Deep Learning Approach towards Multi-label Classification of Crisis Related Tweets

Alan Aipe
IIT Patna
alan.me14@iitp.ac.in

Mukuntha N S
IIT Patna
mukuntha.cs16@iitp.ac.in

Asif Ekbal
IIT Patna
asif@iitp.ac.in

Sadao Kurohashi
Kyoto University
kuro@i.kyoto-u.ac.jp

ABSTRACT

Micro-blogging sites like Twitter, over the last decade, have evolved into a proactive communication channel during mass convergence and emergency events, especially in crisis stricken scenarios. Extracting multiple levels of information associated with the overwhelming amount of social media data generated during such situations remains a great challenge to disaster-affected communities and professional emergency responders. This valuable data, segregated into different informative categories, can be leveraged by government agencies, humanitarian communities as well as citizens to bring about faster response in areas of necessity. In this paper, we address the above scenario by developing a deep Convolutional Neural Network (CNN) for multi-label classification of crisis related tweets. We augment the deep CNN with several linguistic features extracted from tweets, and investigate their usages in classification. Evaluation on a benchmark dataset shows that our proposed approach attains the state-of-the-art performance.

Keywords

Deep learning, Multi-label classification, Social media, Crisis response

INTRODUCTION

Emergency situations bring along with them innumerable challenges to professional humanitarian communities, various government agencies as well as the citizens, and demand fast and effective decision-making capabilities from the response authorities. There has been phenomenal growth in the social media information available during the last few years. Numerous platforms have been introduced, and many new ones are constantly being created. This has created an opportunity to build socially intelligent systems exploiting the large volume of textual contents generated through social media platforms. Effective communication at different layers, for e.g. from person to person, person to government agencies, and government to people etc. can be efficiently managed through these various information sources. Information gathered from Twitter, one of the most popular social media platforms, can be leveraged as an aid to improve overall situational awareness, thereby taking informed decisions (Castillo, 2016; Vieweg, Castillo, et al., 2014).

Extracting valuable insights from a large chunk of twitter data tends to be challenging and complex owing to the character limit and informality of the shared information. Users often employ abbreviations, spelling variations and colloquial words while using Twitter platform for sharing information (Han et al., 2013). Despite advances in natural language processing (NLP), interpreting the semantics of short informal texts remains itself a hard problem. Here we propose an efficient multi-label classification model, which can be generalized to multiple crisis situations. This would prove highly beneficial to various sectors (citizens, humanitarian communities, government agencies etc.) as an initial step towards harnessing the information flow during crisis situations. In recent times, deep learning models have shown tremendous success in solving several problems in image processing, computer vision and natural language processing. In this paper we propose a deep learning model augmented with domain-specific features for classifying the crisis-specific twitter data.

Problem Definition

The main objective of this paper is to propose a scalable deep learning architecture for the multi-label classification of disaster specific tweets into discrete informative categories. As discussed in Dembczyński et al. (2012), a straightforward approach for multi-label classification is the Binary Relevance (BR) method. Here one binary classifier is trained for each label and used to predict whether, for a given test instance, this label is present (relevant) or not.

For better understanding, the goal, as mentioned above, can be modularized into a set of research questions. By combining qualitative, quantitative analysis and sleek architecture design, we seek to answer the following questions.

RQ1: Can we develop a deep binary relevance classifier for multi-label classification, which performs well in multiple crisis domains, using tweet text as the only input to the model?

RQ2: What kind of enhancements do twitter-centric features like hashtags, user-mentions and keywords extracted from embedded URLs bring to the model?

By answering the above mentioned questions, we intend to create a system which can classify a given crisis specific tweet into one or more informative categories. For example, “Death toll rises to 112. I pray for the affected families. #NapaEarthquake” should be classified into *Casualties category* as well as *Sympathy* and *Emotion* category. Hence, the problem is posed as a multi-label classification problem.

Motivation

Due to its growing ubiquity, rapidity of communication, and cross-platform accessibility, Twitter is increasingly being considered as a means for emergency communication during and after crisis related events. During a sudden outburst of emergency, the affected people tend to share information about personal whereabouts and the general conditions of their immediate environment, in anticipation of quick humanitarian response. Identification and classification of such tweets targeted at disaster events would help the crisis-response community in deep analysis of escalating situations and in efficient allocation of available resources, as discussed in Imran, Castillo, Diaz, et al. (2015). Classification enables more effective re-routing of necessary information to the concerned specialized units on a real-time basis without having to stroll through the huge collection.

Most of the existing research works carried out in this area treat the above mentioned problem as a binary or multi-class classification problem (Nguyen, S. R. Joty, et al., 2016; Imran, Mitra, and Castillo, 2016). Necessity for a multi-label classification model arises owing to the observation that a majority of tweet instances do actually belong to multiple classes of information. People often mix personal and informative contents pertaining to multiple categories while sharing messages over social media platforms. For example, “I would like to donate $100 to #NapaEarthquake relief fund. Around 1k casualties reported ☹” contains information regarding both *donation* and *reported casualties*. Thus, a multi-label classifier would assist in better and accurate information extraction as well. Finding relevant features that help in an online setting, and improving the model's ability to generalize to multiple crisis domains are of utmost importance. With the recent success of deep neural networks, we hypothesize that general relationships (e.g. lexical, syntactic as well semantic relationships) in this complex domain can be effectively captured by these models.

Contributions

In this paper, we propose a Convolutional Neural Network (CNN) based deep learning architecture for multi-label classification of disaster related tweets into seven distinct informative categories. To the best of our knowledge, no prior research work has dealt with deep neural network (DNN) based multi-label classifier in the crisis response domain. Therefore, we intend to open this niche with this paper. Another important characteristic of our proposed approach is that we investigate the effect of twitter specific linguistic knowledge to the deep CNN.

The proposed model is generic in the sense that we do not necessarily train and evaluate the model with the datasets of similar domains. Rather we train and evaluate our proposed approach over a single dataset containing tweets shared during crises belonging to multiple domains (for example, earthquakes, floods, typhoons etc.). Experimental results, recorded in subsequent sections, prove that the model generalizes satisfactorily to multiple disaster domains, which share the same information taxonomy. In an online setting, the model yileds as output an evolving set of hashtags and user-mentions, befitting the disaster, which can be fed back into the tweet collecting module (for example, AIDR(Imran, Castillo, Lucas, et al., 2014) to fetch disaster related tweets with higher accuracy.

RELATED WORKS

Classification of disaster related tweets and accurate Information Extraction from it have gained significant attention to the government and other stakeholders due to the growing need for building an automated crisis response system. The recent advancement in Natural Language Processing, Machine Learning and Data Mining techniques have facilitated this process.

Here we present a very brief summary of the most relevant research works carried out in this domain. Vieweg, Hughes, et al. (2010) proposed a methodology and disaster ontology to identify tweets that provide situational awareness. Kongthon et al. (2012) developed classifiers to analyze tweets from the Thailand floods in 2011. Purohit and Sheth (2013) developed Twitris with semantic enrichment, classification and geotagging as its main capabilities. Robinson et al. (2013) visualized heat-maps based on geotags in order to extract valuable information. Caragea, McNeese, et al. (2011) used twitter and SMS data generated during crisis events for the classification purpose. Shallow machine learning approaches towards multi-class classification of disaster-specific tweets were proposed in Imran, Mitra, and Castillo (2016) and Imran, Elbassuoni, et al. (2013). They trained three different models, namely Support Vector machine (SVM), Naive Bayes (NB) and Random Forest (RF), for the classification of tweets into 9 different informative categories, namely *Caution and Advice, Displaced people and evacuations, Donation needs or offers, Infrastructure and Utilities damage, Injured or dead people, Missing or trapped or found people, Sympathy or emotional support, Other useful information* and *Irrelevant case*. A naive Bayes classifier for cross-lingual crisis domain adaption was explored by Imran, Mitra, and Srivastava. (2016). A multi-label naive-Bayes classifier has been proposed in Imran, Elbassuoni, et al. (2013) to extract the most relevant information. The survey as presented in Imran, Castillo, Diaz, et al. (2015) provides a very informative description of various crisis response and analysis systems developed to address information flow during emergency situations.

Prior research work focusing on application of deep learning in NLP has shown that neural networks mimic biological information processing and thus outperform shallow ML models in grasping general and complex relationship existing in various datasets. A unified Deep Neural Network (DNN) architecture for solving various NLP tasks including part-of-speech tagging, chunking, named entity recognition and semantic role labeling was presented in Collobert et al. (2011) and S. Joty and Hoque (2016). In case of crisis domain, Caragea, Silvescu, et al. (2016) used convolutional neural network (CNN) to segregate informative and non-informative messages during disasters. A MLP-CNN based multi-class classification approach to classify tweets into different informative classes was proposed by Nguyen, S. R. Joty, et al. (2016) and Nguyen, Al-Mannai, et al. (2017). In our current work we propose a multi-label deep learning model for classifying disaster-specific tweets, and to the best of our knowledge this is the very first attempt in this direction.

PROPOSED METHOD

In this section we describe our proposed approach for multi-label classification of disaster specific tweets.

Deep CNN based Architecture

At first we describe the preliminary architecture for deep CNN based multi-label classification of disaster related tweets with respect to the 7-class taxonomy (as discussed in Section Crisis Information Taxonomy). This addresses our first research question (i.e. **RQ1** of Problem Definition Section). An elevated view of the overall architecture is shown in Figure 1. The overall system consists of seven similar classifiers (represented as B_i), each of which takes in a similar input and produces an unique binary output, predicting whether or not a particular label is relevant.

The seven binary outputs are independent to each other, and each output corresponds to one of the seven categories associated with the information taxonomy used in this paper. Moreover, inter-label dependency between the informative categories varies significantly with the type of crisis under consideration. For example, there is a very high inter-label dependency between 'Casualties and Public Impact' and 'Infrastructure Damage' in case of physical disasters like earthquakes while negligible dependency exists in case of biological crises like epidemics.

Our proposed architecture is more generic in the sense that rather than training our model on any specific type of disaster related tweets, we make use of a dataset containing tweets of multiple crises. We thereafter train seven separate neural network classifiers (for each category), rather than a single neural network with seven output neurons, so that each classifier learns features corresponding to only its category ignoring inter-label dependency.

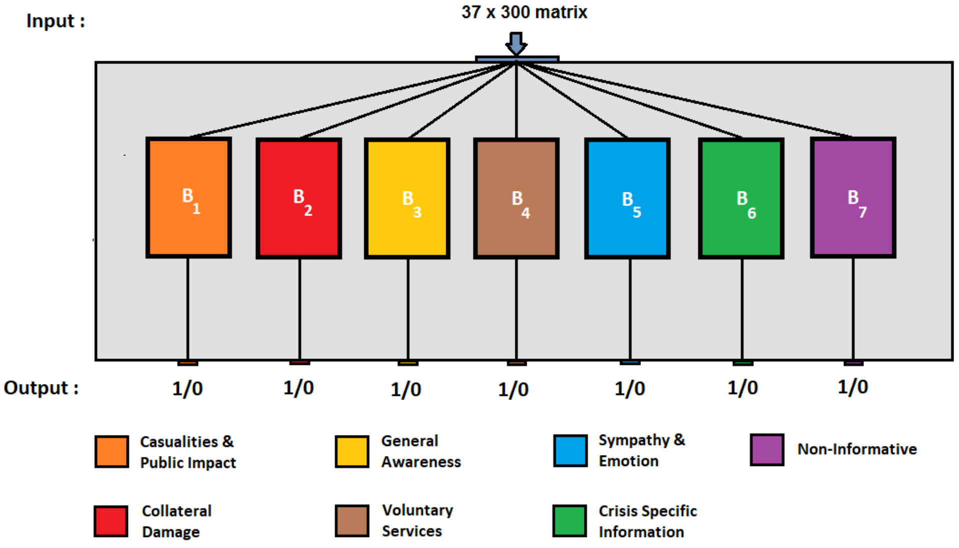

Figure 1. An overview of the proposed architecture

The seven classifiers, represented as B_i in Figure 1, have identical neural net design as shown in Figure 2, but vary in their hyper-parameter values. The CNN is chosen for this classification problem owing to its ability to use distributed representation of words as well as learn key features automatically at different levels of abstractions. The matrix representation of tweet text (as discussed in the Dataset and Experimental Setup Section), generated using pre-trained word2vec model (Mikolov et al., 2013), is fed into a convolutional layer generating 250 feature maps of kernel or filter sizes {2, 3, 4}. These feature maps are max-pooled, followed by global max-pooling and concatenation to be inputted into a fully connected dense layer which mimics auto-encoder neural net properties. All neurons until this level have Rectified Linear Unit (ReLU) as their activation function and corresponding weights are initialized using uniform random distribution. Output from the final layer neurons of auto-encoder network is then fed into a single neuron with sigmoid as activation function. In order to avoid overfitting, dropout probabilities are applied to the final three layers of the model.

Augmenting Twitter-Centric Linguistic Features to CNN

We propose a modified version of the baseline model (i.e. the CNN as described above) by augmenting twitter-centric features to the deep CNN. We closely analyze the influence of these features on the overall system performance. We explore the features based on hashtags, user-mentions and keywords extracted from the embedded URLs in the tweet message (addressing the research question **RQ2**). All these features are fed into the first fully connected layer as shown in *Figure 3*. It is found that integration of twitter-centric textual features, indeed, has a positive influence on the performance of the classifier (discussed in more details in Table 4). Hence, our final architecture comprises of seven classifiers, corresponding to the 7-class taxonomy, integrated with those combinations of twitter-centric features which resulted in the highest Area Under ROC curve (AUC) values during 5-fold cross validation evaluation.

Hashtags: A hashtag is a keyword or a phrase used in a Tweet message to describe a topic or a theme. This is denoted by a special character '#' at the beginning of a word. First, an unified set of hashtags collected from all the tweets belonging to a specific label from the training set is created (i.e., 7 sets, each pertaining to a specific label). Thereafter, we compute the Jaccard similarity between the set of hashtags present in each tweet and the set of hashtags contained in the category-specific unified set. This is then fed as an input to the first fully connected layer of the corresponding classifier as shown in Figure 3. The Jaccard Similarity J for any two sets A and B is given by the equation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

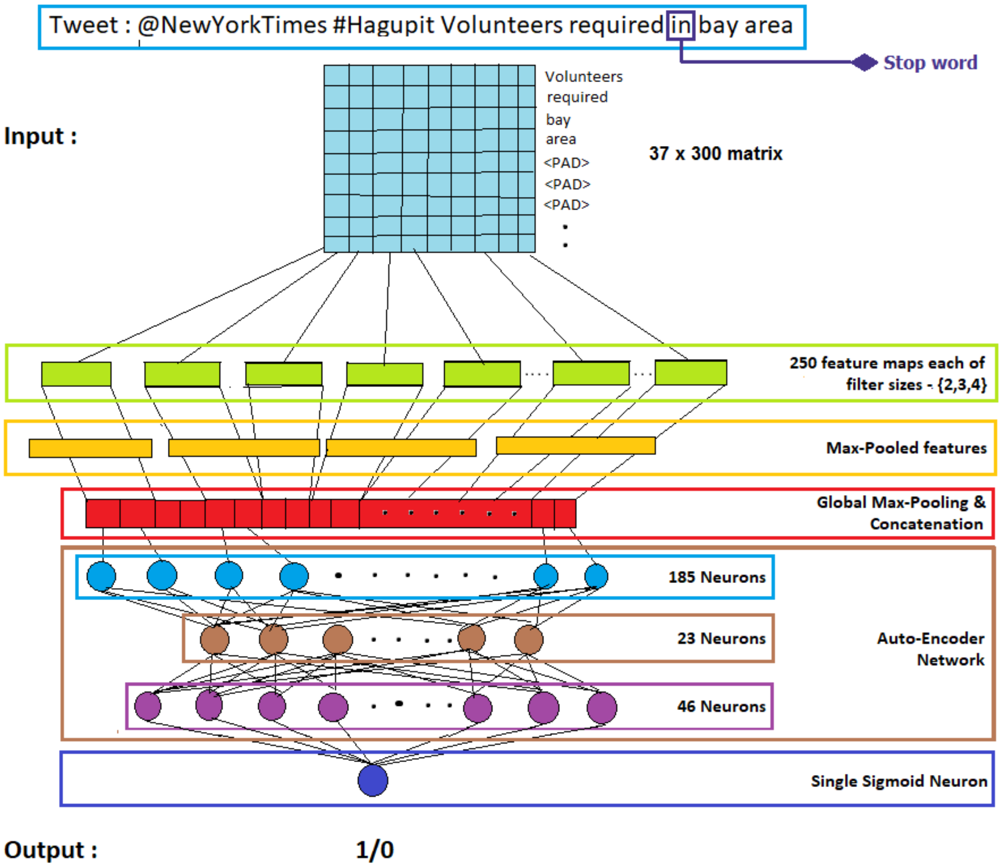

Figure 2. Deep CNN based architecture design (baseline) of classifier ($\mathbf{B}_i$)

User-mentions: An user-mention corresponds to a word used to refer to a user account in a tweet, identified by the use of special character '@' at the beginning of the word. Like with hashtags, the Jaccard similarity between the set of user-mentions present in each tweet and the unified set of user-mentions for that particular label from the training data is fed as an input to the first fully connected layer as shown in Figure 3.

Keywords extracted from Embedded URLs: The top six keywords, having the highest Tf-Idf values, are extracted from the content of the webpage (Content refers to the text inside <p> tags, the webpage title and the meta description) pointed to by the embedded URLs in a tweet. Similar to the other features, Jaccard similarity between the set of extracted keywords and the global set corresponding to that particular label is fed into the first fully connected layer of the network as shown in Figure 3.

Online Setting: The proposed architecture, if deployed in an online setting, would generate an evolving set of hashtags, user-mentions and URL keywords over the time for each category. The union of all these can be fed back into the tweet collecting module for better efficiency of disaster-specific data retrieval.

DATASET AND EXPERIMENTAL SETUP

In this section we present the datasets, taxonomy and the various setups that we use for the evaluation.

Tweet Dataset

The dataset used for our research was obtained from CrisisNLP[1]. It consists of of 49,143 tweet IDs, each corresponding to a crisis-related message from Twitter posted during 14 different crises that occurred from 2013 to 2015. Each tweet was annotated by the paid CrowdFlower[2] workers or volunteers, with a label corresponding to a category belonging to the subset of annotations used by the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA). A category is assigned to a tweet if and only if at least three different annotators agree on that particular label. **Table 1** shows the distribution of tweet instances corresponding to each crisis type.

[1] http://crisisnlp.qcri.org/lrec2016/lrec2016.html
[2] https://www.crowdflower.com/

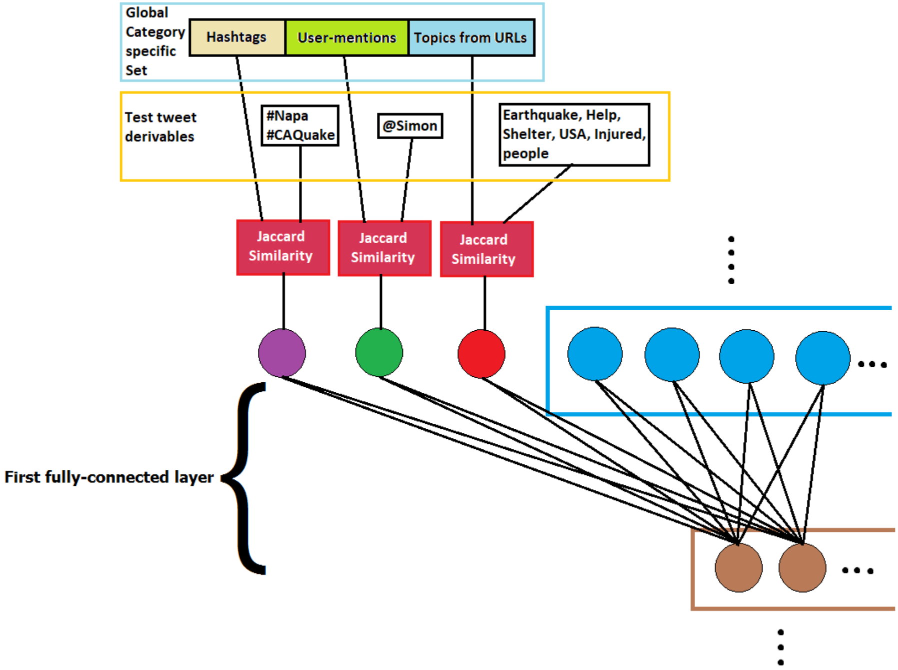

Figure 3. Proposed deep CNN Architecture with augmented twitter-centric features

Crisis Name	X	Y
California Earthquake	2014	184
Chile Earthquake	2014	441
Hurricane Odile	2014	184
Iceland Volcano Eruption	0	417
Airline MH370 crisis	0	135
MERS	2018	383
Typhoon Hagupit	2014	9676
Cyclone Pam	2014	601
Nepal Earthquake (2015)	3019	9472
Landslide Worldwide	0	4493
Pakistan Flood	2014	0
India Flood	2004	0
Ebola Virus Epidemic	2018	0
Pakistan Earthquake (2013)	2014	0
Total	**23157**	**25986**

Table 1. Number of Crowdflower annotated tweets (X) and Volunteer annotated tweets (Y), corresponding to 14 different disasters that happened between 2013 and 2015, present in the original single-label CrisisNLP dataset.

Crisis Information Taxonomy

Our analysis of the dataset provides us more insight about the type of information that flows during crisis situations. The labeling instructions for annotation of the tweets vary with respect to the crisis under consideration. This implies that the guidelines for a given category which is informative to one crisis may be insignificant to the another. For example, while the category *Infrastructure Damage* is significant in case of geophysical disasters like earthquakes or tsunamis it is insignificant to a biological crisis like an epidemic. Similarly, a category named *Disease Prevention* is significant in biological scenarios whereas it is not applicable to geophysical situations. The taxonomy that we use for classification of disaster tweets comprises of 7 classes which are largely derived from the one reported in Imran, Mitra, and Castillo (2016). All the classes from the CrisisNLP dataset are mapped to one of these seven labels. The crisis information taxonomy used in this paper is as follows:

- *Casualties and Public Impact*: This denotes to the reports of affected people in the form of death, injury, displacement etc. This is the class that we obtain by merging the classes *Injured or dead people*, *Missing*

trapped or found people and *Displaced people and evacuations* in the original dataset. These classes were merged as (i). the number of instances for the *Missing trapped or found people* classs were relatively low, and (ii). the ambiguity betwen these two classe was very high-with almost all containing the similar contents.
Example:- 25 people dead in the hurricane #latestReports #BBC

- *Collateral Damages*: This corresponds to the reports of indirect or utility damages caused by the crisis. This class is derived from *Infrastructure and utilities damage* in the original dataset.
Example:- Bridge collapses due to 7.2 richter scale #napaquake

- *General Awareness*: This class denotes the tweets regarding public notices or situational awareness. This is created from the *Caution and advice* class of the original dataset.
Example:- Rescue missions are continuing at the affected areas. Everyone is advised to stay calm.

- *Voluntary Services* : This class denotes to the tweets showing willingness to donate services like money, food, clothes, manpower etc. This class is derived from the *Donation needs or offers* of the original dataset.
Example:- I would like to donate $1000 to #NapaReliefFund.

- *Sympathy and Emotion*: This particular class denotes to the tweets showing emotional support.
Example:- My prayers are with families of ten fellow beings who lost their lives at bay area today.

- *Crisis-specific Information*: This class denotes the tweets corresponding to crisis-specific categories which cannot be included in any of above categories. This is derived from the *Other useful information* class of the original dataset.
Example:- Palmer Tech at the verge of launching medicine which can fight ebola #Epidemic

- *Non-informative* : Tweets which do not impart any significant information.
Example:- Chilling out in the hurricane stricken beach.. #GonnaEnjoyToday

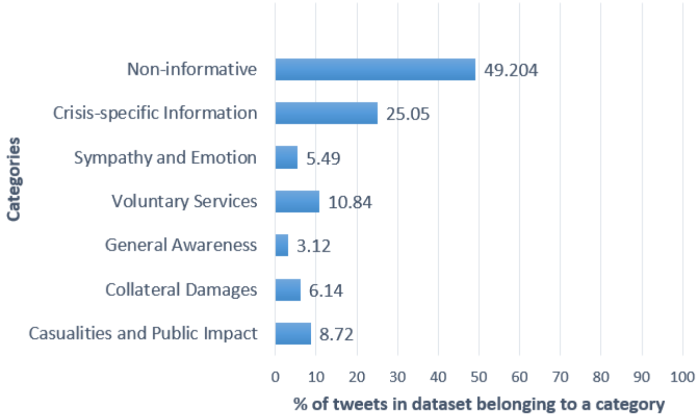

Figure 4. Class distribution of tweets in the original single label CrisisNLP dataset with respect to the 7-class taxonomy system.

We show the class-wise distribution of instances in Figure 4.

Word Embeddings

Capturing semantic similarity between target texts is an important step towards accurate classification. For this reason, word embeddings play a pivotal role. We use three different 300 dimensional pre-trained word2vec (Mikolov et al., 2013) word embeddings in our current work – CrisisNLP word embedding[1] trained over 52 million tweet messages provided by the CrisisNLP team, Google News word embedding[3] trained by the skip-gram model on part of Google news dataset containing about 100 billion words and Wiki fastText word embedding[4]. The methodology and rationale behind using three different word embeddings is further discussed in Tweet Vectorization.

[3] https://github.com/mmihaltz/word2vec-GoogleNews-vectors
[4] https://s3-us-west-1.amazonaws.com/fasttext-vectors/wiki.en.vec

Tweet Text Normalization

Use of colloquial and abbreviated forms of words while messaging in social media platforms, often poses a challenge in automated tex analysis. An Out-Of-Vocabulary (OOV) word dictionary contains the mapping between the OOV words and its corresponding normalized wordforms. We use the dictionary as reported in Imran, Mitra, and Castillo (2016), which contains 1,415 pairs of OOV wordforms.

Tools used for Experiments

The codebase, during experimentation, is written in Python (version 3.6) with external libraries – namely *keras*[5] for neural network design, *sklearn*[6] for evaluation of baseline and the proposed model, *pandas*[7] for easier access of data in the form of tables (or, data frames) during execution, *twython*[8] for fetching tweets corresponding to IDs in CrisisNLP dataset, *NLTK*[9] for textual analysis and *pickle*[10] for saving and retrieving input-output of different modules from the secondary storage devices.

Preprocessing

The class distribution of the collected data from the original CrisisNLP dataset mapped onto our 7-class taxonomy, is shown in *Figure 4*. The pre-processing phase comprises of the removal of non-ASCII characters, stop words and handling of non-alphanumeric characters followed by tokenization. Tokens of size (number of characters) less than 3 were removed with an intution that these do not contribute significantly in classification. Tweet text normalization was carried out using an OOV dictionary. Out-of-vocabulary terms such as slangs, place names, abbreviations, misspellings, etc. were replaced with their corrections or normalized forms, using the OOV dictionary. Tweets with identical tokens were combined together so as to form a multi-label k-hot encoding corresponding to each tweet. Many informative tweets were represented by re-tweets as well in the dataset, and had been labeled multiple times. It was observed that several of these tweets therefore contained multiple labels assigned to themall being the valid labels. This phase resulted in 27,150 data instances with the corresponding k-hot vector. Shrinkage of the number of data instances from 49,143 to 27,150 points to the dominance of multi-label instances in the dataset. Since the original dataset was a single label dataset, it is evident that a dataset prepared with multiple-labels in mind could potentially be a better indicator.

Tweet Vectorization

In order to feed the tokenized tweet into the proposed architecture, the corresponding 2-D matrix should be generated. Three pre-trained word2vec models as discussed earlier are used for this purpose. Each token is converted into a 300-dimensional vector using each of CrisisNLP, Google and Wiki trained word2vec models. Usage of these models ensured that 99.15% of the tokens are present in the combined vocabulary. The overall vector of every token is the weighted average of these three vectors in 4:2:1 ratio. For OOV words, we define a 300-dimensional zero vector. Vectors corresponding to all tokens of a given tweet are stacked together and zero-padded to form a 2-D matrix of the desired size (37 x 300). The number 37 was chosen, since it was the maximum number of tokens found in a preprocessed tweet from the training set.

EXPERIMENTS, RESULTS AND ANALYSIS

In this section we report the details of our experiments, show the evaluation results with necessary analysis.

Experiments: Training & Hyper-parameter Tuning

The baseline CNN as well as the modified architecture discussed in the previous section are trained on the instances obtained after preprocessing. The system is tuned using 5-fold cross validation. We report in *Table 2* the values of optimal hyper-parameters with respect to each component during the experiments. Batch size is set to 32. For training we use the various combinations of twitter-centric features, used along with the word embedding features (as discussed in the Section Augmenting Twitter-Centric Linguistic Features to CNN). We choose the final model architecture based on that particular configuration, which produces the best performance in terms of the AUC value.

Algorithm 1 Online Learning Evaluation

```
Initialize model by training on known crisis data except that of Nepal Earthquake (21750 data instances)
N is set to 1
for minibatch $d_i = \{d_1 \dots d_{n-2}\}$ do
    if i is not divisible by 10 then
        Continue model training on $d_i$
    else
        Train model on previously seen data (21750 + 10*N mini-batches)
        Increment N
    end if
    Evaluate model on $d_{i+1} \cup d_{i+2}$
end for
```

Category	N	D		
Casualties and Public Impact	6	0.25	0.38	0.5
Collateral Damages	5	0.2	0.4	0.5
General Awareness	6	0.25	0.31	0.5
Voluntary Services	6	0.1	0.25	0.5
Sympathy and Emotion	5	0.25	0.38	0.5
Crisis-specific Information	7	0.25	0.38	0.5
Non-informative	3	0.0	0.25	0.5

Table 2. Optimal hyper-parameter values used during experiment. N denotes number of epochs for training and D corresponds to the set of dropout values to be applied to final three neural layers of each component

Online Learning Evaluation

We evaluated our model in an online environment with an algorithm similar to that discussed in Nguyen, S. R. Joty, et al. (2016), as shown in Algorithm 1. The model was initialized on the available CrisisNLP dataset except the Nepal Earthquake. We thereafter divide the incoming dataset D (containing around 5400 tweets) for the Nepal Earthquake into mini-batches d_i of equal size (set to 100). We trained our model on every d_i, and was evaluated on the next 2 mini-batches, $d_{i+1} \cup d_{i+2}$. After N-th interval of 10 successive mini-batches, the model was retrained over the whole training dataset (i.e. 21750 + 10N mini-batches) to avoid the loss of information corresponding to the past instances. Evaluation on future mini-batches in this way could help predict the model's practical performance in a disaster-prone situation. The model was evaluated using the metrics, Hamming Loss and AUC (AUC was calculated for every label). Hamming Loss is a widely used evaluation metric for multi-label classification, and is computed by,

$$\text{HammingLoss}(x_i, y_i) = \frac{1}{|D_t|} \sum_{i=1}^{|D_t|} \sum_{j=1}^{|L|} \frac{xor(x_{i,j}, y_{i,j})}{|L|}, \tag{2}$$

where $|D_t|$ is the number of samples in the test dataset, $|L|$ is the total number of labels, $x_{i,j}$ is the prediction (0 or 1) and $y_{i,j}$ is the ground truth (0 or 1) for the i^{th} data instance and the j^{th} label. It reflects the fraction of the wrong labels to the total number of labels.

Results and Analysis

The model is trained on a dataset of 27,150 unique instances. Results of 5-fold cross validation for the baseline model are shown in *Table 3*. We demonstarte the evaluation results with different combinations of linguistis features in *Table 4*. In *Table 5* we show the evaluation results in terms of F1-Score and AUC values for the best model (i.e.

[5] https://keras.io/

[6] http://scikit-learn.org/

[7] http://pandas.pydata.org/

[8] https://twython.readthedocs.io/

[9] http://www.nltk.org/

[10] https://docs.python.org/3/library/pickle.html

Category	F1-Score	AUC
Casualties and Public Impact	0.9221	0.811
Collateral Damages	0.958	0.776
General Awareness	0.978	0.787
Voluntary Services	0.934	0.746
Sympathy and Emotion	0.953	0.712
Crisis-specific Information	0.828	0.743
Non-informative	0.816	0.812

Table 3. Results of the baseline CNN in the offline setting on the dataset containing tweets related to multiple crises. AUC refers to Area Under Curve metric.

Set of Features	A	B	C	D	E	F	G
Hashtags	0.8201	0.7961	0.8034	0.7688	**0.7294**	**0.7604**	0.8341
User-mentions	0.817	0.7798	0.8058	0.7604	0.7197	0.7516	0.8311
URL Keywords	**0.8253**	0.7959	**0.808**	**0.774**	0.7273	0.7583	0.834
Hashtags + User-mentions	0.8111	0.7809	0.7996	0.7619	0.7273	0.7488	0.8317
Hashtags + URL Keywords	0.8223	**0.7985**	0.8082	0.7684	0.7278	0.7583	**0.8347**
User-mentions + URL Keywords	0.8156	0.7801	0.8008	0.7663	0.7175	0.7471	0.8316
Hashtags + User-mentions + URL Keywords	0.8127	0.783	0.7964	0.7611	0.7222	0.7482	0.8333

Table 4. Results (AUC values) of the proposed architecture with different combinations of twitter-centric textual features along with deep CNN. Labels A – G represent informative categories, as discussed in the Crisis Information Taxonomy Section, in the same order. Highlighted values represent models with highest AUC values for a given category.

Category	F1-Score	AUC
Casualties and Public Impact	0.9544	0.8253
Collateral Damages	0.9683	0.7985
General Awareness	0.9837	0.808
Voluntary Services	0.9427	0.774
Sympathy and Emotion	0.9717	0.7294
Crisis-specific Information	0.8516	0.7604
Non-informative	0.7502	0.8347

Table 5. Results of the proposed model in an offline setting on the dataset containing tweets related to multiple crises: Showing both AUC and F1-Score.

the model that produces the highest peformance) in each category. This model is thereafter used for the evaluation in an online setting. We report these results in Figure 5 and Figure 6.

A close scrutiny of the results as shown in Table 3 and 4 demonstrates that there is, indeed, a positive influence of twitter-centric features on the performance of the system. Even though all the three features contribute in improving the performance, the highest peak is seen with the 'Hashtags' and 'URL keywords' features. Such a behavior can be accounted for by the fact that all crises are events. In an environment like Twitter, hashtag is an indicative notion to the topic or a trending event. Prior research works in event detection rely on hashtags as an important feature. Moreover, due to the character limits imposed on tweets, users usually embed URLs in tweets pointing to the other web sources. In a crisis situation, such URLs can be considered as a rich source of information, especially in case of 'Casualties and Public Impact' and 'General Awareness' categories where government, humanitarian and media communities share information and awareness through their official websites. User-mentions are usually used for replying to specific users on Twitter. Thus, it is possible that the usage of user-mentions does not reveal much about the informativeness of a given message. This might be the reason behind the performance degradation when all three twitter-centric features are used altogether. F1-score is sensitive to the threshold at the output layer while AUC is not. In Table 5, we observe that F1 score for the non-informative category is lower with twitter centric features, but the AUC is still higher. Currently we round off values at the last layer to the nearest integer (ie. 0 or 1)

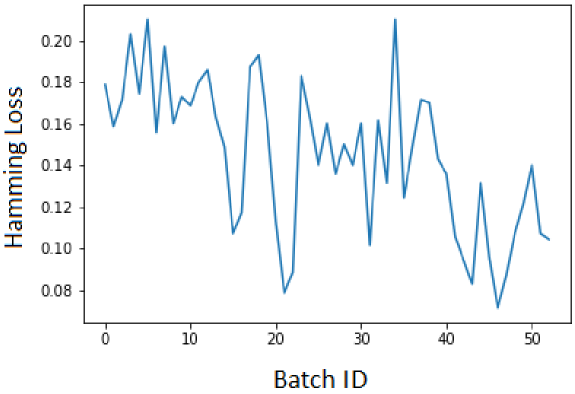

Figure 5. Hamming loss with respect to incoming batches of batch size 200 in online setting, as explained in Online Learning Evaluation

using 0.5 as threshold to obtain the label. The higher AUC shows that with different values of the threshold, it is possible to obtain a higher F1 score as well, and the overall performance of the classifier is still better with twitter centric features. Moreover, training and testing are done on a dataset containing tweets belonging to multiple crises. Thus, the good performance as exhibited by the architecture proves that the proposed model generalizes well to multiple crisis situations.

Results obtained in the online setting show a decrease in hamming loss with incoming batches (as shown in Figure 5). The noisy nature of the AUC curves could be due to the small number of tweets per label in some minibatches. Results show an overall increment in the AUC values for the majority of informative categories. These observations establish that the proposed architecture performs well in an online streaming scenario, making it a perfect fit for the practical purposes.

A quick stroll through the dataset also revealed that annotators were confused between non-informative personal tweets and the tweets belonging to ‘Casualties and Public Impact’ category. One possible reason for such a pattern is that while expressing suffering faced during the crisis, users tend to make it personal. For example, ’Lucky to be one among 25 people to be saved #hagupit’ - This tweet is dominantly personal but it still gives minuscule information that 25 other people were also affected by the crisis. Thus, a tweet being informative and non-informative at the same time introduces ambiguity for annotation, and this affects the performance of the designed system. The unbalanced class distribution is yet another hurdle faced by the model. In our proposed model we use binary cross entropy as the loss function. Designing a category specific cost-sensitive loss function might improve the performance. Moreover, the proposed architecture uses twitter-centric textual features like hashtags, user-mentions and keywords extracted from embedded URLs. Exploring the effect of twitter-centric non-textual features like retweet count, presence of media files etc. can open tremendous opportunities for better feature engineering, and thereby improves the performance as well as the generalization capability of the model.

Comparison with existing models

Most of the prior works were mainly focused on binary and multi-class classification using classical supervised models. There have been a few attempts for multi-class classification using deep learning models. But, to the best of our knowledge none of these models dealt with a multi-label deep learning model. We introduce a model based on deep CNN architecture that is capable of multi-label classification. Moreover, we deal with a single dataset containing tweets associated with multiple crises, while majority of the existing systems tend to use data of a single crisis for both training and evaluation. The results obtained in the online setting by the CNN based binary classifier (informative vs. non-informative), as proposed in (Nguyen, S. R. Joty, et al., 2016), can be compared with that of our proposed model as both make use of the identical settings. Both the models were trained and tested on a dataset of multiple crises. Average AUC value of their model was around 0.74 – 0.75 while our proposed model attains an AUC value of 0.8347 (approx. 11.3% increase). This difference can be attributed to a variety of reasons. Our model is much more robust since it has an additional auto-encoder network which helps in dimensionality reduction and a faster learning process, along with twitter centric feature integration. Also, since they had modeled the problem as a single-label problem, the existence of multiple labels for the same tweet text could have confused the classifier. Imran, Elbassuoni, et al. (2013) mentioned about the development of a multi-label naive Bayes classifier, which was evaluated on the Joplin 2011 tornado dataset with a 4-class information taxonomy, namely ’Caution & Advice’, ’Donation & Offers’, ’Casualty’ and ’Information Source’. However, we can not compare this model’s performance

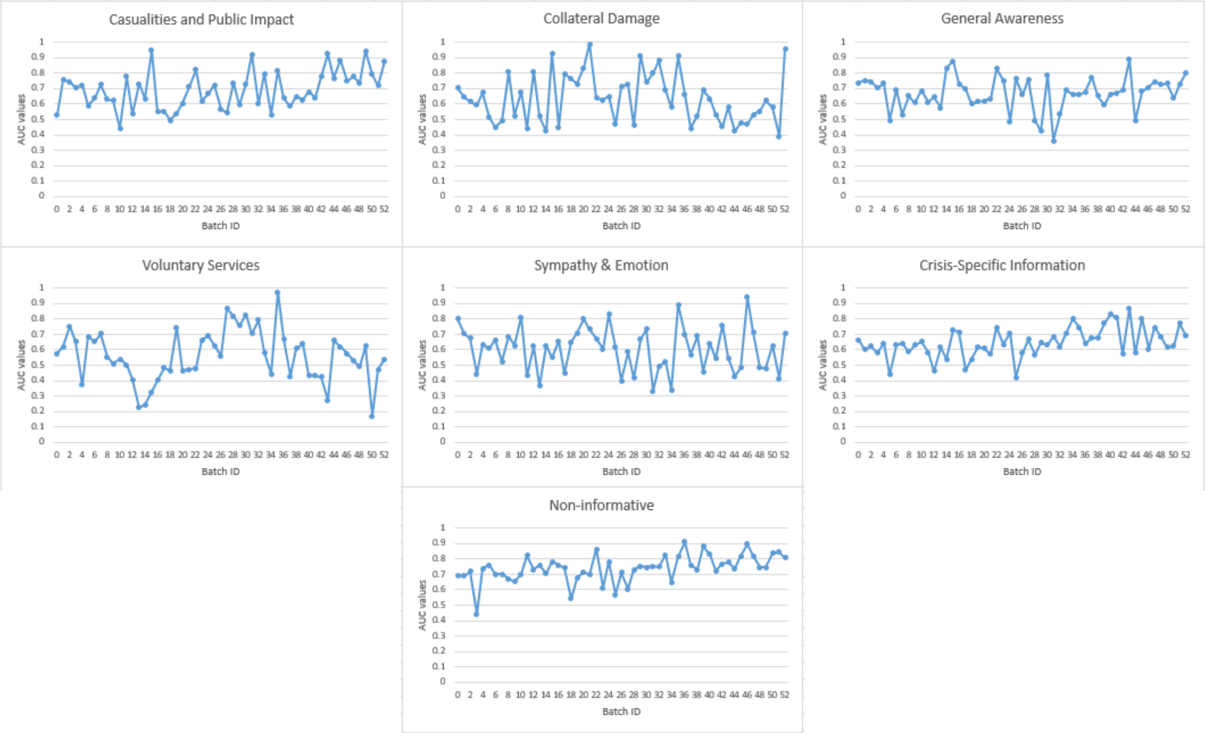

Figure 6. AUC values with respect to incoming batches of batch size 200 experimented with final architecture in online setting, as explained in Online Learning Evaluation

with the results obtained in our proposed model as the evaluations were carried out on two completely different experimental setups – we have used a dataset containing information from multiple disasters, and also a different taxonomy. Our proposed model achieves good performance on multiple crises (as shown in Table 4).

Experiments also show that in an online setting, our proposed architecture yields as output an evolving set of hashtags, user-mentions and relevant topics for each category. The resultant set obtained by taking the union of all such topics can be fed back into a tweet collection module that helps in better filtering of disaster-specific tweets from the public stream of messages.

CONCLUSION AND FUTURE WORK

In this paper, we have established the importance of viewing the classification of disaster-specific tweets as a multi-label problem, and have proposed a deep CNN-based architecture for this purpose. Twitter centric-features derived from the hashtags, user-mentions and keywords extracted from embedded URLs were also explored. Experimental results show that we can achieve significantly better performance with our proposed system compared to the existing state-of-the-art models. Results obtained from both the offline and online settings showcase the ability of our model to adapt and perform well in a practical scenario. The results show that our architecture is a step forward in solving the problem of harnessing the overwhelming flow of information through social media in a crisis situation.

Detailed performance analysis of the model has been performed, pointing at sites of future improvement which includes formulation of a category specific cost-sensitive error function and studying twitter-centric non-textual features which can be used to increase the classification efficiency. We hope to accomplish the above mentioned enhancements as part of our future work. Moreover, modeling information extraction corresponding to the categories to which a given tweet belongs to, is yet another research area that needs to be explored. In conclusion, we hope that our work would prove beneficial to the humanitarian community and foster further research in this field.

ACKNOWLEDGEMENT

Asif Ekbal gratefully acknowledges the JSPS Invitation Fellowship, Government of Japan; and Young Faculty Research Fellowship (YFRF) Awrad, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

REFERENCES

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). "Identifying informative messages in disaster events using convolutional neural networks". In: *International Conference on Information Systems for Crisis Response and Management*, pp. 137–147.

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., Wu, D., Tapia, A., Giles, L., Jansen, B., et al. (2011). "Classifying text messages for the Haiti earth-quake". In: *Proceedings of the 8th International ISCRAM Conference–Lisbon (Vol. 1)*.

Castillo, C. (2016). "Introduction". In: *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, pp. 1–17.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). "Natural language processing (almost) from scratch". In: *The Journal of Machine Learning Research*, pp. 2493–2537.

Dembczyński, K., Waegeman, W., Cheng, W., and Hüllermeier, E. (2012). "On label dependence and loss minimization in multi-label classification". In: *Machine Learning* 88.1, pp. 5–45.

Han, B., Cook, P., and Baldwin, T. (2013). "Lexical Normalization for Social Media Text". In: *ACM Trans. Intell. Syst. Technol.* 4.1, 5:1–5:27.

Imran, M., Mitra, P., and Srivastava., J. (2016). "Cross-language domain adaptation for classifying crisis-related short messages." In: *International Conference on Information Systems for Crisis Response and Management*.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Comput. Surv.* 47.4, 67:1–67:38.

Imran, M., Castillo, C., Lucas, L., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial Intelligence for Disaster Response". In: *Proceedings of the Tenth International Conference on World Wide Web (WWW)*. Seoul, Korea.

Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media". In: *Proc. of ISCRAM, Baden-Baden, Germany*.

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

Joty, S. and Hoque, E. (2016). "Speech Act Modeling of Written Asynchronous Conversations with Task-Specific Embeddings and Conditional Structured Models". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1746–1756.

Kongthon, A., Haruechaiyasak, C., Pailai, J., and Kongyoung, S. (2012). "The role of Twitter during a natural disaster: Case study of 2011 Thai Flood". In: *Technology Management for Emerging Technologies (PICMET), 2012 Proceedings of PICMET'12:* IEEE, pp. 2227–2232.

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

Nguyen, D. T., 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. arXiv: 1610.01030.

Nguyen, D. T., Al-Mannai, K., Joty, S. R., Sajjad, H., Imran, M., and Mitra, P. (2017). "Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks." In: *ICWSM*, pp. 632–635.

Purohit, H. and Sheth, A. (2013). "Twitris v3: From citizen sensing to analysis, coordination and action". In: *Proceedings of ISCRAM*, pp. 918–922.

Robinson, A., Savelyev, A., Pezanowski, S., and MacEachre, A. (2013). "Understanding the utility of geospatial information in social media". In: *Proceedings of the 10th International ISCRAM Conference – Baden-Baden, Germany*.

Vieweg, S., Castillo, C., and Imran, M. (2014). "Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters". In: *Social Informatics*. Ed. by L. M. Aiello and D. McFarland. Cham: Springer International Publishing, pp. 444–461.

Vieweg, S., Hughes, A., Starbird, K., and Palen, L. (2010). "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In: *Proceedings of the SIGCHI'10 Conference on Human Factors in Computing Systems*, pp. 1079–1088.