Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison

Nilani Algiriyage[*]
Massey University
r.nilani@massey.ac.nz

Rangana Sampath
Massey University
r.sampath@massey.ac.nz

Raj Prasanna
Massey University
r.prasanna@massey.ac.nz

Emma E H Doyle
Massey University
e.e.hudson-doyle@massey.ac.nz

Kristin Stock
Massey University
k.stock@massey.ac.nz

David Johnston
Massey University
d.m.johnston@massey.ac.nz

ABSTRACT

Social media applications such as Twitter and Facebook are fast becoming a key instrument in gaining situational awareness (understanding the bigger picture of the situation) during disasters. This has provided multiple opportunities to gather relevant information in a timely manner to improve disaster response. In recent years, identifying crisis-related social media posts is analysed as an automatic task using machine learning (ML) or deep learning (DL) techniques. However, such supervised learning algorithms require labelled training data in the early hours of a crisis. Recently, multiple manually labelled disaster-related open-source twitter datasets have been released. In this work, we collected 192, 948 tweets by combining a number of such datasets, preprocessed, filtered and duplicate removed, which resulted in 117, 954 tweets. Then we evaluated the performance of multiple ML and DL algorithms in classifying disaster-related tweets in three settings, namely “in-disaster”, “out-disaster” and “cross-disaster”. Our results show that the Bidirectional LSTM model with Word2Vec embeddings performs well for the tweet classification task in all three settings. We also make available the preprocessing steps and trained weights for future research.

Keywords

Tweet classification, machine learning, deep learning, disasters.

INTRODUCTION

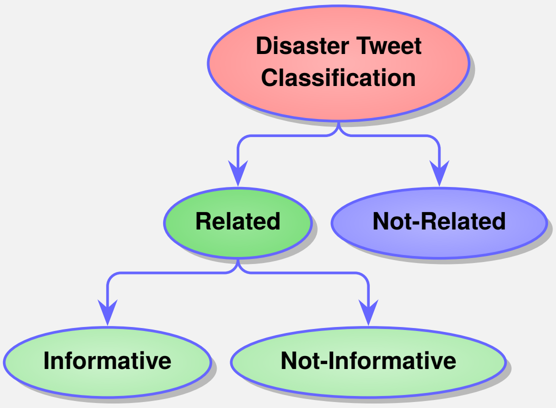
Social media (SM) platforms play an important role in providing a quick understanding of the situation as it unfolds during disasters. Research has found that the general public use SM applications during disasters to communicate information regarding urgent needs, infrastructure damage, injured or dead people, volunteering or donation efforts, and situational updates (Kumar et al. 2019; Madichetty and Sridevi 2019; O’Keefe and Alrashdi 2018; Alam, Joty, et al. 2018). Timely access to SM data can be leveraged for emergency response in the first few hours to significantly reduce both human loss and economic damage (Alam, Joty, et al. 2018).

One of the main challenges for utilizing SM in crisis situations is the reliable detection of useful messages in a massive amount of streaming data. A straight forward method for collecting disaster-related tweets is to use disaster-keyword filtering. For example, tweets can be filtered using a dictionary with relevant keywords (e.g.,

[*] corresponding author

“flood”, “earthquake”) or specific hashtags (e.g.,“#NepalQuake”, “#boulderflood”, “#coloradoflood”). However, these descriptive keywords are diverse and ambiguous, and the hashtags chosen by individual users are often not consistent over time (Wiegmann et al. 2020; Ning et al. 2019). As a result, a significant portion of the collected tweets can be irrelevant. Therefore, detecting disaster-related tweets is commonly modelled as an automatic classification task and tackled with Machine Learning (ML) algorithms and more recently with Deep Learning (DL) algorithms (Wiegmann et al. 2020; Alam, Joty, et al. 2018; Neppalli et al. 2018).

Tweet classification for disaster response is a text classification task that aims to determine if a tweet is related to a particular type of predefined informative class (O’Keefe and Alrashdi 2018). Olteanu et al. (2015) showed that crisis related tweets can be broadly categorised into: *related and informative*, *related but not informative*, and *not related*. For example see Figure 1 representing the Olteanu categorization using tweets extracted during 2013 Colorado floods.



1. “RT @CUBoulder: SEEK HIGHER GROUND IMMEDIATELY: Wall of water coming down Boulder Canyon. Move away from Boulder Creek! #BoulderFlood”
2. “Sending our thoughts and prayers to those affected by the Colorado flooding”
3. “Will people understand the value of longterm data? I hate to see @USGS streamgages being discontinued due 2 lack of funding #coloradoflood”

Figure 1. The Olteanu catergorization for tweets, and three example tweets from the 2013 Colorado floods (Olteanu, Vieweg, et al. 2015). The first tweet is categorised as “Related and informative”, second as “Related - but not informative” and third as “Not related”

A vast majority of the existing literature has focused on classifying tweets of the same event type and mostly used the CrisisLexT26 dataset for training classifiers (Acerbo and Rossi 2017; Gata et al. 2019; Burel and Alani 2018; Khare et al. 2018). The CrisisLexT26 contains around 250K tweets posted during 26 crisis events in 2012 and 2013 (Olteanu, Vieweg, et al. 2015). Communication patterns of people might change over the years and, therefore, classification accuracy using classifiers trained on older datasets may not be high for future events (Graf et al. 2018). Furthermore, supervised learning algorithms work well with more and complete training data covering the full spectrum of inputs that the model is supposed to handle during the classification task. Therefore, there is a timely need to test classifiers for new and more comprehensive datasets. To the best of the authors’ knowledge, to date, there exists no research for large scale ML and DL model evaluation in identifying disaster-related tweets combining multiple datasets. Therefore, during this research, we address the following research question.

- What ML or DL model has the best performance for a disaster-related tweet classification task?

To answer this question, we conduct experiments in the following three settings:

- In-disaster: training and test data belong to the same disaster type.
- Out-disaster: training and test data belong to different disaster types.
- Cross-disaster: training set consists of tweets of various disaster types.

During in-disaster experiments, we take both train and test data belonging to the same disaster type. The currently labelled datasets belong to a few disaster categories, including flood, earthquake, hurricane and biological. However, in reality, there are many more disasters (e.g., landslides, volcanic eruptions, droughts and tsunami) where people use tweets to communicate. Therefore, we wanted to explore how accurate a model can be if they are applied outside the domain. As a result, we explore a setting where train and test datasets belonging to different disaster types. We train models for a combination of disaster types and test on individual types during the cross-disaster experiment. Altogether, we carry out 540 train, test experiments.

Our contributions can thus be summarized as follows:

1. Evaluation of a large-scale ML, DL model for disaster-related tweet classification
2. Evaluation of three state-of-the-art word embedding models
3. Publication of all the learning weights so that the response agencies can quickly adopt the trained models for an ongoing disaster[1]

The rest of this article is organized as the follows: The Related Work section reviews the literature related to disaster tweet classification. In the Methodology section we discuss the technical architecture and algorithms developed. The Results section provides results and critique of the findings. Finally, the Conclusion gives a brief summary and some directions for future research.

RELATED WORK

In the crisis domain, useful information retrieval is an early step in processing data from SM platforms (Basu et al. 2020; Ghosh et al. 2019; Zheng and Sun 2019; Meurisch et al. 2019). A large and growing body of literature has investigated this as an automatic tweet classification problem (Parilla-Ferrer et al. 2014; Toriumi and Baba 2016; Alam, Joty, et al. 2018; Gata et al. 2019; Neppalli et al. 2018). These studies can be divided into three categories; related tweet classification, informative tweet classification and specific topical classifications. Related tweet classification focuses on identifying whether a tweet is related to a crisis event or not (Graf et al. 2018). The concept of "Informativeness" is subjective, which heavily depends on the receiver of the information. However, generally "informative" tweets can be defined as tweets that provide valuable information to anyone in the scene of a disaster (e.g., a victim, supporter or responder). In comparison, "non-informative" tweets can be defined as tweets which do not convey any useful content in the scene of a disaster (Neppalli et al. 2018). Research on specific topical classifications group tweets into multiple categories such as injured or dead people, sympathy and emotional support, affected people, caution and advice, missing people and donation needs (D. T. Nguyen et al. 2016). A summary of the closely related work is presented in Table 1.

The vast majority of literature has considered classifying useful tweets using ML algorithms such as Naïve Bayes (NB) (Parilla-Ferrer et al. 2014), Random Forest (RF) (Kaufhold et al. 2020), Logistic Regression (LR) (D. Nguyen et al. 2017), Artificial Neural Networks (ANNs) (Caragea et al. 2016) and Support Vector Machines (SVMs) (Khare et al. 2018). More recent attention has focused on using deep neural networks such as Convolutional Neural Networks (CNN) (Neppalli et al. 2018; Ning et al. 2019) and Long Short-Term Memory Networks (LSTM) (O'Keefe and Alrashdi 2018) to address the disaster-related tweet classification task. Supervised ML or DL algorithms require labelled data to train classifiers that can be further used for classifying new data. Labelling the training data is typically carried out manually and is, therefore, a time-consuming and expensive process. This poses a major challenge when attempting to use supervised learning algorithms to assist disaster response in the event of a new disaster, as the time and effort needed to label tweets from the disaster prevent timely use of classifiers. However, recently multiple research work made manually labelled datasets such as CrisisNLP, CrisisLex and CrisisMMD freely available online (Imran, Elbassuoni, et al. 2013; Alam, Joty, et al. 2018; Olteanu, Vieweg, et al. 2015; Olteanu, Castillo, et al. 2014). The review article by Kruspe et al. (Kruspe et al. 2020) summarises the details of such datasets. Furthermore, the Incident Streams of Text REtrieval Conference (TREC-IS) editions were designed to provide annotated datasets and bring together academia and industry to research automatically processing social media streams (*TREC-Incident Streams* n.d.).

Word embedding is a key factor in improving the performance of a DL model for a text classification task. Multiple general-purpose word embeddings such as GloVe (Pennington et al. 2014), fastText (Bojanowski et al. 2017) and Word2Vec (Mikolov et al. 2013) and domain-specific word embeddings such as Crisis embedding (D. T. Nguyen et al. 2016) have been proposed. However, there is not much work done to examine the effectiveness of different

[1] Trained weights of the models, Disaster_Tweet_Classification

Table 1. The summary tweet classification studies.

Reference	Classification	Dataset	#Size	Algorithm	Features	Category			Best Accuracy
						In-disaster	Out-disaster	Cross-disaster	
(Stowe, Paul, et al. 2016)	Relatedness	Own data	7,490	SVM*, MaxEnt, NB	uni-grams, Word2Vec	✓			0.72
(Burel and Alani 2018)	Relatedness	CrisisLexT26	28,000	NB, CART, SVM,CNN*	TF-IDF, Word2Vec			✓	0.83
(Khare et al. 2018)	Relatedness	CrisisLexT26	5,931	SVM	Semantic	✓		✓	0.86
(Wiegmann et al. 2020)	Relatedness	7 datasets; AIDR, CrisisLexT6, CrisisLexT26, CrisisNLP, CrisisMMD, Epic Annotations, collection by McMinn et al.	123,166	feed-forward NN*,CNN	BERT, USE	✓		✓	0.98
(D. Nguyen et al. 2017)	Relatedness	CrisisNLP3, CrisisLex, AIDR	21,021	SVM, LR, RF, CNN*	TF-IDF, Word2Vec and Cisis embedings	✓		✓	0.94
(To et al. 2017)	Relatedness	CrisisLexT26, Crowd-Flower10K	10,876	LR	TF-IDF, Word2Vec				
(Madichetty and Sridevi 2019)	Informativeness	CrisisMMD	4,434	SVM,CNN, CNN and ANN*	n-grams	✓			0.76
(Win and Aung 2017)	Informativeness	CrisisLexT26, AIDR	6,780	RF, SVM, NB, LibLinear classifier*	n-grams				0.75
(Parilla-Ferrer et al. 2014)	Informativeness	Own data	4,000	NB, SVM*	BOW	✓			0.87
(Caragea et al. 2016)	Informativeness	CrisisLexT26 (flooding only)	5,577	ANN,SVM,CNN*	n-grams	✓		✓	0.78
(Acerbo and Rossi 2017)	Informativeness	CrisisLexT26		Random Forest			✓		0.76
(Ning et al. 2019)	Informativeness	CrisisLexT26		CNN				✓	0.81
(D. T. Nguyen et al. 2016)	Topical	CrisisNLP		CNN, Bi-LSTM*		✓			0.62

* The model having the best accuracy.
# Total number of tweets in the dataset.

deep learning architectures and different word embeddings in improving tweet classification models (O'Keefe and Alrashdi 2018).

Many approaches for tweet classification focus on particular disaster types (Acerbo and Rossi 2017; Gata et al. 2019; Caragea et al. 2016). For example, Caragea et al. (2016), use flooding datasets extracted from CrisisLexT26 as for both training and testing datasets. Similarly, Gata et al. (2019) train SVM and NB models to detect tweets related to earthquake events. However, only a few studies comprehensively test classification across various disaster types (Cresci et al. 2015; Graf et al. 2018; Wiegmann et al. 2020). Closer to our objective is the work by Graf et al. (Graf et al. 2018) and Wiegmann et al. (Wiegmann et al. 2020). Graf et al. (2018) introduce a cross-domain informativeness classifier based on SVM classifier. The study by Wiegmann et al. (2020) compares the effectiveness of three state-of-the-art machine learning models, namely CNN and two transformer models: BERT and Universal Sentence Encoder (USE) for the related tweet classification task. However, they explicitly consider only cross-disaster types. Also, these approaches have been mostly pursued in academic contexts and have not been made available to the public and responding organisations through easily accessible and integrable tools (Burel and Alani 2018).

METHODOLOGY

We conduct experiments under three settings to evaluate twelve ML models and two DL models with three different word embeddings for the disaster-related tweet classification task.

Dataset

We extracted tweets from Disaster Data Corpus 2020 created by Wiegmann et al. 2020, that includes data from seven repositories, namely, CrisisLex T26 (Olteanu, Vieweg, et al. 2015), CrisisLex T6 (Olteanu, Castillo, et al. 2014), CrisisNLP - RESOURCE # 1 (Imran, Mitra, et al. 2016), CrisisNLP - RESOURCE # 2 (Imran, Elbassuoni, et al. 2013), CrisisNLP - RESOURCE # 5 (Alam, Ofli, et al. 2018), Epic Annotations (Stowe, Palmer, et al. 2018), and the dataset collected by (McMinn et al. 2013). Furthermore, we downloaded additional non event-tagged Kaggle ("Real or Not? NLP with Disaster Tweets") dataset[2] that was originally created by figure-eight [3], and Appen Disaster Response Messages[4] and Kaggle ("Disasters on social media") dataset [5]. Table 2 lists the 46 disasters contained in the datasets considered in this study and the number of Related and Not-Related tweets available for each of them before the preprocessing steps. We assigned each disaster to one of 8 disaster types, based on the work by Wiegmann et al. (2020). The combined dataset has 192, 948 labelled tweets in total.

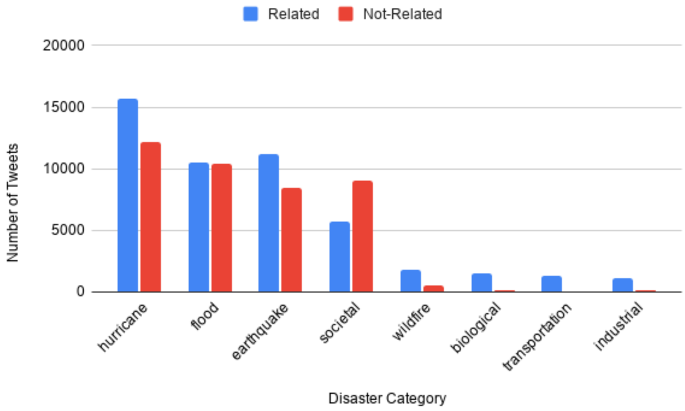


Figure 2. Related and Not-related tweets by category.

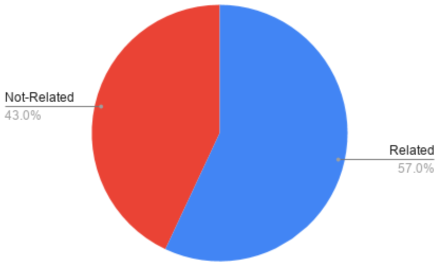


Figure 3. Related and Not-related tweets in the dataset.

During the noise reduction and preprocessing steps, we removed all non-English tweets, duplicate tweets and re-tweets (e.g., RT @username:) using string manipulations in Python Pandas library [6]. Furthermore, all URLs, hashtags, special characters, emoticons, and emojis were removed. Also, we removed stop words, words having less than three characters and sentences having less than three words. After these steps, there were 117, 954 tweets,

[2] Kaggle "Real or Not? NLP with Disaster Tweets" dataset, https://www.kaggle.com/c/nlp-getting-started/overview

[3] Appen Datasets Resource Center, https://www.figure-eight.com/data-for-everyone/

[4] Appen Disaster Response Messages, https://appen.com/datasets/combined-disaster-response-data/

[5] Kaggle "Disasters on social media" dataset, https://www.kaggle.com/jannesklaas/disasters-on-social-media

[6] Python pandas library: https://pandas.pydata.org/

Table 2. Related and Not-Related labelled tweets grouped by disaster category

Name	No of Tweets		Name	No of Tweets	
	Related	Not-Related		Related	Not-Related
Flood (9)			**Wildfires (3)**		
2012 Philipinnes	906	88	2012 Colorado	953	238
2013 Sardinia	926	68	2013 Australia	949	242
2013 Manila	921	47	2014 California	1,245	344
2013 Alberta	6,172	4,856	**Societal (2)**		
2013 Queensland	11,332	10,259	2013 Boston bombing	6,577	4,416
2013 Colorado	925	70	2013 LA airport shootings	912	87
2014 India	1,322	498	**Industrial (4)**		
2014 Pakistan	1,744	25	2012 Venezuela refinery explosion	339	58
2017 Sri Lanka	367	655	2013 West-Texas explosion	6,157	4,825
Earthquake (12)			2013 Brazil nightclub fire	952	40
2012 Costarica	909	399	2013 Savar building collapse	1,141	75
2012 Guatemala	940	108	**Transportation (4)**		
2012 Italy	940	50	2013 Glasgow helicopter crash	918	177
2013 Bohol	969	31	2013 New York train crash	999	0
2013 Pakistan	1,569	312	2013 Spain train crash	991	6
2013 California	1,595	106	2013 LA train crash	966	31
2013 Chile	1,590	342	**Hurricane (11)**		
2015 Nepal	10,583	5,801	2011 Joplin Tornado	1,756	976
2017 Mexico	1,030	350	2012 Hurricane Sandy	6,138	3,870
2017 Iraq and Iran	493	104	2012 Hurricane Pablo	907	68
2018 Nepal	3,410	2,820	2013 Typhoon Yolanda	940	71
Biological (2)			2013 Oklahoma Tornado	5,165	4,827
2014 Ebola	1,559	215	2014 Typhoon Hagupit	1,778	232
2014 Mers	1,331	27	2014 Hurricane Odile	1,219	43
Other (3)			2015 Cyclone Pam	1,508	496
2013 Russia meteor impact	1,133	271	2017 Hurricane Harvey	3,329	1,105
2013 Singapore haze	933	46	2017 Hurricane Maria	2,843	1,713
Kaggle and Appen datasets	24,800	14,725	2017 Hurricane Irma	3,548	956

reducing around 40% of the tweets. We also applied lemmatization to convert words into their root forms as it improves the classification accuracy (Acerbo and Rossi 2017). Figure 2 illustrates the number of Related and Not-Related tweets in each disaster category after the preprocessing stages.

We combined the Related and Informative and Related but not Informative into the *Related* class, and Not Applicable into the *Not-Related* class. For the datasets where there were topical classes, we combined them into *Related* class (e.g., "affected people", "missing trapped or found people"). These two classes were then used for distinguishing crisis-related content from unrelated content for creating binary text classifiers (Khare et al. 2018). Also, we combined the same disaster events across different datasets (e.g., Queensland Floods in CrisisLex and CrisisNLP). Figure 3 presents the distribution of total Related and Not-related tweets in the dataset.

To avoid classification bias towards the majority class, we balanced the data from each category by matching the number of Related tweets with Not-Related ones. For example, after preprocessing, the number of related and not-related tweets of earthquake category were 6,946 and 4,650, respectively. We randomly selected not-related tweets from other categories except for earthquake category and made the dataset such as having 6,946 related and 6,946 not-related tweets.

Models

We selected five supervised ML algorithms that have been mostly explored for disaster tweet classification tasks namely Logistic Regression (LR), Decision Tree (DT), SVM, NB, ANN and RF (Parilla-Ferrer et al. 2014; Kaufhold et al. 2020; Khare et al. 2018). In addition, six more ML models were selected that have rarely been studied for disaster tweet classification tasks in literature such as Gradient Boosting Classifier (GB), RidgeClassifier, AdaBoost, k-Nearest Neighbors (KNN), xgboost, and catboost. All the algorithms were implemented in Python scikit-learn

(Pedregosa et al. 2011), using the default parameters. Furthermore, we used two DL algorithms, namely CNN and Bi-directional LSTM (Bi-LSTM).

Text documents have to be converted into numerical vectors for the machine learning task. Single-dimensional bag-of-word (BOW) model with Term Frequency–Inverse Document Frequency (TF-IDF) representations have been widely adopted for traditional ML algorithms (D. Nguyen et al. 2017), whereas word embeddings such as Word2Vec and GloVe have been used for DL models (O'Keefe and Alrashdi 2018; Burel and Alani 2018). We extracted word-level unigrams from tweets as features for our ML models and converted to TF-IDF vectors by considering each tweet as a document.

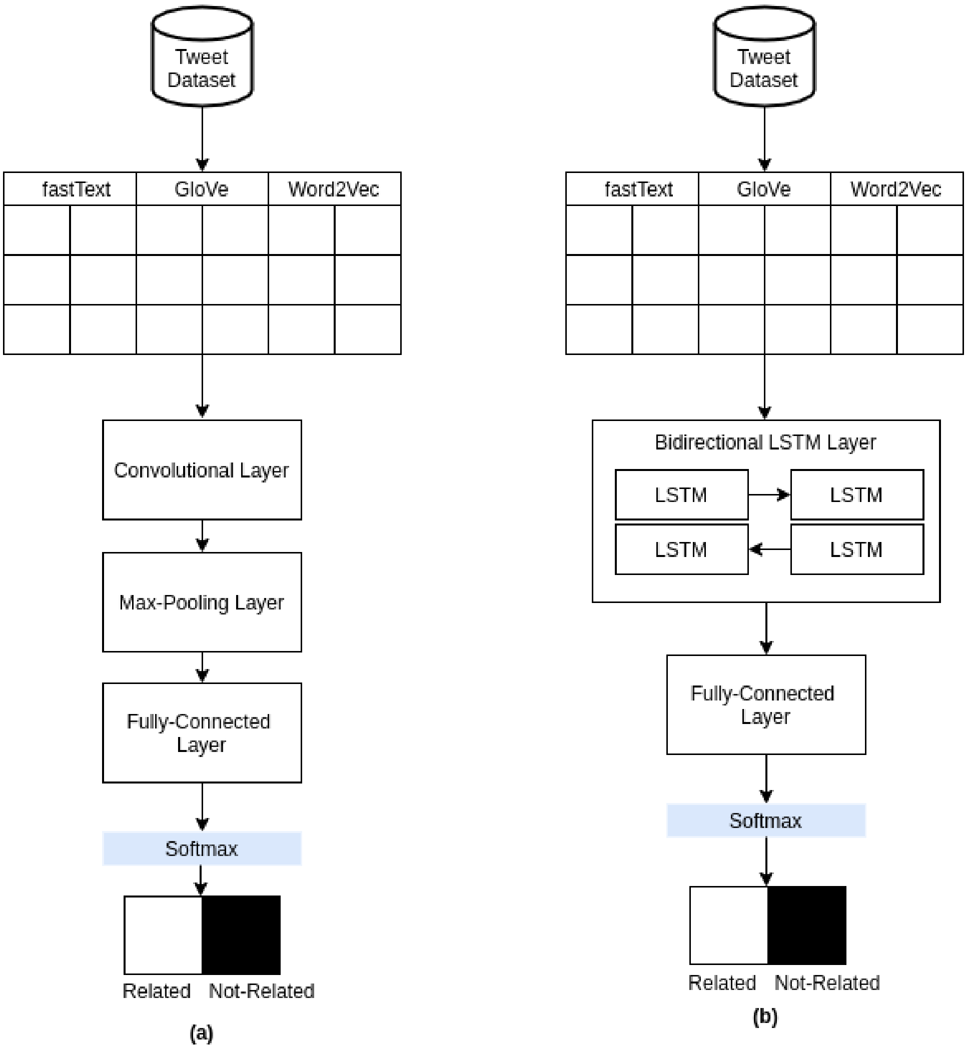


Figure 4. Illustration of (a) CNN and (b) Bidirectional LSTM of twitter text classification task.

Word embeddings generate a vectorized representation of words by mapping words to vectors instead of a one-dimensional space. Therefore, semantically close words should have a similar vector representation instead of a distinct representation. We used pre-trained Word2Vec model of Google News dataset about 100 billion words [7], pre-trained fastText model of Wikipedia 2017, UMBC web base corpus having 999,995 word vectors [8] and pre-trained GloVe embeddings having 2 196,016 vectors [9] as features for our DL models. When embedding, each tweet is represented as a matrix of size $n * k$, where n is the maximum length of a tweet text (number of words) in the training data and k is the embedding vector dimension. We used ($k = 300$) for all three embedding models and applied zero-sequence padding for the tweet texts having the number of words less than n. Kim et al. (2014) described a CNN architecture for text classification tasks and has been mostly adopted in disaster tweet

[7] word2vec pre-trained word vectors, https://code.google.com/archive/p/word2vec/
[8] fastText pre-trained word vectors, https://fasttext.cc/docs/en/english-vectors.html
[9] GloVe pre-trained word vectors, https://nlp.stanford.edu/projects/glove/

classification studies (Kim 2014; Neppalli et al. 2018; Burel and Alani 2018). We adopted a similar model with a single convolution layer followed by a max-over-time pooling layer and a fully connected layer where the softmax function is applied to predict the document classes. Furthermore, our dropout rate was set to 0.5 for regularisation and ran the model for 100 epochs. However, early stopping was used to terminate the execution based on validation accuracy. The Long-Short Term Memory (LSTM) is a specialized version of Recurrent Neural Network (RNN) capable of learning long term dependencies. While LSTM can only see and learn from past input data, Bidirectional LSTM (Bi-LSTM) runs input in both forward and backward directions. This bidirectional feature of Bi-LSTM is critical for the various applications involved with understanding complex language (O'Keefe and Alrashdi 2018). ALRashdi et al. (2019) described a Bi-LSTM model for disaster tweet classification, and we adopted a similar architecture for our experiments. Figure 4 illustrates the architectures of the CNN and Bi-LSTM networks for the tweet classification task.

Experiments

We carry out experiments under the following three settings.

1. In-disaster balanced training dataset (disasters considered: Earthquake, Flood, Hurricane and Societal)
2. Out-disaster balanced training dataset (disasters considered: Earthquake, Flood, Hurricane and Societal)
3. Cross-disaster balanced training dataset (disasters considered: Earthquake, Flood, Hurricane, Societal, Wildfire, Industrial, Transportation and Biological)

Table 3. In-disaster, out-disaster and cross-disaster experimental datsets

In-Disaster	
Train Dataset	Test Dataset
Earthquake	Earthquake (2018 Nepal)
Flood	Flood (2017 Sri Lanka)
Hurricane	Hurricane (2017 Maria)
Societal	Societal (2013 LA airport shootings)

Cross-Disaster	
Train Dataset	Test Dataset
All data	Earthquake (2017 Iraq and Iran)
All data	Flood (2018 Nepal)
All data	Hurricane (2017 Maria)
All data	Societal (2013 LA airport shootings)
All data	Biological (2014 MERS)
All data	Transportation (2013 LA train crash)
All data	Wildfire (2014 California)
All data	Industrial (2013 Brazil nightclubfire)

Out-Disaster	
Train Dataset	Test Dataset
Earthquake	Flood (2017 Sri Lanka)
Earthquake	Hurricane (2017 Maria)
Earthquake	Societal (2013 LA airport shootings)
Flood	Earthquake (2018 Nepal)
Flood	Hurricane (2017 Maria)
Flood	Societal (2013 LA shootings)

Out-Disaster	
Train Dataset	Test Dataset
Hurricane	Earthquake (2018 Nepal)
Hurricane	Flood (2017 Sri Lanka)
Hurricane	Societal (2013 LA shootings)
Societal	Earthquake (2018 Nepal)
Societal	Flood (2017 Sri Lanka)
Societal	Hurricane (2017 Maria)

We formulated our experiments for all three settings such that the tests are applied to the newest disaster dataset. For example, we selected the most recent disasters from each category and used that as the test dataset. In the case of multiple disasters in the same year, we chose the disaster with the fewest tweets as the test dataset. Therefore, our test datasets were; 2017 Sri Lanka floods, 2018 Nepal earthquake, 2014 MERS, 2014 California wildfires, 2013 LA airport shootings, 2013 Brazil nightclub fire, 2013 LA train crash and 2017 hurricane Maria. Hence, before training the algorithms, we removed those entire test datasets from each category to test the models for unseen data. Table 3 lists the training and testing datasets considered for three experiments. We used 10 fold stratified sampling for cross-validation [10] while training the models. The performance of algorithms were measured

[10] Stratified ShuffleSplit cross-validator, `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html`

using average F1-score. Altogether we carried out 36 model training and evaluations for in-disaster category. To reduce the number of training and evaluations, we selected the top three ML models and best DL model based on average F1-score for the out-disaster and cross-disaster experiments. All experiments were executed in the Google Collaboratory [11] environment.

RESULTS AND DISCUSSION

Table 4. Average F1-scores of the DL and ML models for the in-disaster experiments (The best scores are highlighted in grey, and the three best performing ML models and the best performing DL model are underlined).

Algorithm	Hurricane	Societal	Earthquake	Flood
Linear SVM	0.742	0.565	0.810	0.866
RidgeClassifier	0.718	0.571	0.813	0.800
Logistic Regression	0.743	0.576	0.799	0.819
Decision Tree	0.695	0.613	0.769	0.741
k-Nearest Neighbors	0.492	0.512	0.523	0.516
Gradient Boosting Classifier	0.687	0.491	0.676	0.727
NB	0.729	0.737	0.795	0.854
AdaBoost	0.715	0.605	0.731	0.792
Random Forest	0.524	0.687	0.740	0.789
Perceptron	0.628	0.632	0.751	0.753
xgboost	0.716	0.613	0.756	0.767
catboost	0.696	0.537	0.723	0.734
LSTM-fastText	0.770	0.690	0.794	0.914
CNN-fastText	0.791	0.779	0.769	0.905
LSTM-GloVe	0.776	0.787	0.753	0.911
CNN-GloVe	0.766	0.707	0.799	0.898
LSTM-Word2Vec	0.793	0.795	0.820	0.925
CNN-Word2Vec	0.787	0.764	0.804	0.900

Table 5. Average F1-scores of the ML and DL models (LSTM-fastText (DL^{1}), CNN-fastText (DL^{2}), LSTM-GloVe (DL^{3}), CNN-GloVe (DL^{4}), LSTM-Word2Vec (DL^{5}) and CNN-Word2Vec (DL^{6})) for the Out-disaster experiments. The best scores are highlighted in grey.

Algorithm	Earthquake-Flood	Earthquake-Hurricane	Earthquake-Societal	Flood-Earthquake	Flood-Hurricane	Flood-Societal	Hurricane-Earthquake	Hurricane-Flood	Hurricane-Societal	Societal-Earthquake	Societal-Flood	Societal-Hurricane
SVM	0.632	0.531	0.492	0.597	0.530	0.504	0.653	0.827	0.510	0.522	0.519	0.507
LR	0.644	0.532	0.525	0.639	0.539	0.504	0.713	0.831	0.521	0.524	0.518	0.504
NB	0.759	0.522	0.551	0.844	0.641	0.581	0.822	0.841	0.630	0.729	0.789	0.616
DL^{1}	0.844	0.593	0.600	0.833	0.612	0.658	0.732	0.896	0.610	0.677	0.762	0.427
DL^{2}	0.782	0.495	0.602	0.784	0.680	0.622	0.793	0.903	0.566	0.743	0.801	0.469
DL^{3}	0.839	0.610	0.624	0.682	0.624	0.665	0.808	0.911	0.621	0.783	0.730	0.567
DL^{4}	0.837	0.606	0.678	0.742	0.649	0.634	0.783	0.886	0.605	0.735	0.810	0.525
DL^{5}	0.864	0.653	0.657	0.856	0.739	0.677	0.826	0.907	0.644	0.795	0.819	0.634
DL^{6}	0.794	0.583	0.626	0.733	0.740	0.616	0.776	0.886	0.618	0.660	0.730	0.608

Table 4 shows the F1-score for ML and DL algorithms for in-disaster experiments, where scores for the models range from 0.49 to 0.92. The Bi-LSTM model with Word2Vec features performs the best while the KNN algorithm produces the worst results. From the data in Figure 4, it is apparent that the DL algorithms outperform ML

[11] Google Collaboratory, https://colab.research.google.com/

Table 6. Average F1-scores of the ML and DL models (LSTM-fastText (DL^1), CNN-fastText (DL^2), LSTM-GloVe (DL^3), CNN-GloVe (DL^4), LSTM-Word2Vec (DL^5) and CNN-Word2Vec (DL^6)) for the Out-disaster experiments. The best scores are highlighted in grey.

Algorithm	Wildfire	Hurricane	Industrial	Societal	Transport	Biological	Earthquake	Flood
SVM	0.625	0.727	0.535	0.590	0.744	0.797	0.712	0.891
LR	0.616	0.729	0.543	0.642	0.718	0.781	0.638	0.894
NB	0.606	0.612	0.607	0.606	0.674	0.836	0.668	0.898
DL^1	0.793	0.741	0.412	0.513	0.752	0.628	0.729	0.919
DL^2	0.781	0.712	0.563	0.518	0.771	0.720	0.779	0.911
DL^3	0.795	0.743	0.553	0.499	0.771	0.729	0.814	0.902
DL^4	0.763	0.689	0.593	0.559	0.711	0.583	0.708	0.916
DL^5	0.798	0.766	0.608	0.644	0.775	0.804	0.816	0.928
DL^6	0.770	0.716	0.588	0.622	0.770	0.794	0.805	0.896

algorithms having F1-scores over 0.69. It can also be seen from the data that across all the experiments flood tweet dataset has achieved higher F1-values. A possible explanation for this can be the larger number of common words among the flood datasets considered.

Regarding the out-disaster experiments, any DL or ML model trained on hurricane dataset and tested on flood dataset has performed the best while the models trained on societal and applied on hurricane has performed the worst (see Table 5). Overall, ML/DL model's performance applied for an out domain has obtained lower average F1-values, with scores ranging from 0.42 to 0.91. This finding implies that out-disaster experiments need to be carefully designed. Furthermore, choosing DL models over ML models yields better results. Among the DL models, the Bi-LSTM model with Word2Vec embeddings has performed the best.

Table 6 illustrates the results of cross-disaster experiments, where average F1-scores ranging from 0.41-0.93. The Bi-LSTM model with Word2Vec features has achieved the highest F1-scores while KNN algorithm performing the worst. Overall, a model trained on a combined disaster dataset applied on flood data performs the best while industrial and societal categories perform poorly.

The current state-of-art for relatedness classification can be found in (Stowe, Paul, et al. 2016; Burel and Alani 2018; Khare et al. 2018; Wiegmann et al. 2020; D. Nguyen et al. 2017; To et al. 2017). The accuracy scores reported by them are as 0.72 for in-disaster experiments in (Stowe, Paul, et al. 2016), 0.83 for cross-disaster experiments in (Burel and Alani 2018), 0.98 for cross-disaster experiments in (Wiegmann et al. 2020) 0.86 for cross-disaster experiments in (Khare et al. 2018) and 0.94 for in-disaster experiments in (D. Nguyen et al. 2017). It is important to note that these experimental settings are significantly different from ours. For example, in (D. Nguyen et al. 2017) in-disaster experiments contained tweets only from Cyclone PAM and in (Wiegmann et al. 2020) cross-disaster experiments contained data from same disaster category.

In summary, from these results, it is interesting to note that DL models outperform the traditional ML algorithms. This finding supports previous research by (D. Nguyen et al. 2017) and (Burel and Alani 2018) who showed that DL classifiers performed better than all non-DL classifiers. It seems possible that word embedding performs well than the BOW and TF-IDF representations. However, the training time for DL algorithms were higher than the classical ML models. Moreover, the difference in classification time for both approaches was negligible. Among the DL models, the Bi-LSTM model with Word2Vec features has performed the best across all three experimental settings. Another important finding is that with the default parameters, the KNN algorithm has performed the worst for all three experiments. We have considered data-rich disasters (having more than 25,000 tweets in the training dataset) for in-disaster and out-disaster categories while cross-disaster category having a combination. From the data in Figures 4 and 6 it is visible that there is no significant deviation among the results of in-disaster and cross-disaster results. An implication of this is the possibility of using a cross-disaster dataset if the training data unavailable. However, the F1-scores of the out-disaster category are generally lower except for DL models. Therefore, out-disaster experiments have to be carefully designed. The generalisability of these results is subject to certain limitations. For instance, we kept the default parameters for all our ML algorithms. As indicated in the literature, parameter tuning yields better results (Derczynski et al. 2018). Therefore, future research has to be done to identify the best parameters for the ML models. Furthermore, our training datasets are of different sizes as we wanted to explore the maximum possible performance of the individual classifier.

CONCLUSION

This study investigated the identification of the best-performing ML or DL models, to classify disaster-related tweets in three settings: in-disaster, out-disaster and cross-disaster. The research findings suggest that generally, DL models outperform traditional ML models for the tweet classification task. The use of different embedding plays a significant role in text classification. It was shown that the Bi-LSTM model with Word2Vec features performing the best for all three experimental settings considered, namely, in-disaster, out-disaster and cross-disaster.

This is the largest study so far, evaluating ≈0.2 million labelled tweet dataset. The evidence from this study suggests that classifiers can be trained to identify disaster-related tweets in all three categories. However, these findings are limited by using default parameters for the ML algorithms and considering only English tweets. Therefore, in our future work, we plan to identify the best parameters for the experimented models. Furthermore, we will study the capacity of studied models for reproducing or replicating for future researchers.

REFERENCES

Acerbo, F. S. and Rossi, C. (2017). "Filtering informative tweets during emergencies: a machine learning approachfacerbo2017filtering". In: *Proceedings of the First CoNEXT Workshop on ICT Tools for Emergency Networks and DisastEr Relief*, pp. 1–6.

Alam, F., Joty, S., and Imran, M. (2018). "Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.

Alam, F., Ofli, F., and Imran, M. (2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. AAAI Press, pp. 465–473.

Basu, M., Ghosh, K., and Ghosh, S. (2020). "Information Retrieval from Microblogs During Disasters: In the Light of IRMiDis Task". In: *SN Computer Science* 1.1, pp. 1–10.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Burel, G. and Alani, H. (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, Rochester, NY, USA, May 20-23, 2018*. Ed. by K. Boersma and B. M. Tomaszewski. ISCRAM Association.

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.

Cresci, S., Tesconi, M., Cimino, A., and Dell'Orletta, F. (2015). "A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages". In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 1195–1200.

Derczynski, L., Meesters, K., Bontcheva, K., and Maynard, D. (2018). "Helping Crisis Responders Find the Informative Needle in the Tweet Haystack". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. Ed. by K. Boersma and B. M. Tomaszewski. ISCRAM Association.

Gata, W., Amsury, F., Wardhani, N. K., Sugiyarto, I., Sulistyowati, D. N., and Saputra, I. (2019). "Informative Tweet Classification of the Earthquake Disaster Situation In Indonesia". In: *2019 5th International Conference on Computing Engineering and Design (ICCED)*. IEEE, pp. 1–6.

Ghosh, S., Rudra, K., Ghosh, S., Ganguly, N., Podder, S., Balani, N., and Dubash, N. (2019). "Identifying Multi-Dimensional Information from Microblogs During Epidemics". In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pp. 224–230.

Graf, D., Retschitzegger, W., Schwinger, W., Pröll, B., and Kapsammer, E. (2018). "Cross-domain informativeness classification for disaster situations". In: *Proceedings of the 10th international conference on management of digital ecosystems*, pp. 183–190.

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*, pp. 1021–1024.

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, Portorož, Slovenia, May 23-28, 2016*. Ed. by N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, et al. European Language Resources Association (ELRA).

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, p. 102132.

Khare, P., Burel, G., and Alani, H. (2018). “Classifying crises-information relevancy with semantics”. In: *European Semantic Web Conference*. Springer, pp. 367–383.

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

Kruspe, A., Kersten, J., and Klan, F. (2020). “Detection of informative tweets in crisis events”. In: *Natural Hazards and Earth System Sciences Discussions*, pp. 1–18.

Kumar, A., Singh, J. P., and Saumya, S. (2019). “A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification”. In: *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)*. IEEE, pp. 222–227.

Madichetty, S. and Sridevi, M. (2019). “Detecting informative tweets during disaster using deep neural networks”. In: *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE, pp. 709–713.

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*, pp. 409–418.

Meurisch, C., Hamza, Z., Bayrak, B., and Mühlhäuser, M. (2019). “Enhanced Detection of Crisis-Related Microblogs by Spatiotemporal Feedback Loops”. In: *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 1. IEEE, pp. 507–512.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). “Efficient Estimation of Word Representations in Vector Space”. In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Ed. by Y. Bengio and Y. LeCun.

Neppalli, V. K., Caragea, C., and Caragea, D. (2018). “Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters”. In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. Ed. by K. Boersma and B. M. Tomaszewski. ISCRAM Association.

Nguyen, D., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (2017). “Robust classification of crisis-related data on social networks using convolutional neural networks”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1.

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.

Ning, X., Yao, L., Benatallah, B., Zhang, Y., Sheng, Q. Z., and Kanhere, S. S. (2019). “Source-Aware Crisis-Relevant Tweet Identification and Key Information Summarization”. In: *ACM Transactions on Internet Technology (TOIT)* 19.3, pp. 1–20.

O’Keefe, S. E. M. and Alrashdi, R. M. M. (2018). “Deep learning and word embeddings for tweet classification for crisis response”. In: *The 3rd National Computing Colleges Conference*. York.

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 Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. Ed. by E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh. The AAAI Press.

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*, pp. 994–1009.

Parilla-Ferrer, B. E., Fernandez, P., and Ballena, J. (2014). “Automatic classification of disaster-related tweets”. In: *Proc. International conference on Innovative Engineering Technologies (ICIET)*. Vol. 62.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

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)*, pp. 133–143.

Stowe, K., Paul, M., Palmer, M., Palen, L., and Anderson, K. M. (2016). “Identifying and categorizing disaster-related tweets”. In: *Proceedings of The fourth international workshop on natural language processing for social media*, pp. 1–6.

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 international conference on multimedia big data (BigMM)*. IEEE, pp. 330–337.

Toriumi, F. and Baba, S. (2016). “Real-time tweet classification in disaster situation”. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 117–118.

TREC-Incident Streams (n.d.). http://dcs.gla.ac.uk/~richardm/TREC_IS/.

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (2020). “Analysis of Detection Models for Disaster-Related Tweets”. In: *ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management*. ISCRAM Association.

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

Zheng, X. and Sun, A. (2019). “Collecting event-related tweets from twitter stream”. In: *Journal of the Association for Information Science and Technology* 70.2, pp. 176–186.