Information Processing and Management 57 (2020) 102132

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning

Marc-André Kaufhold[a,b,*], Markus Bayer[a], Christian Reuter[a]

[a] Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Germany
[b] Institute for Information Systems, University of Siegen, Germany

ARTICLE INFO

Keywords:
Crisis management
Information overload
Relevance classification
Social media
Supervised machine learning

ABSTRACT

The research field of crisis informatics examines, amongst others, the potentials and barriers of social media use during disasters and emergencies. Social media allow emergency services to receive valuable information (e.g., eyewitness reports, pictures, or videos) from social media. However, the vast amount of data generated during large-scale incidents can lead to issue of information overload. Research indicates that supervised machine learning techniques are suitable for identifying relevant messages and filter out irrelevant messages, thus mitigating information overload. Still, they require a considerable amount of labeled data, clear criteria for relevance classification, a usable interface to facilitate the labeling process and a mechanism to rapidly deploy retrained classifiers. To overcome these issues, we present (1) a system for social media monitoring, analysis and relevance classification, (2) abstract and precise criteria for relevance classification in social media during disasters and emergencies, (3) the evaluation of a well-performing Random Forest algorithm for relevance classification incorporating metadata from social media into a batch learning approach (e.g., 91.28%/89.19% accuracy, 98.3%/89.6% precision and 80.4%/87.5% recall with a fast training time with feature subset selection on the European floods/BASF SE incident datasets), as well as (4) an approach and preliminary evaluation for relevance classification including active, incremental and online learning to reduce the amount of required labeled data and to correct misclassifications of the algorithm by feedback classification. Using the latter approach, we achieved a well-performing classifier based on the European floods dataset by only requiring a quarter of labeled data compared to the traditional batch learning approach. Despite a lesser effect on the BASF SE incident dataset, still a substantial improvement could be determined.

1. Introduction

As the work of professional bodies, volunteers, and others is increasingly mediated by computer technology, and more specifically by social media,[1] research on crisis management has become more common (Hiltz, Diaz, & Mark, 2011; Palen & Hughes, 2018; Reuter, Hughes, & Kaufhold, 2018). The emerging research field of *crisis informatics* has revealed interesting and important real-world

* Corresponding author.
E-mail address: kaufhold@peasec.tu-darmstadt.de (M.-A. Kaufhold).

[1] We follow the definition of social media as a "group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content" (Kaplan and Haenlein, 2010).

https://doi.org/10.1016/j.ipm.2019.102132
Received 8 April 2019; Received in revised form 6 September 2019; Accepted 21 September 2019

uses for social media (Soden & Palen, 2018). Coined by Hagar (2007), crisis informatics is 'a multidisciplinary field combining computing and social science knowledge of disasters; its central tenet is that people use personal information and communication technology to respond to disasters in creative ways to cope with uncertainty' (Palen & Anderson, 2016).

During disasters and emergencies, it is necessary for emergency services to obtain a comprehensive situational overview for coordination efforts and decision making (Imran, Castillo, Diaz, & Vieweg, 2015; Vieweg, Hughes, Starbird, & Palen, 2010). In such situations, social media are increasingly used for the exchange of information (Hughes & Palen, 2009) while emergency services encounter the issue of information overload, amongst others (Hughes & Palen, 2014; Mendoza, Poblete, & Castillo, 2010; Plotnick & Hiltz, 2016). Research indicates that supervised machine learning techniques are suitable for identifying relevant messages and filter out irrelevant messages, thus mitigating information overload (Habdank, Rodehutskors, & Koch, 2017). Besides the potential of improving the performance of algorithms for relevance classification, supervised machine learning techniques require a considerable amount of labeled data, which constitutes an issue due to the time-sensitive nature of disasters and emergencies (Imran, Castillo, Diaz, & Vieweg, 2018). Furthermore, clear criteria for relevance classification are required, a usable interface to facilitate the labeling process (Stieglitz, Mirbabaie, Fromm & Melzer, 2018) and a mechanism to rapidly deploy retrained classifiers.

Based on a communication matrix (Reuter et al., 2018) and social media analytics framework (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018), we designed and evaluated a system featuring active and online learning to support the information flow of *integration of citizen-generated content* featuring social media. Thereby, we seek to answer the following research questions:

- What are suitable criteria for relevance classification and labeling in disasters and emergencies (RQ1)?
- How can existing supervised machine learning techniques for relevance classification be improved for use in real disaster and emergency environments (RQ2)?
- How can the amount of labeled data required for relevance classification be reduced by active incremental learning and transparent visualization of the classifier's quality (RQ3)?
- How can the dynamic retraining of relevance classifiers be supported by user feedback performance-wise using batch learning with feature subset selection (RQ4)?

To reflect the methodology used in our project, the paper is structured as follows: First, based on the analysis of related work on relevance classification in social media during disasters and emergencies (Section 2), we present the architecture of a system for social media analysis and relevance classification (Section 3). Thereafter, we present an approach and evaluation of relevance classification via batch learning (Section 4) as well as an approach and preliminary evaluation of relevance classification via active and online learning (Section 5). Finally, we discuss the results, outlining practical implications and theoretical contributions, to draw conclusions on how relevance classification in social media might be further improved during disasters and emergencies (Section 6).

This paper contributes to the research areas of machine learning and social media analytics with (1) a system for social media monitoring, analysis, and relevance classification, (2) the definition of criteria (content, keywords, language, length, links, location of author and post, media, named entities, retweets, time) for relevance classification in social media during disasters and emergencies, (3) the evaluation of a well-performing classifier for relevance classification using batch learning, (4) an approach for relevance classification including active and online learning, (5) an approach for reclassifying wrongly classified messages using batch learning with feature subset selection, as well as (6) the preliminary evaluation and outlook for relevance classification using active, incremental and online learning.

2. Literature review

In order to get insights into **RQ1 ("What are suitable criteria for relevance classification and labeling in disasters and emergencies?")**, we conducted a literature review whose method (Section 2.1), results (Sections 2.2–2.4) and research gap (Section 2.5) are described in the following.

2.1. Method

Considering the search scope framework of vom Brocke et al. (2015), this literature review followed a sequential search *process (I)*. The literature review motivates the use of social media in disasters and relevance classification to mitigate information overload (Section 2.2), it discusses abstract and precise criteria for relevance classification (Section 2.3), outlines artificial intelligence and social media analytics approaches for relevance classification (Section 2.4) and formulates a research gap (Section 2.5). In terms of *sources (II)*, the bibliographic databases of IEEE Xplore and Google Scholar were searched to achieve a representative *coverage (III)* of the topics by using the *technique (IV)* of keyword search.

However, since we want to establish a more detailed understanding on definitions and criteria of relevance in social media during disasters and emergencies, in Section 2.2, we strived for a comprehensive coverage. This part of the literature review is structured into two phases to firstly identify abstract and interpretative relevance criteria, whereby we understand relevance as a component of information quality, and secondly precise and factual relevance criteria. Since keyword searches such as *"(relevance OR relevant OR importance OR important) AND (social media OR social network OR twitter) AND (crisis OR disaster OR emergency)"* did not yield a relevant number of useful results, we also conducted backward, based on works such as Habdank et al. (Habdank et al., 2017), and forward searches. The review outlined that most publications do not discuss their definition or understanding of relevance comprehensively, only discussing some criteria and indicators selectively.

2.2. Social media in disasters and relevance classification as means to mitigate information overload

Motivated by citizen behavior during both natural and human-induced (large-scale) incidents, such as 2012 Hurricane Sandy (Hughes, St. Denis, Palen, & Anderson, 2014), the 2013 European Floods (Albris, 2017), or the 2016 Brussels bombings (Stieglitz, Bunker, Mirbabaie, & Ehnis, 2017), crisis informatics has examined potentials and challenges of social media usage in disasters and emergencies by both authorities and citizens (Imran et al., 2015; Reuter & Kaufhold, 2018; Soden & Palen, 2018). Amongst others, social media might enable crowdsourcing of specific tasks (Dittus, Quattrone, & Capra, 2017; Ludwig et al., 2017), communication between authorities and citizens (Reuter & Spielhofer, 2017; Reuter, Ludwig, Kaufhold, & Spielhofer, 2016), co-ordination among citizens and mobilization of unbound digital or real volunteers (Kaufhold & Reuter, 2016; Starbird & Palen, 2011; White, Palen, & Anderson, 2014), (sub-)event detection (Pohl, Bouchachia, & Hellwagner, 2015; Sakaki, Okazaki, & Matsuo, 2010) or improved situational awareness (Imran et al., 2015; Vieweg et al., 2010).

However, considering the emergency services' use of social media, according to a study with the US public sector, the major barriers to social media use are organizational rather than technical (Hiltz, Kushma, & Plotnick, 2014). Research suggests that human factors are crucial for effective emergency management, but also technology for conducting respective emergency tasks (Kim, Sharman, Cook-Cottone, Rao, & Upadhyaya, 2012). However, once organizational willingness is established, technical systems are needed to make sense of the large amount of data. For instance, research has identified barriers and challenges in the authorities' use of social media, such as credibility, liability (Hughes & Palen, 2014), reliability and overload of information (Mendoza et al., 2010) as well as lack of guidance, policy documents, resources, skills and staff within the organization (Plotnick & Hiltz, 2016).

In this paper, we present an approach for relevance classification, which potentially mitigates the issue of information overload by filtering out irrelevant messages. For the conceptual framing, we refer to the crisis communication matrix by Reuter et al. (2018) and the social media analytics framework by Stieglitz et al. (2018). The crisis communication matrix distinguishes authorities (A) and citizens (C), both as sender and receiver, respectively, to derive four communication flows of crisis communication (A2C), self-help communities (C2C), (inter-)organizational crisis management (A2A) and integration of citizen-generated content (C2A) (Reuter et al., 2018).

To leverage social media information such as eyewitness reports, pictures, and videos taken with mobile phones as a basis for authorities' decision-making, they are not only required to *integrate citizen-generated content (C2A)*, i.e., monitoring social media, while managing the vast amount of data (Olshannikova, Olsson, Huhtamäki, & Kärkkäinen, 2017). When tens of thousands of social media messages are generated during large-scale emergencies, authorities have to deal with the issue of information overload which is traditionally defined as '[too much] information presented at a rate too fast for a person to process' (Hiltz & Plotnick, 2013, p. 823). Referring to the information overload problem from the field of visual analytics, Keim, Andrienko, Fekete, Carsten, and Melan (2008) highlight the danger of getting lost in data which may be firstly irrelevant to the current task at hand and secondly processed and presented in an inappropriate way. Considering the human capacity of information processing, Miller (1956) suggests 'organizing or grouping the input into familiar units or chunks' (p. 93) to overcome such limitations. Accordingly, functionalities such as filtering and grouping potentially assist in overcoming the issue of information overload (Moi et al., 2015; Plotnick, Hiltz, Kushma, & Tapia, 2015; Tucker, Ireson, Lanfranchi, & Ciravegna, 2012). This is supported by a survey of 477 U.S. county-level emergency managers, which revealed that perceived information overload negatively influences the adaptation of social media, while the 'chunking' or grouping of social media messages by specific tools positively influences the intention to use social media during emergencies (Rao, Plotnick, & Hiltz, 2017). Since this paper focuses on relevance classification for noise reduction via supervised machine learning, we will introduce abstract and precise relevance criteria (Section 2.3) and existing social media analytics approaches for relevance classification (Section 2.4) before outlining a research gap (Section 2.5).

2.3. Abstract and precise criteria for relevance classification

Firstly, Saracevic defines relevance as an intuitive understanding: "Intuitively, we understand quite well what relevance means. It is a primitive "y' know" concept, as is information for which we hardly need a definition" [41, p. 324]. With regard to information science, he postulates that "relevance is considered as a measure of the effectiveness of the contact between a source and a destination in a communication process", including the aspects of subject knowledge and literature, other linguistic or symbolic representations, source and destination (of files), information systems, environment, realities, functions and values (Saracevic, 1975). Accordingly, Schamber and Eisenberg introduce a "user-centric" approach to relevance, emphasizing subjective realities in assessing relevance (Schamber & Eisenberg, 1988), which is extended in a later publication: "As based on the Sense-Making approach, the locus of relevance is within individuals' perceptions of information and information environment – not in information as represented in a document or some other concrete form. […] Relevance, then, is a *dynamic* concept that depends on users' individual judgments of the quality of the relationship between information and information need at a certain point in time" [43, p. 770].

In their publication about quality of information, Rohweder, Kasten, Malzahn, Piro, and Schmid (2011) define that a piece of information is relevant if it contains information *necessary* for the user. Necessity in this context means that the respective information facilitates reaching a certain achievement. Other than relevance, the authors name the terms *immediacy* and *appropriate scope* as quality of information. During the literature review, these terms have shown to be frequently seized upon. Furthermore, Shankaranarayanan, Iyer, and Stoddard (2012) point out that for defining *relevance* the respective context of use has to be considered. Saracevic adds that "In information science, we consider relevance as a relation between information or information objects […] on the one hand and contexts, which include cognitive and affective states and situations (information need, intent, topic, problem, task; […]) on the other hand, based on some property reflecting a desired manifestation of relevance (topicality, utility, cognitive match;

[…])" [(Saracevic, 2007), p. 1918]. Accordingly, Agarwal and Yiliyasi (2010) define *contextual relevance* regarding quality of information in social media. Data from social media might be relevant for certain actors while being irrelevant for others. They describe *relevance* as a degree which determines how useful data are for a certain task. Specifically referring to crisis situations, Jensen (Jensen, 2012) defines *relevance* as *useful* and moreover *adequate* and *valuable* information. This is also in line with Eisenberg, how adds that "two common choices for definitions seem to be *topicality* and *usefulness*" [(Eisenberg, 1988), p. 387].

Borlund differentiates between multidimensional, dynamic and situational relevance [(Borlund, 2003), p. 913]. Regarding crisis situations, Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010) describe a classifier of Twitter posts relating to news, events, opinions, offers and private messages. They define events as "*something that happens at a given place and time*" (Sriram et al., 2010). Verma et al. (2011) develop a classifier which recognizes relevant Twitter posts regarding situational awareness in crisis situations. They describe a Tweet as relvant if it contains tactical and usable information. Other than in the thesis at hand, the authors limit themselves to textual features. Their result is that posts, which show situational awareness, tend to exhibit objective, impersonal and formal features. Objective posts are based on factual information, do not express an opinion and do not include emotional language. Impersonality is characterized by an emotional distance between user and event. Formal tweets are such which are grammatically coherent and express complete thoughts (Verma et al., 2011). Vieweg (2012) describes in her dissertation the usage behavior of Twitter users in crisis situations. She uses the aforementioned classifier developed by Verma et al. (2011) to minimize the dataset to a manageable amount of posts. Vieweg describes how Tweets that contribute to situational awareness in mass events can be recognized. The posts are divided into three categories:

1 "O: Off-topic"
2 "R: On-topic, relevant to situational awareness"
3 "N: On-topic, irrelevant to situational awareness"

O means that none of the information contributes to the event as such. The *N*-division relates to posts that refer to the event but do not add to the relevance. As an example, she mentions persons who call for donations or express their sympathy. All posts which „*contain information that provides tactical, actionable information that can aid people in making decisions, advise others on how to obtain specific information from various sources, or offer immediate postimpact help to those affected by the mass emergency*" [53, p. 164] are allocated to the *R*-Classification. In the dissertation, the author does not offer any specific criteria to explain how the relevance filters were labeled. Instead she refers to other classification of the data. In the first part of their publication, Imran, Elbassuoni, Castillo, Diaz, and Meier (2013) also filter the relevance regarding situational awareness in crisis situations. First, they refer to the definition of relevance Vieweg (2012) employs for the *R*-Classification but also offer their own definition of relevance. They differentiate Twitter posts as follows:

1 "Personal Only"
2 "Informative (Direct or Indirect)"
3 "Other"

A post is personal ("Personal Only") if it only relates to the author himself and close friends or family and does not disclose any useful information to persons unbeknownst to the author. Informative posts ("Informative") include information which might be useful for persons who do not know the author. Informative posts are divided into direct, if the post was written by an eye-witness, and indirect, if the post was written by a person based on information from news, radio or television. All Tweets not written in English are categorized as "Other" (Imran et al., 2013, 2013). Table 1 summarizes the afore-presented relevance criteria.

Besides abstract and interpretative criteria, precise criteria to classify relevance in crisis situations were also identified.

Table 1
Abstract and interpretative relevance criteria.

Criteria	Description	Sources
Subjective	depending on the individuals' perception of information	(Schamber & Eisenberg, 1988; Schamber, Eisenberg, & Nilan, 1990)
Necessary	information that facilitate reaching a certain achievement	(Rohweder et al., 2011)
useful, adequate, valuable	–	(Jensen, 2012)
context-related	depending from the actual use	(Agarwal & Yiliyasi, 2010; Shankaranarayanan et al., 2012)
tactical	–	(Verma et al., 2011; Vieweg, 2012)
usable	–	(Verma et al., 2011)
objective	objective, do not express an opinion and do not include emotional language	(Verma et al., 2011)
impersonal	emotional distance between user and event	(Imran et al., 2013; Verma et al., 2011)
formal	complete and grammatically coherent	(Verma et al., 2011)
contributing to decision making	–	(Imran et al., 2013, 2012)
assisting and advisory	–	(Imran et al., 2013, 2012)
informative	not only relating to close friends or family of the author	(Imran et al., 2013, 2013)

Abel, Hauff, and Stronkman (2012) present "Twitcident", a filter system for crisis situation, and describe that *keyword-based filters* are one of the two steps towards the recognition of relevant posts in their system. Within the scope of their research project "Alert4All", Johansson, Brynielsson, and Quijano (2012) work on the importance of terminology regarding events. Furthermore, Vieweg (2012) provides a list of terms that are not allowed in the tweets regarding a specific dataset. She removes all posts which include terms regarding donations, sympathies or anger.

Vieweg, Hughes, Starbird, and Palen (2010) point out in their crisis-specific Twitter analysis that the *geographic position* of the Twitter user and references to a certain location can be an indication for relevant posts, especially in terms of situational awareness. Further publications emphasize the relevance of these criteria for authorities (Ludwig, Reuter, & Pipek, 2015; Reuter, Ludwig, Ritzkatis, & Pipek, 2015; Sriram et al., 2010; Verma et al., 2011; Vieweg, 2012). de Albuquerque, Herfort, Brenning, and Zipf (2015) analyze the importance of the geographic position on the basis of the flooding of the river Elbe in 2013. They point out that tweets which are close to the event (up to 10 km) have a higher likelihood to be involved in the flooding.

Rohweder et al. (2011) identify the chronological correlation as a criterion for the quality of information. Moreover, Ludwig et al. (2015) emphasize the importance of *currency* and temporal proximity as criterion for relevance. Sriram et al. (2010) and Palen et al. (2010) also mention timely information regarding the event as an indication for relevance. Similarly, for example Vieweg (2012), almost any analysis of user-generated data in crisis situations limits the dataset on the basis of time without explicitly mentioning this as a filter for relevance.

Starbird and Palen (2004) show that posts which were *retweeted* have a stronger connection to the crisis event. Furthermore, it became evident that mainly persons who were close to the event used the retweet function. The authors therefore conclude that a focus on retweets might help to minimize uncertainties in the dataset of crisis situations. Reuter, Heger, and Pipek (2013) also observe the behavior of Twitter users in crisis situations. They find that 22.32% of all retweets during an event were sent by only 51 users and therefore argue that the information are highly relevant. The papers of Uysal and Croft (2011) as well as Mendoza et al. (2010) describe informally that the retweet behavior of users points to relevance and are therefore in accordance with the aforementioned publications.

Reuter et al. (2013) observe, besides the relevance of retweets, that *links* to other webpages are a particularity. Accordinlgy, 39% of the collected posts during the "Super Outbreak" of 2011 contain links. Uysal and Croft (2011) suggest another approach to filtering relevant posts on Twitter but their publication is abstracted from crisis situations. Through machine learning they show that the presence of an URL improves the exactness of the classification. In addition, Habdank et al. (2017) write that links in the form of pictures or videos are an important indication for relevance. Pictures, URLs and videos offer the opportunity to extend postings trough external content (Ludwig et al., 2015).

The evaluation of the machine relevance classification done by Imran et al. (2013) underlines that the length of a tweet contributes to its relevance. According to Rohweder et al. (2011), the criterion of a reasonable length has to be considered. Furthermore, Sriram et al. (2010) and Abel et al. (2012) eliminate in their preselection of relevant posts those which are too short. The treshold is three words in the case of Sriram et al. (2010) whereas Abel et al. (2012) eliminate all posts with less than 100 characters. In addition to that, Sriram et al. (2010) and Imran et al. (2013) only consider English tweets as relevant. Abel et al. (2012) specify *language* as criterion for relevance; accoringly, the tool "Twitcident" is able to translate the posts. In Table 2 these characteristics are compiled with a general description as well as their sources.

2.4. Machine learning and social media analytics for relevance classification

As public interfaces (APIs) enable the retrieval and processing of high volume datasets, 'systems, tools and algorithms performing social media analysis have been developed and implemented to automatize monitoring, classification or aggregation tasks' (Pohl, 2013). Here, *social media analytics* is defined as the process of social media data collection, analysis and interpretation in terms of actors, entities and relations (Stieglitz, Dang-Xuan, Bruns, & Neuberger, 2014). Accordingly, Stieglitz et al. (2018) differentiate between the steps of discovery, tracking, preparation and analysis of social media data. Social media data, sometimes referred to as

Table 2
Precise and factual relevance criteria.

Criteria	Description	Sources
keywords	inclusion and removal of posts with specific terms	(Abel et al., 2012; Johansson et al., 2012, 2012)
geographic location and referenced positions	actual location of a Twitter user or included positions in the post's text	(de Albuquerque et al., 2015; Ludwig et al., 2015; Reuter et al., 2015; Sriram et al., 2010; Starbird & Palen, 2004; Verma et al., 2011; Vieweg, 2012, 2010)
currency	temporal correlation between Twitter post and event	(Ludwig et al., 2015; Palen et al., 2010; Rohweder et al., 2011; Sriram et al., 2010)
retweet behavior	exact repetition of a Twitter post by other authors	(Mendoza et al., 2010; Reuter et al., 2013; Starbird & Palen, 2004; Uysal & Croft, 2011)
linking	links in the form of URL, pictures or videos	(Habdank et al., 2017; Reuter et al., 2013; Uysal & Croft, 2011)
length of tweets	–	(Abel et al., 2012, 2013; Rohweder et al., 2011; Sriram et al., 2010)
language	–	(Abel et al., 2012, 2013; Sriram et al., 2010)

Table 3
Overview of works examining textual classification problems in disasters or emergencies.

Authors	Classification problems	Used methods
(Habdank et al., 2017)	relevance (binary)	NB, DT, **RF**, SVM, NN
(Imran et al., 2013)	informative/relevant posts (multi-class), direct (eyewitnesses) and indirect posts (binary), different information types (multi-class)	**NB**
(Verma et al., 2011)	situational awareness/relevance (binary)	NB, **ME**
(Purohit, Castillo, Diaz, Sheth, & Meier, 2014)	help request or offer (multi-class), types of help (multi-class), resources for help (multi-class)	**NB, RF**
(Caragea et al., 2011)	different information types (multi-label)	SVM
(Markham & Muddiman, 2016)	relevance	NB
(Imran, Mitra, & Castillo, 2016)	different information types (multi-class)	SVM, RF, NB
(Ashktorab, Brown, Nandi, & Culotta, 2014)	damage reports (binary)	KNN, DT, NB, **LR**
(Abel et al., 2012, 2012)	relevance (binary)	RBLM, JS
(Li et al., 2015)	relevance (binary), help offer (binary), emotions for victims (binary)	NB
(Imran et al., 2016, 2017)	different information types (multi-class)	RF
(Li et al., 2017)	relevance (binary)	NB, RF, SVM, LR
(Caragea et al., 2016)	informative posts (binary)	**CNN**, ANN, SVM
(Nguyen et al., 2016)	informative/relevant posts (binary), different information types (multi-class)	LR, RF, SVM, **CNN**

If a comparison was conducted, the best performing method is marked bold; abbreviations: Artificial/Convolutional Neuronal Networks (ANN/CNN), Decision Trees (DT), Jaccard Similarity (JS), *k*-Nearest Neighbors (KNN), Logistic Regression (LR), Maximum Entropy (ME), Naïve Bayes (NB), Neuronal Networks (NN), Relevance-Based Language Models (RBLM), Random Forest (RF), Support Vector Machines (SVM)

big social data, includes the characteristics of *high-volume* (large-scale), *high-velocity* (high speed of data generation), *high-variety* (heterogeneous data with a high degree of complexity due to the underlying social relations) and *highly semantic* (manually created and highly symbolic content with various, often subjective meanings) data (Olshannikova et al., 2017). These characteristics pose challenges for emergency services who need their own analytical concepts.

In this paper, we focus on machine learning as a subset of artificial intelligence, which can be applied as a technique (e.g., automatically classifying messages as relevant or irrelevant) to solve problems (e.g., mitigating information overload of emergency managers by reducing the volume of incoming irrelevant data) in the domain of social media analytics. Many researchers already examined the classification of posts in disaster scenarios; Table 3 illustrates a comprehensive selection of research papers. In most cases, the motivation of supervised relevance classification algorithms lies in reducing the large volume of noisy data, e.g., to facilitate analysis by emergency services (Habdank et al., 2017; Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013). Relevance classification is a binary problem, where a post is either marked as 'relevant' or 'not relevant'. For instance, in a study of Habdank et al. (2017), the Random Forest classifier outperformed a Support Vector and Neuronal Network classifier, achieving an accuracy of 88%, precision of 86%, and recall of 89%. However, in the work of Caragea, Caragea, and Herndon (2017) and Caragea, Silvescu, and Tapia (2016) it was shown that Convolution Neural Networks seem to be well suited for classification of disaster posts with the potential to outperform Random Forests.

However, the issue of existing approaches is that well-performing classifiers require a considerable amount of labeled data, which is often not available at the beginning of a disaster or emergency (Imran, Mitra, & Srivastava, 2017; Li et al., 2015). Thus, in using *domain adaptation*, multiple approaches incorporate labeled data from past disasters in a new situation so that few or no new labels have to be created (Imran et al., 2017; Imran, Mitra, & Srivastava, 2016; Li et al., 2017; 2015). These strategies worked well, although different factors such as language may reduce the usefulness of labeled data from past events (Imran et al., 2016).

Most approaches apply *offline learning*, where the phases of labeling data, training a hypothesis and prediction (Fig. 1) are performed disjointedly (Khouzam, 2009). A version of offline learning is called batch learning, where a batch of data is labeled and then provided for the learning algorithm (Sebastiani, 2002). Due to direct availability of the whole dataset, the algorithm is able to analyze the data in greater detail (Kulessa, 2015). However, research outlines the potentials of *online learning*, which "helps models dynamically adapt to new change and patterns in the data" [14, p. 510]. In case of online learning, data is not provided in a single batch but as a sequential stream of data, allowing for updating the model continuously as new (labeled) data becomes available. This has the potential to help with overcoming the sparsity of labeled data at the beginning of a disaster and increasing the classifier's accuracy over time (Imran et al., 2017; Li et al., 2015). Furthermore, *incremental learning* describes the capability of an algorithm to be enhanced sequentially (Khouzam, 2009). The benefits of incremental algorithms are that they can be trained rapidly and the training data does not have to be loaded into the central memory completely (Ma, Saul, Savage, & Voelker, 2009). However, since they require the capability of processing an indefinite stream of data rapidly, it is possible that the classification quality decreases (Kulessa, 2015).

Finally, *active learning* may help to reduce the amount of labeled data required for a well-performing classifier. The basic idea is that "a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns" (Settles, 2010). Using active learning instead of labeling random data, the classifier proposes data whose labeling is most likely to increase the classifier's accuracy (Fürnkranz, 2018). Thus, active learning strives for increasing the classifier's accuracy with least effort in terms of the amount of labeled data (Yang & Loog, 2017).

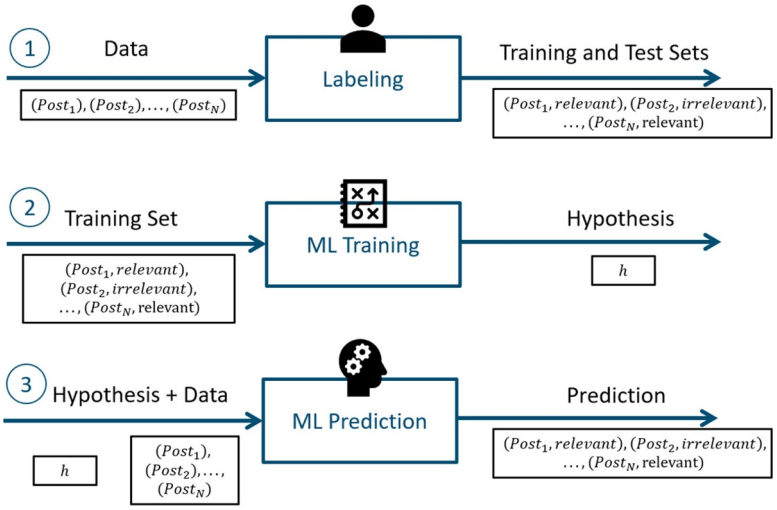

Fig. 1. Supervised machine learning steps.

2.5. Research gap

During disasters and emergencies, social media allow for the creation of large but potentially noisy volumes of citizen-generated content, often referred to as big social data (Moi et al., 2015; Olshannikova et al., 2017). Emergency services may extract actionable or situational information from eyewitness reports, photos or videos (Zade et al., 2018), yet due to limited personal or organizational resources (Reuter et al., 2016) the issue of information overload constitutes a severe problem (Plotnick & Hiltz, 2016). Here, technology might assist in overcoming information overload, both in terms of algorithmic quality as well as usable and tailorable user interfaces (Plotnick & Hiltz, 2018; Stieglitz et al., 2018). From an algorithmic point of view, research examined the potentials of alert generation (Adam, Eledath, Mehrotra, & Venkatasubramanian, 2012; Avvenuti, Cresci, Marchetti, Meletti, & Tesconi, 2014; Cameron, Power, Robinson, & Yin, 2012; Párraga Niebla et al., 2011; Purohit, Castillo, Imran, & Pandey, 2018; Reuter, Amelunxen, & Moi, 2016), event summarization (Nguyen, Kitamoto & Nguyen, 2015; Rudra, Ghosh, Ganguly, Goyal, & Ghosh, 2015, 2018), precise search keyword selection (Abel et al., 2012; Johansson et al., 2012, 2012), and relevance classification (Abel et al., 2012, 2012; Habdank et al., 2017; Li et al., 2017, 2015; Markham & Muddiman, 2016; Nguyen et al., 2016; Verma et al., 2011) for mitigating information overload.

Although research developed a variety of relevance classification algorithms, these face practical issues in real-world disaster and emergency scenarios: at the beginning of such events, there is a lack of labeled data which is required in considerable quantity for supervised machine learning algorithms to perform well (Imran et al., 2017; Li et al., 2015). Research suggests that active and online learning potentially mitigate that issue: Active learning is capable of reducing the required amount of labeled data (Settles, 2010), thus reducing the effort for emergency services, while online learning facilitates the continuous improvement of the classifier's accuracy, including the rapid adaptation to changes and new patterns in the data (Imran et al., 2018). To harness these potentials, we implemented a real-time evaluation mechanism using online and active learning (Section 5.1.1) and a feedback classification mechanism which uses a batch learner with feature subset selection for fast classifier retraining (Section 5.1.2).

From a user interface point of view, the literature revealed criteria whose consideration and display could assist the labeling and classification of relevant messages. These include currency (Ludwig et al., 2015; Palen et al., 2010; Rohweder et al., 2011; Sriram et al., 2010), geolocation (de Albuquerque et al., 2015; Ludwig et al., 2015; Reuter et al., 2015; Sriram et al., 2010; Starbird & Palen, 2004; Verma et al., 2011; Vieweg, 2012, 2010), keywords (Abel et al., 2012; Johansson et al., 2012, 2012) and language (Abel et al., 2012, 2013; Sriram et al., 2010), amongst others. Also, usable interfaces are required to exploit the potentials of active and online learning. Thus, we integrated our algorithms into a user interface (Section 3.4).

3. Technological basis: social data management and analysis

To facilitate relevance classification, we integrated active, incremental and online learning into our social media platform whose design method (Section 3.1) and architecture (Section 3.2), comprising the Social Media API (SMA) as a backend (Section 3.3) and the Social Media Observatory (SMO) as frontend (Section 3.4), are described in the following as relevant prerequisites for the understanding of our evaluations in Sections 4 and 5.

3.1. Method

The architecture underwent several iterations of development before integrating machine learning components. The development of the SMA is not documented in a research publication, since it was primarily driven by the requirements of the SMO. However, we

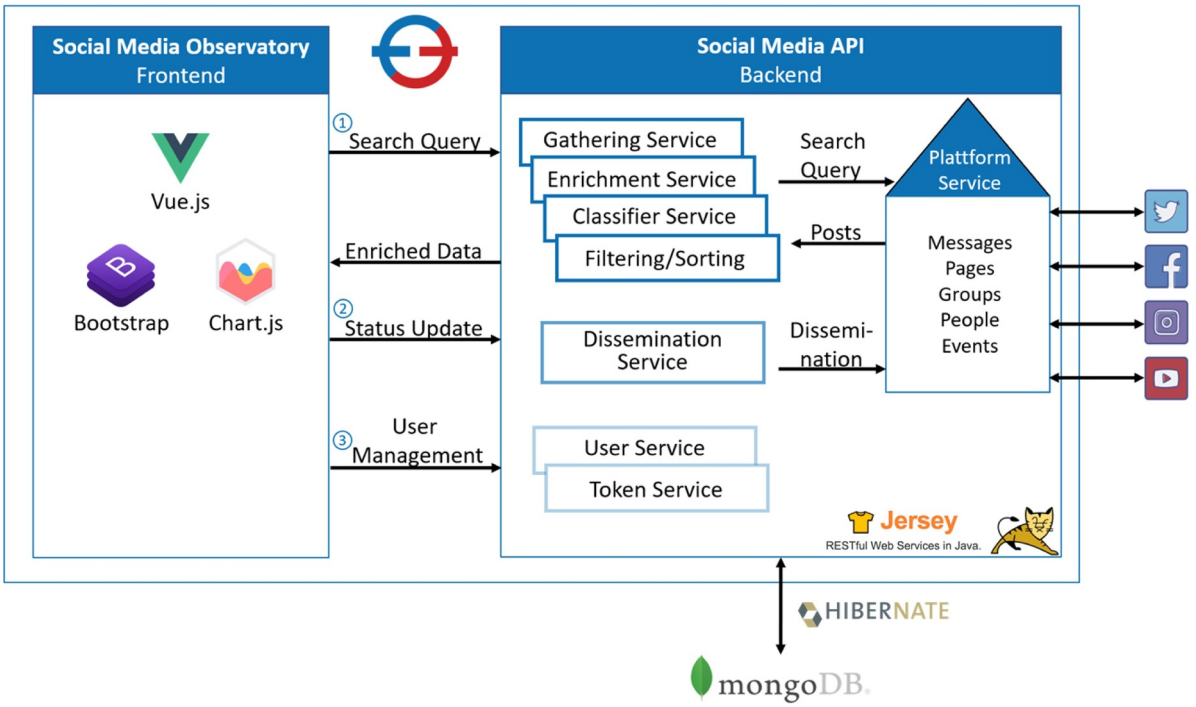

Fig. 2. Social Media Architecture comprising Social Media Observatory and Social Media Service.

provide a detailed description of its design, also to convey the necessary steps of pre-processing required for the implementation of our classifiers. The SMO underwent three iterations of development based on *user-centered evaluations* which are documented in a research paper (Reuter et al., 2016). Based on the outlined design requirements, a fourth iteration was conducted to (1) implement the design requirements, (2) upgrade it with state-of-the-art libraries and technologies as well as to (3) improve its hedonic and pragmatic quality. Besides its basic functionality, we describe the integration of active real-time data labeling and feedback classification, which are evaluated in Section 5, into the graphical user interface.

3.2. Overall architecture

The overall architecture (Fig. 2) comprises a frontend called *Social Media Observatory* (SMO) and a backend called *Social Media API* (SMA). The SMO is a web application based on *Vue.js* as the overall framework, *Bootstrap* for responsive design and *Chart.js* for data visualization. All actions of the SMO, such as searching for posts in social media, disseminating posts to social media or managing users, are forwarded to the SMA. The SMA is realized as a service following the paradigm of a web-based and service-oriented architecture (SOA). It is a Java Tomcat application using the Jersey Framework for RESTful Web services and the MongoDB database via Hibernate Object/Grid Mapper (OGM) for document-oriented data management. The implementation allows for using a local or remote instance of MongoDB. Several libraries facilitate the integration of social media source APIs such as Facebook Graph API or Twitter Search API. To overcome the diversity of data access and structures, all gathered social media entities are processed and stored according to the ActivityStreams 2.0 Core Syntax in JavaScript Object Notation (JSON). In the following subsections, we will give a brief overview of the Social Media Service (Section 3.3) and Social Media Observatory (Section 3.4).

3.3. Backend: the social media service

The *Social Media API* (SMA) allows for the gathering, processing, storing, and re-querying of social media data. Based on underlying social media, the SMA contains different services that are used by several client applications. Although it was developed as enabling technology for crisis management applications, its implementation allows for supporting a variety of use cases in different fields of application, e.g., to examine the impact of a product image within the field of market research. To enable access to social big data and allow for subsequent analysis, our first step was to specify a service for gathering and processing social media content. With *gathering*, we refer to the ability to uniquely or continuously collect social media activities (e.g., messages, photos, videos) from different sources (Facebook, Google+, Twitter, and YouTube) in a unified manner using multiple searches or filter criteria. *Processing* means that the SMA is able to access, disseminate, enrich, manipulate and store social media activities.

3.3.1. Endpoint overview and pre-processing

SMA comprises five main services, each providing a multitude of service functions: The *Gathering Service* contains endpoints for

Table 4
Required and optional query parameters.

Parameters	Type	Description
keyword	String	Required. The search keywords.
sources	String	Required. A list (Facebook, Google+, Instagram, Twitter, YouTube).
since/ until	Long	Search Service. Lower/upper bound of the searched timeframe (Unix time).
start/ end	String	Crawl Service. Starting/termination point of the crawl job (Unix time).
latitude/ longitude	Double	Latitude/longitude for geo search (decimal degree).

gathering and loading social media activities. The main components are the Search Service, enabling one-time search requests, and Crawl Service, which continuously queries new social media activities across a specified timeframe. Using the *Enrichment Service*, gathered social media activities are enriched with further computed and valuable metadata. For the initialization of a crawl job, a POST call is sent to the *crawlService* endpoint, which contains a payload matching the *application/json* Content-Type header. In its basic configuration, a keyword, at least one source and an interval value to determine the timeframe between each gathering request are required; further configuration parameters may be specified to filter the scope of the crawl job (Table 4). To query gathered results of a certain crawl job, a GET call is sent to the *crawlService/{crawljobId}* endpoint with *{crawljobId}* being a concrete instance of an identifier. The identifier may be retrieved from the response of the crawl job's initialization or from the list of the *crawlService/ allJobs* endpoint.

The *Classifier Service* realizes our supervised machine learning components. Before the classification can take place (see Sections 4 and 5), several steps of pre-processing are required. Firstly, the removal of specific characters is realized with simple Java operations. The CISTEM algorithm is used to stem German words (Weißweiler & Fraser, 2017) and the lemmatization is realized via the LanguageTool library (LanguageTool 2019). During the training phase and for the translation of words into machine-readable characters from all training tweets, a modified Bag-of-Words (BoW) is created. In Java, this is realized as a hash map, where every word as a key is mapped to a number. With regard to efficiency and Inverse Document Frequency (IDF), a word in the training BoW is mapped to the number of documents that comprise the respective word. For each word in the BoW, a feature for each post is acquired. Thus, for each tweet the algorithm iterates through the BoW and generates a vectorized number for each of these words. There, classifier supports two approaches to compute these numbers. Firstly, the *normalized term frequency* (TF), where for each word its frequency in a tweet is divided through the maximum frequency of any word in the tweet. Secondly, this normalized term frequency can be extended to TF.IDF by multiplying the TF with IDF. To determine whether a geolocation is available in a post, the German version of the CoreNLP toolkit (Manning et al., 2014) is used.

Furthermore, the algorithm evaluated in Section 4.2 requires the computation of geographical and temporal distances between any post and the event's, e.g., a disaster or emergency, location and time. The geographical distance is based on the latitude and longitude of the authors and posts location: *distance in kilometers* $= sqrt(dx^2 + dy^2)$, *whereby* $dx = 71.5 * (\text{longitude}_1 - \text{longitude}_2)$ *and* $dy = 111.3 * (\text{latitude}_1 - \text{latitude}_2)$. Since the geolocation of an author is often not represented in latitude and longitude but rather in a textual description, the geocoding REST interface of “here” is used (here 2019). The temporal distance is computed by the *Duration.between()* function of the *Java.Time* library. Finally, the Naïve Bayes and Random Forest classifiers are implemented using the Java library Weka (Hall et al., 2009). All processed data is stored in *Attribute-Relation File Format (ARFF)* files, which can be processed by Weka learning algorithms. Such a file comprises a list of all used features as well as their internal representation (see Table 6).

The *Dissemination Service* allows for the publishing of messages in social media. Finally, the *Source Service* constitutes the interface to individual social media sources and allows for the integration of social media in a standardized manner.

3.3.2. Data specification using activity streams 2.0

The Activity Streams 2.0 Core Syntax (AS2) defines that “an activity is a semantic description of potential or completed actions” (World Wide Web Consortium 2016), which has at least a verb (the type of activity, e.g., like, post, share), an actor (e.g., the creator) and an object (e.g., an image or message object). There are already many verbs and object types defined within a specification, for instance, a place object may contain the attributes latitude, longitude, and altitude (Table 5). Although the specification allows modeling the activities of liking, sharing and so on, there are no attributes designated to carry information like “20 users liked this post”. While the specification may be extended with own verbs and object types such as our “enrichedData” object, foreign implementations possibly have not enough knowledge to process them in an intended way. Activity objects must be encapsulated in a collection object before returning them as a JSON object.

3.3.3. Data storage using HibernateOGM/MongoDB

For storing and retrieving the collected data in and from a MongoDB database instance, we deploy the Java framework Hibernate OGM. We selected MongoDB as a document-oriented NoSQL solution due to its good performance in reading, writing and deleting operations on large datasets and, compared to SQL solutions, flexible document schemas and the option of *sharding*, a method for distributing data across multiple instances or machines (Li & Manoharan, 2013). Furthermore, with the aid of the OGM tools (object-grid mapper), we could operate without direct database commands because they are encapsulated in the framework, e.g., as save, update or delete functions. The generation of the database scheme and the storage of the corresponding instances of the objects are done automatically. Only the annotations of the appropriate classes and their attributes are needed for Hibernate to transform the Java classes into database query commands. Moreover, we use a compound unique index based on the activity's source and ID to

Table 5
ActivityStreams with EnrichedData object.


```
{
    "actor": {
      "content": "48, 2 Kinder, Sarkasmus, private Meinung",
      "displayName": "anonymised",
      "id": "twitter:844,424,271″,
      "type": "person",
      "url": " https://goo.gl/QqV2q6"
},
    "object": {
      "content": "RT @bzberlin: #Debüt mit 1:0 gegen @SERCWildWings https://t.co/UNlq698PlJ",
    "enrichedData": {
      "absFearFactor": 0,
      "absHappinessFactor": 0,
      "embeddedUrls": ["https://t.co/UNlq698PlJ"],
      "language": "de",
      "tags": ["Debüt"],
      "media": [{
        "mediaType": "image/jpeg",
        "type": "photo",
        "url": "https://goo.gl/QqV2q6"
      }],
      "mentions": ["bzberlin", "SERCWildWings"],
      "numOfCharacters": 133,
      "numOfSentences": 7,
      "numOfWords": 11,
      "numRetweets": 3,
    },
    "id": "twitter:823724465664883940″,
    "location": {
      "displayName": "Neunkirchen, Deutschland",
      "latitude": 50.78506988,
      "longitude": 8.00512706,
      "type": "place"
    },
    "startTime": "2017-02-01T10:30:47.000+01:00″,
    "type": "post",
    "url": "https://goo.gl/QqV2q6"
  }
}
```


prevent duplicate activities on database level.

3.4. Frontend: the social media observatory

The *Social Media Observatory* (SMO) is a user interface which utilizes the SMA for social media monitoring, analysis, and relevance classification (Fig. 3). More specifically, it facilitates the creation of continuous search jobs (Crawljob tab) and single-time searches (Search tab), management of users (Admin tab), creation of classifiers (Classifier tab) as well as dissemination of messages (Dissemination tab).

3.4.1. Data gathering and visualization

After entering the SMO, the dashboard view (Fig. 3) visualizes the number of posts crawled, crawljobs created, classifiers learned, users created and comprises an overview of functionality.

After entering the Crawljob tab, the user has an overview of his own crawljobs, (Fig. 4), each displaying the search keyword, user-based description, author of the job, location, sources, results and creation time, and is able to create new crawljobs (Fig. 5) based on keywords, description, selection of social networks, the definition of geographic boundaries (optional) and a time period (optional). The user can visit the results as a list view (Fig. 8), each entity containing the time of creation, username, content (text, photos, videos), and metadata (language, location, retweets). Furthermore, the Search tab enables the user to perform single searches based on keywords and selected social media. The Admin tab allows administrators to manage existing users or create new users with a name, password and role (e.g., admin or user).

3.4.2. Data labeling and filtering

For each crawljob, the user may label postings according to their relevance (Fig. 6). In the beginning a single post with its content, multimedia files and metadata is loaded into the view. After the user labeled it as ‘relevant’ (green button) or ‘not relevant’ (red button) the next post is loaded. After each fiftieth post the estimated accuracy of the classifier (see Section 5.1.1 for the specification

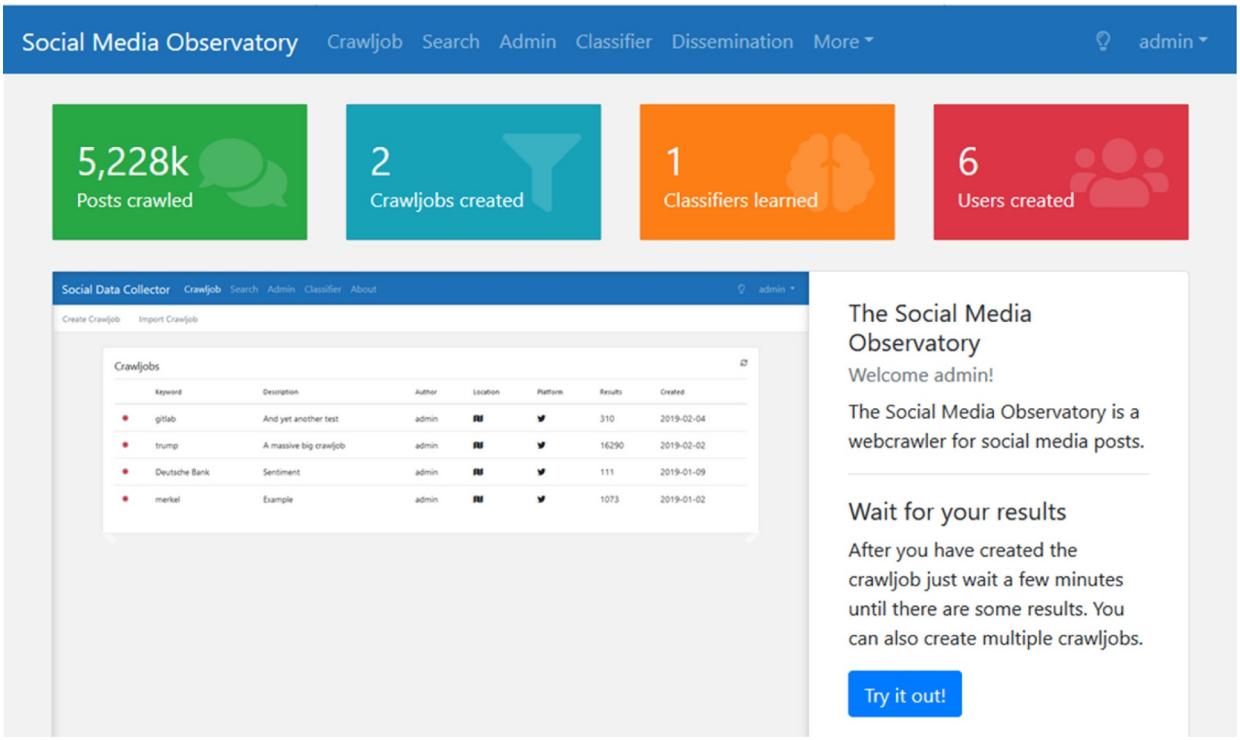

Fig. 3. Dashboard overview for administrators.

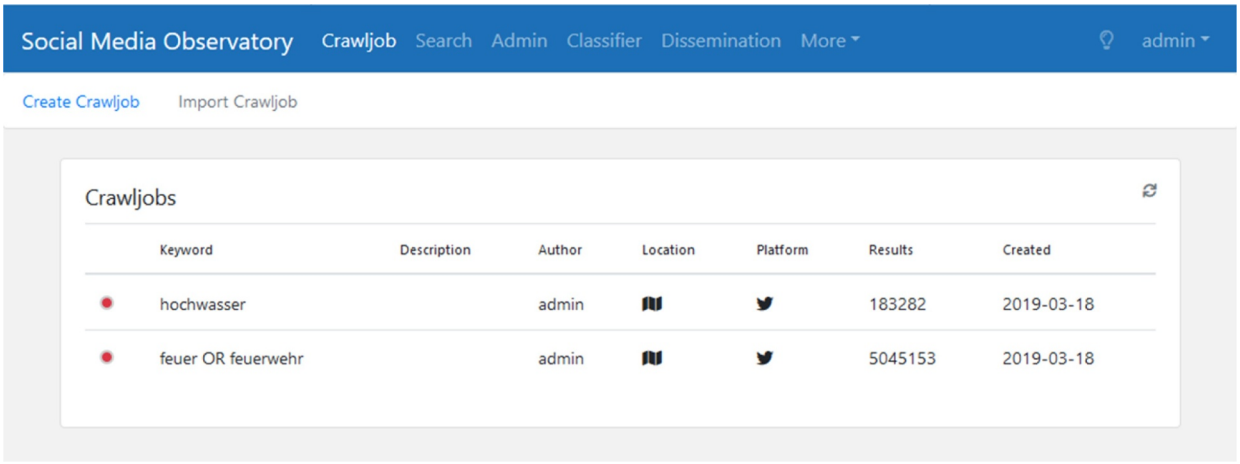

Fig. 4. Overview of crawljobs.

of the real-time evaluation approach) is indicated by a time chart, highlighting the development of the classifier's accuracy over time as well as a pie chart, which shows the classifier's accuracy in percentage but also its precision and recall on mouseover. This way the user can track whether a learning process takes place or if the learning curve converges.

After labeling a sufficient number of posts the user may create a classifier based on a crawljob (Fig. 7). The user is required to enter a name, description and the underlying crawljob of the classifier. Optionally, he can define an event location and event date as geographical and temporal features for the classifier.

By entering the post list of a crawljob the user is able to apply the trained classifier (Fig. 8). In the default configuration 'not relevant' posts are greyed out. However, if the user disagrees with the classification, he can mark a 'relevant' post as 'not relevant' (by a tick in the top-right corner of each posting) and vice versa. As soon as the user leaves the list view or navigates to the next page of results the classifier is retrained accordingly (see Section 5.1.2 for the specification of the feedback classification approach).

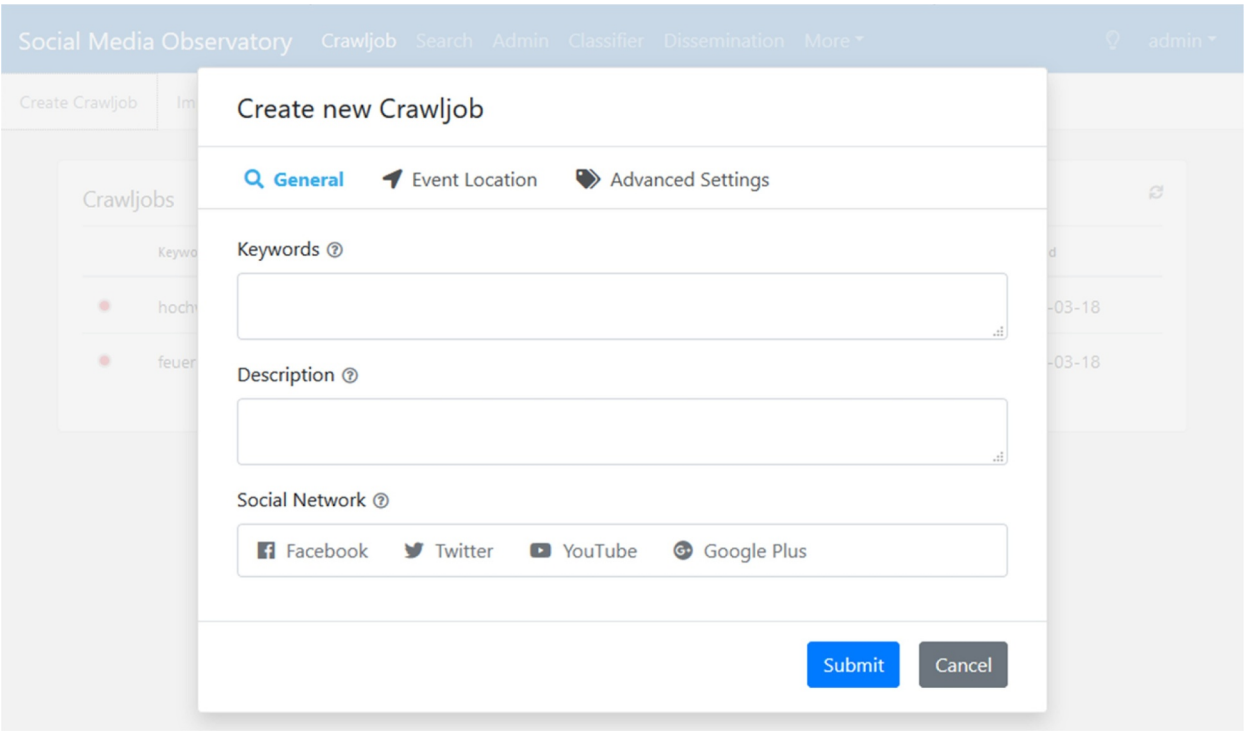

Fig. 5. Creation of a crawljob.

Fig. 6. Relevance labeling for a crawljob.

4. Evaluation I: relevance classification via batch learning

To achieve insights into **RQ2 ("How can existing supervised machine learning techniques for relevance classification be improved for use in real disaster and emergency environments?")**, we conducted our first evaluation whose approach (Section 4.1) and conduction (Section 4.2) are described in the following.

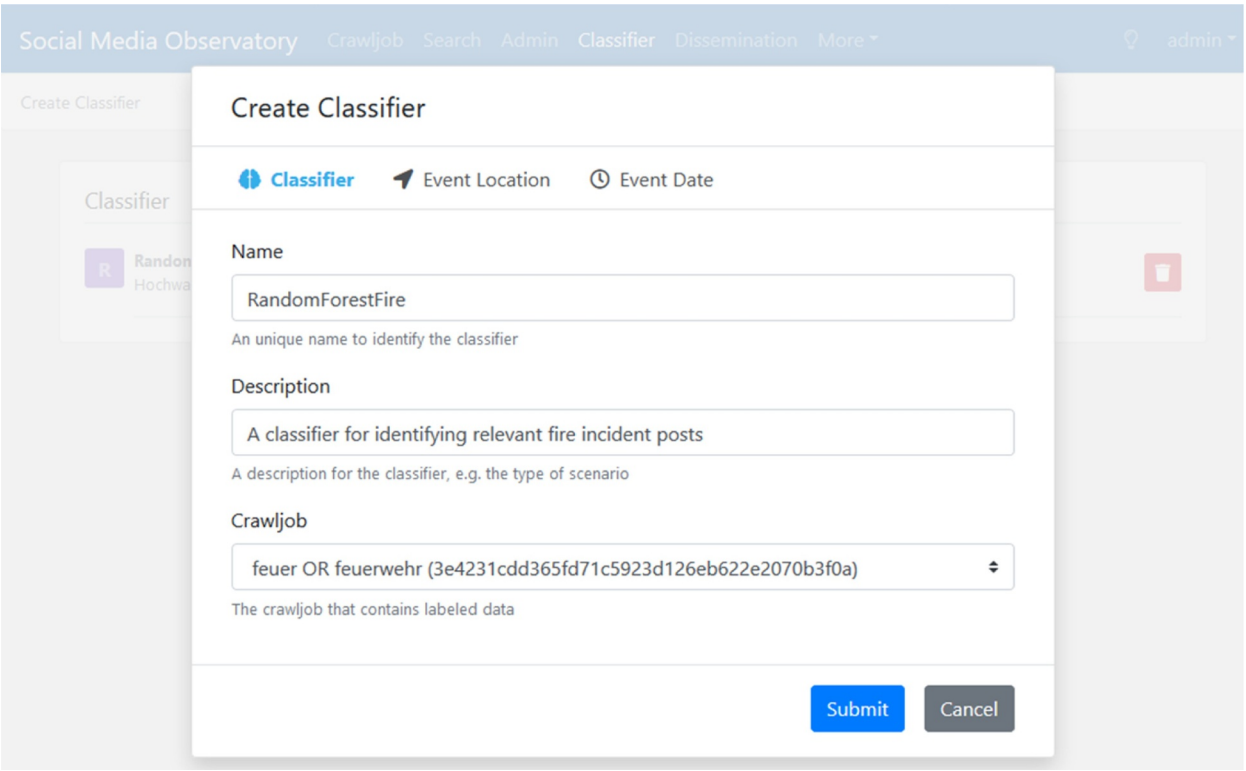

Fig. 7. Create a relevance classifier based on labeled crawljob posts.

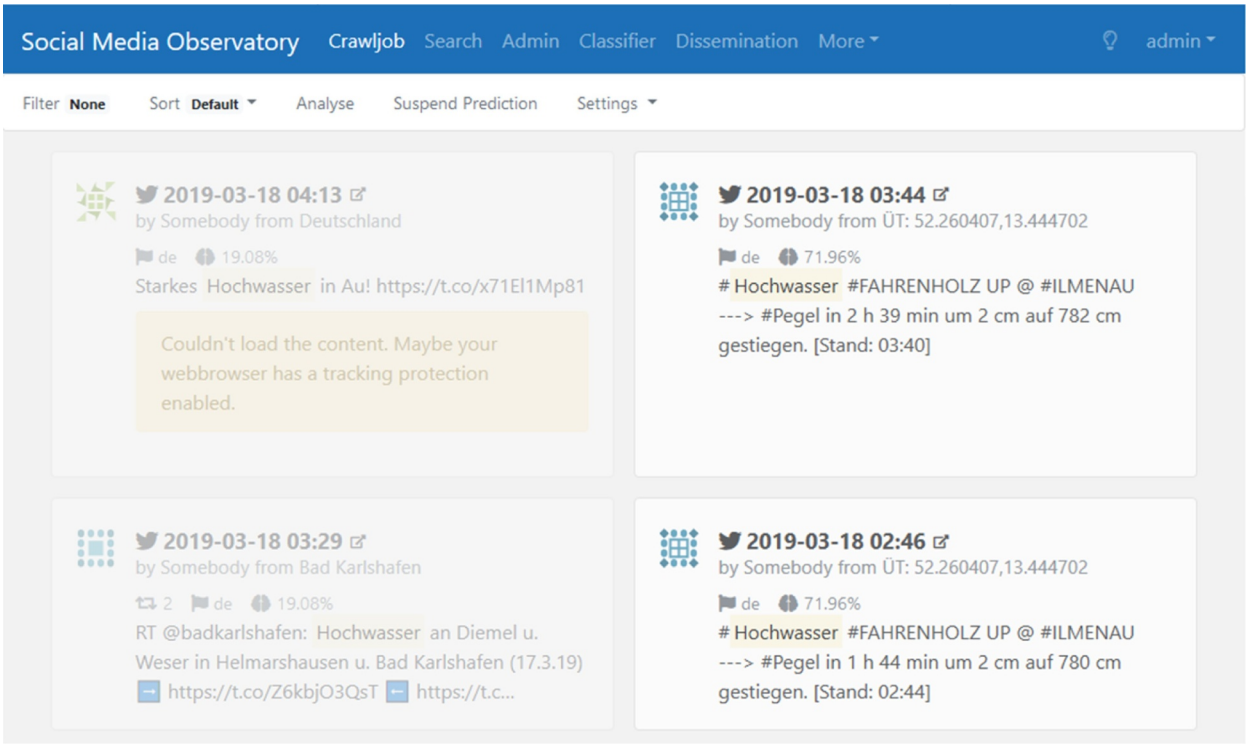

Fig. 8. Filtered List of crawljob posts.

4.1. Approach

Information is relevant when it is useful, valuable, appropriate and necessary (Jensen, 2012; Rohweder et al., 2011). To do justice to these characteristics, one must consider the exact context (Agarwal & Yiliyasi, 2010). In the following and based on the definitions and criteria from Section 2.3, the contextual relevance related to crises or emergency situations serves as the foundation for labeling and the classification features of the machine learning algorithm. As a result, the term relevance as defined in this paper refers to C2A content, which contributes to information acquisition of and decision support for emergency services. Content of this kind is often up-to-date, refers to the location of the incident, contains facts and references, is retweeted several times and may contain hyperlinks. It is to be stressed that relevance, as described in this paper, must be separated from an all-encompassing information quality. To obtain a high density of high-quality information, customizable filters are of greater use after the classification of relevance. For example, Reuter, Ritzkatis, and Ludwig (2014) present a customizable rating service in which end-users can adjust weightings of several characteristics regarding their information quality. The classification of relevance can thus be a first step towards maintaining information quality.

As the labeling is done by humans, an interpretative and analytical approach is possible (in contrast to the classification feature selection). Therefore, the criteria from Tables 1 and 2 can be used for defining relevance in the labeling process (Section 4.1.1). The labelers are provided with an abstract definition and precise criteria regarding relevance in disasters. By contrast, an algorithmic relevance classification (Section 4.1.2) can only contain exact and measurable criteria, which can be found in Table 2. Ideally, the algorithm, based on the training data of the labeling process and the elaborated classification features, should predict a post as relevant if it suffices the requirements of relevance as described in this paper.

4.1.1. Relevance features for labeling

The specification and explanation of relevance is important for two aspects of the labeling process for training data. On the one hand, this is the only way to ensure a comprehensive insight into the classifier developed here. On the other hand, several persons who have reached a consensus regarding this clarification can participate in the creation of training data.

In their analysis of tweet behavior in disasters, Verma et al. (2011) found that objective, impersonal and formal information characteristics tend to be included in their definition of relevant content. However, such a classification would disregard a large number of important tweets. For example, a tweet can be written in a personal or in an informal style and contain an opinion but still provide important information for emergency services. A witness of an incident might not have the time to write a formal post but might be able to share important information due to the short distance and high timeliness. Cheong and Lee (2011) also point out that the civilian sentiment during terror attacks can be of great use for decision-makers and authorities. The concept of relevance as defined in this paper, in contrast, complies with the description of Vieweg (2012). The labelers label posts as relevant when information can contribute to decision-making, contain advice or when important sources are referenced. A further approach of determining relevance is the distinction between informative, personal and other tweets by Imran et al. (2013). According to them, a tweet is considered to be informative when it is of interest to people beyond the author's immediate circle. Hence, their labeling process sorts out all tweets which are of interest to the author and his/her immediate circle of family/friends only and do not provide any useful information to people who do not know the author.

Imran et al. (2013) further describe that all tweets not written in English are sorted out. However, language is not perceived to be a restriction for the concept of relevance in this paper. For example, Wilson, Stanek, Spiro, and Starbird (2017) analyze online rumoring behaviors in crises based on languages. They point out that the main language of the affected population, independent of English, is important in the context of the event. Finally, they state that language is an aspect that requires further consideration when conducting studies of social media. Imran et al. (2013) further describe that an exact distinction between informative and personal content was not always clear to the labelers. In this paper, a tweet is always considered to be relevant in case of doubt. For the emergency services to use the data in a beneficial way, an all-encompassing and strong reduction of the number of posts is helpful but can lead to a loss of relevant content. As in the scenario of Markham and Muddiman (2016), a certain amount of posts incorrectly classified as relevant is acceptable. Furthermore, tweets are tested to see whether or not they actually contribute to situational awareness in crises. As outlined by Vieweg (2012), all tweets containing condolences or calls for donation only are classified as irrelevant. Content is further classified as relevant if the information it contains is of use or important for the Virtual Operation Support Teams (VOST). The role of virtual communities, as described by Reuter et al. (2013) as well as Kaufhold and Reuter (2016), should also be taken into account. Helpers and reporters directly or indirectly provide important and up-to date information on the incident. Retweeters and repeaters distribute the most important information (which was generated by other users), thus filtering important information that should not be overlooked. Such content is often associated with facts. Labelers ensure that all types of facts related to a crisis are labeled as relevant. False facts can also provide important information for authorities. They can identify the false information and correct it using A2C communication (Reuter et al., 2018). Furthermore, labelers label posts in which help is requested or offered as relevant.

For each tweet, the assessors are provided with an overview of various characteristics of it, which do not have to be extracted from the Twitter message itself (Section 3.4.2, Fig. 6). These characteristics may be helpful in identifying relevance but should not be seen as exclusive criteria. From the definition of an event by Sriram et al. (2010), the criteria geographical location, referenced location and timeliness of a post emerge. All three criteria are important for filtering the relevance for VOSTs. While the referenced location in a tweet indicates a connection of the author to the incident, the local proximity can even point to witnesses, helpers and people in need of help at the incident's location. Currency is important because, for example, messages clearly sent before the incident can be irrelevant. Therefore, the assessors are given the location of a tweet and the calculated local proximity to the incident.

However, as only 10% of all Twitter users worldwide allow the locating of a tweet, as described by Ludwig et al. (2015), the assessors are additionally advised to analyze the tweet for referenced locations. Furthermore, the example dataset from Section 4.2.1 shows that many users indicate the federal state they live in their user profile. As this can also indicate local proximity to the post, this information is also shown. In addition, the assessor is provided with the date and time of the tweet as well as with the temporal distance of the post from the incident. It is especially important to determine whether a tweet was written before or after the incident, as in some emergency situations no information is available shortly before or until the emergency services arrive.

As shown in Table 2, the number of retweets of each post is made available to the assessors. The role of retweeters, more specifically the observation that retweet behavior may indicate relevance, as described by Reuter et al. (2013), is supported by examinations of Starbird and Palen (2004), Uysal and Croft (2011) and others. However, the following tweet from the 2017 Las Vegas shooting illustrates that the assessors should not treat this criterion as an exclusive indication for relevance: *"[..] As I #prayforlasvegas I pray for us all. Find each other out there.…* https://www.instagram.com/p/BZwx8oVle7s/" (Perry, 2017).

Katy Perry published this post on October 3, 2017, to communicate her condolences to the victims and their families of the Las Vegas shooting. So far, the tweet has received about 2.800 retweets and can thus be considered as a very often shared post of this incident. However, in terms of content, this tweet is not relevant for the emergency services. The high number of retweets can be explained by the enormous fan base of Katy Perry. In addition to the number of retweets, the number of likes/favorites is also submitted to the assessors. To like a tweet does, however, not necessarily mean that the person likes the content of it.

As described by Gorrell and Bontcheva (2016), Twitter users also use a like/favorite, for example, to say "thank you" or to create a reminder for themselves. Assessors also see extended content in the form of pictures, videos and URLs. This extension allows Twitter users to bypass the 280-character limitation on Twitter and is therefore likely to be of importance for the emergency services, as described in Section 2.3. For example, pictures and videos can provide important information about burning buildings. URLs can refer to blogs or other posts in social media which might as well provide necessary information.

Finally, the number of words and letters is shown to the assessors, whereby the lower bound to label a tweet is selected by the assessors themselves. Abel et al. (2012) chose a very high lower limit of at least 100 characters per relevant post. For example, the following tweet contains only 88 characters, but according to Verma et al. (2011), it is a relevant post related to the Red River floods in 2009: *„Country residents outside of Fargo are surrounded by flood waters. Some R being rescued"* (Wise Bitch, 2009). The minimum requirement of Sriram et al. (2010) of three words per Twitter message seems to be a well-chosen parameter regarding the classification of relevance in crises.

4.1.2. Relevance features for machine learning

For machine learning, abstract and interpretative relevance criteria are not practicable. Precise and measurable criteria must be defined for the algorithm to be able to conduct a classification of relevance. The selected classification features form the resulting classifiers and define the concept of relevance for machine learning. To form a basis of this concept, the criteria from Table 2 are discussed. Also, additional features are considered which do not necessarily have to indicate relevance, as the selected algorithm in Section 4.2 is evaluated for various classification feature set combinations.

Keywords or terms are essential for identifying relevance in crises. Search-filtering takes place before the actual relevance classification so that the amount of user-generated content is reduced to a manageable subset. The further textual content of a tweet is particularly important for identifying relevance for humans as well as for machines. As in the publication of Vieweg (2012), a sorting out of tweets containing certain terms, for example, relating to condolences, could be considered. However, it was decided for this paper that the tweet's content or the sequence of words which the algorithm can weight as positive or negative is sufficient as a feature.

The classification results are also used to test whether the results of the classifier can be improved by text-based *Named Entity Recognition* (NER). Named Entities are phrases that contain the names of persons, organizations or places (Tjong Kim Sang & De Meulder, 2003). Especially the identification of the location can be of great importance. Even though the algorithm is given the distance of the tweet to the incident, tweets rarely contain location information. Here, NER can help to identify further locations and thus support the computation of geographical distances of the author and posts in relation to the incident's location. Furthermore, the collection of Twitter posts takes place within a time frame. However, as collected messages may lie in the past and the time frame may not always match, the difference between the creation of the tweet and the time of the incident is another criterion which has to be taken into account by the classifier. In the classification, the retweet behavior can also indicate the relevance of a tweet. Yet, since the procedure described in this paper is based on many other characteristics, the tweet example of Katy Perrey would probably still be excluded. Moreover, links in the form of URLs, videos or picture can be important for the emergency services. This information is versatile and exceeds the limitation of 280 characters per tweet. An analysis of the content of pictures, videos or Internet pages is, however, not included in the classifier. Furthermore, the length of a tweet is of great importance for the machines and should not be underestimated. Too short tweets (based on the number of letters) are classified as less relevant by the algorithm. Also, as mentioned by Abel et al. (2012), there is the assumption that the underlying language of a tweet may be relevant. In view of the globalized world, a number of important issues could be overlooked. The algorithm is given the original language as an independent classification feature but the actual text is translated if it was not written in German. Table 6 shows the previously mentioned classification features of the algorithm in a compact design.

4.2. Evaluation

In the evaluation, various classification features of a classifier detecting a tweet's relevance in disasters and emergencies are

Table 6
Features, description and internal representation in the classifier (* means optional, [] means that it is no parameter feature).

Feature	Description	Internal representation
[Keyword filtering]	Pre-filtering by keywords of the incident	Social Media API Crawljob – Collections
[Temporal filtering]	Pre-filtering with regard to the time of the dataset	Social Media API Crawljob – Collections
Content	Word sequence based on unigrams	TF and TF-IDF values based on a Bag of Words of the training data – Double (per word)
NER	Extraction of persons, names, and locations from text	Availability of a geolocation - Boolean
Post distance*	Geographical distance between the post and the event	Kilometer – Double (−1.0, if no data available)
Author distance*	Geographical distance between the author and the event	Kilometer – Double (−1.0, if no data available)
Temporal distance	Temporal difference between the post and the event	Minutes – Double
Retweet count	Number of repetitions of a tweet by other authors	#Retweets – Integer
URLs	The availability of links	Availability – Boolean
Media	The availability of images or videos	Availability – Boolean
Length of post	–	#Characters – Integer
Language	–	Country code – Nominal

tested. Thus, different pre-processing steps (Section 3.3.1) and subsets of all classification features and methods have to be considered. In the process, approaches from previous work are combined to achieve a high-performance solution. For example, Habdank et al. (2017) and Verma et al. (2011) evaluate various algorithms with purely textual classification features, while Imran et al. (2013) test one algorithm with several different classification features. In the evaluation procedure, two different algorithms with several classification features are tested for effectiveness and – in contrast to previous work – efficiency. The test is based on an event that is used as a proxy for further scenarios. The evaluations are performed on a computer with Windows 10, Intel i7-7500 with two physical and two logical cores at 2.90 GHz and 12 GB Ram in Java.

4.2.1. Method

As classifiers require training data in the form of labeled or coded text (from which they learn how to distinguish between different types of discourse), all tweets are pre-processed and stored in an ARFF file as a training dataset at the beginning of each evaluation step. Based on this file, a classifier is trained. Depending on what is to be tested, each evaluation step inherently adapts the set of characteristics, parameters of the algorithm or the algorithm itself.

4.2.1.1. Datasets. This study is based on two datasets, which were qualitatively assessed and labeled as relevant or irrelevant by a single expert per dataset. The first underlying dataset for the test is based on the 2013 European floods but focusing on German data in this case (Reuter, Ludwig, Kaufhold, & Pipek, 2015). As reported by the German Federal Agency for Civic Education (bpb 2013), there was a lot of commitment from volunteers. Most notably, social media were used to disseminate information and offer help. According to this information, many appeals for donations were made via the Internet, which were labeled as irrelevant.

A total of 3923 Twitter messages sent over a period from 30 May to 28 June 2013 were collected and classified. This corresponds approximately to the amount of data analyzed by Habdank et al. (2017) and Imran et al. (2013). There are 1626 relevant and 2297 irrelevant tweets. In this case, all texts were written in German. However, content not written in German was translated. As can be seen from the dataset, a large number of Twitter users provided status updates on the current water level in addition to offers of and requests for help.

For another evaluation and interpretation of the different phases, we use a dataset from the incident of BASF SE in Ludwigshafen on October 17th, 2016. A facility of BASF caught fire because of working on a pipeline route. The fire lasted 10 h and was extended by two explosions (Habdank et al., 2017). Five people lost their lives and 28 people were seriously injured.

This dataset was already used for relevance classification by Habdank et al. (2017). They only relied on textual information of the posts. It contains 3790 posts with 1816 being relevant to the accident. Unfortunately, about 5% of the posts only have textual information and no metadata linked to them. There are even more posts that are missing only some of the metadata.

4.2.1.2. Cross-validation. As in Verma et al. (2011) and Habdank et al. (2017), a 10-fold cross-validation was used for analysis. Whereas in a regular validation the data is partitioned into a training set (to train the model) and a test set (to evaluate it), in a 10-fold cross-validation the data is randomly portioned into ten equal subsets (Habdank et al., 2017). Of the ten subsets, a single subset is retained as the validation data for testing the model and the remaining 10–1 subsets are used as training data. The cross-validation process is then repeated ten times, with each of the ten subsets used exactly once as the validation data (Hastie, Tibshirani, & Friedman, 2009, pp. 241–249). This prevents certain patterns from occurring which could falsify the classification quality.

4.2.1.3. Criteria. All evaluation combinations are evaluated using the criteria *accuracy, recall, precision* and *time (in seconds)*. Accuracy indicates the percentage of tweets correctly classified as relevant or irrelevant by the classifier. This means that all tweets correctly predicted as relevant (true positives (TP)) and all tweets correctly predicted as irrelevant (true negative (TN)) are divided by all tweets available (Habdank et al., 2017).

Recall is the ratio of tweets predicted as relevant to all tweets classified as relevant (POWERS, 2011). The number of all tweets classified as relevant can be expressed by the number of TP plus the number of tweets incorrectly predicted as irrelevant (false negatives (FN)).

Precision, in contrast, describes the ratio of all TP to all tweets predicted as relevant. The number of all tweets predicted as relevant can be expressed by the number of TP plus the number of tweets falsely predicted as relevant (false positives (FP)) (POWERS, 2011). This measure is important to minimize the number of FP. In the context of crisis informatics, many false positives mean that many irrelevant tweets were reported to the emergency services. Since this would result in a greater time consumption, the recall unit should be weighted higher in the following examination. The higher the recall, the fewer tweets were falsely predicted as irrelevant (Habdank et al., 2017). Minimizing the FN is particularly important as otherwise valuable information could be lost; in the context of crisis informatics, this could cost lives (Habdank et al., 2017).

The time consumption of classification is particularly important for the use in a productive system. Emergency services should be able to conduct a relevance classification as quickly as possible without being hindered by a long, internal pre-processing or classification process. In the following, time data refer to a complete analysis run including pre-processing (Processing-Duration (PD)), training (Training-Duration (TD)) and 10-fold cross-validation (Validation-Duration (VD)). Thus, time data do not indicate the training or classification in particular. However, conclusions can be drawn for its use in a productive system.

4.2.1.4. Steps. Given the numerous possibilities to achieve a high quality and temporal efficiency of the algorithm, a systematic approach is necessary. In the text processing phase, the text processing with individual options is tested. The feature set phase consists of a classification feature analysis in which a classifier is tested on various feature combinations. The parameter optimization phase is constitutive for the setting of the learning algorithm. In the algorithm phase, two classification algorithms are compared. Finally, in the fifth phase, a filter approach is used to minimize the vast amount of classification features to the most important subset of it.

- **Text Processing Phase:** In the first phase, text processing options are tested with a Random Forest. The vectorizations Term Frequency (TF) and Term Frequency–Inverse Document Frequency (TF.IDF) are compared, stemming is compared to the more complex lemmatization and the appropriateness of Named Entity Recognition (NER) is discussed. As the NER process takes place in the text processing phase, the results are discussed in this phase and not in the classification feature set phase.
- **Classification Feature Set Phase:** In this phase, different classification feature combinations are tested. The best algorithm of the first phase is used as a reference for the text-based Random Forest. Based on this, a combinatorial adding of the classification features number of retweets, length, geographical distance (author distance and tweet distance), temporal distance and the existence of media and URLs follows. In the last test run, the non-textual characteristics are tested to see whether they indicate relevance without the underlying tweet's text.
- **Parameter Optimization Phase:** The best Random Forest of the last phase is optimized regarding various parameters. For example, it is possible to determine the depth or the number of decision trees. Furthermore, the so-called “threshold-moving” is particularly suitable to compensate for an imbalance of recall and precision. The percentage threshold from which a tweet is classified as relevant or irrelevant is changed (Zhou & Liu, 2006).
- **Algorithm Phase:** The best result of the last two phases is used to compare a Naïve Bayes algorithm with Random Forest. The decision for these two algorithms is justified by the fact that they stand out in multiple reference works (Habdank et al., 2017; Imran et al., 2013; Markham & Muddiman, 2016; Verma et al., 2011). We most importantly focus on Random Forests because already other authors, like Habdank et al. (2017), showed that these are well performing. Even if Convolution Neural Networks are also gaining popularity in this research area, we are not considering them for our production system, since they are hard to train and adapt for new datasets.
- **Feature Subset Selection Phase:** The last phase serves as an outlook on possible reductions of the large and complex classification feature set. The Bag of Words (BoW) is generated based on all words of the training dataset. Most of the several thousand words are not substantial for the actual classification and can thus be eliminated. This could potentially result in fewer adaptations in the learning process and a faster classification process. A so-called filter approach is used here: a heuristic ranking algorithm is used to find the best subset of classification features (John, Kohavi, & Pfleger, 1994). An analysis of the resulting classification features can yield information on the intelligence of the algorithm.

4.2.2. Results of the 2013 European floods dataset

It has to be noted that the following results are only substantial for this paper's dataset and the corresponding scenarios. The use of other scenarios and other tweets, e.g., tweets that are not exclusively German, could lead to different results. Comparing the results with related work, the Random Forest of Habdank et al. (2017) achieved the very good results within a similar setting in this research field with an accuracy of 88.7%. Using a Naïve Bayes classifier, Spielhofer, Greenlaw, Markham, and Hahne (2016) achieved an accuracy of 77.2%. However, it is worth mentioning that all algorithms were evaluated using different datasets. For example, the inclusion of many retweets in the test dataset could greatly improve the results due to the analogousness of tweets. Information about the occurrence of retweets was not provided by any author of the related work.

4.2.2.1. Text processing phase. All steps of this phase are evaluated with a Random Forest and exclusively textual classification features. In each run, the words are processed according to the pre-processing explained in Section 3.3.1. Characters, like “\n”, “\r”, double whitespaces and characters that could not be encoded properly, are removed and the tweets are tokenized. The content is not translated since the dataset contains German tweets only. The tests showed that the stop word list of Porter (2019) was most suitable.

Table 7
Classification quality of TF and TF.IDF vectorizations based on a textual Random Forest.

	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
TF	90.8	91.3	81.1	**850.271**
TF.IDF	**90.84**	**91.4**	**81.2**	957.981

Therefore, it is used in all following procedures. If no other information is given, the classifiers are tested with simple stemming and TF.

4.2.2.2. Comparison between TF and TF.IDF. The two vectorizations TF and TF.IDF are implemented according to the specifications in Section 3.3.1. Table 7 shows the classification quality with the percentage results for accuracy, precision and recall as well as the time consumption in seconds. With the TF.IDF unit, the classifier achieves a classification quality minimally higher than with the TF unit with an Accuracy of 90.84%. However, this difference is considered as not substantial. The use of the more complex vectorization takes almost two minutes longer than using the simple TF method. Both approaches would thus prove to be practicable during real-world application. As the TF has proven to be faster and easier to implement, the following test steps are only given in this unit.

4.2.2.3. Comparison between stemming and lemmatization. The classification quality for both word stem creations is also not substantially different, as Table 8 shows. However, as the time for stemming of approximately 14 min is substantially better than the time for lemmatization of approximately 32 min, stemming is preferred in the following tests.

4.2.2.4. Appropriateness of NER. The inclusion of the NER classification feature produces the best classifier so far considering the quality, as shown in Table 9. However, a 0.3% improvement of the recall does not justify a time of approximately 132 min. In all three test procedures, the standard implementation with simple TF, stemming and without NER processing proves to be the most useful for the use in a productive system. The classification quality is not substantially worse. For example, the accuracy does not differ by more than 0.15% in all methods. Thus, an additional expenditure in time is not compensated.

4.2.2.5. Feature set phase. The various feature set combinations are created based on the results of the text processing phase. Thus, if word characteristics occur in the combination of the set of characteristics, character removal, tokenization, stemming, and TF are used.

Table 10 shows a section of the evaluation with various classification feature combinations specified in the first column. The highlighted values represent the best result of the respective column. As in the previous test procedures, the classification quality is shown with the results of accuracy, precision and recall; the time in seconds represents the effectiveness and efficiency in a productive system. As can be seen, all classification features from Table 10 have been combined with the classification feature words at least one time, except for NER and language (for reasons mentioned above). Further combinations have been listed as an example or if they could improve the algorithm. Slight fluctuations in time can be explained by internal factors of the computer, e.g., a more efficient use of the cache.

Adding classification features to the existing textual classification features slightly improves the quality of the algorithm in every case. Adding classification features further, however, can have a negative impact on the classification quality, as can be seen in rows 3 and 4 with the example of the number of retweets. This is probably due to an overfitting resulting from too much information being available to the algorithm. In this dataset, the Random Forest has reached the best classification quality when the classification features words, geographical distance, temporal distance and length are combined. The accuracy of 91.123% exceeds the accuracy of the purely textual classifier by 0.323%. This means that by adding these features, 17 tweets previously misclassified could now be classified correctly. However, an additional time of about four minutes is required.

The best performance in time can be achieved without textual classification features (see Table 10, last row). Although this procedure results in the worst classification quality with an accuracy of 84.35%, it confirms the assumption that the non-textual classification features of Table 10 point to relevance without the underlying tweet's text. Nonetheless, the result is still more than acceptable as algorithms of related work achieved even worse results.

The following phases are based on the Random Forest with the classification features words, geographical distance, temporal distance and length as this combination has proven to achieve the best classification quality.

4.2.2.6. Parameter optimization phase. The first runs of the standard implementation of the Random Forest in Weka show discrepancies in the recall and precision value (Table 11). This imbalance could be inherent in the dataset but is improved by

Table 8
Classification quality and time of the word stem creations stemming and lemmatization on the basis of a textual Random Forest.

	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
Stemming	90.8	91.3	81.1	**850.271**
Lemmatization	**90.87**	**91.4**	**81.2**	1936.385

Table 9
Classification quality and time with and without the NER classification feature based on a textual Random Forest.

	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
Without NER	90.8	91.3	81.1	**850.271**
With NER	**90.95**	**91.4**	**81.5**	7952.865

Table 10
Classifications with different feature set combinations. A Random Forest with the respective classification features is used.

Classification Features Used	Accuracy	Precision	Recall	Time (s)
Words	90.8	91.3	81.1	850.271
Words + Number of Retweets	90.82	91.4	81.1	851.14
Words + Length	90.89	91.4	81.3	862.69
Words + Number or Retweets + Length	90.85	91.4	81.1	841.78
Words + Temporal Distance	90.93	91.4	81.3	901.22
Words + Geographical Distance (Author Distance and Tweet Distance)	91.03	91.6	81.3	1021.663
Words + Geographical Distance (Author Distance and Tweet Distance) + Temporal Distance	91.21	**91.8**	81.4	1078.092
Words + Distance (Author Distance and Tweet Distance) + Temporal Distance + Length	**91.23**	**91.8**	**81.5**	1110.276
Words + URLs	90.9	91.4	81.4	850.22
Words + Media	91	91.5	**81.5**	860.12
All Classification Features	91	91.6	81.1	1071.79
No Words + All Other Classification Features	84.35	84.4	75.1	**281.14**

Table 11
Modification of the threshold value (threshold-moving) of the Random Forest with the classification features words, geographical distance, temporal distance and length.

	Accuracy [%]	Precision [%]	Recall [%]
Threshold: 0.5	91.23	**96.9**	81.5
Threshold: 0.3	**91.64**	94	**85.2**

changing the threshold from which a result is classified as relevant or irrelevant (Zhou & Liu, 2006). We used a threshold selector meta classifier in combination with the RF classifier in weka. Other default settings, such as the number or depth of decision trees, were tested within a random search. The default values proved to be good for the classification problem at hand.

In the standard implementation of a Random Forest in Weka, the threshold value is set to 0.5. As can be seen in Table 11, a threshold value of 0.3 has proven to be useful, as recall and precision converge and accuracy increases to 91.64%. While the recall value increases as well, the precision value decreases. According to Habdank et al. (2017), this change is to be assessed as positive. Even though a decreased precision value means that the emergency services are shown more irrelevant tweets, which may result in an additional time consumption, a decreased recall value implies that more relevant tweets are incorrectly classified as irrelevant, which could have severe consequences in crises.

4.2.2.7. Algorithm phase. Since the combination of the classification features words, geographical distance, temporal distance and length have provided the best results, it is used in the following comparison of the Naïve Bayes classifiers and the Random Forest. As in the last phase, the Naïve Bayes classifier was tested for various parameters. A threshold-moving is not necessary because the recall and precision value are very balanced. As in the Random Forest classification, other parameters prove to be useful in the default setting. For the Random Forest classifier, the threshold of 0.3 was transferred from the last phase.

Table 12 illustrates that the Random Forest performs better than the Bayes classifier in every aspect of the classification quality. The important recall value of the Random Forest is at 85.2% and the recall value of the Naïve Bayes implementation is at 82.7%. Thus, the Naïve Bayes approach incorrectly classifies 282 tweets as irrelevant, whereas the Random Forest approach produces only 240 FN. However, the pre-processing, training and 10-fold cross-validation of the Naïve Bayes classification with about 4 min is much faster than the Random Forest with 18 min.

For completeness and to make an all-encompassing statement, both algorithms have to be tested on all classification feature

Table 12
Classification quality and time of the Naïve Bayes and Random Forest classifier based on the classification features words, geographical distance, temporal distance and length.

	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
Random Forest	**91.64**	**94**	**85.2**	1120.22
Naïve Bayes	85.19	81.8	85.2	**263.749**

Table 13
Classification quality and time with and without a feature set elimination procedure based on the classification features words, geographical distance, temporal distance and length.

	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
10,153 Features	**91.64**	94	**85.2**	1120.22
148 Features	91.28	**98.2**	80.4	**204.326**

combinations and all pre-processing steps. As the results are very clear and in accordance with those of Habdank et al. (2017), this dataset assumes that the Naïve Bayes classifier is to be used for faster classification and the Random Forest for a higher classification quality of relevance.

4.2.2.8. Feature subset selection phase. The Bag-of-Words (BoW) of this dataset comprises 10,149 words, with all of them included as classification features in the classification. In the following, a filter approach is applied to the Random Forest algorithm in an attempt to reduce the time required. The filter procedure is applied directly to the dataset independently of the classifier using information gain as evaluator with a threshold of 0.005.

In this way, the feature set was reduced substantially. Without the filter approach, 10,149 words in combination with geographical distance, length and temporal distance were previously used for classification. Using the feature set search, the number can be reduced to 148. These 148 classification features are already decisive for the classification of relevance. Moreover, the resulting classifier still achieves a high classification quality compared to the best algorithm of the previous phase and, as expected, the required time is lower (Table 13). With about 3 min and 15 s this algorithm is also faster than the Naïve Bayes approach of the algorithm phase.

When analyzing these 148 classification features, it can be seen that the algorithm selected all four features gained from the feature set phase for the relevance classification. Furthermore, the set contains words that provide direct conclusions about the term relevance as defined in Section 2.3. This includes words such as “water level” (German: “Pegel”), “alarm level” (German: “Alarmstufe”) and “cm” (short for “centimeter”) which point to information and facts of the tweets. The terms “help” (German: “Hilfe”) and “helpers” (German: “Helfer”) speak in favor of the term relevance as defined by Vieweg (2012). Terms such as “donation” (German: “Spende”) are in the end probably weighted negatively by the algorithm as they indicate irrelevance.

4.2.3. Results of the 2016 BASF SE incident dataset

For a second evaluation of the before stated phases, we use the BASF incident dataset. The best combination of features is given by taking the tweet length and the temporal distance into account. With these features the random forest classifier reached a very good accuracy of 90.3%. The shortened results can be seen in Table 14. Adding the distance features, as in the flooding scenario, did not improve the result. We expect that this could be due to the incomplete dataset. Only about 50% of the tweets have an author location specification, whereas in the flooding dataset we were able to calculate the author distance of about 85%. This sparsity implies that the entropy of these features is naturally higher, causing the decision trees in the end to use other features for building the trees.

Threshold-moving did not change the result in a noticeable way, since recall and precision are already good balanced.

In the feature reduction phase, we were able to successfully reduce the size of the features without diminishing much of the quality of the classifier. The filter approach in combination with a threshold of 0.005 reduced the initial 9164 features to remaining 173 also containing temporal difference and length. When inspecting these values, it can be seen that information gain ranked the time difference as the highest feature. Furthermore, there are several words that are expressive for relevance in this fire and explosion situation. “Explosion” (German: “Explosion”), “toxic alarm” (German: “gift-alarm”) and “safety note” (German: “Sicherheitshinweis”) are words that are probably involved in tweets containing facts and advisory information.

4.2.4. Summary

In all phases of the evaluation, the classification quality for this dataset was continuously improved and the required time reduced. Each of the steps covered with a different sub-area. Hence, the resulting findings are different. The *text processing phase* showed that the TF vectorization, the stemming and the exclusion of the NER classification feature are preferable to the complex procedure TF.IDF, lemmatization and the use of the NER classification feature because they offer higher efficiency. The additional required time does not justify the small increase in classification quality. In the *feature set phase*, the use of the classification features word, geographical distance, temporal distance and length proved to be the best combination regarding classification quality in the flooding scenario. For both datasets, the exclusion of textual classification features and the use of all other classification features

Table 14
Classification quality and time of two different additional features sets and a feature reduction on the BASF incident dataset.

Classification features used	Accuracy [%]	Precision [%]	Recall [%]	Time [s]
Words	90.08	**90.1**	89.2	1109.04
Words + Temporal Distance + Length	**90.32**	89.8	**90.0**	1224.10
Words + Temporal Distance + Length [Reduced]	89.13	89.6	87.5	**143.00**

results in an acceptable classification quality and a considerably better (lower) time consumption.

In the case of the flooding scenario, the *parameter optimization phase* showed that the use of the Random Forest with a threshold value of 0.3 instead of 0.5 improves the classification quality by increasing accuracy and recall. According to the results of the *algorithm phase*, Random Forests achieve a better classification quality than Naïve Bayes classifiers. However, Naïve Bayes classifiers are preferable when aiming for a lower time consumption. In the *feature subset selection phase*, the time consumption is considerably reduced when using only the most important classification features. The classification quality decreases minimally. Examining the remaining classification features, it can be seen that these reflect the terminology of relevance as defined in this paper.

5. Evaluation II: relevance classification via active and online learning

Finally, for insights into **RQ3 ("How can the amount of labeled data required for relevance classification be reduced by active incremental learning and transparent visualization of the classifier's quality?")** and **RQ4 ("How can the dynamic retraining of relevance classifiers be supported by user feedback performance-wise using batch learning with feature subset selection?")**, we conducted our second evaluation whose approach (Section 5.1) and conduction (Section 5.2) are described in the following.

5.1. Approach

As already indicated by a variety of publications, active learning units can substantially reduce the amount of labeled data required to reach a certain accuracy threshold in different application domains of machine learning (Bernard, Zeppelzauer, Lehmann, Müller, & Sedlmair, 2018; Imran et al., 2017; Settles, 2010). Since time is limited in disaster scenarios, active learning could be of use for creating suitable classifiers. For instance, Imran et al. (2017) demonstrate an active learning unit which supports the labeling of events by suggestions based on past events. Motivated by this potential, this section conceptualizes and evaluates an active learning unit which directly works with the dataset of an event, requiring no previously labeled data or events.

The active learning is realized via *uncertainty sampling*. According to Lewis and Catlett (2014), it is especially reasonable to label the posts where the classifier has the lowest confidence. The learning unit occurs in a *pool-based sampling* scenario using the *least confidence* measure. Settles (Settles, 2010) describes pool-based sampling as an environment in which the algorithm checks the set or subset of the non-labeled data with regard to its information content and returns the most reasonable datum. By applying the least confidence measure to a binary classification problem, the instance is returned where the classifier's prediction confidence is nearest to 50% (Settles, 2010).

For the presented approach it is obvious that classifiers, which are able to compute prediction probabilities, are already required at the labeling stage. For the execution of the active learning unit it is important that these classifiers can be rapidly adapted by newly labeled posts to improve their quality and to allow an application in real-time. Lewis and Catlett note that "uncertainty sampling requires the construction of large numbers (perhaps thousands) of classifiers which are applied to very large numbers of example" (Lewis & Catlett, 2014).

The aim is to develop an online learning environment where sequentially labeled data incrementally enhance the classifier. The RF classifier of Section 4 is not suitable for this task since it is a batch learner which requires a completely new training run for each additional datum. Incremental learning, then again, allows the sequential integration of new posts without looking at the previously labeled data again. To clarify this aspect, we compared the RF batch classifier with different incremental methods regarding retraining time. Weka supports multiple incremental learning approaches and Hoeffding Tree (HT) (Ren, Lian, & Zou, 2014), Incremental Naïve Bayes (iNB) (Hulten, Spencer, & Domingos, 2001) and k-Nearest Neighbor (IBk) (Aha, Kibler, & Albert, 1991) are candidates for our scenario. In case of IBk, we additional use a k-D-tree structure to improve the search for the $k = 50$ neighbors (Moore, 1991). We use the dataset of Section 4.2.1 for comparison, dividing it into 3902 posts for initial training and 20 posts for retraining. As seen in Table 15, iNB and HT require less than a second for retraining. IBk with $k = 50$ requires three seconds for the task and still seems suitable for the use in SMO. However, the RF classifier requires more than two minutes for the same task.

In addition, we have found that incremental methods for our purposes produce worse classification grades than RF (Section 5.2). However, the maximization of classification accuracy is not required during the labeling process using an active learning unit as, thereafter, a different algorithm can be used for the creation of the classifier based on the actively created dataset. As Lewis and Catlett (Lewis & Catlett, 2014) outline, for uncertainty sampling it is suitable to use simple or 'cheap' classifiers for the active selection of data to be labeled and to create complex or 'expensive' classifiers based on this. Despite the heterogeneity of the type of simple or complex classifiers, a reduction of the required labeled data is possible.

In the labeling process of SMO, we use IBk classifiers with k-D-trees that are extended incrementally with each new labeled post. The process is visualized in Fig. 9. After each third to fifth labeling, the classifier is used to compute an active learning request (marked with a *-sign and in yellow in comparison to the black non-active labeled posts). In this case, the classifier predicts the

Table 15
Time for the retraining of the classifier with Random Forest Naïve Bayes, Hoeffding Tree, and k-Nearest Neighbor (IBk).

	RF	NB	HT	IBk
Time [s]	129.294	0.017	0.19	3

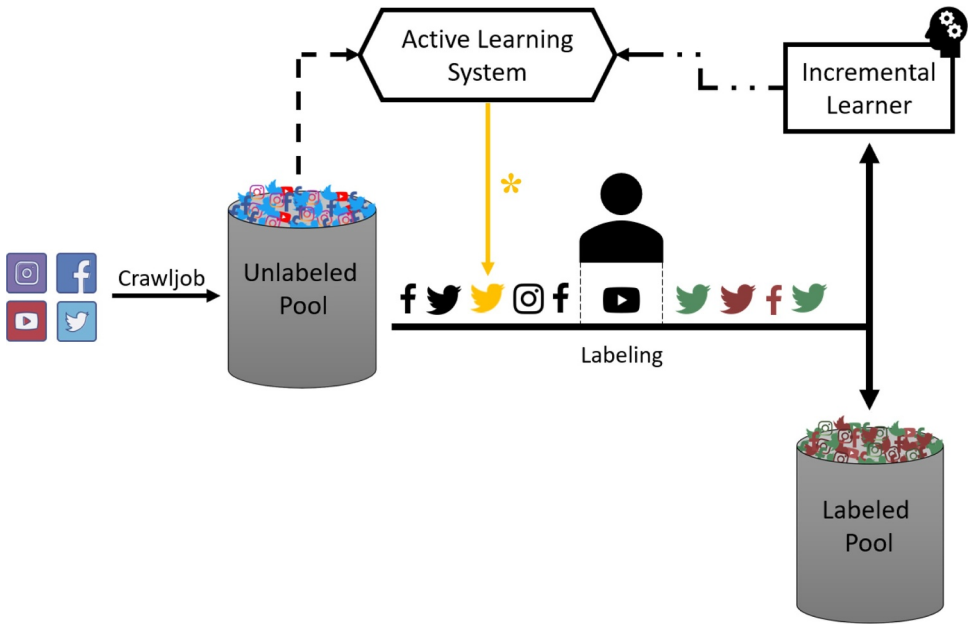

Fig. 9. Incremental active learning in the labeling process of SMO.

probability with regard to the relevant class of 50 to 100 random unlabeled posts (unlabeled pool) und returns the post that is most ambiguous, i.e., nearest to the 0.5 value (50%). By only using each fifth post for the active learning unit, bigger amounts of the pool can be included and excessive influence of the first increments can be prevented. The indicated values are well-suited for real-time use but are dependent on the performance of the underlying machine. After the labeling process, RF classifiers are trained to achieve the best classification accuracy based on the labeled posts (labeled pool).

5.1.1. Real-time evaluation using active and incremental learning

In the labeling process, labelers and researchers ask themselves the question of how much data needs to be labeled to achieve a classifier with an acceptable quality. To get one step closer to a potential answer, we propose real-time evaluation during the labeling that displays an estimated quality, in terms of accuracy, precision, and recall, for the resulting classifier.

For this, we can utilize the fact that the active learning unit from Section 5.1 already conducts online learning during the labeling process. Due to incremental learning, a new classifier is available after each labeled post. Although the classification quality is lower than in the case of using a RF classifier, an approximation regarding the final classifier is sufficient in this case since the main aim is to convey an idea of the classifier's learning curve based on the labeled data. However, it is still a problem to guarantee its real-time application. A cross-validation is not recommended since it requires retraining the classifier and does not utilize the incremental characteristics. Thus, we propose to skip every fourth or fifth labeled post for the training of the classifier in order to move it into a test set. Thus, roughly 75% (80%) of the posts constitute the training set and 25% (20%) constitute the test set. This strategy reflects the holdout method, which is frequently used as an evaluation method besides cross validation. The extension to the labeling process is marked with a *-sign and depicted in blue in Fig. 10. The integration of real-time evaluation is described in Fig. 6 of Section 3.4.2.

5.1.2. Feedback classification using feature subset selection

Active learning indicates the potential of labeling posts where the classifier is unconfident. Another (even better) option would be to select and correct the posts which were misclassified by the classifier. This cannot be achieved proactively, but a reactive method, comparable to spam detection, would be necessary to reduce false positive and false negatives after final classifier creation. Based on the system of a European research project, Markham and Muddiman describe feedback classification as a potential extension, since continuous feedback “leads to a system that adapts and improves over time” [75, p. 3]. The correction of false predictions has a considerable impact on the classifier's quality and allows for the adaptation of the classifier along the event's process.

For the practical implementation, again, the issue emerged that batch RF classifiers, which performed better than incremental classifiers, require too much time for retraining since they are not extensible with new data. We address this issue by the use of supervised Feature Subset Selection (FSS), as most of the thousands of features are not decisive for the actual classification and can be eliminated (see Section 4.2.2 for the feature subsection selection phase of the first evaluation). As a consequence, the training is performed much faster and less overfitting occurs during learning. As a side note, this approach is not suitable for real-time evaluation and active learning (Section 5.1.1), because supervised FSS requires a considerable amount of labeled data. However, for this problem we apply a filter approach. By using a search algorithm the best subset of features is searched (John et al., 1994). To reduce the time of this approach, we perform a bottom-up-search with the upper bound of 200 features. The time evaluation revealed that the RF classifier of Section 4.2.2 requires less than five seconds for retraining with 20 posts after application of FSS (Table 16). This is

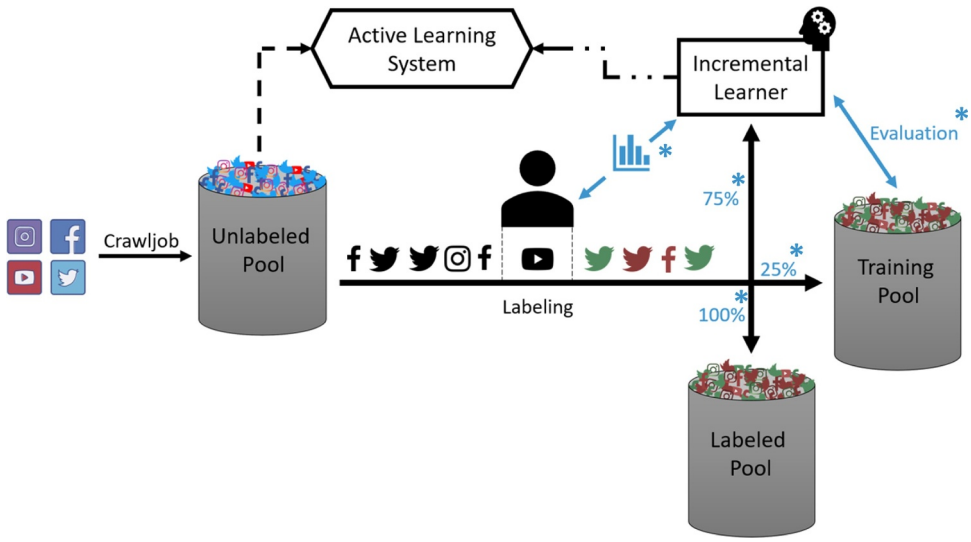

Fig. 10. Real-time evaluation during the labeling process in SMO.

Table 16
Comparison of time for the recreation of the RF classifier with and without Feature Subset Selection (FSS) using the 2013 European floods dataset.

	RF without FSS	RF with FSS
Time [s]	129.294	4.704

still too slow for the labeling process but suitable for the correction of false classifications. The feedback classification process is marked with a *-sign and depicted in green in Fig. 11. The integration of feedback evaluation is described in Fig. 8 of Section 3.4.2.

To summarize the overall approach, Fig. 11 also comprises the labeling process with the active learning and real-time evaluation methods. The active learning unit allows to use a small pool of labeled data to create a classifier of good quality. The real-time evaluation based on online learning supports the labeler to get an idea at which time a sufficient amount of posts was labeled. Subsequently, a RF classifier using FSS can be trained. Furthermore, the false predictions of this classifier can be corrected by the user, e.g., the emergency manager, reactively; based on the feedback, a new classifier is trained in the background. All units are designed for multithreading and are thread safe so that delays constitute no problems, allowing all units to work consistent and as intended.

5.2. Preliminary evaluation

5.2.1. Method

In this section, we simulate a real application of the active learning unit in combination with the real-time evaluation component using online learning. We use the collected data from the 2013 European floods and the BASF SE incident in 2016. We assume a scenario where we have no labeled data available but intend to build a relevance classifier. The first phase of the evaluation describes the comparison of three incremental classification methods on the flooding scenario. The second phase compares the learning success with and without active learning on both datasets. In summary, several datasets, each comprising 1000 posts, were labeled within the SMO. The evaluations are performed on a computer with Windows 10, Intel i7-7500 with two physical and two logical cores at 2.90 GHz and 12GB Ram in Java.

5.2.2. Results

The first phase reveals which incremental classifier achieves good results, so that the SMO can provide accurate estimations on the classifier's quality and suggest useful posts for active learning. For that purpose, the SMO labeling process was conducted three times for Incremental Naïve Bayes (iNB[2]), Hoeffding Trees (HT[3]) and a *k*-Nearest Neighbor algorithm (IBk[4]) based on the 2013 European floods dataset and using the active learner. The area under the ROC curve (AUC) of the real-time evaluation was noted down every 50

[2] Using standard parameters from Weka.
[3] Using standard parameters from Weka.
[4] As indicated earlier, we used a *k*-D-tree as representation model to facilitate a fast classification. Furthermore, 50 neighbours were selected as parameter for *k*.

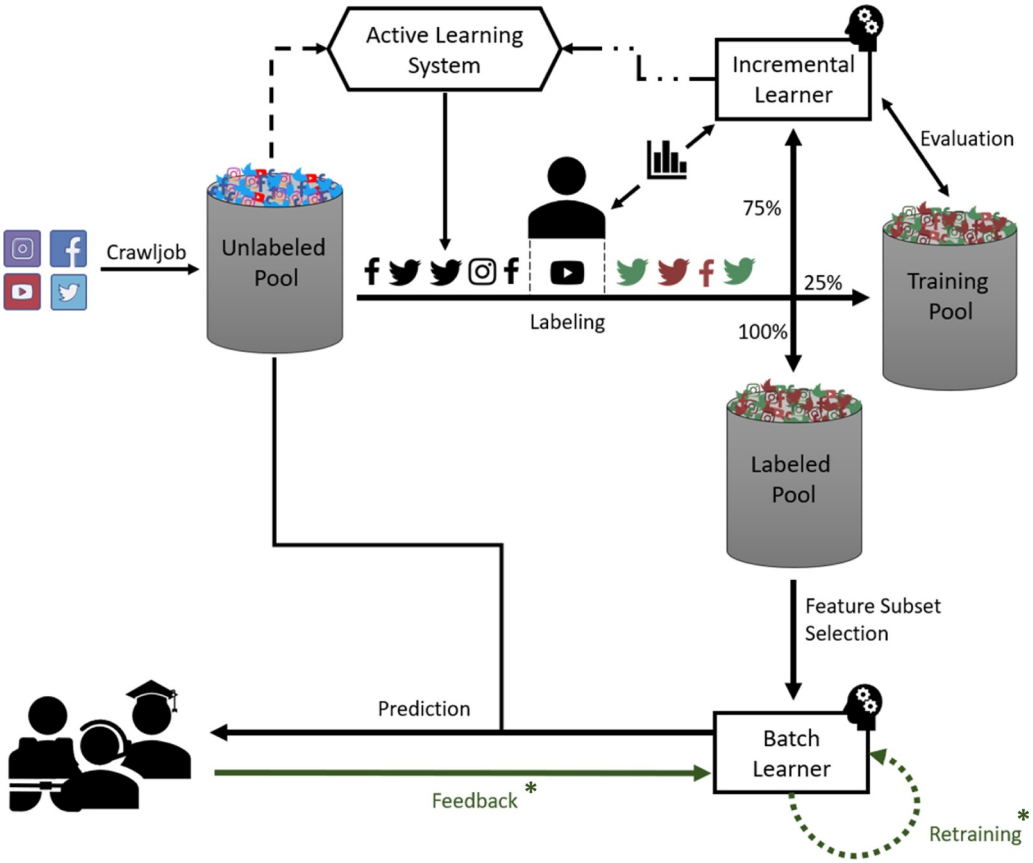

Fig. 11. The classification process comprising active learning, realtime evaluation and feedback classification.

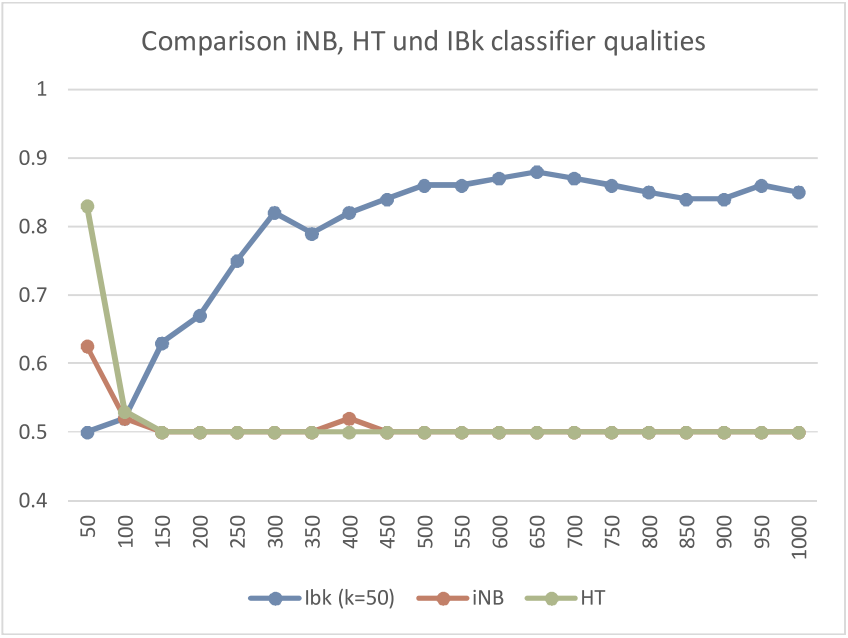

Fig. 12. Comparison of the incremental learning methods iNB, HT, and IBk using AUC values from the real-time evaluation in SMO.

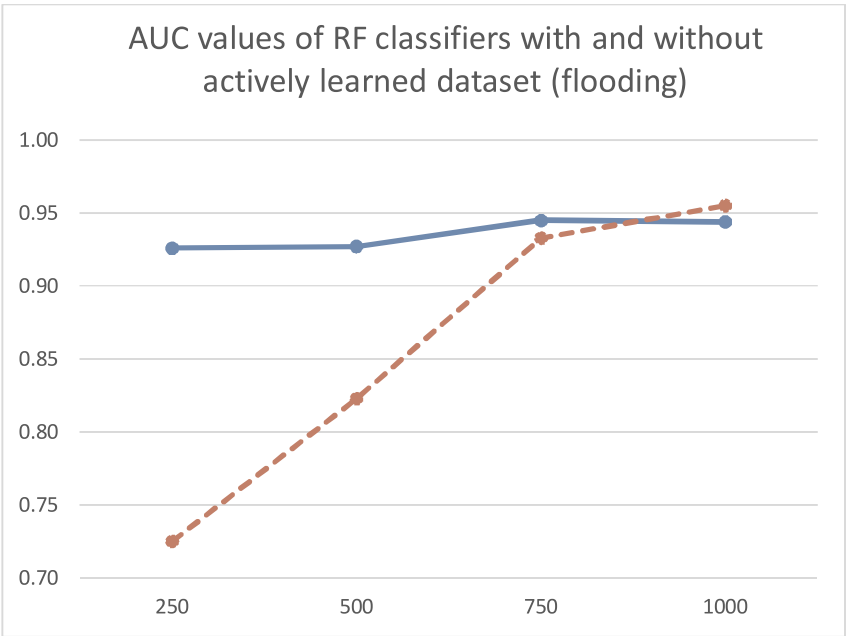

Fig. 13. Comparison of AUC values of RF classifiers on the flooding dataset with (solid) or without (dashed) active learning.

labels. The results are displayed in Fig. 12. The evaluation revealed that iNB and HT are not suitable for the present classification problem. After 200 labeled posts, both algorithms score consistently low AUC values around 0.5, thus showing a random classification behavior. Besides low suitability for real-time evaluation, these algorithms are not able to suggest useful posts for active learning. In contrast, IBk is suitable for the present classification problem, showing a good learning curve. After labeling 1000 posts, an accuracy of 84.5% was achieved, which missed the accuracy of the resulting RF at that point (89.3%) by only about 5%.

In the second phase the dataset labeled with IBk and active learning is compared to a dataset that was labeled without active and incremental learning. On the flooding dataset, after each 250 labeled posts, a RF classifier was trained and evaluated with 10-fold cross validation. The AUC values are displayed in Fig. 13. The evaluation revealed that the RF which was supported by active learning

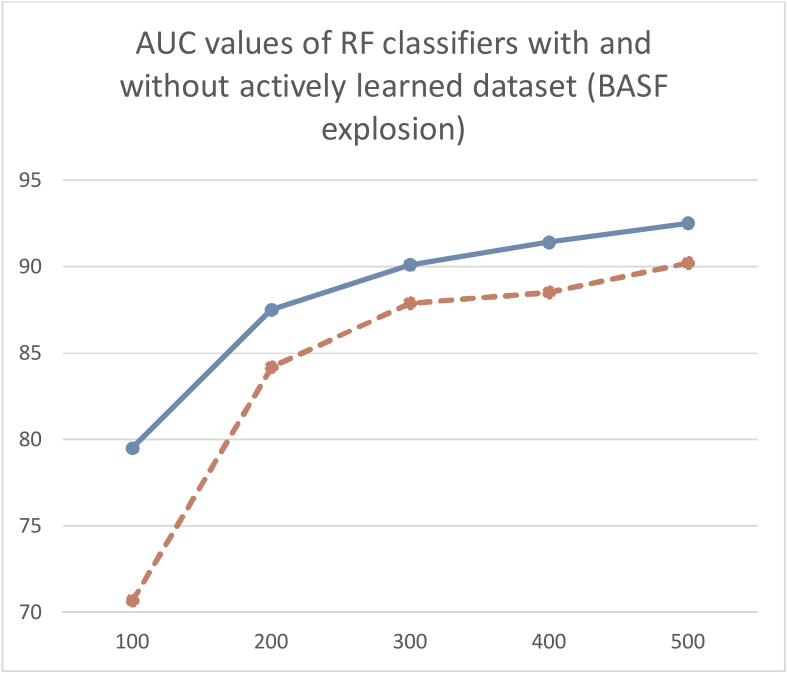

Fig. 14. Comparison of AUC values of RF classifiers on the BASF SE dataset with (solid) or without (dashed) active learning.

reaches excellent AUC values earlier. Already after 250 labeled posts, the AUC is 0.2 better than the RF without active learning. This implies the usefulness of active learning for labeling posts in disaster scenarios, reducing the amount of required labeled data.

Within the BASF dataset we trained a RF classifier every 100 labeled posts. Even if the values are not as striking as before, in this explosion scenario, the active learning approach also seems to be useful. The AUC values can be seen in Fig. 14. During the evaluation we noticed two particularities, which could also be the reason for the differing deviations of the two datasets. The "normal" classification already leads to good results after 200 posts, meaning that even without active learning it is fast in finding a good separation of the two classes. We also noticed that the IBk classifier performs slightly worse on this dataset, which could result in an inaccurately behavior when choosing the active posts.

6. Discussion and conclusion

Research in the field of crisis informatics revealed that social media enable emergency services to receive valuable information (e.g., eyewitness reports, pictures, or videos) contributing to situational awareness during disasters and emergencies (Imran et al., 2015; Reuter & Kaufhold, 2018). However, the vast amount of data generated during large-scale incidents can lead to the issue of information overload (Plotnick & Hiltz, 2016). Research indicates that supervised machine learning techniques are suitable for identifying relevant messages and filter out irrelevant messages, thus mitigating information overload (Habdank et al., 2017). However, in order to train an accurate classifier, clear criteria for relevance labeling and classification, a considerable amount of labeled data and a usable interface to facilitate the labeling process are necessary. In the following, we outline our results with regard to our research questions (Section 6.1) as well as limitations and future work (Section 6.2).

6.1. Results

In order to address the above-mentioned issues, we presented (1) a system for social media monitoring, analysis and relevance classification, (2) criteria for relevance labeling and prediction in social media during disasters and emergencies, (3) the evaluation of a well-performing algorithm for relevance classification using batch learning as well as (4) an approach and preliminary evaluation for relevance classification including active and online learning to (4.1) reduce the amount of required labeled data by real-time evaluation and (4.2) correct misclassifications of the algorithm by feedback classification. Our results contribute to answering the following research questions.

Which are suitable criteria for relevance classification and labeling in disasters and emergencies (RQ1)? Our literature review revealed that there is already a multitude of approaches dealing with relevance classification problems, which are summarized in Table 3. However, we found out that they often lack a clear definition of relevance as well as a set of criteria to determine relevance. Thus, we identified on the one hand abstract and interpretative (Table 1) and on the other hand clear and precise (Table 2) relevance criteria. While both types of relevance criteria are useful for labeling only the latter ones are suitable for relevance classification using supervised machine learning. The determined criteria also served as an input for the design and evaluation of a batch learning RF classifier (Section 4).

How can existing supervised machine learning techniques for relevance classification be improved for use in real disaster and emergency environments (RQ2)? In this paper, we evaluated supervised machine learning classifiers which do not only take textual features but also relevance criteria into account. In our evaluation based on a flooding dataset, the combination of textual content (words), geographical distance (author and tweet), temporal distance and length reached the best classification quality (Table 10). With almost 92% accuracy, 94% precision and 85% recall the Random Forest classifier reached an excellent quality after the parameter optimization phase. Also, the classifier based on the BASF incident dataset research better results than in Habdank et al. (2017) with almost 90% accuracy, 90% precision and 90% recall. Thus, the classifiers reached a better accuracy than those in related work because the insights of these studies were incorporated into this one. For instance, Imran et al. (2013) and Spielhofer et al. (2016) used Naïve Bayes classifiers, which yielded worse results for the present classification problem. Although Habdank et al. (2017) also used Random Forests, they are only based on textual features. Then again, Spielhofer et al. (2016) and Habdank et al. (2017) reached better recall values. However, the recall value could be improved by the further adaptation of threshold values during the parameter optimization phase.

Furthermore, many machine learning studies only focus on the evaluation itself, thus neglecting the importance of the efficiency (training time) of a classifier. However, in a productive system, the balance between effort and quality is important. Accordingly, the evaluation of the classifier also considered time expose for training (Section 4.2.2). The previously outlined RF classifier required acceptable 18 min of time consumption for pre-processing, learning and 10-fold cross validation. However, if a faster classification is required, the feature subset selection phase outlined in the case of a flooding dataset, that a classifier using 148 features (instead of 10,153) achieved 91.3% accuracy, 98.3% precision and 80.4% recall with a time expose of only 3 min, also reducing the memory consumption of the classifier.

How can the amount of labeled data required for relevance classification be reduced by active incremental learning and transparent visualization of the classifier's quality (RQ3)? We proposed methods to enhance research regarding classifying relevant messages in social media. At its core, we looked at the dilemma that, on the one hand, classifiers require a lot of training data for the present classification problem to achieve a considerable quality but, on the other hand, at the beginning of a disaster, often no labeled data is available and data has to be labeled in a limited amount of time. To address this issue, multiple researchers describe and evaluate approaches based on *domain adaptation* that use datasets from past events to reduce the amount of required labeled data from the current event (Imran et al., 2016, 2017; Li et al., 2017, 2015). In this work, however, we discuss approaches for improving

the labeling process without incorporating data from past events.

To reduce the amount of required labeled data we implemented an active learning unit. This required rapidly trained classifiers that adapt with each new post during the labeling process. Thus, we realized it in an online learning environment which was realized via incremental classifiers (Section 5.1). However, without further support, the labeling process is not transparent, i.e., the labeler is not able to assess the quality of the classifier during the labeling process and does not know when to finish the labeling process. Thus, we implemented an approach for predicting the classifier's quality in terms of accuracy, precision and recall, on frontend (Section 3.4.2) and backend (Section 5.1.1) level. Based on the merits of incremental learning, we proposed an approach for real-time evaluation.

The evaluation revealed that not all incremental learning algorithms are suitable for the present classification problem (Section 5.1.2). However, IBk classifiers reached an acceptable quality for real-time predictions and constitute a successful learning unit for active learning. It was furthermore shown that the active learning unit was able to substantially reduce the number of labeled posts required for a well-performing classifier.

How can the dynamic retraining of relevance classifiers be supported by user feedback performance-wise using batch learning with feature subset selection (RQ4)? Even if the classifier achieves an excellent quality, misclassifications still occur. End-users, such as emergency services, might be interested in correcting these misclassifications while performing their analytical tasks to improve the quality of the classifier based on their feedback (Markham & Muddiman, 2016). Thus, we implemented a feedback classification mechanism on frontend (Section 3.4.2) and backend (Section 5.1.2) level. However, this kind of a posteriori classification is subject to the slow classification time of offline or batch learning algorithms, which are required to maximize the classifier's quality. In order to allow a rapid adaptation, we implemented a feature subset selection approach.

6.2. Practical implications

As a practical contribution, we created a ready-to-use system for potential end-users, such as emergency mangers, and as a foundation for conducting further machine learning evaluations. Based on this, we derived the two following major practical implications.

Incorporate metadata of social media to improve the quality of relevance classification (P1). From a practical point of view, our results indicate that the consideration of metadata, as identified and summarized as precise relevance criteria (see RQ1), yield in better relevance classification results in comparison to classifiers that only consider textual features (Habdank et al., 2017; Markham & Muddiman, 2016) (see RQ2). However, the best classification results were achieved with a different set of metadata between our two datasets. Thus, characteristics of datasets, such as the number of specified locations, should be considered for determining the best classifier that still performs well also with regard to training time. To assist this process, future research could determine social media guidelines (Kaufhold, Gizikis, Reuter, Habdank, & Grinko, 2019) for the interpretation of dataset metadata and its implications for the selection of classifier features. Furthermore, since professional roles, such as incident or public relation managers, may have different conceptualizations of relevant information, different classifiers using distinct metadata may be deployed to improve the actionability of information (Zade et al., 2018).

Use active and online learning to reduce classifier training time during disasters and emergencies (P2). The promising results of our active learning unit highlight that active learning can significantly reduce the training time of a well-performing classifier (see RQ3). Its combination with an approximate real-time evaluation of the classifier's quality during the labeling process assists the user in determining the stage where the classifier's quality is good enough for a first deployment during an emergency. After initial deployment, the labeling of data could be continued to deploy improved revisions of the classifier over time, also considering changes in language during developing events, such as emerging hashtags or words during collective sense-making processes (Stieglitz, Mirbabaie, & Milde, 2018) that correlate with relevant or irrelevant content. However, the overall process must be designed time as efficient as possible due to scarcity of time of emergency managers (Reuter et al., 2016), which suggests combining domain adaptation approaches to combine labeled data from past similar events with new data of the current event (Imran et al., 2018).

Further practical implications, especially with regard to the deployment of the architecture in real-world environments, have to be derived by the evaluation of the system also including the dynamic retraining of relevance classifiers (RQ4), with expert users from emergency services, which is discussed in Section 6.4.

6.3. Theoretical contributions

Our paper contributes to the area of textual content analysis, applied to the domains of emergency management and social media, and which comprises the areas of relevance classification, information quality assessment, sentiment analysis, clustering and summarization, humanitarian classification, topic modeling and named entity recognition, amongst others (Alam, Ofli, & Imran, 2019; Gründer-Fahrer, Schlaf, Wiedemann, & Heyer, 2018; Imran et al., 2018; Kaufhold, Rupp, Reuter, & Habdank, 2019). More specifically, in the area of textual relevance classification, we made two contributions.

Identification of abstract and precise relevance criteria for labeling and classifier training (T1). In order to improve the foundations of relevance labeling and classification in future research, we identified and compiled a variety of abstract and factual criteria from existing research publications. While abstract and interpretative (summarized in Table 1) criteria help to develop labeling guidelines for datasets, which might vary according to the needs and structures of different personnel, roles or organizations (Hughes et al., 2014; Reuter et al., 2016), our results show that classifiers considering factual and precise relevance criteria

(summarized in Table 2) are able to outperform those who focus on textual features only but not considering metadata (Habdank et al., 2017).

Novel concept for rapid relevance classification of social media posts in disasters and emergencies (T2). On the other hand, as a means of reducing information overload for emergency managers especially in large-scale emergencies, we proposed a novel concept for rapid relevance classification of social media posts, as summarized in Fig. 11. Firstly, it comprises data labeling component with an approximate real-time evaluation of the expected classifier quality via online learning to support the user in quality assessment already in the labeling process. While research starts recognizing the value of white-box approaches to increase the users' trust and understandability of (transparent) algorithmic decisions (Kaufhold et al., 2019), this component contributes by making the approximate quality of the classifier transparent already during the labeling phase.

Secondly, it incorporates active learning to reduce the amount of labeled data required to achieve good quality classifiers in time-constraint settings, such as emergencies. Due to the promising results, active learning should be considered in analytical frameworks and the labeling and evaluation of more datasets is required to identify criteria determining the performance of active learning in the domains of emergency management and social media. Furthermore, a combination of active learning and domain adaptation seems worth researching too allow the use of already labeled data from similar past events and actively learn from new event data (Imran et al., 2018).

Finally, it allows the retrospective retraining of the classifier based misclassified instances. Despite this functionality allows the user to further tune the performance of the classifier, its effect is subject to evaluation in future. Again, we envision that a white-box approach, helping the user to understand why a specific message was classified as relevant or irrelevant, would help to identify patterns to increase the users' competence in training classifiers appropriate and according to their needs, which might result in gathering more actionable information from social media (Zade et al., 2018).

While this concept is a first theoretical contribution, further research from human-computer interaction, machine learning and social media analytics has the potential to distill this concept into a human-centered and white-box analytical framework for rapid relevance labeling and classification.

6.4. Limitations and future work

This paper is subject to limitations and future research potentials, alongside those outlined in the two sections before. Firstly, our evaluations are based on two dataset of the 2013 European floods, which was the biggest disaster event in Germany in terms of social media use (Reuter et al., 2015), and the 2016 BASF SE incident. Future work should examine the approach with further datasets and in different scenarios, such as small-scale emergencies (e.g., crime, house fires or multiple collisions), major events (e.g., violent demonstrations) or large-scale disasters (e.g., hurricanes, terrorist attacks or wildfires). Furthermore, the datasets only contained posts from Twitter. However, our architecture allows the creation and analysis of cross-source datasets, which could be examined in the future. Our approach, combining active and online learning, has further research potentials on both algorithmic and user interface level.

With regard to the algorithmic perspective, further offline and online machine learning algorithms as well as their impact on the classifier's quality and the active unit, considering the time effort of labeling, should be examined. For instance, Wang, Wan, Cheng, and Li (2009) propose approaches for incremental Random Forest algorithms. Furthermore, it was observed that *Stochastic Gradient Descent* classifiers yield similarly good results as do IBk classifiers. It is also plausible that HT and iNB classifiers achieved bad classification results due to an inability to deal with the large number of features. As our evaluation revealed, well-performing classifiers are still possible by only incorporating metadata-based features, omitting the content or words of a post. A feature vector based on metadata of posts could improve the success probability of both incremental approaches. Furthermore, it is possible to use active learning approaches beyond uncertainty sampling, as indicated by the work of Párraga Niebla et al. (Settles, 2010). The works from Caragea et al. (2016) and Nguyen et al. (2016) also show that *Convolutional Neuronal Networks* (CNN) are able to achieve excellent classification results. Currently, the SMO is limited to the processing of the posts' content and metadata but CNNs are also able to assist in image processing to assess the damage caused by a disaster, as demonstrated by Nguyen, Ofli, Imran, and Mitra (2017).

From a user interface perspective, a system such as the Social Media Observatory (SMO) allows to evaluate our algorithmic approach empirically from the perspective of Human-Computer Interaction (HCI) (Wobbrock & Kientz, 2016). The deployment of SMO for emergency services, such as fire services and police, in experimental or real-world settings would allow for the evaluation of several aspects: a) the hedonic and pragmatic quality criteria of the overall interface (Hassenzahl, Burmester, & Koller, 2003), b) the quality of the labeling interface (Fig. 6), also with regard to the displayed relevance criteria for labeling assistance, as well as c) the perceived usefulness of both the proactive real-time evaluation of the classifier's quality during labeling and reactive feedback classification to correct misclassifications (Fig. 8). Although the SMO is able to show data in both static list and visualization views separately, we envision an customizable, interactive, integrated and real-time visual analytics dashboard (Keim et al., 2008; Onorati, Díaz, & Carrion, 2018) to enhance emergency services' situational awareness. Here, the application of relevance classification, including the proposed feedback classification approach, could help to mitigate information overload. Based on this, its integration with other means of reducing information overload, such as alert generation (Adam et al., 2012; Avvenuti et al., 2014; Cameron et al., 2012; Párraga Niebla et al., 2011; Purohit et al., 2018; Reuter et al., 2016), event summarization (Nguyen et al., 2015; Rudra et al., 2015, 2018) and precise search keyword selection (Abel et al., 2012; Johansson et al., 2012, 2012), should be examined.

At its current state, the overall architecture is optimized for a single system since the focus of this study was on algorithmic feasibility. Despite it works performant on our single-node server, in terms of scalability, it would be possible in future work to

implement multiple processing instances for parallelization and database “sharding”, i.e., using a distributed database. On the other hand, data access is limited by social media tokens, e.g. Twitter allows 450 queries per 15 min on their search API, whereof each query contains a maximum of 100 messages (up to 45,000 messages per 15 min effectively) (Reuter & Scholl, 2014). Using multiple tokens for one application must be carefully checked against different social media terms of services.

Acknowledgments

This research was funded within the research group “KontiKat” of the German Federal Ministry of Education and Research (BMBF No. 13N14351), the National Research Center for Applied Cybersecurity “CRISP” by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education (HMWK), and the Collaborative Research Centre “MAKI” (SFB 1053) of the German Research Foundation. We would like to thank Habdank et al. (2017) for providing the labeled data of the 2016 BASF SE incident.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., Tao, K., & Stronkman, R. (2012a). Semantics + filtering + search = twitcident exploring information in social web streams categories and subject descriptors. *Proceedings of the ACM conference on hypertext and social media* (pp. 285–294). .

Abel, F., Hauff, C., & Stronkman, R. (2012b). Twitcident: Fighting fire with information from social web streams. *Proceedings of international conference companion on world wide web* (pp. 5–8). .

Adam, N., Eledath, J., Mehrotra, S., & Venkatasubramanian, N. (2012). Social media alert and response to threats to citizens (SMART-C). *Proceedings of the international conference on collaborative computing: networking, applications and worksharing (Collaboratecom 2012)* (pp. 181–189). .

Agarwal, N., & Yiliyasi, Y. (2010). Information quality challenges in social media. *Proceedings of the international conference on information quality (ICIQ-2010). 2010. Proceedings of the international conference on information quality (ICIQ-2010)* (pp. 234–248).

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning, 6*(1), 37–66.

Alam, F., Ofli, F., & Imran, M. (2019). Descriptive and visual summaries of disaster events using artificial intelligence techniques: Case studies of Hurricanes Harvey, Irma, and Maria. *Behaviour and information technology (BIT),* 1–31.

Albris, K. (2017). The switchboard mechanism: How social media connected citizens during the 2013 floods in Dresden. *Journal of contingencies and crisis management (JCCM), 26*(3), 350–357.

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. *Proceedings of the international conference on information systems for crisis response and management (ISCRAM)* (pp. 354–358). .

Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., & Tesconi, M. (2014). EARS (Earthquake alert and report system): A real time decision support system for earthquake crisis management. *Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1749–1758). .

Bernard, J., Zeppelzauer, M., Lehmann, M., Müller, M., & Sedlmair, M. (2018). Towards user-centered active learning algorithms *37* (3), 121–132.

Borlund, P. (2003). The concept of relevance in information retrieval. *Journal of the american society for information science and technology, 54*(10), 913–925.

bpb. (2013). Hochwasser in Deutschland 2013. *Bundeszentrale für Politische Bildung,* 1–3.

Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency situation awareness from twitter for crisis management. *Proceedings of the international conference companion on world wide web* (pp. 695–698). .

Caragea, C. (2011). Classifying text messages for the Haiti earthquake. *Proceedings of the international conference on information systems for crisis response and management (ISCRAM)* (pp. 1–10). .

Caragea, C., Silvescu, A., & Tapia, A. H. (2016). Identifying informative messages in disasters using convolutional neural networks. *International conference on information systems for crisis response and management (ISCRAM)*.

Cheong, M., & Lee, V. C. S. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter. *Information systems frontiers, 13*(1), 45–59.

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.

Dittus, M., Quattrone, G., & Capra, L. (2017). Mass participation during emergency response: Event-centric crowdsourcing in humanitarian mapping. *Proceedings of the ACM conference on computer-supported cooperative work and social computing (CSCW)* (pp. 1290–1303). .

Eisenberg, M. B. (1988). Measuring relevance judgments. *Information processing and management, 24*(4), 373–389.

Fürnkranz, J. (2018). Introduction to machine learning, TU-Darmstadt data mining und Maschinelles Lernen 2018-2019, Präsentation.

Gorrell, G., & Bontcheva, K. (2016). Classifying Twitter favorites: Like, bookmark, or thanks? *Journal of the association for information science and technology, 67*(1), 17–25.

Gründer-Fahrer, S., Schlaf, A., Wiedemann, G., & Heyer, G. (2018). Topics and topical phases in German social media communication during a disaster. *Natural language engineering, 24*(2), 221–264.

Habdank, M., Rodehutskors, N., & Koch, R. (2017). Relevancy assessment of tweets using supervised learning techniques mining emergency related tweets for automated relevancy classification. *Proceedings of the international conference on information and communication technologies for disaster management (ICT-DM)*.

Hagar, C. (2007). The information needs of farmers and use of ICTs. In B. Nerlich, & M. Doring (Eds.). *From Mayhem to meaning: Assessing the social and cultural impact of the 2001 foot and mouth outbreak in the UK*. Manchester, United Kingdom: Manchester University Press.

Hall, M. A., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The Weka data mining software: An update. *SIGKDD explorations: newsletter of the Special Interest Group (SIG) on knowledge discovery & data mining. 11. SIGKDD explorations: newsletter of the Special Interest Group (SIG) on knowledge discovery & data mining* (pp. 10–18).

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. *Mensch & Computer 2003: Interaktion in Bewegung* (pp. 187–196). .

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. *Elements, 1,* 337–387.

here (2019). Geocoder API here.

Hiltz, S. R., Diaz, P., & Mark, G. (2011, December). Introduction: Social media and collaborative systems for crisis management. *ACM tranactions on computer-human interaction (ToCHI), 18*(4), 1–6.

Hiltz, S. R., Kushma, J., & Plotnick, L. (2014). Use of social media by US public sector emergency managers: Barriers and wish lists. *Proceedings of the international conference on information systems for crisis response and management (ISCRAM)* (pp. 600–609). .

Hiltz, S. R., & Plotnick, L. (2013). Dealing with information overload when using social media for emergency management: Emerging solutions. *Proceedings of the information systems for crisis response and management (ISCRAM)* (pp. 823–827). .

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *Proceedings of the information systems for crisis response and management (ISCRAM). 6*.

Hughes, A. L., & Palen, L. (2014). *Social media in emergency management: Academic perspective.* Washington, DC: Federal Emergency Management Agency.

Hughes, A. L., St. Denis, L. A., Palen, L., & Anderson, K. M. (2014). Online public communications by police & fire services during the 2012 hurricane sandy. *Proceedings of the conference on human factors in computing systems (CHI)* (pp. 1505–1514). .

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 97–106). .
Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). A processing social media messages in mass emergency: A survey. *ACM Computing Surveys, 47*(4).
Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2018). Processing social media messages in mass emergency: Survey summary. *Companion proceedings of the web conference 2018* (pp. 507–511). .
Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013a). Extracting information nuggets from disaster-related messages in social media. *Proceedings of international conference on information systems for crisis response and management (ISCRAM)* (pp. 791–800). .
Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013b). Practical extraction of disaster-relevant information from social media. *Proceedings of the international conference on world wide web companion* (pp. 1021–1024). .
Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013c). Extracting information nuggets from disaster-related messages in social media. *Proceedings of the international conference on information systems for crisis response and management (ISCRAM)*.
Imran, M., Mitra, P., & Castillo, C. (2016a). *Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages.* arXiv preprint.
Imran, M., Mitra, P., & Srivastava, J. (2016b). *Cross-language domain adaptation for classifying crisis-related short messages.* arXiv preprint.
Imran, M., Mitra, P., & Srivastava, J. (2017). Enabling rapid classification of social media communications during crises. *International journal of information systems for crisis response and management, 8*(3), 1–17.
Jensen, G. E. (2012). *Key criteria for information quality in the use of online social media for emergency management in New Zealand*.
Johansson, F., Brynielsson, J., & Quijano, M. N. (2012). Estimating citizen alertness in crises using social media monitoring and analysis. *Proceedings of the european intelligence and security informatics conference* (pp. 189–196). .
John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Machine learning proceedings* (pp. 121–129). .
Kaplan, A. M., & Haenlein, M. (2010, January). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68.
Kaufhold, M.-A., Gizikis, A., Reuter, C., Habdank, M., & Grinko, M. (2019a). Avoiding chaotic use of social media before, during, and after emergencies: Design and evaluation of citizens' guidelines. *Journal of contingencies and crisis management (JCCM), 27*(3), 197–279.
Kaufhold, M.-A., & Reuter, C. (2016). The self-organization of digital volunteers across social media: The case of the 2013 European floods in Germany. *Journal of homeland security and emergency management, 13*(1), 137–166.
Kaufhold, M.-A., Rupp, N., Reuter, C., & Habdank, M. (2019b). Mitigating information overload in social media during conflicts and crises: Design and evaluation of a cross-platform alerting system. *Behaviour and information technology (BIT),* 1–24.
Keim, D., Andrienko, G., Fekete, J., Carsten, G., & Melan, G. (2008). Visual analytics: Definition, process and challenges. *Information visualization - human-Centered issues and perspectives,* 154–175.
Khouzam, B. (2009). Incremental decision trees.
Kim, M., Sharman, R., Cook-Cottone, C. P., Rao, H. R., & Upadhyaya, S. J. (2012, December). Assessing roles of people, technology and structure in emergency management systems: A public sector perspective. *Behaviour and information technology, 31*(12), 1147–1160.
Kulessa, M. (2015). Online-Lernen von zufälligen Entscheidungsbäumen.
LanguageTool (2019). LanguageTool.
Lewis, D. D., & Catlett, J. (2014). Heterogeneous uncertainty sampling for supervised learning. *Machine learning proceedings*.
Li, H. (2015). Twitter mining for disaster response: A domain adaptation approach. *12th international conference on information systems for crisis response and management (ISCRAM)* (pp. 1–7). .
Li, H., Caragea, D., Caragea, C., & Herndon, N. (2017). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of contingencies and crisis management (JCCM), 26*(1), 16–27.
Li, Y., & Manoharan, S. (2013). A performance comparison of SQL and NOSQL databases. In: *Proceedings of the IEEE pacific RIM confonference on communications, computers, and signal processing* (pp. 15–19). .
Ludwig, T. (2017). Situated crowdsourcing during disasters: Managing the tasks of spontaneous volunteers through public displays. *International journal of human-computer studies (IJHCS), 102*(C), 103–121.
Ludwig, T., Reuter, C., & Pipek, V. (2015). Social haystack: Dynamic quality assessment of citizen-generated content during emergencies. *Transactions on human-computer interaction (ToCHI), 21*(4) 17:1-17:27.
Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Identifying suspicious URLs : An application of large-scale online learning. *Proceedings of the annual international conference on machine learning* (pp. 681–688). .
Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60). .
Markham, D., & Muddiman, A. (2016). *EmerGent deliverable 4.4: Specification of mining methods to develop, version 2*.
Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under crisis: Can we trust what we RT ? *Proceedings of the first workshop on social media analytics* (pp. 71–79). .
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97 American Psychological Association, US.
Moi, M., Friberg, T., Marterer, R., Reuter, C., Ludwig, T., Markham, D., Hewlett, M., & Muddiman, A. (2015). Strategy for processing and analyzing social media data streams in emergencies. *Proceedings of the international conference on information and communication technologies for disaster management (ICT-DM)* (pp. 1–7). .
Moore, A. W. (1991). An intoductory tutorial on kd-trees. *Efficient memory-based learning for robot control*.
Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2016). Rapid classification of crisis-related data on social networks using convolutional neural networks. *Proceedings of the AAAI conference on web and social media*.
Nguyen, D. T., Ofli, F., Imran, M., & Mitra, P. (2017). Damage assessment from social media imagery data during disasters. *Proceedings of the 2017 IEEE/ACM international conference on advances in social network analysis and mining* (pp. 569–576). .
Nguyen, M.-T., Kitamoto, A., & Nguyen, T.-T. (2015). TSum4act: A framework for retrieving and summarizing actionable tweets during a disaster for reaction. In T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, & H. Motoda (Eds.). *Advances in knowledge discovery and data mining. PAKDD 2015. Lecture notes in computer science* (pp. 64–75). Springer.
Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing big social data. *Journal of big data, 4*(1), 1–19.
Onorati, T., Díaz, P., & Carrion, B. (2018). From social networks to emergency operation centers: A semantic visualization approach. *Future generation computing systems, 95*, 829–840.
Palen, L. (2010). A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. In: *Proceedings of the ACMBCS visions of computer science conference* (pp. 1–12). .
Palen, L., & Anderson, K. M. (2016). Crisis informatics: New data for extraordinary times. *Science, 353*(6296), 224–225.
Palen, L., & Hughes, A. L. (2018). Social media in disaster communication. In H. Rodríguez, W. Donner, & J. E. Trainor (Eds.). *Handbook of disaster research* (pp. 497–518). Cham: Springer International Publishing.
Párraga Niebla, C. (2011). Alert4All : An integrated concept for effective population alerting in crisis situations. *Proceedings of the international conference on information systems for crisis response and management (ISCRAM)*.
Perry, K. (2017) "As I #prayforlasvegas I pray for us all. Find each other out there.…https://www.instagram.com/p/BZwx8oVle7s/ [Tweet].
Plotnick, L., & Hiltz, S. R. (2018). Software innovations to support the use of social media by emergency managers. *International journal of human-computer interaction, 34*(4), 367–381.
Plotnick, L., Hiltz, S. R., Kushma, J., & Tapia, A. (2015). Red tape: Attitudes and issues related to use of social media by U.S. county-level emergency managers. *Proceedings of the information systems for crisis response and management (ISCRAM)*.
Plotnick, L., & Hiltz, S. R. (2016, January). Barriers to use of social media by emergency managers. *Journal of homeland security and emergency management, 13*(2), 247–277.

Pohl, D. (2013). Social media analysis for crisis management: A brief survey. Available: http://stcsn.ieee.net/e-letter/vol-2-no-1/social-media-analysis-for-crisis-management-a-brief-survey [Accessed: 25 May 2014].
Pohl, D., Bouchachia, A., & Hellwagner, H. (2015, June). Social media for crisis management: Clustering approaches for sub-event detection. *Multimedia tools and applications, 74*(11), 3901–3932.
Porter, M. (2019). Snowball.
POWERS, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of machine learning technologies, 2*(1), 37–63.
Purohit, H., Castillo, C., Diaz, F., Sheth, A., & Meier, P. (2014). Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday, 19*(1), 1–7.
Purohit, H., Castillo, C., Imran, M., & Pandey, R. (2018). Ranking of social media alerts with workload bounds in emergency operation centers. *Proceedings of the IEEE/WIC/ACM international conference on web intelligence (WI)* (pp. 206–213). .
Rao, R., Plotnick, L., & Hiltz, S. R. (2017). Supporting the use of social media by emergency managers: Software tools to overcome information overload. *Proceedings of the hawaii international conference on system sciences (HICSS)* (pp. 304–312). .
Ren, S., Lian, Y., & Zou, X. (2014). Incremental naïve bayesian learning algorithm based on classification contribution degree. *Journal of computers, 9*(8), 1967–1974.
Reuter, C., Amelunxen, C., & Moi, M. (2016a). Semi-automatic alerts and notifications for emergency services based on cross-platform social media data – evaluation of a prototype. *Informatik 2016: Von Menschen für Menschen*.
Reuter, C., Heger, O., & Pipek, V. (2013). Combining real and virtual volunteers through social media. *Proceedings of the information systems for crisis response and management (ISCRAM)* (pp. 780–790). .
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.
Reuter, C., & Kaufhold, M.-A. (2018). Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics. *Journal of contingencies and crisis management, 26*(1), 41–57.
Reuter, C., Ludwig, T., Kaufhold, M.-A., & Pipek, V. (2015a). XHELP: Design of a cross-platform social-media application to support volunteer moderators in disasters. *Proceedings of the conference on human factors in computing systems (CHI)* (pp. 4093–4102). .
Reuter, C., Ludwig, T., Kaufhold, M.-A., & Spielhofer, T. (2016b). Emergency services attitudes towards social media: A quantitative and qualitative survey across Europe. *International journal of human-computer studies, 95*, 96–111.
Reuter, C., Ludwig, T., Kotthaus, C., Kaufhold, M.-A., von Radziewski, E., & Pipek, V. (2016c). Big data in a crisis? Creating social media datasets for emergency management research. *i-com: journal of interactive media, 15*(3), 249–264.
Reuter, C., Ludwig, T., Ritzkatis, M., & Pipek, V. (2015b). Social-QAS: Tailorable quality assessment service for social media content. *Proceedings of the international symposium on end-user development (IS-EUD)* (pp. 156–170). .
Reuter, C., Ritzkatis, M., & Ludwig, T. (2014). Entwicklung eines SOA - basierten und anpassbaren bewertungsdienstes für inhalte aus sozialen medien. *Informatik 2014 - Big Data - Komplexität meistern* (pp. 977–988). .
Reuter, C., & Scholl, S. (2014). Technical limitations for designing applications for social media. *Mensch & Computer: Workshopband* (pp. 131–140). .
Reuter, C., & Spielhofer, T. (2017). Towards social resilience: A quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. *Journal of technological forecasting and social change, 121*, 168–180.
Rohweder, J. P., Kasten, G., Malzahn, D., Piro, A., & Schmid, J. (2011). Informationsqualität - definitionen, dimensionen und Begriffe. *Daten und informationsqualität - auf dem Weg zur Information Excellence* (pp. 25–45). .
Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., & Ghosh, S. (2015). Extracting situational information from microblogs during disaster events: A classification-summarization approach. *Proceedings of the ACM international on conference on information and knowledge management* (pp. 583–592). .
Rudra, K., Goyal, P., Ganguly, N., Mitra, P., & Imran, M. (2018). Identifying sub-events and summarizing disaster-related information from microblogs. *Proceedings of the international ACM SIGIR conference on research & development in information retrieval* (pp. 265–274). .
Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In: *Proceedings of the international conference on world wide web* (pp. 851). .
Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the american society for information ccience, 26*(6), 321–343.
Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the american society for information science and technology, 58*(13), 1915–1933.
Schamber, L., & Eisenberg, M. B. (1988). Relevance: The search for a definition. *Annual meeting of the American Society for Information Science* (pp. 17). .
Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A *re*-examination of relevance: Toward a dynamic, situational definition. *Information processing and management, 26*(6), 755–776.
Sebastiani, F. (2002). Machine {Learning} in {Automated} {Text} {Categorization}. *ACM Computing Surveys, 34*(1), 1–47.
Settles, B. (2010). *Active learning literature survey*. *15*. *Active learning literature survey* (pp. 201–221). Madison: University of Wisconsin.
Shankaranarayanan, G., Iyer, B., & Stoddard, D. (2012). Quality of social media data and implications of social media for data quality. *Proceedings of the international conference on information quality (ICIQ)* (pp. 311–325). .
Soden, R., & Palen, L. (2018). Informating crisis: Expanding critical perspectives in crisis informatics. *Proceedings of the ACM on human-computer interaction*.
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. *Proceedings of the international conference on information and communication technologies for disaster management (ICT-DM)* (pp. 1–6). .
Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in Twitter to improve information filtering. In: *Proceedings of the ACM SIGIR conference on research and development in information retrieval SE* (pp. 841–842). .
Starbird, K., & Palen, L. (2004). Pass it on?: Retweeting in mass emergency. *Proceedings of the information systems for crisis response and management (ISCRAM)*. *2010*. *Proceedings of the information systems for crisis response and management (ISCRAM)* (pp. 1–10).
Starbird, K., & Palen, L. (2011). Voluntweeters: Self-organizing by digital volunteers in times of crisis. *Proceedings of the conference on human factors in computing systems (CHI)*.
Stieglitz, S., Bunker, D., Mirbabaie, M., & Ehnis, C. (2017). Sense-making in social media during extreme events. *Journal of contingencies and crisis management (JCCM)*.
Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social media analytics: An interdisciplinary approach and its implications for information systems. *Bususiness and information systems engineering (BISE), 6*(2), 89–96.
Stieglitz, S., Mirbabaie, M., Fromm, J., & Melzer, S. (2018a). The adoption of social media analytics for crisis management - challenges and opportunities. *Proceedings of the european conference on information systems (ECIS)*.
Stieglitz, S., Mirbabaie, M., & Milde, M. (2018b). Social positions and collective sense-making in crisis communication. *International journal of human-computer interaction, 34*(4), 328–355.
Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018c). Social media analytics – challenges in topic discovery, data collection, and data preparation. *International journal of information management, 39*, 156–168.
Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CONLL-2003 shared task. *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003 -*. *4*. *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003 -* (pp. 142–147).
Tucker, S., Ireson, N., Lanfranchi, V., & Ciravegna, F. (2012). 'Straight to the information I need': Assessing collational interfaces for emergency response. *Proceedings of the international conference on information systems for crisis response and management (ISCRAM)*.
Uysal, I., & Croft, W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. *Proceedings of the ACM international conference on information and knowledge management* (pp. 2261). .
Verma, S. (2011). Natural language processing to the rescue? extracting 'Situational awareness' tweets during mass emergency. *Proceedings of the 5th international AAAI*

conference on weblogs and social media (pp. 385–392). .
Vieweg, S. (2012). Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications, 1–300.
Vieweg, S. (2012b). Twitter communications in mass emergency. In: *Proceedings of the conference on computer supported cooperative work companion* (pp. 227). .
Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010a). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. *Proceedings of the conference on human factors in computing systems (CHI)* (pp. 1079–1088). .
Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010b). Microblogging during two natural hazards events. In: *Proceedings of the conference on human factors in computing systems (CHI)* (pp. 107–1088). .
vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the association for information systems, 37*, 205–224.
Wang, A., Wan, G., Cheng, Z., & Li, S. (2009). An incremental extremely random forest classifier for online learning and tracking. *Proceedings of the international conference on image processing*. ICIP.
Weißweiler, L., & Fraser, A. (2017). Developing a stemmer for German based on a comparative analysis of publicly available stemmers. *Proceedings of the international conference of the german soceity for computational linguistics and language technology*.
White, J. I., Palen, L., & Anderson, K. M. (2014). Digital mobilization in disaster response: The work & self - organization of on-line pet advocates in response to hurricane sandy. *Proceedings of the conference on computer supported cooperative work (CSCW)* (pp. 866–876). .
Wilson, T., Stanek, S. A., Spiro, E. S., & Starbird, K. (2017). Language limitations in rumor research? comparing french and English tweets sent during the 2015 Paris attacks. *Proceedings of the information systems for crisis response and management (ISCRAM)* (pp. 546–553). .
Wise Bitch. (2009). Country residents outside of Fargo are surrounded by flood waters. Some R being rescued [Tweet].
Wobbrock, J. O., & Kientz, J. A. (2016). Research contribution in human-computer interaction. *Interactions, 23*(3), 38–44.
World Wide Web Consortium. (2016). Activity streams 2.0. *W3C Recommendation*. [Online]. Available https://www.w3.org/TR/activitystreams-core/.
Yang, Y., & Loog, M. (2017). Active learning using uncertainty information. *Proceedings of the international conference on pattern recognition*.
Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., & Starbird, K. (2018). From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proceedings of the ACM on human-computer interaction*.
Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering, 18*(1), 63–77.