Ensemble learning for the classification of social media data in disaster response

Hafiz Budi Firmansyah*
Citizen Cyberlab, CUI, University of Geneva, Switzerland
hafiz.firmansyah@unige.ch

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

Jesus Cerquides
IIIA-CSIC, Barcelona, Spain
cerquide@iiia.csic.es

ABSTRACT

Social media generates large amounts of almost real-time data which has proven valuable in disaster response. Specially for providing information within the first 48 hours after a disaster occurs. However, this potential is poorly exploited in operational environments due to the challenges of curating social media data. This work builds on top of the latest research on automatic classification of social media content, proposing the use of ensemble learning to help in the classification of social media images for disaster response.

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Experimental results show that ensemble learning is a valuable technology for the analysis of social media images for disaster response,and could potentially ease the integration of social media data within an operational environment.

Keywords

Ensemble learning, image classification, social media, disaster response

INTRODUCTION

As mentioned by the United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA)[1], the first 72 hours after a disaster are crucial; response must begin during that time to save lives. Gathering accurate and timely information about the damaged areas just after a disaster is essential for emergency response teams.

Over the last decade, social media has demonstrated as an effective and efficient channel for getting situational awareness in the event of a disaster. Social media offers some main benefits such as providing information promptly generated by a crowd in a nearly real-time manner (Havas et al. 2017). However, the curation of the information ensuring its accuracy within time constraints is still a challenge even using the latest technologies in data science.

Recent research in the field of machine learning is continuously improving the accuracy of the classifications of social media data for disaster response. Namely, computer vision for analysing photos, and Natural Language processing for analysing text have been considered major technologies for processing social media information. In recent years, neural network architectures such as ResNet, VGG16, and DenseNet have proven to be able to filter relevant data and improve classification accuracy.

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

Both images and textual information from social media have demonstrated to be valuable in disaster response. Additionally, multi-modal learning combining both sources of information, textual and visual, has demonstrated to outperform those methods using just images or texts.

By leveraging different deep learning strategies such as transfer learning, researchers have produced significant results on Social media data classification in disaster response.

Ensemble learning (Rokach 2010) uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Several works in the literature have applied ensemble learning to the text part of tweet messages for disaster response.

On the other hand, the publication of datasets containing social media data related to disasters response has allowed researchers to compare different approaches and quickly move forward the state of the art. This paper proposes the use of ensemble learning for the classification of social media images in disaster response. The performance of this strategy is evaluated using the well-known CrisisMMD dataset (Alam, Ofli, et al. 2018).

The contribution of ensemble learning is compared with five existing approaches from the literature which have demonstrated to outperform using the CrisisMMD dataset.The experimental results demonstrate that ensemble learning is a valuable technology for the analysis of social media images in emergencies. In addition, different directions to improve the result of using ensemble learning are provided in this paper, which position ensemble learning as a key machine learning approach for the use of social media data in operational environments.

The main contributions of the paper are:

- This is the first time in which ensemble classifiers are proposed for the analysis of social media images for disaster response.
- Our experimental results show that simple ensembles provide better quality predictions than current state-of-the-art deep learning models, even if they incorporate the text of the tweet.

RESEARCH BACKGROUND

One of the major contribution in the field of automatic classification of social media content, namely textual and images content has been the creation of publicly available datasets. Repositories such as crisisNLP [2] and CrisisLex [3], and datasets such CrisisMMD (Alam, Ofli, et al. 2018) have enabled the comparison of many different methods and algorithms for automatic classification of social media content in the field of disaster management.

Previous studies have demonstrated the potential of both images (Nguyen et al. 2017; Alam, Imran, et al. 2017; Daly and Thom 2016; Barozzi et al. 2019) and textual information (Stefan et al. 2019; Imran et al. 2015; Lorini et al. 2019) from social media in humanitarian aids.

The combination of both textual and visual information using multi-modal deep learning have demonstrated to provide useful information and to outperform the accuracy of the classification (Madichetty et al. 2021; Agarwal et al. 2020; Gautam et al. 2019).

A major groundbreaking technology for the automatic classification of social media information has been Deep Learning (LeCun et al. 2015). Deep Learning architectures have evolved extremely fast over the last seven years. Recent deep learning architecture such as VGG-16 (Simonyan and Zisserman 2015), ResNet50 (He et al. 2017), InceptionV3 (Szegedy et al. 2015), or DenseNet (Madichetty et al. 2021) have taken advantage of features from ImageNet dataset creating pre-trained architectures which are later on adapted for the field of Disaster Response using transfer learning. Transfer learning and new deep learning architectures have demonstrated to outperform the state of the art using CrisisMMD dataset (Ofli et al. 2020; Yang et al. 2021; Gautam et al. 2019; Madichetty et al. 2021). As far as we know the best performance achieved so far has been reported by the DenseNet architecture (Madichetty et al. 2021).

[2] https://crisisnlp.qcri.org/
[3] https://crisislex.org/

Ensemble Learning

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone (Rokach 2010; L. I. (I. Kuncheva 2014). An ensemble classifier is built by training a set of base classifiers and later on combining their predictions.

Existing ensemble techniques provided an increased robustness and accuracy by learning base classifiers in different subsamples of the data. For instance, bagging (Breiman 1996) learns classifiers on different samples with replacement from our training data. Alternatively, boosting (Schapire 1999) incrementally builds new weighted samples of the training data assigning larger weight to those instances with larger prediction error.

Alternatively, ensembles of heterogeneous classifiers, with each classifier using a different machine learning algorithm, can also be built. The combination of their predictions can be done through different techniques (L. I. Kuncheva 2002): hard-voting or majority voting assigns each classifier a single indivisible vote and selects the class that has obtained the larger number of votes; soft-voting or weighted voting allows each classifier to split its single vote fractionally into the different classes and provides as output the average of the votes received. More complex alternatives such as stacking or Bayesian model averaging have been proposed.

Ensemble learning for disaster response

Several works in the literature have applied ensemble learning to the text part of tweet messages. Kejriwal and Zhou (Kejriwal and Zhou 2019; Kejriwal and Zhou 2020) use a weighted ensemble of three base classifiers, built respectively from local embeddings, manual features and Wiki embeddings to create an early detection system for disasters. Nalluru et al. (Nalluru et al. 2019) apply feature ensembling to join domain-agnostic and domain-specific word embeddings to classify tweet relevance in a disaster scenario. Priya et al. (Priya et al. 2020) tackle the distribution drift in the source and target data along with the class imbalance in the target data by building an ensemble of three different models. Two models separately learn the event invariant and specific features of a target data from a set of source and target data. The third model is an adversarial model with the objective of improving the prediction accuracy of the other two models. Pekar et al. (Pekar et al. 2016) create an ensemble which combines multiple classifiers specific to each emergency type to classify previously unseen texts.

Talukdar et al. (Talukdar et al. 2020) use bagging to model the susceptibility of flood in the Teesta River basin in Bangladesh, using geographical information, climatic data and satellite images. However, they do not use photographs or images obtained from social media.

Thus, as far as we can tell, this is the first work in which an ensemble of classifiers on images obtained from social media is proposed to help in disaster response.This paper proposes the use of ensemble learning which build on top of the line of works of Alam et. al.(Alam, Ofli, et al. 2018), Ofli et al.(Ofli et al. 2020), and Madichetty et al. (Madichetty et al. 2021) which use the CrisisMMD dataset to evaluate the use of social media for disaster response.

SCIENTIFIC QUESTIONS AND EXPERIMENTAL DESIGN

Social media data has been proven to be potentially valuable in disaster response, especially for assessing damage just after a disaster occurs. However, this potential is poorly exploited by practitioners in operational environments. A major reason is a limitation of providing an accurate dataset within time constraints.

Our most relevant scientific question can be stated as follows:

Can ensemble learning be a valuable technology for improving the accuracy of the classification of social media images for disaster response?

To provide an answer to this question, this section presents the experimental design implemented in this work.

Experimental design

The experimental design is composed of three major phases: (1) the preparation of the datasets used for training and validating the approaches which are included in the experiment. This phase aims at enabling the comparison with previous work and easing the reproducibility of the results. (2) Selection and training of different deep learning algorithms. This phase selects the promising deep learning algorithm which has proven good results in the literature and training them using datasets from the previous phase. (3) Implementation and evaluation of the ensemble learning proposed in this work.

Step 1. Preparation of datasets

As described in the related work, the CrisisMMD dataset (Alam, Ofli, et al. 2018) is becoming a *de facto* standard for the evaluation of new computational methods using social media for disaster response. Consequently, our experimental design relies on CrisisMMD. The dataset contains several thousands of tweets including texts and images collected during seven major disasters including earthquakes, hurricanes, wildfires, and floods.

The dataset defines three different tweet classification tasks:

1. Task 1: Informative vs Not informative. Tweets should be classified as either `Informative` or `Not Informative` for disaster response.
2. Task 2: Humanitarian categories. Tweets should be classified as in one out of a set of seven categories depending on the information that the tweet is providing.
3. Task 3: Damage severity assessment. Tweets should be classified in one out of four severity categories: `Don't know or can't judge`, `Little or no damage`, `Mild damage`, and `Severe damage`.

As explained in the related work, ensembles have been tested (and proven valuable) on the text component of the tweets. Thus, here we are specifically interested in its potential contribution to the image component.

A common strategy for the evaluation of machine learning algorithms is to split the available data into three different disjoint parts, namely training, validation and testing. The model parameters are initially fit on the training part. The fitted model is used to predict the responses for the observations in the validation part. The testing part is used to provide an unbiased evaluation of a final model fit on the training data set.

In order to ease reproducibility and comparability with previous results, CrisisMMD provides the two different scenarios for which a training, validation and testing splits are provided. The first of the two scenarios (namely `all`) incorporates all of the labeled images. Training, validation, and testing are provided for each of the three tasks, resulting in 3 different datasets, namely `informative_all`,`humanitarian_all`, and `damage_all`. The second scenario was introduced in (Ofli et al. 2020) to deal with the multimedia classification of tweets and contains only those images in which the image classification agrees with the text classifications (namely `agreed`). Training, validation, and testing are provided for tasks 1 and 2, resulting in two additional datasets, namely `informative_agreed`, and `humanitarian_agreed`. Datasets are summarized in Table 1.

Table 1. Datasets

Scenario	Task	Dataset
`all`	Informative vs Not informative	`informative_all`
	Humanitarian categories	`humanitarian_all`
	Damage severity assessment	`severity_all`
`agreed`	Informative vs Not informative	`informative_agreed`
	Humanitarian categories	`humanitarian_agreed`

Step 2. Selection and training of different deep learning architectures

For each of the five datasets described in the previous paragraph and summarized in Table 1, we investigate whether the application of ensemble learning provides any benefits. To that purpose, we have selected a set of base classifiers each of them with a neural network architecture which has proven itself valuable for the classification of images in the ImageNet (Deng et al. 2009) database. The five base classifiers are VGG-16 (Simonyan and Zisserman 2015), InceptionV3 (Szegedy et al. 2015), ResNet50 (He et al. 2017), MobileNetV2 (Sandler et al. 2018), and DenseNet201 (Huang et al. 2017). To train each of these classifiers we rely on transfer learning. That is, for each of the models, we take the parameters that provided a good quality classification of images on ImageNet and keep them frozen. The rationale for this is that these parameters are creating a rich set of features on which many image classification problems can be solved. Training is only performed at the so-called *top* of the network, which connects the features to the output of the model, being that output the specific categories in which images are classified in each of the tasks.

We train each of the base classifiers with the objective to minimize the cross entropy loss. Cross entropy is a measure of how well a classifier approximates the probabilities of its predictions. The lower the cross entropy, the

Table 2. Crossentropy of the base classifiers

Dataset	VGG16	DenseNet201	MobileNetv2	InceptionV3	ResNet50
informative_agreed	0.382	**0.375**	0.412	0.450	0.488
humanitarian_agreed	0.608	**0.535**	0.658	0.730	0.805
damage_all	0.792	**0.781**	0.826	0.878	0.869
informative_all	0.442	**0.425**	0.465	0.526	0.538
humanitarian_all	0.805	**0.743**	0.859	0.959	1.031

Table 3. Accuracy of the base classifiers

Dataset	VGG16	DenseNet201	MobileNetv2	InceptionV3	ResNet50
informative_agreed	83.246	**83.507**	81.943	79.791	78.292
humanitarian_agreed	76.126	**80.524**	76.754	72.356	69.005
damage_all	65.028	**66.730**	63.516	63.138	61.626
informative_all	78.721	**80.823**	77.738	74.832	72.150
humanitarian_all	70.228	**73.402**	70.049	66.026	63.523

better the classifier is estimating the probabilities. We train along 50 epochs on the training set, automatically reducing the learning rate to obtain a better fit. Out of the 50 classifiers obtained (one per epoch), we select the one with a smaller cross entropy on the validation set.

Step 3. Evaluation of the ensemble learning approach

To evaluate whether ensemble learning provides any benefits, we compare the results of these five base classifiers with the results of all the possible ensembles of two classifiers. Since the current objective is to validate the proof of concept, the ensemble technique used is simply averaging the predictions of the two classifiers. The technique only computes the average of previously trained models, without performing additional training or adjustments. Thus, the prediction of an ensemble of two models (say A and B) is obtained by averaging the prediction of A (p_A), and the prediction of B (p_B). Averaging is simply computed as the product, so the probability assigned by the ensemble A,B to class c is computed as $p_{A,B}(c) = p_A(c) \cdot p_B(c)$.

For each of the classifiers (both base and ensemble) we evaluate its performance using two measures: its cross entropy in the testing set, which provides an idea of how well the method is estimating its probability of success, that is, how aware the classifier is of what it knows and what it does not know, and its accuracy on the testing set, which provides an estimate of the average number of images that the method will correctly classify.

EXPERIMENTAL RESULTS

The evaluation of the ensemble learning approach is implemented in two parts. Firstly, the performance of the five base classifiers introduced in the experimental design section is compared. Secondly, we analyse the contribution of the ensemble learning approach by comparing it with the individual performance of previous classifiers.

Base classifier comparison

Table 2 shows the cross entropy for the five base classifiers with the best classifier marked in boldface. We can see that DenseNet201 provides the best results, i.e. the lowest cross entropy in each of the datasets.

Table 3 shows that DenseNet201 is also the classifier with the highest accuracy on all five datasets. The fact that DenseNet201 is the best on both measures and on every dataset tested could reduce the potential value of ensembling since ensembling is usually most valuable when there is no such a clear winner.

Contribution of ensemble learning

Despite the fact that DenseNet201 is consistently the winning base model, we can still improve significantly by using an ensemble. Out of the different ensembles tested, two of them appear as especially promising. Both of them incorporate DenseNet201 as one of the components of the ensemble, being the other one VGG16 in one case (which

Table 4. Crossentropy of the base and ensemble classifiers

Dataset	VGG16	DenseNet201	MobileNetv2	D+V	D+M
`informative_agreed`	0.382	0.375	0.412	**0.349**	**0.361**
`humanitarian_agreed`	0.608	0.535	0.658	**0.530**	0.541
`damage_all`	0.792	0.781	0.826	**0.7511**	**0.753**
`informative_all`	0.442	0.425	0.465	**0.406**	**0.415**
`humanitarian_all`	0.805	0.743	0.859	**0.708**	**0.730**

Table 5. Accuracy of the base and ensemble classifiers

Dataset	VGG16	DenseNet201	MobileNetv2	D+V	D+M
`informative_agreed`	83.246	83.507	81.943	**84.615**	83.441
`humanitarian_agreed`	76.126	80.524	76.754	80.314	**80.628**
`damage_all`	65.028	66.730	63.516	**66.918**	66.162
`informative_all`	78.721	80.823	77.738	**81.090**	**81.448**
`humanitarian_all`	70.228	73.402	70.049	**74.117**	**73.581**

we name D+V) and MobileNetV2 in the other (which we name D+M). Table 4 shows how these two ensembles compare with the base classifiers in terms of cross entropy. The ensembles are boldfaced when they improved over the best base classifier, which is always DenseNet201. D+V provides better cross entropy than the state-of-the-art method in all the datasets and D+M in four out of five of them.

Table 5 shows how the two best ensembles compare with the base classifiers in terms of accuracy. The ensembles are boldfaced when they improved over the best base classifier, which is always DenseNet201. In every dataset, it is an ensemble method the one that reaches maximum accuracy. D+V provides better accuracy than the state-of-the-art method in four out of five datasets, whilst D+M does it in three out of five datasets. Furthermore, D+V turns out to be better than DenseNet201 in the informative agreed scenario, but not on the humanitarian agreed scenario. We argue that this could be explained by the poor performance of VGG16 on humanitarian agreed (with a 4% decrease in accuracy with respect to dense net. It is usually the case that an ensemble of two classifiers is better than both of them. However, when one of the classifier is considerably worse than the other, the so-called ensemble effect will not be enough to compensate the loss of prediction quality. In colloquial terms, we could say that asking for a second opinion is good as long as the individual you are asking it to is at about your level of knowledge for the topic. If the person you are asking to is clearly less informed than you, you probably are better off not considering his opinion.

The corresponding entries in Table 5 show an accuracy of 83.507 for the best base classifier and accuracy of 84.615 for the D+V ensemble, i.e. a 1.108 improvement. To get a better understanding of the improvement provided by ensembles, we highlight here the recent research by Ofli et al (Ofli et al. 2020) where they report an accuracy of 83.3 for their image classifier (VGG16) and an accuracy of 84.4 for a multi-modal classifier which takes into account not only image but also the text of the tweet, that is a 1.1 improvement is obtained by additionally considering the text. This means that, at least for that dataset, the increase in accuracy obtained by using ensembles could be similar to the one reported by including the tweet text into the classifier, i.e. combining textual and visual information.

The experiment clearly shows that the ensembles shows a better performance on binary classification tasks (i.e informative agreed and informative all) than on multiclass classification tasks (i.e humanitarian agreed, damage all, humanitarian all). This could happen since binary tasks are usually less complex than multiclass tasks provided we are given the same amount of data for learning.

DISCUSSION

The use of computational methods for automatic classification of social media information for disaster response in an operational environment is still a challenge. The results presented in this paper motivated the use of ensemble learning, instead of the use of a single pre-trained model in an operational environment. There are two major motivation factors: On one hand, the proposed ensemble learning approach produces better quality predictions in both binary and multi-class tasks in comparison with a single pre-trained model. On the other hand, the ensemble learning approach could potentially mitigate deficiencies on new models reducing the effort of migrating to new technologies, thus, ensembles could provide a more robust approach for the incorporation of social media data in disaster response.

CONCLUSION AND FUTURE WORK

Analyzing images from social media is not a trivial task for the first responders of disaster. Despite research had been conducted to automatize that process, the intended result had not been achieved with the desired accuracy.

This study proposed the use of ensemble learning for classifying disasters images from social media. Our main contributions on this research are i) the ensemble model produces better model accuracy both of binary and multi-class task in comparison with single pre-trained model ii) illustrating how different disaster dataset to perform on various pre-trained model and ensemble model respectively.

This work provides the ground to improve the performance of automatic classification of social media content for disaster response with two different directions: (1) Adopting existing methods for aggregating the results of the different deep learning models participating in the ensemble learning process, and (2) Adopting new deep Learning architectures such as EfficientNetV2 (Tan and Le 2021) which recently presented outstanding results using ImageNet dataset but it has never been applied in the field of disaster response. In addition, it would be relevant to evaluate the adaptability of existing approaches when they have to classify social media content from a disaster they have never seemed before, and analyze whether ensemble learning presents better adaptation to completely new disasters than using any other deep learning algorithm alone.

ACKNOWLEDGEMENTS

This work was partially 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

Agarwal, M., Leekha, M., Sawhney, R., and Shah, R. R. (Apr. 2020). "Crisis-DIAS: Towards Multimodal Damage Analysis - Deployment, Challenges and Assessment". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.01, pp. 346–353.

Alam, F., Imran, M., and Ofli, F. (July 2017). "Image4Act: Online Social Media Image Processing for Disaster Response". In: *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 601–604.

Alam, F., Ofli, F., and Imran, M. (2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*. USA.

Barozzi, S., Fernandez-Marquez, J. L., Shankar, A. R., and Pernici, B. (2019). "Filtering images extracted from social media in the response phase of emergency events". In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*. Ed. by Z. Franco, J. J. González, and J. H. Canós. ISCRAM Association.

Breiman, L. (1996). "Bagging predictors". In: *Machine learning* 24.2, pp. 123–140.

Daly, S. and Thom, J. A. (2016). "Mining and Classifying Image Posts on Social Media to Analyse Fires". In: *Proceedings of the 13th International Conference on Information Systems for Crisis Response and Management*. ISCRAM Association, p. 14.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (June 2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Gautam, A. K., Misra, L., Kumar, A., Misra, K., Aggarwal, S., and Shah, R. R. (Sept. 2019). "Multimodal Analysis of Disaster Tweets". In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 94–103.

Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J., Van Achte, T., Zeug, G., Mondardini, M., Grandoni, D., et al. (Nov. 29, 2017). "E2mC: Improving Emergency Management Service Practice through Social Media and Crowdsourcing Analysis in Near Real Time". In: *Sensors* 17.12, p. 2766.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (June 2015). "Processing Social Media Messages in Mass Emergency: A Survey". In: *ACM Computing Surveys* 47.4, 67:1–67:38.

Kejriwal, M. and Zhou, P. (2019). "Low-supervision urgency detection and transfer in short crisis messages". In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 353–356.

Kejriwal, M. and Zhou, P. (2020). "On detecting urgency in short crisis messages using minimal supervision and transfer learning". In: *Social Network Analysis and Mining* 10.1, pp. 1–12.

Kuncheva, L. I. (2002). "A theoretical study on six classifier fusion strategies". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.2, pp. 281–286.

Kuncheva, L. I. (I. (2014). *Combining pattern classifiers : methods and algorithms / Ludmila I. Kuncheva.* eng. Second edition. Hoboken, New Jersey: Wiley.

LeCun, Y., Bengio, Y., and Hinton, G. (May 2015). "Deep learning". en. In: *Nature* 521.7553, pp. 436–444.

Lorini, V., Castillo, C., Dottori, F., Kalas, M., Nappo, D., and Salamon, P. (2019). "Integrating Social Media into a Pan-European Flood Awareness System: A Multilingual Approach". In: *CoRR* abs/1904.10876. arXiv: 1904.10876.

Madichetty, S., Muthukumarasamy, S., and Jayadev, P. (2021). "Multi-modal classification of Twitter data during disasters for humanitarian response". In: *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15.

Nalluru, G., Pandey, R., and Purohit, H. (2019). "Classifying Relevant Social Media Posts During Disasters Using Ensemble of Domain-agnostic and Domain-specific Word Embeddings". In: *arXiv preprint arXiv:1911.05165*.

Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (2017). "Robust classification of crisis-related data on social networks using convolutional neural networks". In: *Eleventh international AAAI conference on web and social media*.

Ofli, F., Alam, F., and Imran, M. (2020). "Analysis of social media data using multimodal deep learning for disaster response". In: *arXiv preprint arXiv:2004.11838*.

Pekar, V., Binner, J., and Najafi, H. (2016). "Detecting Mass Emergency Events on Social Media: One Classification Problem or Many?" In: *Proceedings of the International Conference on Data Science (ICDATA)*. The Steering Committee of The World Congress in Computer Science, Computer . . ., p. 31.

Priya, S., Upadhyaya, A., Bhanu, M., Kumar Dandapat, S., and Chandra, J. (2020). "EnDeA: Ensemble based Decoupled Adversarial Learning for Identifying Infrastructure Damage during Disasters". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1245–1254.

Rokach, L. (2010). "Ensemble-based classifiers". In: *Artificial intelligence review* 33.1, pp. 1–39.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520.

Schapire, R. E. (1999). "A brief introduction to boosting". In: *Ijcai*. Vol. 99, pp. 1401–1406.

Simonyan, K. and Zisserman, A. (Apr. 2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv:1409.1556 [cs]*.

Stefan, I., Rebedea, T., and Caragea, D. (2019). "Classification of eyewitness tweets in emergency situations". In: *RoCHI*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (Dec. 2015). "Rethinking the Inception Architecture for Computer Vision". In: *arXiv:1512.00567 [cs]*.

Talukdar, S., Ghose, B., Shahfahad, Salam, R., Mahato, S., Pham, Q. B., Linh, N. T. T., Costache, R., and Avand, M. (Dec. 2020). "Flood susceptibility modeling in Teesta River basin, Bangladesh using novel ensembles of bagging algorithms". en. In: *Stochastic Environmental Research and Risk Assessment* 34.12, pp. 2277–2300.

Tan, M. and Le, Q. V. (June 2021). "EfficientNetV2: Smaller Models and Faster Training". In: *arXiv:2104.00298 [cs]*.

Yang, X., Jiu, M., and Sun, Y. (2021). "Multimodal Deep Kernel Learning: Application to Disaster Response in Social Media Network". In: *2021 4th International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 560–564.