How Social Media Text Analysis Can Inform Disaster Management

Sabine Gründer-Fahrer[1](✉), Antje Schlaf[1], and Sebastian Wustmann[2]

[1] Institute for Applied Informatics, Hainstraße 11, 04109 Leipzig, Germany
gruender@uni-leipzig.de, antje.schlaf@informatik.uni-leipzig.de
[2] CID GmbH, Gewerbepark Birkenhain 1, 63579 Freigericht, Germany
s.wustmann@cid.de

Abstract. Digitalization and the rise of social media have led disaster management to the insight that modern information technology will have to play a key role in dealing with a crisis. In this context, the paper introduces a NLP software for social media text analysis that has been developed in cooperation with disaster managers in the European project *Slandail*. The aim is to show how state-of-the-art techniques from text mining and information extraction can be applied to fulfil the requirements of the end-users. By way of example use cases the capacity of the approach will be demonstrated to make available social media as a valuable source of information for disaster management.

1 Introduction

The emerging field of *crisis informatics* (e.g., Palen et al. (2010)) is driven by the insight that, in the digital age, the ability to efficiently access and process huge amounts of unstructured data is crucial to situational awareness, knowledge building, and decision-making of organizations responsible for saving lives and property of people affected by a crisis. Disaster events like hurricane Katrina, 9/11, the Haiti earthquake, or the Central-European Flooding 2013 have demonstrated that there is urgent need to understand how information is shared during a crisis and to improve strategies and technologies for turning information into relevant insights and timely actions. Within crisis informatics, social media offer an interesting new opportunity for improvement of disaster management by providing fast, interactive communication channels and enabling participation of the public (Starbird and Palen 2011). However, social media data are *big data* in terms of volume, velocity, variety and veracity, and, accordingly, the demands and challenges with respect to the development of appropriate information technologies are especially high.

The paper presents possibilities for social media analysis that arise within a disaster management software that has been developed as part of the *Slandail* project (Slandail 2014), funded by the European community. *Slandail* deals with data in different modalities (texts and images) and languages (English, German and Italian) as well as with the integration of cross-lingual and cross-cultural

G. Rehm and T. Declerck (Eds.): GSCL 2017, LNAI 10713, pp. 199–207, 2018.
https://doi.org/10.1007/978-3-319-73706-5_17

aspects of crisis communications and has a special focus on issues related to the legal and ethical correctness of data use. End-users from Ireland, Germany, and Italy have been involved in the development of the system from design to testing.

The focus of the paper is on text analysis functionalities using NLP methods from the fields of text mining and information extraction that have been contributed by the two German partner organizations in cooperation with the disaster control authorities Landeskommando and Bezirksverbindungskommando in Saxony. The prototype of the software (Topic Analyst) has been implemented at CID and further developed in cooperation with InfAI during the course of the *Slandail* project. The software module is currently under consideration by German authorities for future use in German disaster management.

2 Approach

Computational methods from NLP offer a wide variety of possibilities for systematically and efficiently searching, filtering, sorting and analyzing huge amounts of data and thereby can enable end-users from disaster management to face the problem of information overload posed by social media. In this section, we describe how interests on the side of disaster management have guided our choice of the methods used and our way to apply them in context of our software.

2.1 End-User Requirements

Aspects: First of all, disaster managers want to find structured information on what is happening in a crisis situation (‘what?’). Equally important aspects are the place (‘where?’) and the time of the event (‘when?’). Further relevant information may concern the organizations involved in the event (‘who?’).

Perspectives: Beside the current state of the event with respect to all of these aspects, disaster managers are interested in current changes of state and in the development of the event over time in order to detect hot spots or trends.

Combinations of Filters: Taking into account the variety of possible circumstances and different roles disaster managers may have to play in context of a crisis, the analysis tool must allow for great flexibility in combining all aspects just mentioned (e.g., ‘how did a certain aspect of the situation develop at a certain location?’).

Granularity: Similarly, since disaster managers are interested in an overview as well as in special details of the situation, there has to be the possibility of zooming in and out and looking at the event with different levels of granularity.

Relevance: A special case of guiding the attention of the end-users is the filtering out of irrelevant or wrong information.

Usability: Finally, in context of an application in disaster management, efficiency and user-friendliness of software are of high importance.

2.2 Implementation of Requirements

Aspects: For the first aspect of the analysis ('what?'), we referred to topic model analysis on basis of the HDP-CRF algorithm (Teh and Jordan 2010). In an unsupervised setting, topic modelling reveals the latent thematic structure in huge collections of documents. Furthermore, we applied hashtag statistics and keyword extraction by comparison of term distributions between the target collection and a reference corpus (*differential analysis*). For keywords or hashtags, co-occurrences analysis can reveal relations between concepts or entities.

Regarding the second aspect ('where?'), we either referred to meta data information or conducted location extraction using a list of location markers from *OpenStreetMap* (OpenStreetMap 2016) together with rule-based and context-sensitive techniques. By means of related longitude latitude coordinates, locations can be projected on a map.

Temporal information ('when?'), is provided by social media meta data. Names of organizations ('who?') got extracted by an NER approach that combines machine learning, rule based and context-sensitive techniques.

Perspectives: To take into account the different possible temporal perspectives, we not only provided means to summarize but also aggregate measurements of the various aspects over time. Additionally, we enabled calculation of growth or shrink from one interval to the next for all aspects.

Combination of Filters: All aspects of analysis as well as all meta data can be used as filter criteria and can be applied separately or in combination to create different sub-collections of data as input for analysis in line with special interests.

Granularity: The software offers possibilities for zooming in and out of a situation within the dimension of each aspect. Beside this, it integrates the mentioned statistically based *distant reading* procedures for entire collections or sub-collections of text with possibilities of manual *close reading* of single documents.

Relevance: As a provisional indication of the relevance of a message, we used the number of shares or retweets it received. On the one hand, the fact that many people found a message relevant, may really prove its relevance, on the other hand, even if the shared or retweeted message was not really relevant or even wrong, it may gain relevance from the point of view of disaster management because many people read it.

Usability: Beside the performance of the software in real-time or near-real-time, its easy handling and the intuitive visualization of analysis results have been in focus of our work. The software is accessible by an interactive graphical web interface with filter panels, drag-and-drop functionality, clickable graphs and configurable dashboards. For analysis of data, there are available two main modules – monitoring (*dashboard*) and analysis (*browser*). While the dashboards in the monitoring module are supposed to give a continuous overview over some predefined fields of interest, the analysis module allows for specific ad hoc investigations and close-reading of documents.

3 Examples

In this section, we demonstrate main functionalities of our software by means of example. As test data sets we used Facebook and Twitter data that had been created during the Central European flooding in June 2013 in Germany and Austria. For the Facebook flood corpus, we collected data from public pages or groups containing the words ‘Hochwasser’ or ‘Fluthilfe’ in their names via the public API (about 36k messages). For the Twitter flood corpus (about 354k tweets), we retrieved the current version of the research corpus of the QuOIMA project (QuOIMA 2011), that had been collected from the API filtering by disaster-related hash tags as well as by names of manually chosen public accounts connected to disaster management and flood aid (ibid.).

The example use case we present will be built around the topic extraction functionality. The dashboard in Fig. 1 gives an overview of topics and topic proportions for the Facebook flood corpus for the entire period of the event.

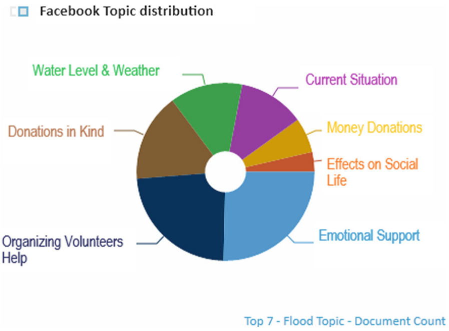


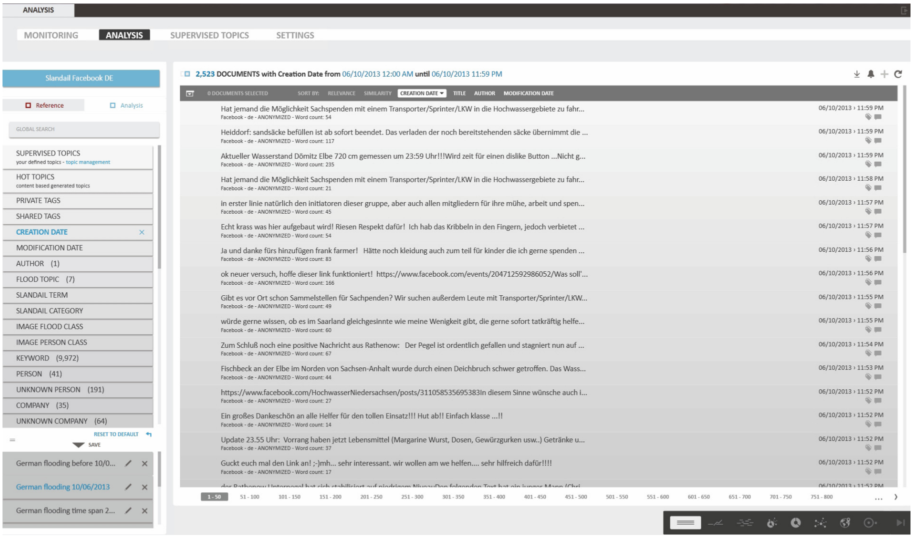
Fig. 1. Topics and topic proportions Facebook flood

By clicking on a name of a topic, it is possible to change to the analysis mode and to close-read or further analyze the messages belonging to this topic. The analysis view is shown in Fig. 2.

In the analysis mode, one could get an overview of the content of messages in a certain topic by showing typical topic words, for instance. Figure 3 includes typical words for the volunteering topic. By touching one of the words with the mouse, its co-occurring terms will get connected to it by edges to form a graph.

The dashboard in Fig. 4 changes temporal perspective and analyses the development of topics over the time of the event for the Twitter flood corpus. Again, clicking on the dots on the graph lines gives access to the messages showing the respective topic at the respective day for inspection or further analysis.

The dashboard in Fig. 5 gives an example of filtering for relevance of messages by number of their retweets. The peak around 20th June is connected to heavy rainfalls and thunderstorms that made alarm levels and subjective worries of the people rise anew but finally did not cause mayor new floodings. By only showing messages retweeted more than 6 times, a disaster manger searching for



2,523 DOCUMENTS with Creation Date from 06/10/2013 12:00 AM until 06/10/2013 11:59 PM

Hat jemand die Möglichkeit Sachspenden mit einem Transporter/Sprinter/LKW in die Hochwassergebiete zu fahr... 06/10/2013 › 11:59 PM
Facebook - de - ANONYMIZED - Word count: 54

Heiddorf: sandsäcke befüllen ist ab sofort beendet. Das verladen der noch bereitstehenden säcke übernimmt die ... 06/10/2013 › 11:59 PM
Facebook - de - ANONYMIZED - Word count: 117

Aktueller Wasserstand Dömitz Elbe 720 cm gemessen um 23:59 Uhr!!!Wird zeit für einen dislike Button ...Nicht g... 06/10/2013 › 11:59 PM
Facebook - de - ANONYMIZED - Word count: 235

Hat jemand die Möglichkeit Sachspenden mit einem Transporter/Sprinter/LKW in die Hochwassergebiete zu fahr... 06/10/2013 › 11:58 PM
Facebook - de - ANONYMIZED - Word count: 21

in erster linie natürlich den initiatoren dieser gruppe, aber auch allen mitgliedern für ihre mühe, arbeit und spen... 06/10/2013 › 11:57 PM
Facebook - de - ANONYMIZED - Word count: 45

Echt krass was hier aufgebaut wird! Riesen Respekt dafür! Ich hab das Kribbeln in den Fingern, jedoch verbietet ... 06/10/2013 › 11:57 PM
Facebook - de - ANONYMIZED - Word count: 54

Ja und danke fürs hinzufügen frank farmer! Hätte noch kleidung auch zum teil für kinder die ich gerne spenden ... 06/10/2013 › 11:56 PM
Facebook - de - ANONYMIZED - Word count: 83

ok neuer versuch, hoffe dieser link funktioniert! https://www.facebook.com/events/204712592986052/Was soll'... 06/10/2013 › 11:56 PM
Facebook - de - ANONYMIZED - Word count: 166

Gibt es vor Ort schon Sammelstellen für Sachpenden? Wir suchen außerdem Leute mit Transporter/Sprinter/LKW... 06/10/2013 › 11:55 PM
Facebook - de - ANONYMIZED - Word count: 49

würde gerne wissen, ob es im Saarland gleichgesinnte wie meine Wenigkeit gibt, die gerne sofort tatkräftig helfe... 06/10/2013 › 11:55 PM
Facebook - de - ANONYMIZED - Word count: 60

Zum Schluß noch eine positive Nachricht aus Rathenow: Der Pegel ist ordentlich gefallen und stagniert nun auf ... 06/10/2013 › 11:54 PM
Facebook - de - ANONYMIZED - Word count: 67

Fischbeck an der Elbe im Norden von Sachsen-Anhalt wurde durch einen Deichbruch schwer getroffen. Das Wass... 06/10/2013 › 11:53 PM
Facebook - de - ANONYMIZED - Word count: 44

https://www.facebook.com/HochwasserNiedersachsen/posts/311058535695383In diesem Sinne wünsche auch i... 06/10/2013 › 11:52 PM
Facebook - de - ANONYMIZED - Word count: 27

Ein großes Dankeschön an alle Helfer für den tollen Einsatz!!! Hut ab!! Einfach klasse ...!! 06/10/2013 › 11:52 PM
Facebook - de - ANONYMIZED - Word count: 14

Update 23.55 Uhr: Vorrang haben jetzt Lebensmittel (Margarine Wurst, Dosen, Gewürzgurken usw..) Getränke u... 06/10/2013 › 11:52 PM
Facebook - de - ANONYMIZED - Word count: 37

Guckt euch mal den Link an! ;-)mh... sehr interessant. wir wollen am we helfen.... sehr hilfreich dafür!!!! 06/10/2013 › 11:52 PM
Facebook - de - ANONYMIZED - Word count: 17

der Rathenow Unterpegel hat sich stabilisiert auf niedrigem NiveauDen folgenden Text hat ein junger Mann (Chri... 06/10/2013 › 11:52 PM

Fig. 2. Analysis modus with document view and filter panel

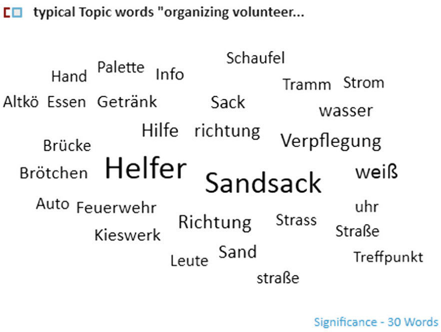


Fig. 3. Typical words for topic 'organizing volunteer's help'

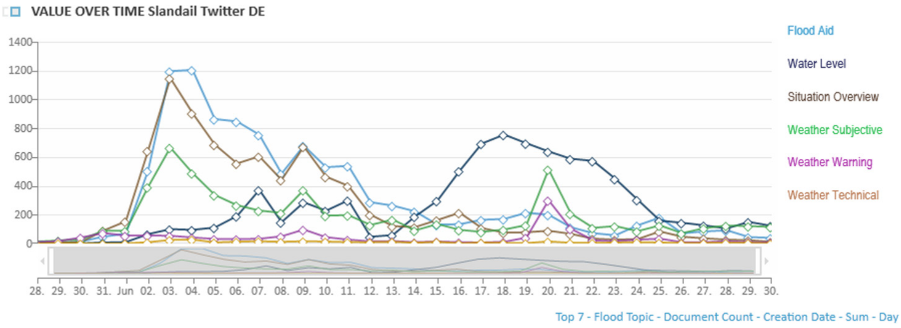


Fig. 4. Development of topics over time for Twitter

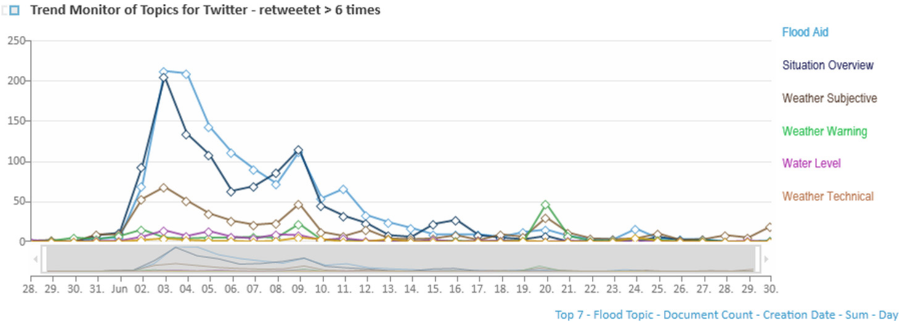


Fig. 5. Development of topics over time for Twitter tweets retweeted >6 times

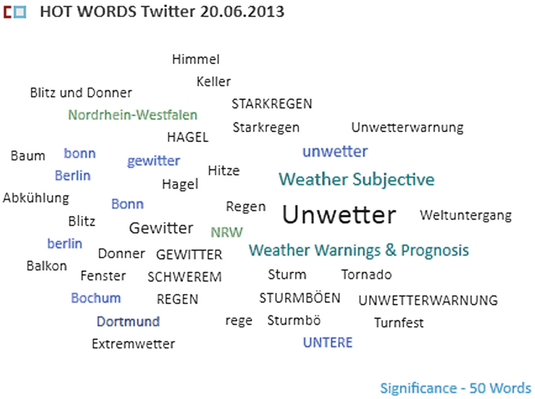


Fig. 6. Significant words and topics at a day in Twitter

practically relevant information can filter out precaution and pure worries as 'noise'.

To get an idea of what is happening at the 20th June to cause these emotional reactions, one could either change to the close-reading modus for the relevant topics or extract a situational overview by the help of differential analysis to show keywords that were significantly more frequent at this day (target corpus) than they had been before (reference corpus), see Fig. 6.

The third possible temporal perspective focuses on changes in topic prominence at one day or interval in comparison with the day or interval before. Figure 7 illustrates a possible outcome of analysis for Facebook.

On basis of this insight into hot topics, one could, again, ask further questions. For instance, one could be interested in the most popular organizations involved in donations in kind, or want to find out locations where many volunteering activities are organized. Figures 8 and 9 reveal the results of the respective analyses.

While our examples were developing from the point of view of the topic aspect ('what?'), each other aspect can equally well serve as a starting point

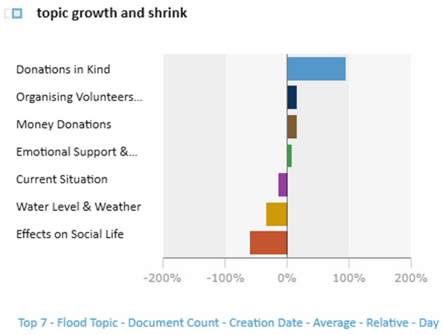


Fig. 7. Change of topic prominence for Facebook for a day

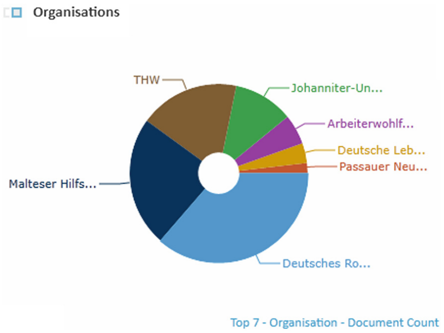


Fig. 8. Prominent organizations in topic ‘donations’ at a day

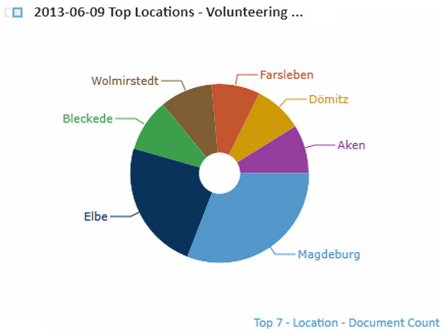


Fig. 9. Prominent locations in topic ‘volunteering’ at a day

for filtering and further analysis. For instance, the location aspect (‘where?’) is often of high interest for disaster managers. As before, the temporal perspective can either reveal an overview (as in Fig. 9), significant changes or the temporal development of the aspect. In Fig. 10, geographical hot spots are identified, while Fig. 11 shows the geographical unfolding of the flood event over its lifetime.

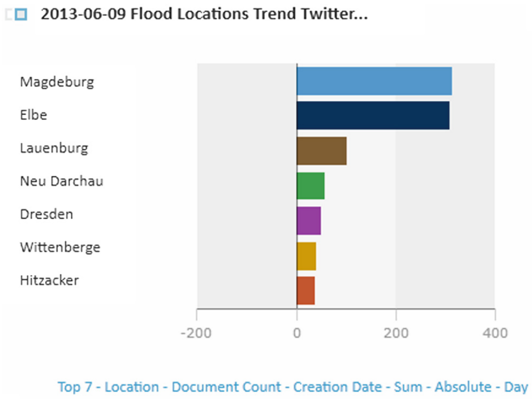


Fig. 10. Change of location prominence for Twitter for a day

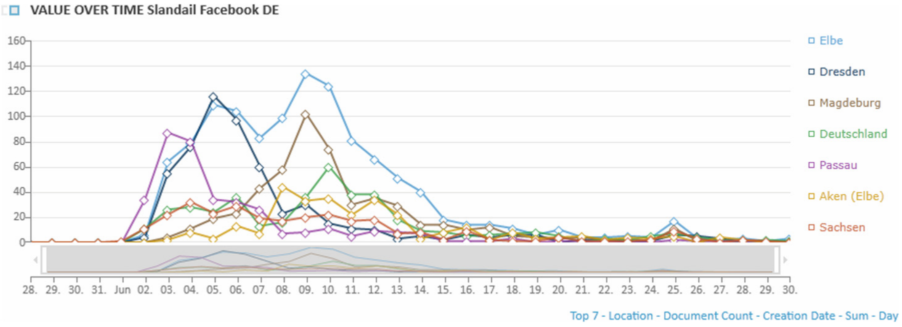


Fig. 11. Prominent locations over time for Facebook

4 Conclusion

In the work presented in this paper we were bridging between the fields of informatics and disaster management in order to design and create a social media text analysis software suitable for information gathering and knowledge acquisition in context of a crisis. Our first focus was on showing which methods can be chosen and how they can be applied in a system as to meet the special interests and requirements on the end-user side. Following this, various example use cases illustrated our general approach and demonstrated the capacity of our software to extract information useful for disaster management from huge collections of social media data. An approach along these lines can help to meet the challenges and make use of the opportunities that digitalization and the rise of social media have brought to disaster management.

Acknowledgments. The research leading to these results has received funding from the European Union's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL).

References

OpenStreetMap: OpenStreetMap - Deutschland (2016). https://www.openstreetmap.de. Accessed 08 June 2017

Palen, L., Anderson, K., Mark, G., Martin, J., Sicker, D., Palmer, M., Grunwald, D.: A vision for technology-mediated support for public participation & assistance in mass emergencies and disasters. In: Proceedings of the ACM-BCS Visions of Computer Science (2010)

QuOIMA: QuOIMA Open Source Integrated Multimedia Analysis (2011). www.kiras.at/projects. Accessed 08 June 2017

Slandail: Slandail - Security System for language and image analysis (2014). http://slandail.eu. Accessed 08 June 2017

Starbird, K., Palen, L.: "Voluntweeters": self-organizing by digital volunteers in times of crisis. In: Proceedings of the ACM-BCS Visions of Computer Science (2011)

Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: Hjort, N.L., et al. (eds.) Bayesian Nonparametrics, pp. 114–133. Cambridge University Press, Cambridge (2010)