# Multimodal Semi-supervised Learning for Disaster Tweet Classification

Anonymous ACL submission

## Abstract

During natural disasters, people often use social media platforms, such as Twitter, to post information about casualties and damage produced by disasters. This information can help relief authorities gain situational awareness in nearly real time, and enable them to quickly distribute resources where most needed. However, annotating data for this purpose can be burdensome, subjective and expensive. In this paper, we investigate how to leverage the copious amounts of unlabeled data generated by disaster eyewitnesses and affected individuals during disaster events. To this end, we propose a semi-supervised learning approach to improve the performance of neural models on several multimodal disaster tweet classification tasks. Our approach shows significant improvements, obtaining up to 3.5% F1 performance gain at no additional annotation cost.

## 1 Introduction

The upswing of text and image sharing on social media platforms, such as Twitter, during mass emergency situations has led to numerous opportunities to gain timely access to valuable information that can help disaster relief authorities act quicker and more efficiently. Specifically, as a disaster unfolds, information shared on social media can provide insights into the infrastructure and utility damage, casualties, and missing people. Recent studies have focused on collecting and manually annotating disaster data with respect to such situational awareness categories, followed by training machine learning classifiers to automatically identify situational awareness information, useful for relief operations (Alam et al., 2018b; Ashktorab et al., 2014).

However, disaster events produce large amounts of user-generated data, of which only a small fraction can be annotated, due to the time-sensitive nature of the problem, together with high annotation costs, and also inherent subjectivity associated with annotating data (e.g., tweets).

To address this limitation, we propose a semi-supervised multimodal approach that can leverage the copious amounts of unlabeled data to improve the performance on various multimodal tasks. Specifically, we extend the *FixMatch* (Sohn et al., 2020) algorithm proposed for semi-supervised image classification to a multimodal setting. To account for subjective annotations and potentially overlapping labels, we use soft pseudo-labels instead of the original hard pseudo-labels. We apply the adapted *FixMatch* to the CrisisMMD labeled dataset and tasks (Alam et al., 2018b), to improve the performance of supervised baselines through the use of unlabeled data. We use 122K unlabeled tweets, containing both text and images, collected automatically using text queries about disasters that occurred during the year of 2017. Experimental results show that our proposed approach produces performance improvements on all three CrisisMMD tasks. To our knowledge, we are the first to propose a semi-supervised method for multimodal data using FixMatch and text-based searches for collecting a large unsupervised dataset. While our experiments focus on disaster tweets, the method can be easily generalized. Finally, we provide an extensive error analysis of our models. We analyze how the supervised model's predictions change with the introduction of unlabeled data and reinforce the importance of our improved version of *FixMatch*.

Our contributions are as follows:

**(1)** We extend *FixMatch* algorithm to a multimodal scenario and offer two extensions to the original approach relevant for text and multimodal datasets. **(2)** We show that inexpensive unlabeled data gathered using text queries and basic preprocessing can be leveraged by our multimodal FixMatch to improve performance on 3 classification tasks. **(3)** We provide a detailed analysis into the predictions of the semi-supervised approaches, and compare them to their supervised counterparts.

## 2 Related Work

**Semi-supervised learning.** Semi-supervised learning is the approach of combining labeled data with large amounts of unlabeled data during training. *MixMatch* (Berthelot et al., 2019b) uses a sharpening technique, and guesses low-entropy labels for augmented unlabeled data. Next, it employs *MixUp* (Zhang et al., 2017) to blend the labeled and unlabeled examples. *FixMatch* (Sohn et al., 2020) combines two standard semi-supervised techniques: consistency regularization (Rasmus et al., 2015; Sajjadi et al., 2016; Tarvainen and Valpola, 2017) and pseudo-labeling (Lee et al., 2013). The pseudo-labels are generated using the current model's predictions on weakly-augmented unlabeled images. Next, the model tries to predict the pseudo-labels for strongly augmented versions of the same images. *Noisy Student Training* (Xie et al., 2020) first trains a teacher model on the labeled data to predict pseudo-labels for the unlabeled examples. Next, it trains a larger student model on all the data (i.e. labeled and unlabeled) using augmentation and dropout. The teacher model is then replaced by the student, and the process is repeated until convergence. Text and image methods are usually related: *MixText* (Chen et al., 2020) is an adaptation of *MixMatch* for text, while *UDA* (Xie et al., 2019) is introduced both for images and text.

**Disaster tweet classification.** A significant body of research focuses on the benefits of social media information for improving disaster relief efforts. Some of these studies focus solely on the analysis of textual data (e.g., tweets) (Imran et al., 2015; Kryvasheyeu et al., 2016; Li et al., 2018a; Enenkel et al., 2018; Alam et al., 2018a), while others focus only on the analysis of images (Bica et al., 2017; Nguyen et al., 2017; Li et al., 2019; Weber et al., 2020). However, many tweets posted during disasters contain both text and images, which if studied jointly, can provide a better portrayal of the damage produced by disasters, or the needs of the affected individuals. Therefore, it is not surprising that multimodal models in the disaster space have recently started to gain popularity (Mouzannar et al., 2018; Rizk et al., 2019; Gautam et al., 2019; Nalluru et al., 2019; Agarwal et al., 2020; Abavisani et al., 2020; Xukun and Caragea, 2020; Hao and Wang, 2020).

These existing approaches, however, do not use the large amounts of unlabeled multimodal data generated during disasters. In this paper, we propose a semi-supervised approach to leverage this data to improve the multimodal disaster tweet classification. Our approach extends *FixMatch* (originally proposed for image classification) to the multimodal setting and introduces two enhancements.

## 3 Methods

**Baseline Modeling.** First, we experiment with an image-only model, *ResNet*-152 (He et al., 2016), on top of which we add a linear layer for classification. Next, we use a *Multimodal Bitransformer* (*MMBT*) (Kiela et al., 2019) to leverage both the image and text for disaster tweet classification, as it already showed good results on this task (Sosea et al., 2021). We randomly crop and rescale the input images to 224x224, a common size for these types of networks, and also perform a standard horizontal flip and shift augmentation. We denote these approaches by *ResNet Aug* and *MMBT Aug*.

**Semi-supervised learning.** To leverage the large amounts of data generated during disaster events, we adapt the *FixMatch* (Sohn et al., 2020) algorithm to the multimodal setting. *FixMatch* obtains impressive performance on several Computer Vision tasks by combining consistency regularization (Sajjadi et al., 2016; Laine and Aila, 2016) and pseudo-labeling (McLachlan, 1975). *FixMatch* computes the overall loss $l$ as a weighted sum of two loss terms $l = l_s + \lambda_u l_u$, where $\lambda_u$ is a weighting parameter, $l_s$ is the loss on labeled data, and $l_u$ is the loss on unlabeled data. Specifically, in the multimodal setting, the labeled loss is defined as:

$$l_s = \frac{1}{B}\sum_{b=1}^{B} H(p_b, p_m(\alpha(x_b^{img}), \beta(x_b^{txt})))$$

where $B$ is the batch size, $H$ is the cross-entropy loss, $p_b$ is the one-hot encoding of the true label of a multimodal tweet $(x_b^{img}, x_b^{txt})$, and $p_m$ is the model's prediction (i.e., probability distribution over possible classes $y$) on a weakly augmented image, $\alpha(x_b^{img})$, and weakly augmented text, $\beta(x_b^{txt})$. The unlabeled loss is defined as:

$$l_u = \frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbb{1}_\tau(q_b) H(\hat{q}_b, p_m(\mathcal{A}(u_b^{img}), \mathcal{B}(u_b^{txt})))$$

where $\mu$ is the ratio between the number of labeled and unlabeled examples in a batch, and $q_b = p_m(\alpha(u_b^{img}), u_b^{txt})$ is the probability distribution over classes $y$, for the unlabeled example $(u_b^{img}, u_b^{txt})$. The function $\mathbb{1}_\tau(q_b)$ is used to filter out examples for which the prediction confidence, i.e., $\max_y(q_b)$, is less than a threshold, $\tau$. For the remaining examples, the prediction is converted to a pseudo-label using $\hat{q}_b = \arg\max_y(q_b)$. Finally, the cross-entropy loss is computed between the one-hot encoding of this pseudo-label and the prediction of the model on a strongly augmented version of the current image, $\mathcal{A}(u_b^{img})$, and the corresponding augmented text, $\mathcal{B}(u_b^{txt})$. The strong augmentations for image use either RandAugment (Cubuk et al., 2020) or CTAugment (Berthelot et al., 2019a). For text augmentation we experiment with EDA (Wei and Zou, 2019) and back-translation (Edunov et al., 2018). We offer more details about our text augmentation methods in Appendix F.

In this paper, we apply the *FixMatch* algorithm to our multimodal disaster domain, using *MMBT* as the base model. To understand the benefits of the multimodal representation, we also apply *FixMatch* on images only, using *ResNet-152* as the base model. We denote these methods by *MMBT FixMatch* and *ResNet FixMatch*, respectively.

**FixMatch Enhancements.** We propose two key enhancements to the unlabeled loss computation. First, we use *soft* pseudo-labels ($q_b$) instead of the hard labels ($\hat{q}_b$) used in the original paper:

$$l_u^{LS} = \frac{1}{\mu B} \sum_{b=1}^{\mu B} H(q_b, p_m(\mathcal{A}(u_b^{img}), \mathcal{B}(u_b^{txt})))$$

We argue that, in the disaster domain, there can be significant semantic overlap between two labels. For instance, in Figure 1e, which is labeled with *Rescue, volunteering, or donation effort* for the humanitarian task, there is a destroyed building in the background. By using soft labels, we can also incorporate information about the *Infrastructure and utility damage* class instead of stirring the model towards confidently predicting the example into the *Rescue, volunteering, or donation effort* class.

Second, we consider a variable weighting scheme for the loss, $l$. Originally, *FixMatch* employed a fixed weighting between the labeled and unlabeled loss (e.g., $\lambda_u = 1$). We argue that the predictions of the model during the first few epochs are not qualitative, hence using the predicted labels of unlabeled data can hurt the performance. To prevent that, we employ a linear growth of the unlabeled loss. Starting with 0 in the first epoch, we increase this loss in steps of 2 each epoch. Our loss becomes $l^{LS} = l_s + \lambda_u(t) l_u^{LS}$, where $\lambda_u(t) = 2t$, and $t$ is the epoch number. We denote the corresponding *MMBT* semi-supervised model by *MMBT Fixmatch LS*, while the corresponding *ResNet-152* model is denoted by *Resnet Fixmatch LS*.

## 4 Experiments

**Labeled Data.** We evaluate our semi-supervised multimodal approach on CrisisMMD (Alam et al., 2018b), a multimodal Twitter dataset from natural disasters. The dataset contains $18,000$ tweets with both text and images extracted during disasters such as the Iraq-Iran Earthquakes or Hurricanes Irma, Harvey and Maria. CrisisMMD was manually labeled for three classification tasks: *(1) Informativeness:* A tweet is labeled as *Informative* or *Not Informative*, depending on whether the tweet is useful for humanitarian aid purposes or not useful. *(2) Humanitarian:* We use the 5-class version of this data (Ofli et al., 2020) to alleviate the skewed label distribution. *(3) Damage Assessment.* We use a 2-class version of this data, similar to prior works (Li et al., 2018b). Each tweet image is labeled as depicting *Damage* or *No Damage*.

**Unlabeled Data.** We show that, by using text queries and preprocessing for collecting the unlabeled corpus, the performance of FixMatch can be improved even though the two datasets are not sampled from the same distribution. We used the Twitter Streaming API with a list of relevant keywords for the text in the training dataset. Then we selected $122k$ unique tweets containing both text and images that do not overlap with CrisisMMD. We provide more details in Appendix D.

**Experimental Setup.** To separately assess the impact of using multimodal data and of introducing text augmentations, we conduct our experiments in two stages. First, to ensure a fair comparison with the ResNet-based models, which only use the image modality, we experimented with versions of MMBT-based models where no text augmentation is used ( $\mathcal{B}$ is the identity function). Second, we analyze the impact of augmenting each modality separately or performing both text and image augmentations. We propose the following Fixmatch adaptations: **1)** $FixMatchLS_{img}$ solely augments the image, **2)** $FixMatchLS_{eda}$ only augments the text using EDA, **3)** $FixMatchLS_{img+eda}$ augments both modalities, using EDA for text augmentation, and **4)** $FixMatchLS_{img+bt}$ augments both modalities, using back-translation for text augmentation.

All hyperparameters and model setups are available in Appendix A. To attain statistically significant results, we ran each experiment $5$ times and report the average of the results. To improve reproducibility, we will release the splits (see Appendix B) for each task alongside our code.

| MODEL | INFORMATIVE P | R | F1 | DAMAGE P | R | F1 | HUMANITARIAN P | R | F1 |
|---|---|---|---|---|---|---|---|---|---|
| RESNET AUG | 0.767 | 0.767 | 0.766 | 0.861 | 0.863 | 0.858 | 0.804 | 0.812 | 0.806 |
| RESNET FIXMATCH | 0.793 | 0.793 | 0.793 | 0.886 | 0.887 | 0.886 | 0.820 | 0.820 | 0.816 |
| RESNET FIXMATCH LS | 0.804 | 0.804 | 0.804 | **0.887** | **0.888** | **0.887** | 0.829 | 0.825 | 0.819 |
| MMBT AUG | 0.786 | 0.785 | 0.785 | 0.865 | 0.867 | 0.865 | 0.865 | 0.862 | 0.863 |
| MMBT FIXMATCH | 0.808 | 0.806 | 0.806 | 0.882 | 0.882 | 0.882 | 0.865 | 0.865 | 0.864 |
| MMBT FIXMATCH LS | **0.820** | **0.820** | **0.820** | 0.885 | 0.882 | 0.883 | **0.873** | **0.872** | **0.872** |

Table 1: Results on CrisisMMD tasks using image augmentations - best results for each task are highlighted in **bold**.

| MODEL | INFORMATIVE 250/CLASS P | R | F1 | INFORMATIVE 500/CLASS P | R | F1 | HUMANITARIAN P | R | F1 |
|---|---|---|---|---|---|---|---|---|---|
| $MMBT(supervised)$ | 0.666 | 0.667 | 0.666 | 0.713 | 0.704 | 0.705 | 0.865 | 0.862 | 0.863 |
| $FixMatchLS_{img}$ | 0.695 | 0.688 | 0.689 | 0.741 | 0.730 | 0.730 | 0.873 | 0.872 | 0.872 |
| $FixMatchLS_{eda}$ | 0.687 | 0.673 | 0.673 | 0.741 | 0.731 | 0.722 | 0.878 | 0.877 | 0.877 |
| $FixMatchLS_{img+eda}$ | 0.701 | 0.702 | 0.701 | 0.759 | 0.756 | 0.756 | **0.885** | **0.881** | **0.881** |
| $FixMatchLS_{img+bt}$ | **0.744** | **0.742** | **0.743** | **0.772** | **0.759** | **0.760** | 0.880 | 0.879 | 0.878 |

Table 2: Results on CrisisMMD tasks after adding text augmentations - best results are highlighted in **bold**.

## 5 Results

**Disaster Tweet Classification.** We show experimental results using the previously described approaches in Tables 1 and 2. As it can be seen in Table 1, our enhanced *FixMatch* models, which use soft-labels and a linear schedule for weighting the unlabeled loss, consistently outperform all the other models on all tasks. On the *Informative* task, *MMBT FixMatch LS* improves the F1 performance of the supervised *MMBT Aug* model by as much as 3.5%. Interestingly, on the *Humanitarian* task, the *MMBT FixMatch* approach, which uses hard labels and a constant loss weighting, obtains similar performance to *MMBT Aug*, which uses no unlabeled data. We attribute this to the nature of the humanitarian task, where the boundary between classes may not be well defined, i.e., an example annotated with class $y_1$ can exhibit characteristics specific to a different class $y_2$. We argue that the use of the "hard labeling" mechanism for these types of tasks can lead to poor model performance. On the other hand, the *MMBT FixMatch LS* manages to prevent this shortcoming, and obtains an F1 increase of 1% over the *MMBT Aug* model. Finally, on the *Damage* task, we observe that the *ResNet* and the *MMBT* perform similarly, which is not surprising, given that the examples in this task were annotated based only on the image in the tweet. However, similar to the *Informative* task, the best semi-supervised approach outperforms the other method by as much as 2.9% F1. Table 2 shows the improvement obtained for the best model so far (*MMBT FixMatch LS*) when also introducing text augmentation. Here, to test the limit of our approach, we also experiment with few labeled examples (250/500) per class on the informative task. Our results show that while there is no clear winner when augmenting only one modality ($FixMatchLS_{img}$ performs better than $FixMatchLS_{eda}$ on Informative task, but worse on Humanitarian), it is clear that augmenting both modalities is always the best option. Using back-translation instead of EDA gives better results on the Informative tasks, but there is a slight decrease in performance on the Humanitarian task.

All improvements of the enhanced *FixMatch* over baselines are statistically significant, according to a t-test with $p < 0.01$. These results show the feasibility of our proposed *FixMatch* variant: using *cheap to acquire* unlabeled data, the performance of supervised models is significantly improved.

**Error Analysis** We also investigate common errors of the supervised models, which are corrected by our *FixMatch* approach. We explain a few patterns and provide supporting examples in Appendix E. Our proposed *FixMatch* variant is able to correct these types of errors. Moreover, the *FixMatch* model is confident in its predictions, usually assigning a probability over 90% to the correct class.

## 6 Conclusion

We extended *FixMatch* to multimodal data and proposed two improvements. We applied the improved *FixMatch* on three disaster-centric multimodal tweet classification tasks, and showed that the approach can leverage large unlabeled data to improve supervised model performance. Our semi-supervised approach is general enough and can be easily applied to other datasets, being at the same time very efficient as it does not add any inference complexity to the base model.

# References


Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689.

Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 346–353.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018a. Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In *Twelfth International AAAI Conference on Web and Social Media*.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018b. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Zahra Ashktorab, Christopher Brown, Manojit Nandi, and Aron Culotta. 2014. Tweedr: Mining twitter to inform disaster response. In *ISCRAM*, pages 269–272.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2019a. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. 2019b. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059.

Melissa Bica, Leysia Palen, and Chris Bopp. 2017. Visual representations of disaster. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1262–1276.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. *arXiv preprint arXiv:2004.12239*.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Markus Enenkel, Sofia Martinez Saenz, Denyse S Dookie, Lisette Braman, Nick Obradovich, and Yury Kryvasheyeu. 2018. Social media data analysis and feedback for advanced disaster risk management. In *Social Web in Emergency and Disaster Management*.

Akash Kumar Gautam, Luv Misra, Ajit Kumar, Kush Misra, Shashwat Aggarwal, and Rajiv Ratn Shah. 2019. Multimodal analysis of disaster tweets. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 94–103. IEEE.

Haiyan Hao and Yan Wang. 2020. Leveraging multimodal social media data for rapid disaster damage assessment. *International Journal of Disaster Risk Reduction*, page 101760.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4):67.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. 2016. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3).

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018a. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27.

Xukun Li, Doina Caragea, Huaiyu Zhang, and Muhammad Imran. 2018b. Localizing and quantifying damage in social media images. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 194–201. IEEE.

Xukun Li, Doina Caragea, Huaiyu Zhang, and Muhammad Imran. 2019. Localizing and quantifying infrastructure damage using class activation mapping approaches. *Social Network Analysis and Mining*, 9(1):44.

Geoffrey J McLachlan. 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369.

Hussein Mouzannar, Yara Rizk, and Mariette Awad. 2018. Damage identification in social media posts using multimodal deep learning. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2018)*, Rochester, NY.

Ganesh Nalluru, Rahul Pandey, and Hemant Purohit. 2019. Relevancy classification of multimodal social media streams for emergency services. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 121–125. IEEE.

Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. 2017. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 569–576. ACM.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. In *17th International Conference on Information Systems for Crisis Response and Management*. ISCRAM.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. *arXiv preprint arXiv:1507.02672*.

Yara Rizk, Hadi Samer Jomaa, Mariette Awad, and Carlos Castillo. 2019. A computationally efficient multimodal classification approach of disaster-related twitter images. In *Proceedings of the 34th ACM/SIGAPP symposium on applied computing*, pages 2050–2059.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Tiberiu Sosea, Iustin Sirbu, Cornelia Caragea, Doina Caragea, and Traian Rebedea. 2021. Using the image-text relationship to improve multimodal disaster tweet classification.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.

Ethan Weber, Nuria Marzo, Dim P Papadopoulos, Aritro Biswas, Agata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. 2020. Detecting natural disasters, damage, and incidents in the wild. *arXiv preprint arXiv:2008.09188*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Li Xukun and Doina Caragea. 2020. Improving disaster-related tweet classification with a multimodal approach. In *ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management*.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

## A Fixmatch Setup

First, we tried to find the best *FixMatch* setup for our experiments (without our extension). To achieve this, we experimented with a variety of setups, by manually tuning the *FixMatch* hyperparameters and choosing the values that yield the best F1 score:

- For the ratio $\mu$ between unlabeled and labeled examples we tried values from the set $\{3, 5, 7\}$. We observed that setting $\mu$ to 7 produced the best results. We did not try values bigger that 7 due to computation limitations. However, 7 is the reported best $\mu$ in the original *FixMatch* paper, too.

- For the weight of the unlabeled loss, $\lambda_u$, we experimented with values in the set $\{1, 10, 50, 100\}$, and obtained the best results with value 1 (similar to the original paper).

- For image preprocessing, we cropped and rescaled all images to 224x224 size. We also tried to reduce the size of the images to 96x96 to improve computational performance, but the results were heavily affected.

- For image augmentation we used random horizontal flip as weak augmentation and *RandAugment* as strong augmentation in all our experiments.

- Initially, the original paper used no dropout, but we observed that adding 0.2 dropout improved the results.

- Exponential moving average (EMA) with decay 0.999 was kept as in the original paper. We experimented with a smaller decay or without EMA, but this negatively impacted the performance.

- Instead of SGD and *cosine learning rate schedule*, we used *Adam* with a *ReduceOnPlateau schedule*, which improved the results.

- We experimented with learning rates from the set $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$, and picked $10^{-5}$ as the optimal value.

- For the confidence threshold $\tau$, we found that 0.7 was the best for our tasks. This is compatible with the value chosen in the original paper on the *ImageNet* dataset. We experimented with values in the set $\{0.5, 0.7, 0.85, 0.95\}$.

- Due to computation limitations, we used a batch size of 8 with 40 gradient accumulation steps in all our experiments.

We apply the best hyperparameters found for the classic *FixMatch* algorithm to our extended *FixMatch LS* version. Our changes are:

- we used *soft labels* instead of hard pseudo-labels for the unlabeled data

- we used a *linear schedule* for the unlabeled loss weight $\lambda_u$

Note that replacing pseudo labels with soft labels for the unlabeled data completely removes the confidence threshold parameter, $\tau$. However, introducing the linear schedule $\lambda_u(t) = c * t$ for the unlabeled loss adds one extra parameter, $c$. This is the only hyperparameter tuned for *FixMatch LS*. After experimenting with values in the set $\{1, 2, 3\}$, we choose $\lambda_u(t) = 2 * t$ to be our weight in all the experiments.

In order to attain statistically significant results, we ran each experiment 5 times and report the average of the results. The training process took 20 days to complete on a system with 4 Nvidia V100 GPUs, each experiment running for roughly 20 hours on a single GPU.

## B Splits

We show the number of examples from the train, development, and test sets for the 3 tasks in CrisisMMD in Table 3. Moreover we provide the class distributions in Table 4.

## C Predictions

We show comparisons between predictions of the *MMBT Aug* and the *FixMatch LS* model in Tables 5 and 6. We show the input samples and the ground truths in Figure 1.

## D Unlabeled Data

We collected data from Twitter during disasters that happened in 2017: California Wildfires, Mexico Earthquake, and Hurricanes Harvey, Irma, and Maria. The tweets were crawled using the Twitter streaming API (keywords such as #hurricaneharvey, #harvey, #hurricane) during the following disasters: Hurricane Harvey, Hurricane Irma, Hurricane Maria, Mexico Earthquake, and Chiapas

| DATASET | SIZE | TRAIN | DEV | TEST |
|---|---|---|---|---|
| INFORMATIVE | 13494 | 10795 (80%) | 1349 (10%) | 1350 (10%) |
| DAMAGE | 6089 | 4262 (70%) | 913 (15%) | 914 (15%) |
| HUMANITARIAN | 8079 | 6126 (75.8%) | 998 (12.4%) | 955 (11.8%) |

Table 3: Data splits for each task

| DATASET | INFORMATIVE | DAMAGE | HUMANITARIAN |
|---|---|---|---|
| Labels | *uninformative* (55%) | *no damage* (70%) | *not humanitarian* (53%) |
| | *informative* (45%) | *damage* (30%) | *other relevant information* (22%) |
| | | | *rescue volunteering or donation effort* (15%) |
| | | | *infrastructure and utility damage* (9%) |
| | | | *affected individuals* (1%) |

Table 4: Labels distribution for each task

Earthquake. This collection was filtered for disaster relevance using a Naive Bayes classifier trained on CrisisLexT6 to ensure that it mostly contained tweets relevant to disasters. Subsequently, duplicate tweets, retweets and non-English tweets were removed. Finally, we selected only tweets that contained both an image and text.

In addition, we used several methods to clean and filter out duplicates between our dataset and CrisisMMD. This is done in order to make sure that test samples (from CrisisMMD) are not seen during training, not even as unlabeled examples (as part of our unlabeled dataset). First, we removed all retweets (tweets with the "RT" token), and normalized the texts removing characters repetitions (all consecutive identical characters of size > 2 are reduced to only 2 characters) and user mentions. Next, we removed duplicates using the drop_duplicates function from the pandas library.

The resulting unlabeled corpus will be made publicly available.

## E Error Analysis

We investigate common errors of the models that use no unlabeled data, which are corrected by our *FixMatch* models. To this end, we first sample 20 such examples for each CrisisMMD task, followed by manually inspecting the output probabilities and the contents of the image and text. We show some examples in Figure 1, and provide the full model predictions in Appendix C. We observed a few patterns. First, we spotted some erroneous predictions due to semantic disparities between the textual and the image modalities (i.e., the image and text pinpoint to different labels, hence the final label is subjective). An example is shown in Figure 1b. Second, we encountered a significant number of examples where the image modality is distorted, or contains noise. For instance, in Figure 1c, the photo contains perturbations (i.e., the rain drops) that hinder the capability to observe the main focus of the picture: a *collapsed huge crane*. Third, we observe some examples which contain characteristics specific to more than one class. In Figure 1e, even though the main focus of the tweet is on *Rescue and volunteering* efforts, the image also exhibits traits of the *Infrastructure and utility damage* class: a destroyed building.

Our proposed *FixMatch* variant is able to correct these types of errors. Moreover, the *FixMatch* model is confident in its predictions, usually assigning a probability over 90% to the correct class.

## F Text Augmentation

For the text augmentation, we explore with two different techniques:

- easy data augmentation (EDA (Wei and Zou, 2019)), which consists of randomly applying 4 possible operators: syonoym replacement, random insertion of a word, random swap of 2 words, random deletion of a word. The longer a sentence is, the more transformations will be applied to it, as we used the EDA framework for applying these transformations on 10% of the words in each text.
- back-translation ((Edunov et al., 2018)), as described in UDA (Xie et al., 2019) and MixText (Chen et al., 2020); it consists of translating a sentence to another language and than

| IMAGE | MODEL | LABEL | |
|---|---|---|---|
| | | *informative* | *not informative* |
| (a) | MMBT AUG | 0.71 | 0.29 |
| | FIXMATCH LS | 0.09 | 0.91 |
| (c) | MMBT AUG | 0.24 | 0.76 |
| | FIXMATCH LS | 0.98 | 0.02 |

Table 5: Examples of predictions for the Informative Task

| IMAGE | MODEL | LABEL | | | | |
|---|---|---|---|---|---|---|
| | | *not hum.* | *other* | *rescue* | *damage* | *affected* |
| (b) | MMBT AUG | 0.36 | 0.06 | 0.04 | 0.51 | 0.09 |
| | FIXMATCH LS | 0.89 | 0.01 | 0.02 | 0.07 | 0.01 |
| (d) | MMBT AUG | 0.02 | 0.03 | 0.16 | 0.78 | 0.01 |
| | FIXMATCH LS | 0.03 | 0.03 | 0.90 | 0.01 | 0.03 |
| (e) | MMBT AUG | 0.01 | 0.01 | 0.02 | 0.95 | 0.01 |
| | FIXMATCH LS | 0.01 | 0.01 | 0.93 | 0.04 | 0.01 |

Table 6: Examples of predictions for the Humanitarian Task

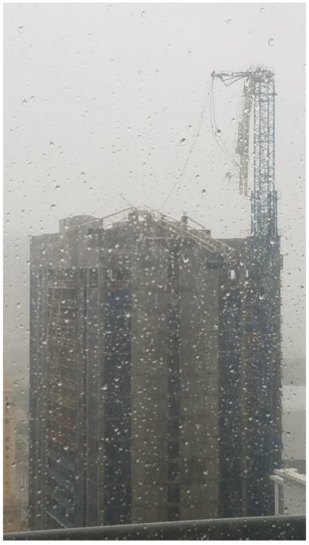
back to the original language, thus obtaining a new sentence having the same meaning. Inspired by MixText(Chen et al., 2020), we use FairSeq(Ott et al., 2019) with Russian as an intermediate language and random sampling with $0.9$ temperature instead of beam search in order to ensure the diversity of the augmentations.

## G Limitations

While our approach provides significant improvements on all CrisisMMD tasks, we also have to acknowledge the limitations of the proposed method. As it generally is the case with semi-supervised approaches, the training time is significantly increased, as more data needs to be passed through the model until convergence, comparing to a supervised approach. Regarding our method of collecting unalabeled data by searching for relevant keywords, although it is generic and could be applied to datasets from other domains, it is limited for datasets containing tweets. For other types of datasets, obtaining a relevant unlabeled corpus in the same manner could be more challenging.

(a) *This 4 BD/ 2 BA in Mora MUST be seen. Call, text or direct message me for more info!*

(b) *St. Augustine bed & breakfast picking up the pieces after Hurricane Irma*



(c) *A huge crane just collapsed on top of building in down town Miami*

(d) *Irma update: Free roof help available*

(e) *Magnitude 6.1 aftershock hits Mexico as search for people and pets continues*

Figure 1: Examples of errors of the *MMBT* model that are corrected by *FixMatch* on the *Informativeness* and *Humanitarian* CrisisMMD tasks: **(a)** MMBT: *informative*; True: *not informative* **(b)** MMBT: *infrastructure and utility damage*; True: *not humanitarian* **(c)** MMBT: *not informative*; True: *informative* **(d)** MMBT: *infrastructure and utility damage*; True: *rescue, volunteering, or donation effort* **(e)** MMBT: *infrastructure and utility damage*; True: *rescue, volunteering, or donation effort*