LLM-guided Semi-Supervised Approaches for Social Media Crisis Data Classification

Jacob Ativo*
Department of Computer Science
California State University, East Bay
Hayward CA 94542, USA
jativo@horizon.csueastbay.edu

Bharaneeshwar Balasubramaniyam*
Department of Computer Science
Kansas State University
bharanibala@ksu.edu

Anh Tran*
Independent Researcher
anhtranst@gmail.com

Khushboo Gupta
Department of Computer Science
University of Illinois at Chicago
kgupta27@uic.edu

Hongmin Li
Department of Computer Science
California State University, East Bay
Hayward CA 94542, USA
hongmin.li@csueastbay.edu

Doina Caragea
Department of Computer Science
Kansas State University
dcaragea@ksu.edu

Cornelia Caragea
Department of Computer Science
University of Illinois at Chicago
cornelia@uic.edu

ABSTRACT

Semi-supervised learning approaches have been investigated as a means to enhance the analysis of social media data in disaster management contexts. In this work, we present the first empirical evaluation of large language model (LLM) guided semi-supervised learning for crisis related tweet classification. We compare two recent LLM assisted semi-supervised methods, VerifyMatch and LLM guided Co-Training (LG-CoTrain), against established semi-supervised baselines. Our results show that LG-CoTrain significantly outperforms classical semi-supervised approaches in low resource settings with 5, 10 and 25 labeled examples per class, achieving the highest averaged Macro F1 across events. VerifyMatch achieves competitive performance while also demonstrating strong calibration properties. As the number of labeled examples increases, the performance gap narrows and Self Training emerges as a strong baseline. We further observe that compact semi-supervised models can, in some cases, outperform very large LLMs operating in zero-shot settings. This finding highlights the potential of transferring knowledge from LLMs into smaller and more deployable models through LLM guided semi-supervised learning, offering a practical pathway for real world disaster response applications. Our project repository on Github is here.

Keywords

Semi-supervised learning, large language model, social media crisis data, model calibration, disaster response

*These authors contributed equally to this work.

INTRODUCTION

During emergency events, individuals increasingly turn to social media platforms such as X (formerly Twitter), Reddit, and Instagram to seek information and share updates. From a communication perspective, these platforms function bidirectionally: authorities disseminate critical disaster response information (e.g., warnings or evacuation orders) to the public, while the public also provides firsthand reports and situational updates that can be mined to enhance situational awareness (Reuter and Kaufhold 2018; Reuter 2022; Wolbers et al. 2021). Consequently, both researchers and practitioners recognize the substantial value of such user-generated content for crisis response. However, effectively integrating social media streams into real-time operations remains challenging due to information overload characterized by high volume, velocity, and varying levels of veracity (Purohit et al. 2025).

To address these challenges, extensive research over the past decade has focused on applying Machine Learning (ML) and Natural Language Processing (NLP) techniques to automatically classify social media data into actionable categories, such as infrastructure damage or requests for rescue. A wide range of ML models have been proposed for these classification tasks, including statistical learning approaches and supervised deep learning models (Starbird et al. 2010; Imran, Elbassuoni, et al. 2013; C. Caragea et al. 2014; Nguyen et al. 2016; Burel and Alani 2018; Kersten et al. 2019; Ghafarian and Yazdi 2020). However, supervised models typically require substantial amounts of high-quality human-labeled data to achieve strong performance, which is often scarce in the time-sensitive context of disaster response.

To mitigate this limitation, researchers have explored domain adaptation, transfer learning, and semi-supervised learning approaches. Domain adaptation methods leverage labeled data from previous disaster events to alleviate label scarcity in newly emerging events (Li, D. Caragea, C. Caragea, and Herndon 2018; Alam, Joty, et al. 2018a). In contrast, semi-supervised learning methods aim to train effective models by combining a small amount of labeled data with a large volume of unlabeled data through pseudo-labeling strategies. In a typical teacher–student semi-supervised learning framework, a teacher model trained on the limited labeled data first generates pseudo-labels for the unlabeled instances, and these pseudo-labeled examples are subsequently used to train a student model (Li, D. Caragea, and C. Caragea 2021; H. P. Zou et al. 2023; Gupta et al. 2025). In general, the performance of semi-supervised approaches depends heavily on the quality of the pseudo-labels. Therefore, a central research question in semi-supervised learning is how to effectively leverage unlabeled data to generate high-quality pseudo-labels that improve downstream model performance.

With the rapid advancement of Large Language Models (LLMs), recent studies have explored leveraging LLMs to improve pseudo-labeling of semi-supervised models built on smaller pre-trained language models such as BERT (Devlin et al. 2018), particularly for text classification tasks (Park and C. Caragea 2024; Rahman and C. Caragea 2025). In the context of social media crisis data analysis, there has been a surge of work employing LLMs in zero-shot (i.e., making predictions using the LLM without task-specific labeled examples), few-shot (i.e., provide the LLM a small number of labeled examples in the prompt), and fine-tuning (i.e., updating a small sized LLM model parameters on task-specific labeled data) to identify informative social media content for disaster management (Imran, Ziaullah, et al. 2025; Taghian Dinani et al. 2024; McDaniel et al. 2024; Yin et al. 2025; Shrestha 2025; Salfinger and Snidaro 2024; Lei et al. 2025; Guo et al. 2025). However, to the best of our knowledge, no prior work has investigated semi-supervised models guided by LLMs in the crisis domain.

To this end, we study two BERT-based semi-supervised models enhanced with LLM-generated pseudo-labels for social media crisis classification: (1) VerifyMatch (Park and C. Caragea 2024), originally proposed for Natural Language Inference (NLI), and (2) LLM-guided Co-training (LG-CoTrain) (Rahman and C. Caragea 2025), designed for general text classification. Following the experimental protocol of Gupta et al. (2025), which evaluates several semi-supervised methods on 10 disaster events from the HumAID dataset (Alam, Qazi, et al. 2021), we experiment with VerifyMatch and LG-CoTrain, enhanced with GPT-4o pseudo labels, on the same benchmark.

Specifically, we evaluate the performance of the VerifyMatch and LG-CoTrain approaches using the Macro-F1 score, as well as the Expected Calibration Error (ECE), and compare the results with those of the existing baselines from Gupta et al. (2025) to form a more comprehensive study for semi-supervised learning algorithms on social media crisis data classification. To summarize, our main contributions are as follows:

- We evaluate two semi-supervised approaches, VerifyMatch and LG-CoTrain, using zero-shot pseudo-labels generated by GPT-4o on 10 disaster events from the HumAID dataset, a benchmark of disaster-related tweets annotated with humanitarian categories such as damage, injured people, and requests or urgent needs. We further compare these models against all semi-supervised methods examined by Gupta et al. (2025).
- Our experimental results show that LG-CoTrain significantly outperforms other approaches in low-resource settings (e.g., 5 or 10 labeled examples per category). Moreover, it demonstrates good model calibration.

However, as the amount of labeled data increases (e.g., 50 labeled examples per category), the performance gap between LG-CoTrain and other semi-supervised models narrows, and Self-training emerges as a competitive baseline that is difficult to surpass.

- Semi-supervised models based on smaller pre-trained language models outperform zero-shot GPT-4o only on a subset of disaster events. This may be due to limited and potentially unrepresentative unlabeled data in the HumAID benchmark, including missing class examples in the sampled unlabeled sets—an issue that can also arise in real-world scenarios. Larger and more representative unlabeled datasets could help mitigate these limitations.

Still all these findings highlight the potential of transferring knowledge from LLMs into smaller and more deployable models through LLM guided semi-supervised learning, offering a practical pathway for real world disaster response applications.

RELATED WORK

There is a vast body of literature on semi-supervised learning (SSL) in machine learning. In this section, we first provide an overview of SSL approaches, and then review prior work applying SSL and large language models (LLMs) to social media disaster data analysis.

SSL Overview. A wide range of SSL approaches have been proposed for text classification, beginning with the original idea of Self-Training and pseudo-labeling (Scudder 1965). In self-training, a teacher model is first trained on limited labeled data and then used to generate pseudo-labels for unlabeled instances. These pseudo-labeled examples are subsequently incorporated into training a student model, often in an iterative manner. Two critical design choices in this framework are (1) how to select pseudo-labeled examples for inclusion in training and (2) whether to use hard labels (the most likely class per example) or soft labels (predicted class probabilities). Incorporating low-confidence pseudo-labels may lead to error propagation and degrade student model performance.

Various pseudo-label selection strategies have been proposed. For example, FixMatch (Sohn et al. 2020) and related self-training methods adopt fixed confidence thresholds to filter pseudo-labels, while Uncertainty-Aware Self-Training (UST) (Mukherjee and Awadallah 2020) employs more sophisticated uncertainty estimation techniques grounded in probability theory. However, threshold-based filtering may restrict the student model's access to potentially informative unlabeled data. To alleviate this limitation, methods such as MixMatch (Berthelot et al. 2019) and SoftMatch (Chen et al. 2023) were introduced. SoftMatch retains all pseudo-labeled samples but assigns lower weights to low-confidence instances during training, thereby balancing data quantity and label quality. MixMatch similarly leverages soft pseudo-labels and further incorporates MixUp (Zhang et al. 2017), which interpolates pseudo-labeled and human-annotated examples to generate smoother and potentially higher-quality training signals.

AUM-based Self-Training (AUM-ST) (Sosea and C. Caragea 2022) takes a different perspective by filtering low-quality pseudo-labeled examples through tracking training dynamics using the Area Under the Margin (AUM). Building upon this idea, AUM-based MixUp in Self-Training (AUM-ST-Mixup) (Gupta et al. 2025) integrates MixUp and additional confidence-tracking mechanisms on top of AUM-ST to further enhance pseudo-label reliability. Confidence-based Mixup in Self-Training (Conf-ST-Mixup) (Gupta et al. 2025) enhances pseudo-labeling by defining prediction confidence as the probability gap between the top two classes, where a larger gap indicates higher confidence, allowing the model to distinguish easy-to-learn (reliable) from hard-to-learn (ambiguous) samples. It applies mixup across labeled, high-confidence, and low-confidence pseudo-labeled data to regularize training, reduce error propagation, and promote smoother decision boundaries.

Despite these improvements, methods relying on a single model for pseudo-label generation remain vulnerable to reinforcing incorrect high-confidence predictions, particularly in early training stages (Rahman and C. Caragea 2025). To mitigate this issue, VerifyMatch (Park and C. Caragea 2024) incorporates LLM-generated pseudo-labels alongside a verifier model, enabling cross-validation of pseudo-label quality. Combined with MixUp, VerifyMatch achieves competitive performance in low-resource settings. Finally, LLM-guided Co-Training (LG-CoTrain) (Rahman and C. Caragea 2025) integrates LLM-generated pseudo-labels within a dual-model co-training framework, where two models iteratively learn from each other while incorporating LLM guidance. Unlike MixUp-based approaches, LG-CoTrain retains all pseudo-labeled data without modification. LG-CoTrain outperforms the zero-shot Phi-3 and other SSL approaches and achieves state-of-the-art performance on four out of five text classification benchmark datasets.

SSL for Social Media Crisis Data Analysis. Several studies have applied SSL techniques to social media crisis data analysis. For example, Alam, Joty, et al. (2018b) proposed a graph-based semi-supervised CNN model for

Twitter data from two disaster events. Li, D. Caragea, and C. Caragea (2021) applied self-training with BERT and CNN models to the CrisisLexT6 and CrisisLexT26 datasets (Olteanu, Castillo, et al. 2014; Olteanu, Vieweg, et al. 2015), which contain tweets from various disaster events. CrisisLexT6 is annotated for whether tweets are related to the disaster, while CrisisLexT26 includes coarse- and fine-grained humanitarian labels similar to those in the HumAID dataset. Sirbu et al. (2022) extended FixMatch by incorporating soft-labeling for multimodal disaster tweet classification (text and images) on the CrisisMMD dataset (Alam, Ofli, et al. 2018). H. P. Zou et al. (2023) proposes CrisisMatch which differs from MixMatch by using hard pseudo-labeling for entropy maximization instead of sharpening for text classification on the HumAID dataset. Meanwhile, H. Zou et al. (2023) proposes a novel approach by using memory bank — DeCrisisMB (H. Zou et al. 2023) — to addresses the bias in SSL which assigns disproportionate pseudo-labels for more occurring instances in highly imbalanced datasets such as crisis-related tweet classification.

More recently, Gupta et al. (2025) proposed Confidence-based and AUM-based MixUp with Self-Training (AUM-ST-MixUp) and conducted a systematic evaluation across 10 disaster events in the HumAID dataset, comparing these methods with several SSL baselines discussed above. In addition, Gupta et al. (2025) employed Expected Calibration Error (ECE) to measure model calibration. ECE quantifies the extent to which predicted probabilities align with observed outcomes. A well-calibrated model produces confidence estimates that align with empirical correctness rates; for example, predictions assigned 85% confidence should be correct approximately 85% of the time. Calibration is particularly important for interpreting model outputs and supporting reliable decision-making in high-stakes contexts such as disaster response.

LLMs for Social Media Crisis Data Analysis. With the rapid development of LLMs, increasing attention has been devoted to applying them to social media data analysis (Lei et al. 2025; Sánchez et al. 2025). For instance, Guo et al. (2025) compared fine-tuned Llama 3.2 11B models with zero-shot GPT-4o and prior approaches, demonstrating that fine-tuned Llama models achieve state-of-the-art performance on multimodal crisis classification using the CrisisMMD dataset. Most closely related to our work is Imran, Ziaullah, et al. (2025), which systematically evaluated the robustness of several LLMs—including GPT-3.5, GPT-4, GPT-4o, Llama-2 13B, Llama-3 8B, and Mistral 7B—on the HumAID dataset under zero-shot and few-shot settings. Their results indicate that few-shot prompting does not consistently improve performance, and GPT-4 achieved the strongest overall results among the evaluated models.

In this work, we compare LLM-guided SSL approaches against the SSL methods evaluated in Gupta et al. (2025) with respect to model prediction accuracy, as well as calibration error (ECE metric).

DATASET

We use the same 10 disaster events from the HumAID dataset as our benchmark following Gupta et al. (2025), as shown in Table 1. HumAID is a human-annotated Twitter dataset comprising 77,196 tweets from 19 disaster events, categorized into 11 classes. Excluding the "Don't know or can't judge" category, the remaining 10 primary humanitarian categories are: *1. Caution and advice; 2. Sympathy and support; 3. Requests or urgent needs; 4. Displaced people and evacuations; 5. Injured or dead people; 6. Missing or found people; 7. Infrastructure and utility damage; 8. Rescue, volunteering, or donation effort; 9. Other relevant; 10. Not humanitarian*.

Table 1 and 2 also reports the statistics of the training, validation, and test splits, number of classes for each event as well as class distribution. For detailed class distributions within each event, we refer readers to Gupta et al. (2025). Consistent with their experimental setup, we adopt the same data split configuration, in which the training set is further divided into labeled and unlabeled subsets, as described in the Experimental Setup section.

METHODS

We study semi-supervised learning (SSL) for crisis-related tweet classification under limited labeled data. All methods operate on a shared experimental setting consisting of a small labeled subset $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{n_L}$ and a larger unlabeled subset $\mathcal{D}_U = \{x_j\}_{j=1}^{n_U}$ drawn from the same disaster event. Our goal is to learn a classifier $f_\theta(x)$ that generalizes well to held-out test data while maintaining calibrated confidence estimates. Unless otherwise specified, all neural models use BERTweet as the encoder backbone, due to its overall good supervised performance (Gupta et al. 2025), followed by a task-specific classification head.

Supervised Baselines

We include two supervised baselines to contextualize SSL performance.

- **Limited-Label Supervision - Lower Bound.** A BERTweet classifier trained solely on $\mathcal{D}_L$ serves as a lower-bound reference, representing performance achievable without unlabeled data.

Disaster Event/Data Split	C	Train	Val	Test	5 lb/cl		10 lb/cl		25 lb/cl		50 lb/cl	
					L	*U*	*L*	*U*	*L*	*U*	*L*	*U*
California Wildfires 2018	10	5163	752	1461	50	5113	100	5063	250	4913	500	4663
Canada Wildfires 2016	8	1569	228	445	40	1529	80	1489	189	1380	364	1205
Cyclone Idai 2019	10	2753	401	779	50	2703	100	2653	238	2515	453	2300
Hurricane Dorian 2019	9	5329	776	1508	45	5284	90	5239	225	5104	442	4887
Hurricane Florence 2018	9	4384	639	1241	45	4339	90	4294	225	4159	438	3946
Hurricane Harvey 2017	9	6378	929	1805	45	6333	90	6288	225	6153	450	5928
Hurricane Irma 2017	9	6579	954	1862	45	6534	90	6489	225	6354	450	6129
Hurricane Maria 2017	9	5094	742	1442	45	5049	90	5004	225	4869	450	4644
Kaikoura Earthquake 2016	9	1536	224	435	45	1491	90	1446	217	1319	417	1119
Kerala Floods 2018	9	5588	814	1582	45	5543	90	5498	225	5363	439	5149

Table 1. Data distribution and splits under different labels-per-class (lb/cl) settings for each of the 10 events in the HumAID dataset, where *C* stands for the number of classes, *L* stands for the number of labeled instances and *U* stands for the number of unlabeled instances.

Disaster Event	**Classes**	**Class Distribution**									
		1	2	3	4	5	6	7	8	9	10
California Wildfires 2018	10	97	330	55	258	1362	125	295	991	727	923
Canada Wildfires 2016	8	74	113	14	266	0	0	176	653	218	55
Cyclone Idai 2019	10	62	338	100	40	303	13	248	1308	285	56
Hurricane Dorian 2019	9	958	758	125	561	42	0	571	691	1011	612
Hurricane Florence 2018	9	917	330	38	446	208	0	224	1034	445	742
Hurricane Harvey 2017	9	379	444	233	482	488	0	852	1976	1237	287
Hurricane Irma 2017	9	429	397	88	528	626	0	1317	1113	1651	430
Hurricane Maria 2017	9	154	470	498	92	211	0	999	1384	1097	189
Kaikoura Earthquake 2016	9	345	302	17	61	73	0	218	145	218	157
Kerala Floods 2018	9	97	585	413	39	254	0	207	3005	669	319

Table 2. Disaster events and the corresponding number of classes, number of tweets in train split per event with the class distribution

- **Full-Supervision - Upper Bound.** A BERTweet model trained on the complete labeled training split (including $\mathcal{D}_L$ and $\mathcal{D}_U$ subsets) provides an approximate upper bound for in-domain performance (Table 3 BERTweet All).

Zero-Shot LLM Baseline

Zero-shot setting: To contextualize LLM-guided SSL performance, we evaluate GPT-4o in a zero-shot classification setting, where the LLM directly predicts class labels for test instances without any labeled training example and no task-specific fine-tuning. This baseline measures the standalone capability of LLMs relative to compact supervised and semi-supervised models. Specifically, we experimented with GPT-4.1, GPT-4o, GPT-4o mini, and GPT-5.1, and among them, GPT-4o consistently achieved better performance on both training and test splits. Therefore, we use GPT-4o to generate pseudo-labels for the entire unlabeled training set. We attempted to reproduce the best overall result obtained with zero-shot GPT-4 by Imran, Ziaullah, et al. (2025), however, due to API deprecation, exact replication was not possible. Nevertheless, our zero-shot GPT-4o results are comparable (slightly better) on average to the zero-shot GPT-4o reported by Imran, Ziaullah, et al. (2025) as shown in Table 3. While GPT-4o-mini produced competitive results, it generated a number of out-of-source (OOS) predictions.

Prompt engineering: We evaluate three prompts with GPT-4o on the validation splits of all 10 events, ranging from simple category definitions from the HumAID dataset to more detailed descriptions. As performance remains nearly identical across prompts (Macro-F1 range: 0.601–0.613) with the simplest version being the best, we adopt the simplest version. The detailed prompt

Classical Semi-Supervised Learning

We also used representative SSL approaches previously evaluated on the HumAID dataset (Gupta et al. 2025).

- **Self-Training (ST).** A teacher model trained on $\mathcal{D}_L$ generates pseudo-labels $\hat{y}_j$ for unlabeled examples in $\mathcal{D}_U$. High-confidence pseudo-labeled instances are iteratively incorporated into training.

- **Uncertainty-Aware Self-Training (UST).** UST refines pseudo-label selection by incorporating uncertainty estimation, reducing the influence of noisy high-confidence predictions.
- **MixMatch.** MixMatch integrates soft pseudo-labeling with MixUp interpolation between labeled and unlabeled examples, encouraging smoother decision boundaries.
- **AUM-based Self-Training (AUM-ST) and AUM-ST-MixUp.** AUM-ST tracks training dynamics via Area Under the Margin (AUM) to filter unreliable pseudo-labels and AUM-ST-MixUp combines this filtering with MixUp-based regularization.
- **Confidence-based MixUp in Self-Training (Conf-ST-MixUp).** Conf-ST-MixUp enhances pseudo-labeling by defining prediction confidence as the probability gap between the top two classes, where a larger gap indicates a more reliable pseudo-label. It then applies MixUp across labeled, high-confidence, and low-confidence pseudo-labeled data to regularize training and reduce error propagation.

LLM-Guided Semi-Supervised Learning

We investigate two SSL frameworks that incorporate zero-shot pseudo-labels generated by an LLM, in our case, GPT-4o. These pseudo-labels are used to augment or guide the training of smaller, task-specific models.

- **VerifyMatch.** VerifyMatch integrates LLM-generated pseudo-labels with a verifier model that cross-validates predictions before incorporating them into training. Combined with confidence-aware MixUp, this mechanism aims to reduce confirmation bias and mitigate overconfident errors, two common issues in pseudo-labeling. Confirmation bias refers to the tendency of a model to reinforce its own incorrect pseudo-labels during self-training, for example, repeatedly assigning the same wrong label to similar inputs. Overconfident errors refer to incorrect predictions made with high confidence, for example, assigning a wrong label with near-certain probability.
- **LLM-Guided Co-Training (LG-CoTrain).** LG-CoTrain employs a dual-model co-training architecture in which two classifiers iteratively exchange pseudo-labels while incorporating LLM guidance. Unlike MixUp-based methods, LG-CoTrain retains all pseudo-labeled examples and relies on cross-model agreement to stabilize learning. This framework is particularly effective in extremely low-resource settings, where model-generated pseudo-labels alone may be unreliable.

Overall, the evaluated methods differ along three dimensions: (1) pseudo-label generation source (model-based vs LLM-based), (2) pseudo-label filtering strategy (confidence thresholding, uncertainty-aware weighting, verification, or co-training), and (3) representation regularization (none, MixUp, or cross-model consistency). This structured comparison enables analysis of both predictive performance and calibration behavior under label scarcity.

EXPERIMENTAL SETUP

We run experiments with all the approaches described in the Methods section. For each event, we use the same train/validation/test splits as Gupta et al. (2025). We simulate low-resource settings by selecting a fixed number of labeled examples per class (lb/cl) from the training split. We evaluate four label budgets: 5, 10, 25, and 50 lb/cl. The remaining training instances are treated as unlabeled and are used by SSL methods according to their respective learning objectives. This split configuration is kept consistent across all methods to ensure a fair comparison.

We use the same metrics as in Gupta et al. (2025), specifically, Macro-F1 and the Expected Calibration Error (ECE) averaged across the 10 disaster events, to evaluate all methods. Macro-F1 captures balanced performance across classes, while ECE quantifies how well predicted probabilities align with empirical correctness.

Hyperparameter tuning was performed using Weights & Biases and the Optuna package over learning rate, batch size, number of epochs, and additional some main model-specific parameters. Detailed configurations will be released in the project repository. For Weights & Biases, we employed Bayesian sweeps to automate the search; however, we observed occasional optimization instability. Prior work (Liu and Wang 2021) shows that, in low-resource transformer fine-tuning, automated hyperparameter optimization may fail to outperform simple grid-search under limited search budgets due to overfitting and instability. Moreover, repeated tuning on a small validation set can lead to meta-overfitting, where configurations that perform well on the development set do not generalize to test performance or calibration. This may also explain small discrepancies between our results and those reported in Gupta et al. (2025).

RESULTS AND DISCUSSION

We show our experimental results in Table 3 and Table 4, as well as Figure 1. Table 3 reports the zero-shot performance of GPT-4o, along with the performance of the fully supervised BERTweet upper bound. Table 4 summarizes the Macro-F1 and ECE scores averaged across the 10 disaster events under varying label budgets.

Model	Zero-shot GPT-4o (Imran, Ziaullah, et al. 2025)	Zero-shot GPT-4o train	Zero-shot GPT-4o test	BERTweet - All
F1 ↑	0.612	0.628	0.641	0.678
ECE ↓	-	-	-	0.110

Table 3. Performance results for Zero-shot GPT-4o and supervised BERTweet trained on the whole training split

Method/Metric	F1 ↑				ECE ↓			
# Label	5	10	25	50	5	10	25	50
BERTweet	0.423	0.517	0.563	0.606	0.206	0.222	0.247	0.256
ST	0.448	0.548	0.625	**0.655**	0.305	0.231	0.184	0.165
UST	0.465	0.546	0.609	0.641	0.342	0.271	0.225	0.191
MixMatch	0.459	0.553	0.624	0.647	0.374	0.297	0.246	0.228
AUM-ST	0.424	0.505	0.572	0.595	0.264	0.206	0.204	0.194
Conf-ST-MixUp	0.421	0.533	0.623	0.643	0.408	0.321	0.244	0.246
AUM-ST-MixUp	0.476	0.532	0.611	0.639	0.190	**0.069**	**0.057**	**0.064**
VerifyMatch	0.463	0.549	0.616	0.644	**0.127**	0.086	0.083	0.100
LG-CoTrain	**0.608**	**0.619**	**0.631**	0.645	0.174	0.160	0.122	0.108

Table 4. BERTweet and SSL performance on the 10 HumAID disaster events. The results are reported in terms of Macro-F1 and ECE values averaged over the 10 events in the dataset. The best result for each setup is highlighted in bold.

Model Comparisons

Zero-Shot GPT-4o vs. BERTweet-all (upper bound). Table 3 shows that GPT-4o achieves an averaged Macro-F1 of 0.628 on the training split and 0.641 on the test split across the 10 disaster events, slightly outperforming the GPT-4o results reported by Imran, Ziaullah, et al. (2025). These results provide an overall indication of the pseudo-label quality generated by GPT-4o. Analyzing performance by category over the combined training splits of all 10 events shows that GPT-4o performs well on clear and concrete categories, achieving Macro-F1 scores above 0.6. For instance, it reaches 0.885 on *Injured or dead people* and 0.827 on *Rescue, volunteering, or donation effort*. However, GPT-4o struggles on three categories (F1 < 0.6), including *Other relevant information* (0.276), *Requests or urgent needs* (0.526), and *Not humanitarian* (0.569), which are broader or less precise. For the remaining categories, the performance is 0.634 for *Caution and advice*, 0.739 for *Sympathy and support*, 0.766 for *Displaced people and evacuations*, 0.698 for *Missing or found people*, and 0.704 for *Infrastructure and utility damage*. A detailed breakdown is provided in the project repository. Overall, categories with more accurate pseudo-labels are expected to benefit more from LLM-guided SSL approaches such as LG-CoTrain.

Despite its relative good zero-shot performance, GPT-4o remains below the fully supervised BERTweet model trained on the complete training set, which achieves a Macro-F1 of 0.678 and serves as an approximate upper bound. This comparison suggests that while zero-shot LLMs offer competitive performance without task-specific training, further gains can be achieved through supervised or semi-supervised adaptation.

VerifyMatch vs. classical SSL approaches. Table 4 shows that VerifyMatch yields competitive performance relative to the classical SSL baselines but lower ECE, especially at low 25–50 lb/cl budgets, indicating that combining LLM pseudo-labels with an explicit verification mechanism can improve robustness to noisy pseudo-labels. Compared to threshold-based or training-dynamics filtering strategies, VerifyMatch provides a more conservative but stable learning signal.

LG-CoTrain vs. other SSL approaches. Table 4 also shows that under the most challenging conditions, LG-CoTrain performs best among all SSL approaches: 0.608 Macro-F1 at 5 lb/cl and 0.619 at 10 lb/cl. This is a substantial improvement over classical SSL baselines such as ST, UST, and MixMatch, and also surpasses the best-performing baseline from Gupta et al. (2025) (Conf-ST-MixUp, AUM-ST-MixUp) and VerifyMatch in terms of Macro-F1 at both 5 and 10 lb/cl. These results suggest that incorporating LLM pseudo-labels within a co-training framework is particularly effective when the labeled set is too small for reliable self-training.

At higher label budgets, the performance gap between LG-CoTrain and other SSL baselines narrows, but with the average Macro F1 of 0.631 at 25 lb/cl, LG-CoTrain still outperforms the other SSL approaches. Notably, Self-Training becomes a strong baseline and achieves the best Macro-F1 value of 0.655 at 50 lb/cl. In this regime,

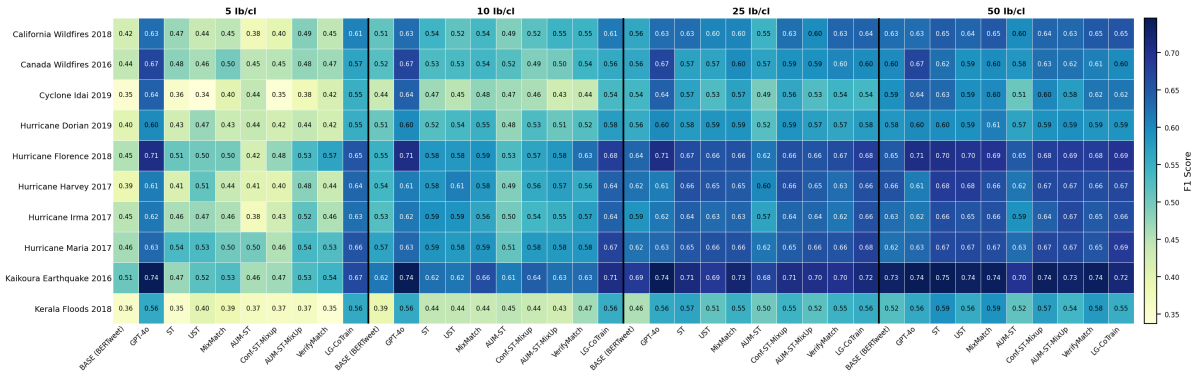

Figure 1. Per-event Macro-F1 scores for all methods across the 10 HumAID disaster events under different label budget settings. Values are rounded for readability.

the advantage of LG-CoTrain diminishes, suggesting that once enough labeled data are available, the marginal benefit of LLM-guided pseudo-labeling is smaller. This may be due to the fact that the overall F1 obtained with GPT-4o is around 0.630-0.640. Furthermore, this can also be attributed to the limited amount of unlabeled data in the HumAID benchmark, and the fact that it may not be fully representative of minority classes, an issue that can also be encountered in real-world scenarios. Larger and more representative unlabeled datasets could help mitigate these limitations.

GPT-4o vs. LLM-guided SSL approaches. In low-resource settings, LG-CoTrain provides consistent gains over classical SSL methods on many events, which explains its strong averaged performance. While zero-shot GPT-4o remains highly competitive across events, LG-CoTrain surpasses its performance in several cases. In particular, for the Hurricane Harvey, Irma, and Maria events under the 10, 25, and 50 labeled examples per class settings, LG-CoTrain achieves higher Macro-F1 scores than GPT-4o. Even under the most constrained setting of 5 labeled examples per class, LG-CoTrain still outperforms GPT-4o on Hurricane Harvey and Hurricane Maria. This observation reinforces an important takeaway: when deployment constraints allow direct LLM inference, zero-shot LLMs can serve as strong baselines; however, when practitioners require a smaller deployable model, LLM-guided SSL can distill and transfer some of the LLM's strengths into a compact classifier that can be executed effieciently and repeatedly at scale.

The amortized training paradigm of LLM-guided SSL approach such as LG Co-Train is particularly advantageous in crisis-response settings, where streaming data volumes are high and rapid, low-latency and accurate decisions are required. Thus, the value of LLM-guided SSL lies not merely in marginal performance gains, but in enabling scalable, controllable, and cost-efficient deployment.

SSL approaches vs. BERTweet (lower bound). Also worth noting, all SSL approaches show better performance than the BERTweet trained only on the limited amount of labeled data, suggesting that SSL approaches represent a good option for classifying social media crisis data in a low-data regime.

Calibration Behavior. Beyond classification accuracy, calibration is critical in crisis-response applications, where confidence scores may influence downstream triage and operational decisions. Table 4 shows that AUM-ST-MixUp achieves the lowest ECE at 10, 25 and 50 lb/cl, reflecting strong calibration under severe label scarcity. VerifyMatch also yields consistently low ECE values across all budgets, indicating that its verification mechanism can help control overconfident errors. LG-CoTrain prioritizes Macro-F1 improvements in the lowest-resource settings, while its calibration improves as more labeled data are available (ECE decreases from 0.174 at 5 lb/cl to 0.108 at 50 lb/cl).

Overall, the results highlight a practical trade-off: the most accurate method under extreme label scarcity is not always the best calibrated, and practitioners may need to balance performance and confidence reliability depending on operational needs.

Per-event Analysis

Figure 1 presents per-event Macro-F1 scores for all methods across label budgets, providing additional insight beyond the averaged results. Performance varies substantially across disasters, reflecting differences in event characteristics and class distributions (shown in Table 2).

Similar to Imran, Ziaullah, et al. (2025), who find that LLMs struggle with flood events, we observe that both zero-shot GPT-4o and the SSL models perform relatively worse on the Kerala Floods 2018 event compared to others. We hypothesize that this variation may be attributable not only to the disaster type, but also largely to two factors: class imbalance and pseudo-label quality. For events where a single class dominates the training set, such as Kerala Floods (53.8% of training tweets belong to *rescue_volunteering_or_donation_effort*) and Cyclone Idai (largest class 47.5%), the models consistently underperform because rare classes have too few examples for reliable classification, and Macro-F1 penalizes poor recall on any single class equally. Conversely, for more balanced events such as Kaikoura Earthquake (largest class 22.5%) and Hurricane Florence (largest class 23.6%), the models achieve better scores overall as compared to the scores for other events. GPT-4o pseudo-label quality also plays a role: events where GPT-4o achieves higher zero-shot F1 tend to produce better LG-CoTrain results, since higher-quality pseudo-labels provide a stronger training signal for co-training. The number of active classes also plays a role: events with only 8 classes (Canada Wildfires) or with extremely rare classes, such as Cyclone Idai where *missing_or_found_people* has only 13 training tweets, create near-zero F1 on those classes, which leads to a disproportionately reduced macro average.

Ablation Study

To quantify the contribution of the LLM in the LLM-guided SSL approaches, we compare LG-CoTrain with GPT-4o pseudo-labels against the same co-training approach without LLM pseudo-labels, a variant called Self-guided CoTrain (SG-CoTrain). SG- CoTrain replaces GPT-4o pseudo-labels with pseudo-labels from the BERTweet teacher model trained on the small labeled set. We run (SG-CoTrain) with 5 labeled examples per class across 10 events with 3 runs for each event. Table 5 reports per-event Macro-F1 on the test set for these two experiments.

Event	LG-CoTrain	SG-CoTrain	Delta
California Wildfires 2018	0.608	0.408	0.200
Canada Wildfires 2016	0.568	0.355	0.213
Cyclone Idai 2019	0.552	0.314	0.237
Hurricane Dorian 2019	0.554	0.409	0.145
Hurricane Florence 2018	0.655	0.319	0.336
Hurricane Harvey 2017	0.636	0.400	0.236
Hurricane Irma 2017	0.626	0.422	0.204
Hurricane Maria 2017	0.655	0.456	0.199
Kaikoura Earthquake 2016	0.669	0.501	0.168
Kerala Floods 2018	0.560	0.378	0.182
Average	**0.608**	**0.396**	**0.212**

Table 5. Ablation study for co-training with 5 labels for class, with LLM pseudo-labels (LG-CoTrain) and without LLM pseudo-labels (SG-CoTrain). The Delta improvement obtained from the LLM pseudo-labels is also shown.

LG-CoTrain consistently outperforms SG-CoTrain across all events, with gains ranging from 0.145 to 0.336 in F1 (average improvement of 0.212). Since both models share the same co-training pipeline and hyperparameter search space, this performance gap can be attributed primarily to differences in pseudo-label quality. This result is expected: a BERTweet teacher trained on only 50 labeled examples (5 lb/cl) produces pseudo-labels that are too noisy for effective co-training and even degrade performance through error propagation. In contrast, GPT-4o's zero-shot predictions are sufficiently accurate to provide a strong initial signal, enabling more effective learning.

To further examine SG-CoTrain, we also run an simple experiment using fixed hyper-parameter for all events with filtering pseudo-labeled data from the BERTweet teacher by retaining only the 50 most confident predictions per class. That preliminary results show that applying confidence filtering improves performance, despite using fewer pseudo-labeled samples. And we also observe the gap between LG-CoTrain and non-LLM variants tends to narrow as the amount of labeled data increases, which suggests that with sufficient labeled data, confidence-filtered self-guidance can partially substitute for LLM guidance. But we will leave the verification of this hypothesis and more comprehensive analysis to future work.

Deployment Considerations

In our experiments, GPT-4o zero-shot classification of the evaluation set (6,463 tweets across 10 events) cost $17.83 (via the OpenAI Batch API). The unlabeled training pool totals 44,373 tweets (approximately 4,400 per event on average), so we estimate pseudo-labeling cost at roughly $12 per event at this scale. Including hyperparameter tuning and model training on a dataset of this size, the full pipeline for a single new disaster event requires under one hour on an 8-GPU NVIDIA H100 cluster, cost of which could start from $24 depending on the chosen cloud

platform. The total deployment cost is therefore approximately $36 per event. More affordable GPU may reduce this cost but would require more training time. These estimates are based on our dataset with an average tweet length typical of Twitter/X posts; actual costs will vary with the number of tweets, their average length, as well as the LLM API cost per token.

Taken together, our experimental results suggest three practical implications. First, LLM-guided SSL (especially LG-CoTrain) is most beneficial when labeled data are extremely scarce (5–10 lb/cl), although the cost of using LLMs should be carefully considered in real-world settings with large volumes of noisy social media data. Second, as the amount of labeled data increases, simpler SSL methods such as Self-Training become highly competitive and can even outperform more complex approaches, making them strong baselines in moderate-resource settings. Third, calibration varies substantially across methods, and approaches such as AUM-ST-MixUp and VerifyMatch may be preferred when reliable confidence estimates are important.

In practice, this suggests the following. When only a very small labeled set is available (e.g., 5 lb/cl) and rapid deployment is needed, a cost-effective strategy is to generate a limited amount of pseudo-labeled data with an LLM and then train an LG-CoTrain model to filter actionable social media content. When a moderate amount of labeled data is available (e.g., 50 lb/cl), it may be preferable to avoid LLM costs and instead rely on task-specific SSL methods such as Self-Training. Prior work such as Guo et al. (2025) also shows that task-specific models are often preferred when feasible, as they can outperform LLMs in zero-shot settings. Finally, when interpretability and well-calibrated confidence estimates are critical, methods such as AUM-ST-MixUp provide a more suitable choice.

CONCLUSION AND FUTURE WORK

This paper provides the first empirical evaluation of LLM-guided semi-supervised learning for social media crisis classification on the HumAID benchmark. We compare two recent LLM-assisted SSL methods, VerifyMatch and LG-CoTrain, against widely used SSL baselines under multiple label budgets and evaluate the results in terms of predictive performance (Macro-F1) and as well as reliability (ECE).

Our results show that LG-CoTrain achieves the strongest Macro-F1 in extremely low-resource settings (5–10 labeled examples per class), demonstrating that LLM pseudo-labels can meaningfully improve SSL when labeled data are severely limited. VerifyMatch provides competitive classification performance and strong calibration, while AUM-ST-MixUp remains a strong choice when calibration is the primary concern. As the label budget increases, the performance advantage of LLM-guided SSL decreases and Self-Training becomes a difficult-to-beat baseline, emphasizing that method choice should reflect the available annotation budget and deployment requirements. We discuss corresponding recommendations for practical deployment.

Future work will explore three directions. First, we will investigate how unlabeled data scale and representativeness affect LLM-guided SSL, especially under missing-class or distribution-shift scenarios that can arise in real operations. Second, we will study calibration-aware training objectives and pseudo-label filtering strategies tailored to crisis informatics, aiming to improve both accuracy and reliability. Finally, we will extend the evaluation to multimodal crisis datasets and cross-event generalization settings, where LLMs may provide even greater benefits as a source of transferable knowledge.

ACKNOWLEDGMENT

This work is supported by a collaborative CAHSI-Google Institutional Research Program award.

REFERENCES

Alam, F., Joty, S., and Imran, M. (2018a). *Domain Adaptation with Adversarial Training and Graph Embeddings*. arXiv: 1805.05151 [cs.LG].

Alam, F., Joty, S., and Imran, M. (2018b). "Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 12. 1.

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

Alam, F., Qazi, U., Imran, M., and Ofli, F. (2021). "Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks". In: *Proceedings of the International AAAI Conference on Web and social media*. Vol. 15, pp. 933–942.

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.

Burel, G. and Alani, H. (2018). "Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media". In: *15th International Conference on Information Systems for Crisis Response and Management*.

Caragea, C., Squicciarini, A. C., Stehle, S., Neppalli, K., and Tapia, A. H. (2014). "Mapping moods: Geo-mapped sentiment analysis during hurricane sandy". In: *11th Proceedings of the International Conference on Information Systems for Crisis Response and Management, University Park, Pennsylvania, USA, May 18-21, 2014*. Ed. by S. R. Hiltz, L. Plotnick, M. Pfaf, and P. C. Shih. ISCRAM Association.

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. (2023). *SoftMatch: Addressing the Quantity-Quality Trade-off in Semi-supervised Learning*. arXiv: 2301.10921 [cs.LG].

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805.

Ghafarian, S. H. and Yazdi, H. S. (2020). "Identifying crisis-related informative tweets using learning on distributions". In: *Inf. Process. Manag.* 57.2, p. 102145.

Guo, D., Anh, T., Xiao, X., Li, H., and Caragea, D. (2025). "Multimodal Disaster-related Tweet Classification with Parameter-Efficient Fine-Tuning of Large Language Models". In.

Gupta, K., Gautam, N., Sosea, T., Caragea, D., and Caragea, C. (2025). "Calibrated Semi-Supervised Models for Disaster Response based on Training Dynamics". In: *Proceedings of the International ISCRAM Conference*.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Practical extraction of disaster-relevant information from social media". In: *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*. Ed. by L. Carr, A. H. F. Laender, B. F. Lóscio, I. King, M. Fontoura, D. Vrandecic, L. Aroyo, J. P. M. de Oliveira, F. Lima, and E. Wilde. International World Wide Web Conferences Steering Committee / ACM, pp. 1021–1024.

Imran, M., Ziaullah, A. W., Chen, K., and Ofli, F. (2025). "Evaluating Robustness of LLMs on Crisis-Related Microblogs across Events, Information Types, and Linguistic Features". In: *Proceedings of the ACM on Web Conference 2025*. WWW '25. Sydney NSW, Australia: Association for Computing Machinery, pp. 5117–5126.

Kersten, J., Kruspe, A. M., Wiegmann, M., and Klan, F. (2019). "Robust filtering of crisis-related tweets". 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.

Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q. R., and Li, J. (2025). "Harnessing large language models for disaster management: A survey". In: *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14528–14551.

Li, H., Caragea, D., and Caragea, C. (2021). "Combining Self-training with Deep Learning for Disaster Tweet Classification". In: *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*. ISCRAM Digital Library, pp. 719–730.

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.

Liu, X. and Wang, C. (Aug. 2021). "An Empirical Study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, pp. 2286–2300.

McDaniel, E., Scheele, S., and Liu, J. (2024). *Zero-Shot Classification of Crisis Tweets Using Instruction-Finetuned Large Language Models*. arXiv: 2410.00182 [cs.CL].

Mukherjee, S. and Awadallah, A. (2020). "Uncertainty-aware self-training for few-shot text classification". In: *Advances in Neural Information Processing Systems* 33, pp. 21199–21212.

Nguyen, D. T., Al-Mannai, K., Joty, S. R., Sajjad, H., Imran, M., and Mitra, P. (2016). "Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks". In: *CoRR* abs/1608.03902.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *Proceedings of the Eighth International Conference on Weblogs*

and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014. Ed. by E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh. The AAAI Press.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. Vancouver, BC, Canada: ACM, pp. 994–1009.

Park, S. Y. and Caragea, C. (2024). "VerifyMatch: A semi-supervised learning paradigm for natural language inference with confidence-aware MixUp". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19319–19335.

Purohit, H., Buntain, C., Hughes, A. L., Peterson, S., Lorini, V., and Castillo, C. (2025). *Engage and Mobilize! Understanding Evolving Patterns of Social Media Usage in Emergency Management*. arXiv: 2501.15608 [cs.HC].

Rahman, M. M. and Caragea, C. (Nov. 2025). "LLM-Guided Co-Training for Text Classification". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Reuter, C. (2022). "A European perspective on crisis informatics: citizens' and authorities' attitudes towards social media for public safety and security". PhD thesis. Radboud University Nijmegen, Netherlands.

Reuter, C. and Kaufhold, M.-A. (2018). "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 41–57.

Salfinger, A. and Snidaro, L. (2024). "Probing the Consistency of Situational Information Extraction with Large Language Models: A Case Study on Crisis Computing". In: *IEEE Conference on Cognitive and Computational Aspects of Situation Management, CogSIMA 2024, Montreal, QC, Canada, May 7-10, 2024*. IEEE, pp. 91–98.

Sánchez, C., Abeliuk, A., and Poblete, B. (May 2025). "Large Language Models in Crisis Informatics for Zero and Few-Shot Classification". In: *ACM Trans. Web*.

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

Shrestha, T. (2025). "Extracting actionable requirements from crisis event tweets for requirements engineers". MA thesis. Calgary, Canada: University of Calgary.

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.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. (2020). "Fixmatch: Simplifying semi-supervised learning with consistency and confidence". In: *Advances in neural information processing systems* 33, pp. 596–608.

Sosea, T. and Caragea, C. (2022). "Leveraging training dynamics and self-training for text classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4750–4762.

Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. (2010). "Chatter on the red: what hazards threat reveals about the social life of microblogged information". In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010, Savannah, Georgia, USA, February 6-10, 2010*, pp. 241–250.

Taghian Dinani, S., Caragea, D., and Gyawali, N. (2024). "Disaster Tweet Classification Using Fine-Tuned Deep Learning Models Versus Zero and Few-Shot Large Language Models". In: *Data Management Technologies and Applications*. Cham: Springer Nature Switzerland, pp. 73–94.

Wolbers, J., Kuipers, S., and Boin, A. (2021). "A systematic review of 20 years of crisis and disaster research: Trends and progress". In: *Risk, Hazards & Crisis in Public Policy* 12.4, pp. 374–392.

Yin, K., Liu, C., Mostafavi, A., and Hu, X. (2025). *CrisisSense-LLM: Instruction Fine-Tuned Large Language Model for Multi-label Social Media Text Classification in Disaster Informatics*. arXiv: 2406.15477 [cs.CL].

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412*.

Zou, H., Zhou, Y., Zhang, W., and Caragea, C. (2023). "Decrisismb: Debiased semi-supervised learning for crisis tweet classification via memory bank". In: *Findings of the association for computational linguistics: EMNLP 2023*, pp. 6104–6115.

Zou, H. P., Caragea, C., Zhou, Y., and Caragea, D. (2023). "Semi-supervised few-shot learning for fine-grained disaster tweet classification". In: *Proceedings of the 20th International ISCRAM Conference*. ISCRAM 2023.