Discover Artificial Intelligence

Discover

Research

Social media data analysis framework for disaster response

Víctor Ponce-López[1] · **Catalina Spataru**[1]

Received: 17 February 2022 / Accepted: 25 May 2022
Published online: 06 June 2022

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 handcrafted features and fine-tuning approaches using deep bidirectional Transformer neural networks on these disaster response datasets. With the development of the multiclass classification approach, we compare the state-of-the-art results in one of the benchmark datasets containing the largest number of disaster-related categories. The multiclass classification 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 and 0.70, respectively.

Keywords Disaster response · Machine learning · Text analysis · Message filtering framework

1 Introduction

The recent increase in the scale and scope of natural disasters and armed conflicts in recent 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 social media data (SMD) analysis have contributed to deploying 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 potentially be addressed through text analysis techniques, natural language processing and machine learning methods.

In the context of disaster response, there is both a reduced number of available benchmark datasets and a lack of evaluation approaches to ensure the robustness and the generalisation capabilities of supervised machine learning approaches. These limitations present important tasks to address most of the challenges described earlier. Therefore,

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44163-022-00026-4.

✉ Víctor Ponce-López, v.poncelopez@ucl.ac.uk; Catalina Spataru, c.spataru@ucl.ac.uk | [1]UCL Energy Institute, University College London, London, UK.

further research and development of frameworks become a need to effectively train meaningful and robust models. Indeed, this is a crucial step (1) to filter the massive-unconstrained textual information from SMD collections and (2) to inform automatically about certain message categories which are relevant to the disaster domain. These are the two main research challenges addressed in this work.

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 applied to these benchmark datasets with respect to existing works. We highlight the novelty of this work in four 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 multiclass classification tasks.
- We present new state-of-the-art results with the original form of the largest disaster-response dataset, in contrast to the current evaluations available. This is to provide with a more reliable validation strategy and results for future comparisons of machine learning approaches in the disaster domain.
- We build several binary classifiers for the main disaster category groups using deep bidirectional neural networks of pretrained Transformer models, showing clear benefits in performance boost in comparison to traditional handcrafted feature models. We present a novel use of these deep learning approaches by fine-tuning pre-trained models on these benchmark datasets. 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) [4] and MET (Meteorological Office) [5], which describes extreme events with their multiple severity levels.
- Finally, we present a web interface developed to illustrate the use of the framework with an example of application for the different classification methods.

This paper is organized as follows. Section 2 presents a summary of the most relevant tools for disaster response and their underlying APIs along with their main features and limitations. Bearing this in mind, we highlight the research gap addressed in our proposed framework. In Sect. 3, we present our methodology for the data pre-processing and our approaches for model category learning. Sections 4 and 5 provide a concise description of the benchmark datasets used to train our machine learning models and their evaluation, respectively. Moreover, in Sect. 5 we propose an extension of the evaluation approach for the multiclass classification approach to address a fundamental research gap to train robust multiclass models for the purpose of disaster response. Section 6 describe the models utilised and the results obtained, followed by a discussion of the results presented with respect to the state-of-the-art in Sect. 7. Finally, we show an example of usage of our models with a developed web interface in Sect. 8. Section 9 concludes the paper.

2 Literature survey of key tools and platforms for disaster response

There exists a comprehensive list of tools, platforms, and tutorials[1] for analysing social media data that are beyond the scope of this article. U.S. Department of Homeland Security (DHS) [6] provides an alphabetized list of them that includes product information.

Although demographic data are some of the most common data collected worldwide, in the context of disaster response this type of data is not enough to assist affected populations and to guarantee their availability before, during or after disasters [7]. To address these issues, Poblet et al. [8] 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. Nevertheless, 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).

Due to the large number of studies in the literature of speciality, tools developed recently have been selected, and in line with the main features and scope of our framework, or those having functionalities that complement each other

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

Table 1 Description of tools for natural disaster response

Tool / API	Scope	Main features	Main limitations
yourTwapperkeeper [30] (now in Hootsuite[a])/ 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 [31]/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[b]/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 [32] (DataStreamer[c]/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 [33]/Twitter API Entreprise[d]	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 [34]	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 [35–37]	Dynamic meta-network assessment	Multiplatform. Joint network data import from Facebook accounts and email boxes. Hundreds of social network, 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	Unspecified availability of twitter streaming. Uncertain information about versions, subscriptions, or availability of source code. Number of nodes, agents and organisations is limited on the ORA-LITE version

Table 1 (continued)

Tool / API	Scope	Main features	Main limitations
Social Radar [38, 39], CRAFT, SORASCS [40, 41]	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

[a] https://www.hootsuite.com/

[b] https://websightline.com/

[c] https://www.datastreamer.io/

https://developer.twitter.com/en/products/twitter-api/enterprise

[d] https://craft.co/

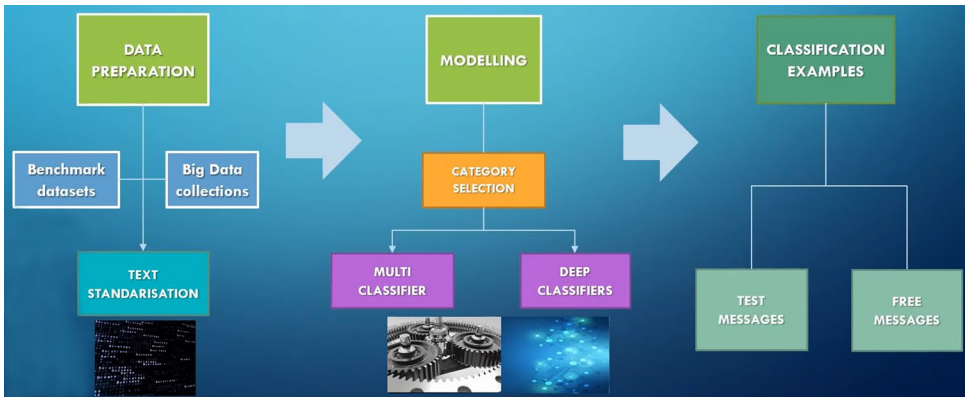

Fig. 1 Schematic Diagram of the proposed Disaster Filtering Framework

to address specific joint applications for disaster response. However, one may notice the literature does not provide a holistic framework for SMD supported by rigorous machine learning evaluations. In this paper, our goal is to address some of the key limitations on SMD analysis and evaluation modelling in a holistic manner, as the main contribution to the state-of-the-art in artificial intelligence for the disaster domain.

3 Methodology

In this section, we present our proposed framework for the disaster filtering approach illustrated in Fig. 1. The data preprocessing is discussed in Sect. 3.1 as part of the data preparation module. Then, in Sect. 3.2, we present our modelling approaches and category selection for binary and multilabel classification. We base on the principles from the Sphere's humanitarian standards [9] to abstract higher-level categories of information needs into simpler categories that are relevant to the disaster domain. This is described in detail as category mining in Sect. 3.2.3. We use the messages that belong to the original and simplified categories for training our multiclass and deep binary models, respectively. These learning approaches are supported by their evaluation discussed in Sect. 5 and model results in Sect. 6. Finally, the classification examples are presented as part of a developed web application, which shows the steps of the framework architecture in Sect. 0.

3.1 Preprocessing data for disaster response and text analysis

First, we use a Twitter preprocessing library [10] for cleaning mentions, hashtags, smileys, emojis, or reserved words such as RT; the BS4 library [11] to parse HTML URLs, and regular expression [12] operations in Python to remove anything that is not a letter of a number in all the messages. Then, we prepare the data into training, validation and testing splits for the learning process as explained in Sect. 5.

In addition to the grouping strategy for some categories explained in Sect. 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 preprocessing the data (Sect. 3.1), first we base on Naïve Bayes (NB) and Support Vector Classification (SVC) approaches for multiclass classification in one of the benchmark datasets with the largest number of disaster categories. Then, we build binary classifiers for different target categories based on traditional handcrafted features followed by the proposed fine-tuned DistilBERT deep learning models.

3.2.1 Multiclass classification

Our multiclass classification approaches for comparison are based on naïve Bayes [13] methods and support vector machines [14] for text classification. To build the models, we use [15] to implement the Pipeline and Grid Search techniques. The Pipeline consists of a structure that uses 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

3.2.2.1 BoW and TF-IDF We base on the approach from [16] 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). These two methods use the NTLK library [17] to perform tokenization by breaking the raw text into words, sentences called tokens, stop words removal filtering out noninformative 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 the number of documents. Based on the approach provided in [16] to train these models for each target class, we then process the input test message into terms and classify it into the class that gives a higher probability of its terms being present.

3.2.2.2 DistilBERT Our deep binary classifier models are based on Distilled BERT [18], which stands for a distilled version of BERT (Bidirectional Encoder Representations from Transformers [19]) and is 60% faster than BERT and 120% faster and smaller than ELMo (Embeddings from Language Models [20]) and BiLSTM (Bidirectional Long-Short Term Memory [21]) networks. DistilBERT is based on the compression technique known as knowledge distillation, from [22, 23], in which a compact learning model is trained to reproduce the behaviour of a larger learned model. These learning and learned models are also called the student and the teacher, respectively, although they can consist of an ensemble of models. In supervised machine learning, a model performing well on the training set will predict an output distribution with high probability on the correct class and with near-zero probabilities on other classes. The knowledge distillation is based on the idea that some of these 'near-zero' probabilities are larger than others and reflect, in part, the generalization capabilities of the model and how well it will perform on the test set.

This lightweight version of BERT representations into DistilBERT models presents clear advantages in terms of speeding up their training processes without a noticeable lose of performance. It is particularly useful in practice when computational resources are limited, but also to ease transferability and reproducibility for both research and development of applications beyond the disaster domain.

We use the Distilled BERT [18] method through the Python library Transformers [24]. This involves end-to-end tokenization, punctuation, splitting, and wordpiece based on the pre-trained BERT base-uncased model. Then, we fine-tune the DistilBERT model transformer with a sequence classification/regression head on top for General Language Understanding Evaluation (GLUE) tasks [25]. Overall, for these tasks DistilBERT has about half the total number of parameters of BERT base and retains 95% of its performances. In our experiments, we set the maximum number of training epochs to 10 for fine-tuning, and the number of global steps to evaluate and save the final models varies depending on the target category for which the models are fine-tuned.

3.2.3 Category mining

The multiclass classification approach utilises the multiple categories provided by the original datasets. On the other hand, our binary classification approaches employ a strategy to abstract higher level categories in two datasets to fine-tune independent deep learning models. Our first deep model aims to distinguish disaster-related messages from nondisaster-related messages. Then, we abstract categories for medical-related information formed by the categories 'aid-related', 'medical products', 'other aid', 'hospitals', and 'aid centres' to fine-tune our second deep binary classifier. Similarly, another higher-level category group for information related to humanitarian standards is formed by the categories 'medical help', 'water', 'food', and 'shelter' to train another deep classification model. We apply this category mining approach

to create these higher-level categories following the grouping criteria of information needs from Sphere's humanitarian standards [9]. Finally, we perform a similar strategy to fine-tune a deep binary classifier of flood events to distinguish mild flooding from moderate and severe flooding.

4 Data

In this section, we specify the differences of all the data used in our analysis. First, we describe the annotated datasets we use to learn models with our methodology for different target categories relevant to disaster and crisis management. The set of messages from annotated datasets is preprocessed at first glance using a similar procedure to that described in Sect. 3.1.

4.1 Annotated datasets

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 have been used and evaluated for purposes related to crisis and disaster response. The following items describe the datasets:

- The Social Media Disaster Tweets (SMDT) [26] dataset consists of 10,876 messages from Twitter. Each message was annotated to distinguish between disaster-related and nondisaster-related messages. There are 4673 messages out of the total, which are labeled as related to disasters, and 6203 messages annotated as not relevant to disasters. No data splits are provided for evaluation.
- The Multilingual Disaster Response Messages (MDRM) [27] 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, superstorm Sandy in the U.S.A. in 2012, and news articles spanning many years and 100 s of different disasters. The data have been encoded with 34 different categories related to disaster response and have been stripped of messages with sensitive information in their entirety. Untranslated messages and their English translations are available, which make the dataset 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 contain 20,316 messages labelled as disaster-related, in which 2,158 are floods. Pre-stablished training, validation and testing splits are provided for this dataset. We employ our category mining approach from Sect. 3.2.3 to identify 16,232 messages related to medical information, and 9010 are related to humanitarian standards' information.
- 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 and 2013 [28]–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. The data were detrended 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 dataset opens a new approach to considering flood occurrence over a long-time scale using reported information. It contains a total of 1821 records with three impact categories of mild, moderate, and severe events. From these total number of records, we found 661 records with descriptive messages—464 for mild events, 69 for moderate events, and 91 for severe events.

5 Evaluation of machine learning models

In the case of the MDRM dataset, training, validation, and testing splits are provided by the dataset [27]. We use these provided splits to train our binary classifiers with handcrafted features and deep bidirectional Transformer neural networks. All 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 multiclass classification approach, we use the entire MDRM dataset to perform the evaluation using random splits of 33% for testing and 66% for training the models. Then, we train our classifiers using fivefold cross validation and compute the average scores for each metric. This is done to make results comparable with [29], given that all the results we found for this dataset do not follow the proposed splits provided by the dataset on the multiclass classification

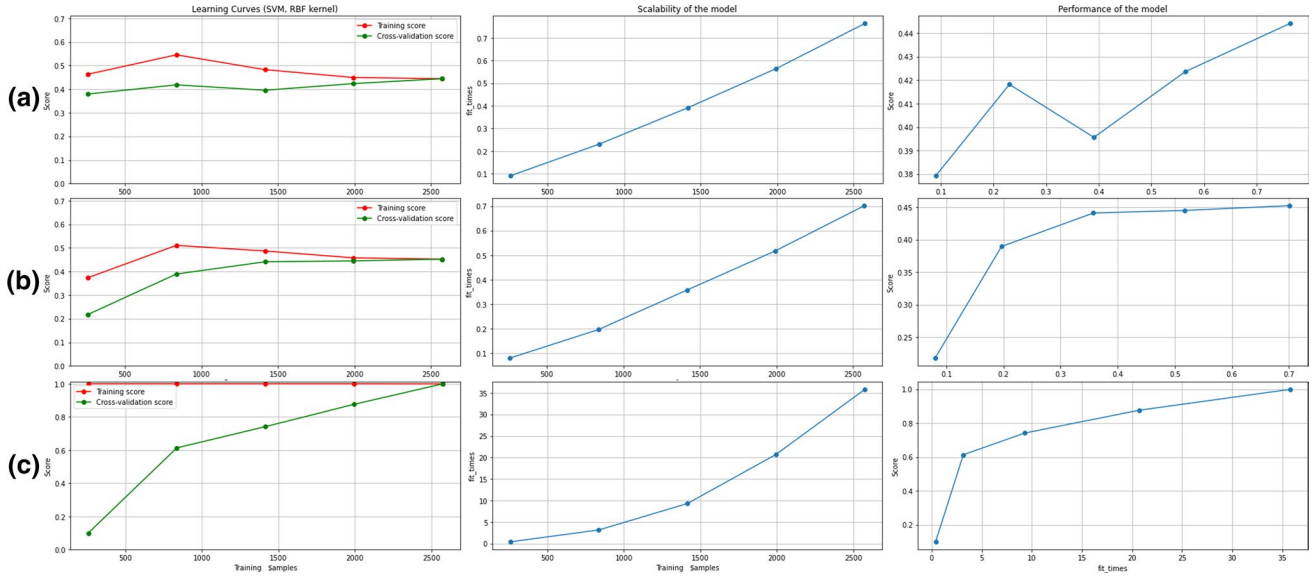

Fig. 2 Learning curve and scalability of the multiclass classification approaches (**a**) Bernoulli naïve Bayes, (**b**) multinomial naïve Bayes, and (**c**) support vector classification, using original splits of the MDRM dataset

task. In addition, 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 reliable results. We use this evaluation only for comparison purposes with existing reported results for this dataset. The performance of the SVC multiclass classification approach with this strategy reach the training process after 8000 samples, with no further improvements after that point when adding more samples.

Therefore, we perform an additional evaluation with the original form for the splits of the MDRM dataset to enhance current results and to allow future research comparisons. To the best of our knowledge, there are no previously published results for this dataset despite of the dataset being also publicly available with its original splits in competition platforms such as Kaggle.[2] Therefore, we provide for the first time a reliable machine learning evaluation for this dataset on the disaster response domain. Figure 2 plots the learning curves to show the performance and scalability of our NB-based approaches and the SVC multiclass classification 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 random splits. In this setting, we see a considerable improvement in the learning process in all our methods, especially in our SVC model. All models are trained using an Intel i7 CPU at 2.60 GHz and 16 GB RAM. The SVC approach took 120 h to train the 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 fine-tuning the binary classifiers. For the MDRM dataset, we fine-tuned each deep bidirectional Transformer neural network using an NVIDIA Tesla V100 GPU and 16 GB RAM, and the processes took between 2 and 5 days depending on the category model. The severity model for the UnifiedCEHMET dataset contains a slightly less number of samples, and its deep bidirectional Transformer neural network was fine-tuned using an NVIDIA GeForce RTX 2060Ti and 16 GB RAM, and the process took 38 h.

The results are provided in terms of precision, recall, F1-score, and binary classifier accuracy. Their values are between 0 and 1 and higher is better. Note that we do not provide results for accuracy in test data in the multiclass classification because when the class distribution is unbalanced, the accuracy metric is considered a poor choice, as it gives high scores to that predict only the most frequent class.

To calculate the above metrics we considered the method developed in [15]. 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 are treated as a collection of binary problems,

[2] https://www.kaggle.com/landlord/multilingual-disaster-response-messages.

Table 2 Comparison of multiclass classification methods using random 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**

Table 3 Comparison of multiclass classification methods using original splits

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

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 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. Microaveraging may be preferred in multilabel settings, including multiclass classification where a majority class is to be ignored.

Therefore, we compute the micro-Average Precision (mAP) from prediction scores as:

$$mAP = \sum_{n} \left(R_n - R_{n-1}\right) P_n \tag{1}$$

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.

For the binary classifiers, we use additionally the accuracy metric to compute the count of correct predictions. Its value is 1.0 when the entire set of predicted labels for a sample strictly match with the true set of labels; otherwise it is 0.0. Assuming $\hat{y}_i$ is the predicted value of the i-th sample and y_i the corresponding true value, then we calculate the accuracy of correct predictions over $n_{samples}$ as:

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

6 Models and results

We use the MDRM dataset described in Sect. 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 [29] based on Bernoulli NB classification[3] on random splits of the dataset. However, we report the results of our fivefold cross validation to provide 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 among all categories for precision, recall and F1-score.

There are multiple versions with similar codes available[4] using NB-based multiclass classification models where classification results have not been reported for this dataset. Although the MDRM dataset has been recently incorporated into the Kaggle platform[7], there are no reporting results on the original splits of the dataset. The results from our evaluation as specified in Sect. 5 are reported in Table 3, again showing 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

[3] Public repository https://github.com/ioslilyng/DisasterResponse.
[4] https://github.com/agpt8/Disaster-Response-Classification.

Table 4 Comparison of binary classification on smdt and mdrm datasets for the disaster category

Target category	Disaster				
Dataset	SMDT		MDRM		
Method/Metric	BoW [16]	TF-IDF [16]	BoW [16]	TF-IDF [16]	Distil BERT [18]
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 [16]	TF-IDF [16]	Distil BERT [18]	BoW [16]	TF-IDF [16]	Distil BERT [18]	BoW [16]	TF-IDF [16]	Distil BERT [18]
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**

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 random and original splits.

In our binary classification, Tables 4 and 5 show a clear outperformance of the DistilBERT deep learning models at identifying the four main disaster categories on both the MDRM and UnifiedCEHMET datasets. We use the resulting models fine-tuned on these datasets for our filtering approach on the remaining data collections. The handcrafted feature methods BoW and TF-IDF are taken as a baseline model to show initial performances in all three datasets. However, 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 the DistilBERT method at learning models despite it 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, this is the first time these datasets have been used to train relevant disaster models for these disaster-related categories via deep learning NLP-based methods.

7 Discussion

From the performance evaluation of our machine learning models and the results, there are several aspects to consider.

In terms of multiclass classification approaches, SVC model outperforms the NB-based models in all metrics when using random splits, with 0.83 of precision compared to 0.59 and 0.79 of precision for the Bernoulli and multinomial NB, respectively. By contrast, the multinomial NB leads the precision scores of 0.91 for precision using the original splits of the datasets. This fact points the multinomial NB as the best choice for classifying the test messages of the MDRM dataset among its 34 disaster-related categories. Nevertheless, the exponential scalability and logarithmic performance of the SVC model indicate the model will potentially outperform again both NB models with increased number of available training samples, leading to better classification messages among the 34 disaster-related categories in longer training periods. As we add more samples to the training data, the NB-based methods present a linear scalability and some unstable performance increases in the case of Bernoulli NB. By contrast, the SVC method generally increases its learning costs in terms of computing time with respect to NB.

(a) (b)

Fig. 3 **a** Multiclass classification approach to learn an SVC model to distinguish the 34 disaster categories of the MDRM dataset. The disaster-related categories are shown along with their distribution percentages of sample messages for each category; **b** Binary classification approach of deep learning models with bidirectional Transformer neural network models. We fine-tune one model for each group of main disaster categories: disaster-related, humanitarian standards, medical, and severity. The deep model of severity is fine-tuned on the Unified-CEHMET dataset, and we show the number of messages for each grouped category of severity levels

In general, we can conclude that the random splits at 66:33 ratio for training and testing provide useful indicative results in terms of model accuracy, as it was done in the previous existing works. For a full performance evaluation, however, the use of original splits provided by the dataset leads to reliable results and hence more robust models. Thus, we encourage future studies on the MDRM dataset to evaluate performance results using the original training, validation, and testing splits. This will lead to more generalization capability of the trained models with potential implications in applications that involve reliable filtering of disaster-related messages.

In terms of the binary classification, we conclude that the DistilBERT models based on deep bidirectional Transformer neural networks outperform the traditional machine learning models in all the classification tasks from our category mining. This is especially noted for the 'severity' classifier, with a difference of improvements in precision around 40%-50% less with respect to the DistilBERT model. This means that we provide with state-of-the-art results for the detection of medical-related information, humanitarian information, and severity information of messages. We provide novel models to classify categories that are found to be relevant in the context of disaster response.

As we described above in our performance evaluation, one should note the complexity and increased costs of fine-tuning the deep bidirectional Transformer neural networks to learn these types of models. Nevertheless, the resulting DistilBERT models are still lightweight and therefore permit transferability and feasible reproducibility at testing times. For this purpose, the next Sect. 8 reflects the online application of these type of methods to classify input messages.

To the best of our knowledge, the combination of data sets and methods used in this work provides the scientific community with new state-of-the-art evaluation and performance comparison to steer future studies and AI-based applications in this field of research.

8 Framework architecture and interface

In this section, we present a web application developed to illustrate the architecture of the framework,[5] the steps of the employed classification approaches, and some examples of usage with input texts. Despite the high complexity of the deep machine learning models trained as part of our research framework, we kept the interface as simple as possible to provide visual insights of our methodology and examples of classification outcomes.

Figure 3 shows the methodology described in Sect. 3.2 for the multiclass classification and binary classification, the latest including the category mining approach. This illustrates the two core parts of our disaster framework, which

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

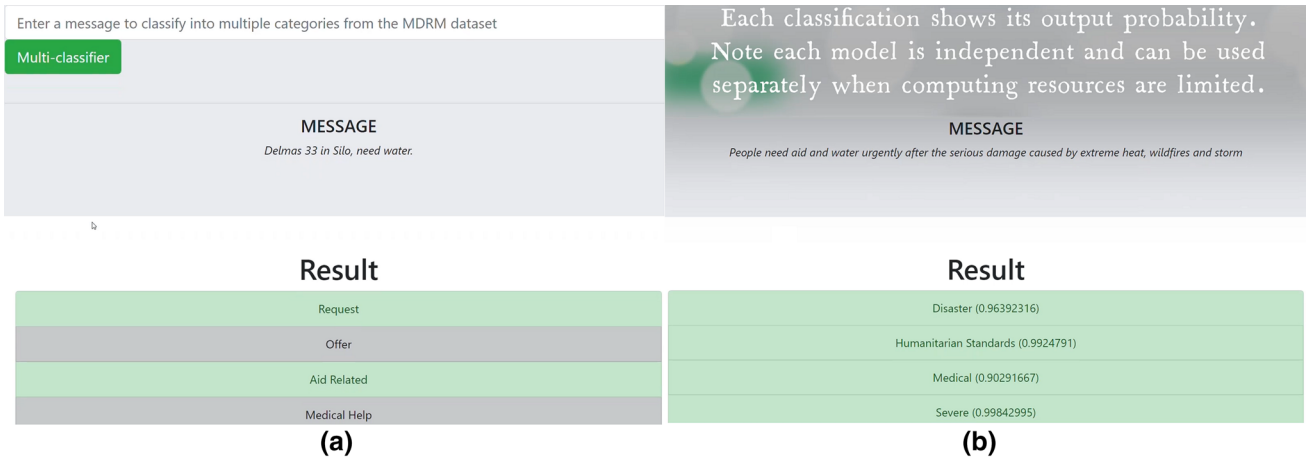

Fig. 4 Examples of (**a**) multiclass classification and (**b**) binary classification for test input messages

precedes the evaluation and results of the models in Sects. 5 and 6, respectively. On the other hand, Fig. 3 shows the list of 34 categories defined by the MDRM dataset along with the distribution of percentages for each category according to their number of sample messages. This is to provide simplified insights about the type of information learned by our classifiers related to disasters, and the uses and applications we can get out of it in the context of disaster response.

Figure 4 shows examples of the web application interface for an input target message we aim to classify into the 34 disaster-related categories and to the four main disaster-related categories, respectively. Note that the deep classification shows the confidence scores of each Transformer model.

9 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 for further evaluation of machine learning classifiers in this domain on benchmark datasets both in multiclass classification and in binary classification settings. We present a new state-of-the-art performance evaluation and results on these datasets achieved with SVC and multinomial NB approaches depending on the settings of the dataset. Furthermore, this study provides a novel approach in the 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 mining 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 the methodology and empirical application results that the score shows the potential to explain informative and influential information in a variety of large-scale data sets. This is worth noting because such information leads to higher prediction performances, thus helping to improve the accuracy of classification methods for message filtering purposes. Finally, we illustrate our web interface to provide simplified insights of this research to motivate future studies and potential applications in humanitarian response, including management of crisis or natural hazards, needs of resource allocation, or emergency assistance. For the next generations of benchmark datasets in the context of disaster response, we consider additional crowdsourcing tasks essential to gather higher amount of quality-diverse data. This will steer to boosted performances in data-hungry models such as deep neural networks, as well as increased robustness and generalization capabilities of the models based on the evaluations presented in this work; hence leading to provide better information for decision makers in the disaster domain.

Author 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.

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), and CNR (Italy). In particular, this research was funded by UKRI grant EP/V002945/1.

Data availability The datasets utilized in this study can be found in the reference works as follows: Social Media Disaster Tweets (SMDT) can be found in [26]. Multilingual Disaster Response Messages can be found in [27]. Unified records of extreme flooding events reported by the UK Centre for Ecology and Hydrology (CEH) and the UK Meteorological Office (MET) between 1884 and 2013 can be found in [28].

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.

Declarations

Competing interests The authors declare no competing interests.

References

1. Leaning J, Debarati GS. Natural disasters, armed conflict, and public health. New Engl J Med Public Health. 2013;369(19):1836–42.
2. Landwehr PM, Carley KM. "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. 2014;1:225–57.
3. Niles MT, Emery BF, Reagan AJ, Dodds PS, Danforth CM. Social media usage patterns during natural hazards. PLoS ONE. 2019;14(2):1–16.
4. CEH, UK Centre for Ecology & Hydrology. https://www.ceh.ac.uk/.
5. Metoffice, UK Meteorological Office, https://www.metoffice.gov.uk/.
6. 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.
7. 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.
8. Poblet M, García-Cuesta E, Casanovas P. 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.
9. 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.
10. Özcan S. Tweet-Preprocessor. Available: https://pypi.org/project/tweet-preprocessor/.
11. Richardson L. Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/.
12. Friedl J. Mastering regular expressions. 3rd ed., O'Reilly Medi, 2009.
13. Manning CD, Raghavan P, Schuetze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008.
14. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers. 1999.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
16. Chavda V. Tweet classification. https://github.com/pointoflight/tweet_classification.
17. Bird S, Klein E, Loper E. Natural language processing with python, O'Reilly Media Inc, 2009.
18. Sanh V, Debut L, Chaumond J, Wolf T, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, In 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS, 2019.
19. Devlin J, Chang MW, Lee K, Toutanova K. 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.
20. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Contextualized Word Representations. In Proceedings of NAACL-HLT, New Orleans, Louisiana, 2018.
21. Wang A, Tenney IF, Pruksachatkun Y, Yeres P, Phang J, Liu H, Htut PM, Yu K, Hula J, Xia P, Pappagari R, Jin S, McCoy RT, Patel R, Huang Y, Grave E, Kim N, Févry T, Chen B, Nangia N, Mohananey A, Kann K, Bordia S, Patry N, Benton D, Pavlick E, Bowman SR. Jiant 1.3: A software toolkit for research on general-purpose text understanding models. 2019.
22. Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. in NIPS Deep Learning and Representation Learning Workshop, 2015.
23. Bucila C, Caruana R, Niculescu-Mzil A. Model compression, in KDD, 2006.

24. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi C, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, Platen Pv, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S, Drame M. Lhoest Q, Rush A. Transformers: State-of-the-Art Natural Language Processing, in 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.
25. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR, GLUE: a multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the International Conference on Learning Representations (ICLR), 2019.
26. Figure Eight. Social Media Disaster Tweets. https://www.figure-eight.com/data-for-everyone/.
27. Appen. Multilingual Disaster Response Messages. https://appen.com/datasets/combined-disaster-response-data/.
28. Stevens AJ, Clarke D, Nicholls RJ. Trends in reported flooding in the UK: 1884–2013. Hydrol Sci J. 2016;61(1):50–63.
29. Ng L A Machine Learning Pipeline for Disaster Response. 2020. https://github.com/lng15/DisasterResponse.
30. Bruns A, Liang YE. Tools and methods for capturing Twitter data during natural disasters. First Monday. 2012;17(4):1–8.
31. Imran M, Ofli F, Alam F. AIDR: Artificial Intelligence for Digital Response, Qatar Computing Research Institute, 2013. http://aidr.qcri.org/.
32. Spinn3r. API Documentation. 2016. http://docs.spinn3r.com/.
33. GNIP. Grand Central Station for the Social Web, ReadWriteWeb. 2008.
34. 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.
35. N. Altman, K. M. Carley and J. Reminga, "ORA User's Guide 2020," CMU-ISR-20–110, 2020.
36. Carley KM. ORA: a toolkit for dynamic network analysis and visualization, RJ (Alhajj R., Ed., New York, NY: Encyclopedia of Social Network Analysis and Mining, Springer, 2014.
37. Ujawary-Gil A. Organizational network analysis: auditing intangible resources. 1st Edition. Routledge., 1st ed., Routledge, 2019.
38. Costa B, Boiney J. Social Radar. MITRE, McLean, Virginia, USA. McLean: The MITRE Corporation; 2012.
39. Mathieu J, Fulk M, Lorber MM, Klein G, Costa B, Schmorrow D. Social Radar Workflows, Dashboards, and Environments. Bedford: The MITRE Corporation; 2012.
40. Schmerl B, Garlan D, Dwivedi V, Bigrigg MW, Carley KM. 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.
41. Garlan D, Schmerl B, Dwivedi V, Bigrigg MW, Carley K. 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.