Analysis of Detection Models for Disaster-Related Tweets

Matti Wiegmann
Bauhaus-Universität Weimar
German Aerospace Center (DLR)
matti.wiegmann@uni-weimar.de

Jens Kersten
German Aerospace Center (DLR)
jens.kersten@dlr.de

Friederike Klan
German Aerospace Center (DLR)
friederike.klan@dlr.de

Martin Potthast
Leipzig University
martin.potthast@uni-leipzig.de

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

ABSTRACT

Social media is perceived as a rich resource for disaster management and relief efforts, but the high class imbalance between disaster-related and non-disaster-related messages challenges a reliable detection. We analyze and compare the effectiveness of three state-of-the-art machine learning models for detecting disaster-related tweets. In this regard we introduce the Disaster Tweet Corpus 2020, an extended compilation of existing resources, which comprises a total of 123,166 tweets from 46 disasters covering 9 disaster types. Our findings from a large experiments series include: detection models work equally well over a broad range of disaster types when being trained for the respective type, a domain transfer across disaster types leads to unacceptable performance drops, or, similarly, type-agnostic classification models behave more robust at a lower effectiveness level. Altogether, the average misclassification rate of 3,8% on performance-optimized detection models indicates effective classification knowledge but comes at the price of insufficient generalizability.

Keywords

Tweet Filtering, Crisis Management, Evaluation Framework.

INTRODUCTION

Social media users share what happens to them and around them, especially when a disaster strikes or when a potential hazard catches their attention. Apart from seeking help and advise they also share eyewitness reports, discuss background information, express sentiment, and connect and coordinate with relevant people. Disaster management and relief efforts are often scarce on real-time information about ongoing disasters, and an analysis of the related social media buzz promises to fill this gap. However, since only a small fraction of all social media messages at any given point in time are disaster-related (Plotnick and Hiltz 2016), the key to tapping this resource for disaster relief is the effective filtering of relevant messages.

A straightforward approach to collect disaster-related messages is filtering, using a dictionary with relevant keywords. This approach fails for cases where the disaster-related terminology is diverse and ambiguous (e.g., “earthquake”, “shaking”, “thoughts and prayers”), and where descriptive terms, such as hashtags (e.g., “#colorado” for the 2013 Colorado floods), are chosen by individual users and are often not consistent over time, while some messages even use incorrect terminology. Moreover, a-priori knowledge of an impending or ongoing disaster is required, since many disaster-indicating words are also used in other situations. Hence, detecting disaster-related messages is commonly modeled as a classification task and tackled with machine learning technology. The existing research studies this classification task mostly in two settings: within-disaster and cross-disaster. In the within-disaster setting, social media messages from a specific disaster, such as the floods in Alberta in 2013, are used for both training and test; this setting is easy to study but lacks practical relevance. In the more realistic cross-disaster setting, one or more sets of messages related to individual disasters are used for training, while messages for other disasters of the same type are used for test, which requires significantly more data and which is therefore studied less often. Most researchers address settings with a single disaster type only.

Given the outlined background we consider the following research questions as highly relevant for practical applications:

1. How well do state-of-the-art models for disaster tweet detection cope with class imbalance?
2. What is the effectiveness discount between disaster-type-specific models and generic, type-agnostic models?
3. How effective are state-of-the-art models in cross-type detection settings?

A literature survey reveals that, despite the many pioneering efforts and the large number of published experiments, these questions have not received the deserved attention. The paper in hand starts closing this gap by evaluating the state of the art with respect to class imbalance, cross-disaster detection, and cross-type detection. The models in our experiments include a recently published Convolutional Neural Network (CNN) (Kersten et al. 2019) and two transformer models, namely BERT (Devlin et al. 2018) and Universal Sentence Encoder (USE) (Cer et al. 2018). We evaluate the models using a six-fold Monte Carlo cross-validation on human-curated tweets for 46 disasters out of 9 disaster types, resulting in 648 experiments per model. The cross-type performance is evaluated within 162 experiments and considers models trained on messages originating from biological hazards, earthquakes, floods, tropical and extra-tropical storms, industrial, societal, and transportation disasters, as well as wildfires. The best-performing models achieve an F_1 of 0.924. We analyze model generalizability by relating the performance loss of a generic classifier trained on all disaster types to the average expected loss of a specialized classifier if applied on disasters from types it was not trained for (altogether 1,620 experiments). A key finding is that the best specialized classifier for biological hazards, earthquakes, and hurricanes is better than a generic classifier by at most 0.046 F_1, and that the average loss of applying an out-of-type classifier is 0.491. The average misclassification rate of all 162 models on a sample of 5 million tweets collected during tranquil periods, unrelated to any disaster, is 4.8%.

Our contributions can thus be summarized as follows:

1. *Large-scale evaluation framework.* We introduce the Disaster Tweet Corpus 2020, the largest corpus of disaster-related tweets to date, consisting of 129,166 tweets sent during 46 disasters covering 9 disaster types.[1]
2. *Systematic evaluation of the state of the art.* We do the most extensive evaluation of state-of-the-art natural language processing algorithms to date within cross-disaster and cross-type settings.
3. *Insights on practical applicability.* We analyze trade-offs and risks attached to type-specific versus generic models when applying them today in realistic settings.

In what follows, the Related Work section surveys the relevant literature. The Methodology section describes the corpus construction, the analyzed models, and the experimental setup. The Results and Discussion section reports on selected outcomes and insights gained, followed by a conclusion and discussion of avenues for future research.

RELATED WORK

Several prior publications study the problem of detecting disaster-related messages, albeit using different terminology and connotations, namely as relevance (Habdank et al. 2017; Kaufhold et al. 2020; Stowe, Palmer, et al. 2018), informativeness (Win and Aung 2017), usefulness (Nguyen et al. 2017), topicality or aboutness (Xu and Chen 2006; Li et al. 2018), and relatedness (Kersten et al. 2019). In contrast to identifying all messages related to a disaster, informativeness, relevance, or usefulness are more applicable to specific applications (i.e. information extraction or damage assessment). Table 1 shows an overview of the recent related work alongside employed methods, covered disasters, conducted experiments, and used datasets.

Although traditional rule-based, keyword-based, and query-based systems have been studied, the most widely employed method for filtering disaster-related social media messages is supervised machine learning. Statistical algorithms based on hand-crafted features are a popular subject of inquiry, with logistic regression employed by

[1] The Disaster Tweet Corpus 2020 is available at `https://doi.org/10.5281/zenodo.3713920`.

Table 1. Overview of the related work proposing detection approaches, grouped by their underlying paradigm, and listing all datasets used. Column "Tests" denotes if the the cross-disaster (CD), cross-type (CT), and filtering (F) is studied. Disaster types: bombing (BO), earthquake (EQ), explosion (EX), flood (FL), hurricane (HU), severe weather (SW), train crash (TC), volcanic eruption (VE), wildfire (WF), and various/other (VA).

Reference	**Disaster Types**										**Tests**			**Datasets**
	BO	EQ	EX	FL	HU	SW	TC	VE	WF	VA	CD	CT	F	
Rules, keywords, or queries														
Abel et al. (2012)	–	–	–	–	–	–	–	–	–	x	–	–	x	TREC (2011), RSS news feeds, news articles
Olteanu, Castillo, et al. (2014)	x	–	x	x	x	x	–	–	–	–	–	x	x	CrisisLex-T6
Zheng et al. (2017)	–	x	–	x	–	–	–	–	–	x	–	–	x	Own data
Machine learning based on feature engineering														
Parilla-Ferrer et al. (2014)	–	–	–	x		––	–	–	–	–	–	–	x	Own data
Stowe, Paul, et al. (2016)	–	–	–	–	x	–	–	–	–	–	–	–	–	Own data
To et al. (2017)	–	x	–	x	–	–	–	–	x	–	–	–	–	CrisisLex-T26 (Olteanu, Vieweg, et al. 2015), Crowdflower (2015), Pfeffer and Morstatter (2016)
Win and Aung (2017)	–	x	–	–	x	–	–	x	x	–	–	–	–	CrisisLex-T6, AIDR (Imran, Castillo, et al. 2014)
Habdank et al. (2017)	–	–	x	–	–	–	–	–	–	–	–	–	–	Own data
Li et al. (2018)	x	–	x	x	x	x	–	–	–	–	x	x	–	CrisisLex-T6
Mazloom et al. (2019)	x	–	x	x	x	x	–	–	–	x	x	x	–	CrisisLex-T6, Schulz and Guckelsberger (2016)
Kejriwal and Zhou (2019)	–	x	–	x	–	–	–	–	–	x	–	x	–	Own data
Kaufhold et al. (2020)	–	–	x	x	–	–	–	–	–	–	–	–	–	Reuter et al. (2015), Habdank et al. (2017)
Neural networks														
Nguyen et al. (2017)	–	x	–	–	x	–	–	–	–	–	–	x	–	CrisisLex-T6, CNLP (Imran, Mitra, et al. 2016), AIDR
Alam, Joty, et al. (2018)	–	x	–	x	–	–	–	–	–	–	–	x	–	CrisisNLP (2018)
Burel and Alani (2018)	–	–	–	–	–	–	–	–	–	x	–	–	–	CrisisLex-T26
Kersten et al. (2019)	–	–	–	x	x	–	–	–	–	x	x	x	x	CrisisLex-T6, CNLP, CMMD (Alam, Ofli, et al. 2018), EPIC (Stowe, Palmer, et al. 2018), McMinn et al. (2013)
Kruspe et al. (2019)	–	–	–	–	–	–	–	–	–	x	x	x	–	CrisisLex-T26, CNLP
Ning et al. (2019)	–	–	–	–	–	–	–	–	–	x	x	x	–	CrisisLex-T26
Snyder et al. (2019)	x	–	–	–	–	–	x	–	x	x	–	–	–	CrisisLex-T26, Own data

Win and Aung 2017, naïve Bayes and support vector machines by Parilla-Ferrer et al. 2014, and random forest decision trees by Kaufhold et al. 2020. These methods typically achieve an accuracy of about 0.75 in cross-disaster experiments, notably outperformed by recent neural network architectures. Ning et al. 2019 demonstrated that convolutional neural networks (CNNs) outperform other methods in the cross-disaster prediction of informativeness on the popular CrisisLex-T26 (Olteanu, Vieweg, et al. 2015) collection, reporting F_1-scores of 0.81. Similarly, Burel and Alani 2018 report F_1-scores of 0.84 for filtering disaster-related tweets in a 5-fold cross-validation on CrisisLex-T26 with a standard CNN, and Kersten et al. 2019 report F_1-scores of 0.83 with a parallel CNN in cross-disaster settings over several collections of disaster-related tweets.

Most approaches presented in the related work are intended for multiple disasters or disaster types, with the most studied ones being earthquakes, floods, and hurricanes, followed by explosions, bombings, and severe weather events, while some also study uncommon disasters or completely different events. However, experiments are often conducted only on single disasters, while the more realistic cross-disaster setting has been gaining traction only recently. The cross-type transfer of models is more frequently investigated, especially using neural networks, although only a few studies comprehensively test the performance and transferability of classifiers across disaster types, like Kersten et al. 2019 for hurricanes and floods. Recently, active learning approaches have been employed, where models are trained or fine-tuned based on data annotated on the fly by citizens during disasters (Kaufhold et al. 2020; Snyder et al. 2019), or based on labeled data from past events used in combination with unlabeled or partially labeled messages from ongoing events (Imran, Mitra, et al. 2016; Li et al. 2018; Mazloom et al. 2019).

Of the various data sources used, the most recurring ones are the collections CrisisLex-T6 (Olteanu, Castillo, et al. 2014) and CrisisLex-T26 (Olteanu, Vieweg, et al. 2015), which cover 6 and 26 different disasters, respectively. Half of all studies rely on a public resource like CrisisLex, with acquiring and annotating own data being a close second. The dominant strategy for data collection involves requesting tweets by keywords taken from general disaster terminology ("earthquake", "ground shaking"), combined with disaster-specific indicator words, phrases, and hashtags, as well as an iterative refinement of queries as described by Olteanu, Castillo, et al. 2014.

Table 2. Human-annotated tweets used in this study, grouped by disaster type, where the number of tweets is what remained after preprocessing, excluding duplicates, very short, and non-English tweets.

Name	Tweets
Earthquake (11)	**27,034**
2012 Costarica	568
2012 Guatemala	390
2012 Italy	268
2013 Pakistan	3,468
2013 Bohol	1,330
2013 California	326
2013 Chile	3,840
2015 Nepal	5,930
2017 Mexico	236
2017 Iraq and Iran	86
2018 Nepal	10,592
Flood (9)	**14,210**
2012 Philipinnes	1,324
2013 Sardinia	182
2013 Manila	1,224
2013 Alberta	9,060
2013 Queensland	6,336
2013 Colorado	1,706
2014 India	3,594
2014 Pakistan	3,514
2017 Srilanka	1,480
Biological (2)	**6,106**
2014 Ebola	3,448
2014 Mers	2,658
Hurricane (9)	**48,922**
2012 Hurricane Sandy	12,530
2012 Hurricane Pablo	1,426
2013 Typhoon Yolanda	1,544
2014 Typhoon Hagupit	3,916
2014 Hurricane Odile	2,458
2015 Cyclone Pam	3,886
2017 Hurricane Harvey	7,706
2017 Hurricane Maria	7,674
2017 Hurricane Irma	7,782
Wildfires (3)	**4,820**
2012 Colorado	182
2013 Australia	1,730
2014 California	2,908
Industrial (4)	**10,166**
2013 West-Texas explosion	8,102
2013 Brazil nightclub fire	478
2012 Venezuela refinery explosion	124
2013 Savar building collapse	1,462
Societal (2)	**11,206**
2013 Boston bombing	9,576
2013 LA airport shootings	1,630
Transportation (4)	**3,850**
2013 Glasgow helicopter crash	1,554
2013 New York train crash	1,498
2013 Spain train crash	704
2013 LA train crash	94
Other (2)	**2,852**
2013 Russia meteor impact	1,524
2013 Singapore haze	1,328

METHODOLOGY

We conducted three experiments to evaluate state-of-the-art approaches with regard to their effectiveness to detect messages related to disasters. The first experiment studies their performance in cross-disaster settings, covering 9 disaster types. More specifically, each model was trained on two selected disasters and then tested on different disasters of the same type. The second experiment compares such type-specific models to a generic model, which is trained on disasters from all types, illustrating the trade-off between specialized and generic models. The third experiment analyzes the effectiveness of classifying unrelated tweets on a large sample of tweets from a tranquil period. As a baseline approach, we examined the performance of a standard list of disaster-related keywords to detect disaster-related tweets. In an auxiliary experiment, we demonstrate the stability of our evaluation strategy with regard to the amount of available training data.

Data

Table 2 lists the 46 disasters considered in this study and the number of tweets available for each of them. The disaster-related tweets originate from 7 collections reviewed in the related work: AIDR (Imran, Castillo, et al. 2014), CrisisLex T6 (Olteanu, Castillo, et al. 2014), CrisisLex T26 (Olteanu, Vieweg, et al. 2015), CrisisNLP (Imran, Mitra, et al. 2016), CrisisMMD (Alam, Ofli, et al. 2018), Epic Annotations (Stowe, Palmer, et al. 2018), and the collection of events from 2012 by McMinn et al. 2013. We assigned each disaster to one of 9 disaster types, based on the taxonomy of disaster types developed by the disaster databases EM-DAT (Guha-Sapir 2019) and Glide (GLIDE 2019). In particular, we grouped all tropical and extra-tropical storms to "hurricanes", added a "societal" type, and merged all uncommon disasters in the "others" type. Additionally, we created a "tranquil period" dataset of 5 million tweets, randomly sampled from all tweets sent since 2011, in order to evaluate the model performance in a detection setting with many negative examples. Not all of the above datasets contain non-disaster-related tweets to an equal amount. To ensure balanced datasets for the evaluation, we removed all negative examples from the datasets and filled them up to balance with tweets from the tranquil sample.

All datasets were preprocessed as follows: (1) removal of all non-English tweets as well as all retweet-indicating prefixes (RT @username:), (2) replacement of all URLs with <URL>, hashtags with <HASHTAG>, user mentions with <USER>, emoticons with <SMILE>, emojis with <EMOJI>, colon-separated numbers with <TIME>, and other numeric strings with <NUMBER>, (3) collapsing of character repetitions and adding <REPETITION>, (4) removal of line breaks and collapsing of white space, and, finally (5) removal of all duplicates and tweets shorter than five characters (excluding the above replacement tags). More than half of the tweets have been removed from the dataset, minimizing side-effects and providing for a sensible collection of tweets.

Models

Three machine learning architectures are used in this study: A parallel CNN previously proposed for detecting disaster-related tweets as baseline (Kersten et al. 2019), a feed-forward neural network based on BERT embeddings[2]

[2] The uncased, 12 layer BERT from TensorFlow Hub: `https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1`.

Table 3. Average F_1 of the respective model when trained on multiple disaster types (MDT) and tested on all eleven test datasets, their average cross-disaster (CD) and cross-type (CT) F_1 scores, and their average miss-classification rate on tranquil tweets (MCR).

Model	MDT	CD	CT	MCR
CNN	0.775	0.740	0.323	0.063
BERT	0.776	0.690	0.415	**0.037**
USE	**0.818**	**0.752**	**0.467**	0.044

and a feed-forward neural network based on a universal sentence encoder (USE)[3]. The CNN represents specialized architectures for detecting disaster-related tweets, BERT and USE represent the current state of the art in many natural language processing applications, performing well in a broad range of language tasks while being less reliant on a lot of training data.

The CNN is built on word embeddings that were specifically trained with Word2Vec for disaster tweets (Nguyen et al. 2017), using three parallel CNN branches with filter sizes 3, 4, and 5, respectively; each branch has 128 filters and a dropout of 0.5. The branches are concatenated and passed to three feed-forward layers with Rectified Linear Units (ReLu) as activation function and Max-Entropy as optimization criterion. The BERT-based model is a 3-layer feed-forward neural network without dropout, ReLUs as activation function, and Max-Entropy as optimization criterion; the input encoding is generated by the pre-trained BERT implementation. Accordingly, the USE-based classifier employs the transformer-based USE encoder to generate input encodings for a feed-forward neural network identical to the BERT-based model. A direct comparison of the average performance difference between these models is shown in Table 3.

Experiment Settings

For the cross-disaster classification experiments, we constructed the training dataset for each type by selecting two disasters from the five types flood, hurricane, earthquake, transportation, and industrial, and one disaster from the three types biological, wildfire, and societal. We randomly sampled 1,500 positive and negative examples each from the respective training disaster to avoid size effects on the models' performance scores. We trained a model on the sampled tweets for each disaster type and tested it against all tweets from all other disasters of the same type not selected for training. In addition, we sampled 3,000 tweets from the "other" category to test against disasters of uncommon type. We repeated the procedure of selecting disasters and sampling training examples in a 6-fold Monte Carlo cross-validation. Altogether we trained 144 models and executed 162 evaluations. To assess model generalizability, we combined all training and test samples for each cross-validation step and trained a generic model for multiple disaster types. We tested all 8 same-type models and the generic model on each of the 10 test sets from the cross-disaster experiments, and the combined test set to get both, the cross-type loss for each model and the performance difference of using the generic model over the specialized one. In this regard, we trained another 18 models and executed another 1,620 evaluations. Finally, the 162 models of the first two experiments are tested on a random sample of 500,000 tweets from a tranquil period in order to evaluate the misclassification rate in a close-to-realistic setting with exclusively negative examples.

For the keyword-based baseline, we used the CrisisLex (Olteanu, Castillo, et al. 2014) dictionary of disaster-related keywords, preprocessed them, and classified each tweet in all datasets as related if it contained one or more of the keywords. To test the training data requirements of our method, we repeated the cross-disaster evaluation for the generic and the specialized models on earthquakes, floods, and hurricane events, leaving out a static test set, and evaluated all models in a 5-fold Monte Carlo cross-validation with 20,000, 2,000, 200, and 20 tweets each for training.

RESULTS AND DISCUSSION

Table 3 shows the average F_1 of the model-architectures CNN, BERT, and USE over all respective models, where USE-based models perform best as generic classifiers, as well as in the cross-disaster and the cross-type scenario, while the BERT-based models perform marginally better with regard to their misclassification rate.

Regarding the cross-disaster detection performance of a model, Table 4 (left) shows the best-performing cross-disaster model for each tested disaster type after cross-validation. The F_1 scores of these models range from 0.869 on societal disasters to 0.980 on biological disasters. There is no significant difference in performance between

[3] The large tranformer-based model from Tensorflow Hub: `https://tfhub.dev/google/universal-sentence-encoder-large/3`.

Table 4. Average F_1, precision, recall, and misclassification rate on the tranquil tweets (MCR) of the best-performing models and the wordlist baseline (WL) on each test disaster. The left half shows the best model trained on tweets from only one type of disaster, the right half shows models trained on data from all disaster types. Bold highlighting indicates the best F_1 in that row, and underlining that the best-performing model diverges in type from the test data.

Test Type	**Best Training Type**						**Cross-type Training Data**					**WL**
	Training Type	F_1	Prec	Rec	MCR	Model	F_1	Prec	Rec	MCR	Model	F_1
All	–	–	–	–	–	–	0.883	0.947	0.826	0.047	USE	–
Biological	Biological	**0.980**	0.977	0.982	0.022	USE	0.965	0.957	0.973	0.047	USE	0.275
Earthquake	Earthquake	**0.937**	0.951	0.926	0.054	USE	0.924	0.958	0.895	0.050	BERT	0.614
Flood	Flood	0.891	0.927	0.863	0.078	USE	**0.916**	0.904	0.929	0.122	CNN	0.691
Hurricane	Hurricane	**0.929**	0.950	0.910	0.062	USE	0.883	0.953	0.826	0.047	USE	0.672
Industrial	Wildfire	0.886	0.969	0.817	0.031	USE	**0.973**	0.971	0.975	0.047	USE	0.799
Societal	Societal	**0.869**	0.950	0.803	0.041	USE	0.864	0.953	0.793	0.047	USE	0.596
Transportation	Transportation	0.937	0.990	0.899	0.013	CNN	**0.984**	0.985	0.983	0.047	USE	0.648
Wildfire	Flood	0.925	0.935	0.916	0.022	BERT	**0.929**	0.950	0.909	0.047	USE	0.497
Other	Earthquake	0.680	0.916	0.543	0.054	USE	**0.739**	0.877	0.655	0.122	CNN	0.258
Average	–	0.888	0.949	0.844	0.038	–	**0.906**	0.945	0.876	0.062	–	–
Wordlist	–	–	–	–	–	–	0.621	0.989	0.453	0.019	–	–

Table 5. Average F_1 over all models and cross-validations for the cross-disaster (CD) and cross-type (CT) experiments, and the average cross-type loss and variance of the F_1 scores when using a cross-type over a cross-disaster model. CT-Neg shows the percentage of tweets that were classified negative in cross-type settings.

Test Data	**CD**	**CT**	**CT-Loss**	**CT-Variance**	**CT-Neg**
Biological	0.953	0.340	0.613	0.057	0.851
Earthquake	0.923	0.556	0.368	0.076	0.743
Flood	0.867	0.511	0.356	0.059	0.779
Hurricane	0.933	0.497	0.436	0.068	0.775
Industrial	0.801	0.631	0.170	0.052	0.704
Societal	0.826	0.380	0.446	0.036	0.843
Transportation	0.904	0.471	0.433	0.056	0.784
Wildfire	0.901	0.544	0.357	0.063	0.748
Average	0.889	0.491	0.397	0.058	0.779

resource-rich disaster types (earthquake, flood, and hurricane) and the newly added ones. For the uncommon disasters collected in the “other” type, the best model in the cross-disaster setting achieves a lower F_1 of 0.680, and therefore remains as a future challenge. When additionally considering the cross-type models, it is notable that the model for wildfires outperforms that for industrial disasters, and that the flood-model outperforms the wildfire model on the respective test data. A successful domain transfer across types is nonetheless the exception, as shown by the low average cross-type F_1 and the low variance between the different cross-type measures displayed in Table 5. The exceptional cases in which type-transfer works warrant further inspection on an individual basis.

Table 4 (right) shows the results for the best-performing generic model, which was trained on tweets from all disaster types. The generic model works better on the rare, unseen events in the “other” type, but also outperforms the specialized models on floods, wildfires, industrial, and transportation disasters. In addition, Table 5 shows the average loss of applying a specialized model on a different type of test disasters, which can be interpreted as the risk of choosing the wrong model. The average cross-type-loss is 0.397 in F_1, which is significant considering that our generic model has a higher average F_1 over all disaster types than the best specialized models. If classification effectiveness is the primary goal, it is advisable to chose the USE-based generic model when filtering for more than one disaster type.

Table 4 reports the average misclassification rate over the tranquil tweets, ranging from 0.006 to 0.122 over all models. The best generic model achieves an average of 0.055 and the best specialized model 0.038. When comparing the average misclassification rate between model architectures (see Table 3), the BERT-based architecture performs best with an MCR of 0.037, followed by USE with 0.044, and lastly the CNN with 0.063. If noise-reduction is the primary goal and cross-type loss is an acceptable risk, the BERT-based specialized models should be chosen.

When comparing the model results to the keyword-based baseline also shown in Table 4, it is notable that the baseline has a very high precision, low misclassification rate on tranquil tweets, and much lower recall than the

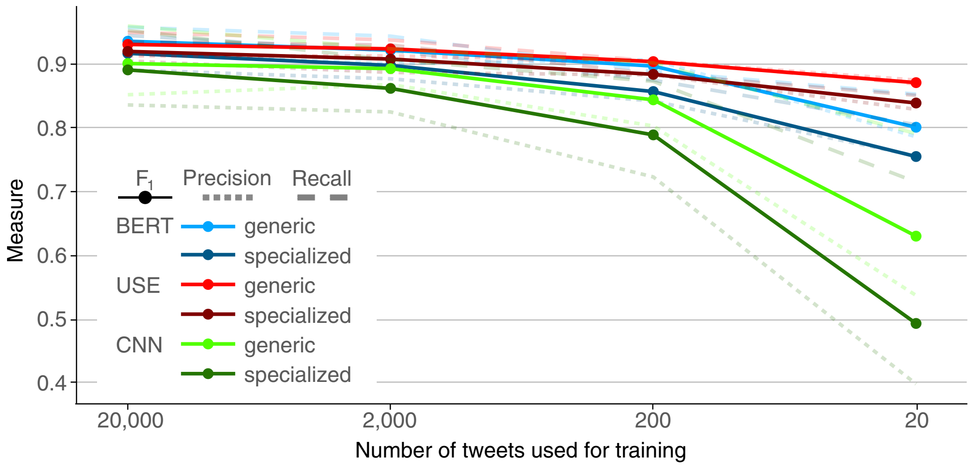

Figure 1. Model performance over training set size as average cross-event F_1 scores for earthquake, flood, and hurricane prediction (specialized), and their combination (generic).

machine learning approaches, which leads to a significantly lower F_1 score on the balanced datasets. The baseline performs especially poor on rare disasters, as can be seen by the low F_1 scores on biological and other disasters. This exposes the major trade-off of keyword-based filters: they are more reliable, less prone to additional noise, and easier to interpret, but frequently miss unforeseen information.

Figure 1 shows the results of training models for earthquakes, floods, and hurricanes with varying amounts of training data on a static test dataset. The best-performing model achieves 0.93 F_1 given 20,000 training examples, with all other models within the 5% range. In a resource-constrained environment, the generic models outperform the specialized ones notably, but these differences become marginal as the number of examples increases. While the models based on pretrained, contextualized word embeddings do not significantly benefit from more training data beyond 2,000 tweets, the CNN relies more heavily on them. This can be explained by the higher number of trainable parameters in the CNN model. While the CNN performs only marginally above random predictions given only 20 tweets, both BERT and USE achieve meaningful classifications results. This may be explained by the limitations of the training data: Since most positive examples were originally collected using keywords, it is plausible that a model with few trainable parameters (like a feed forward neural network using contextualized word embeddings) learns the most important keywords from the given examples and approximates a restricted wordlist for filtering.

Limitations

Some limitations apply to the interpretation and practical application of the results described in this study. Firstly, the results are not representative for the maximum possible performance of the individual classifiers, since specialized classifiers were trained on only 3,000 tweets for reasons of comparability. Especially with regards to the lower recall values, training models with more examples typically increases performance. Secondly, the best-case misclassification rate of 0.038 suggest good performance, but conceivably may still be too high for real-world applications with a common related-unrelated-ratio of 1:10,000 or worse. Thirdly, the low average cross-type performance of 0.491 and the high cross-type loss of 0.397 suggest that specialized models classify tweets related to other disaster types as unrelated and can thus distinguish between different disasters occurring in parallel. This is not necessarily the case, since the average rate of 0.779 for negative predictions from cross-type models is much higher than the misclassification rate suggests. Lastly, the performance reported in this study is not necessarily comparable to those reported in related work. Our method of sampling negative examples is not based on potential keywords but on a statistically sound representation of all unrelated tweets, resulting in a lower lexical similarity between classes which may render the classification task easier. Since the presented models are intended as a detector on an unfiltered social media stream instead of on a stream filtered using keywords, we believe that our methodology reflects a more realistic scenario.

CONCLUSION

We present a benchmark corpus of human-curated tweets related to 46 disasters compiled from related work, and grouped into 9 disaster types, the Disaster Tweet Corpus 2020. With this corpus, we compared the ability of 162 models over 1,944 evaluations to classify tweets as related or unrelated to a disaster with regard to cross-disaster classification, generalizability, and misclassification rate.

The best specialized models achieve an F_1 of 0.888 on average when generalizing to unseen disasters of a known type, but lose an average 0.397 F_1 when generalizing to disasters of an unseen type. Generic models trained on all disasters perform slightly better in terms of F_1 than specialized models, but have weaker misclassification rates on unrelated tweets. As practical recommendation, a generic model is preferable unless only one disaster type is of interest, or unless reducing noise as much as possible is paramount, and the risk of choosing the wrong model is of no concern.

The limitations of our study leave room for further exploration of the topic. Naturally, creating larger and more representative training datasets will improve the performance of the individual classifiers, especially with regard to recall. Although effectiveness is comparable across disaster types on similarly sized training data, improving the detection of tweets from rare and unseen disasters remains a future challenge. Models applied across disaster types are generally not competitive, although the occasionally observed exceptions warrant further investigation on a case-by-case basis.

REFERENCES

Abel, F., Hauff, C., Houben, G., Stronkman, R., and Tao, K. (2012). "Twitcident: fighting fire with information from social web streams". In: *WWW (Companion Volume)*. ACM.

Alam, F., Joty, S., and Imran, M. (July 2018). "Domain Adaptation with Adversarial Training and Graph Embeddings". In: *Proceedings of the 56th ACL*. Melbourne, Australia: ACL.

Alam, F., Ofli, F., and Imran, M. (2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the 12th 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 ISCRAM*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N. L. U., John, R. S., Constant, N., Guajardo-Caspedes, M., Yuan, S., Tar, C., et al. (2018). "Universal Sentence Encoder". In:

CrisisNLP (2018). https://crisisnlp.qcri.org/data/acl_icwsm_2018/ACL_ICWSM_2018_datasets.zip.

Crowdflower (2015). https://d1p17r2m4rzlbo.cloudfront.net/wp-content/uploads/2016/03/socialmedia-disaster-tweets-DFE.csv.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805.

GLIDE (2019). http://glidenumber.net.

Guha-Sapir, D. (2019). *The Emergency Events Database (EM-DAT)*. www.emdat.be.

Habdank, M., Rodehutskors, N., and Koch, R. (2017). "Relevancy assessment of tweets using supervised learning techniques: Mining emergency related tweets for automated relevancy classification". In: *2017 4th ICT-DM*.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: artificial intelligence for disaster response". In: *WWW (Companion Volume)*. ACM.

Imran, M., Mitra, P., and Srivastava, J. (2016). "Enabling Rapid Classification of Social Media Communications During Crises". In: *IJISCRAM* 8.

Kaufhold, M.-A., Bayer, M., and Reuter, C. (2020). "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning". In: *Information Processing & Management* 57.1.

Kejriwal, M. and Zhou, P. (2019). "Low-supervision Urgency Detection and Transfer in Short Crisis Messages". In: *CoRR* abs/1907.06745.

Kersten, J., Kruspe, A., Wiegmann, M., and Klan, F. (2019). "Robust Filtering of Crisis-related Tweets". In: *Proceedings of the 16th ISCRAM, May 19-22*. ISCRAM. Valencia, Spain.

Kruspe, A., Kersten, J., and Klan, F. (2019). "Detecting Event-Related Tweets by Example using Few-Shot Models". In: *Proceedings of the 16th ISCRAM, May 19-22*. Valencia, Spain.

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

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (July 2019). “A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets”. In: *International Journal of Information Systems for Crisis Response and Management* 11.

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 CIKM*. San Francisco, California, USA: ACM.

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 ICWSM*. AAAI.

Ning, X., Yao, L., Benatallah, B., Zhang, Y., Sheng, Q. Z., and Kanhere, S. S. (Aug. 2019). “Source-Aware Crisis-Relevant Tweet Identification and Key Information Summarization”. In: *ACM Trans. Internet Technol.* 19.3.

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 ICWSM*.

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.

Parilla-Ferrer, B. E., Fernandez, P., and T. Ballena IV, J. (Dec. 2014). “Automatic Classification of Disaster-related Tweets”. In: *Proceedings of the ICIET*. Bangkok, Thailand.

Pfeffer, J. and Morstatter, F. (2016). *Geotagged Twitter Posts from the United States: A Tweet Collection to Investigate Representativeness*. Carnegie Mellon University, Arizona State University. URL: http://doi.org/10.7802/1166.

Plotnick, L. and Hiltz, S. R. (2016). “Barriers to Use of Social Media by Emergency Managers”. In: *Journal of Homeland Security and Emergency Management* 13.

Reuter, C., Ludwig, T., Kaufhold, M.-A., and Pipek, V. (2015). “XHELP: Design of a Cross-Platform Social-Media Application to Support Volunteer Moderators in Disasters”. In: *Proceedings of the 33rd Annual ACM CHI*. Seoul, Republic of Korea: ACM.

Schulz, A. and Guckelsberger, C. (2016). http://www.doc.gold.ac.uk/~cguck001/IncidentTweets/.

Snyder, L. S., Lin, Y., Karimzadeh, M., Goldwasser, D., and Ebert, D. S. (2019). “Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness”. In: *IEEE Transactions on Visualization and Computer Graphics*.

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 LAW-MWE-CxG-2018*. Santa Fe, New Mexico, USA: ACL.

Stowe, K., Paul, M. J., Palmer, M., Palen, L., and Anderson, K. (Nov. 2016). “Identifying and Categorizing Disaster-Related Tweets”. In: *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. Austin, TX, USA: ACL.

To, H., Agrawal, S., Kim, S. H., and Shahabi, C. (2017). “On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?” In: *2017 IEEE Third BigMM*.

TREC (2011). https://trec.nist.gov/data/tweets/.

Win, S. S. M. and Aung, T. N. (May 2017). “Target oriented tweets monitoring system during natural disasters”. In: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*.

Xu, Y. (and Chen, Z. (2006). “Relevance judgment: What do information users consider beyond topicality?” In: *Journal of the American Society for Information Science and Technology* 57.7.

Zheng, X., Sun, A., Wang, S., and Han, J. (2017). “Semi-Supervised Event-related Tweet Identification with Dynamic Keyword Generation”. In: *Proceedings of the 2017 ACM CIKM*. Singapore, Singapore: ACM.