Geocoding of Crisis Related Social Media Messages for Assessing Voluntary Help Efforts as a Contribution to Situational Awareness

Sebastian Drost
Bochum University of Applied Sciences, Department of Geodesy
Lennershofstraße 140, 44801 Bochum, Germany
sebastian.drost@hs-bochum.de

Andreas Wytzisk
Bochum University of Applied Sciences, Department of Geodesy
Lennershofstraße 140, 44801 Bochum, Germany
andreas.wytzisk@hs-bochum.de

Albert Remke
52°North Initiative for Geospatial Open Source Software GmbH
Martin-Luther-King-Weg 24, 48155 Münster, Germany
a.remke@52north.org

Abstract

Recent crisis events in Germany have shown that voluntary help efforts increase significantly. Due to this fact, it will be more and more a sophisticated task for disaster control authorities to manage the activities of volunteers at operation site. Since most of these activities are coordinated in social networks like Facebook, those platforms provide useful information that contribute to situational awareness. Hence, this work presents techniques for processing crisis related messages from social media for assessing voluntary help efforts. We demonstrate how to extract relevant messages that mention help efforts from Facebook by the use of machine learning based text classification. Further, we propose a Natural Language Processing based approach for extracting toponyms from social media messages and introduce a linguistic approach for disambiguating those toponyms.

Keywords: Social Media, Text Classification, Machine Learning, Message Geocoding, Situational Awareness, Disaster Management

1 Introduction

During a period from May to June in 2013 heavy rainfalls causes extensive floodings in several German regions. For most of these regions disaster warnings were declared by the German disaster control authorities (BMI, 2013). The different emergency services were supported by many volunteers. Most of these came together in social networks like Facebook to share help requests or to coordinate relief activities (Kaufhold and Reuter, 2016). As a consequence, emergency responders were confronted with a large amount of people that transferred their virtual voluntary activities into the real world. However, in various places there have been some complications which were caused by the increased number of volunteers.

Due to these experiences, the research project KUBAS (Coordination of voluntary helpers to overcome disaster situations) has started in 2016 to address several problems that appeared during the 2013 flooding. The goal of the project is to provide an IT-infrastructure for integrating voluntary help efforts into organized control structures of emergency services (e.g. command systems). Furthermore, new ways of communicating with volunteers will be analysed (Martin-Luther-Universität Halle-Wittenberg, 2016). All of the research in this project aims to find an approach for a better coordination of voluntary activities. With regard to this objective, the present work explores techniques to extract useful information from social media that contribute to situational awareness within the scope of KUBAS by assessing voluntary help efforts. We want to demonstrate how to filter relevant messages out of public social network groups and how to extract location information from those messages by the use of Natural Language Processing tools. Finally, we introduce an approach that utilizes a free geographical database in order to disambiguate toponyms for an automatically geocoding of the social media messages.

2 Related Work

In the past, several studies have analysed the use of social media in crisis situation and explored what kind of information may contribute to situational awareness. The work of Starbird et al. (2010) and Vieweg et al. (2010) about the use of Twitter during natural hazards in USA and Canada in 2009 have shown up the potential of user generated information in crisis situation very early. Thus, a lot of reports from eyewitnesses and help requests from affected people were published. A great part of those messages contained situational updates and geo-location information. Similar observations could be made by Kaufhold and Reuter (2016) in their studies about the use of social media during the German 2013 flooding.

There has also been some work on extracting useful information for crisis response from the vast number of messages by the use of machine learning techniques. Verma et al. (2011) analysed how to handle massive datasets from Twitter by the use of Natural Language Processing and built a classifier for identifying tweets that contribute to situational awareness. Imran et al. (2013) worked on the extraction of valuable "information nuggets" for disaster response. They utilized machine learning techniques for classifying twitter messages into crisis related sub-type categories and extracted structured information like location or time references.

Furthermore, some researches focused on geocoding and mapping strategies for social media messages during emergencies. Dhavase and Bagade (2014) used geoparsing to extract location information from tweets for crime and disaster events in India. The goal of Middleton et al. (2014)

was to develop a web-platform for mapping crisis events in real-time. Their work includes both a geoparsing component and a geocoding component.

3 Filtering Crisis Related Social Media Messages

The task of manually monitoring messages that appear on a social media platform to identify the relevant ones is nearly impossible. Regarding to this problem, we applied text classification based on supervised machine learning for the message extraction. Such a method enables automatically classifying messages into predefined categories. To find an appropriate classification method we collected some crisis related messages, pre-processed those and evaluated different classifiers with a standardized statistical method. Since most of the voluntary coordination activities took place in groups on Facebook, we considered Facebook posts related to the German flooding in 2013 for our studies.

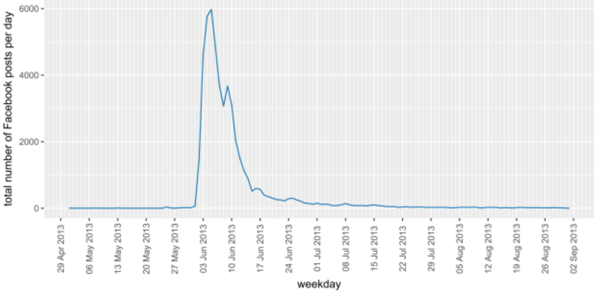
3.1 Prefiltering

For fetching posts from Facebook, we used the *Facebook Graph API*. This is a HTTP-based API that provides access to the platform's content (Facebook, 2017). Due to the fact, that the Facebook Graph API is restricted to public posts but provides no search for those, we had to implement a custom approach for fetching possible relevant messages.

First, we determined public groups and pages that matched one of the keywords we predefined in a list (e.g. “flood”, “flood relief” or “flood helper”). As a result, we identified 225 public groups and 55 pages that were related to flooding events or relief activities.

For each of these groups and pages we extracted the published posts within a time interval from May 2013 to the end of August 2013, which covers the time the events took place, including a small buffer. Finally, we collected nearly 50.000 Facebook posts which were saved in a *MongoDB* database subsequently. We found that the publication frequency of the posts was highest between 1 June 2013 and 5 June 2013 (Figure 1). This is exact the time interval most of the affected regions were evacuated by the disaster control authorities. So, there was some kind of a correlation between the chronology of the events and the activities on social media.

Figure 1: Relevant Facebook post per day during the German flooding 2013



3.2 Preprocessing

In order to train a text classifier, we first created a training dataset and extracted text features. The training dataset was constructed by manually categorizing a small part of the Facebook posts we collected previously. We considered to take those posts that were published in the beginning of the events to assume “realistic” conditions.

Each post from the training dataset was labelled with one of the two classes “relief” and “non relief”, regarding to whether the message content mention help efforts or not.

To reduce the number of text features, we first conducted a filter pipeline on the raw content of the Facebook posts. Therefore, we used the machine learning based natural language processing toolkit *Apache OpenNLP*. First, we removed all non-words like numbers, emojis and URLs and normalized the text tokens by lowercasing all letters. Following, we applied enhanced NLP techniques. Stemming was used to reduce a word to its root form. We also removed stop words to decrease the number of words and to avoid the presence of high frequency terms.

After all, we used the “bag-of-words” approach for text feature representation. This treats each text document as a collection of words without any relative order (Manning, 2009). To generate numeric values for the unigram text features, we used different weighting schemes: binary weighting, term frequency (tf) and term frequency-inverse document frequency (tf-idf). Our intention was to analyse if one of these provides a better feature representation than the others.

3.3 Text Classification Evaluation

There are many supervised machine learning algorithms for text classification. For our studies we tested the widespread methods Decision Trees, Support Vector Machines, Naïve Bayes, Multinomial Naïve Bayes and K-Nearest Neighbors (Sebastiani, 2001). Each classifier was trained and tested several times by the use of the different numeric text features that were generated by the different weighting schemes. For this, we used the Java based open source machine learning library *WEKA* which includes several implementations of the most common machine learning algorithms (Franke et al., 2016).

To compare the classification results, we used the evaluation metrics precision, recall and F_1 score. For this, we defined true positives (TP) as posts that actually belongs to “relief” and were predicted as “relief”, false positives (FP) as posts that actually belongs to “non relief” but were predicted as “relief” and false negatives (FN) as posts that were predicted as “non relief” but actually belong to “relief”.

The evaluation metrics were calculated by the use of a stratified 10-fold-cross-validation. This was done for each classifier several times and each time with a different feature set that was produced by one of the weighting schemes. So, for each classifier we could identify which feature weighting scheme produced the best result. Finally, we compared these results among each other. As shown in Table 1, we found that Support Vector Machines with a binary feature weighting outperforms the other classifiers with a precision of 0.767, a recall of 0.800 and a F_1 score of 0.783.

Table 1: Best results for different classifiers after a 10-fold-cross-validation with different feature weighting schemes

Methods	Weighting	Precision	Recall	F_1
Decision Trees	binary	0.706	0.730	0.718
Naïve Bayes	binary	0.586	0.809	0.680
Multinomial Naïve Bayes	TF	0.525	0.649	0.580
Support Vector Machines	binary	0.767	0.800	0.783
K-Nearest-Neighbor	binary	0.697	0.733	0.715

4 Locating Voluntary Relief Activities

After extracting those messages that mention help efforts, we wanted to identify the locations they refer to. Due to the lack of explicit geotags in the metadata of the Facebook posts (less than 1% were geotagged), a custom approach was needed for relating the messages to its' geographical locations. Hence, this work presents a message geocoding method that exploits place names from the content of Facebook posts.

4.1 Message Geoparsing

Finding textual references to geographic locations, known as toponyms, is also called geoparsing or toponym recognition. A common approach for this task is to use *Named Entity Recognition* (NER), a statistical NLP technique to find mentions of people, places, organizations, times and locations. We utilized the *Name Finder* component from the *Stanford CoreNLP Natural Language Processing Toolkit*. To apply this Named Entity Recognizer on German Facebook posts, we used a model that was trained on German newspaper text corpora and was proposed by Faruqui and Padó (2010). This model provides four entity classes: person (PER), location (LOC), organization and miscellaneous (MISC). To evaluate the effectiveness of the NER, we compared its results with a validation dataset. This was constructed by tokenizing the posts from the text classification training dataset that belongs to "relief" and manually annotating it with one of the four entity classes. Thus, we were able to identify the count of TP, TN and FP and calculate precision, recall and F_1 score. We found that the precision was significantly higher than the recall (Table 2).

The sacrification of recall in favour of precision by NER tools is a problem, that was already addressed by Lieberman (2010). Regarding to his work, we extended the out-of-the-box Stanford NER by a custom approach. First, we normalized the Facebook posts by lowercasing those terms that were completely written in capital letters and then uppercasing the first letter of all terms that were recognized as named entities by the *Stanford Part-Of-Speech (POS) Tagger*. We also looked for those terms that were recognized as nouns by the POS tagger and had a preposition (e.g. "at", "in", "to") in prior position. Following, we applied a look up for the found named entities and nouns in the free *GeoNames* geographical database. Those terms that had a corresponding entry in the database was marked as a location.

Finally, we calculated the evaluation metrics for the extended approach and compared it with the result from the out-of-the-box NER (Table 2). We could significantly increase the recall while the precision was reduced only slightly. This also affected in a higher F_1 score. In conclusion, our results show that an out-of-the-box NER can be improved by applying some additional processing. However, we assume that a NER model that is trained on German Facebook posts instead of newspapers will affect in even better results but will also require time-consuming labelling a large number of text tokens.

Table 2: Comparison of the result of the Stanford out-of-the-box NER and an extended approach on Facebook posts

Methods	Precision	Recall	F_1
Stanford NER out-of-the-box	0.914	0.434	0.589
Stanford NER extended	0.834	0.676	0.747

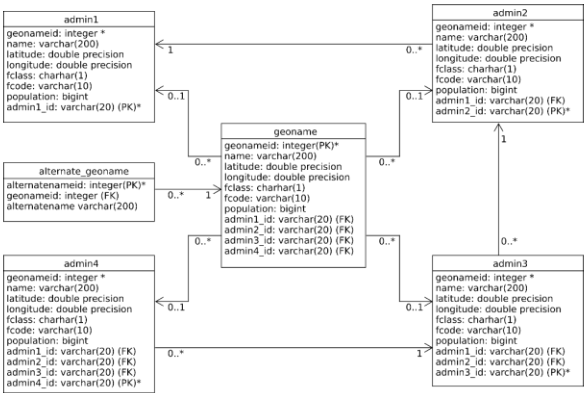
4.2 Message Geocoding

After all, we still had to identify the correct location for each extracted toponym and assign it to geographic coordinates. This task is also known as toponym resolution. The problem for this task is to solve toponym ambiguities since some toponyms may refer to multiple geographic locations that share the same name.

At this point, we want to introduce the *GeoNames Conceptual Density* that is based on the *Conceptual Density* approach. This is a knowledge-based method for disambiguating toponyms which transfers a technique for resolving ambiguous word senses to the geographical domain and exploits the structure of the lexical database *WordNet* for calculating a kind of conceptual density (Buscaldi, 2008).

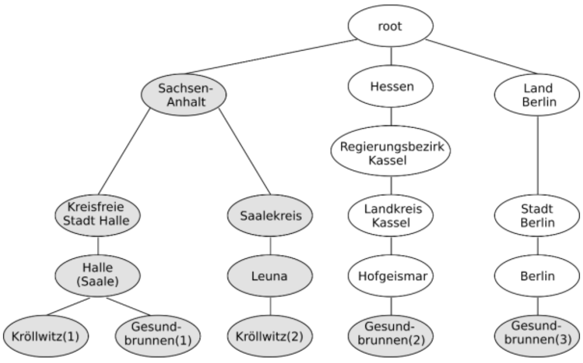
Our approach utilizes the GeoNames database to construct similar relationships between geographical names like the WordNet holonym hierarchies. For this, we imported a GeoNames database dump into a *PostgreSQL* database. As shown in Figure 2, we prepared the tables in a way that each geographical term relates to its higher-level administrative unit. So, each entry in the database can be seen as a synset with its own subhierarchy formed by its holonym relations.

Figure 2: Data model for geographical terms and administrative units from *GeoNames*


admin1
geonameid: integer *
name: varchar(200)
latitude: double precision
longitude: double precision
fclass: charhar(1)
fcode: varchar(10)
population: bigint
admin1_id: varchar(20) (PK)*
admin2
geonameid: integer *
name: varchar(200)
latitude: double precision
longitude: double precision
fclass: charhar(1)
fcode: varchar(10)
population: bigint
admin1_id: varchar(20) (FK)
admin2_id: varchar(20) (PK)*
geoname
geonameid: integer(PK)*
name: varchar(200)
latitude: double precision
longitude: double precision
fclass: charhar(1)
fcode: varchar(10)
population: bigint
admin1_id: varchar(20) (FK)
admin2_id: varchar(20) (FK)
admin3_id: varchar(20) (FK)
admin4_id: varchar(20) (FK)
alternate_geoname
alternatenameid: integer(PK)*
geonameid: integer (FK)
alternatename varchar(200)
admin4
geonameid: integer *
name: varchar(200)
latitude: double precision
longitude: double precision
fclass: charhar(1)
fcode: varchar(10)
population: bigint
admin1_id: varchar(20) (FK)
admin2_id: varchar(20) (FK)
admin3_id: varchar(20) (FK)
admin4_id: varchar(20) (PK)*
admin3
geonameid: integer *
name: varchar(200)
latitude: double precision
longitude: double precision
fclass: charhar(1)
fcode: varchar(10)
population: bigint
admin1_id: varchar(20) (FK)
admin2_id: varchar(20) (FK)
admin3_id: varchar(20) (PK)*
1
0..*
0..1
0..1
0..*
0..*
0..*
1
1
0..*
0..*
0..1
0..1
0..*
0..*
1

We want to demonstrate the toponym disambiguation procedure by an example. The following Facebook post contains two toponyms: “Volunteers are needed in *Kröllwitz* and *Gesundbrunnen* to protect the dams!” The GeoNames database has two entries for *Kröllwitz* and three entries for *Gesundbrunnen*. So, each toponym has multiple senses. With these and its’ higher level administrative units, a holonym tree like the one shown in Figure 3 can be constructed. In this, the relevant synsets for the toponym *Gesundbrunnen*, that form the context, are marked grey.

Figure 3: Holonym tree for geographical terms from the GeoNames database



Following, the *Base Conceptual Density* for each sense of an ambiguous toponym can be calculated by

$$baseCD(M, nh) = \frac{M}{nh} \qquad (1)$$

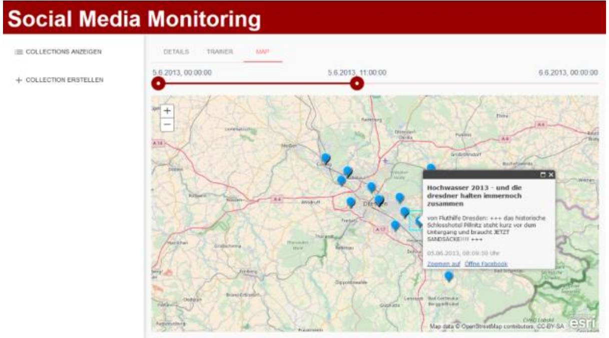
where *M* are the relevant synsets in the subhierarchy and *nh* is the total number of synsets in a subhierarchy. More clearly, *M* gives the number of all candidate toponyms and its higher level administrative units that share a common context. For the first meaning of *Gesundbrunnen* there are two other toponyms that share *Sachsen-Anhalt* as the common context. Altogether, there are 8 synsets in that subhierarchy, so it gives a value of *M*=8. In contrast, the other two meanings of *Gesundbrunnen* don’t share a common context with other candidate toponyms which results in a value of *M=1* for each.

The *baseCD* has to be calculated for each sense of a toponym. The one with the highest score is most likely to be the actual sense for a toponym. For *Gesundbrunnen* the formula in (1) gives the following scores:

- *Gesundbrunnen* (1): $baseCD(M = 8, nh = 8) = 1$
- *Gesundbrunnen* (2): $baseCD(M = 1, nh = 5) = \frac{1}{5}$
- *Gesundbrunnen* (3): $baseCD(M = 1, nh = 4) = \frac{1}{4}$

Hence, the first sense of *Gesundbrunnen* that refers to *Halle (Saale)* would be the one, we are looking for. Thus, the corresponding coordinates that are stored in the *GeoNames* database would be assigned to it.

Figure 4: Situation map of the social media monitoring platform that shows Facebook posts



Hochwasser 2013 - und die dresdner halten immernoch zusammen

von Fluthilfe Dresden: +++ das historische Schlosshotel Pillnitz steht kurz vor dem Untergang und braucht JETZT SANDSÄCKE!!!! +++

05.06.2013, 08:09:30 Uhr

Zoomen auf Öffne Facebook

In conclusion, our *GeoNames Conceptual Density* approach can be utilized for an automatic geocoding of social media messages by the use of textual references to geographic locations. This enables the mapping of those messages. For the purpose of a social media monitoring component within the KUBAS system, we also developed an open source platform, that combines all the above described techniques. As depicted in Figure 4, our platform provides a situation map that shows all Facebook posts that mention help efforts and provides a link to the specific discussion on the Facebook platform where the post appeared.

5 Conclusions

In this paper we presented different techniques and approaches that, in combination, enable an automatically geocoding of crisis related social media messages. We showed how to extract and filter relevant posts that mention relief efforts from Facebook by the use of supervised text classification. We also explained how Natural Language Processing can help to extract textual location information in order to enable geocoding if no explicit geotags exist. In addition, we introduced a linguistic approach, based on the free GeoNames database, to disambiguate ambiguous toponyms.

Finally, we developed a social media monitoring platform that can easily be integrated into the existing system architecture of KUBAS. Emergency responders can use it in crisis situation to locate voluntary activities in a timely manner. The extracted messages provide complementary information for disaster control authorities. For instance, if a certain location is mentioned in several messages related to help efforts, crisis responders may assume, that an increased number of volunteers will be present at operation site. So, the situation map within our platform support emergency responders by providing an easy to understand view on ongoing events. This additional information is very helpful for the coordination of voluntary activities within the context of KUBAS. Therefore, our approach contributes to a specific kind of situational awareness in crisis events

References

Bundesministerium des Innern (BMI) (2013) *Bericht zur Flutkatastrophe 2013: Katastrophenhilfe, Entschädigung, Wiederaufbau*. [Online] Available from: http://www.bmi.bund.de/SharedDocs/Downloads/DE/Broschueren/2013/kabinettbericht-fluthilfe.html [Accessed 3rd April 2018].

Buscaldi, D. and Rosso, P. (2008) A conceptual density-based approach for the disambiguation of toponyms. In: *International Journal of Geographical Information Science*, 22(3), 301–313.

Dhavase, N. and Bagade, A. M. (2014) Location Identification for Crime & Disaster Events by Geoparsing Twitter. In: *International Conference for Convergence of Technology*.

Facebook (2018) *Documentation: Graph API Overview*. [Online] Available from: https://developers.facebook.com/docs/graph-api/overview [Accessed 3rd April 2018].

Faruqui, M., and Padó, S. (2010) Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In: *Proceedings of the Conference on Natural Language Processing 2010*, pp. S. 129–133.

Franke, E., Hall, M. A., and Witten, I. H. (2016) *The WEKA Workbench: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*. [Online] Available from: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf [Accessed 3rd April 2018].

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P. (2013) Extracting Information Nuggets from Disaster-Related Messages in Social Media. In: *Proceedings of the 10th International ISCRAM Conference*.

Kaufhold, M.-A. and Reuter, C. (2016) The Self-Organization of Digital Volunteers across Social Media: The Case of the 2013 European Floods in Germany. In: *Journal of Homeland Security and Emergency Management*, 13(1), 137-166.

Manning, C. D., Raghavan, P., and Schütze, H. (2009) *Introduction to information retrieval* (Reprinted.), Cambridge, Cambridge University Press.

Martin-Luther-Universität Halle-Wittenberg (2016) *Goal of KUBAS*. [Online] Available from: https://kubas.uni-halle.de/en [Accessed 3rd April 2018].

Middleton, S. E., Middleton, L. and Modafferi, S. (2014) Real-Time Crisis Mapping of Natural Disasters Using Social Media. In*: IEEE Intelligent Systems*, 29(2), 9-17.

Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. In: *ACM Computing Surveys*, 34(1), 1-47.

Starbird, K., Palen, L., Hughes, A. L. and Vieweg, S. (2010) Chatter on The Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 241–250.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A. and Anderson, K. M. (2011) Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 385–392.

Vieweg, S., Hughes, A. L., Starbird, K. and Palen, L. (2010) Microblogging during two natural hazards events. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1079–1088.