
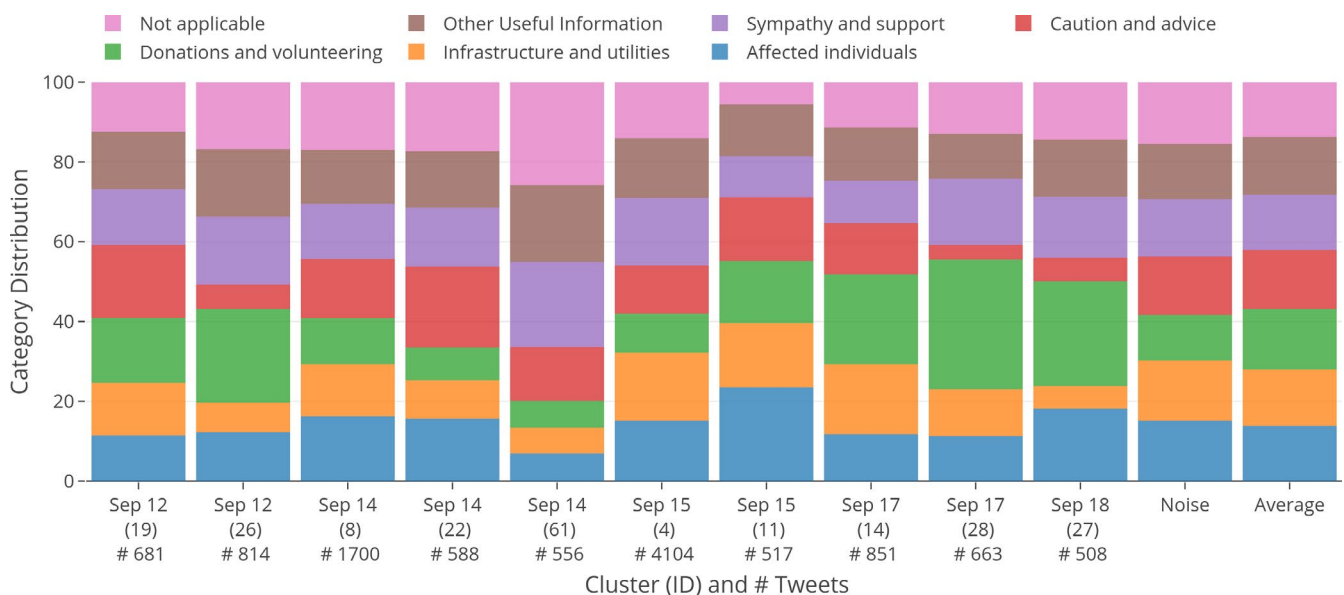

FIGURE 8 CrisisLex class distributions for the ten largest ST-DBSCAN clusters [Colour figure can be viewed at wileyonlinelibrary.com]
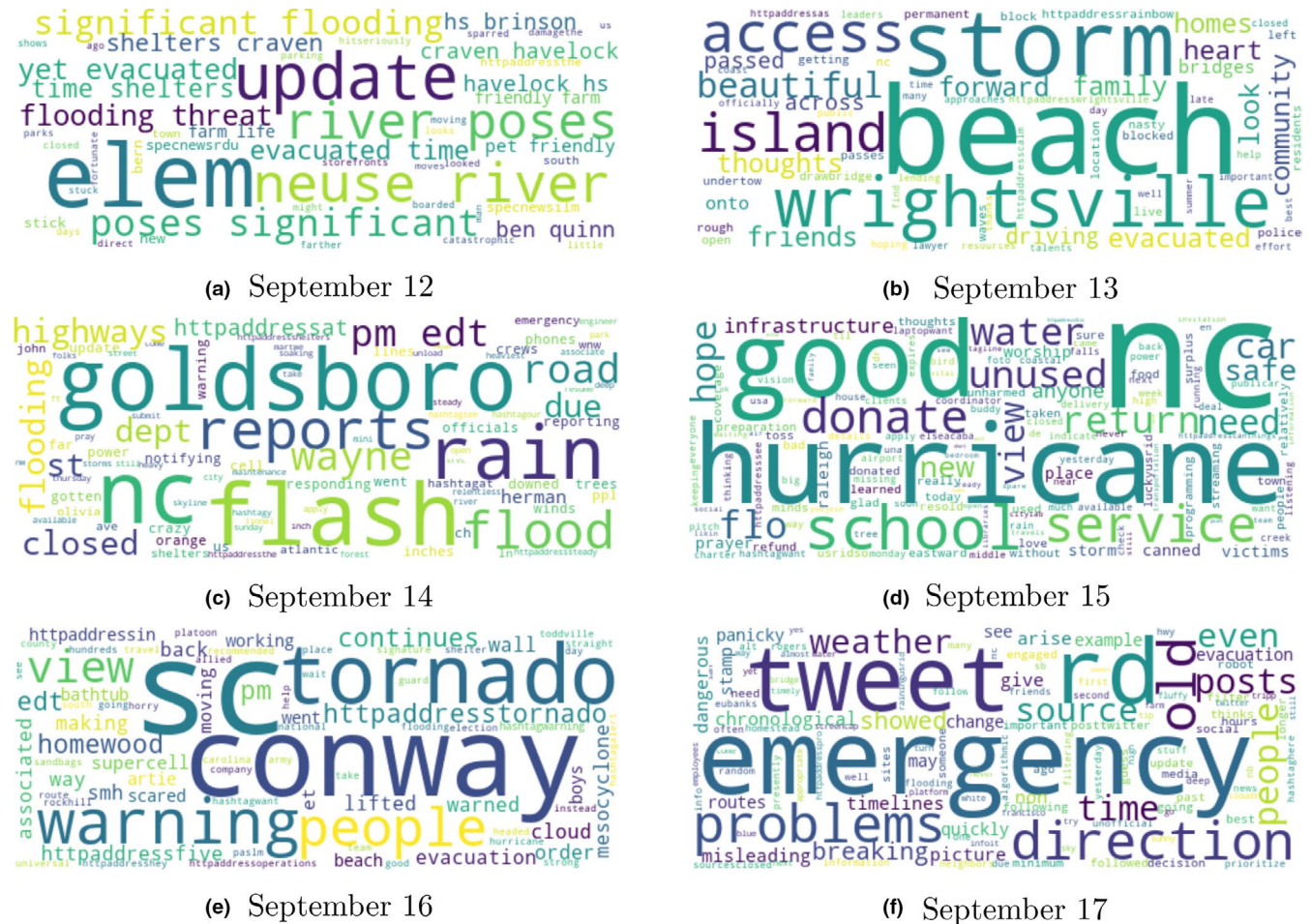
FIGURE 9 Six example word clouds representing ST-DBSCAN clusters. (a) New Bern: Neuse River flood warnings, call to evacuate, shelters available; (b) Wrightsville beach closed; (c) Goldsboro: Flood warnings, roads and highways closed due to flash floods; (d) Cary: Where to donate unused goods, like water; (e) Conway: Tornado warnings; (f) Carrboro: Critical discussion about Twitter [Colour figure can be viewed at wileyonlinelibrary.com]

propose to apply topic modelling to all crisis-related tweets. Mapping and visualizing the locations of mentioned topics in specific time ranges provides a complementary view on the data and may support to investigate the distribution of discussed topics in space and time.

Before tweet encoding, we applied some standard pre-processing steps, such as filtering retweets, lowercasing, removing user names, links, double spacings and numbers, tokenization, stop word removal and stemming. For NMF, we identified the following parameters to be suitable: $k = 20$ topics, regularization factor $\alpha = 0.1$ and a regularization mixing parameter $r = 20$, which ensures a balanced combination of element-wise L1 and L2 regularization. We found 10 of the 20 identified topics to be quite related to human interpretable and crisis-related topics (see Table 1). The remaining topics are rather related to specific places, things people talk about while waiting (e.g. food), about their expectations when they can return back to their homes after being evacuated as well as mixtures of rather non-actionable or not emergency-related topics (according to manual inspection of random samples).

As an example, the spatio-temporal distributions of the topic flood for 15–17 September are shown in Figure 10. Each dot on

the maps represents one or more tweets assigned to the topic. The larger the marker, the more tweets contribute to the topic in the corresponding region. From these plots, clear occurrence peaks in Charlotte and in Durham can be observed on 16 and 17 September, respectively. Besides these rather large-scale events that are obviously visible since a large fraction of the population is affected, also heavy impact flood events represented by only a handful of tweets, for example in Wilmington, are visible. A subsequent clustering and summarization, for example using word clouds, would also in this case help to capture and visualize the present information.

Further relations between the mapped NMF topics to real events reported on local news sites could be identified, for example in case of traffic-related problems, road and highway closures, and power outages.

5 | DISCUSSION

The experimental results demonstrate that the proposed workflow has a great potential to meet our defined goals of capturing local

Topic Interpretation	Top 20 keywords
Help and Support (e.g. donations, shelter)	help, need, us, need_help, donat, help_us, know, commun, way, call, help_,would, let, come, want, support, make, relief, shelter, could, give
Situation Report (official media)	report, co, edt, pm, pm_edt, co_nc, tree, road, counti, flash, flash_flood, report_flash, emerg, tropic, sc, tropic_storm, report_tropic, media, nc_emerg, mngr
Sympathy and Support	safe, stay, stay_safe, everyon, hope, friend, safe_, safe_ everyon, famili, home, pleas_stay, everyon_stay, hope_everyon, weather, keep, dri, weekend, love, stay_home, travel
Weather Report	rain, wind, mph, f, wind_mph, humid, f_humid, humid_wind, cloud, mph_mph, heavi, gust, temperatur, temperatur_f, f_f, weather, rain_wind, current, pressur, wind_rain
Power Supply	power, outag, without, back, line, without_power, tree, power_outag, power_line, lost, lost_power, restor, hour, went, work, area, crew, power_back, power_went, lose
Flood	flood, flash, flash_flood, road, area, river, warn, flood_warn, rain, re- port_flash, due_flood, due, creek, drive, street, feet, flood_ water, close, near, stage
Traffic	traffic, rd, charlotte, rd_, hwi, accid, accid_, durham, near, st, exit, tree, nb, sb, fallen, blvd, raleigh, close, wb, eb
Flood and Water	water, rescu, rise, ncwx, river, flood_water, road, water_ rescu, wral, high, food, bottl, new, level, stand, creek, bottl_water, feet, drive, good
School Closures	today, close, school, open, day, counti, updat, due, weather, shelter, tomorrow, pm, road, monday, friday, hour, septemb, high, cancel, student
Sentiment	thank, us, god, work, keep, good, much, prayer, thank_god, great, pray, appreci, area, safe, back, see, everyon, bless, hard, respond

event dynamics and impacts on population and environment from Twitter data. The initial robust filtering of crisis-related tweets (i.e. without a focus on specific events) enables the application of the subsequent analysis methods to answer event- and user-specific questions. Large-scale and long-term hurricane events, like the one exemplarily studied in this paper, usually resemble characteristics of types of events such as floods, power outages and tornadoes. Thus, we expect that the proposed workflow will generate helpful insights when transferred to other types of crisis events. The workflow can be applied in the aftermath of crisis events or in a repeated (sliding windowed) manner during events.

The involved analysis methods have shown to provide a multi-faceted view on the data with different levels of granularity. Tweet classification is conducted by utilizing a CNN model. State-of-the-art model accuracies of around 0.55 F1 score for the complex task of classifying short messages into 25 crisis-related information classes imply a need for improvements for these models.

Local hotspot analysis and STKDE are methods to reveal broad trends and changes regarding crisis-related Twitter activities and

therefore potentially support the detection of (sub-) events, such as the Durham flooding on September 17. By applying local hotspot analysis on a subset of topics, for example related to damage, more detailed and specific spatio-temporal information can be extracted from the data.

ST-DBSCAN clustering assigns crisis-related tweets to spatio-temporal clusters. This technique has shown to be quite helpful to identify small clusters that often represent real events discussed by a few affected citizens. A meaningful way to facilitate human understanding of a cluster's characteristics is to visualize the class distribution and a word cloud. Analysing class distributions is helpful and representative for the content of large clusters and helps to get a broad overview about the discussed topics. For small clusters containing only a handful of tweets, word clouds help to visualize and immediately capture the corresponding topics. According to the examples in Figure 9, this is an effective approach to reveal detailed information. Furthermore, this approach circumvents errors introduced by the CNN classifier.
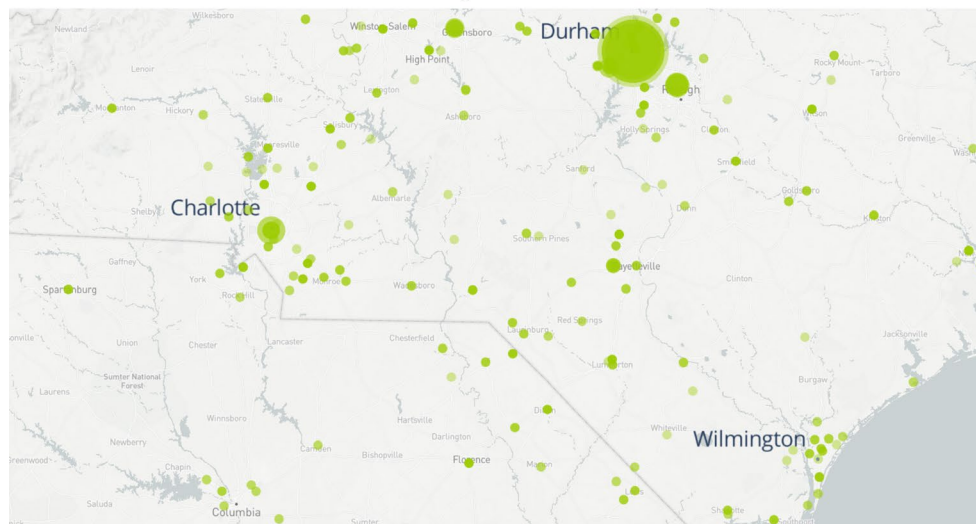
A drawback of ST-DBSCAN is that a large amount of tweets might be labelled as noise. Furthermore, the content of large
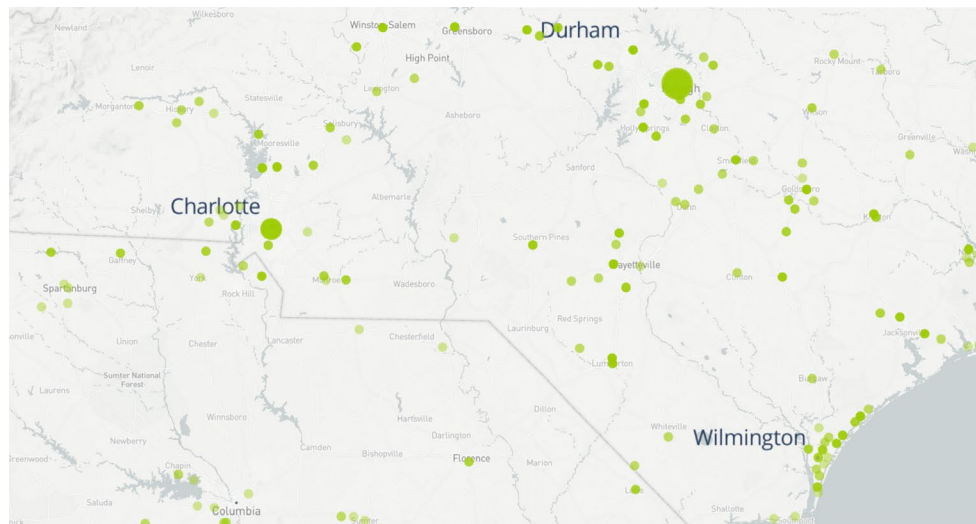
(a) September 15
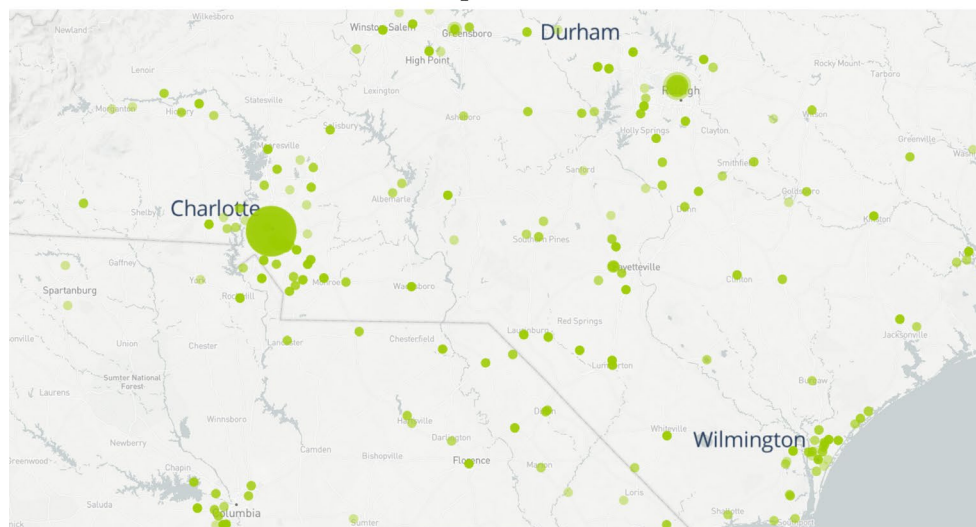
(b) September 16

(c) September 17

FIGURE 10 Exemplary plots of tweet occurrences from NMF topic *Flood*. Small points contain ≤ 10 tweets, the largest point contains 50 tweets. Maps are based on OpenStreetMap data (OpenStreetMap, 2019) [Colour figure can be viewed at wileyonlinelibrary.com]

spatio-temporal clusters tends to be a mixture of various topics difficult to interpret in detail. This motivates the application of topic modelling techniques. Neglecting information about space and time, discussed topics from the collection of all tweets are extracted. The spatio-temporal distribution of topics can then be visualized, for example on daily maps. In our experiments, this daily analysis is applied to Twitter data for the whole region affected by the hurricane. 10 meaningful crisis-related topics, such as *Flood, Traffic* and *Situation Report*, could be identified automatically by NMF (see Table 1). As demonstrated for the class *Flood* (see Figure 10), mapping these topics and observing topic changes over time enables the identification of real events. As a conclusion from our findings, the application of topic modelling is likely to also be a good approach to extract more fine-grained information from large ST-DBSCAN clusters.

Our findings indicate that the proposed workflow is a very effective tool. However, further experiments, improvements and adjustments are necessary. Applying filtering, clustering, topic modelling and hotspot analysis in a varied order is likely to reveal more detailed event- and user-specific information. For instance, a sequence of filtering crisis-related tweets, the classification or topic modelling-based extraction of damage-related tweets followed by a clustering (or hotspot detection) might be a good approach to map potential damages reported via Twitter.

6 | CONCLUSIONS AND OUTLOOK

In this paper, a general workflow to capture large-scale and local natural disaster event dynamics from Twitter data is proposed. Even though the involved methods have shown to provide a meaningful and multifaceted view to the happenings during hurricane Florence in different granularities, this workflow can be seen as a starting point for various improvements, further experiments and interesting research directions. Due to the desired flexibility and transferability of the workflow, another sequence of methods might be more appropriate for specific applications. In turn, this flexibility would enable the mining of more detailed information, for instance, when topic modelling or a further clustering step is applied to a large cluster found by ST-DBSCAN.

Qualitative and exemplary results of STKDE, ST-DBSCAN, hotspot analysis, NMF topic modelling and cluster analysis based on word clouds and semantic class histograms are presented in this work in order to demonstrate the effectiveness and value of our workflow. However, more systematic and quantitative experiments are required in order to evaluate each workflow component in detail. Related to this, an interesting and important topic for further research would be how spatio-temporal analysis results can be meaningfully evaluated since ground truth information, in particular for very local events, is often not available and distributed over various data sources (news articles, feeds and blogs). Furthermore, workflow transferability to different event types has to be evaluated.

Two interesting but contrary research directions are possible for the workflow development. One possibility would be to add more interactivity, workflow flexibility and visualization options, for instance with a graphical user interface allowing to inspect results and to adjust method parameters according to specific user requirements. The other one would be to increase the degree of automation in order to obtain a fully automated and near real-time event detection and tracking system. For the latter, modules for continuous data acquisition, information localization, event detection, data storage and others have to be added to the workflow. Additional available information, such as tweet metadata and the directed network of "@username" mentions on Twitter, could also be exploited for data analysis.

With respect to the summarized open research directions in the introduction, the extraction and ranking of important actionable messages should be further investigated. For this, it would be helpful to gain more insights on the specific language characteristics for these types of short messages. Furthermore, the combination and fusion of analysis results with other information sources should be addressed in future research, since data solely obtained from a single source are likely to only represent a specific subgroup of citizens affected by a crisis and thus reveal just a partial view on the event.

ORCID
Jens Kersten https://orcid.org/0000-0002-4735-7360
Friederike Klan https://orcid.org/0000-0002-1856-7334

ENDNOTES
[1] https://www.tensorflow.org/

[2] https://keras.io/

[3] https://scikit-learn.org/stable/

[4] https://github.com/maovieira/localGi

[5] https://www.qgis.org

[6] https://github.com/eubr-bigsea/py-st-dbscan

REFERENCES
Abel, F., Hauff, C., Houben, G., Stronkman, R., & Tao, K. (2012). Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*, ACM, 305–308.

Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, *7*, 31883–31902. https://doi.org/10.1109/ACCESS.2019.2903568

Alam, F., Imran, M., & Ofli, F. (2019). Crisis data processing services. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain, May 19–22, 2019*.

Alam, F., Ofli, F., Imran, M., & Aupetit, M. (2018). A twitter tale of three hurricanes: Harvey, Irma, and Maria. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Rochester, NY, USA, May 20–23, 2018*.

Berlingerio, M., Calabrese, F., Lorenzo, G. D., Dong, X., Gkoufas, Y., & Mavroeidis, D. (2013). Safercity: A system for detecting and analyzing incidents from social media. In *2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 1077–1080).

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*, 10. https://doi.org/10.1088/1742-5468/2008/10/p10008

Burel, G., & Alani, H. (2018). Automatic Detection and Classification of Crisis-related Content on Social Media. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Rochester, NY, USA, May 20–23, 2018*.

Burel, G., Saif, H., Fernandez, M., & Alani, H. (2017). On semantics and deep learning for event detection in crisis situations. In *Workshop on Semantic Deep Learning (SEMDEEP), ESWC 2017*.

Casalino, G., Castiello, C., Buono, N. D., & Mencar, C. (2018). A framework for intelligent Twitter data analysis with non-negative matrix factorization. *IJWIS*, *14*, 334–356. https://doi.org/10.1108/IJWIS-11-2017-0081

Cheng, T., & Wicks, T. (2014). Event detection using Twitter: A spatio-temporal approach. *PLoS One*, *9*(6), 1–10. https://doi.org/10.1371/journal.pone.0097807

Cichocki, A., & Phan, A.-H. (2009). Fast local algorithms for large scale non-negative matrix and tensor factorizations. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *E92-A*(3), 708–721. https://doi.org/10.1587/transfun.E92.A.708

Dhillon, I. S., & Sra, S. (2005). Generalized non-negative matrix approximations with Bregman divergences. In *Proceedings of the 18th International Conference on Neural Information Processing Systems* (283–290). NIPS'05, Vancouver, British Columbia, Canada: MIT Press. https://doi.org/10.5555/2976248.2976284

Epanechnikov, V. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158. https://doi.org/10.1137/1114019

ESRI (2018). *Population Density by Congressional District*. Retrieved from https://www.arcgis.com/home/item.html?id=ff48bbae433442a38f6c635b8c7baf72.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). KDD'96. Portland, Oregon: AAAI Press.

Europe Media Monitor (2019). *Europe Media Monitor NewsBrief Service*. Retrieved from https://emm.newsbrief.eu/

GDELT Project (2019). *GDELT Project Page*. Retrieved from https://www.gdeltproject.org/.

Google Maps. (2019). V 7.3.2.5495, North Carolina, USA, lat 35.25, lon -78.91, eye alt 588.2 km. SIO, NOAA, U.S. Navy, NGA, GEBCO. Landsat, Copernicus, Google 2019, accessed at September 12, 2019.

Hasan, M., Orgun, M. A., & Schwitter, R. (2018). A survey on real-time event detection from the Twitter data stream. *Journal of Information Science*, *44*(4), 443–463. https://doi.org/10.1177/0165551517698564

Hu, Y., Wang, F., Guin, C., & Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography*, *99*, 89–97. https://doi.org/10.1016/j.apgeog.2018.08.001

Huang, Y., Li, Y., & Shan, J. (2018). Spatial-temporal event detection from geo-tagged tweets. *ISPRS International Journal of Geo-Information*, *7*(4), 150. https://doi.org/10.3390/ijgi7040150

Humanity Road. (2019). Retrieved from https://www.humanityroad.org.

Hunt, L., & Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(4), 352–361. https://doi.org/10.1002/widm.33

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2018). Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the Web Conference 2018*(pp. 507–511). WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). AIDR: Artificial intelligence for disaster response. In *WWW (companion volume)* (pp. 159–162). ACM.

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. In *Proceedings of the 10th international conference on language resources and evaluation (LREC)*. Portoroz, Slovenia: European Language Resources Association (ELRA).

Jinyin, C., Huihao, H., Chen, J., Shanqing, Y., & Zhaoxia, S. (2017). Fast density clustering algorithm for numerical data and categorical data. *Mathematical Problems in Engineering*, *2017*, 1–15. https://doi.org/10.1155/2017/6393652

Kersten, J., Kruspe, A., Wiegmann, M., & Klan, F. (2019). Robust filtering of crisis-related tweets. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain, May 19–22, 2019*.

Kruspe, A., Kersten, J., & Klan, F. (2019). Detecting event-related tweets by example using few-shot models. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Valencia, Spain, May 19–22, 2019*.

Kruspe, A., Kersten, J., Wiegmann, M., Stein, B., & Klan, F. (2018). Classification of incident-related tweets: Tackling imbalanced training data using hybrid CNNs and translation-based data augmentation. In *Proceedings of the 27th Text REtrieval Conference (TREC 2018), Gaithersburg, Maryland, November 14–16, 2018*.

Lee, J., Gong, J., & Li, S. (2017). Exploring spatiotemporal clusters based on extended kernel estimation methods. *International Journal of Geographical Information Science*, *31*(6), 1154–1177.

Liang, C., Lin, G., & Peng, J. (2018). Discover the spatio-temporal process of typhoon disaster using micro blog data. *MATEC Web Conference*, *246*, 3013. https://doi.org/10.1051/matecconf/201824603013

MacMillan, K., & Wilson, J. D. (2017). Topic supervised non-negative matrix factorization. CoRR, abs/1706.05084. arXiv: 1706.05084. Retrieved from http://arxiv.org/abs/1706.05084

Martin, Y., Li, Z., & Cutter, S. L. (2017). Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of hurricane Matthew. *PLoS One*, *12*(7), 1–22. https://doi.org/10.1371/journal.pone.0181701

McCreadie, R. (2019). TREC-IS V2 Metrics. Retrieved from http://dcs.gla.ac.uk/~richardm/TREC_IS/2019/TREC_IS_Metrics.pdf

McCreadie, R., Macdonald, C., & Ounis, I. (2016). EAIMS: Emergency analysis identification and management system. In *SIGIR* (pp. 1101–1104). ACM.

McCreadie, R., & Soboroff, I. (2018). TREC overview paper - Incidents streams track. In *26th Text REtrieval Conference (TREC)*, 14th November 2018 (p. 9).

Ngamassi, L., Malik, A., Zhang, J., & Edbert, D. (2017). Social media visual analytic toolkits for disaster management: A review of the literature. In *14th Proceedings of the International Conference on Information*

Systems for Crisis Response and Management (ISCRAM), Albi, France, May 21–24, 2017.

Nguyen, D. T., Al-Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the 11th international conference on web and social media, ICWSM 2017* (pp. 632–635). AAAI press.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 994–1009). Vancouver, BC, Canada: ACM.

OpenStreetMap (2019). *OSM Planet Dump*. Retrieved from https://planet.osm.org.

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

Ramachandran, D., & Ramasubramanian, P. (2018). Event detection from Twitter – a survey. *International Journal of Web Information Systems*, *14*(3), 262–280. https://doi.org/10.1108/IJWIS-11-2017-0075

Resch, B., Uslaender, F., & Havas, C. (2018). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362–376. https://doi.org/10.1080/15230406.2017.1356242

Reuter, C., Hughes, A. L., & Kaufhold, M.-A. (2018). Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human-Computer Interaction*, *34*(4), 280–294. https://doi.org/10.1080/10447318.2018.1427832

Santa, F., Henriques, R., Torres-Sospedra, J., & Pebesma, E. (2019). A statistical approach for studying the spatio-temporal distribution of geolocated tweets in urban environments. *Sustainability*, *11*(3), 595. https://doi.org/10.3390/su11030595

Shi, Y., Deng, M., Yang, X., Liu, Q., Zhao, L., & Lu, C.-T. (2016). A framework for discovering evolving domain related spatio-temporal patterns in Twitter. *ISPRS International Journal of Geo-Information*, *5*, 193. https://doi.org/10.3390/ijgi5100193

Steiger, E., de Albuquerque, J. P., & Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Transactions in GIS*, *19*(6), 809–834. https://doi.org/10.1111/tgis.12132

Thom, D., Bosch, H., Koch, S., Woerner, M., & Ertl, T. (2012). Spatiotemporal anomaly detection through visual analysis of geolocated Twitter messages. In *2012 IEEE Pacific Visualization Symposium*(pp. 41–48). https://doi.org/10.1109/PacificVis.2012.6183572

Thomas, C., McCreadie, R., & Ounis, I. (2019). Event tracker: A text analytics platform for use during disasters. In *SIGIR* (pp. 1341–1344). ACM.

Twitter (2019). *Twitter streaming API overview*. Retrieved from https://developer.twitter.com/en/docs/tweets/filter-realtime/overview.html

Walther, M., & Kaisser, M. (2013). Geo-spatial event detection in the Twitter stream. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, S. Ilya, & E. Yilmaz (Eds.), *Advances in information retrieval* (pp. 356–367). Berlin, Heidelberg: Springer, Berlin Heidelberg.

Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, *32*(1), 49–72. https://doi.org/10.1080/13658816.2017.1367003

Wang, Z., Ye, X., & Tsou, M.-H. (2016). Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, *83*(1), 523–540. https://doi.org/10.1007/s11069-016-2329-6

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., & Stein, B. (2020). Analysis of Detection Models for Disaster-Related Tweets. In *Proceedings of the 17th ISCRAM Conference*, *May 2020.*

Wikipedia (2019). *Wikipedia*. Retrieved from https://en.wikipedia.org/

Zhao, L., Chen, F., Dai, J., Hua, T., Lu, C.-T., & Ramakrishnan, N. (2014). Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PLoS One*, *9*(10), 1–12. https://doi.org/10.1371/journal.pone.0110206

Zhou, X., & Xu, C. (2017). Tracing the spatial-temporal evolution of events based on social media data. *ISPRS International Journal of Geo-Information*, *6*(3), 88. https://doi.org/10.3390/ijgi6030088