Single or ensemble model ? A study on social media images classification in disaster response

Hafiz Budi Firmansyah
University of Geneva
Switzerland
hafiz.firmansyah@unige.ch

Jose Luis Fernandez-Marquez
University of Geneva, Switzerland
Switzerland
joseluis.fernandez@unige.ch

Jesus Cerquides
IIIA-CSIC, Barcelona
Barcelona, Spain

Giovanna Di Marzo Serugendo
University of Geneva
Switzerland

ABSTRACT

Social media generate large amounts of almost real-time data which can turn out extremely valuable in an emergency situation, especially for providing information within the first 72 hours after a disaster event. Despite abundant state-of-the-art machine learning techniques to automatically classify social media images, the operational problem in the event of a new disaster remains unsolved. In this study, we evaluate the adaptability of a machine learning model when tested with a completely new disaster. The experimental result showed that a single model trained on the data from different disasters obtained better performance than an ensemble of models, with one model for each individual disaster.

KEYWORDS

Ensemble learning, Adaptability, Image Classification, Social Media, Disaster Response

ACM Reference Format:
Hafiz Budi Firmansyah, Jose Luis Fernandez-Marquez, Jesus Cerquides, and Giovanna Di Marzo Serugendo. 2023. Single or ensemble model ? A study on social media images classification in disaster response. In *The 10th Multidisciplinary International Social Networks Conference (MISNC 2023), September 04–06, 2023, Phuket, Thailand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3624875.3624884

MISNC 2023, September 04–06, 2023, Phuket, Thailand

ACM ISBN 979-8-4007-0817-6/23/09.
https://doi.org/10.1145/3624875.3624884

1 INTRODUCTION

1.1 Research Background

The first 72 hours after a disaster are extremely critical. During a short period, the first responders should take prompt action to save people's lives [1]. The emergency team should also conduct situational awareness of disaster scenes. Moreover, obtaining fast and accurate information is challenging due to time limits and location complexity. Meanwhile, United Nations Member States have agreed to adopt the Sendai Framework For Disaster Reduction which embraces the use of social media in disaster risk communication [14]. To be more precise, the framework stated "strengthen the utilization of media, including social media, traditional media, big data, and mobile phone networks, to support national measures for successful disaster risk communication".

Over the last decade, previous works have demonstrated that the publicly shared information on social media platforms in disaster events covers important information such as early warnings and infrastructure damage (i.e. roads, electricity, water leaks, and ruined buildings [5, 13]). While the curated dataset is provided by the research communities, analyzing the social information and maintaining its accuracy in a timely manner still leaves a challenge even using the latest technologies in machine learning. Furthermore, most of the information on social media is irrelevant or has a bare minimum informativeness in terms of data quality [8]. Consequently, it can affect the reliability and validity of data [23]. In the event of a completely new disaster, where a new kind of disaster or similar disaster in a different place happens, we have a limited time to prepare a machine learning model. Then, adaptive machine learning becomes inevitable, especially to overcome geographical complexity [27]. However, the adaptability performance of the model to analyze unseen types of crisis data coming from different training data is not yet widely evaluated [15].

This works aim at evaluating the adaptability of the machine learning model toward a new crisis. To evaluate the adaptability, we incorporated real world dataset which covers several types of disasters, including hurricanes, earthquakes, floods, and wildfires.

Our work contributes to the area of machine learning and social media analysis for disaster. The contributions of this work are summarized as follows:

[1] https://www.unocha.org/story/five-essentials-first-72-hours-disaster-response

(1) Building an understanding of the adaptability of ensemble model on various unseen disaster images (not only floods and earthquakes but also covering hurricanes and wildfires)
(2) Introducing a machine learning model training approach for handling huge amounts of social media images from a completely new disaster

1.2 Research Questions

With the complexity of understanding visual content on social media data, the emergency team needs to get a better perception of the current disaster situation to achieve situational awareness immediately. Having those challenges in our mind, the relevant research questions are particularly outlined below:

- How well does a of machine learning model perform in unseen disasters? **(RQ1)**
- Is an ensemble model or a single model a better technique to classify a social media image? **(RQ2)**

To answer **RQ1**, we incorporate various pre-trained models and use an ensemble learning approach. Specifically, we use a deep learning approach to extract features and classify the images. Popular performance metrics are used to evaluate model performance. Furthermore, we answer **RQ2** by evaluating our approach using a confusion matrix.

2 RELATED WORKS

2.1 Social media data classification for disaster management

Studies on the use of social media data classification for disaster management have attracted many researchers. From the point of view of data pre-processing, Authors in [3] applied a filter to select relevant images and reduce the number of irrelevant images from social media data. They also measured the impact of the filtering approach with precision and recall metrics. The experimental results demonstrated that filtering produced higher recall and precision values.

From the perspective of disaster response, artificial intelligence has proven useful to support emergency responders in getting situational awareness. The work presented in [16] incorporated several text vectorization techniques and various traditional machine learning algorithms to assess damage severity from an earthquake into four different levels (no damage, slight damage, moderate damage, and heavy damage) during the response phase. The authors claimed that the experiment presented the value of social media data as the source to conduct rapid damage assessment for the earthquake case.

In terms of methodology, several studies [4, 6] have demonstrated the use of machine learning techniques such as supervised, unsupervised learning, and probabilistic framework using various types of datasets. However, we found that most existing works have mostly focused on text modality. Some works mentioned the analysis of social media images. But, the majority did not focus on the adaptability quantification dimension.

2.2 Ensemble method on image classification

Recent studies on the ensemble method have demonstrated valuable results in image classification for disaster management. Authors in [9] explored an ensemble of pre-trained models to classify social media image informativeness. They addressed the approach of an ensemble of technology that outperformed a single pre-trained model. For more details, Table 1 summarizes the previous works on crisis image classification using the ensemble method and their limitations. Generally, the previous research used various sources of data sets such as websites, satellites, and social media. Additionally, the previous works attempted to improve state-of-the-art technology. Despite extensive research on the use of ensemble methods for classifying images, in this article, we focus on assessing the adaptability of an ensemble of deep-learning classification models on unseen social media image disasters.

Table 1: Overview of ensemble method research on image classification for crisis events

Paper	Discussion Topic	Limitations
[20]	Use deep stacked ensemble model which combined BCDU-Net, DeepWaterMap, and U-net to identify water surface	The experiment adopted Landsat images collected from AICrowd site
[11]	Develop an ensemble model by combining VGG16, ResNet, and bagging ensemble fusion strategy	The works used flood related dataset, not including other disaster types (earthquake, wildfires, and hurricane)
[26]	Implement of bagging approach to model the susceptibility of flood in the Teesta River basin in Bangladesh, using geographical information, climatic data, and satellite images	Using satellite images as image source

3 DATASET AND METHODS

3.1 Dataset

To conduct the experiment, we used a real-world disaster dataset named CrisisMMD [1]. It is the most widely used dataset for disaster management research, and it includes seven major disaster events in the world, namely Irma hurricane, Harvey hurricane, Maria hurricane, Iraq Iran earthquake, Mexico earthquake, California Wildfires, and Srilanka floods. The dataset consists of two different modalities: text and images. The dataset encompasses three different tasks : (1) Damage severity - Multiclass classification (2) Informativeness - Binary classification (3) Humanitarian categories - Multiclass classification. For this experiment, we considered focusing on informativeness binary classification. The informative tweets cover all the information that is useful for humanitarian aid, while not informative tweets do not possess that attribute. Table 2 shows the distribution of data for each disaster event. In this experiment, we only focused on the image modality and ignored the text modality.

Figure 1: The overview of experimental methodology. a) Single model approach b) Ensemble model approach

3.2 Methodology

This paper aims at solving the problem of model adaptability quantification. We evaluate the performance of single and ensemble models tested with a new set of disaster images. We define a new image as a completely new picture in terms of place and disaster or containing the same disaster but happening in a different place. Figure 1 illustrates our methodological approach from data preparation to model evaluation.

Fundamentally, the experiment investigated two different approaches: single model and ensemble learning. A single model is simply a model trained on a dataset, while an ensemble model is a combination of several individual models which was trained on a different category of disaster (hurricane, wildfires, earthquake, and floods) respectively. As a comparison, we trained a pre-trained model using data training from the same distribution. For example in Irma hurricane, we used the training, validation, and test data which included images of Irma hurricane. We run the methodology for both single model and ensemble learning. We used supervised deep learning algorithms to classify image informativeness with the three following steps.

- **Data preparation.** Initially, we partitioned the data into three different splits: 70 % training, 15 % validation, and 15 % test. To get an unbiased result, we excluded the test data from the distribution. As an example, in the single model, we just included Harvey hurricane, Maria hurricane, Mexico earthquake, Iran Iraq earthquake, California wildfires and Sri lanka floods in the training dataset for testing Irma hurricane. For the ensemble model case, we ignored the tested disaster, similar to the single model. Next, we conducted a data augmentation technique. The data augmentation technique enabled us to generate batches of tensor image data with real-time data augmentation. As a pre-processing function, we used torch mode that scaled image pixels within the range 0 and 1 and normalize each channel with respect to the ImageNet [7] dataset. Next, we extracted the features using the initial layer of pre-trained model.
- **Model training.** We relied on training a last layer of a pre-trained model for addressing small data training. To find the best single pre-trained model, we trained and evaluated five state-of-the-art pre-trained models. For the ensemble model, we initially trained the base model. Then, we combined a group of six base models respectively. We used a simple averaging fusion strategy to fusion the models. Simple averaging fusion incorporates two or more models. The prediction of the ensemble model calculated the average of the prediction from all models. Averaging is computed as the product of the probabilities reported by each of the base models. As an example, the probability assigned by the ensemble of A and B to class c is computed as equation 1:

$$p_{A,B}(c) = p_A(c) \cdot p_B(c) \tag{1}$$

- **Model evaluation.** The evaluation step has two axes: model performance evaluation and model adaptability evaluation. We used accuracy, recall, precision, and F-1 score as the performance metrics. We identified that there are at least two disasters having imbalanced data, namely Iraq Iran earthquake, and the Srilanka floods, we considered using an F-1 score. To evaluate the model's adaptability, we get rid of the testing data from the training data. For instance, if we test the model for Irma Hurricane, we include all the disasters except Irma Hurricane. Finally, the confusion matrix aims to show the performance of the classifier in different disaster and model settings.

Table 2: Crisis events data distribution. Informative class means that the image has meaningful information for humanitarian aid, while not informative contains banners, logos, and cartoons

Disaster	Informative	Not Informative	Total
Irma hurricane	2208	2296	4504
Harvey hurricane	2457	1977	4434
Maria hurricane	2231	2325	4556
Iraq Iran earthquake	400	197	597
Mexico earthquake	841	539	1380
California wildfires	985	604	1589
Srilanka floods	252	770	1022

3.3 Experimental Settings

3.3.1 Network architecture. The initial step of the work involved how to choose the appropriate artificial neural network architecture.

This subsection explains detailed steps of experimental settings. We trained various pre-trained models in this experiment. We rely on a transfer learning approach that enabled us to use the model features and the parameters from ImageNet [7]. In the experiment, we conducted an experiment using five different pre-trained models, namely VGG16 [24], Densenet201 [18], MobilenetV2 [22], InceptionV3 [25] and Resnet50 [12]. The goal of this work was to find the best pre-trained model which would be determined as the base classifier with the CrisisMMD dataset.

The pre-trained models use Convolutional Neural Networks (CNN) as the underlying architecture. Basically, CNN consists of three different layers: convolutional layer and ReLU, pooling layer, and fully connected layer. In the fully connected layer, we can find the softmax layer which has a dimension of 1000 because they were trained for 1000 classes images. Since we only need two classes: informative and not-informative, we changed the dimension from 1000 to 2.

3.3.2 Configuration. As a programming environment, we used Python and Keras application programming interface (API) [2]. In particular, we trained the models using Adam optimizer with sparse categorical cross entropy as the loss function and learning rate 0.00001. To create an image augmentation, we implemented ImageDataGenerator from Keras. In addition, we set the number of epochs as 50. Finally, The python code is executed in a Jupyterlab development environment on AMD EPYC 16 core using 16 GB RAM and NVidia 3090 24GB graphics card. To facilitate reproducible research, the code for this experiment is available on https://github.com/hafizbudi/ensemble_supermodel_crisis

4 EXPERIMENTAL RESULTS

The goal of the experiment is to analyze the adaptability of two different machine learning models tested with completely a new disaster. As the first step, we evaluated the best pre-trained model. Then, we conducted two different experiments separately. Those two experiments are the single model and the ensemble model. For each experiment, we run the methodology mentioned in Section 3. The performance of same-disaster, single model, and ensemble disasters was reported in the following subsection. We defined the same disaster as a model which is trained and tested within the same event and type of disaster. A single model is defined as all disasters in one dataset using one model, while an ensemble model is an ensemble of different disasters (hurricanes, earthquakes, wildfires, and floods).

4.1 Choosing the best pre-trained model

Initially, we run an experiment to evaluate the best pre-trained model among the state-of-the-art pre-trained models. The justification behind this step is that we need to select one pre-trained model among five state-of-the-art pre-trained models. The selected pre-trained model is used to implement the experiment. Table 3 describes the performance of each pre-trained model. We observed that Densenet201 produced the best accuracy. However, the only exception was for two disaster events namely Iraq Iran earthquake, and Mexico earthquake. Those disasters demonstrated slightly better performance when trained on VGG16. Based on the experiment, we considered using Densenet201 as the base model since it demonstrated the best result in five out of seven disasters.

4.2 Develop same-disaster model

After choosing the best pre-trained model, we developed a same-disaster model using Densenet201 pre-trained model. The same disaster is a model which is trained and tested with the same event and disaster. The goal of running the same-disaster was to provide an ideal case as a comparison.

4.3 Quantify single and ensemble model performance degradation

As methodology evaluation, we quantified the distance between the same disaster with single and ensemble model. In general, Figure 2 illustrates the performance of the single model and ensemble model. The horizontal axis indicates the name of the disaster. Meanwhile, the vertical axis shows the accuracy of the model. The different colors represent the different experimental approaches.

It was obvious that the same disaster shows the best performance compared to the single and ensemble models. The reason is that the same-disaster testing and training data come from the same distribution. The second best model was a single model. Finally, the ensemble model was less adaptive. More explanation about the performance is detailed below.

For a single model, we trained a Densenet201 with one event together. The overall performance results show that they were lower than same-disaster. However, we noticed that the case of the Irma and Harvey hurricane were the exception. Irma and Harvey hurricane gained slightly better accuracy because they took advantage of Maria hurricane data for instance the appearance of water and fallen tree. Maria Hurricanes provided 4556 images which could considerably also help increase the accuracy.

For the ensemble model, we joined several Densenet201 trained on various different disasters respectively. Figure 2 illustrates that the overall performance of the ensemble model was the lowest. The case of California wildfires had the worst adaptability where the difference between ensemble and same-disaster was about 2.5 %. While the smallest gaps were identified for Irma Hurricane and Maria Hurricane which had a difference of 0.3 % respectively.

Table 4 presents the confusion matrix of two different approaches evaluated using various types of disasters. We observed that a single model mostly excels in the number of true positive or informative images. While the ensemble model showed a better performance to indicate the true negative or not-informative images. Table 5 presents the experiment results of classification comparing ensemble and single model approach using accuracy, recall, precision, and F-1 score as performance measures.

From Table 5, it is obvious that about six out of seven disaster events showed single model superiority in accuracy and F-1 score in comparison with the ensemble model. The exceptional result was Mexico earthquake. For single model accuracy, the improvement for Irma hurricane was 4.4 in absolute percentages, 6.4 in absolute percentage for Harvey hurricane, 2.8 in absolute percentage for Maria hurricane, 7.4 in absolute percentage for Iraq Iran earthquake, 2.6

[2] https://keras.io/

Table 3: Base classifier results using accuracy as a measure in different disaster datasets (V for VGG16, D for Densenet201, M for MobilenetV2, I for InceptionV3, R for Resnet50)

Disaster	V	D	M	I	R
Irma Hurricane	0.722	**0.742**	0.680	0.698	0.690
Harvey Hurricane	0.764	**0.788**	0.748	0.733	0.715
Maria Hurricane	0.769	**0.782**	0.756	0.711	0.720
Iraq Iran Earthquake	**0.827**	0.777	0.765	0.753	0.777
Mexico Earthquake	**0.792**	0.774	0.774	0.686	0.728
California Wildfires	0.822	**0.839**	0.796	0.757	0.740
Srilanka Floods	0.769	**0.782**	0.756	0.711	0.720

in absolute percentage for California wildfires, 0.7 in absolute percentage for Srilanka floods. Similarly for F-1 score measure, single model outperformed ensemble model with a simple average data fusion strategy. To conclude, for all disaster events, the ensemble and same-disaster average distance produced 9.2 %. Furthermore, the distance between the single model and same-disaster model was 5.8 % on average. Hence, our single model induced reasonable results with a difference of no more than 6 % in terms of accuracy.

4.4 Model classification results

This subsection aims at showing the correctly classified and misclassified result. We only focus on showing qualitative results for the single model. Figure 3a shows an example of an image that was correctly classified as informative by single model approach. This image depicts the electricity infrastructure damage during the Irma hurricane which was classified as informative according to [1]. In Figure 3b, we present an example of an image that was correctly classified as not informative by the classifier. The image shows the candle with a sentence written above it. Our classifier has successfully recognized that the image is not useful for humanitarian action.

The image in Figure 4a was wrongly classified as informative, while the ground truth agreed that the image was not informative. This classification result could happen because the single model may detect some relevant objects for instance water and people as the features that represent an informative image.

An image in Figure 4b was classified as not informative by a single model. It should be determined as informative since there was a donation or volunteering effort depicted by some objects for example the volunteer and donation package. However, we observed that the person in the image was not completely represented which might confuse the classifier to detect the person.

5 DISCUSSION

The massive quantity of social images might decrease the classification performance of existing machine learning models. Hence, it

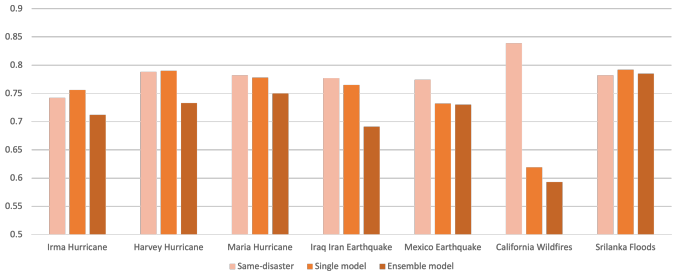
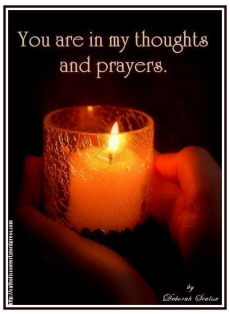


Figure 2: The comparison of model accuracy using three different approaches

(a) Classified correctly as informative

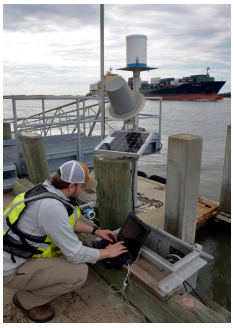


(b) Classified correctly as not informative

Figure 3: Example of images correctly classified by single model

Table 4: Confusion matrix of ensemble and single model classification performance. The highest results are indicated with boldface

Disaster	Model	TP	TN	FP	FN
Irma Hurricane	Ensemble	210	**290**	69	**133**
	Single	**258**	273	**86**	85
Harvey Hurricane	Ensemble	259	**234**	48	**131**
	Single	**311**	225	**57**	79
Maria Hurricane	Ensemble	220	**298**	55	**117**
	Single	**276**	261	**92**	61
Iraq Iran Earthquake	Ensemble	28	**28**	4	**21**
	Single	**36**	26	**6**	13
Mexico Earthquake	Ensemble	92	**71**	16	**38**
	Single	**95**	64	**23**	35
California Wildfires	Ensemble	58	**79**	11	**83**
	Single	**66**	77	**13**	75
Srilanka Floods	Ensemble	**33**	88	**26**	7
	Single	**33**	**89**	25	7



(a) Wrongly classified as informative

(b) Wrongly classified as not informative

Figure 4: Example of images wrongly classified by single model

Table 5: Performance metrics results

Disaster	Model	Accuracy	Recall	Precision	F1
Irma Hurricane	Ensemble	0.712	0.612	**0.752**	0.675
	Single	**0.756**	**0.752**	0.750	**0.751**
Harvey Hurricane	Ensemble	0.733	0.664	0.843	0.743
	Single	**0.797**	**0.797**	**0.845**	**0.820**
Maria Hurricane	Ensemble	0.750	0.652	**0.800**	0.718
	Single	**0.778**	**0.818**	0.750	**0.782**
Iraq Iran Earthquake	Ensemble	0.691	0.571	**0.875**	0.691
	Single	**0.765**	**0.734**	0.857	**0.791**
Mexico Earthquake	Ensemble	**0.751**	0.707	**0.851**	**0.773**
	Single	0.732	**0.730**	0.805	0.766
California Wildfires	Ensemble	0.593	0.411	**0.840**	0.552
	Single	**0.619**	**0.468**	0.835	**0.600**
Srilanka Floods	Ensemble	0.785	**0.825**	0.559	0.666
	Single	**0.792**	**0.825**	**0.568**	**0.673**

is critical to develop an adaptive machine-learning model. In this study, we tried to address that problem.

5.1 Methodological contribution

Single model has better adaptability compared to ensemble model with simple averaging data fusion. Previous studies [9, 19] showed that the ensemble model with simple averaging fusion tested with visual data and textual social media data extracted from the same distribution improved the accuracy performance. However, our experiment showed that ensemble model is less beneficial to improve the accuracy of unseen disasters. Generally, the single model reached an accuracy near the same-disaster model. We argue that the reason behind the result is that a single model has a lower number of parameters. In contrast, ensemble learning consisted of several base models has a larger number of parameters. Given that the deep learning model consists of millions of parameters, an increase in the number of pre-trained models will increase the model's complexity.

When tested with unseen information, a single model produces better performance in comparison with an ensemble with simple average data fusion. The literature review indicated that ensemble learning can improve model performance, which is explained in the Section 2. Despite many works praising ensemble learning since it excelled in performance improvement compared to a single deep learning model [11, 20], we found that ensemble learning did not improve adaptability performance in the case of classifying a completely new disaster both of place and type of disaster. We observed that class imbalance and noisy data are suspected as two main factors why the ensemble model demonstrated low performance in terms of adaptability [21].

5.2 Practical contributions

This work introduced several contributions for the practitioners, mainly for the first responders. From a practical perspective, some practical contributions are introduced to support helping the work of various stakeholders for instance decision makers and humanitarian organizations to get an adaptive machine learning model. The further details of the practical contribution are explained below.

5.2.1 Machine learning model adaptability assessment. We evaluated the adaptability of two machine learning approaches, namely single and ensemble models. From the experiment, a single model demonstrated better performance in terms of adaptability tested with a new disaster image. This knowledge is potentially helpful to help disaster management technology in developing an automated machine learning classifier.

5.2.2 Single model deployment could provide a reasonably good result for classifying a new disaster image. Previous studies [2, 17] demonstrated that the deep learning approach could accelerate the disaster response process by analyzing the information, for instance, social media image filtering and real-time face recognition. However, training a deep learning model from scratch will burden the resource during the disaster. The pre-trained models played a significant role to provide initial weight to the neuron in each layer in a deep learning architecture. We claimed that using transfer learning and single model creation would play an important role in helping the first responders to conduct a classification task with data collected from social media.

5.3 Future study

This study presents two limitations. Firstly, since only one ensemble learning was implemented, expected improvement could be

achieved by adopting recent results on advanced fusion mechanisms for instance Bayes optimal classifier and super learner [10]. Secondly, the size of the dataset is dramatically reduced when we only focus on a single type of disaster that could potentially affect the performance of the ensemble strategy. Future research will focus on larger datasets for example incidentsdataset [28].

6 CONCLUSION AND FUTURE WORKS

In the event of a disaster, social media offers advantageous content to enable a faster response. While automatic classification techniques have been demonstrated to achieve significant accuracy in filtering non-relevant information and classifying the severity of the damage, still in the presence of new events and new disasters the models dramatically reduce their performance.

This study proposes a methodology to assess the issue of adaptability using a well-established social media dataset for disaster response, called CrisisMMD. The research demonstrated that the pre-trained model reduces their performance when used in a disaster that was not part of the learning process. Even though the problem of adapting the pre-trained model to the new event is mentioned in the literature, this research quantified the performance degradation and evaluated two different strategies to mitigate it.

Experimental results show that creating a model including images from different disasters (e.g. floods, earthquakes, hurricanes, and wildfires) performed better than individual models which are trained for each type of disaster, and later combined to classify unseen images.

As a future work, we see two main directions. First, the implementation and evaluation of other ensemble learning approaches, for instance, bagging, boosting, and stacking might be considered for future research. Second, we would suggest measuring the impact of the proposed methodology in the real-world case scenario.

Acknowledgement

This work was funded by the Indonesia Endowment Fund For Education (LPDP), the H2020 EU Project Crowd4SDG under the grant agreement No 872944, the H2020 EU Project Humane-AI-net under the grant agreement No 952026, and the project CI-SUSTAIN funded by the Spanish Ministry of Science and Innovation (PID2019-104156GB-I00).

REFERENCES

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)* (2018).

[2] Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. 2020. Deep learning benchmarks and datasets for social media image classification for disaster response. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (2020). https://doi.org/10.1109/asonam49781.2020.9381294

[3] Sara Barozzi, Amudha Ravi Shankar, Jose Luis Fernandez-Marquez, and Barbara Pernici. 2019. Filtering Images Extracted from Social Media in the Response Phase of Emergency Events. *ISCRAM 2019 Conference Proceedings – 16th International Conference on Information Systems for Crisis Response and Management* (2019).

[4] Moumita Basu, Anurag Shandilya, Prannay Khosla, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations. *IEEE Transactions on Computational Social Systems* 6, 3 (2019), 604–618. https://doi.org/10.1109/tcss.2019.2914179

[5] Carlos Castillo. 2019. *Big Crisis Data: Social Media in disasters and time-critical situations.* Cambridge University Press.

[6] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet. *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10* (2010). https://doi.org/10.1145/1871437.1871535

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Xuefan Dong and Ying Lian. 2021. A review of social media-based public opinion analyses: Challenges and recommendations. *Technology in Society* 67 (2021), 101724. https://doi.org/10.1016/j.techsoc.2021.101724

[9] Hafiz Budi Firmansyah, Jesus Cerquides, and Jose Luis Fernandez-Marquez. 2022. Ensemble learning for the classification of social media data in disaster response. *ISCRAM 2022 Conference Proceedings – 19th International Conference on Information Systems for Crisis Response and Management* (2022).

[10] M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. 2022. Ensemble deep learning: A Review. *Engineering Applications of Artificial Intelligence* 115 (2022), 105151. https://doi.org/10.1016/j.engappai.2022.105151

[11] Muhammad Hanif, Muhammad Atif Tahir, and Muhammad Rafi. 2021. Vrbaggednet: Ensemble based Deep Learning Model for Disaster Event Classification. *Electronics* 10, 12 (2021), 1411. https://doi.org/10.3390/electronics10121411

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017). https://doi.org/10.1109/ICCV.2017.322

[13] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency. *Comput. Surveys* 47, 4 (2015), 1–38. https://doi.org/10.1145/2771588

[14] Ilan Kelman. 2015. Climate change and the Sendai Framework for Disaster Risk Reduction. *International Journal of Disaster Risk Science* 6, 2 (2015), 117–127. https://doi.org/10.1007/s13753-015-0046-5

[15] Prashant Khare, Grégoire Burel, and Harith Alani. 2018. Classifying crises-information relevancy with semantics. *The Semantic Web* (2018), 367–383. https://doi.org/10.1007/978-3-319-93417-4_24

[16] Lingyao Li, Michelle Bensi, Qingbin Cui, Gregory B. Baecher, and You Huang. 2021. Social media crowdsourcing for rapid damage assessment following a sudden-onset natural hazard event. *International Journal of Information Management* 60 (2021), 102378. https://doi.org/10.1016/j.ijinfomgt.2021.102378

[17] Fang Liu, Yeting Guo, Zhiping Cai, Nong Xiao, and Ziming Zhao. 2019. Edge-Enabled Disaster Rescue: A Case Study of Searching for Missing People. *ACM Trans. Intell. Syst. Technol.* 10, 6, Article 63 (dec 2019), 21 pages. https://doi.org/10.1145/3331146

[18] Sreenivasulu Madichetty and M. Sridevi. 2021. A novel method for identifying the damage assessment tweets during disaster. *Future Generation Computer Systems* 116 (2021), 440–454. Publisher: Elsevier.

[19] SreeJagadeesh Malla and Alphonse P.J.A. 2021. Covid-19 outbreak: An Ensemble pre-trained deep learning model for detecting informative tweets. *Applied Soft Computing* 107 (2021), 107495. https://doi.org/10.1016/j.asoc.2021.107495

[20] Kaveh Moradkhani and Abdolhossein Fathi. 2022. Segmentation of waterbodies in remote sensing images using deep stacked ensemble model. *Applied Soft Computing* 124 (2022), 109038. https://doi.org/10.1016/j.asoc.2022.109038

[21] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery* 8, 4 (2018). https://doi.org/10.1002/widm.1249

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4510–4520.

[23] Tomer Simon, Avishay Goldberg, and Bruria Adini. 2015. Socializing in emergencies—A review of the use of social media in emergency situations. *International Journal of Information Management* 35, 5 (2015), 609–619. https://doi.org/10.1016/j.ijinfomgt.2015.07.001

[24] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (April 2015). http://arxiv.org/abs/1409.1556 arXiv: 1409.1556.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]* (Dec. 2015). arXiv: 1512.00567 version: 3.

[26] Swapan Talukdar, Bonosri Ghose, Shahfahad, Roquia Salam, Susanta Mahato, Quoc Bao Pham, Nguyen Thi Thuy Linh, Romulus Costache, and Mohammadtaghi Avand. 2020. Flood susceptibility modeling in Teesta River basin, Bangladesh using novel ensembles of bagging algorithms. *Stochastic Environmental Research and Risk Assessment* 34, 12 (Dec. 2020), 2277–2300. https://doi.org/10.1007/s00477-020-01862-5

[27] Jiting Tang, Saini Yang, and Weiping Wang. 2021. Social Media-based disaster research: Development, trends, and obstacles. *International Journal of Disaster Risk Reduction* 55 (2021), 102095. https://doi.org/10.1016/j.ijdrr.2021.102095

[28] Ethan Weber, Dim P. Papadopoulos, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. 2022. INCIDENTS1M: A large-scale dataset of images with natural disasters, damage, and incidents. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), 1–14. https://doi.org/10.1109/tpami.2022.3191996