Information Systems 57 (2016) 88–110

Contents lists available at ScienceDirect

Information Systems

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

A rapid-prototyping framework for extracting small-scale incident-related information in microblogs: Application of multi-label classification on tweets

Axel Schulz [a,*], Eneldo Loza Mencía [b], Benedikt Schmidt [a]

[a] Telecooperation Lab, TU Darmstadt, Hochschulstrasse 10, 64289 Darmstadt, Germany
[b] Knowledge Engineering Group, TU Darmstadt, Hochschulstrasse 10, 64289 Darmstadt, Germany

ARTICLE INFO

Article history:
Received 29 November 2014
Received in revised form
12 August 2015
Accepted 6 October 2015
Available online 14 November 2015

Keywords:
Mining microblogs
Disaster management
Multi-label classification
Incident type detection

ABSTRACT

Small scale-incidents such as car crashes or fires occur with high frequency and in sum involve more people and consume more money than large and infrequent incidents. Therefore, the support of small-scale incident management is of high importance.

Microblogs are an important source of information to support incident management as important situational information is shared, both by citizens and official sources. While microblogs are already used to address large-scale incidents detecting small-scale incident-related information was not satisfyingly possible so far.

In this paper we investigate small-scale incident reporting behavior with microblogs. Based on our findings, we present an easily extensible rapid prototyping framework for information extraction of incident-related tweets. The framework enables the precise identification and extraction of information relevant for emergency management. We evaluate the rapid prototyping capabilities and usefulness of the framework by implementing the multi-label classification of tweets related to small-scale incidents. An evaluation shows that our approach is applicable for detecting multiple labels with an match rate of 84.35%.

1. Introduction

Improving situational awareness is a major prerequisite for efficient decision making in urban emergency management. This is gained based on information about the environment within the volume of time and space affected by a crisis [1]. Currently, urban emergency management gains situational awareness from information provided by bystanders and the city infrastructure. While there is no structured way of collecting information from bystanders, the city infrastructure is built from cost-intensive sensors, thus, both sources have their drawbacks and limitations. Therefore, using additional information sources for acquiring incident information to improve situational awareness is highly desirable.

With the advent of social media new means to collect information that contributes to situational awareness have emerged. Twitter messages have been shown as valuable information source during incidents. Real-time detection of earthquakes and tracking of diseases [2,3], as well as the detection of fires and floods [4] are but a few examples that demonstrate how valuable such approaches can be in the course of crisis mitigation.

* Corresponding author.
E-mail addresses: schulz.axel@gmx.net (A. Schulz), eneldo@ke.tu-darmstadt.de (E. Loza Mencía), benedikt.schmidt@tk.informatik.tu-darmstadt.de (B. Schmidt).

http://dx.doi.org/10.1016/j.is.2015.10.010
0306-4379/

Decision makers in crisis management could greatly benefit from this new source of information, if appropriate, and reliable information from citizens could be retrieved in time. However, this information source remains unused. One reason for this is the sheer amount of information created every day, which results in an *information overload* that is not manageable for humans. Also, automatic processing is not easily applied, because of the heterogeneous and unstructured nature of the data. Thus, inferences on all available information are not easily drawn and potentially valuable situational information remains unused. Nevertheless, it is beyond doubt that a massive stream of user-generated content contains pieces of highly relevant information that is not known to the decision maker. Harvesting this information can contribute to a better situational picture finally leading to an improved situational awareness compared to a situation where this information is not available at all.

The aforementioned examples focus on large-scale incidents. This type of incidents is well researched. Still, people contribute much valuable information about crisis small-scale incidents using social media, as well. This information is currently not taken into account by decision makers in emergency management. One reason for this is the increased complexity of filtering: small scale incidents have a smaller duration and affect a smaller area than large scale incidents. Therefore, small scale incidents are reported less frequently, resulting in a difficult challenge with respect to identification and information extraction [5]. Due to this specific reporting behavior, methods used to extract information about large-scale incidents are not immediately applicable to the small-scale incident domain. Thus, the development of dedicated processing methods to extract information from tweets that address small-scale incidents is required. To develop such methods, a detailed analysis of the usage behavior of social media for sharing small-scale incident information is required.

In this paper we address the aforementioned challenges and provide the following contributions to the extraction of incident-related information using social media for improving situational awareness.

Analysis of small-scale incident reporting behavior: We analyze information sharing behavior of people and organizations during small-scale incidents in social media. Our analysis shows that social media is in fact used to report small-scale incident-related information. Furthermore, we show that the shared information is generated by valuable sources and potentially complements the existing channels for situational information (bystanders and the city infrastructure).

Rapid prototyping framework for information extraction: We introduce a rapid prototyping framework for extracting incident-related information from tweets. The framework implements the processing steps required for the evaluation of different information extraction methods in the domain of emergency management. Therefore, the framework simplifies method development and evaluation because tedious, re-occurring tasks like preprocessing and evaluation setup are offered by the framework.

Application of multi-label classification for extracting incident-related information: Supervised classification is one means for extracting incident-related information contained in social media. However, up to now relevant information is only detected using multi-class classification, i.e., tweets are labeled with exactly one label out of a predefined label set (e.g., “fire” or “crash”). During our research we found that assigning only one label would result in the loss of important situational information for decision making in crisis management. For example the tweet “THIS CAR HIT THE FIRE HYDRANT AND CAUGHT FIRE....SOMEONE HOLIDAY ALTERED” not only contains information about a car crash but additional information about fire and injury are included. With this paper we extend our earlier work on multi-label classification of incident-related tweets and show its realization with our framework [6].

The paper is structured as follows. First, we report the specific characteristics of user-generated content. Second, we analyze value and frequency of incident-related information shared. Third, our framework for the rapid prototyping of techniques to extract small-scale incident-related information from microblogs is introduced. As a showcase motivated from the analysis, we apply and evaluate the framework with a multi-label classification task. Finally, we discuss related work and conclude the paper.

2. Background

The main interest of this paper is the extraction of incident-related information from social media. In the following, we introduce the definition of the terms *event* and *incident* and give an overview of the basic characteristics of user-generated content.

2.1. Specification of events and incidents

Events generally refer to occurrences in the real world. Up to now, there is no consensus on the definition of an event [7]. Here, we follow the most basic and general definitions of the term. In the first Topic Detection and Tracking (TDT) challenge, Allen et al. [8] defined an event as “some unique thing that happens at some point in time”. This definition was later refined by Yang et al. [9] to: “an event identifies something (non-trivial) happening in a certain place at a certain time”. Both definitions show that an event is clearly characterized by spatial and temporal dimensions. Furthermore, events can be regarded as “instances of topics” [9].

Based on these initial definitions, we define an event as follows:

- An *event* is something that is happening in the real world at a certain place, at a certain time, and which has a thematic dimension to be captured by a topic name.

Furthermore, following the definitions provided in related work, an event is characterized by three information dimensions:

- A topic (i.e., a thematic dimension).
- A location (i.e., a spatial dimension).
- A specific time period (i.e., a temporal dimension).

Let us consider the following tweet *“Unhappy :(Traffic jam bcs. of car crash. H5”*. The tweet refers to an event with the

topic (thematic dimension) to be called *car crash*. The event happened at a specific time, on March 12, 2014 (tweet send time), at a specific location, on the highway A5.

Throughout this paper, we focus on incidents (or accidents) as a specific type of event that is specified by the same three properties. Following the definition of an incident by the Federal Emergency Management Agency (FEMA) [10], we define an incident as: an incident is an unexpected event in the real world typically resulting in a damage or injury that happens at a certain place, at a certain time, and which can be described by a topic.

Incidents can be further distinguished with respect to their impact as large-scale incidents and small-scale incidents. We assume that information sharing behavior for small-scale incidents differs from the well-researched large-scale incident information sharing. This assumption grounds on the nature of small-scale incidents:

- *Spatial and temporal extent*: Small-scale incidents affect a limited spatial extent. Furthermore, these incidents generate effects that mostly last for a very limited amount of time.
- *Commonplace nature*: Small-scale incidents occur frequently and everyday; thus, a large number of singular incidents exist.
- *Limited information*: The amount of information shared during small-scale incidents is low compared to large-scale incidents.

Consider a car crash as an example for these characteristics: The crash takes place at an intersection of two streets, affects the traffic for a limited amount of time, e.g., for two hours until the car is removed. People passing by might use social media and mention it, but due to the specific place, the limited time and the ordinary nature of the incident one could assume that the situational information shared is rather limited.

2.2. Background on user-generated content and twitter

User-generated content is defined by [11] as "various forms of media content that are publicly available and created by end-users." Social media was built upon the principles of the Web 2.0 and allows the sharing and creation of user-generated content. Different types of social media [12] can be differentiated: social networking sites such as Facebook[1] and LinkedIn[2] allow connecting with friends, whereas social media data in the form of videos, audio files, or photos is shared on YouTube[3] or Flickr.[4] Textual content is mostly shared in blogs such as Engadget[5] or with limited content on microblogging sites such as Twitter[6] or Tumblr.[7]

[1] https://www.facebook.com/
[2] https://www.linkedin.com/
[3] https://www.youtube.com/
[4] https://www.flickr.com/
[5] http://www.engadget.com/
[6] https://twitter.com/
[7] https://www.tumblr.com/

In this paper, we focus on textual content as there has been a rapid growth of text data in social media [13]. Furthermore, we focus on Twitter as one very prominent platform on which information is shared every day and by a variety of people. In 2013, Twitter had about 240 million active users [14], who shared more than 500 million messages per day [15]. This huge amount of data provides a wide base of information for a variety of topics.

On Twitter, users can post short messages of up to 140 characters in length called *tweets*. These *microposts* are either sent from mobile devices, from third-party applications, or from web applications. For each user, a stream of microposts is displayed as a *microblog*, which is the reason why Twitter is often referred to as a microblogging platform. Twitter is also a social network as users are able to follow each other's microblogs. Furthermore, users can forward or *retweet* each other's messages.

While communicating, people use a variety of Twitter-specific symbols [16]. Placenames or user names are referenced using the "@" symbol. Also, Twitter allows us to use the hashtag "" symbol to specify a number of keywords or a topic of a tweet. For instance, "swineflu" was introduced for the trending news event. However, there is no common convention on how to name these topics [17]. Furthermore, hashtags are not necessarily unique and are highly dependent on how they are used in the whole social network [18].

2.3. Characteristics of user-generated content

Social media data is significantly different compared to other information sources. It has different characteristics that complicate answering research questions. As outlined in the introduction, social media data such as tweets is inherently noisy and unstructured. In Listing 1, an example tweet illustrates the unstructuredness of textual information in social media.

Listing 1. Example tweet showing the unstructuredness of textual information.

```
RT: @People  0noe friday afternoon in heavy
    traffic, car crash on I-90, right lane
    closed
```

First, Twitter-specific annotations such as @-mentions and retweets are used. Second, abbreviations such as "0noe" are present. Third, very short sentences are written due to the restricted length of a tweet. However, the information density is high as, for example, the position, and the type of incident is mentioned.

The characteristics that user-generated content shares are described in the following:

- *Vast amount of information*: The amount of social media data created every day is further increasing [13]. This results in an information overflow, which is difficult to handle. There is a lack of time to analyze the incoming flood of data, especially for time critical decisions.
- *Heterogeneity*: The types of social media data differ. Social media content might be audio or video files, images, or

textual content. This content is not necessarily interlinked. Furthermore, it is shared across various platforms.
- *Dynamism*: Information in social media is changing frequently. For instance, people update their current location or their current status. Furthermore, interests change rapidly as trends evolve. Thus, user-generated content has a very dynamic nature.
- *Reliability*: Social media platforms are used by companies, domain experts, as well as a variety of regular users. Also, these platforms are spammed by automatic bots and people alike. This results in a high variety of quality, which makes the identification of relevant and reliable information much harder.
- *Interconnectedness*: Compared with traditional texts, textual data in social media is not independent and identically distributed (i.i.d.) [13]. For instance, people annotate their content with specific annotations such as hashtags, which are used to refer to a certain topic. Also, users share URLs that refer to external websites. Furthermore, users themselves are interlinked with each other via friendship or follower relationships.

In particular, textual content shared in social media has special properties that pose new challenges to our research goal:

- *Unstructuredness of textual content*: Text shared in social media is inherently unstructured. Users tend to use abbreviations or nonstandard vocabulary in their posted content. This is even increased through the diversity of authorship; thus, many different styles of writing can be found. Some users such as domain experts post-information carefully, while other users do not.
- *Length of textual content*: In most social networks, the length of each posting is limited. For instance, messages on Twitter are limited to 140 characters. Thus, short messages consist of only few phrases or sentences.
- *Regional variation*: Words and phrases used in social media texts are interconnected to the location where a text was created. Thus, mechanisms that apply for one city may not necessarily apply just as precisely for data of a different city.

3. Preliminary study

Small-scale incidents affect only few people, because they are ordinary events in the urban environment and have a limited duration. Considering that user-generated content follows rules of awareness one can assume that the amount of information related to small-scale incidents is rather small. This is a basic assumption that potentially has important effects on the techniques to interact with incident information. First, it is necessary to understand information sharing behavior. Second, existing processing techniques for user-generated content need to be assessed for small-scale incidents and the development of new, use-case specific methods might be desirable. In the following, we will investigate these aspects in detail in a study of small-scale incident-related information shared on Twitter. This information is required to (1) understand the type of incidents covered in social media, to (2) get an overview of incident-related situational information provided, and to (3) understand the identity of the user groups incident-related information originates from. We assume that studies on large-scale incidents are not appropriate to answer our questions due to the described differences.

In the following, we report a study which analyzes small-scale incident reports in tweets for the purpose of providing a first insight into the three mentioned information requirements.

3.1. Data collection and coding of tweets

Before introducing the results, we elaborate on how the data was collected, selected, and labeled, in order to extract manageable datasets.

Our analysis of small-scale incident-related information shared on Twitter is based on a dataset of 7.5 million tweets collected from 11/19/12 to 02/07/13 that contains the following number of tweets: 213 tweets related to car accidents, 212 tweets related to fire incidents, 231 tweets related to shooting incidents, and 544 not incident-related tweets. Thus, we could identify 656 incident-related tweets in our dataset.

The tweets were collected in a 15 km radius around the city centers of Seattle, WA and Memphis, TN. For the collection, we used the Twitter Search API. Though we know about the limitations of this API, prior work shows [4] the appropriateness for the task at hand as the Search API not only provides explicitly geotagged tweets, but also tweets that have been geocoded by Twitter (e.g., using the user profile[8]). We decided not to use the Streaming API as we wanted to have a fixed benchmark dataset for our evaluation. Also, labeling costs are expensive, thus, we needed to restrict to this dataset.

Seattle and Memphis were chosen because our pre-studies among US cities showed that the tweets for those cities contained a useful amount of incident-related information. While the use of two cities investigates the transferability of the results among cities that already have a relevant amount of incident-related information, the transfer to cities with less incident-related tweets is open for future work.

Because a smaller set was sufficient and accounting for the cost of manual labeling processes we reduced the dataset by applying the incident-keyword filtering as presented in [19]. The resulting 1200 keyword-filtered tweets were manually labeled by five researchers of our department. All researchers have experience in emergency management and data labeling. Every tweet was labeled by each researcher. To assign the final coding four out of five coders had to agree on a label. If no agreement could be achieved, the final label was resolved in a group discussion.

The following coding schemes were applied:

Incident type label assignment: We focus on three diverse incident types throughout the paper in order to differentiate tweets contributing to situational awareness.

[8] https://dev.twitter.com/docs/using-search

Three classes have been chosen, because we identified them as the most common incident types in Seattle using the Seattle Real Time Fire Calls dataset,[9] which is a frequently updated source for official incident information. Thus, for our study we focused on three classes consisting of very common and distinct incident types and one neutral class: “car crash”, “fire”, “shooting”, and “not incident related”. The coding scheme was designed to enable the clear differentiation of different types of incidents.

Situational information tag assignment: Each tweet was classified with respect to the coverage of detailed incident information. Each incident-related tweet could be annotated in an online survey by the five researchers with free tags to describe the tweet content. The maximal length for tags was limited to three words.

For instance, the tweet “1 killed, 1 injured in South Memphis crash on I-240: One person was killed Monday morning in a crash on Interstate...” was annotated with “1 killed”, “1 injured”, “crash”, and “interstate”. Overall, 1299 tags were assigned to the 656 incident-related tweets.

3.2. Study

Our analysis of situational information contained in incident-related tweets focuses on three different aspects. First, we analyze the user groups who generate incident-related information. Then we analyze situational information contained in the tweets in a quantitative and a qualitative content analysis.

3.2.1. Exploration of user types

In the first part of the study, we analyze the origin of incident-related tweets in terms of user groups. The origin of tweets is crucial to understand whether possibly new, yet unknown information about incidents is captured; if most of the tweets were sent automatically by emergency management systems, the information would not be new, thus, not valuable. Furthermore, related work shows that information quality highly depends on the user group information it originates from [20].

Approach: To analyze which user groups contribute small-scale incident related information, we analyzed the origin of the incident-related tweets in our dataset. Using the description of the users' Twitter profiles, two researchers from our department manually coded all users into different groups. Following the approach described by Choudhury et al. [21], we identified five user groups. Official organizations like the Seattle Fire Department are categorized as *emergency management organizations* (EMO). Organizations not related to emergencies, like magazines, are considered as *other organizations* (ORG). Furthermore, we found specialized traffic reporters or journalists, which are represented as *emergency management journalists/bloggers* (EMJ), in contrast to *other journalists/bloggers* (JOU). Users not present in these groups are categorized as *individual users* (I).

Results: We were able to identify 246 unique users that are sharing incident-related tweets. The first bar of each stacked cluster in Fig. 1 shows the distribution of the number of users for each category according to the different types of incidents. We can observe that a high proportion of the users (196 of 246) reporting about the different incident types are individual users. This finding is important as it shows that many different people share incident-related information. On the other hand, only few emergency management organizations (11) and focused journalists (2) are publishing incident-related tweets, which is obvious as their number is limited w.r.t. a city.

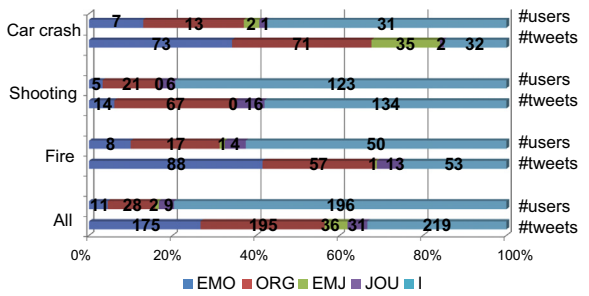

Fig. 1. Distribution of number of users and number of tweets from different user categories by incident type in the dataset.

The second bar of each stacked cluster in Fig. 1 shows the overall number of tweets shared by each user category for the different types of incidents. We can notice that even though the number of the emergency management organizations and other organizations is significantly smaller than the number of individual users, most of the tweets are shared by users belonging to these organizations (overall 56%). The individual users share 33.3% of the tweets, though, the number of tweets by individual users regarding the shootings is much higher. The reason for this might be that shootings are more of public interest compared to car crashes and fires. Furthermore, the results indicate that individual users contribute only one or at most two tweets regarding small-scale incidents.

3.2.2. Quantitative analysis of situational information

In this part of the study, we report a quantitative analysis of the situational information contained in our dataset.

Approach: For analyzing quantitative information, we performed several automatic steps. On the one hand, the detection of URLs in tweets was easily achieved using regular expressions. On the other hand, we are also interested in understanding the usage of temporal and spatial information in incident-related tweets. For extracting temporal information automatically, we adapted the HeidelTime [22] framework for temporal expression recognition. Spatial information was extracted using the approach described in [23]. It is important to note that we did not analyze the metadata of a tweet, but only the information provided in the message itself.

Results: The automatic data coding allows us to examine the differences in terms of characteristics of content posted by different user categories. We automatically counted the number of spatial and temporal mentions in the tweets, as they provide information about the location and the time of an incident. Furthermore, as URLs are often

[9] http://data.seattle.gov

Table 1
Content characteristics of tweets differentiated by user type.

Type	Car Crash			Fire		
	Location (%)	Time (%)	URL (%)	Location (%)	Time (%)	URL (%)
EMO	98.63	15.07	4.11	97.73	38.64	11.36
ORG	91.55	12.68	56.34	89.47	40.35	100.00
EMJ	100.00	11.43	0.00	100.00	0.00	0.00
JOU	100.00	50.00	100.00	84.62	38.46	84.62
I	59.38	15.63	18.75	49.06	39.62	37.74
Type	Shooting			All Incident Types		
	Location (%)	Time (%)	URL (%)	Location (%)	Time (%)	URL (%)
EMO	92.86	28.57	42.86	97.71	28.00	10.86
ORG	82.09	26.87	97.01	87.69	25.64	83.08
EMJ	0.00	0.00	0.00	100.00	11.11	0.00
JOU	87.50	0.00	37.50	87.10	19.35	61.29
I	62.69	22.39	36.57	58.90	25.57	34.25

posted as references to describe pictures or additional descriptions, we analyzed the numbers of URLs based on regular expressions.

In Table 1 we show the results of our analysis. We can conclude that, at least for our collected data, organizations and journalists always tend to mention spatial locations, while only around half of the tweets shared by individual users contain location mentions. As location mentions can be on country, city, or even street level, we will have a look into the level of detail of the mentioned location information in the following qualitative analysis section. Regarding temporal information, no clear differences between the user types can be found. Most of the temporal mentions are shared during fires or shootings compared to less mentions during car crash incidents.

Most of the links are posted by organizations, journalists and individual users. In contrast, the EMO category of users usually does not include URLs in the tweet. A possible reason is the tendency of EMOs to tweet early about incidents while there is still no web content to be referenced.

Finally, we compared the differences in terms of characteristics between incident-related tweets and tweets not related to incidents. We randomly chose 219 tweets that were coded as not incident-related during our user study and which appeared in the same time period as the incident-related tweets. As shown in Table 2, incident-related tweets contain twice as much location mentions compared to not incident-related tweets.

Regarding the temporal mentions as well as the URLs we could not find clear differences between incident and not incident-related tweets. Nevertheless, the number of tweets with temporal mentions is quite high. Summarized, during small-scale incidents large amount of valuable situational information is shared. In most cases, spatial information is posted referring to the situation of incident occurrence.

3.2.3. Qualitative analysis of situational information

Following the previous study, we present the results of a qualitative analysis of situational information shared in our dataset. For example, as we have shown that location mentions are commonly present, we wanted to find out how precise location information in tweets is. Furthermore, we show that several other important situational updates are shared.

Approach: For a qualitative analysis, the lead authors identified and organized situational information into categories following the approach of [4]. We identified and coded situational information into the following categories: *Precise Incident Type* is a more fine-grained description of the incident type; *Affected Objects* refers to affected things such as buildings or cars that were damaged; *Damage/Injury Reports* are information describing the condition of involved people; *Road Conditions* is a description of the road conditions; *Precise Location* is a description of the location on street-level.

The categories were identified based on the thematic coding described in the Method section. Following [4], each type of information that appeared more than five times was given a category name. All other types were not analyzed in this study and discarded. Each tweet may be assigned to none, one or more categories. E.g., the following tweet provides information about possible injuries, road conditions, as well as precise location information: “Traffic: Still dealing w/ MAJOR delays *BOTH* directions on I-240 (Midtown) near S Pkwy due to early injury crash! WREG MEMtraffic.” Finally, the annotators assigned the categories to all 656 incident-related tweets based on the tags provided in the online survey.

Results: In Table 3 the percentages for each user type and each category are shown as well as the overall percentage of the appearance of each category in all incident-related tweets.[10] Overall around 10% of all incident-related tweets contain information about the precise incident type, which might be helpful for fine-grained differentiation of the situation at hand. Most of those tweets are

[10] Note that the numbers do not necessarily sum up to 100% as categories with less than five tweets are not present and tweets can also belong to more than one category.

Table 2
Content characteristics of incident-related tweets and tweets not related to incidents.

Type	Location (%)	Temporal (%)	URL (%)
Incident	81.40	25.15	41.92
No Incident	43.84	19.18	43.84

Table 3
Overview of different categories of situational information for each user type and in relation to the overall number of tweets per user type (in brackets).

Precise incident type	
All	9.58%
EMO	23.80% (10.0%)
ORG	52.38% (17.93%)
EMJ	1.58% (2.94%)
JOU	4.76% (11.11%)
I	17.46% (4.98%)
Affected objects	
All	21.46%
EMO	7.80% (7.33%)
ORG	37.59% (28.80%)
EMJ	2.13% (8.82%)
JOU	10.64% (55.56%)
I	41.84% (26.70%)
Damage/Injury reports	
All	21.16%
EMO	10.07% (9.33%)
ORG	48.92% (36.96%)
EMJ	0.00% (0.00%)
JOU	2.88% (14.81%)
I	38.13% (23.98%)
Road conditions	
All	7.31%
EMO	27.08% (8.67%)
ORG	43.75% (11.41%)
EMJ	25.00% (35.29%)
JOU	0.00% (0.00%)
I	4.17% (0.90%)
Precise location	
All	19.93%
EMO	19.08% (16.67%)
ORG	32.82% (23.37%)
EMJ	9.92% (38.24%)
JOU	6.87% (33.33%)
I	31.30% (18.55%)

posted by organizations, compared to a rather low percentage by individual users.

Information about affected objects is shared quite often in incident-related tweets. Most of those tweets are contributed by ORGs and individual users. As it is highly important for emergency managers to know if a school, a chemistry plant, or a truck carrying flammable liquids is on fire, the early reporting by individual users in combination with this information can be very helpful. Also around 21% of the incident-related tweets contain information about the people involved and the amount of injured persons.

With around 7% the road condition information is rather uncommon in incident-related tweets, although it is highly relevant for rescue squads. Actually, ORG institutions and EMJs were the group of users who most frequently shared road conditions. Precise location information, which is mostly accurate on street and intersection level, is shared in 20% of the tweets. In this case, one third of the location information is provided by individual users. This information could be leveraged for geolocalization of the incident.

Summarized, important situational information such as precise location information, information about the type of event, affected objects and injured people is shared in incident-related tweets. Thus, new techniques are necessary to make use of this source of information.

3.3. Discussion about small-scale incident coverage in tweets

Our analysis of information about small-scale incidents contained in microblogs provides important insights: (1) a variety of individuals are sharing information about small-scale incidents, information that is not necessarily available for decision makers, (2) incident-related tweets contain important situational information that could enrich the situational picture. Most important: precise location information is present in the text, which enables decision makers to easily geolocalize the location of an incident. As only around 2% of all tweets are explicitly geotagged, extracting this spatial information from the tweet message could be helpful. Affected objects such as buildings or cars and much more important information about potentially injured persons is shared. This information is especially valuable as it allows better planning of response measures.

Finally, different types of situational information such as road conditions and the number of injured people are shared as well. This is an important finding as not only the presence of an incident is mentioned, but also background information about the event. Differentiating the incident types as well as the different types of background information would allow a more fine-grained sharing of information to the corresponding emergency management agency, i.e., information about the number of injured people could directly be forwarded to medical centers. In this paper, we will explicitly address this issue, i.e., the extraction of a variety of different types of situational information in one extraction step.

The study showed that microblogs are an important source of small-scale incident information. Based on this insight, the remainder of this paper introduces a framework that facilitates the rapid prototyping of techniques to extract small-scale incident information contained in tweets.

4. Architecture of a framework for extraction of incident-related information

In this section, we specify a framework for the rapid prototyping of techniques to extract incident-related information from microblogs. Following the incident definition introduced in this paper, information about topics, locations and time of reported incidents needs to be extracted from social media. As discussed earlier the development of methods for the extraction of the relevant information types is challenging not only due to the complex task of selecting an appropriate information extraction technique and its evaluation but also due to the need for complementary steps. Such complementary steps include preprocessing, labeling and filtering of the data.

With a rapid prototyping framework we want to facilitate the development of techniques for the extracting incident-related information from user-generated content. Therefore, we focus on providing a fully configurable analysis process including relevant steps like preprocessing and crowdsourced data labeling, filtering and evaluation. Thus, the framework supports the development of extraction techniques by reusing framework components. To assess newly developed methods specific consideration is given to an evaluation that allows the simple comparison of different methods based on statistics, e.g., using the significance of the evaluation results. In a nutshell we present a flexible, highly configurable framework for the rapid prototyping of methods for the extraction of incident-related information from microblogs.

The section is structured as follows. First, we elaborate on the extraction of incident-related information and specify resulting requirements for a framework. Second, we discuss the extraction process implemented in the framework.

4.1. Requirements

For the rapid prototyping of methods to extract incident-related information different processing steps and capabilities should be available. In the presented framework we have synthesized our experience with processing pipelines.

Following our definition of event and incident, a framework to analyze tweets with respect to incident-related information needs to provide capabilities for the analysis of the following three dimensions:

- The framework should provide capabilities for the analysis of the *thematic* dimension of an information item.
- The framework should provide capabilities for the analysis of the *spatial* dimension of an information item.
- The framework should provide capabilities for the analysis of the *temporal* dimension of an information item.

We also have identified several requirements for a general-purpose system for extracting incident-related information from microblogs that guided the development of the system.

First, the framework needs to offer support for relevant processing steps and, second, provide means for configuring and reusing those processing steps with limited effort. We have decided for a generic pipeline that comprises collection and filtering, preprocessing, human-based classification, machine-based classification, aggregation, refinement, presentation and usage. Based on a preprocessing of text, feature generation needs to be supported, the availability of different machine and data mining techniques need to be offered. The interconnection of these components was designed to be as flexible as possible as such as system should be able to cope with different machine learning algorithms and feature extraction tasks depending on the task at hand.

Rapid prototyping requires an assessment of the success as well as the repeatability of the results. Therefore evaluation of the developed methods is required. Furthermore, many evaluation scenarios require labeled datasets. Frequently the labeling of data for specific purposes is an inherent part of the rapid prototyping process. Therefore, the labeling should be supported by the framework.

4.2. Architecture of a framework for small-scale incident detection

In the following subsection, the framework for detecting small-scale incident-related information based on user-generated content is presented. The framework relies on two approaches for analyzing a large amount of data: *crowdsourcing* (i.e., the engagement of humans for manual filtering of user-generated content) and *machine learning* for automatic extraction of useful information. The combination of both approaches is necessary as on the one hand, manual analysis of user-generated content is prone to errors. Furthermore, crowdsourcing might result in untrustworthy information [24]. Also, the application of crowdsourcing in time-critical situations as emergencies is not always possible. On the other hand, training and validation is needed for machine learning algorithms in order to adjust to a specific problem domain. For this, commonly, annotated training data is needed. Also, one model trained on one city may not be applicable on data of a different city because of the nature of social media data and the resulting diversity. Thus, already trained models need to be refined to changing conditions. For our framework, we decided to combine both approaches to overcome the limitations of each individual one.

The six steps defined in the framework are summarized in the following (see Fig. 2 for a connection of each of the steps).

4.2.1. Collection and filtering

In the first and initial step, user-generated content is collected. Apart from social media, where valuable information is directly provided, additional information from on-site bystanders can be collected using mobile applications [25]. Twitter provides two major APIs for collecting tweets. First, the Streaming API[11] allows the crawling of real-time Twitter data. This API provides access to a 1% real-time stream of all tweets created worldwide. Second,

[11] https://dev.twitter.com/docs/streaming-apis/streams/public

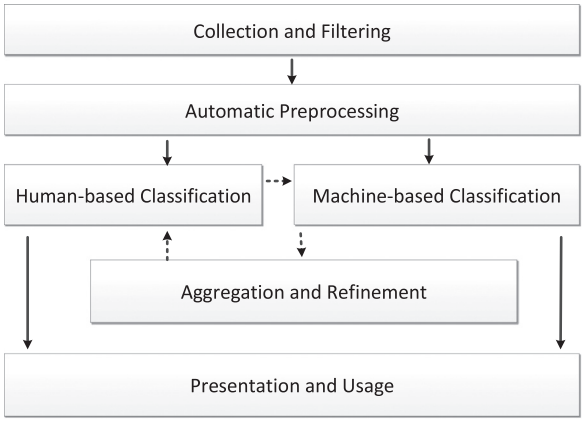

Fig. 2. Overview of the process steps covered by the framework. The dashed errors indicate repeating refinement and training steps, whereas the other steps are typically performed once.

the Search API[12] can be used to get tweets related to certain keywords or a location. It allows specifying a search query containing multiple keywords and GPS coordinates as well as a radius. Using this API, it is possible to collect a stream of tweets for a single city. The Search API provides not only explicitly geotagged tweets but also tweets that have been geocoded by Twitter (e.g., using the user profile). However, the results provided by this API are not complete sets of all tweets, but they are prefiltered by Twitter.[13]

As the amount of data collected in this step is large, a prefiltering of the information base can be applied. Thus, all incoming data is filtered according to certain conditions such as the presence of certain keywords as shown in [19]. As a result of this step, a large amount of unstructured information is collected, which needs to be further processed.

4.2.2. Automatic preprocessing

As the information obtained in the previous step is usually very short and contains noise, applying automatic processing steps such as machine learning is difficult. Thus, in the second step of the framework, several automatic preprocessing steps are conducted. First, the unstructured information base is converted to a structured information base using Natural Language Processing. Second, named entities and temporal expressions are identified by various means to be used in subsequent steps.

Textual preprocessing: First, we remove all re-tweets as these are just duplicates of other tweets and do not provide additional information. Second, @-mentions of Twitter users are removed from the tweet message as we want to prevent overfitting towards certain user tokens. Before further processing is applied, the text is converted to Unicode, as some tweets contain non-Unicode characters. Third, abbreviations are resolved using a dictionary of abbreviations based on the data provided by the Internet Slang Dictionary&Translator.[14] Then, we identify and replace URLs with a common token "URL". As a next step, stopwords are removed. This is important as very frequent words have limited influence when it comes to classifying tweets due to their relative frequency. Based on the resulting text, we conduct tokenization. Thus, the text is divided into discrete words (tokens) based on different delimiters such as white spaces. Every token is then analyzed and non-alphanumeric characters are removed or replaced. Also, lemmatization is applied to normalize all tokens. Additionally to the common NLP processing steps, we identify and replace location mentions such as "Seattle" with a common token to allow semantic abstraction. For this, we use the approach we presented in [19] to detect named entities referring to locations (so-called location mentions) in tweets in order to replace them with two tokens "LOC" and "PLACE".

Extracting temporal information: As a second preprocessing step, the temporal dimension for each information item is derived automatically to infer the point in time of an event mentioned in a tweet. For example, the tweet shown in Listing 2 contains the temporal expression "friday afternoon" referring to the point in time when an accident occurred.

For identifying temporal expressions in tweets, we decided to adapt the HeidelTime [22] framework. The framework has been chosen because the system showed good performance on various datasets [26]. We extended the standard HeidelTime tagging functionality to mark up temporal expressions such as dates and times[15] with two annotations: "DATE" and "TIME". As a result, the temporal expression in the example tweet is replaced with our annotation.

The annotated temporal expressions are also used to provide an estimation of the point in time when an event mentioned in a tweet occurred. This is important as using the creation date of a tweet is not always correct as people also report on incidents that occurred in the past. For estimating this point in time, we use the creation date of a tweet as the base for our estimations. Using the extension, all temporal expressions are extracted and combined with the creation date to calculate the date when the event could have occurred. Though our approach takes different time zones into account, we do not use the actual geographic location of a user to add an additional offset to our calculation. The result of estimating the point in time is finally returned in a machine-readable format.

Listing 2. Replaced temporal expression in example tweet.

```
RT: @People  Onoe [friday afternoon] in heavy
    traffic, car crash on I-90, right lane
    closed

RT: @People  Onoe [@TIME] in heavy traffic,
    car crash on I-90, right lane closed
```

[12] https://dev.twitter.com/docs/api/1.1/get/search/tweets
[13] https://dev.twitter.com/docs/faq
[14] http://www.noslang.com
[15] Durations are not used as they are not valuable for detecting the time when an incident occurred.

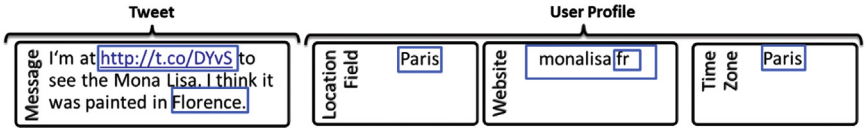

Fig. 3. Example for spatial indicators in tweets and user profiles.

Extracting spatial information: As a third step of the automatic preprocessing, we deal with the problem of how to infer the spatial dimension of a tweet. To this end, we extended our geolocalization approach presented in [23] to allow street-level geolocalization of tweets. A brief example of the extracted spatial information is shown in Fig. 3. Geolocalization of tweets without attached geoposition is highly important because only a limited number of about 1–2% of all tweets are explicitly geotagged. The applied geolocalization approach uses an estimation of the city and country where a tweet was sent from and additionally considers location mentions extracted from the tweet message. The process works as follows. As a first step, we identify toponyms, i.e., the mentions of locations, in the tweet message. Based on this information, we create triples of combinations of consecutive words (word-n-grams) to determine likely location names. We decided to use triples, as most street names can be represented using word-3-grams. We then use geocoding APIs such as the MapQuest Nominatim API[16] to map each n-gram to a location in the corresponding city. This results in several sets of coordinate pairs for each n-gram. Based on these pairs, we create a polygon. As a last step, we remove redundant polygons as some n-grams refer to the same location. Once all polygons are determined, they are stacked one over the other. The highest area in the height profile is identified, and its polygon outline is determined as intersection of the contributing polygons. In this case, the polygon is used as estimation of the location where an event mentioned in a tweet has happened. This approach allows a street-level geolocalization of the message focus of tweets. In a preliminary evaluation we found that this approach is able to estimate the location of an event described in a tweet with a median distance of 250 m.

As a result of this step, unstructured data is prepared in a way that it can be used as structured data for applying machine learning. Furthermore, we inferred spatial and temporal information for each tweet.

4.2.3. Human-based classification

In this part of the framework, manual classification is applied to infer the thematic dimension of an incident. For this, we apply crowdsourcing to classify incoming information in a way that relevant information is identified, which afterward can be provided to the subsequent steps. In [25] we presented our approach of having the command staff asking questions in order to articulate a particular information need such as "Which tweet is incident related?" or "Is this tweet related to a fire incident?".

In this step, a set of information items is presented to a user base, i.e. crowdsourcing users or domain experts, which then analyzes the information and classifies it according to the relevancy for a specific question. Thus, user-generated content is classified according to predefined incident types. The result of the step is a dataset containing the information items that have been assigned a thematic dimension (i.e., an incident type).

4.2.4. Machine-based classification

As outlined before, crowdsourcing is limited when it comes to timely information retrieval on a large amount of user-generated content. Thus, automatic approaches for classifying social media data are a necessity. We use supervised learning for obtaining models which are able to infer the thematic dimension of a tweet. For training these models, the preclassified information provided in the *Human-Based Classification and Aggregation* phase is used. The trained models can be used to automatically infer the thematic dimension of an information item for a large amount of data.

For evaluating the models, the framework can be configured using simple JSON-based config files. With this config file the end-user is enabled to specify the *Pipeline type*, the *Feature sets*, the *Classification type*, as well as the *Statistical tests* to be performed.

Pipeline type: Our framework is able to support the following three pipeline types: Train/Test, Cross-Validation (CV), and Multiple Cross-Validation. The simplest pipeline available is the Train/Test pipeline, which uses the specified train and test datasets. Each dataset is then used for training and testing, respectively. The CV pipeline utilizes only one given labeled dataset to generate separate folds for training and testing. Finally, the CV pipeline may be repeated several times using the Multiple CV Pipeline. By this approach, experiments gain a higher validity as folds change for each run of the CV pipeline.

For running the different pipelines, we integrated support for different learning and classification algorithms as provided by the commonly used WEKA framework [27] and Mulan, an open-source library for multi-label classification based on Weka [28].

Feature sets: Every pipeline requires a feature set as specified for the machine-based learning and classification step. To identify the features that provide the best classification performance, our framework allows for configuring different feature sets that need to be evaluated. As the optimal selection of features is mostly not known in advance, all features that may be valuable can be specified and are automatically evaluated using the power set of all possible feature combinations. Furthermore, as some experiments require a common feature set to compare against, we allow for specifying a baseline feature set. Other approaches for feature subset selection and dimensionality reduction and can easily be implemented for the framework in the case of a respective demand.

[16] http://developer.mapquest.com/web/products/open/nominatim

Classification type: For performing the actual machine learning, our framework provides capabilities to perform single-label, multi-label, as well as regression learning. The corresponding approaches are freely configurable, i.e., the learning algorithm to use and the parameters that need to be specified can easily be set in the configuration files. Each pipeline will then handle all combinations of classifiers and feature spaces internally, resulting in one single evaluation per dataset. The combinations are evaluated in parallel by the framework in cases where this is possible in order to reduce the runtime of each single evaluation.

With the model generated in this step, we are able to classify a large amount of user-generated content with respect to the situational information shared.

4.2.5. Aggregation

Finally, based on the spatial, temporal, and thematic information derived, each individual information item can clearly be related to a real-world incident. Based on this, new incidents can be detected. Also, information related to the same incident is automatically clustered to provide a set of relevant information to a decision maker.

The design of our approach presented in [29] follows the assumption that every incident-related information is either related to a specific real-world event or not. Thus, we propose to cluster all instances based on the three dimensions that define an event: temporal and spatial extent as well as the event type. As a result, each instance is aggregated to a cluster. As we use the properties of real-world events, it is much easier to identify those tweets that might be helpful for training.

If a micropost lies within the spatial, temporal, and thematic extent of another micropost, then the new micropost is assumed to provide information about the same event. This assertion can be formalized as a triple of the form $\{event\_type, radius, time\}$. The spatial extent is a radius in meters drawn around the spatial location of the event. The temporal extent is a timespan in minutes calculated from the creation time of the initial event. The thematic extent is the type of an event. For example, for our approach we use the rule $\{Car\_Crash, 200\text{ m}, 20\text{ min}\}$, which asserts that each incoming micropost of the event type *Car Crash* is aggregated to a previously reported incident if it is of the same type, within a range of 200 m, and within a time of 20 min. The parameters that specify the rule were proposed by emergency management staff, but could be varied according to the individual needs. Clearly, altering the radius or the time will have a strong effect on the final clustering. However, as experts suggested to use the specified values we did not change them. Then, with the help of these three assertion types, a rule engine computes whether microposts are clustered as they describe the same event or not.

Microposts containing no thematic information are assigned the *unknown_event* type. Missing spatial information is replaced with a common spatial center, e.g., the center of the city for which the microposts are used. Missing temporal information is replaced with the creation date of the micropost. Thus, even with one or two missing dimensions, we are still able to build clusters.

Similar clustering techniques are used for the task of topic sensing, where the goal might be to detect particular events or trending topics in a flow of social messages, e.g. incoming results from the individual states in the US Elections [30]. Note, that we explicitly try to exploit the particularities of small-scale incidents, namely the characterization by type, location and time, in order to obtain more accurate results than by applying more open and explorative methods.

4.2.6. Supervision and refinement

For assessing the overall performance of our framework, several statistical tests can be conducted based on performance metrics such as precision or recall. For the statistical evaluation and reporting of results, we developed and used the Open Source framework STATSREP-ML [31].[17]

After each pipeline run, performance metrics are generated. Our framework provides a statistical evaluation module that processes all generated performance measures. The module covers both parametric and non-parametric tests and checks their assumptions where appropriate. These tests are not implemented directly in Java, but executed in R, the free software environment for statistical computing and graphics.[18] This makes it possible to rely on the wide range of validated statistical packages available on CRAN, the Comprehensive R Archive Network. The rJava framework[19] facilitates the communication between Java and R. The currently integrated tests satisfy the particular requirements of the machine learning domain, and were chosen in accordance to the state of the art literature [32–35]. R is also used for plotting diagrams to support the interpretation of the results.

Those indicators will also hint to refinement requirements. Refinement might be needed since the topics of interest and the terms and style of communication of content produced on social media platforms is subject to an ongoing transformation process. For instance, semi-supervised learning approaches might be applied to label new instances for training as shown in [36].

4.2.7. Presentation and usage

Finally, a large amount of previously unprocessed and unstructured data can now be represented as a structured information base. Hereby, it can be used for decision making. Furthermore, data derived in the preceding steps can be used as new input for the framework. For instance, new information is collected in the human-based classification step. As a result of this framework, a structured information base that enhances the situational awareness of a decision maker is created. The information can now be consumed and used for taking decisions. Furthermore, the resulting information can be fed again into the framework.

[17] The framework is available at http://cguckelsberger.github.io/statistical-evaluation-for-machine-learning

[18] http://www.r-project.org/

[19] http://rforge.net/rJava/

5. Application of multi-label classification for extracting incident-related information

Our preliminary study has shown that tweets contain a variety of incident-related information which could be potentially useful for emergency services in real life. Hence, the following survey focuses on events which relate to emergency situations and which require the attention of emergency services. More specifically, we try to detect incidents which involve fire, shootings, (car) crashs or injuries. In addition to the knowledge of occurrence and place of an relevant incident in a timely manner, the precise knowledge of the type of incident is a very valuable information to emergency services and for the preparation of an emergency operation.

However, the shortcomings of the chosen categorization scheme used in the preliminary study also became evident. For instance, 544 of 1200 tweets could not be assigned to one of the predefined classes *car crash*, *shooting* or *fire*. Hence they were categorized as *non-incident*, although this negation only refers to the three predefined incident types. This type of categories collecting all kinds of non-identifiable or non-matching cases can often be found in the literature and results from the requirement imposed by the multiclass classification that every object has to be associated to exactly one class. Furthermore, the multiclass setting does also not cover the case of multiple classes assigned to one event, as the following example tweet will show:

Listing 3. Example of a tweet whose event type assignment cannot be satisfactorily reflected by a multiclass but by a multi-label classification scheme.

```
1 killed, 1 injured in South Memphis crash on
    I-240: One person was killed Monday
    morning in a crash on Interstate...
```

Annotators assigned the class *crash* to this event, although they also deemed relevant to tag the tweet with the label "1 injured" (among others, see Section 3.1), which is certainly a valuable information. If the classification scheme was extended by this type of event, both the assignments to only *crash* or only *injury* would miss relevant information. The multi-label problem setting in contrast allows us to assign an object to an arbitrary number of classes. Hence, we propose instead to use multi-labeled data in order to train multi-label classifiers which are able to automatically classify incident-related tweets into incident types [6].

Moreover, considering tweets as multi-label data adds one interesting aspect which could be exploited in order to improve the classification performance, namely the correlation between incident types. For instance, as we will also see further on from the collected data, it is very likely that there are injured people when there was a shooting. Thus, certainty about a *shooting* event could help learning algorithms to also accurately predict the *injury* incident type. We will investigate the used approaches particularly under the aspect whether it is possible to exploit this type of incident type dependencies.

The remainder of this paper investigates the applicability of our framework for the analysis of thematic properties of incident-related tweets. This section introduces multi-label classification of tweets as a showcase for the developed pipeline and it extends an earlier work of the authors with an expanded analysis [6]. This is the first work known to the authors on multi-label classification of microblogs and particularly on the exploitation of label dependencies on tweets (see also Section 6).

In the following, we introduce multi-label classification and the application in our framework. We also describe the collected data, the used methodology and finally the obtained empirical results.

5.1. Multi-label classification approaches

Multi-label classification refers to the task of learning a function that maps instances $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,a}) \in \mathcal{X} \subseteq \mathbb{R}^a$ to label subsets or label vectors $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n}) \in \{0,1\}^n$, where $\mathcal{L} = \{\lambda_1, \ldots, \lambda_n\}$, $n = |\mathcal{L}|$ is a finite set of predefined labels and where each label attribute y_i corresponds to the absence (0) or presence (1) of label λ_i. Thus, in contrast to multi-class classification, alternatives are not assumed to be mutually exclusive, such that multiple labels may be associated with a single instance.

This makes multi-label data particularly interesting from the learning perspective, since, in contrast to binary or multi-class classification, label dependencies and interconnections present in the data can be detected and exploited in order to obtain additional useful information or just better classification performance. Some examples for multi-labeled tweets were already shown throughout the paper (see Sections 1, 3.2.2, 5). As we will see further on, around 11.6% of the total amount of tweets and 22.7% of the incident-related tweets in the collected multi-labeled data were assigned to more than one category (cf. Table 4). Thus, we believe that it is not unusual to encounter tweets with several possible labels, so that in our opinion the view of microblogs as multi-labeled data seems more natural, more realistic, and more general. Nonetheless, previous work usually focuses on the multi-class labeling of tweets and this is the first work known to the authors which tries to exploit the multi-label topic characteristics of tweets.

In the following, we will describe commonly used approaches for multi-label classification: Binary Relevance (BR), Label Powerset (LP), and Classifier Chains (CC). All described techniques are based on the decomposition or transformation of the original multi-label problem into single-label binary problems, as most multi-label techniques do [37]. An illustration of these techniques is presented in Fig. 4. This has the advantage that we can use state-of-the-art text classification algorithms for learning the binary problems such as support vector machines [38,39]. We will also have a closer look at each classification approach with respect to taking dependencies between labels into account. Two of the used approaches are specifically tailored in order to cope with such dependencies.

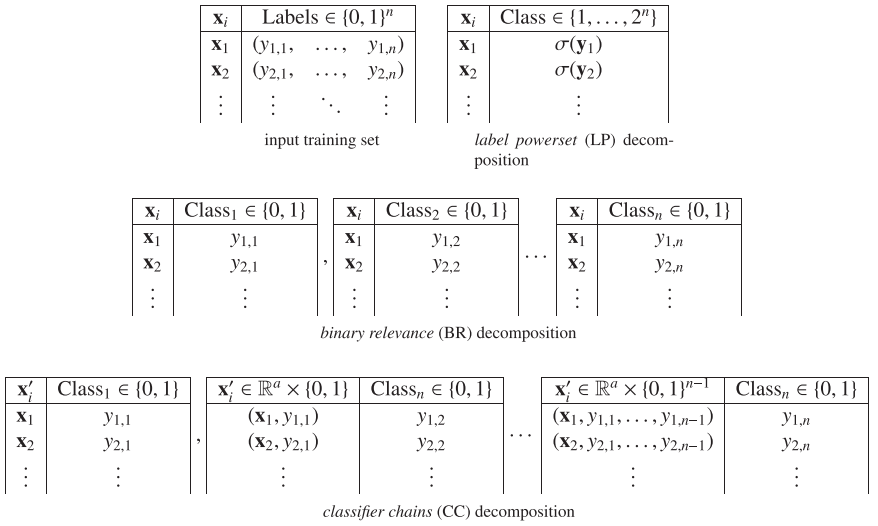

Fig. 4. Decomposition of multi-label training sets into multiclass (LP) or binary (BR, CC) problems. $\mathbf{x}_i'$ denotes the augmented instance. During prediction, $y_{i,1}, y_{i,2}, \ldots$ in the extended input space is replaced by the predictions by $h_1^{CC}, h_2^{CC}, \ldots$ (see text).

Table 4
Distribution of the 10 label combinations occurring in the 2000 tweets of the dataset.

Label combination	Number of tweets
{}	971
{Fire}	313
{Shooting}	184
{Crash}	268
{Injury}	32
{Crash, Fire}	2
{Injury, Crash}	47
{Injury, Shooting}	149
{Injury, Fire}	33
{Injury, Fire, Crash}	1

5.1.1. Binary relevance

The most common approach for multi-label classification is to use an ensemble of binary classifiers, where each classifier predicts if an instance belongs to one specific class or not. The union of all classes that were predicted is taken as the multi-label output. This approach is comparable to classical one-against-all for a multi-class problem. Formally, we convert a training example pair $(\mathbf{x}_i, \mathbf{y}_i)$ into n separate pairs $(\mathbf{x}_i, y_{i,j})$, $j = 1 \ldots n$, one for each of the n base classifiers h_j. The predicted labels $\hat{y}_j$ for a test instance $\mathbf{x}$ are then the result of $h_j(\mathbf{x}) \in \{0, 1\}$.

This method is fast and simple, however, it is not able to take label dependencies into account since each base classifier is trained independently from the other classifiers. As was recently stated by Dembczynski et. al [40], this is not necessarily a disadvantage if the objective is to obtain good label-wise predictions, such as measured by the Hamming loss (cf. Section 5.2). Therefore, BR serves as a fairly good performing baseline for our experiments.

5.1.2. Label powerset

The basic idea of this algorithm is to transform multi-label problems into a multi-class classification problem by considering each member of the powerset of labels in the training set as a single class. Hence, each training example is converted into $(\mathbf{x}_i, \sigma(\mathbf{y}_i))$ with σ, σ^{-1} denoting a bijective function that maps between the label powerset of $\mathcal{L}$ and a set of 2^n meta-classes. The classifier h^{LP} is trained e.g. with one-against-all (like in our setting), and the prediction for $\mathbf{x}$ is obtained with $\sigma^{-1}\left(h^{LP}(\mathbf{x})\right)$.

LP takes label dependencies into account to some extent, as each distinct occurrence of a label pattern is treated as a new class. It is hence able to model the joint label distribution, but not explicitly and directly specific dependencies (correlations, implications, etc.) between labels. As a consequence, LP is tailored towards predicting exactly the correct label combination. As it is pointed out in [40] and contrary to what one may believe at first, this stays usually in contrast to predicting correctly each label individually (BR), i.e. we usually have a trade-off between both objectives.

In addition to the obvious computational costs problem due to the exponential grow of meta-labels, the sparsity of some label combinations, especially with an increasing number of labels, often causes that some classes contain only few examples. This effect can also be observed in our data, cf. Table 4.

5.1.3. Classifier chains

As stated before, in a BR approach the correlation between labels is ignored. However, in our dataset we often encounter co-occurrences and inter-dependencies of tweets which could potentially be exploited for classification. For instance, it is very likely in our dataset that injured people are mentioned when also any incident type is mentioned (200 of 967 cases). On the other hand, it seems almost a matter of course that there was an incident

if there is an injured person. Although this only happens in 200 out of 232 cases in our data we consider it relevant for larger datasets. The classifier chains approach (CC) of Read et al. [41] is able to directly capture such dependencies and has therefore become very popular recently.

The idea of this approach is to construct a chain of n binary classifiers h_j^{CC}, for which (in contrast to BR) each binary base classifier h_j^{CC} depends on the predictions of the previous classifiers $h_1^{CC} \ldots h_{j-1}^{CC}$. More specifically, we extent the feature space of the training instances for the base classifier h_j^{CC} to $((x_{i,1} \ldots x_{i,a}, y_{i,1} \ldots y_{i,j-1}), y_{i,j})$. Since the true labels y_i are not known during prediction, CC uses the predictions of the preceding base classifiers instead. Hence, the unknown y_j are replaced by the predictions $\hat{y}_j = h_j^{CC}(\mathbf{x}, \hat{y}_1 \ldots \hat{y}_{j-1})$.

This shows up one problematic aspect of this approach, namely the order of the classifiers in the chain. Depending on the ordering, CC can only capture one direction of dependency between two labels. More specifically, CC can only capture the dependencies of y_i on $y_1, \ldots, y_{i-1}$, but there is no possibility to consider dependencies of y_i on $y_{i+1}, \ldots, y_n$. Recovering our example from the beginning, we can either learn the dependency of the label *incident* given *injury* or the other way around, but not both. In addition, the effect of error propagation caused by the chaining structure may also depend on the label permutation. We will evaluate the effect of choosing different orderings for our particular dataset later on in Section 5.3.

Furthermore, CC has advantages compared to LP. CC considers to predict the correct label-set, such as LP [40], but unlike LP, CC is able to predict label combinations which were not seen beforehand in the training data. In addition, the imbalance between positive and negative training examples is generally lower than for LP.

Note that the CC method naturally allows us to create ensembles of classifier chains (ECC) by simply training several classifiers using different label sequences. In addition to improving the prediction quality due to the ensemble effect, this method could also alleviate the problem of the direction of dependencies. However, in contrast to the base CC, the aggregation strategy used (majority vote for each label) is not tailored towards predicting the correct label-sets but each label independently. Thus, in this work we exhaustively explore all possible label sequences in order to analyze the potential of the base method itself.

5.2. Approach and framework

In this subsection, we describe the application of our framework for the multi-label classification task. For this, we specify the goal and the applied methods for the data extraction task.

- *Goal:* Our goal is the extraction of incident-related information from tweets. We focus on two important types of situational information identified in the pre-study of this paper, namely the precise incident type and injury reports. The incident type is the most important information as it helps differentiating noise from incident-related tweets. Also, injury reports provide very helpful information that is often provided. We do not cope with affected objects, as the variety is too large to allow manual labeling with respect to this information type. Also, road conditions and precise location is not a classification task, but can be detected using different extraction techniques as we showed in [23].
- *Technique:* We use multi-label classification techniques to identify information contained in tweets. We focus on three different incident types in order to identify incident-related tweets. These classes have been chosen because we identified them as the most common incident types using the Seattle Real Time Fire Calls dataset already mentioned in Section 3.1. We included also *injury* as an additional label. This results in four labels consisting of very common and distinct incident types and the injury label: Fire, Shooting, Crash, and Injury. We have implemented the following algorithms for multi-label classification: Binary Relevance, Classifier Chains and Label Powerset. Furthermore, a pure keyword based approach is realized as a simple and intuitive process that serves as a baseline for our evaluation.

In the following, we describe the usage of our framework to realize multi-label classification.

Collection and filtering: As ground truth data, we make use of the 7.5M tweets collected for the pre-study. Also for this evaluation, the dataset was reduced by conducting the keyword-filtering.

Automatic preprocessing: We apply the aforementioned preprocessing steps to convert the unstructured information base to a structured information base. This includes the necessary steps such as abbreviation resolution, tokenization, and lemmatization. As we are only interested in the thematic dimension, we do not apply temporal and spatial localization.

Human-based classification: Based on the filtered dataset provided by the collection and filtering step, we randomly selected 20,000 tweets. The selected tweets have been labeled manually by one researcher of our department with respect to their incident-relatedness and suitability of a multi-label classification problem. Out of these tweets, we randomly selected 2000 tweets for further re-labeling for our multi-label classification problem. Those tweets were manually examined by five researchers using an online survey. To assign the final coding, we differentiated between two types of agreement: (1) if four out of five coders agree on one label, only this label is assigned, and (2) if less than four coders agree on one label, all labels which at least two coders assumed as correct are assigned as possible labels and further verified in a group discussion.

Dataset characteristics: The final labeled dataset consists of 10 different label combinations (out of 16 possible ones). The distribution for every combination is outlined in Table 4. The distribution indicates that around 15% (232) of all tweets in our dataset have been labeled with multiple labels. Another observation is that almost 50% of the tweets do not have any label assigned, which is rather unusual compared to typically used and analyzed multi-

label datasets.[20] In addition, the label cardinality, i.e., the average number of labels assigned to an instance, is around 0.59, whereas common datasets have at least more than 1 assigned. On the other hand, this is mainly due to the low number of total labels, since the label density (the average percentage of labels which are true) is 15%, which is a relatively high value. From a multi-label learning perspective, this is an interesting property of this dataset since it is not clear how commonly used techniques will behave under these circumstances. For example, many algorithms ignore instances without any label given.

Machine-based classification: We performed our experiments with the 10-fold CV multi-label Classification Type module. We used two base learners for our evaluation as these are commonly used in the related work where they commonly showed a good or the best performance. First, we use the LibLinear implementation of support vector machines with linear kernel [42] as our base learner. We use the default settings, as we found that additional parameter optimization was not beneficial for improving the overall classification results. Second, we used the Weka implementation of Naive Bayes.

As we were interested which multi-label classification algorithm performs best, we evaluated all three algorithms as implemented in the Mulan framework, i.e., Binary Relevance, Classifier Chains, and Label Powerset. With our framework, we are able to evaluate the different base learners and multi-label algorithms easily, by specifying them in the corresponding configuration file.

Feature sets: For training the models, we evaluated several Feature Sets that were extracted from the tweets that are used for training a classifier. We conducted a comprehensive feature selection using our framework and the two base learners, analyzing the value of each feature for the overall classification performance. We compared word-n-grams, char-n-grams, TF-IDF [43] scores as well as syntactic features such as the number of explanation marks, question marks, and upper case characters. We found that the following features are the most beneficial for our classification problems:

- *Word 3-gram extraction*: We extract word three-grams from the tweet message. Each 3-gram is represented by two attributes. One attribute indicates the presence of the 3-gram and another attribute indicates the frequency of the 3-gram.
- *Sum of TF-IDF scores*: For every document we calculate the accumulated TF-IDF (term-frequency inverse-document-frequency) score based on the single TF-IDF scores of each term in the document. The rational behind this is to create a similarity score that is not as strict as traditional TF-IDF scores, but allows forming clusters of similar documents.
- *Syntactic features*: Along with the features directly extracted from a tweet, several syntactic features are expected to improve the performance of our approach. People might tend to use a lot of punctuations, such as explanation marks and question marks, or a lot of capitalized letters when they are reporting some incident. In this case, we extract the following features: the number of '!' and '?' in a tweet and the number of capitalized characters.
- *Spatial features*: As location mentions are replaced with a corresponding token, they appear as word unigrams in our model and can therefore be regarded as additional features.

More sophisticated approaches, which specifically consider multi-label data [44,45] or focus on the evaluation of combination of features [46,47], could also be considered in the future.

Aggregation and refinement: To find the best feature combinations and classification results for our problem, we evaluated according to several multi-label specific metrics, which are the following:

Exact match: Exact match is the percentage of the m test instances for which the label-sets were exactly correctly classified (with $[[z]]$ being indicator function returning 1 if z is true, otherwise 0)

$$ExactMatch(h) = \frac{1}{m}\sum_{i=1}^{m}\left[[y_i = h(\mathbf{x}_i)]\right] \quad (1)$$

Hamming loss: The instance-wise Hamming loss [48] is defined as the percentage of wrong or missed labels compared to the total number of labels in the dataset. In this case, it is taken into account that an incorrect label is predicted and that a relevant label is not predicted. As this is a loss function, the optimal value is zero.

Recall, precision and F1: We use micro-averaged precision and recall measures to evaluate our results, i.e., we compute a two-class confusion matrix for each label ($y_i = 1$ vs. $y_i = 0$) and eventually aggregate the results by (component-wise) summing up all n matrices into one global confusion matrix (cf. [37]). Recall and precision is computed based on this global matrix in the usual way, F1 denotes the unweighted harmonic mean between precision and recall. In Table 8, we also report recall, precision and F1 for each label using the label-wise confusion matrices.

Overall, we could easily adapt our framework to the specifics needed for applying multi-label classification on tweets in the domain of emergency management. By just exchanging feature sets and classification types in a single configuration file, we were able to evaluate a variety of different combinations. These evaluation results are presented in the following.

5.3. Evaluation of multi-label classification of tweets

In the following section, we provide the evaluation results for the presented multi-label classification approaches on our dataset. To underline the need for an automatic and intelligent extraction approach, we also present the results for a keyword-based approach as a simple way for conducting multi-label classification.

[20] We refer to the repository at http://mulan.sourceforge.net/datasets.html for an overview of the statistics of the commonly used benchmark datasets in multi-label classification.

Table 5
Overview of real-world incident types used for extraction of incident-related keywords and the number of extracted keywords for keyword-based classification approach.

	Fire	Shooting	Crash	Injury
	Fire in building Fire in single family res Automatic fire alarm resd Auto fire alarm	Assault w/Weap Assault w/Weapons aid	Motor vehicle accident Motor vehicle accident freeway Medic response freeway Car fire Car fire freeway	–
# of Keywords	148	36	73	23

5.3.1. Results for keyword-based filtering

As mentioned before, we use a keyword-based pre-filtering for selecting an initial set of tweets that is suitable for labeling. A first and simple approach for detecting incident-related tweets is to use these keywords for classification.

In Table 5, the real-world incident types from the Seattle Real Time Fire Calls dataset and the corresponding number of extracted keywords are shown. For the injury class, no specific type in the Seattle dataset could be found, thus, we extended the set with a manually created list of keywords and their direct hyponyms.

The results for classifying each individual class are shown in Table 6. The results indicate that precision as well as recall is rather low. Only for the fire class a high recall could be achieved.

Furthermore, if the keywords were used for applying multi-label classification, a precision of 32.22% and a recall of 64.90% would be achieved, which is a rather bad result. Also the exact match rate (28.45%) and Hamming loss (27.08%) would not be satisfactory. Thus, we conclude that with simple keyword-based filtering, multi-label classification cannot be done accurately.

5.3.2. Results for multi-label classification

In the following, we present the result of our multi-label classification problem. As a first step, we coped with the question if correlation between labels is taken into account and beneficial for the classification results. Thus, we evaluated all different label sequences using the classifier chains algorithm for our labels. The values for exact match for each sequence are shown in Fig. 5 (using an SVM as our base learner).

Table 6
Precision and recall for each individual label when applying keyword-based classification.

Measure	Shooting (%)	Fire (%)	Crash (%)	Injury (%)
Precision	31.59	54.12	15.04	63.64
Recall	68.77	95.99	49.37	37.40

The results indicate that the label sequence has indeed an influence on the classification performance. In our case, we get a difference of 1% between the best sequence Shooting, Crash, Fire, Injury and the worst Injury, Crash, Fire, Shooting. Also, we see that the Injury label is best used after incident labels have been classified – for the best cases even as one of the last labels in the sequence. It is also remarkable that classifying Shooting as first label followed up by either Crash or Fire is always a good option. This can be explained on the one hand by the generally good individual prediction performance for Shooting (cf., Table 8), hence leading to low error propagation, and on the other hand by the resulting label dependencies given the Shooting label is known: for instance, we can see from Table 4 that we can safely exclude annotating a tweet as referring to a Crash or Fire if there was a Shooting, since the labels Shooting and Crash, or Shooting and Fire, respectively, never occur together. This shows that our initial assumption, that taking correlation between labels into account can be beneficial for the automatic classification, is indeed true for the present dataset.

Based on the respective best (MAX) and the worst sequence (MIN), we compared CC to the multi-label approaches with the two different base learners. In Table 7 these evaluation results are shown. The first observation is that Naive Bayes is not adequate for classifying the tweets at hand, since though it achieves the best recall values using CC, this is in exchange of very low results on the remaining metrics and approaches. We will therefore focus on the results obtained by applying LibLinear as base learner. The results show that if there is the

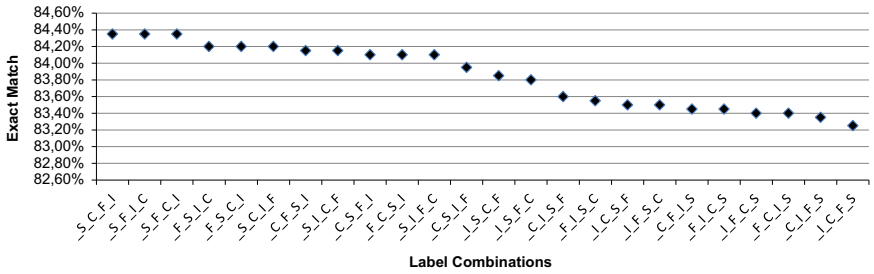

Fig. 5. Percentages of exact matches for all label combinations for the labels Fire (F), Shooting (S), Crash (C), and Injury (I).

Table 7
Results for the different multi-label approaches binary relevance (BR), label powerset (LP) and classifier chains (CC) and base learners obtained by cross-validation.

Measure	Naive Bayes				SVM			
	BR (%)	LP (%)	CC-MIN (%)	CC-MAX (%)	BR (%)	LP (%)	CC-MIN (%)	CC-MAX (%)
Exact Match	59.60	66.95	71.15	**72.45**	83.85	83.05	83.25	**84.35**
H-Loss	15.02	14.08	9.400	**9.175**	4.688	5.313	4.900	**4.588**
F1	52.19	55.37	72.90	**73.61**	83.55	81.53	82.80	**84.02**
Precision	52.40	55.34	66.84	**67.92**	**93.61**	90.28	92.75	93.46
Recall	51.98	55.39	79.63	**80.35**	75.44	74.35	74.72	**76.47**

Table 8
Precision and recall for each individual label.

Label	BR (SVM)		LP (SVM)		CC (SVM)	
	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)
Shooting	95.7	79.3	92.0	76.9	95.7	79.3
Fire	94.7	82.0	90.3	83.0	93.3	83.7
Crash	90.8	77.4	88.0	78.3	90.9	78.3
Injury	92.9	59.5	91.1	54.6	93.0	61.0

opportunity of pre-optimizing the ordering of the labels, e.g., by performing a cross-validation on the training data, then classifier chains is able to slightly outperform the other approaches, which is most likely because the label correlation is valuable. This is also reflected in the good performance with respect to exact match, where the worst CC even outperforms LP, which is particularly tailored towards matching the exact label combination. Note also that LP is a common approach used for circumventing the need for a multi-label classification by creating meta-classes, as already mentioned in the introduction. However, this approach is always inferior to the compared approaches, which demonstrates the need for more advanced techniques in this particular use case.

We can also observe that improving the prediction of the exact label combinations may come at the expense of reducing the performance on label-wise measures: Contrasted to BR, taking into account the predicted labels by CC generally leads to higher potential deteriorations (MIN) (compared to BR) than potential improvements (MAX) for Hamming loss, recall, precision and F1. However, for exact match the largest improvement (+0.50%) is similar to the largest deterioration (−0.60%) and hence this relationship is not that clear.

As a last evaluation step, we evaluated the accuracy of each approach for every individual label. This is important as we want to understand how well a classifier performs for each label. Table 8 depicts the accuracy of individual labels using SVM with the best label order.

The results show that the precision for individual labels is high with about 90–95% for each label, which is much better compared to the keyword-based classification. There are similar performance values for all considered approaches. Therefore, all approaches seem to be appropriate for classifying the individual labels. However, the recall drops significantly, depending on the label type. For instance, injuries often remain undetected. In this case, classifier chains show the best results for precision and recall. Note that the results for BR and CC on Shooting are the same, since the first classifier in the CC ordering is exactly trained like the corresponding BR classifier (cf. also Fig. 4). This is also shown along the chain: CC slightly reduces the good precision of BR in exchange of improved recall.

5.3.3. Discussion

Though the results show the advantage of multi-label classification, we wanted to understand the limitations of our approach. Thus, we first created a confusion matrix for the classifier chains approach with the best label order (see Table 9). The matrix shows that most misclassifications occur due to an assignment of instances to the "no incident" label combination {}. The other wrong classifications are mostly a result of not detecting the injury label or of predicting it wrongly.

The following misclassified tweets show examples for such wrongly classified instances:

Listing 4. Example of misclassified tweets with the respective predictions of the classifier. The notation is *real→predicted*.

```
Tacoma Fire Department replaces 3 fire
    engines with pickup trucks: TACOMA
    Cutbacks within the Tacoma Fire... http
    ://t.co/jPe2kuKG ({} -> {F})

This girl is on fire. This girl is on fire.
    She's walking on fire. This girl is on
    fire - Alicia Keys #deep ({} -> {S})

NeoMemphis News: Massive fire at factory in
    Ripley: Action News 5 is on the scene of
    a factory fire at ... http://t.co/
    brfnVbWp #memphis ({F} -> {F,I})
```

The examples show that certain words such as "fire" or digits in the message might lead to wrong classifications.

Table 9
Confusion matrix. The rows indicate the true label combinations and the columns the predicted ones.

	∅	F	C	F,C	I	F,I	C,I	F,C,I	S	F,S	I,S
∅	924	16	24	0	0	0	0	0	3	0	4
F	49	261	0	0	0	3	0	0	0	0	0
C	54	0	213	1	0	0	0	0	0	0	0
F,C	1	1	0	0	0	0	0	0	0	0	0
I	16	0	1	0	11	0	0	0	1	0	3
F,I	5	10	0	0	1	17	0	0	0	0	0
C,I	8	0	12	0	3	0	23	0	0	0	1
F,C,I	1	0	0	0	0	0	0	0	0	0	0
S	33	4	0	0	1	0	0	0	142	0	4
F,S	0	0	0	0	0	0	0	0	0	0	0
I,S	26	0	0	0	5	0	0	0	22	0	96

This could be avoided by adding additional features or with a larger training set.

We can also observe that the algorithm is able to detect instances with no label associated (∅) with an accuracy of 95.16%, whereas instances with one label are predicted with an accuracy of 78.67% and instances with more than one associated label with 58.62%.[21] Obviously, the accurate detection of incidents becomes harder the more incident types are referred to in a tweet. The reason could be the higher complexity of multi-label classification tasks in general, but also the reduced space available in micro-postings with increasing incident complexity in order to describe or express an incident type. This result hence additionally confirms us in our hypothesis that specialized learning and feature extraction methods are needed for classifying multi-labeled tweets.

5.4. Summary

In this section, we have shown how we applied our framework for multi-label classification for extracting incident-related information from tweets. We first showed the concrete implementations and configuration steps we made to conduct our evaluations. Next, we showed that a simple keyword-based classification approach is not suitable for multi-label classification, underlining the need for automatic extraction approaches. Third, we presented results of state-of-the-art multi-label classification approaches and we showed that these perform quite well for classifying incident-related tweets. Compared to current approaches for the classification of microblogs (see Section 6), which rely on assigning only one label to an instance, the results show that it is possible to infer important situational information, i.e. the incident type and the number of affected people, with only *one* classification step. The results also indicate that the label sequence has an influence on the classification performance, thus, this factor should be taken into account for future approaches. Loza et al. [49] describe a method which is able to automatically consider labels and label predictions in the most appropriate ordering. In addition, the resulting rule based classification models allow us to inspect (label and feature) dependencies and relationships in a natural way.

[21] Note however, that these numbers do not include partially correct predictions.

6. Related work

Following the insight of incident-related information contained in tweets, we review related work that copes with identifying information contained in tweets automatically. The related work in the domain focuses on the classification of user-generated content. Related approaches are differentiated with respect to the corpus used for incident type classification and the scale of the incident type addressed. Most notably the used classification techniques do not offer means to assign multiple labels to a tweet. Furthermore, approaches differ in the learning approach that is used and the number of classes that are detected. Also, different feature groups are used.

We also investigate the use of multi-label classification on short and unstructured texts such as tweets. Furthermore, we show that multi-label classification has not been applied in the domain of emergency management.

6.1. Approaches for extracting incident-related information from microblogs

An overview of related approaches is given in Tables 10 and 11.

Sakaki et al. [2] used an SVM classifier to detect earthquakes as a type of large-scale incident. The SVM was trained using three features extracted from tweets specifically referring to earthquakes: the number of words occurring in the tweets, statistical features (the number of words in a tweet and the position of keywords), and word context features (the words before and after the earthquake-related keyword). They used a dataset of 597 earthquake-related tweets and showed that their approach had a precision of 63.64% and recall of 87.50% for detecting earthquake-related tweets and differentiating them from non-related ones.

Becker et al. [50] presented a system for event detection. Based on cosine similarity of the TF-IDF scores of each tweet to a cluster, a preclustering of tweets was performed [59]. Afterward, each cluster is assigned a label whether it is an event cluster. For this, they use a combination of temporal (i.e., prominent terms), social (i.e., interaction such as retweets and replies), topical features (i.e., common terms), and Twitter-centric features (i.e., presence of hashtags). Based on this, an SVM classifier is trained. An evaluation was conducted on 374 manually annotated event clusters consisting of tweets from New York City and showed an F1 score of 83.7%.

The previous approaches focus on large-scale incidents. In contrast, other state-of-the-art approaches focus on the detection of small-scale incidents.

Hua et al. [51] presented STED, a system for small-scale event detection. Such as [50], they apply text classification for classifying preclustered tweets. Compared with other approaches, named entities are discarded before calculating TF-IDF scores, which are the only features used in their approach. An SVM was trained for a specific event type

Table 10
Overview of related approaches for incident type classification.

Approach	Corpus	Scale of incident		Classifier	# Classes	Multi-label
		Large	Small			
[2]	Tweets	x		SVM	2	
[50]	Cluster	x		SVM	2	
[51]	Cluster		x	SVM	2	
[52]	Tweets		x	NB, SVM	2	
[53]	Tweets		x	Keyw.	2	
[54]	Tweets		x	unknown	2	
[55]	Tweets		x	SVM	2	
[56]	Tweets		x	Keyw., SVM	2	
[57]	Cluster		x	JRip	2	
[58]	Tweets		x	SVM	6	
Our approach	Tweets		x	Keyw., NB, SVM	5	x

Table 11
Overview of related approaches for incident type classification with respect to the used feature groups (Named Entities = NEs).

Approach	N-Grams	NEs	URLs	TF-IDF	Twitter	Other
[2]	x					Contextual
[50]	(x)				x	Buzzy terms
[51]				x		
[52]	x	x	x			
[53]						
[54]		x			x	
[55]	x					
[56]	x				x	
[57]						Sentiment
[58]	x				x	
Our approach	x	x	(x)	x		

and applied on the clusters. The approach was tested on (an undefined number of) tweets collected in Latin America and shows a precision of 72% and recall of 74% for classifying the clusters.

Agarwal et al. [52] proposed an approach for classifying tweets related to a fire in a factory. As a first step, their system detects incident-related messages using a combination of a NB and an SVM classifier. As features, they use the number of occurrences of certain named entities such as locations, organizations, or persons that are extracted using the Stanford NER toolkit. Furthermore, the occurrence of numbers and URLs is used as a feature. Also, word occurrences remaining after stopword filtering are used. The approach was tested on 1400 tweets and shows that they are able to detect tweets related to factory fires with up to 80% accuracy. Furthermore, they showed that NB outperforms the SVM classifier. A possible reason for this might be the use of an untuned SVM.

Wanichayapong et al. [53] focused on extracting traffic information in tweets from Thailand. Their approach mainly relied on a dictionary-based approach. First, tweets are prefiltered using traffic-related keywords. Second, traffic-related keywords in combination with location-related keywords are used to classify traffic tweets. An evaluation of 1249 Twitter messages shows that this simple approach is able to give a precision of 91.39% and a recall of 87.53%.

Li et al. [54] introduced a system for the searching and visualization of tweets related to small-scale incidents based on keyword, spatial, and temporal filtering. Compared to other approaches, they iteratively refine a keyword-based search for retrieving a higher number of incident-related tweets. Based on these tweets a (not named) classifier is built upon text features and Twitter-specific features, such as hashtags, @-mentions, URLs, and the number of spatial and temporal mentions. They report an accuracy of 80% for detecting incident-related tweets, although they do not provide any information about their evaluation approach and the classifier used.

Carvalo et al. [55] evaluated an automatic classification of traffic-related tweets. Compared with other work, they conducted no initial labeling but used a set of tweets from official sources as ground truth data. An SVM classifier was trained based on this and (manually) evaluated on the rest of the tweets. As features, they used simple word unigrams, after stopword and punctuation removal. Furthermore, they showed that an SVM with linear kernels gives the same performance as other kernels. Finally, they achieved an F-measure of approximately 23%.

Power et al. [56] analyzed how to detect tweets related to fire incidents. In a preliminary evaluation, they showed that a simple keyword-based approach using the observed frequency of a word compared with historical frequency gave an accuracy of 48%. In a second evaluation, an SVM

with a linear kernel function was trained. They analyzed several feature combinations based on the number of words, user mention count, hashtag count, hyperlink count, unigram occurrences, and bigram occurrences. They found that a combination of both unigram occurrences and user mention count gave the highest performance with an F1 score of 83.1% on 794 tweets.

Walther and Kaisser [57] presented an approach for small-scale event detection. However, their goal was not to annotate a single tweet but to identify an event based on a set of tweets. As textual features, they used sentiment features, binary weighting of most frequent terms, and several dictionary-based feature groups. From these, they used a semantic dictionary, which contains a list of terms related to higher-level event categories such as "sport events". Their approach has been evaluated with 1000 manually labeled events (they do not provide the overall number of tweets) and evaluated using JRip. They achieved a precision of 85.8% and a recall of 85.6% for classifying the cluster of tweets.

Karimi et al. [58] tried to classify tweets according to six incident type classes. They relied on unigrams and bigrams as well as Twitter-specific features such as hashtags and @-mentions. The approach was evaluated on 5747 tweets and showed an accuracy of up to 90% when using 90% of the data as a training set. Precision and recall were not provided. However, compared with other approaches, they did not conduct cross validation but time-split evaluation. Thus, older data is used for training to deal with the dynamism of user-generated content. Furthermore, they showed that the best results could be achieved by using an SVM classifier.

Power et al. [60] introduce a fire monitoring tool which allows us to represent current fires in a map of e.g. a city. The information is obtained by filtering tweets by keywords and GPS information and a post-step filtering using an SVM.

Ritter et al. [61] analyze the problem of detecting computer security events, such as a distributed denial of service attacks on companies or account hijacking of persons or organizations. They focus on the case where only few tweets for a category of security attack are given or known in the beginning, referred to as seed instances. By using weekly supervised techniques such as expectation regularization or one-class SVM classifiers they were able to outperform systems which assume unlabeled examples to be negative examples. This approach could be useful in our framework in order to handle very rare or newly defined incident types.

6.2. Multi-label classification

The reviewed approaches assign a single label to each micropost and cluster the posts respectively, thus, only one single piece of information is detected. However, as we showed in the qualitative study, the identification of multiple situational information at once is much more desirable. Therefore, we consider multi-label classification. Techniques of multi-label classification have been applied to domains such as text categorization [62,63], music genre detection [64], or tag recommendation [65]. These application domains address long texts, images, or audio information.

Text is probably one of the oldest domains in which the demand for categorization appeared, particularly multi-label categorization [38], with the first multi-label dataset (*Reuters-21578*) used in machine learning research being from the year 1987 [66–68].

Moreover, data is easily accessible and processable as well as vastly available. Hence, text classification was also one of the first research fields for multi-label classification and continues to be the most represented one among the commonly available benchmark datasets.[22]

Applying multi-label learning on very short texts is a topic of open research. Only two previous respective examples are known to the authors: Sajnani et al. [69] and Daxenberger et al. [70]. Sajnani et al. provided a preliminary analysis of multi-label classification of Wikipedia barnstar texts. Barnstars can be awarded by Wikipedia authors and contain a short textual explanation why they have been awarded. In this case, labels for seven work domains have to be differentiated. The authors show which features can be extracted from short texts for multi-label classification and evaluate several multi-label classification approaches. Daxenberger et al. categorize individual article edits into non-exclusive classes such as *vandalism*, *paraphrase*, etc. Our previous work [6] was the first work known to the authors to consider microblogs for multi-label classification and to particularly analyze the dependencies between labels. Only very recently, Liu and Chen [71] used multi-label classification for the sentiment analysis of tweets about two major incidents in China. In their extensive experimental evaluation, the authors mainly focused on the selection of features and feature sets for sentiment classification.

In summary, although many related approaches cope with multi-class classification of short texts such as microblogs, multi-label classification is an open research issue. Especially for the domain of crisis management, no prior research on this topic exists.

7. Discussion

In this paper we have contributed to the field of information extraction from user-generated content to improve situational awareness during small-scale incidents. Our work focuses on three areas to be discussed in the following.

Analysis of small-scale incident reporting behavior: We provided the first analysis of information about small-scale incidents contained in microblogs. A quantitative and a qualitative study showed important first insights: (1) a variety of individuals are sharing information about small-scale incidents, information which is not necessarily available for decision makers, (2) incident-related tweets contain important situational information which could enrich the situational picture. Most importantly: precise location information is present, which enables decision

[22] Cf. http://mulan.sourceforge.net/datasets.html

makers to easily geolocalize the location of an incident. Also, affected objects such as buildings or cars and much more important information about potentially injured persons is shared. This information is especially valuable as it allows better planning of response measures. Overall, microblogs seem to be an important source of small-scale incident information.

Rapid prototyping framework for information extraction: We have introduced a framework which facilitates the development and assessment of extraction methods. The framework integrates crowdsourcing and machine learning to realize information extraction for thematic, location and temporal information. Based on a flexible configuration, different preprocessing steps, pipeline and classification types as well as different feature sets can easily be evaluated. The evaluation capabilities offer a variety of means to compare different approaches.

Application of multi-label classification for extracting incident-related information: We showed how to apply our framework for multi-label learning and applied it on social media data for classification of incident-related tweets. Furthermore, we were able to identify multiple labels with an exact match rate of 84.35%. This is an important finding since the automatic assignment of multiple labels provide important information about the situation at-hand which is not possible to obtain with the previously used multi-class classification approaches. Our study shows the need for multi-label classification techniques, and the effectiveness of existing state-of-the-art approaches. Furthermore, we have shown that the natural relation of labels, which represents for instance the relation between incidents and injuries in the real-world, could be used and exploited by classification approaches in order to obtain better results.

8. Conclusion

The framework introduced in this paper enables the rapid prototyping of methods for the extraction of information from incident-related tweets (see Section 4). Testing and validating a variety of different methods and configuration parameters enable the quick development of models to be directly used in the emergency management domain and grounded on a solid empirical foundation.

The focus on small-scale incidents was supported by our initial study of the information shared in social media (see Section 3). Specific consideration was given to the generation of labeled test data in the framework and to the statistical valid analysis of the results.

The showcase for multi-label classification in Section 5 gives a detailed overview of the steps and activities involved in the framework usage. A variety of different approach for multi-class classification (e.g. label powerset, classifier chains) is compared and the results are analyzed in detail. Overall we see that the application of Classifier Chains with an SVM brings the best results with an exact match of 84.35%.

The framework generates and analyzes models (e.g., for classification or clustering) which can be directly used in applications. Still, we want to enhance the support for out-of the box usage in real work scenarios additionally by adding capabilities for online learning, incremental and active learning.

In the future we plan to investigate further multi-label classification methods to extract situational information from social media. We aim to add costs to our classifications. For instances, not detecting incident labels should be heavily punished compared to misclassifying the incident type. Furthermore, we aim to improve the overall performance of our approach by taking into account different features and a larger training set. We plan to offer the framework to the public and extend it on an "as needed" basis.

Acknowledgments

This work has been partly funded by the German Federal Ministry for Education and Research (BMBF, 01 — S12054).

References

[1] M.R. Endsley, Design and evaluation for situation awareness enhancement, In: Proceedings of the Human Factors Society 32nd Annual Meeting, Aerospace Systems: Situation Awareness in Aircraft Systems, vol. 32, Human Factors and Ergonomics Society, New York, NY, USA, 1988, pp. 97–101.

[2] T. Sakaki, M. Okazaki, Earthquake shakes twitter users: real-time event detection by social sensors, In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, Raleigh, North Carolina, USA, ACM, New York, NY, USA, 2010, pp. 851–860.

[3] A. Signorini, A.M. Segre, P.M. Polgreen, The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza a h1n1 pandemic, PLoS ONE 6 (5) (2011).

[4] S. Vieweg, A. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: what twitter may contribute to situational awareness, In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI'10, Atlanta, Georgia, USA, ACM, New York, NY, USA, 2010, pp. 1079–1088.

[5] P.C. Shih, K. Han, J.M. Carroll, Community incident chatter: informing local incidents by aggregating local news and social media content, In: Proceedings of the International Conference on Information Systems for Crisis Response and Management, 2014, pp. 770–774.

[6] A. Schulz, E. Loza Mencía, T.T. Dang, B. Schmidt, Evaluating multi-label classification of incident-related tweets, In: Proceedings of the Making Sense of Microposts (Microposts2014) at WWW14, CEUR-WS.org, 2014.

[7] A.J. McMinn, Y. Moshfeghi, J.M. Jose, Building a large-scale corpus for evaluating event detection on twitter, In: Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management, CIKM '13, San Francisco, California, USA, ACM, New York, NY, USA, 2013, pp. 409–418.

[8] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, Topic detection and tracking pilot study: final report, In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194–218.

[9] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. Archibald, X. Liu, Learning approaches for detecting and tracking news events, IEEE Intell. Syst. 14 (4) (1999) 32–43.

[10] W. Blanchard, Select Emergency Management-related Terms and Definitions, Online (accessed: 01.04.2014), 2006.

[11] A.M. Kaplan, M. Haenlein, Users of the world, unite! the challenges and opportunities of social media, Bus. Horiz. 53 (1) (2010) 59–68.

[12] P. Gundecha, H. Liu, Mining social media: a brief introduction, INFORMS 9 (2012) 1–17.

[13] C.C. Aggarwal, C.X. Zhai, Mining Text Data, Springer-Verlag, New York, USA, 2012.

[14] I. Twitter, Twitter Reports Fourth Quarter and Fiscal Year 2013 results, Online (accessed: 04.03.2014), 2014.

[15] Securities, E. Commission, Form s-1—securities and exchange commission—twitter, inc., Online (accessed: 20.03.2014), 2013.
[16] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW '10, Raleigh, North Carolina, USA, ACM, New York, NY, USA, 2010, pp. 591–600.
[17] H.-C. Chang, A new perspective on twitter hashtag use: Diffusion of innovation theory, In: Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem, ASIS&T '10, vol. 47, American Society for Information Science, New York, USA, 2010, pp. 1–4.
[18] D.M. Romero, B. Meeder, J. Kleinberg, Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter, In: Proceedings of the 20th International Conference on World Wide Web, WWW '11, Hyderabad, India , ACM, New York, NY, USA, 2011, pp. 695–704.
[19] A. Schulz, P. Ristoski, H. Paulheim, I see a car crash: Real-time detection of small scale incidents in microblogs, In: The Semantic Web: ESWC 2013 Satellite Events, Lecture Notes in Computer Science, vol. 7955, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 22–33.
[20] J. Hurlock, M.L. Wilson, Searching twitter: Separating the tweet from the chaff., In: Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM 2011), 2011.
[21] M. De Choudhury, N. Diakopoulos, M. Naaman, Unfolding the event landscape on twitter: classification and exploration of user categories, In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW'12, ACM, New York, NY, USA, 2012, pp. 241–244.
[22] J. Strötgen, M. Gertz, Multilingual and cross-domain temporal tagging, Lang. Resour. Eval. 47 (2) (2012) 269–298.
[23] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, M. Mühlhäuser, A multi-indicator approach for geolocalization of tweets, In: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM'13, AAAI Press, Palo Alto, California, 2013.
[24] A. Tapia, K. Bajpai, B. Jansen, J. Yen, Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations? In: Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management, ISCRAM'11, ISCRAM, 2011.
[25] A. Schulz, H. Paulheim, F. Probst, Crisis information management in the web 3.0 age, In: Proceedings of the 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM'12, ISCRAM, 2012.
[26] M. Verhagen, R. Saurí, T. Caselli, J. Pustejovsky, Semeval-2010 task 13: Tempeval-2, In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 57–62.
[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The weka data mining software: an update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18, http://dx.doi.org/10.1145/1656274.1656278.
[28] G. Tsoumakas, E. Spyromitros Xioufis, J. Vilcek, I.P. Vlahavas, Mulan: a java library for multi-label learning, J. Mach. Learn. Res. 12 (2011) 2411–2414, software available at http://mulan.sourceforge.net/.
[29] A. Schulz, B. Schmidt, T. Strufe, Small-Scale Incident Detection based on Microposts, In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, ACM New York, NY, USA, 2015.
[30] L.M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, Y. Kompatsiaris, A. Jaimes, Sensing trending topics in twitter, IEEE Trans. Multimedia 15 (6) (2013) 1–15.
[31] C. Guckelsberger, A. Schulz, STATSREP-ML: Statistical Evaluation & Reporting Framework For Machine Learning Results, Technical Report, TU Darmstadt, tuprints. URL ⟨http://tuprints.ulb.tu-darmstadt.de/4294/⟩, 2015.
[32] M. Jakowicz, Shah Nathalie, Evaluating Learning Algorithms. A Classification Perspective, Cambridge University Press, Cambridge, 2011.
[33] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[34] F. Herrera, An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons, J. Mach. Learn. Res. 9 (2008) 2677–2694.
[35] M.J.A. Eugster, T. Hothorn, F. Leisch, Exploratory and Inferential Analysis of Benchmark Experiments, Technical Report 030, 2008.
[36] A. Schulz, F. Janssen, P. Ristoski, J. Fürnkranz, Event-based clustering for reducing labeling costs of event-related microposts, In: International AAAI Conference on Weblogs and Social Media, 2015.
[37] G. Tsoumakas, I. Katakis, I.P. Vlahavas, Mining multi-label data, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer, US, New York, Philadelphia, 2010, pp. 667–685, http://dx.doi.org/10.1007/978-0-387-09823-4_34.
[38] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002) 1–47.
[39] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, In: C. Nédellec, C. Rouveirol (Eds.), Proceedings of 10th European Conference on Machine Learning (ECML-98), Springer-Verlag, Chemnitz, Germany, 1998, pp. 137–142.
[40] K. Dembczynski, W. Waegeman, W. Cheng, E. Hllermeier, On label dependence and loss minimization in multi-label classification, Mach. Learn. 88 (1-2) (2012) 5–45, http://dx.doi.org/10.1007/s10994-012-5285-8.
[41] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (3) (2011) 333–359, http://dx.doi.org/10.1007/s10994-011-5256-5.
[42] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[43] C.D. Manning, P. Raghavan, H. Schtze, An Introduction to Information Retrieval, Cambridge University Press, 2009, 32 Avenue of the Americas, New York, NY, 10013-2473, USA, pp. 117–120.
[44] N. Spolaôr, G. Tsoumakas, Evaluating feature selection methods for multi-label text classication, In: Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering (BioASQ), a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013), 2013.
[45] M.-L. Zhang, J.M. Peña, V. Robles, Feature selection for multi-label naive Bayes classification, Inf. Sci. 179 (19) (2009) 3218–3229, http://dx.doi.org/10.1016/j.ins.2009.06.010.
[46] N. Spolaor, E. Alvares Cherman, M. Monard, H. Lee, Relieff for multi-label feature selection, In: 2013 Brazilian Conference on Intelligent Systems (BRACIS), 2013, pp. 6–11. http://dx.doi.org/10.1109/BRACIS.2013.10.
[47] J. Lee, D.-W. Kim, Feature selection for multi-label classification using multivariate mutual information, Pattern Recognit. Lett. 34 (3) (2013) 349–357.
[48] R.E. Schapire, Y. Singer, BoosTexter: a boosting-based system for text categorization, Mach. Learn. 39 (2/3) (2000) 135–168.
[49] E. Loza Mencía, F. Janssen, Stacking label features for learning multilabel rules, In: S. Džeroski, P. Panov, D. Kocev, L. Todorovski (Eds.), Proceedings of the 17th International Conference on Discovery Science, DS 2014, Bled, Slovenia, October 8–10, 2014, Lecture Notes in Computer Science, vol. 8777, Springer, Gewerbestrasse 11, CH-6330 Cham (ZG), Switzerland, 2014, pp. 192–203. http://dx.doi.org/10.1007/978-3-319-11812-3_17.
[50] H. Becker, M. Naaman, L. Gravano, Beyond trending topics: real-world event identification on twitter, In: Fifth International AAAI Conference on Weblogs and Social Media, ICWSM'11, AAAI Press, Menlo Park, California, 2011.
[51] T. Hua, F. Chen, L. Zhao, C.-T. Lu, N. Ramakrishnan, Sted: Semi-supervised targeted-interest event detection in twitter, In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, ACM, New York, NY, USA, 2013, pp. 1466–1469.
[52] P. Agarwal, R. Vaithiyanathan, S. Sharma, G. Shroff, Catching the long-tail: extracting local news events from twitter, In: Proceedings of the Sixth International Conference on Weblogs and Social Media, ICWSM 2012, AAAI Press, Palo Alto, California, 2012.
[53] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, P. Chaovalit, Social-based traffic information extraction and classification, In: Proceedings of the 11th International Conference on ITS Telecommunications, ITST'11, IEEE Computer Society, Washington, DC, 2011, pp. 107–112.
[54] R. Li, K.H. Lei, R. Khadiwala, K.C.-C. Chang, Tedas: a twitter-based event detection and analysis system, In: Proceedings of the 28th International Conference on Data Engineering, ICDE'12, IEEE Computer Society, Washington, DC, 2012, pp. 1273–1276.
[55] L.S. Carvalho, R. Rossetti, Real-time sensing of traffic information in twitter messages, In: Proceedings of the 4th Workshop on Artificial Transportation Systems and Simulation ATSS, ITSC'10, IEEE Computer Society, Washington, DC, 2010, pp. 19–22.
[56] D.R. Robert Power, Bella Robinson, Finding fires with twitter, In: Australasian Language Technology Association Workshop, Association for Computational Linguistics, 2013, pp. 80–89.
[57] M. Walther, M. Kaisser, Geo-spatial event detection in the twitter stream, In: Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 356–367.

[58] S. Karimi, J. Yin, C. Paris, Classifying microblogs for disasters, In: Proceedings of the 18th Australasian Document Computing Symposium, ADCS '13, ACM, New York, NY, USA, 2013, pp. 26–33.
[59] H. Becker, M. Naaman, L. Gravano, Beyond Trending Topics: Real-world Event Identification on Twitter, Technical Report, Columbia University, 2011.
[60] R. Power, B. Robinson, J. Colton, M. Cameron, A Case Study for Monitoring Fires with Twitter.
[61] A. Ritter, E. Wright, W. Casey, T. M. Mitchell, Weakly supervised extraction of computer security events from twitter, In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015, 2015, pp. 896–905. http://dx.doi.org/10.1145/2736277.2741083.
[62] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Mach. Learn. 39 (2–3) (2000) 135–168.
[63] A. McCallum, Multi-label text classification with a mixture model trained by EM, In: AAAI'99 Workshop on Text Learning, 1999, pp. 1–7.
[64] C. Sanden, J. Z. Zhang, Enhancing multi-label music genre classification through ensemble techniques, In: Proceedings of the 34th International ACM SIGIR conference on Research and development in Information Retrieval, ACM, New York, NY, USA, 2011, pp. 705–714.
[65] I. Katakis, G. Tsoumakas, I. P. Vlahavas, Multilabel text classification for automated tag suggestion, In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge, Antwerp, Belgium, 2008.
[66] P.J. Hayes, S.P. Weinstein, CONSTRUE/TIS: A system for content-based indexing of a database of news stories, In: A.T. Rappaport, R.G. Smith (Eds.), Proceedings of the 2nd Conference on Innovative Applications of Artificial Intelligence (IAAI-90), May 1–3, 1990, Washington, DC, USA, IAAI '90, AAAI Press, Chicago, IL, USA, 1991, pp. 49–64.
[67] D.D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task, In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992, pp. 37–50.
[68] D.D. Lewis, Reuters-21578 Text Categorization Test Collection Distribution 1.0, README file (V 1.3), May 2004.
[69] H. Sajnani, S. Javanmardi, D.W. McDonald, C.V. Lopes, Multi-label classification of short text: a study on wikipedia barnstars., In: Analyzing Microtext, 2011.
[70] J. Daxenberger, I. Gurevych, A corpus-based study of edit categories in featured and non-featured wikipedia articles, In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), 2012, pp. 711–726.
[71] S.M. Liu, J.-H. Chen, A multi-label classification based approach for sentiment classification, Expert Syst. Appl. 42 (3) (2015) 1083–1093. http://dx.doi.org/10.1016/j.eswa.2014.08.036.