Social media as a rainfall indicator

Yu Feng
Leibniz University of Hannover
Institute of Cartography and Geoinformatics
Appelstraße 9a
Hannover, Germany
yu.feng@ikg.uni-hannover.de

Monika Sester
Leibniz University of Hannover
Institute of Cartography and Geoinformatics
Appelstraße 9a
Hannover, Germany
monika.sester@ikg.uni-hannover.de

Abstract

Social media is nowadays more and more used for detecting events and hot topics among the crowds. Detection of events related to extreme weather and disasters is helpful, especially to find eyewitnesses for disaster research and risk analysis. The posts come in real time and contain rich information, not only the texts, but also sometimes the locations, photos and videos. In this paper, we propose a method to filter rainfall relevant Tweets and detect events using spatiotemporal clustering. The main contribution of this paper is that we improved the previously widely used approach - keyword filtering - with a trained language model to extract rainfall events relevant Tweets. We trained this language model based on historical weather records. This method shows a higher precision for filtering rainfall event relevant Tweets, which improved the data quality for the further spatiotemporal event detection.
Keywords: social media, volunteered geographic information, text classification, disaster management, information retrieval.

1 Introduction

Strong urban rainfalls nowadays lead to severe problems and challenges for big cities. Although various weather monitoring systems are already built in a great range and relatively high density in populated areas (Muller et al., 2013), there is still a lack of high systems recording the information at high temporal resolution. Thus, the real time information posted voluntarily by the users on the social media platforms has a large potential for detecting such event efficiently and estimating the rough inundation area with low effort.

The detection of disaster events using social media may sometimes perform better than the traditional public services. For instance, Sakaki et al. (2010) reported an earthquake detection system which was able to send alert emails much earlier than the public announcements from the authority.

When extreme weather such as heavy rainfall or lightening occurs, many people want to share their feeling and experience via social media platforms such as Twitter, Instagram, Facebooks or Sina Weibo with their friends and families. The posts, which are relevant in the context of heavy rainfall events, are important eyewitnesses for researchers to extract relevant disaster information.

Firstly, they can discover the occurrence of flood and its probable inundation area when such event gets more people's attention. Secondly, these posts are informative documents for risk and loss analysis, as well as for the forecast model validation. This information currently is gathered using rather time consuming telephone interviews. The extracted rainfall relevant Tweets could be used for a more efficient telephone interview.

2 State of the art

Hundred millions of Tweets are sent by social media users every day. They are related to a large number of topics. To extract the Tweets related to specific natural events, such as earthquake, flooding, keyword filtering is the most frequently used approach.

Li et al. (2012) set up a Twitter based event Detection and Analysis system (TEDAS) to detect Crime and Disaster related Events (CDE). The query of CDE was based on spatial range, temporal period and user defined keywords. Fuchs et al. (2013) applied a density-based spatiotemporal clustering OPTICS (Ankerst et al., 1999) on the Tweets including both "Hoch" and "Wasser" (flooding corresponds to "Hochwasser" in German) to extract spatiotemporal clusters and they detected large scale flooding events in Germany in 2012. Dittrich and Lucas (2014) applied different groups of keywords in 43 languages to detect various natural disasters such as floods, earthquake, and volcanic eruptions. Sakaki et al. (2010) used also keywords and applied Support Vector Machine (SVM) with linear kernel to classify Tweets into positive (earthquake relevant) and negative categories by preparing 597 labelled datasets. Statistical and contexts features were used in this research, such as number of words of a Tweet, the position of the query word within a Tweet, and the words before and after the query word.

Currently, most of the proposed event detectors, which aim at detecting a specific type of disaster events from Tweets, have applied keyword filtering. However, filtering with a user predefined keywords list is not an optimal strategy, because of the ambiguities of words. For instance, words such as 'shaking', 'storm' are not always related to disasters, but also frequently used in daily life. If we include them in the keywords list, it may lead to miss-detection. If not, we may miss lots of information for the events we would like to extract and document. Thus, to determine whether a word belongs to the keywords is often a trade-off between precision and recall. To improve the event detector, in this paper we propose a language model, which is a trained classifier, to improve the detection of rainfall relevant events in Twitter data.

3 Methods

The approach proposed in this paper is a supervised classification procedure. The collected Tweets are firstly pre-processed and filtered with rainfall relevant keywords – similar to other methods described in the state of art. Each filtered Tweet is then labeled according to the weather records. Tf-idf (term frequency - inverse document frequency) features (Manning et al. 2008) are generated from the raw texts and used for training. With a training of the labeled data, the model is then used for prediction whether an unseen Tweet is rainfall relevant. At last, spatiotemporal clustering was applied to generate events and their affected areas.

3.1 Preprocessing

The Twitter Streaming API offers us an access to the real time Tweets posted by users (Twitter, 2017). However, only 1% of all current Tweets can be crawled by the API according to the reports of many Streaming API users (Dittrich and Lucas, 2014). Therefore, instead of using a crawler globally, we collected Tweets only in several small areas such as west Europe, east Europe, west United States, east United States, using several crawlers, to avoid this 1% limitation.

Since we currently focused on Europe as our test area, we set a bounding box (as shown in Figure 1) from -24.54652 to 18.5000 in longitude and from 27.636311 to 71.185509 in latitude to collect Twitter data. Subsequently, we saved the geo-tagged Tweets in a MongoDB database.

Figure 1: Test area in Europe used for collecting Tweets

Source: Basemap - OpenStreetMap.

Each Tweet contains multiple fields, but for this research only a subset of the fields were extracted, namely the creation time, coordinates, source, user's screen name, language and text, i.e. the message itself. Before further analysis, the text of each Tweet was pre-processed and filtered.

For the texts in Tweets, we firstly removed all the punctuations, numbers, URL and emoticons. Since we used the bag-of-words assumption in the classification later, the relations to the neighboring words are not taken into account. Therefore, for the rest of words, we removed the stopwords to avoid the most common function words such as pronouns, prepositions. Subsequently, we conducted stemming to extract the base form of each word using the Python implementations in the Natural Language Toolkit (ntlk) library for multiple languages (Bird et al., 2009). Both processes were conducted according to the language field given by each Tweet respectively.

3.2 Keyword filtering

The preprocessed and reduced list of words in the Tweets is then filtered with a list of predefined rainfall relevant keywords. It contains concepts such as "flood", "inundation", "rain" and "storm" in English, French, German, Italian, Spanish, Portuguese and Dutch, which are the frequently used languages in the test area and also well supported by the ntlk library used for stemming.

After the keywords filtering, we found that many Tweets are posted by automatic weather services, also called Twitter robot. As many Tweets sent by automatic weather services differentiate with each other only at the numbers and URL. After the preprocessing, many of the documents sent by the robots shared the same words, even the same orders. With this characteristic, we searched the duplicates in the corpus list and checked its user name. When more than 3 Tweets with the same corpus and user name were found, the Tweets with this user name were filtered. With this approach, most of the weather services were found, which in a great range eliminated the noise in the datasets.

3.3 Automatic labeling

In this supervised learning approach, the data was classified into the binary classes positive and negative, which correspond to rainfall relevant and non-relevant. One of the great challenges for a supervised learning algorithm is to provide the necessary amount of labeled training data.

Manually labeling is time consuming. It needs great efforts to achieve a sufficient amount of training data for a supervised classification. Instead of labeling all the Tweets manually, we used extra data sources, such as weather data as rainfall indicator. With the increasing development of earth observation systems, environmental monitoring systems offer us large amounts of data for various events.

In this approach, a crawler was applied to collect the historical data from the public online weather websites, such as Weather Underground (https://www.wunderground.com/). The keyword filtered Tweets were then labeled according to the corresponding weather records on that day at that location. If the weather description at the Tweet's location on that day includes words related to rainfall, this Tweet was then labeled as positive. If not, they were labeled as negative. As the Tweets have been extracted by keyword filtering (using rain-related keywords), most of the Tweets belong to the class positive (rain), there was an imbalance of positive and negative labels. Thus, we randomly selected Tweets which did not contain any of the predefined keywords, and labeled them as negative. By this means, we achieved an equal partition into the positive and negative labels, and thus balanced the training data.

3.4 Text classification

Similar to Sakaki et al. (2010), we assumed that the combination of words in Tweets may indicate whether the Tweet is topic related or not. However, statistical features stated in the research above are not sufficient and have less influence on the relevance to the event. Therefore, we considered only the context features and applied the general approach in natural language processing.

Features of the raw texts were generated using the widespread approach tf-idf weighting with the help of the scikit-learn library (Pedregosa et al., 2011). This weight indicates the importance of a word to a document in a corpus. By this means, the whole training dataset can be represented as a matrix, where the columns correspond to the unique words in the whole corpus and the rows correspond to the documents. The values in this matrix are the weights of the words respect to the document and corpus.

In this research, 10% of the documents were selected beforehand as test dataset and the rest 90% were used for the training. The input for the training was the matrix generated using tf-idf weighting and their corresponding labels. As we considered only the occurrence and combination of words to indicate the event, it corresponds exactly to the bag-of-words assumption, which disregards grammar and text order. We tested the classifiers such as Naïve Bayes classifier, Logistic Regression, Support Vector Machine (SVM) with linear kernel and SVM with RBF (radial basis function) kernel on the training dataset. These methods were frequently used for text classification tasks. Python implementations of these methods in scikit-learn library (Pedregosa et al., 2011) were used for this comparison. Moreover, we picked the one which performed best on test dataset to be the best suited method.

3.5 Spatiotemporal clustering

With the extracted rainfall relevant Tweets, a density based clustering ST-DBSCAN (Birant and Kut, 2007) was applied to extract spatiotemporal clusters, which indicates the probable location, time range and inundation area of a rainfall event. The spatial density of the Tweets does not indicate the intensity of the rainfall event but the density of eyewitnesses about this event.

4 Results

4.1 Filtering and text classification

Totally, we collected, keyword-filtered and labeled 12629 Tweets ranging from Jun. 1 to Jul. 30, 2016. On average, 71% of the keywords filtered Tweets were labeled according to the weather records as rainy day.

Then we applied the text classification with the methods described in Section 3.4. As shown in Figure 2 and Table 1, the methods were evaluated based on accuracy and F1-score. F1-score conveys the balance between the precision and recall, and can be calculated by the harmonic mean of precision and recall. Both SVM with linear kernel and SVM with RBF kernel perform well and can achieve an accuracy and F1-score over 80%. This means most of the Tweets can be correctly predicted as rainfall relevant or non-relevant with these binary classifiers.

During the training with different methods, SVM with RBF kernel was found to be computationally much more expensive than the other three methods. This is because RBF uses a non-linear kernel. Thus, we chose SVM with linear kernel as the best suited classifier for this language model.

Figure 2: Evaluation on test datasets.

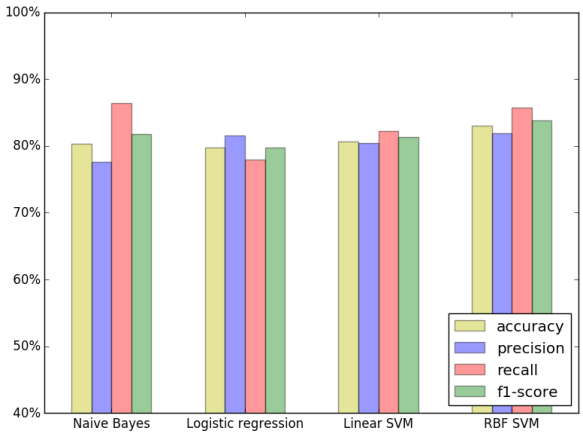


Table 1: Evaluation on test datasets.

Methods	Accuracy	F1-score
Naïve Bayes	80.28%	81.77%
Logistic regression	79.72%	79.73%
Linear SVM	80.67%	81.31%
RBF SVM	83.06%	83.82%

4.2 Event detection

After the classification, ST-DBSCAN was applied to detect spatiotemporal clusters on the filtered Tweets. This method needs three parameters: *eps1* in time dimension, *eps2* in spatial dimension and a minimum number of Tweets *minPts*. As rainfall events normally have big affected area and may last for several hours, in this case we defined them as one hour, 50km and three Tweets respectively. This means, the Tweet locations (points) are considered as a cluster when there are at least three neighbors within one hour time differences and 50km in Euclidian distance. As we found some of the users may send several posts at the same place and same time with similar contents, instead of using minimum number of Tweets, we changed it to the number of different users for event detection to improve credibility.

4.3 Case study

On Jun. 23 2016, extreme rainfall events happened in England and the Benelux nations. On that day, we collected 90604 Tweets in the test area. We applied the SVM with linear kernel on these Tweets without any keyword filtering and yielded 955 Tweets classified as rainfall relevant. The

extracted Tweets using our approach are visualized as a heat map as shown in Figure 3. The highest intensities are the south of England, the Netherlands and Belgium, where there are many eyewitnesses of this event.

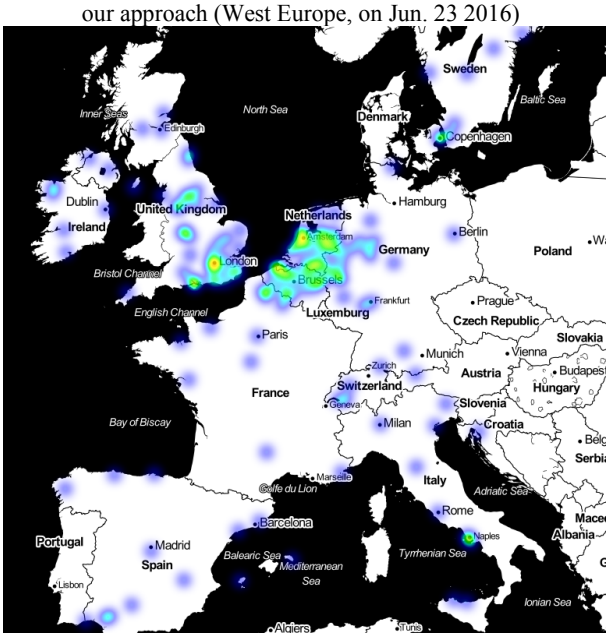
For comparison, we applied a direct keyword filtering according to the list mentioned in Section 3.2 to examine whether our approach has improved the filtering results. This lead to 911 Tweets categorized as rainfall relevant. 825 of the detected Tweets by both methods are the same. We then compared the difference of the filtered Tweets using both methods and applied a manual examination only on these different Tweets, assuming that a confirmation by both methods corresponds to a true classification. The results as shown in Table 2 indicate that, comparing with keyword filtering, our approach has extracted more event relevant Tweets and improved the filtering precision.

ST-DBSCAN was then applied on the Tweets which were filtered with our approach. Spatiotemporal clusters were detected as shown in Figure 4. The most extraordinary cluster was found near London (as shown is Figure 5), which lasted almost the whole day with 374 eyewitnesses.

Table 2: Manual examination on rainfall relevant Tweets detected by the two methods

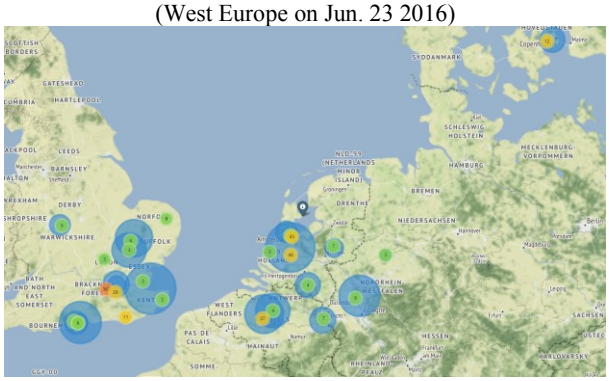
Methods	Same	Difference		Total
		Relevant	Non-relevant	
Keyword filtering	825	41	45	911
Our approach	825	81	49	955

Figure 3: Heat map of rainfall relevant Tweets extracted by our approach (West Europe, on Jun. 23 2016)



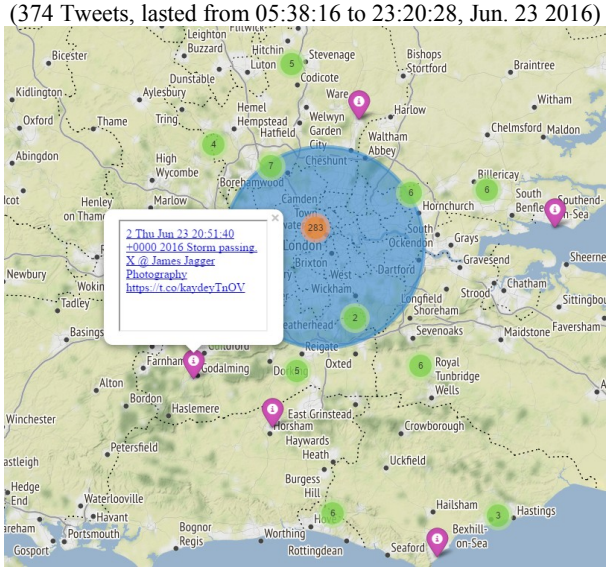
Source: Basemap - OpenStreetMap.

Figure 4: Extracted Spatiotemporal Clusters (West Europe on Jun. 23 2016)



Source: Basemap - OpenStreetMap.

Figure 5: The most significant ST-cluster near London (374 Tweets, lasted from 05:38:16 to 23:20:28, Jun. 23 2016)



Source: Basemap - OpenStreetMap.

5 Conclusions

In this paper we have proposed a method for extracting rainfall events from Twitter data, including a supervised classification. With this infrastructure, we could discover rainfall events and document the majority of Tweets relevant to rainfall events. A filtering accuracy over 80% was achieved, when using the SVM with linear kernel as the classifier. This trained language model could well describe the rainfall relevant events and achieve a higher precision than the previously used keyword filtering. In this way, we improved the quality of the results. With the help of spatiotemporal clustering, events were detected for early warning and the rainfall relevant Tweets were documented for the further risk analysis.

In this work, we used the sparse tf-idf features, which treated each word independently. Such as the words “rainfall” and “rain”, they were treated as different words even though they have similar meaning. In the next step, we plan to apply

word embedding, such as word2vec model (Mikolov et al., 2013), to introduce the latent semantic relation between words into the text classification.

As we are focusing on extreme or rare events, in the next step, we also plan to learn the level or strength of rainfall events based on the use of specific words. For instance, if more users would employ "heavy rain" instead of "shower" or "drizzle", it should have a higher probability that it rains more heavily. On the other hand, we would like take the population density into consideration as a further criterion for choosing spatiotemporal clustering parameters to achieve a more robust clustering result.

Acknowledgement

The work presented in this paper was supported by the project EVUS - Real-Time Prediction of Pluvial Floods and Induced Water Contamination in Urban Areas (BMBF, 03G0846B).

References

Ankerst, M., Breunig, M. M., Kriegel, H. P. and Sander, J. (1999) OPTICS: ordering points to identify the clustering structure. In: *Proceedings of the ACM SIGMOD Conference*, pp. 49-60.

Birant, D. and Kut, A. (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data. In: *Data and Knowledge Engineering*, 60(1), pp. 208–221.

Bird, S., Klein, E. and Loper, E. (2009) *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media.

Dittrich, A. and Lucas, C. (2014) Is This Twitter Event a Disaster? In: *Connecting a Digital Europe through Location and Place. Proceedings of the AGILE'2014 International Conference on Geographic Information Science, Castellón, June, 3–6, 2014*.

Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S. and Stange, H. (2013) Tracing the German centennial flood in the stream of tweets: first lessons learned. In: *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*, pp. 31-38.

Li, R., Lei, K. H., Khadiwala, R. and Chang, K. C. C. (2012) TEDAS: A twitter-based event detection and analysis system. In: *2012 IEEE 28th International Conference on Data Engineering*, pp. 1273-1276.

Manning, C. D., Raghavan, P. and Schütze, H. (2008) *Introduction to information retrieval*, pp. 107-109. Cambridge, Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*, pp. 3111-3119.

Muller, C. L., Chapman, L., Young, D. T., Grimmond, C. S. B. and Cai, X. (2013) Sensors and The City: A Review of Urban Meteorological Sensor Networks. *International Journal of Climatology*, 33, 1585-1600.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th international conference on World wide web*, pp. 851-860.

Twitter (2017) *Streaming APIs*. [Online] Available from: https://dev.twitter.com/streaming/overview [Accessed 10th April 2017].