# Robust Filtering of Crisis-related Tweets

**Jens Kersten***
German Aerospace Center, Jena, Germany†
jens.kersten@dlr.de

**Anna Kruspe**
German Aerospace Center, Jena, Germany
anna.kruspe@dlr.de

**Matti Wiegmann**
German Aerospace Center, Jena &
Bauhaus-Universität Weimar, Germany
matti.wiegmann@uni-weimar.de

**Friederike Klan**
German Aerospace Center, Jena, Germany
friederike.klan@dlr.de

## ABSTRACT

Social media enables fast information exchange and status reporting during crises. Filtering is usually required to identify the small fraction of social media stream data related to events. Since deep learning has recently shown to be a reliable approach for filtering and analyzing Twitter messages, a Convolutional Neural Network is examined for filtering crisis-related tweets in this work. The goal is to understand how to obtain accurate and robust filtering models and how model accuracies tend to behave in case of new events. In contrast to other works, the application to real data streams is also investigated. Motivated by the observation that machine learning model accuracies highly depend on the used data, a new comprehensive and balanced compilation of existing data sets is proposed. Experimental results with this data set provide valuable insights. Preliminary results from filtering a data stream recorded during hurricane Florence in September 2018 confirm our results.

### Keywords

Filtering, Convolutional Neural Networks, Natural Disasters, Twitter, Model Transferability.

## INTRODUCTION

Several studies and use cases have demonstrated the great value and importance of analyzing social media streams, e.g. for rescue and helping activities in case of natural disasters and technical accidents (Reuter and Kaufhold 2018). Due to the huge and higly active user community distributed over the world, the ability to post messages from many sources as well as the accessibility of data streams via official APIs, research mainly focuses on Twitter. Furthermore, various labeled Twitter data sets related to crises and natural disasters are available.

As tweets posted during disasters are extremely varied, an automatic system needs to start by filtering out messages that do not contribute valuable information (Imran, Elbassuoni, et al. 2013). Existing approaches for filtering are keyword- or location-based, or utilize methods of natural language processing (NLP), visual analytics, or machine learning (Imran, Castillo, et al. 2015). An emerging availability of various labeled data sets recently enabled the successful application of deep learning to the problem of filtering (Burel and Alani 2018) as well as categorization of tweets into crisis-related information classes (Burel, Saif, et al. 2017, Nguyen, Al-Mannai, et al. 2017).

In this work, the robustness and accuracy of deep learning models is investigated for the task of filtering disaster-related tweets. Formulating filtering as a, preferably unambiguous, binary classification problem allows for experiments related to the impact of training data on the model quality. As investigated in (Stowe et al. 2018), a careful definition of thematic classes in filtering and classification is a crucial step. The authors distinguish between *related* and *unrelated* tweets, where the first class represents messages that are somehow related to disasters, and the

*corresponding author
†http://DLR.de/dw

latter class contains all off-topic posts. Since inter-annotator agreements around 92 % were obtained for this binary classification problem, this definition is also used here.

To develop robust filtering models, we compiled a new comprehensive and balanced dataset, which is one of the main contributions. More specifically, we are interested in performance differences of models covering various different event types (e.g. hurricanes, floods, earthquakes etc.) compared to models exclusively trained on one specific event type. The goal is to understand how to obtain accurate and robust filtering models and in particular how model accuracies tend to behave in case of new events. In contrast to other works, the application to real data streams is also investigated here.

In the first step, a balanced training data set covering several disaster types and containing approximately 66,000 tweets for both classes is compiled from several existing sets. Various event-specific and unspecific subsets of this training data set are then used to train and test a state-of-the-art Convolutional Neural Network (CNN). The experimental results show that the filtering task in case of new events can be tackled by global models covering several event types with acceptable F1-scores of 0.8 and better, while the standard deviation tends to be slightly higher than for models specifically trained for single event types. Preliminary results from filtering a data stream recorded during hurricane Florence in September 2018 further confirm the observed model behaviors.

## RELATED WORK

It is easier than ever to search the web for information about disasters, but filtering out off-topic discussions from the ocean of online content remains difficult (Landwehr and Carley 2014). A potentially huge fraction of messages collected for the purpose of deriving information related to a given crisis will not be related to the considered event. This fraction depends on the specific collection method used and on other factors, such as the presence of off-topic messages using the same tags or keywords as the on-topic ones (Imran, Castillo, et al. 2015).

The official Twitter streaming API provides filtering by location, keywords, author, and others. An alternative for real-time stream filtering is the use of tools based on location or keywords, for example TweetTracker (Kumar et al. 2011). Following the approach of keyword-based search, a lexicon of crisis-related terms (*CrisisLex*), intended to be used for collecting and filtering microblogged communications in crises, was developed (Olteanu, Castillo, et al. 2014). Since also unrelated messages containing the keywords will be retrieved, this approach usually requires post-filtering, for example using human labeling, crowdsourcing or automatic classification (Imran, Castillo, et al. 2015). Furthermore, a significant fraction of relevant data might mistakenly be discarded if the keywords are not appropriate, e.g. when the event is new or has changed over time. In location-based filtering, a large percentage of tweets will be discarded because only a small fraction of around 2 % is usually geo-referenced (Burton et al. 2012).

In order to mitigate the aforementioned problems of keyword-based approaches, methods for analyzing the information contained in tweets have been investigated in several studies. Analyzing tweets using traditional NLP approaches is known to be challenging, for example because of the limitation to 280 characters, abbreviations, misspellings, emoticons, informal language with grammatical errors and varied vocabulary, improper sentence structure, and mixed languages (Ramachandran and Ramasubramanian 2018). In (Win and Aung 2017), the tasks tweet filtering, feature extraction and classification using a LIBLINEAR classifier are addressed. In case of training and testing tweets from the same event, maximum accuracies of around 91 % were obtained. For test tweets from unseen events (cross-domain experiment), significantly lower accuracies around 75 % were reported. In (Parilla-Ferrer et al. 2014) Naïve Bayes (NB) and Support Vector Machines (SVM) were used to automatically detect informative disaster-related tweets. While NB outperformed SVM in precision, SVM yielded a higher average F-measure (87.3 %) than NB (62.1 %). Even though methods from the area of NLP offer a wide range of possibilities for systematically and efficiently searching, filtering, sorting and analysing huge amounts of data, their application to social media data is a relatively new topic of research and confronts researchers with many questions and challenges (Gruender-Fahrer et al. 2018).

In (Kim 2014), CNNs were applied to the general task of sentence classification. Compared to 14 machine learning methods, a superior CNN performance was achieved for four of seven data sets, indicating a good model transferability. The approach of Kim was adapted to the problem of rapid classification of crisis-related data in (Nguyen, Al-Mannai, et al. 2017). The results demonstrate that classifying tweets in case of new events, i.e., when no event-specific data is available, is still possible with a decrease of AUC-scores in a range of 5 – 19 %. In (Burel and Alani 2018), Kim's CNN was applied to tweets for tackling the problems of filtering (*related* vs. *unrelated*) as well as classifying event types and information categories using data from the *CrisisLexT26* dataset (Olteanu, Vieweg, et al. 2015). F1-scores in the range of 80 – 84 % were obtained for the filtering task.

The results of recent works in the deep learning domain confirm the importance of high-quality data for training, also identified as one of the major challenges in social media analytics in (Stieglitz et al. 2018). Furthermore,

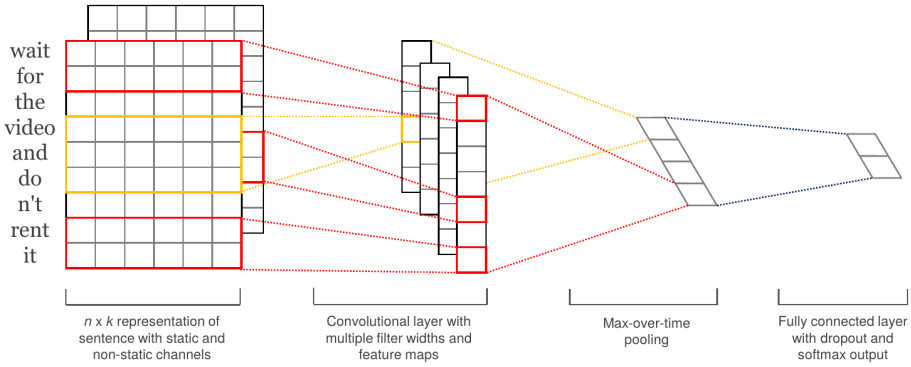



**Figure 1. CNN for text classification as proposed in (Kim 2014)**

the problem definition as well as other factors, like thematic class overlaps (Stowe et al. 2018) and the coverage of multiple event types (Periñán-Pascual and Arcas-Túnez 2018) may significantly affect model accuracy and robustness as well as its transferability to new events. Results of the studies discussed in this section only partially confirm these observations. With a focus on applying deep learning filtering models to real Twitter streams, the above-mentioned dependencies on data used for training is comprehensively and systematically investigated in this work. Since the CNN proposed by Kim has shown a great potential for crisis-related tweet filtering and classification, this model is utilized in our experiments.

## CONVOLUTIONAL NEURAL NETWORKS FOR TWEET FILTERING

Motivated by the promising results in (Burel and Alani 2018) for tweet filtering (*related* vs. *unrelated*), per-tweet event type categorization and per-tweet information type classification, the same CNN architecture (Kim 2014) is utilized in this work (see figure 1). The approach was specifically developed for classifying sentences into diverse categories, e.g. question types or sentiments. At first, each message is pre-processed by applying a set of commonly used standard operations, i.e., lowercasing, multiple space replacing, normalizing URLs, usernames and hashtags, removing special characters, normalizing digits as well as tokenization. Each word of a tweet is then transformed into a real-valued vector of 300 elements.

Commonly used pre-trained embeddings are based on *word2vec* (Mikolov et al. 2013). In contrast, we use the embedding proposed in (Imran, Mitra, et al. 2016), which was specifically trained on 52 million crisis-related twitter messages. Hence, each tweet is represented as a matrix of size $n \times k$, where $n$ is the maximum number of words occurring in the training data and $k = 300$ is the embedding vector length. In case the number of words is lower than $n$, zero padding is applied. Then, a convolutional layer applies $m$ kernels of different widths $w$ in parallel to the input matrix. We use standard values $m = 128$ and $w = (3, 4, 5)$, because parameter variation did not affect the results significantly. Global max-pooling is performed after each of these convolutional layers, and the results are concatenated. This new embedding is then fed into a fully connected layer with a dropout rate of 0.5 to determine the final class.

## DATASET

Various comprehensive labeled crisis-related Twitter data sets are available, enabling the application and evaluation of machine learning methods. Besides a categorization of event types and information classes, labels for relatedness, relevance or informativeness are usually provided or can be derived from the data, for example by merging all information class labels to a *related* class. However, from a machine learning point of view, a significant imbalance of available class instances in nearly all datasets can be observed. Since crisis-related tweets are often pre-selected by using specific keywords or hashtags, the number of related tweets is usually higher than the number of unrelated tweets. Furthermore, not all common types of disaster events are covered by all data sets. To overcome these issues, we compile a new comprehensive and balanced training data set for filtering tweets related to disasters of different types. In this work, the labels from five common data sets are transformed to comply with our definition (*related* vs. *unrelated*). To increase the low number of *unrelated* tweets, the *Events2012* (McMinn et al. 2013) data set is utilized. This set contains relevance judgments for 150,000 tweets covering more than 500 non-crisis events from a broad range of topics. An overview of the resulting data set is provided in table 1.

**Table 1. Composition of the proposed data set**

| *Data set* | *Tweet count related* | *Tweet count unrelated* | *Covered event types* |
|---|---|---|---|
| *Events2012* (McMinn et al. 2013) | 3,612 | 48,531 | Disasters and accidents, 500 other topics, like sports, arts, culture and entertainment |
| *CrisisLexT26* (Olteanu, Vieweg, et al. 2015) | 24,581 | 3,352 | Earthquake, flood, wildfire, meteor, typhoon, flood, train crash, explosions, building collapse, bombings |
| *CrisisNLP* (Imran, Mitra, et al. 2016) | 22,381 | - | Earthquake, hurricane, flood, typhoon, cyclone, ebola, MERS |
| *CrisisMMD* (Alam et al. 2018) | 11,524 | - | Hurricane, earthquake, wildfire, flood |
| *Epic* (Stowe et al. 2018) | 4,578 | 14,809 | Hurricane |
| *Sum* | 66,692 | 66,692 | |

According to our definition, the *CrisisLexT26* classes *related and informative* as well as *related - but not informative* are merged to the target class *related*. From *CrisisMMD* (Alam et al. 2018), only tweets labeled as *informative* are used as *related*. Since tweets from class *not informative* may or may not be related to crisis events, they are not used here. This also applies to *CrisisNLP* (Imran, Mitra, et al. 2016), where the class *not related or irrelevant* might still contain crisis-related tweets. *Epic* (Stowe et al. 2018) uses the same problem definition as this work. From *Events2012*, tweets related to disasters and accidents are used for the target class *related*. In order to achieve perfect class balance, randomly sampled tweets from all other non-crisis topics of this data set are used for the class *unrelated*.

## EXPERIMENTAL RESULTS

In this section, the goal is to evaluate the impact of the proposed data set compared to other related works as well as to investigate the performance behavior and robustness of models covering a broad range of event types (in the following denoted as "global models") compared to models specifically trained on single events (in the following denoted as "local models"). Local models individually trained on hurricane or flood event data sets are evaluated and compared to global models trained with the complete data set. Hurricanes and floods are chosen because of the their frequent occurence as well as the availability of data. Furthermore, the transferability of local and global models to new events is investigated.

The following experiments are conducted: First, global model performances are compared to a benchmark method as well as to local models trained for specific event types. In contrast to other studies, where accuracy measures obtained during cross-validation are reported (e.g. in (Burel and Alani 2018)), we use independent test data excluded from cross-validation. Model transferabilty is evaluated by using test data from unseen events. Secondly, we investigate how well local models perform in case of other event types. To this end, local flood models are applied to hurricane data and vice versa. Third, since available labeled data sets usually represent a pre-selected subset, we evaluate the transferability of our models to new Twitter data streams. First results for filtering ~1,120,000 geo-located tweets from the impact area during hurricane Florence (September 12-19, 2018) as well as after the event (January 08-15, 2019) are presented.

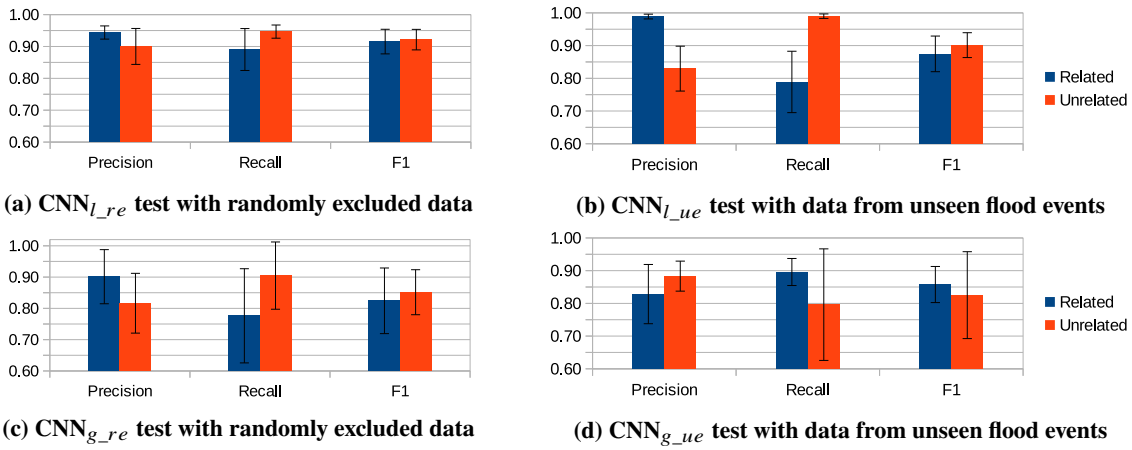
### Flood events: Global versus local models

The five data sets listed in table 1 cover the following six flood events: Phillippines 2012, Alberta 2013, Manila 2013, Queensland 2013 (all from *CrisisLexT26*, ~ 1,000 tweets per event), Pakistan 2014 (*CrisisNLP*, ~ 2,000 tweets) and Sri Lanka 2017 (*CrisisMMD*, ~ 800 tweets). Whereas global models are trained by using the full proposed data set, local models are trained solely based on the above-mentioned flood data sets. Perfect class balance is achieved by randomly sampling the required amount of *unrelated* tweets from the *Events2012* data set. In addition to typically used 10-fold cross-validation (CV), we exclude 10 % randomly sampled tweets from CV for independent testing. In order to test the the models' transferability, each of the events is successively excluded for training local and global models, and is instead used for testing. Hence, a total number of 120 models is trained and evaluated in this experiment, where for each omitted flood event 10-fold CV for a local and a global model is conducted. Besides commonly used averaged results for precision, recall and F1-measure, the standard deviation of these scores is analyzed here as well.

In figure 2, the obtained results for the local and global models averaged over all CVs as well as the six flood events are presented. In case of local model validation with randomly excluded data (figure 2a), F1-scores around 0.92 are obtained. As expected, with test data from unseen events, the local models perform slightly worse (figure 2b). Nonetheless, the average F1-scores are 0.88 and 0.90 for the classes *related* and *unrelated*, respectively. In contrast, the global models yield F1-values between 0.82 and 0.85 and therefore perform worse than the local models. Furthermore, the standard deviations for global models tended to be higher. In table 2, the results are compared to those reported in (Burel and Alani 2018), where the same model and problem definition with *CrisisLexT26* data were used ($\text{CNN}_{base}$). To the best of our knowledge, this represents the best results obtained for filtering (*related* vs. *unrelated*) up to now.

**Table 2. Local and global models tested with randomly excluded and unseen event data, where $l$=local, $g$=global, $re$=randomly excluded data, $ue$=unseen event data**

| | | *Related* | | | *Unrelated* | | |
|---|---|---|---|---|---|---|---|
| *Event* | *Model* | *Precision* | *Recall* | *F1* | *Precision* | *Recall* | *F1* |
| *Flood* | $\text{CNN}_{l\_re}$ | 0.944±0.021 | 0.891±0.066 | 0.915±0.039 | 0.900±0.056 | 0.947±0.021 | 0.922 ±0.032 |
| | $\text{CNN}_{l\_ue}$ | 0.989±0.007 | 0.789±0.094 | 0.875±0.054 | 0.830±0.069 | 0.990±0.007 | 0.901 ±0.038 |
| | $\text{CNN}_{g\_re}$ | 0.901±0.087 | 0.776±0.151 | 0.824±0.105 | 0.816±0.096 | 0.905±0.108 | 0.852 ±0.072 |
| | $\text{CNN}_{g\_ue}$ | 0.828±0.091 | 0.896±0.041 | 0.858±0.055 | 0.883±0.046 | 0.796±0.171 | 0.825 ±0.133 |
| *Hurricane* | $\text{CNN}_{l\_re}$ | 0.900±0.061 | 0.846±0.188 | 0.865±0.134 | 0.892±0.122 | 0.927±0.033 | 0.906 ±0.078 |
| | $\text{CNN}_{l\_ue}$ | 0.755±0.013 | 0.802±0.273 | 0.747±0.181 | 0.895±0.048 | 0.753±0.112 | 0.811 ±0.038 |
| | $\text{CNN}_{g\_re}$ | 0.904±0.095 | 0.775±0.152 | 0.825±0.108 | 0.810±0.102 | 0.905±0.124 | 0.846 ±0.087 |
| | $\text{CNN}_{g\_ue}$ | 0.766±0.222 | 0.863±0.124 | 0.805±0.185 | 0.892±0.151 | 0.803±0.179 | 0.839 ±0.147 |
| | $\text{CNN}_{base}$ | 0.839 | 0.838 | 0.838 | | | |




**(a) $\text{CNN}_{l\_re}$ test with randomly excluded data**

**(b) $\text{CNN}_{l\_ue}$ test with data from unseen flood events**

**(c) $\text{CNN}_{g\_re}$ test with randomly excluded data**

**(d) $\text{CNN}_{g\_ue}$ test with data from unseen flood events**

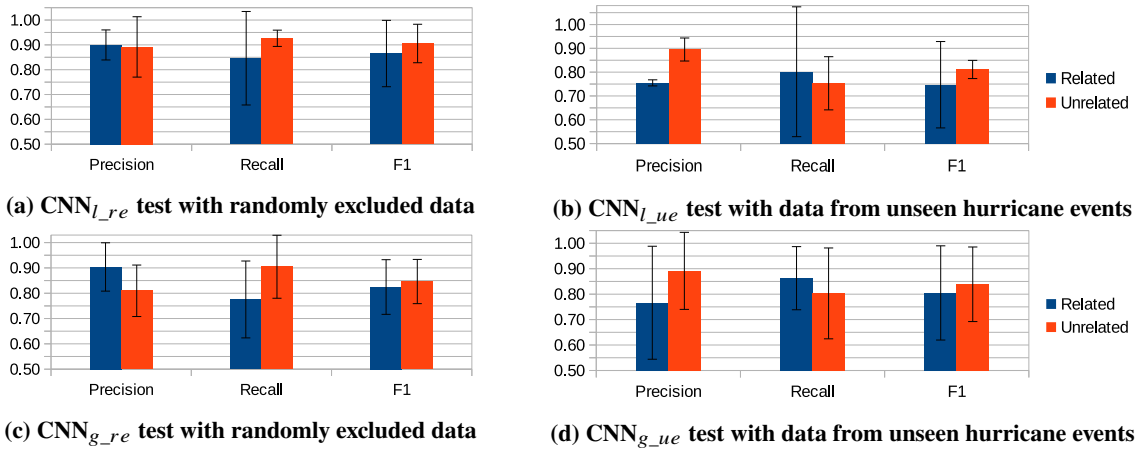
**Figure 2. Averaged results of local and global models tested with randomly excluded and new flood event data**

### Hurricane events: Global versus local models

The data sets listed in table 1 cover the following five different hurricane events: Sandy 2012 (*Epic*, ~ 20,000 tweets), Odile 2014 (*CrisisNLP*, ~ 2,000 tweets), Harvey 2017, Irma 2017, Maria 2017 (all from *CrisisMMD*, ~ 4,000 tweets per event). A total number of 100 models is trained and evaluated in the same manner as described above in this experiment. According to figure 3 and table 2, local and global models tested with randomly excluded data behave similarly to flood-related models. Testing with different unseen hurricane-related data sets results in F1 drops between 0.007 and 0.12. An intersting observation is that global models yield higher F1-scores than local models in case of unseen events.

### Local model transferability to other events

Even though local models tend to be more accurate and robust in our experiments, a practical application might be challenging. Models trained on specific event types would have to be applied to a data stream in parallel. The per-tweet model votes can then be used for filtering as well as a decision regarding the relatedness to specific




**(a) CNN$_{l_re}$ test with randomly excluded data**

**(b) CNN$_{l_ue}$ test with data from unseen hurricane events**

**(c) CNN$_{g_re}$ test with randomly excluded data**

**(d) CNN$_{g_ue}$ test with data from unseen hurricane events**

**Figure 3. Averaged results of local and global CNNs tested with randomly excluded and new hurricane event data**

event types. Even though this would provide valuable additional information, a decision cannot be made in case of ambiguous model votes. In this context, the cross-event performance of local flood and hurricane models is investigated in this experiment. Averaged results and standard deviations for all local models applied on other event types are shown in table 3. Compared to the results of local models applied to data from the same event type,

**Table 3. Scores for cross-event application of local flood (F) and hurricane (H) models**

| | *Related* | | | *Unrelated* | | |
|---|---|---|---|---|---|---|
| *Model→ Event* | *Precision* | *Recall* | *F1* | *Precision* | *Recall* | *F1* |
| *F→ H* | 0.907±0.124 | 0.491±0.146 | 0.629±0.139 | 0.710±0.081 | 0.975±0.017 | 0.819 ±0.050 |
| *F→ F* | 0.989±0.007 | 0.789±0.094 | 0.875±0.054 | 0.830±0.069 | 0.990±0.007 | 0.901 ±0.038 |
| *H→ F* | 0.809±0.101 | 0.785±0.165 | 0.779±0.092 | 0.807±0.106 | 0.791±0.113 | 0.789 ±0.062 |
| *H→ H* | 0.755±0.013 | 0.802±0.273 | 0.747±0.181 | 0.895±0.048 | 0.753±0.112 | 0.811 ±0.038 |

obtained F1-scores are up to 0.25 lower, where hurricane models applied to flood data perform better than vice versa. For *related* tweets, especially in case of flood models applied to hurricane data, the recall drops significantly.

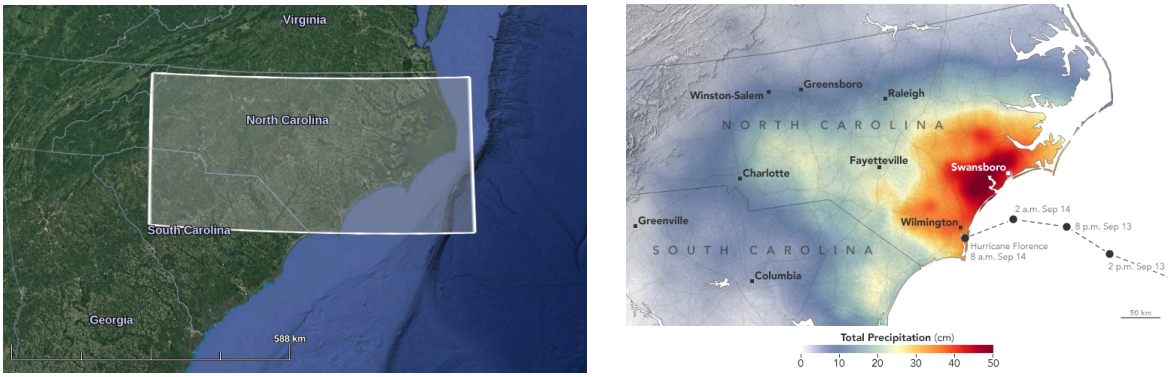
### Hurricane Florence: Global model transferability

Analyzing the tweet content without preliminary keyword-based filtering, as proposed in this work, implies a significantly higher occurence of unrelated tweets, which is usually not represented in labeled data sets. We therefore acquired $\sim 600,000$ tweets during hurricane Florence from September 12th to 19th, 2018, using the Twitter API. In order to avoid effects of keyword-based filtering, we applied location-based filtering by defining an area of interest (AOI) as shown in figure 4a. Forecast services estimated that this would be the mainly affected area for which heavy rainfall was expected as well (figure 4b).

Since only a fraction of all tweets is known to be geo-referenced, this approach might again introduce bias. However, we assume that sampling geo-referenced tweets provides a more representative subset of all tweets than sampling by keywords. Another advantage of this approach is that tweets from directly affected individuals are collected, whereas tweets from users who are not directly involved, but contribute to discussions are discarded.

The results of our global model applied to the Florence data allow for several further analyses. As a first result, the temporal distribution of *related* and *unrelated* tweets binned into intervals of two hours is shown in figure 5. A daily tweet activity pattern with maximum values of $\sim 10k - 12k$ until around midnight can be observed. 14.9 % of all tweets are classified as *related* to a disaster event. Due to the current lack of labels for this data set, the results are analyzed qualitatively in order to gain first insights (see discussion).

The same model is applied to a second data set acquired for the AOI in figure 4a during a non-disaster period (January 08-15, 2019). The corresponding histogram shown in figure 6 has a similar pattern as the one generated for the event period. During the disaster, $\sim 88,400$ tweets were identified as *related*, whereas in the second period around half of this amount was detected ($\sim 43,700$ tweets). With a total number of $\sim 526,000$ tweets recorded in January 2019, 8.3 % of these are classified as *related*. Selected examples are listed in table 4. The highest-ranked




**(a) Defined AOI** **(b) Rainfall and hurricane trajectory (Sep. 13-14)**

**Figure 4. Defined AOI (Google Maps 2018) and precipitation map (NASA Earth Observatory 2018)**

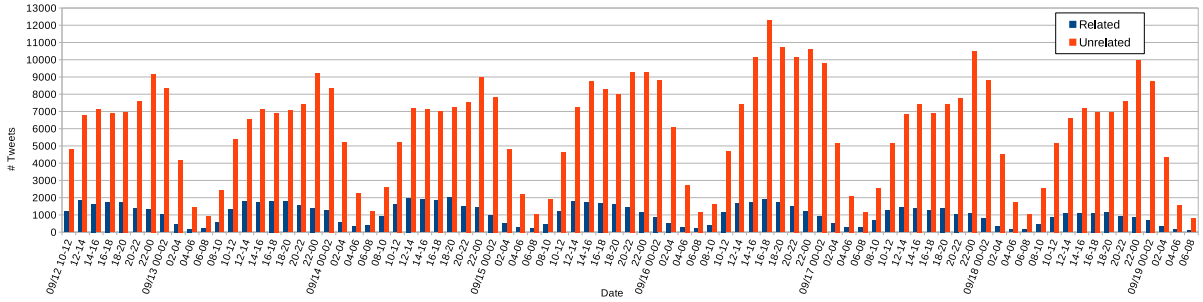



**Figure 5. Filtering results during hurricane Florence, September 12-19, 2018**

*related* tweets demonstrate that even though Florence seemed not to be a huge topic during January, other general crisis-related aspects as well as the heavy 2018 California wildfires are discussed. The *related* examples with a likelihood of 0.9 indicate that many *unrelated* tweets might be misclassified. Classification examples for class *unrelated* with softmax values between 1.0 and 0.5 in turn indicate a large likelihood that these are correctly classified.

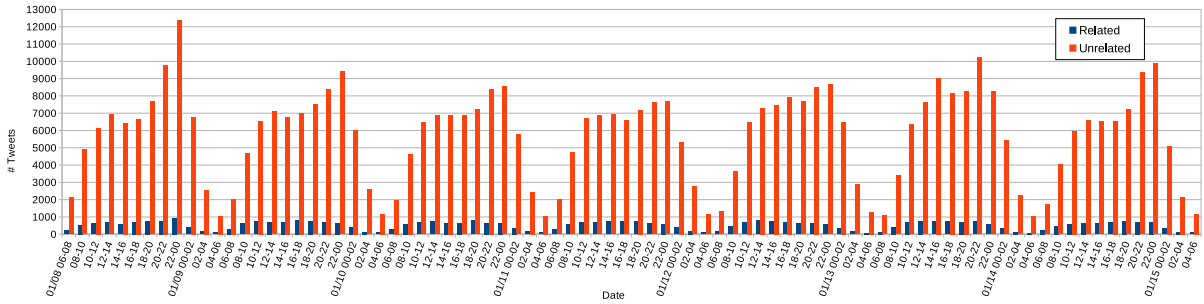



**Figure 6. Filtering results during a non-crisis period, January 08-15, 2019**

## DISCUSSION

In this study, our comprehensive and balanced data set leads to better results for tweet filtering compared to other studies, in which the same CNN architecture and similar hyperparameters were used. We choose a binary and objective classification problem without further semantic interpretations, like the informativeness or usefulness of related tweets. Taking into account that inter-annotator agreement of around 92 % was obtained for this definition in (Stowe et al. 2018), our F1-scores in a range of 75 – 92 % for local models as well as 80 – 86 % for global models are satisfying.

Event type-specific differences of model accuracies can partially be explained by the complexity of disaster types. The fact that flood events represent one of several different sub-event types typically occuring during or after hurricanes might be the main reason for lower scores and higher standard deviations for hurricane models.

**Table 4. Selected filtering results for the non-crisis data set, January 2019**

| | *Classified as related* | |
|---|---|---|
| *Rank* | *Tweet text* | *Softmax* |
| 1 | @userID I think the fact he's threatened to do it for several weeks, negates an emergency. Fires, hurricanes, storms, climate change, threat to democracy, stopping russian influence.....those are emergencies | |
| 2 | @userID FEMA helps victims of fires. You don't know anything about forest fires. | |
| 3 | #hash #hash #hash Trump Halts FEMA Funds for Calif. Wildfires, Forest Management URL #hash via @userID | 1.0 |
| 4 | How's the beaches in Puerto Rico Chuck? | |
| 5 | After the fire: Helping firefighters cope with tragedy. URL | |
| ... | ... | ... |
| 4946 | The golf walls are taking shape. URL | |
| 4947 | How is it getting through TSA? I kinda wanna get the agents like a card or something lol | |
| 4948 | @userID @userID Okay for Comey to travel to all the FBI offices, but Whitaker going to USA offices? Such a low priority for Nadler | 0.9 |
| 4949 | @userID So the Democrats are concerned about the government being re-opened and the 800K plus ancillary people also impacted - can get back to work and paid. Meanwhile, Baby Trump / GOP are only concerned with his 'wall'. Use the damned money already available for border security. | |
| 4950 | Council is hearing an update about installing sidewalks on Oxford Road. We've got a couple of folks who aren't happy about the idea. This was first brought up in 2015. #hash URL | |
| | *Classified as unrelated* | |
| *Rank* | *Tweet text* | *Softmax* |
| 1 | Loyalty Loyalty Loyalty Loyalty Loyalty Loyalty Loyalty ... That's all I ask for | |
| 2 | @userID @userID Total manure! Dynasty? Playoffs? Playoffs? Dynasty? | |
| 3 | *sings* loyalty loyalty loyalty | 1.0 |
| 4 | MVP MVP MVP MVP MVP MVP [EMO] | |
| 5 | POLLS ....??....POLLS ...? or PLAYOFFS ?...SMH URL | |
| ... | ... | ... |
| 482201 | Slow traffic in #hash on I-40 WB between Lk Wheeler Rd - Exit 297 and Gorman St - Exit 295, accident at Gorman St - Exit 295 #hash | |
| 482202 | When @userID comes on.....you be on some other Shut [EMO] #hash [EMO] #hash #hash #hash #hash @ Burlington, North Carolina URL | 0.5 |
| 482203 | EXCUSE ME?!? Why is blue hair hitting like this? | |
| 482204 | To2 w/ @userID for CWL Fort Worth be on to play rn Under @userID Have 600+ pps. We need a Main ICR, 2 Saugs. Be good don't waist our time!!!!! Rt if you the homie [EMO] | |
| 482205 | I'm at @userID Home Improvement in Cary, NC URL | |

According to figures 2a-2c, the flood-related local and global models show a general accuracy pattern, in which recall is lower than precision for class *related* and vice versa for class *unrelated*. A high precision for *related* tweets indicates a low percentage of *unrelated* misclassified tweets. A lower recall for *related* tweets indicates that misclassifications for *related* tweets are more likely. An interesting observation is that the relations are intensified for local models tested with unseen event data, i.e., precision and recall for *related* are around 0.99 and 0.80, respectively, and vice versa for *unrelated*. F1 standard deviations of $0.03 - 0.05$ for the local flood models indicate a high robustness. Furthermore, the averaged scores are higher compared to the baseline model.

For the corresponding hurricane-related models (figures 3a-3c, table 2), other relations as well as lower scores and higher standard deviations can be observed, where the latter two can be explained by the aforementioned heterogeinity of possible sub-events in case of hurricanes. Compared to a baseline value of 0.84 for precision, recall and F1, corresponding average values of 0.90, 0.81 and 0.85 for models tested with randomly excluded data and 0.77, 0.83 and 0.78 for models tested with data from new hurricane events were obtained. This indicates a good performance if test data from the same event is used and slightly worse results for unknown hurricane events. Class-related patterns for precision and recall can be observed with respect to the data used for testing. Whereas the pattern is similar as described for flood models, a difference is observed for testing local models with new events, i.e., a lower precision and higher recall for *related*.

For the global models tested with unseen event data (figures 2d and 3d, table 2), the same pattern with lower precision and higher recall for *related* can be observed. This means that we can expect more false negatives and fewer false positives. A reason for this similar trend in both global models might be the fact that training is done using a large set of tweets from various event types, where only a very small fraction (a single new event) is used for testing. As expected and similar to results, for example reported in (Imran, Elbassuoni, et al. 2013; Win and Aung 2017), a drop of accuracies in case of new events can be observed. However, F1-scores between 0.81 and 0.86 indicate good accuracy compared to our baseline. Note that the scores discussed here were not obtained by testing with data excluded during CV, but by using data from completely different events. With respect to the 92 %

inter-annotator agreement reported in (Stowe et al. 2018), these results are quite satisfactory, even though higher performance variations can be expected with increasing event complexity.

Hypothetically, the application of local models to data from other event types should result in lower scores. The results in table 3 partially show a different behavior. Applying flood models to hurricane data results in 0.49 recall for *related*, indicating a higher complexity of hurricane events covering various sub-events. On the other hand, a high recall of around 0.78 in case of hurricane models applied on flood data reflects the fact that floods are sub-events of hurricanes. We therefore apply a global model trained with our proposed balanced data set to the Florence data.

Regarding the Florence data, we investigate whether our global models tend to provide the same accuracies as reported in the preceding experiments. According to figures 2d and 3d and table 2, we expect more false negatives than false positives as well as F1-scores of $0.80 \pm 0.19$ and better. This should result in a high percentage of correctly classified *unrelated* tweets. The $\sim 502,000$ tweets recorded during Hurricane florence that are classified as *unrelated* are therefore quantitatively analyzed by searching for crisis-related keywords (e.g. flood, hurricane, thunderstorm, tornado, and florence), as well as by manually evaluating random samples. A small fraction of $\sim 10,000$ wrongly classified tweets identified with this approach provides a hint that the model tends to behave as expected. A further indicator for this is the observation of a significantly higher percentage of *unrelated* tweets misclassified as *related*. A thorough review of randomly sampled tweets classified as *related* indicates a strong relationship between the CNN softmax outputs and misclassifications. In case of softmax values between $0.6 - 1.0$, the percentage of wrong classifications seems to be low and increases with decreasing softmax values below 0.6. The results obtained for the second data set from 2019 provide further evidence for these trends.

## CONCLUSIONS AND OUTLOOK

In this work, the task of filtering crisis-related tweets is investigated. A strong dependency on chosen keywords for filtering motivates an approach for analyzing full tweet content with CNNs. Since this is intended to be the first analysis step for several crisis-related applications, a binary classification (*related* vs. *unrelated*) is used here.

In order to cover a broad range of different crisis event types as well as to overcome the often-observed problem of imbalanced training data, a new data set is compiled based on existing sets. Compared to other works, in which the same CNN was utilized, we obtain robust and even better results. In terms of model transferability to six flood and five hurricane events, our event-specific models are able to obtain F1-scores of $0.75 - 0.90$. However, since the usage of models for single event types may be ambiguous in case of similar events (e.g. floods often occur during hurricanes), we recommend to use models covering a broad range of event types. For new hurricane and flood events, an average F1-score of $0.83 \pm 0.13$ is obtained for these models.

However, our main focus is the application of models to real data streams. We are therefore currently investigating if the model behavior for data representing a realistic Twitter stream is similar to our experimental results reported in this work. First qualitative results with with a data set recorded during hurricane Florence indicate strong similarities. For a non-crisis period, only 8.3 % of all tweets were classified as *related*, further confirming the promising results.

In future works, we plan to (partially) label the Florence data set for further quantitative analyses. With the goal of model enhancement in case of new events, methods for domain adaptation, such as those proposed by (Li et al. 2018) for NB classifiers or online deep learning as proposed in (Nguyen, Joty, et al. 2016) could be further investigated.

## REFERENCES

Alam, F., Ofli, F., and Imran, M. (2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*. Stanford, CA, USA.

Burel, G. and Alani, H. (May 2018). "Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, p. 12.

Burel, G., Saif, H., Fernandez, M., and Alani, H. (May 2017). "On Semantics and Deep Learning for Event Detection in Crisis Situations". In: *Workshop on Semantic Deep Learning (SemDeep), at ESWC 2017*.

Burton, S. H., Tanner, K. W., Giraud-Carrier, C. G., West, J. H., and Barnes, M. D. (2012). "'Right Time, Right Place' Health Communication on Twitter: Value and Accuracy of Location Information". In: *Journal of Medical Internet Research* 14.6:e156.

Google Maps (2018). *V 7.3.2.5495, North Carolina, USA, lat 35.25, lon -78.91, eye alt 705 feet. SIO, NOAA, U.S. Navy, NGA, GEBCO. Landsat, Copernicus, Google 2018, accessed at December 19, 2018*.

Gruender-Fahrer, S., Schlaf, A., Wiedemann, G., and Heyer, G. (2018). "Topics and topical phases in German social media communication during a disaster". In: *Natural Language Engineering* 24.02, pp. 221–264.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Practical extraction of disaster-relevant information from social media". In: *Proceedings of the 22nd International Conference on World Wide Web (WWW) Companion*. Rio de Janeiro, Brazil: ACM Press, pp. 1021–1024.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Computing Surveys* 47.4, pp. 1–38.

Imran, M., Mitra, P., and Castillo, C. (May 2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia: European Language Resources Association (ELRA).

Kim, Y. (2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1746–1751.

Kumar, S., Barbier, G., Abbasi, M. A., and Liu, H. (2011). "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief". In: *Proceedings of the 5th International AAAI Conference an Weblogs Social Media (ICWSM)*, pp. 661–662.

Landwehr, P. M. and Carley, K. M. (2014). "Social Media in Disaster Relief - Usage Patterns, Data Mining Tools, and Current Research Directions". In: *Data Mining and Knowledge Discovery for Big Data*. Ed. by W. W. Chu. Vol. 1. Berlin, Heidelberg: Springer, pp. 225–257.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (Mar. 2018). "Disaster response aided by tweet classification with a domain adaptation approach". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 16–27.

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). "Building a Large-scale Corpus for Evaluating Event Detection on Twitter". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)*. San Francisco, California, USA: ACM, pp. 409–418.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. arXiv: 1301.3781.

NASA Earth Observatory (2018). https://earthobservatory.nasa.gov/images/92775/florence-inundates-the-carolinas, accessed: December 19, 2018.

Nguyen, T. D., Al-Mannai, K. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (Jan. 2017). "Robust classification of crisis-related data on social networks using convolutional neural networks". In: *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*. AAAI press, pp. 632–635.

Nguyen, T. D., Joty, S. R., Imran, M., Sajjad, H., and Mitra, P. (2016). "Applications of Online Deep Learning for Crisis Response Using Social Media Information". In: *CoRR* abs/1610.01030.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *Proceedings of the 8th International AAAI Conference on Web and Social Media*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. Vancouver, BC, Canada: ACM, pp. 994–1009.

Parilla-Ferrer, B. E., Fernandez, P., and T. Ballena IV, J. (Dec. 2014). "Automatic Classification of Disaster-Related Tweets". In: *International conference on Innovative Engineering Technologies (ICIET)*. Bangkok, Thailand.

Periñán-Pascual, C. and Arcas-Túnez, F. (June 2018). "The Analysis of Tweets to Detect Natural Hazards". In: *Intelligent Environments 2018 - Workshop Proceedings of the 14th International Conference on Intelligent Environments (IE), Rome, Italy, 25-28 June 2018*, pp. 87–96.

Ramachandran, D. and Ramasubramanian, P. (2018). "Event detection from Twitter - a survey". In: *Int. J. of Web Information Systems* 14.3, pp. 262–280.

Reuter, C. and Kaufhold, M.-A. (2018). "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 41–57.

Stieglitz, S., Mirbabaie, M., Ross, B., and Neuberger, C. (2018). "Social media analytics - Challenges in topic discovery, data collection, and data preparation". In: *Int. J. of Information Management* 39, pp. 156–168.

Stowe, K., Palmer, M., Anderson, J., Kogan, M., Palen, L., Anderson, K. M., Morss, R., Demuth, J., and Lazrus, H. (2018). "Developing and Evaluating Annotation Procedures for Twitter Data during Hazard Events". In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 133–143.

Win, S. S. M. and Aung, T. N. (May 2017). "Target oriented tweets monitoring system during natural disasters". In: *Proceedings of the IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 143–148.