Social Media Data Analysis Framework for Disaster Response

Víctor Ponce-López (✉ vponcelopez@ucl.ac.uk)

University College London

Catalina Spataru

University College London

Research Article

Keywords: disaster response, machine learning, text analysis, message filtering framework

Posted Date: February 23rd, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1370942/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

This paper presents a social media data analysis framework applied to multiple datasets. The method developed uses machine learning classifiers, where filtering binary classifiers based on deep bidirectional neural networks are trained on benchmark datasets of disaster responses for earthquakes and floods and extreme flood events. The classifiers consist of learning from discrete hand-crafted features and fine-tuning approaches using deep bidirectional Transformer neural networks of these datasets of disaster response. With the development of the multiclassification approach we compare the state-of-the-art results in one of the benchmark datasets containing the largest number of disaster-related categories. The multiclassification approaches developed in this research with support vector machines provide a precision of 0.83 and 0.79 compared to Bernoulli Naïve Bayes which are 0.59 and 0.76 and multinomial Naïve Bayes which are 0.79 and 0.91, respectively. The binary classification methods based on the MDRM dataset show a higher precision with deep learning methods (DistilBERT) than BoW and TF-IDF, while in the case of UnifiedCEHMET dataset show a high performance for accuracy with the deep learning method in terms of severity, with a precision of 0.92 compared to BoW and TF-IDF method which has a precision of 0.68 of 0.70, respectively.

1 Introduction

The recent increase in scale and scope of natural disasters and armed conflicts in the last years has motivated public health interventions in the humanitarian response to considerable gains in equity and quality of emergency assistance [1]. The use and integration of social media in people's daily lives as a new resource to broadcast messages, and the social media data (SMD) analysis, have contributed to deploy new technologies for disaster relief [2]. These contributions mostly refer to the detection of certain patterns usage in people's activity during disasters and natural hazards [3] such as activity volume, recovery information, frequent terms for preparation and recovery, questions about the disaster events, search for safety measures, or situational expressiveness. The identification of these patterns can be potentially addressed through text analysis techniques, natural language processing and machine learning methods.

In this paper, we present a filtering framework approach with its elements to extract raw messages from social media data and to detect those that belong to relevant disaster categories via machine learning classifiers. These classification models are trained on benchmark datasets of disaster response and extreme events. We show a comparison of quantitative results for the different methods on these benchmark datasets with respect to existing works. We highlight the novelty of this work into three key contributions:

- We build a new classification model from benchmark datasets containing the largest number of multiple disaster-related categories relevant to the disaster response domain. To the best of our knowledge, our model outperforms their existing results on multiclassification tasks. Our

multiclassification approaches for comparison are based on different Naïve Bayes [4] methods and Support Vector Machines [5] for text classification.

- We build several binary classifiers for the main disaster-category groups using deep bidirectional neural networks of pre-trained Transformer models, showing clear benefits in performance boost in comparison to traditional hand-crafted feature models. Our deep binary classifier models are based on Distilled BERT [6], which stands for a distilled version of BERT (Bidirectional Encoder Representations from Transformers [7]) and is 60% and 120% faster and smaller than BERT and smaller than ELMo (Embeddings from Language Models [8]) and BiLSTM (Bidirectional Long-Short Term Memory [9]) networks. We present a novel use of these deep learning approaches by fine tuning pretrained models on these benchmark datasets. Moreover, we extend their current evaluation method to be more relevant for the purpose of managing disaster response. This procedure is also applied for the first time to train a deep learning classifier on a large historic dataset of unified records from UK CEH (Centre for Ecology & Hydrology) [10] and MET (Meteorological Office) [11], which describes extreme events with their multiple severity levels.
- Finally, we present a web interface developed to show the use of the framework and its application for the different classification methods. We show an example of the application of the framework for identifying key messages in social media data.

2 Literature Survey Of Key Tools And Platforms For Disaster Response

There exists a comprehensive list of tools, platforms, and tutorials¹ for analysing social media data that are beyond the scope of this article. U.S. Department of Homeland Security (DHS) [12] provides an alphabetized list of them which include product information.

Although demographic data are some of the most common data collected world-wide, in the context of disaster response this type of data is not enough to assist affected population and to guarantee their availability before, during or after disasters [13]. To address these issues, Poblet, Garcia and Casanovas [14] describe a high-level taxonomy to classify the different platforms and apps depending on the phase of the management disaster cycle, availability of the tool and its source code, the main core functionalities, and crowdsourcing role types. Still, the variety of tools in the context of disaster response is considerably high. In this section, we refer to recent key tools suitable for natural disaster response along with their main scope, features, and limitations (Table 1).

Table 1: Description of tools for natural disaster response

Tool / API	Scope	Main features	Main limitations
yourTwapperkeeper [15] (now in Hootsuite ²) / Twitter API	Crisis communication	Open Source, low-cost, simple, flexible and scalable. Retrieves content from search and streaming APIs.	In the new integrated Hootsuite platform, there is a risk of missing early tweets from the streaming API being no longer retrievable.
AIDR [16] / Twitter API & Custom API	Humanitarian response	Open Source. Multiple real-time filtering via geolocation, machine learning classification, and keyword-based. Training examples are domain-customizable.	Web application limits the download to 50,000 tweets per data collection. Specific-trusted training examples are required by the system.
WebSightLine ³ / Custom API	Customized	Multiplatform. Full metadata and indexing. Duplicate detection. Add exclusion and customizable filtering on any field.	Limited source code and uncertain module integration.
Spinn3r [17] (DataStream ⁴) / Custom API	Customized	Multiplatform. 95% of the data indexing requirements in real-time streaming. Reduced cost, complexity and infrastructure on analysing unbounded text data. Easy and secure data stream integration. Machine learning models available and extensible.	Relies on WebSightLine. Uncertain integration. Only demo and free trial available. Oriented to faster development towards product delivery. 2-5 min real-time latency, 180-day historical access.
GNIP [18] / Twitter API Enterprise ⁵	Customized	Full stream available. Real-time and historical social data available through data-driven decisions. Insights to understand content performance.	Limited to Twitter social media.
TweetTracker [19]	Humanitarian and Disaster Relief	Filtering based on keyword and location. Tracking according to hashtags, search terms, and location. Activity comparison across different hashtags, search terms, and locations. Multiple metric visualisations for disaster preparedness and emerging disasters monitoring. Can be paired with different analytical tools (e.g. ORA) to provide richer data insights.	Limited to Twitter social media only. Unspecified availability of twitter streaming. Uncertain information about versions, subscriptions, or availability of source code availability for module integration.
ORA [20, 21, 22]	Dynamic meta-network assessment	Multiplatform. Joint network data import from Facebook accounts and email boxes. Hundreds of social network,	Unspecified availability of twitter streaming. Uncertain information about versions, subscriptions, or

dynamic network and trail metrics, procedures for grouping nodes, local patterns identification, comparing and contrasting networks' group contrast. Networks space-time change analysis. Data connection and location with geo-spatial network metrics, and change detection techniques. Identification of key players, groups and vulnerabilities. Model network changes over time.

availability of source code. Number of nodes, agents and organisations is limited on the ORA-LITE version.

Social Radar [23, 24], CRAFT ⁶ , SORASCS [25, 26]	Flexible and Scalable Disaster Response Systems. Social Radar also performs perception and sentiment analysis.	Multiplatform. Interoperation to create flexible disaster response systems and scalable data storage systems that support social media collection and analysis. Specific workflows' preservation, sharing, and modification. Web-based system chaining together third-party tools for sequential data analyses. Interfaces from a crisis responder's perspective. It can be used as components of larger workflows.	Social Radar and CRAFT do not provide facilities to preserve particular workflows for future use. SORASCS is at a different level of application hierarchy than CRAFT and Social Radar: the user is responsible for supplying a database component themselves. SORASCS has weaker user interfaces from a crisis responder's perspective.
---	---	---	--

[1] <https://towardsdatascience.com/how-to-build-a-real-time-twitter-analysis-using-big-data-tools-dd2240946e64>

[2] <https://www.hootsuite.com/>

[3] <https://websightline.com/>

[4] <https://www.datastreamer.io//>

[5] <https://developer.twitter.com/en/products/twitter-api/enterprise>

[6] <https://craft.co/>

3 Methodology

In this section, we present our proposed framework for the disaster filtering approach. The data pre-processing is discussed in Section 3.1 as part of the data preparation module. Then, in Section 3.2 we present our modelling approaches and category selection for binary and multilabel classification, which will be also supported by their evaluation discussed in Section 5 and model results in Section 6. Finally, the classification examples are presented as part of a developed web application, which shows the steps of the framework architecture in Section 7.

3.1 Pre-processing data for disaster response and text analysis

First, we use a Twitter pre-processing library [27] for cleaning mentions, hashtags, smileys, emojis, or reserved words such as RT; the BS4 library [28] to parse HTML URLs, and regular expression [29] operations in Python to remove anything that is not a letter or a number in all the messages. Then, we prepare the data into training, validation and testing splits for the learning process as explained in Section 5.

Besides the grouping strategy for some categories explained in Section 4.1, for the category 'severity' we treat with the class imbalance problem by augmenting the minority subcategories via grouping moderate and severe events into one single class and reducing the majority class by keeping the mild events into another class.

3.2 Learning Category Models

After pre-processing the data (Section 3.1), we use the method developed in [30] for the implementation of Naïve Bayes (NB) methods and Support Vector Classification (SVC) to learn models for multiclassification in one of the benchmark datasets with the largest number of disaster categories. Besides, we base on the approach from [31] to build binary classifiers for the different target categories based on traditional hand-crafted features: Bag of Words (BoW) and Term Frequency Inverse Document Frequency (TF-IDF); and fine-tuning DistilBERT deep learning models.

3.2.1 Multiclassification

To build the models, we use [30] to implement the Pipeline and Grid Search techniques. The Pipeline consists of a structure to use TF-IDF fractional counts from the encoded tokens. Then, Grid Search is used to perform an exhaustive search over specified parameter values for the given estimators including Multinomial Naïve Bayes and Support Vector Classification. We use these techniques to find the relevant parameters for these desired models.

3.2.2 Binary Classification

BoW and TF-IDF

These two methods use the NLTK library [32] to perform tokenization, by breaking the raw text into words, sentences called tokens, stop words removal filtering out non-informative words, and stemming by reducing a word to its word stem or lemma. The BoW method learns probabilistic models able to distinguish between the class of frequent terms appearing in the messages related to the target category and the class of those terms appearing in the rest of messages. The difference with respect to the TF-IDF method is that the TF-IDF learns probabilistic models able to compensate the BoW models by terms that appear in the inverse of number of documents. Based on the approach provided in [31] to train these

models for each target class, we then, process the input test message into terms and classified to the class that gives a higher probability of its terms to be present.

DistilBERT

We use the Distilled BERT [6] method through the Python library Transformers [33]. This involves an end-to-end tokenization, punctuation, splitting, and wordpiece based on the pretrained BERT base-uncased model. Then, we train the DistilBERT model transformer with a sequence classification/regression head on top for General Language Understanding Evaluation (GLUE) tasks [34]. We set the maximum of training epochs to 10, and the number of global steps to evaluate and save the final models varies depending on the target category which the models are fine-tuned for.

4 Data

We consider a set of annotated datasets to learn and evaluate our machine learning models via feature-based and deep learning methods. To the best of our knowledge, this is the first time these datasets are used and evaluated for purposes related to crisis and disaster response. The following items describe the datasets:

- The Social Media Disaster Tweets (SMDT) [35] dataset consists of 10,876 messages from Twitter. Each message was annotated to distinguish between messages whose content is related to disasters and the rest of messages. There are 4,673 messages out of the total which are labelled as related to disasters, and 6,203 messages annotated as not relevant to disasters. No data splits are provided for evaluation.
- The Multilingual Disaster Response Messages (MDRM) [36] dataset contains 30,000 messages drawn from events including an earthquake in Haiti in 2010, an earthquake in Chile in 2010, floods in Pakistan in 2010, super-storm Sandy in the U.S.A. in 2012, and news articles spanning many years and 100s of different disasters. The data have been encoded with 36 different categories related to disaster response and have been stripped of messages with sensitive information in their entirety. The input data in this job contains thousands of untranslated disaster-related messages and their English translations. Upon release, this is the featured dataset of a new Udacity course on Data Science and the AI4ALL summer school and is especially utile for text analytics and natural language processing (NLP) tasks and models. To the best of our knowledge, this is the largest annotated dataset of disaster response messages. The data contains 20,316 messages annotated as related to disasters. Pre-established training, validation and testing splits are provided for this dataset. For these disaster-related messages, 2,158 are related to floods, 16,232 are related to medical information, and 9,010 are related to humanitarian standards' information. The medical-related information groups the categories 'aid-related', 'medical products', 'other aid', 'hospitals', and 'aid centers'. The information related to humanitarian standards groups the categories 'medical help', 'water', 'food', and 'shelter'. This grouping criteria has been done according to the information needs from Sphere's humanitarian standards [37].

- The long-term dataset of unified records of extreme flooding events reported by the UK Centre for Ecology and Hydrology (CEH) and the UK Meteorological Office (MET) between 1884-2013 [38] – from now abbreviated as ‘UnifiedCEHMET’ – is a unique dataset of 100+ year records of flood events and their consequences on a national scale. Flood events were classified by severity based upon qualitative descriptions. There is an increase in the number of reported flood events over time associated with an increased exposure to flooding as floodplain areas were developed. The data were de-trended for exposure, using population and dwelling house data. The adjusted record shows no trend in reported flooding over time, but there is significant decade to decade variability. The study opens a new approach to considering flood occurrence over a long-time scale using reported information (and thus likely effects on society) rather than just considering trends in extreme hydrological conditions. It contains a total of 1821 records with three impact categories of mild, moderate, and severe events. From these total number of records, we can find 661 records with descriptive messages - 464 for mild events, 69 for moderate events, and 91 for severe events.

5 Evaluation Of The Machine Learning Models

In the case of the MDRM dataset, training, validation, and testing splits are provided by the dataset [36]. We use these provided splits to train our binary classifiers with hand-crafted features and deep bidirectional Transformer neural networks. All the splits are created ensuring a suitable balance ratio between positive and negative class samples for the different target categories ‘disaster’, ‘medical’, ‘humanitarian standards’, and ‘severity’, which are controlled by an empirically set parameter.

However, for the multiclassification approach we use the entire MDRM dataset to perform the evaluation using splits of 33% for testing and 66% for training the models. Then, we train our classifiers using 5-fold cross validation and compute the average scores for each metric. This is done to make results comparable with [39], given that all the results we found for this dataset do not follow the proposed splits provided by the dataset on the multiclassification task. Besides, there is no information specified in those results about any validation strategy performed in their evaluation, such as the k -fold cross validation which we do perform to provide with more reliable results. Therefore, we use this evaluation only for comparison purposes with existing reported results for this dataset. Figure 2 plots the learning curve to show the scalability of the SVC multiclassification approach. One can note the quick improvement of the training process when we train the model until 8000 samples, and the saturation with no further improvements when adding more samples after that point.

We perform an additional evaluation with the original form for the splits of the MDRM dataset to enhance current results and to allow future comparisons to users. To the best of our knowledge, there are no previous published results for this dataset in spite of the dataset being also publicly available with its original splits in competition platforms such as Kaggle⁷. Therefore, we provide for the first time a reliable machine learning evaluation for this dataset on the disaster response domain. Figure 3 plots the learning curves to show the performance and scalability of our NB-based approaches and the SVC multiclassification approach. In this case, the improvement is progressively slower during the training

process as we add more samples, until the point where the performance on the validation data is close to the performance on the training data. This is due to the reduced number of validation samples in this setting of original splits of the dataset with respect to the above setting using custom splits. In this setting, we see a considerable improvement in the learning process in all our methods, especially in our SVC model.

Since there are no splits provided for the other datasets, in all remaining settings we generate the splits via random sampling with 40% for testing and 60% for training the binary classifiers.

The results are provided in terms of precision, recall, F1-score, and -for binary classifiers- accuracy. Note we do not provide results for accuracy in test data in the multiclassification because when the class distribution is unbalanced, accuracy metric is considered a poor choice as it gives high scores to models which just predict the most frequent class.

To calculate the above metrics we considered the method developed in [30]. These metrics are essentially defined for binary classification tasks, which by default only the positive label is evaluated, assuming the positive class is labelled '1'. In extending a binary metric to multiclass or multilabel problems, the data is treated as a collection of binary problems, one for each class. There are then several ways to average binary metric calculations across the set of classes, each of which may be useful in some scenario. We select the 'micro' average parameter because "micro" it gives each sample-class pair an equal contribution to the overall metric (because of sample-weight). Rather than summing the metric per class, this sums the dividends and divisors that make up the per-class metrics to calculate an overall quotient. Micro-averaging may be preferred in multilabel settings, including multiclass classification where a majority class is to be ignored.

Intuitively, precision is the ability of the classifier not to label as positive a sample that is negative, and recall is the ability of the classifier to find all the positive samples. The F-measure can be interpreted as a weighted harmonic mean of the precision and recall. A F measure reaches its best value at 1 and its worst score at 0. With F1-score, the recall and the precision are equally important. Their values are between 0 and 1 and higher is better. The average precision (AP) is computed from prediction scores as:

$$AP = \sum_n (R_n - R_{n-1}) P_n(1)$$

where P_n and R_n are the precision and recall at the n -th threshold and are calculated with the true positives, false positives, and false negative predictions [30]. With random predictions, the AP is simply the fraction of positive samples.

The accuracy metric is used to compute either the fraction or the count of correct predictions. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is

1.0; otherwise it is 0.0. If $\hat{y}_i$ is the predicted value of the i -th sample and y_i is the corresponding true value, then the fraction of correct predictions over $n_{samples}$ is defined as:

$$acc(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (2)$$

[7] <https://www.kaggle.com/landlord/multilingual-disaster-response-messages>

6 Models And Results

We use the MDRM dataset described in Section 4.1 to learn classifiers able to discriminate between the samples annotated into the categories of this dataset.

In Table 2, we compare the results of our multiclass models based on Multinomial NB and SVC with existing results [39] based on Bernoulli NB classification⁸ for our custom splits of the dataset. However, we report results of our 5-fold cross validation to provide with a more reliable evaluation, showing that our models still outperform the existing reported results trained on this dataset. For simplicity, we do not show all results for each category, but we report the micro average results amongst all categories for precision, recall and F1-score, respectively.

Table 2
Comparison of multiclassification methods using custom splits

Micro Avg. scores	MDRM dataset		
	Bernoulli NB	Multinomial NB	SVC
Precision	0.59	0.79	0.83
Recall	0.57	0.51	0.57
F1-Score	0.58	0.61	0.67

There are multiple versions with similar codes available⁹ using NB-based multiclassification models where classification results have not been reported for this dataset. Although the MDRM dataset has been recently incorporated into the Kaggle platform¹, there are no reporting results on the original splits of the dataset. Results from our evaluation as specified in Section 5 are reported in Table 3, showing again an outperformance of our method in terms of precision, in this case achieved with the Multinomial NB approach despite the accuracy scores being higher in the learning process with the SVC method. From the outcomes of these results, we encourage future authors who want to train new machine learning models on this dataset for the disaster response domain to use both validation strategies for the purpose of proper model selection using both custom and original splits.

Table 3
Comparison of multiclassification methods using original splits

Micro Avg. scores	MDRM dataset		
	Bernoulli NB	Multinomial NB	SVC
Precision	0.76	0.91	0.79

In our binary classification, Table 4 and Table 5 show a clear outperformance of the DistilBERT deep learning models at identifying the four main disaster categories on both MDRM and UnifiedCEHMET datasets. We use the resulting models fine-tuned on these datasets for our filtering approach on the remainder data collections. The hand-crafted feature methods BoW and TF-IDF are taken as a baseline model to show initial performances in all three datasets. However, note we skip to applying the DistilBERT method to the SDMT dataset due to its reduced number of total samples to learn more generic deep learning models. Thus, the resulting disaster model we use for filtering is the one fine-tuned on the MDRM dataset, and likewise for the medical and humanitarian standards' models. The results of these experiments especially highlight the ability of DistilBERT method at learning models despite it is dealing with smaller datasets. This improvement can be shown the recall metric for the humanitarian standards' category and, especially, all four metrics for the severity category. To the best of our knowledge, it is the first time these datasets are used to train relevant disaster models for these disaster-related categories via deep learning NLP-based methods.

Table 4
Comparison of binary classification on SMDT and MDRM datasets for the disaster category

Target category	Disaster				
Dataset	SMDT		MDRM		
Method / Metric	BoW [31]	TF-IDF [31]	BoW [31]	TF-IDF [31]	Distil BERT [6]
Precision	0.82	0.80	0.89	0.89	0.91
Recall	0.62	0.63	0.89	0.91	0.95
F-score	0.70	0.71	0.89	0.90	0.93
Accuracy	0.77	0.77	0.82	0.83	0.87
* The DistilBERT method is not applied to the SDMT dataset due to its reduced number of total samples.					

Table 5

Comparison of binary classification on MDRM and UnifiedCEHMET datasets for the disaster categories medical, humanitarian standards, and severity

Dataset	MDRM						UnifiedCEHMET		
Target category	Medical			Humanitarian Standards			Severity		
Method / Metric	BoW [31]	TF-IDF [31]	Distil BERT [6]	BoW [31]	TF-IDF [31]	Distil BERT [6]	BoW [31]	TF-IDF [31]	Distil BERT [6]
Precision	0.73	0.67	0.80	0.73	0.84	0.83	0.34	0.39	0.78
Recall	0.73	0.78	0.83	0.38	0.24	0.75	0.39	0.51	0.92
F-score	0.73	0.72	0.81	0.50	0.38	0.78	0.37	0.44	0.84
Accuracy	0.72	0.69	0.80	0.75	0.74	0.86	0.68	0.70	0.92

[8] Public repository <https://github.com/ioslilyng/DisasterResponse>

[9] <https://github.com/agpt8/Disaster-Response-Classification>

7 Framework Architecture For The Web Application

In this section, we present a developed web application which shows the architecture of the framework, visual steps of the employed classification approaches, and some examples of usage with input texts.

Figure 4 and Fig. 5 show the methodology described in section 3.2 for the multi-classification and binary classification and the selection of categories for each approach. This is the core part of our disaster response framework, which is followed by the evaluation and results of the models in Sections 5 and 6, respectively. A list of 34 categories of the MDRM dataset (Fig. 4) has been defined together with the percentage distribution of each category according to their number of sample messages.

Figure 6 and Fig. 7 show examples of the web application interface for an input target message we aim to classify to the 34 disaster-related categories and to the four main disaster-related categories, respectively. Note the deep classification shows the confidence scores of each Transformer model.

[10] An explanatory video of the framework and web application is attached along with this paper as a supplementary material.

8 Conclusion

This paper provides a novel and rigorous framework to analyze and identify key text information revealed in social media related to disasters. We introduce a survey with our selection of tools and platforms for disaster response among the high variety of options. In our framework, we emphasize the need of further evaluation of machine learning classifiers in this domain on benchmark datasets both in multiclassification and in binary classification settings. We present a new state of the art performance evaluations and results on these datasets achieved with SVC and Multinomial NB approaches depending on the settings of the dataset. Furthermore, this study provides a novelty approach in usage of these methods, including deep bidirectional Transformer neural networks, to train several models from benchmark datasets containing the largest annotated messages of disaster-related categories. This paper also delivers a specific categorization strategy for the main disaster categories to train our deep learning models along with their strategy for performance evaluation. Our developed interface shows this combination of deep learning models. We conclude from both methodology and empirical application results that score shows potential to explain informative and influential information in a variety of large-scale data sets. This is worth noticing because such information leads to higher prediction performances, thus helping to improve accuracy of classification methods for message filtering purposes. Finally, we show our web interface to be used for several potential applications in humanitarian response, including management of crisis or natural hazards, needs of resource allocation, or emergency assistance.

Declarations

Data availability

The data sets utilized in this study can be found in the reference works as following:

- Social Media Disaster Tweets (SMDT) can be found in [35].
- Multilingual Disaster Response Messages can be found in [36].
- Unified Records of extreme flooding events reported by the UK Centre for Ecology and Hydrology (CEH) and the UK Meteorological Office (MET) between 1884-2013 can be found in [38].

Code availability

The source codes both for the core of the filtering framework and for the web interface are available at <https://github.com/islandslab/NLP-Disaster>.

Funding

This research was funded by Belmont Forum's first disaster-focused funding Call Belmont Collaborative Research Action 2019: Disaster Risk, Reduction and Resilience (DR32019) which was supported by the Ministry of Science and Technology (MOST) of Chinese Taipei in partnership with funders from Brazil (FAPESP), Japan (JST), Qatar (QNRF), UK (UKRI), US (NSF), CNR (Italy). In particular, this research was funded by UKRI grant EP/V002945/1.

Affiliations:

University College London, UCL, United Kingdom

Dr. Víctor Ponce-López & Prof. Catalina Spataru

Contributions:

Conceptualization VPL and CS; Formal Analysis: VPL with input from CS; Methodology VPL and CS, Software VPL with CS supervision; writing- structure CS and VPL, writing -review and editing VPL and CS, Funding acquisition: CS. Both authors read and approved the final manuscript.

References

1. J. Leaning and G. S. Debarati, "Natural Disasters, Armed Conflict, and Public Health," *New England Journal of Medicine and Public Health*, vol. 369, no. 19, pp. 1836-1842, 2013.
2. P. M. Landwehr and K. M. Carley, "Social Media in Disaster Relief: Usage Patterns, Data Mining Tools, and Current Research Directions," *Data Mining and Knowledge Discovery for Big Data. Studies in Big Data*, vol. 1, pp. 225-257, 2014.
3. M. T. Niles, B. F. Emery, A. J. Reagan, P. S. Dodds and C. M. Danforth, "Social media usage patterns during natural hazards," *PLOS ONE*, vol. 14, no. 2, pp. 1-16, 2019.
4. C. D. Manning, P. Raghavan and H. Schuetze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
5. J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, 1999.
6. V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS*, 2019.
7. J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.
8. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep Contextualized Word Representations," in *Proceedings of NAACL-HLT*, New Orleans, Louisiana, 2018.
9. A. Wang, I. F. Tenney, Y. Pruksachatkun, P. Yeres, J. Phang, H. Liu, P. M. Htut, K. Yu, J. Hula, P. Xia, R. Pappagari, S. Jin, R. T. McCoy, R. Patel, Y. Huang, E. Grave, N. Kim, T. Févry, B. Chen, N. Nangia, A. Mohananey, K. Kann, S. Bordia, N. Patry, D. Benton, E. Pavlick and S. R. Bowman, "jiant 1.3: A software toolkit for research on general-purpose text understanding models," 2019.
10. CEH, "UK Centre for Ecology & Hydrology," [Online]. Available: <https://www.ceh.ac.uk/>.
11. Metoffice, "UK Meteorological Office," [Online]. Available: <https://www.metoffice.gov.uk/>.

12. Space and Naval Warfare Systems Center Atlantic, U.S. Department of Homeland Security (DHS), "Innovative Uses of Social Media in Emergency Management: System Assessment and Validation for Emergency Responders (SAVER)," 2013.
13. National Research Council, Tools and Methods for Estimating Populations at Risk from Natural Disasters and Complex Humanitarian Crises, Washington, DC: The National Academies Press, 2007.
14. M. Poblet, E. C. García and P. Casanovas, "Crowdsourcing Tools for Disaster Management: A Review of Platforms and Methods," in *International Workshop on AI Approaches to the Complexity of Legal Systems*, Berlin, Heidelberg, 2013.
15. A. Bruns and Y. E. Liang, "Tools and methods for capturing Twitter data during natural disasters," *First Monday*, vol. 17, no. 4, pp. 1-8, 2012.
16. M. Imran, F. Ofli and F. Alam, "AIDR: Artificial Intelligence for Digital Response," Qatar Computing Research Institute, 2013. [Online]. Available: <http://aidr.qcri.org/>.
17. Spinn3r, "API Documentation," 2016. [Online]. Available: <http://docs.spinn3r.com/>.
18. GNIP, *Grand Central Station for the Social Web*, ReadWriteWeb, 2008.
19. S. Kumar, G. Barbier, M. Abbasi and H. Liu, "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2021.
20. N. Altman, K. M. Carley and J. Reminga, "ORA User's Guide 2020," CMU-ISR-20-110, 2020.
21. K. M. Carley, ORA: A Toolkit for Dynamic Network Analysis and Visualization, R. J. (. Alhajj R., Ed., New York, NY: Encyclopedia of Social Network Analysis and Mining, Springer, 2014.
22. A. Ujaway-Gil, Organizational Network Analysis: Auditing Intangible Resources. 1st Edition. Routledge., 1st ed., Routledge, 2019.
23. B. Costa and J. Boiney, "Social Radar. MITRE, McLean, Virginia, USA," The MITRE Corporation, McLean, 2012.
24. J. Mathieu , M. Fulk, M. Lorber M, G. Klein, B. Costa and D. Schmorow, "Social Radar Workflows, Dashboards, and Environments," The MITRE Corporation, Bedford, 2012.
25. B. Schmerl, D. Garlan, V. Dwivedi, M. W. Bigrigg and K. M. Carley, "SORASCS: A Case Study in SOA-based Platform Design for Socio-Cultural Analysis," in *Proceedings of the 33rd International Conference on Software Engineering*, Waikiki, Honolulu, 2011.
26. D. Garlan, B. Schmerl, V. Dwivedi, M. W. Bigrigg and K. Carley, "Specifying Workflows in SORASCS to Automate and Share Common HSCB Processes," in *Proceedings of the HSCB Focus 2011: Integrating Social Science Theory and Analytic Methods for Operational Use*, Chantilly, VA, 2011.
27. S. Özcan, "Tweet-Preprocessor," [Online]. Available: <https://pypi.org/project/tweet-preprocessor/>.
28. L. Richardson, "Beautiful Soup Documentation," [Online]. Available: <https://www.crummy.com/software/BeautifulSoup/>.
29. J. Friedl, Mastering Regular Expressions. 3rd ed., O'Reilly Medi, 2009.
30. F. Pedregosa , G. Varoquaux, A. Gramfort, V. Michel, B. Thirion , O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E.

- Duchesnay, "scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
31. V. Chavda, "tweet classification," [Online]. Available: https://github.com/pointoflight/tweet_classification.
 32. S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python*, O'Reilly Media Inc, 2009.
 33. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.
 34. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
 35. Figure Eight, "Social Media Disaster Tweets," [Online]. Available: <https://www.figure-eight.com/data-for-everyone/>.
 36. Appen, "Multilingual Disaster Response Messages," [Online]. Available: <https://appen.com/datasets/combined-disaster-response-data/>.
 37. Sphere Association, *The Sphere Handbook: Humanitarian Charter and Minimum Standards in Humanitarian Response*, fourth edition ed., Geneva, Switzerland: Core Humanitarian Standard on Quality and Accountability © CHS Alliance, 2018.
 38. A. J. Stevens, D. Clarke and R. J. Nicholls, "Trends in reported flooding in the UK: 1884–2013," *Hydrological Sciences Journal*, vol. 61, no. 1, pp. 50-63, 2016.
 39. L. Ng, "A Machine Learning Pipeline for Disaster Response," 2020. [Online]. Available: <https://github.com/lng15/DisasterResponse>.

Figures

Figure 1

Schematic Diagram of the proposed Disaster Filtering Framework.

Figure 2

Learning curve and scalability of the SVC multiclassification approach with confidence scores for the 5-fold cross validation of the MDRM dataset using custom splits.

Figure 3

Learning curve and scalability of the multiclassification approaches (a) Bernoulli Naïve-Bayes, (b) Multinomial Naïve-Bayes, and (c) Support Vector Classification, using custom splits of the MDRM dataset.

Figure 4

Multi-classification approach to learn a SVC model to distinguish 34 disaster categories of the MDRM dataset. The disaster-related categories are described and shown in a piechart, showing a percentage distribution of sample messages for each category.

Figure 5

Binary classification approach of deep learning models with bidirectional Transformer neural network models. Each model is trained for one out of four groups of main disaster categories: disaster-related, humanitarian standards, medical, and severity. The deep model of severity is trained on the UnifiedCEHMET dataset, and we show the amount of messages for each grouped category of severity levels.

Figure 6

Example of multi-classification for a test input message: "Delmos 33 in Silo, need water."

Figure 7

Example of deep classification for a free input message: "People need aid and water urgently after the serious damage caused by extreme heat, wildfires and storm".

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [workflowdisasterfull.mp4](#)