Semi-Supervised Few-Shot Learning for Fine-Grained Disaster Tweet Classification

Henry Peng Zou
University of Illinois Chicago
pzou3@uic.edu

Cornelia Caragea
University of Illinois Chicago
cornelia@uic.edu

Yue Zhou
University of Illinois Chicago
yzhou232@uic.edu

Doina Caragea
Kansas State University
dcaragea@ksu.edu

ABSTRACT

The shared real-time information about natural disasters on social media platforms like Twitter and Facebook plays a critical role in informing volunteers, emergency managers, and response organizations. However, supervised learning models for monitoring disaster events require large amounts of annotated data, making them unrealistic for real-time use in disaster events. To address this challenge, we present a fine-grained disaster tweet classification model under the semi-supervised, few-shot learning setting where only a small number of annotated data is required. Our model, CrisisMatch, effectively classifies tweets into fine-grained classes of interest using few labeled data and large amounts of unlabeled data, mimicking the early stage of a disaster. Through integrating effective semi-supervised learning ideas and incorporating TextMixUp, CrisisMatch achieves performance improvement on two disaster datasets of 11.2% on average. Further analyses are also provided for the influence of the number of labeled data and out-of-domain results.

Keywords

Crisis tweet classification, semi-supervised few-shot learning, pseudo-labeling, TextMixUp.

1. INTRODUCTION

In times of natural disasters, individuals share content, facts, recommendations, and warnings about the disaster in real-time on social media platforms such as Twitter and Facebook. Such information is crucial to help volunteers, emergency managers, and response organizations become more situationally aware and efficient in their rescue activities (Varga et al. 2013; Vieweg et al. 2014).

Although existing works leverage such information to build models to monitor disaster events, many approaches require annotating large amounts of data when disasters happen, which are unrealistic due to the limited response time (C. Caragea et al. 2016; Chowdhury et al. 2020). Current semi-supervised approaches also *over-assume* the number of available labels (e.g., more than 50 labels per class) (Alam, Joty, et al. 2018b; P. Karisani and N. Karisani 2021; Sirbu et al. 2022), which is hard to obtain when the number of classes is large. In addition, while the popular coarse, binary classification identifying whether a tweet is disaster-relevant can be useful, it is more informative to have fine-grained, multi-categorical classifications providing disaster information from different angles (Plotnick et al. 2015; Reuter et al. 2018; Imran, Ofli, et al. 2020; Alam, Qazi, et al. 2021).

To address the above challenges, we investigate the problem of fine-grained disaster tweet classification under the semi-supervised, few-shot learning setting. It aims to classify tweets during a disaster event into fine-grained classes of interest (e.g., injured or dead people, infrastructure and utility damage, caution and advice, rescue volunteering, or donation effort) (Alam, Qazi, et al. 2021). In addition, in our few-shot setting, we only utilize five labeled data points per class, with the rest data being unlabeled and naturally imbalanced. This setting mimics the early stage of a disaster, which is also the most valuable time for rescue and escape.

This paper first studies the effectiveness of various semi-supervised learning components on leveraging few labeled data and large amounts of unlabeled data for disaster tweet classification. Then, we extend the pseudo-labeling algorithm (Lee et al. 2013; B. Zhang et al. 2021) through entropy minimization, data augmentation, and consistency regularization. Particularly, we incorporate TextMixUp (H. Zhang et al. 2018; Chen et al. 2020), which encourages models to behave linearly among samples and avoid overfitting. Then, we propose CrisisMatch, by combining those effective components for fine-grained disaster tweet classification in the semi-supervised few-shot setting. Experimental results show that the proposed CrisisMatch achieves over 11.2% performance improvement on average on two disaster datasets. We also provide further analyses for CrisisMatch on the influence of the number of labeled data and out-of-domain results.

2. RELATED WORK

Semi-supervised learning. Semi-supervised learning is developed to alleviate the reliance on labeled data by leveraging unlabeled data to train more robust models (Lee et al. 2013; Berthelot et al. 2019; Xie et al. 2020; B. Zhang et al. 2021). Self-training adopts the idea of using the output probability of the model as a soft label for unlabeled data (Scudder 1965; McLachlan 1975; Lee et al. 2013; Xie et al. 2020). Pseudo-labeling modifies self-training by using hard labels, instead of soft labels, and confidence thresholding to reduce confirmation bias and select high-quality pseudo-labels for training (Lee et al. 2013; B. Zhang et al. 2021). Mean Teacher (Tarvainen and Valpola 2017) proposes to use the exponential moving average of model weights for label predictions on unlabeled data. MixMatch (Berthelot et al. 2019) utilizes sharpening to encourage low-entropy prediction for unlabeled data and uses MixUp (H. Zhang et al. 2018) to mix labeled and unlabeled data. MixText (Chen et al. 2020) adapts MixUp to text settings by interpolating hidden representations of texts.

Disaster tweet classification. Disaster tweet classification has made huge progress in recent years to improve crisis relief operations. (Imran, Elbassuoni, et al. 2013; Imran, Castillo, et al. 2015) proposed to classify disaster tweets to obtain useful information for disaster understanding and rescue. (Nguyen et al. 2017) introduced Convolutional Neural Networks (CNNs) to classify informative disaster-related tweets. (Kruspe et al. 2019) studied the supervised few-shot learning setting, where only a few labeled data are used for training a disaster tweet classifier. (Li, D. Caragea, and C. Caragea 2021) combined self-training with CNN and BERT pre-trained language models to improve the performance of classifying disaster tweets where only unlabeled data is available. Other works (Alam, Joty, et al. 2018a; Mazloom et al. 2019; Li, D. Caragea, C. Caragea, and Herndon 2018) explored unsupervised domain adaptation in which only unlabeled data is available for the current crisis event while labeled data is available from previous disaster events. However, most prior works either focus on coarse, binary classification or assume that many labeled data is available. In contrast, this work studies fine-grained disaster tweet classification under the few-shot setting, which we believe can benefit the early stage of disaster analysis and rescue.

3. METHOD

This section first reviews several classical semi-supervised learning ideas and components, including self-training, entropy minimization, and consistency regularization, then describes TextMixUp and our CrisisMatch algorithm for the task of fine-grained disaster tweet classification in the few-shot setting.

3.1 Self-Training and Entropy Minimization

In the semi-supervised few-shot learning setting, there are usually only a few labeled data but a large amount of unlabeled data. Self-training takes advantage of unlabeled data by using the model itself to infer predictions on unlabeled data, and then utilizing these predictions as pseudo-labels for training (Scudder 1965; McLachlan 1975; Lee et al. 2013; Xie et al. 2020). However, this may result in the issue of confirmation bias: If these pseudo-labels are incorrect and the model is trained on them, the model can become worse and worse, continually confirming its own incorrect bias. A common strategy to alleviate this problem is to use a threshold to select only pseudo-labels whose largest class probability surpasses the threshold.

Besides, a basic assumption in many semi-supervised learning methods is that data in the same class are clustered together and thus a good classifier's decision boundary between classes should not pass through high-density regions of the data manifold. Entropy minimization achieves this by encouraging the model to produce low-entropy predictions on unlabeled data (Grandvalet and Bengio 2004; Miyato et al. 2018). This is because the model will by necessity output high-entropy predictions for some samples if the decision boundary falls in high-density regions. Here we introduce two ways to implement entropy minimization: Pseudo-labeling and Sharpening.

Pseudo-Labeling (Lee et al. 2013; B. Zhang et al. 2021) uses hard (i.e., one-hot) labels from high-confidence predictions on unlabeled data as the targets for training, which implicitly enforces the model to output low-entropy predictions. Formally, pseudo-labeling minimizes the following loss function:

$$\frac{1}{\mu B}\sum_{b=1}^{\mu B}\mathbb{1}(\max(p_m(y|\hat{u}_b)) > \tau)L(q_m(y|\hat{u}_b), p_m(y|\hat{u}_b)) \tag{1}$$

where μ is the ratio of unlabeled data to labeled data, B is the batch size of labeled data, $\hat{u}_b$ is a ***stochastically*** data augmentation function, u_b represents an unlabeled sample, p_m denotes model's probability prediction, L is L2 loss or cross-entropy loss, τ is a fixed threshold and $q_m(y|\hat{u}_b)$ is the hard one-hot pseudo-label.

Sharpening in MixMatch (Berthelot et al. 2019) uses soft labels and explicitly decreases the entropy of predicted label distribution on unlabeled data by adjusting all probabilities to the power of $1/T$ and normalizing:

$$\text{Sharpen}(p_i, T) = \frac{p_i^{\frac{1}{T}}}{\sum_{j=1}^{C} p_j^{\frac{1}{T}}} \tag{2}$$

where T is the temperature hyperparameter and C is the number of total classes. Lowering the temperature hyper-parameter T encourages the model to produce low-entropy predictions. When $T \to 0$, the predicted label becomes a one-hot hard label.

3.2 Data Augmentation and Consistency Regularization

Data augmentation is a common technique to alleviate overfitting and help regularize models, especially when using smaller datasets. It artificially increases the amount of training data by generating perturbed inputs with transformations that are assumed not to change original class semantics. For instance, synonym replacement, random swap, random insertion and random deletion are convenient and easy data augmentations for text (Wei and Zou 2019). More advanced techniques such as back-translation (Fadaee et al. 2017; Sugiyama and Yoshinaga 2019) and paraphrasing (Kumar et al. 2019) are also proposed to generate more diversified augmented data but are more costly and complicated to implement.

Consistency regularization leverages the idea that a classifier should have similar predictions for data before and after augmentation (Bachman et al. 2014; Sajjadi et al. 2016). A straightforward implementation is to add the loss term:

$$\sum_{b=1}^{\mu B}||p_m(y|\hat{u}_b) - p_m(y|\hat{u}_b)||_2^2 \tag{3}$$

Another implicit way is to use the average prediction of several different augmentations of an unlabeled sample as the common pseudo-label of all augmented data (Berthelot et al. 2019). Formally,

$$q_m(y|\hat{u}_b) = \frac{1}{K}\sum_{k}^{K} p_m(y|\hat{u}_{b,k})) \tag{4}$$

where K is the number of augmentations, $\hat{u}_{b,k}$ is the k-th augmented data, and $q_m(y|\hat{u}_b)$ is the common pseudo-label for all the corresponding augmented unlabeled data.

3.3 MixUp and TextMixUp

MixUp is a technique proposed by (H. Zhang et al. 2018) that can regularize the model to behave linearly among training samples and alleviate overfitting. The idea is to generate numerous new virtual training samples by linearly mixing two randomly sampled input and their one-hot labels:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{5}$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \tag{6}$$

However, applying MixUp directly on text input seems infeasible since the interpolation of discrete text tokens makes no sense.

Algorithm 1 CrisisMatch algorithm.

1: **Input:** Labeled batch $\mathcal{X} = \{(x_b, y_b) : b \in (1, 2, \ldots, B)\}$, unlabeled batch $\mathcal{U} = \{u_b : b \in (1, 2, \ldots, \mu B)\}$, unsupervised loss weight w_u, confidence threshold τ, number of augmentations K.
2: $\hat{\mathcal{U}} = \{\}$
3: **for** $b = 1$ **to** B **do**
4: $\hat{x}_b = \text{Aug}(x_b)$ {*Data augmentation for labeled examples*}
5: **for** $k = 1$ **to** K **do**
6: $\hat{u}_{b,k} = \text{Aug}(u_b)$ {*Data augmentation for unlabeled examples*}
7: **end for**
8: $q_b = \frac{1}{K} \sum_k p_m(y \mid \hat{u}_{b,k})$ {*Compute average prediction across different augmentations of* u_b as guessed probability distribution target for all $\hat{u}_{b,k}$}
9: **if** $\max(q_b) \geq \tau$ **then**
10: $\hat{q}_b = \arg\max(q_b)$ {*Compute one-hot guessed labels from high-confidence predictions*}
11: $\hat{\mathcal{U}} \leftarrow (\hat{u}_{b,k}, \hat{q}_b)$ {*Use guessed labels as training targets for augmented unlabeled examples*}
12: **end if**
13: **end for**
14: $\hat{\mathcal{X}} = \big((\hat{x}_b, y_b); b \in (1, \ldots, B)\big)$ {*Augmented labeled examples and their labels*}
15: $\mathcal{W} = \text{Shuffle}\big(\text{Concat}(\hat{\mathcal{X}}, \hat{\mathcal{U}})\big)$ {*Shuffle all labeled and unlabeled data*}
16: $\tilde{\mathcal{X}} = \big(\text{TextMixUp}(\hat{\mathcal{X}}_i, \mathcal{W}_i); i \in (1, \ldots, |\hat{\mathcal{X}}|)\big)$ {*Apply TextMixUp to labeled data and entries from* $\mathcal{W}$}
17: $\tilde{\mathcal{U}} = \big(\text{TextMixUp}(\hat{\mathcal{U}}_i, \mathcal{W}_{i+|\hat{\mathcal{X}}|}); i \in (1, \ldots, |\hat{\mathcal{U}}|)\big)$ {*Apply TextMixUp to unlabeled data and the rest of* $\mathcal{W}$}
18: $\mathcal{L}_x = \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\tilde{x}, \tilde{y} \in \tilde{\mathcal{X}}} H(\tilde{y}, p_m(y \mid \tilde{x}))$ {*Compute loss for labeled data*}
19: $\mathcal{L}_u = \frac{1}{|\tilde{\mathcal{U}}|} \sum_{\tilde{u}, \tilde{q} \in \tilde{\mathcal{U}}} \|\tilde{q} - p_m(y \mid \tilde{u})\|_2^2$ {*Compute loss for unlabeled data*}
20: **Return:** $\mathcal{L}_x + w_u \mathcal{L}_u$

A more practical method is to interpolate hidden representations of texts at a certain layer and use the mixed representation for future layer and model prediction (Verma et al. 2019; Chen et al. 2020):

$$\tilde{h}_m = \lambda h_m^i + (1 - \lambda) h_m^j \tag{7}$$

$$\tilde{h}_l = g_l(\tilde{h}_{l-1}), l \in [m + 1, L] \tag{8}$$

where h_m^i are the hidden representation of m-th layer for sentence i, $g_l(\cdot)$ is the encoder function.

We refer to the above method as TextMixUp and it has the potential to create many more virtual training samples since it can interpolate representation at any layer of the encoder instead of just the input samples in the original MixUp.

3.4 Our Algorithm: CrisisMatch

In this section, we introduce our algorithm CrisisMatch, which incorporates the components and ideas described above for the task of semi-supervised few-shot disaster tweet classification. The complete algorithm for CrisisMatch is presented in Algorithm 1.

Given a labeled batch $\mathcal{X} = \{(x_b, y_b) : b \in (1, 2, \ldots, B)\}$ and a unlabeled batch $\mathcal{U} = \{u_b : b \in (1, 2, \ldots, \mu B)\}$. We first apply ***data augmentation*** to both labeled and unlabeled d ata. Specifically, we generate one augmented labeled sample $\hat{x}_b$ and K augmented unlabeled samples $\hat{u}_{b,k}$ (algorithm 1, line 4, 6). Then we implicitly enforce ***consistency regularization*** by using the average prediction across different augmentation of u_b as the guessed probability distribution target for all K augmented unlabeled samples $\hat{u}_{b,k}$ (algorithm 1, line 8).

To encourage ***entropy minimization***, we use hard pseudo-labeling by computing one-hot pseudo-labels from unlabeled augmented data that receive high-confidence predictions (algorithm 1, line 1 0). ***TextMixUp*** is then applied on shuffled labeled and unlabeled data to further regularize the model to behave linearly between samples and effectively leverage limited labeled data (algorithm 1, line 1 6). Lastly, we compute cross-entropy loss for labeled data and compute L_2 loss for unlabeled data since L_2 loss is bounded for probabilities and less sensitive to wrong predictions of unlabeled data (algorithm 1, line 18, 19).

Our algorithm CrisisMatch differs from MixMatch(Berthelot et al. 2019) mainly as follows: CrisissMatch uses hard pseudo-labeling for entropy minimization instead of sharpening in MixMatch. We empirically found that hard-pseudo-labeling achieves better results than sharpening on both disaster datasets, as shown in Section 4.5 and

Table 6. Besides, MixMatch uses all sharpened guessed labels on unlabeled for training, while CrisisMatch only selects high-confidence predictions as pseudo-labels for training. We argue that not all guessed labels on unlabeled data should be used since many of them may be incorrect, especially in the few-shot setting, and degenerate model performance. Lastly, CrisisMatch leverages TextMixUp rather than MixUp since TextMixUp can interpolate text hidden representations in any layer of encoder while MixUp only interpolates in input space and is not feasible for interpolating discrete text tokens.

4 EXPERIMENTS

4.1 Datasets

To evaluate the performance of our proposed methods and baseline methods for fine-grained disaster tweet classification, we use three datasets sampled and processed from HumAID((Alam, Qazi, et al. 2021): 1) Earthquake; 2) Wildfires; 3) Floods. We use Earthquake and Wildfires datasets for in-domain evaluation and Floods dataset for out-of-domain evaluation.

These datasets comprise tweets collected in natural disasters that occurred between 2016 and 2019, such as the 2018 California Wildfires and the 2019 Pakistan Earthquake. Originally, there are 10 classes for each disaster type. However, some classes have less than 100 data and even less than 10 data after splitting, and are difficult for effective evaluation. Therefore, we discard those classes with less than 100 data and use the 7-class version of the datasets for a more convincing evaluation. Table 1 shows all labels and class-wise distributions of the datasets used in our experiments. Table 2 shows the data splits for each used disaster dataset. We will release our sampled and processed datasets to make them convenient to use for future researchers.

Labels	**Wildfires**	**Earthquake**	**Floods**
caution_and_advice	245	629	246
infrastructure_and_utility_damage	673	728	453
injured_or_dead_people	1946	1489	409
not_humanitarian	1397	498	865
other_relevant_information	1349	707	1284
rescue_volunteering_or_donation_effort	2349	2049	5401
sympathy_and_support	633	2520	1040
total	8592	8620	9698

Table 1. Labels distribution for each dataset.

Datasets	**Size**	**Train(80%)**	**Dev(10%)**	**Test(10%)**
Wildfires	8592	6874	859	859
Earthquake	8620	6896	862	862
Floods	9698	7758	970	970

Table 2. Data splits for each disaster dataset.

4.2 Evaluation Setting

We use accuracy and macro-F1 as our evaluation metrics. Accuracy is used to measure the overall performance of different approaches on all testing data and macro-F1 is used because it takes class imbalance setting into account and measures the average performance of all classes. All methods are evaluated on the 5-shot setting in default: only 5 labeled data are randomly sampled from the training set and used, and the rest of the training data are treated as unlabeled data. We report accuracy and macro-F1 averaged across three runs with the same three random seeds for all methods.

4.3 Experimental Setup

To test the effectiveness of the proposed semi-supervised learning methods and components discussed in Section 3, we perform experiments for the following methods: 1) Supervised Baseline: supervised baseline that uses

the pre-trained BERT-base-uncased model with one classification layer added and *fine-tuned only on labeled data* for our classification task. For a fair comparison, all methods used the same model architecture with this supervised baseline. 2) PSL: plain pseudo-labeling that uses not only labeled data but also unlabeled data with high-confidence predictions and their hard pseudo-labels for training; 3) PSL++: add data augmentations and consistency regularization to PSL; 4) TextMixUp: apply TextMixUp introduced in Section 3 to the supervised baseline; 5) CrisisMatch: our proposed algorithm, as described in Section 3. Note that Supervised Baseline and TextMixUp use only labeled data (e.g., 5 labeled data per class in default) and other approaches utilize both labeled data and unlabeled data.

All methods use the same model architecture BERT-base-uncased model and hyper-parameters in default. We set batch size as 32, learning rate as 2e-5, and use AdamW as optimizer. The maximum sequence length is 64. λ in TextMixUp are sampled from the Beta distribution with sampling hyper-parameter $\alpha = 0.75$. The weight of the unlabeled loss is searched among $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ and set to 10 in default. We adopt linear ramps-up strategy for the unlabeled loss weight and set ramps-up length as 1000 iterations. Confidence threshold τ is set to 0.75 and sharpening temperature T is set to 0.5. We use synonym replacement and random swap as augmentation methods and set K, the number of augmentations for unlabeled samples, to 2.

4.4 Main Results

Datasets	**Wildfires**		**Earthquake**	
Methods	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**
Supervised Baseline	53.9 ± 1.3	44.2 ± 1.8	51.1 ± 2.0	41.4 ± 1.9
PSL	59.3 ± 0.5	42.2 ± 2.4	57.5 ± 4.9	39.0 ± 3.4
PSL++	60.1 ± 5.3	46.8 ± 2.5	58.2 ± 1.7	41.8 ± 3.6
TextMixUp	56.0 ± 1.7	46.2 ± 1.6	58.1 ± 5.5	46.5 ± 5.9
CrisisMatch	**63.4 ± 1.1**	**51.5 ± 0.9**	**63.0 ± 6.1**	**51.3 ± 5.2**

Table 3. Main results: comparison of all methods on Earthquake and Wildfires datasets. 5 labeled data per class are used for all methods. Methods other than Supervised Baseline and TextMixUp leverages unlabeled data from the rest of training set. Results are averaged across three different runs.

In Table 3, we summarize and compare the results of all methods in Earthquake and Wildfires datasets. We provide detailed analyses and discuss our findings.

Leveraging unlabeled data achieves significantly better accuracy performance with limited labeled data. As shown in Table 3, methods that leveraged unlabeled data, including PSL, PSL++ and CrisisMatch, significantly outperform the supervised baseline, by at least 5.4% accuracy on the Wildfires dataset and 6.4% accuracy on Earthquake dataset. In addition, CrisisMatch surpasses the supervised baselines by 9.5% and 12.9% accuracy on Earthquake and Wildfires datasets. This demonstrates that unlabeled data can be effectively used to boost performance when limited labeled data is available and thus can alleviate reliance on labeled data.

Data augmentation and consistency regularization improve the performance of pseudo-labeling. We compare PSL with PSL++. PSL++ boosts the performance of PSL by 0.8% accuracy and 4.6% macro F1 on Wildfires dataset and 0.7% accuracy and 2.8% macro F1 on Earthquake dataset, justifying the effectiveness of data augmentation and consistency regularization on regularizing the model.

TextMixUp boosts the performance of the supervised baseline. Compared with the supervised baseline, TextMixUP increases 2.1% accuracy, 2% macro F1 on the Wildfire dataset, and 7% accuracy, 5.1% F1 on the Wildfires dataset. This attests that artificially generated mixed samples by TextMixUp alleviate overfitting and regularize the model to perform linearly between data.

CrisisMatch achieves the best performance for fine-grained crisis tweet classification task. The proposed CrisisMatch reaches the best performance of 63.4% accuracy and 51.5% F1 on Wildfires datasets, which obtains +9.5% accuracy and +7.3% F1 huge boosts than the supervised baseline. On the Earthquake dataset, CrisisMatch surpasses the supervised baseline by a large margin of 12.9% accuracy and 9.9% F1. Besides, CrisisMatch consistently performs better than TextMixMatch. These results demonstrate that our proposed CrisisMatch effectively incorporates different ideas and components to handle the fine-grained crisis tweet classification with limited unlabeled data.

4.5 Other Analysis

Influence of number of labeled data

We evaluate our baseline and proposed method using accuracy and macro-F1 with a varying number of labeled data from 1 to 50. As shown in Table 4 and Figure 1, CrisisMatch consistently obtains better accuracy than the supervised baseline in different settings, and also improves macro-F1 in most settings. As the number of labeled data increases, both the CrisisMatch and the supervised baseline achieve better performance in accuracy and macro-F1. For instance, CrisisMatch increases accuracy from 40.7% to 72.5% when the given number of labeled data per class is improved from 1 to 50. In general, the gap between CrisisMatch and the supervised baseline shrinks and both methods become more robust when more labeled data is provided, which we believe is because the supervision signal becomes sufficient.

	Dataset: Wildfires					
Accuracy	1	3	5	10	20	50
Supervised Baseline	36.8 ± 5.0	43.9 ± 4.5	53.9 ± 1.3	59.5 ± 2.5	64.1 ± 1.1	70.9 ± 1.0
CrisisMatch	**40.7 ± 5.3**	**52.5 ± 6.9**	**63.4 ± 1.1**	**66.4 ± 1.2**	**68.3 ± 1.8**	**72.5 ± 0.3**
Macro-F1	1	3	5	10	20	50
Supervised Baseline	28.3 ± 2.9	35.5 ± 3.7	44.2 ± 1.8	50.2 ± 2.3	54.1 ± 1.3	61.4 ± 1.2
CrisisMatch	**31.4 ± 3.1**	**39.3 ± 2.9**	**51.5 ± 0.9**	**52.9 ± 2.2**	**56.2 ± 1.5**	**62.1 ± 1.7**
	Dataset: Earthquake					
Accuracy	1	3	5	10	20	50
Supervised Baseline	34.9 ± 3.0	45.0 ± 4.0	51.1 ± 2.0	63.3 ± 3.5	70.7 ± 1.5	77.5 ± 0.4
CrisisMatch	**43.5 ± 4.4**	**49.4 ± 4.0**	**63.0 ± 6.1**	**74.7 ± 1.2**	**77.8 ± 0.9**	**78.9 ± 0.4**
Macro-F1	1	3	5	10	20	50
Supervised Baseline	25.5 ± 3.7	36.5 ± 2.6	41.4 ± 1.9	51.9 ± 2.3	59.2 ± 1.5	**65.7 ± 0.4**
CrisisMatch	**25.8 ± 5.0**	**38.4 ± 3.2**	**51.3 ± 5.2**	**60.8 ± 0.7**	**63.1 ± 0.7**	65.5 ± 0.1

Table 4. Influence of number of labeled data. Supervised baseline only uses labeled data while CrisisMatch utilizes unlabeled data from the rest of training set.

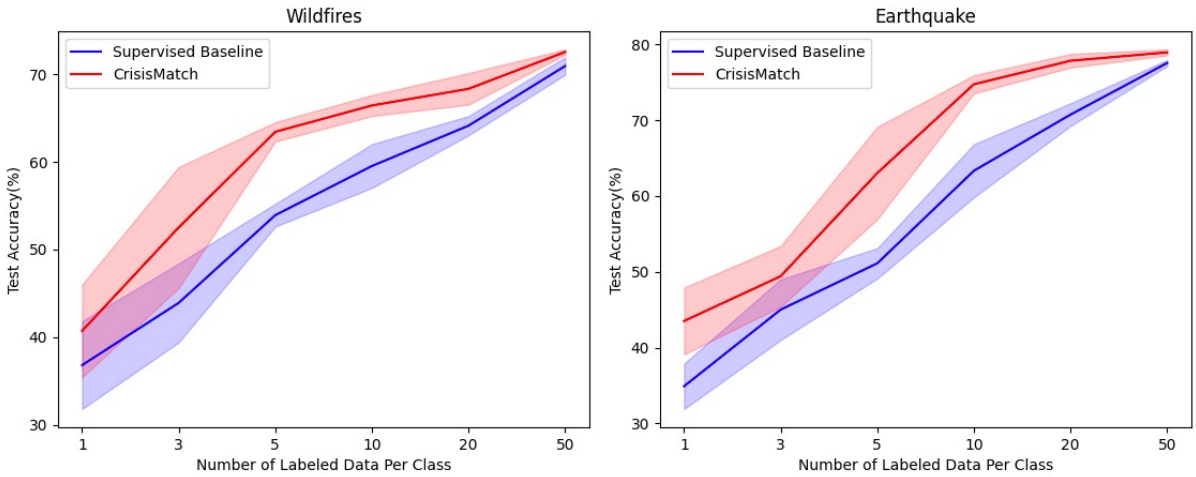

Figure 1. Performance with varying number of labeled data per class.

Out-of-Domain Results

	Source: Earthquake, Target: Floods					
Acc/Macro-F1	5		10		20	
Supervised Baseline	47.3 ± 7.1	31.2 ± 5.1	58.9 ± 6.4	39.8 ± 2.3	64.3 ± 3.6	45.9 ± 4.2
CrisisMatch	**52.9 ± 13.1**	**32.6 ± 7.0**	**66.9 ± 2.1**	**45.0 ± 0.5**	**67.9 ± 0.8**	**47.8 ± 1.0**
	Source: Wildfires, Target: Floods					
Acc/Macro-F1	5		10		20	
Supervised Baseline	**57.7 ± 3.4**	34.7 ± 1.5	61.0 ± 0.3	40.7 ± 3.3	65.9 ± 0.6	46.4 ± 2.0
CrisisMatch	56.4 ± 2.8	**37.6 ± 1.9**	**66.5 ± 3.7**	**44.7 ± 3.3**	**70.6 ± 0.1**	**49.3 ± 2.9**

Table 5. Out-of-domain results.

We investigate the performance of our model on data from out-of-domain (i.e., the distribution of testing data is different from the distribution of training data). Specifically, we use the Earthquake or Wildfires as training data, with each class having 5, 10, or 20 labeled data, and test the model on the Floods dataset. As shown in Table 5, our proposed CrisisMatch generally increases both accuracy and macro-F1 over the supervised baseline. For instance, given 10 labeled data per class on the Earthquake dataset, CrisisMatch outperforms the supervised baseline by 8% accuracy and 5.2% macro-F1 score. This observation confirms the robustness of our proposed CrisisMatch on out-of-domain data.

Entropy minimization: sharpening vs. hard pseudo-labeling.

We experimented with two different approaches to minimize entropy. CrisisMatch with hard pseudo-labeling empirically performs slightly better than sharpening with soft labels. For instance, compared to CrisisMatch with hard pseudo-labeling, CrisisMatch with sharpening decreases the performance by 2.8% macro-F1 on Wildfires and 3.5% accuracy on Earthquake dataset. These results show that hard pseudo-labeling is a more effective method in entropy minimization for our fine-grained crisis tweet classification task.

Datasets	Wildfires		Earthquake	
Methods	Accuracy	Macro-F1	Accuracy	Macro-F1
Sharpening	62.0 ± 7.5	48.5 ± 5.4	59.9 ± 3.1	47.5 ± 3.2
Hard pseudo-labeling	**63.0 ± 6.1**	**51.3 ± 5.2**	**63.4 ± 1.1**	**51.5 ± 0.9**

Table 6. Entropy minimization: sharpening vs. hard pseudo-labeling.

4.6 Ablation Study

Since CrisisMatch comprises various existing mechanisms, we conduct extensive ablation studies to show the effectiveness of each component.

Methods	**Accuracy**	**Macro-F1**
CrisisMatch	63.4 ± 1.1	51.5 ± 0.9
- Unlabeled Data	57.6 ± 3.6	47.4 ± 3.2
- Consistency Regularization	62.7 ± 3.0	49.3 ± 4.2
- TextMixUp	59.3 ± 2.1	48.6 ± 1.4
- All (Supervised Baseline)	53.9 ± 1.3	44.2 ± 1.8

Table 7. Ablation study on Wildfires dataset with 5 labeled data per class.

As illustrated in Table 7, we empirically study the performance of CrisisMatch after removing each of its components at a time. One may notice that the performance drops after stripping each part, demonstrating that all components contribute to CrisisMatch to achieve better results. Furthermore, removing unlabeled data reduces the performance most, suggesting that CrisisMatch can effectively leverage unlabeled data to train a better model. Removing TextMixUp results in the second-largest decrease in performance. This implies the effectiveness of TextMixUp

in providing data augmentation and regularizing the model. In addition, the decrease resulting from removing consistency regularization illustrates the importance of encouraging the model to perform consistently for different augmentations of data.

5 CONCLUSION AND FUTURE WORK

In this paper, we studied several methods on how to effectively leverage unlabeled data for disaster tweet classification in the semi-supervised few-shot setting, where there are only a few labeled data per class but large amounts of unlabeled data are available. Concretely, we introduced different variants of pseudo-labeling by introducing data augmentation, entropy minimization and consistency regularization. Besides, we studied TextMixUp and proposed our algorithm CrisisMatch, which further integrated TextMixUp into pseudo-labeling. Experimental results show that CrisisMatch can surpass the supervised baseline by a significant margin and demonstrate its effectiveness in utilizing unlabeled data. In the future, we plan to deal with the existing data-imbalanced issue and explore adaptive thresholds for different classes on a wider range of datasets and tasks.

REFERENCES

Alam, F., Joty, S., and Imran, M. (July 2018a). "Domain Adaptation with Adversarial Training and Graph Embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1077–1087.

Alam, F., Joty, S., and Imran, M. (2018b). "Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets". In: *Twelfth International AAAI conference on web and social media*.

Alam, F., Qazi, U., Imran, M., and Ofli, F. (2021). "HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks." In: *ICWSM*, pp. 933–942.

Bachman, P., Alsharif, O., and Precup, D. (2014). "Learning with pseudo-ensembles". In: *Advances in neural information processing systems* 27.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). "Mixmatch: A holistic approach to semi-supervised learning". In: *Advances in neural information processing systems* 32.

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). "Identifying Informative Messages in Disasters using Convolutional Neural Networks." In: *ISCRAM*.

Chen, J., Yang, Z., and Yang, D. (July 2020). "MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2147–2157.

Chowdhury, J. R., Caragea, C., and Caragea, D. (2020). "Cross-lingual disaster-related multi-label tweet classification with manifold mixup". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 292–298.

Fadaee, M., Bisazza, A., and Monz, C. (July 2017). "Data Augmentation for Low-Resource Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 567–573.

Grandvalet, Y. and Bengio, Y. (2004). "Semi-supervised learning by entropy minimization". In: *Advances in neural information processing systems* 17.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing social media messages in mass emergency: A survey". In: *ACM Computing Surveys (CSUR)* 47.4, pp. 1–38.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Practical extraction of disaster-relevant information from social media". In: *Proceedings of the 22nd international conference on world wide web*, pp. 1021–1024.

Imran, M., Ofli, F., Caragea, D., and Torralba, A. (2020). *Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions*.

Karisani, P. and Karisani, N. (2021). "Semi-supervised text classification via self-pretraining". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 40–48.

Kruspe, A., Kersten, J., and Klan, F. (2019). "Detecting event-related tweets by example using few-shot models". In: *16th International Conference on Information Systems for Crisis Response and Management*.

Kumar, A., Bhattamishra, S., Bhandari, M., and Talukdar, P. (2019). "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3609–3619.

Lee, D.-H. 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*. Vol. 3. 2, p. 896.

Li, H., Caragea, D., and Caragea, C. (2021). "Combining self-training with deep learning for disaster tweet classification". In: *The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021)*.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2018). "Disaster response aided by tweet classification with a domain adaptation approach". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 16–27.

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (2019). "A hybrid domain adaptation approach for identifying crisis-relevant tweets". In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 11.2, pp. 1–19.

McLachlan, G. J. (1975). "Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis". In: *Journal of the American Statistical Association* 70.350, pp. 365–369.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). "Virtual adversarial training: a regularization method for supervised and semi-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8, pp. 1979–1993.

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*.

Plotnick, L., Hiltz, S. R., Kushma, J. A., and Tapia, A. H. (2015). "Red Tape: Attitudes and Issues Related to Use of Social Media by US County-Level Emergency Managers." In: *ISCRAM*.

Reuter, C., Hughes, A. L., and Kaufhold, M.-A. (2018). "Social media in crisis management: An evaluation and analysis of crisis informatics research". In: *International Journal of Human–Computer Interaction* 34.4, pp. 280–294.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016). "Regularization with stochastic transformations and perturbations for deep semi-supervised learning". In: *Advances in neural information processing systems* 29.

Scudder, H. (1965). "Probability of error of some adaptive pattern-recognition machines". In: *IEEE Transactions on Information Theory* 11.3, pp. 363–371.

Sirbu, I., Sosea, T., Caragea, C., Caragea, D., and Rebedea, T. (2022). "Multimodal Semi-supervised Learning for Disaster Tweet Classification". In: *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2711–2723.

Sugiyama, A. and Yoshinaga, N. (2019). "Data augmentation using back-translation for context-aware neural machine translation". In: *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pp. 35–44.

Tarvainen, A. and Valpola, H. (2017). "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results". In: *Advances in neural information processing systems* 30.

Varga, I., Sano, M., Torisawa, K., Hashimoto, C., Ohtake, K., Kawai, T., Oh, J.-H., and De Saeger, S. (2013). "Aid is out there: Looking for help from tweets during a large scale disaster". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1619–1629.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. (2019). "Manifold mixup: Better representations by interpolating hidden states". In: *International Conference on Machine Learning*. PMLR, pp. 6438–6447.

Vieweg, S., Castillo, C., and Imran, M. (2014). "Integrating social media communications into the rapid assessment of sudden onset disasters". In: *International Conference on Social Informatics*. Springer, pp. 444–461.

Wei, J. and Zou, K. (2019). "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". In: *arXiv preprint arXiv:1901.11196*.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020). "Self-training with noisy student improves imagenet classification". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. (2021). "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling". In: *Advances in Neural Information Processing Systems* 34, pp. 18408–18419.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). "mixup: Beyond Empirical Risk Minimization". In: *International Conference on Learning Representations*.