Calibrated Semi-Supervised Models for Disaster Response based on Training Dynamics

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

Nikita Gautam
Kansas State University
ngautam@ksu.edu

Tiberiu Sosea
University of Illinois at Chicago
tsosea2@uic.edu

Doina Caragea
Kansas State University
dcaragea@ksu.edu

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

ABSTRACT

Despite advancements in semi-supervised learning (SSL) techniques that can be used when labeled data is limited, many SSL approaches still face challenges related to miscalibration. Calibration is crucial for ensuring the accuracy, reliability, and robustness of uncertainty estimates. In this work, we analyze the calibration performance of various SSL methods in the disaster response domain. Our results show that traditional self-training (ST) and mixup-based SSL methods often suffer from high Expected Calibration Error (ECE) despite achieving competitive F1 scores. In contrast, a newly introduced approach in the disaster domain, AUM-ST-Mixup, significantly improves calibration, achieving the lowest ECE across all settings. This improvement suggests that incorporating uncertainty-aware selection via Area Under the Margin (AUM) alongside mixup regularization enhances both predictive performance and model confidence alignment. Our findings highlight the importance of calibration-aware SSL methods, paving the way for more trustworthy model predictions in low-resource settings.

Keywords

Semi-supervised learning, model calibration, regularization, disaster response

INTRODUCTION

In recent years, Natural Language Processing (NLP) techniques have been increasingly used to support disaster response efforts by analyzing social media data contributed by eyewitnesses of disasters. Specifically, text classification has been very important in identifying pertinent information relevant to disaster management and response, including information about casualties, infrastructure damage, etc., from the substantial amount of data posted on social media platforms. Training accurate text classification models requires the availability of a large amount of labeled data, which is usually difficult to obtain in emergency situations.

To address the challenge of data availability, Semi-Supervised Learning (SSL) techniques leverage both labeled and unlabeled data (H. Li et al., 2021; Zou et al., 2023). SSL algorithms use the limited labeled data to train an initial model and utilize a larger set of unlabeled data to improve the model performance. One such popular algorithm is self-training (Rasmus et al., 2015; Scudder, 1965), which uses a teacher model trained on a small amount of labeled data to predict pseudo-labels on the unlabeled set. These pseudo-labeled examples are then used to train the student model. This procedure is iterated until a convergence requirement is met. While self-training offers substantial benefits, its success hinges on the quality of the pseudo-labels. Noisy labels can hinder model

generalization (C. Zhang et al., 2016), especially in deep neural networks, which are prone to overfitting. A major limitation of existing self-training approaches is their reliance solely on the teacher's confidence to determine whether a pseudo-labeled sample should be retained or discarded. This is because the confidence/uncertainty values may not accurately reflect label correctness if the teacher model is poorly calibrated (Guo et al., 2017).

Training well-calibrated classifiers is essential in disaster response and other critical applications. A calibrated model aligns its predictive confidence with its accuracy, enabling systems to quantify the likelihood of error in a prediction. This alignment is very important when AI applications are used in real-world scenarios, as an accurate estimate of confidence can facilitate human intervention in high-stakes situations where model errors could have severe implications, such as the disaster response domain. For instance, Guo et al. (2017) highlighted that current deep neural networks, while highly accurate, are often overconfident in their predictions. This misalignment between confidence and correctness can compromise trustworthiness, a critical issue in safety-critical applications (F. Li et al., 2019; Sarabadani, 2019). To address model calibration, various techniques have been developed, such as temperature scaling to adjust confidence levels (Guo et al., 2017) and label smoothing (Kumar & Sarawagi, 2019; Müller et al., 2019). Additionally, Mixup, a data augmentation method that combines training samples to create new examples, has shown promise for improving performance and calibration by regularizing training data (H. Zhang et al., 2018).

In this work, we propose an SSL approach that employs mixup informed by training dynamics to address the challenge posed by the noise in the form of non-calibrated pseudo-labeled data. The proposed model achieves a good tradeoff between model performance and calibration. Specifically, training dynamics is employed to categorize samples based on learning difficulty (e.g., easy-to-learn versus hard-to-learn samples), using the Area Under the Margin (AUM) (Pleiss et al., 2020) and subsequently the information about easy-to-learn and hard-to-learn examples is used to generate new mixup samples that have a regularization effect.

We provide a comprehensive evaluation of the proposed approach for handling noisy pseudo-labels within the disaster domain both in terms of model performance and calibration. In addition, we also compare the proposed SSL approaches with prior SSL baselines and supervised baselines (specifically, BERT, BERTweet, DBERT, and XLM-R) to gain insights into which models provide the best trade-off between performance and calibration while using a small number of labeled instances.

To our knowledge, our work is the first to systematically compare SSL approaches with supervised baselines in terms of both model performance and calibration. Our experiments specifically assess F1-score and Expected Calibration Error (ECE) metrics, aiming to balance accuracy with reliable probability estimates, which are crucial for decision-making in real-world disaster scenarios. Experimental results demonstrate that incorporating training dynamics-based mixup in SSL leads to a significant reduction in ECE while maintaining competitive classification performance. This highlights the trade-off where a slight decrease in F1 may be acceptable in favor of better-calibrated predictions, ensuring more reliable confidence estimates in real-world applications.

To summarize, using the proposed SSL approach for disaster post classification, which leverages the training dynamics combined with mixup strategies, our research seeks to answer the following key research questions:

RQ1 How does model calibration compare across different SSL approaches in low-resource disaster post classification?

RQ2 How does the proposed SSL approach that handles noisy pseudo-labels compare with other SSL approaches in terms of model performance and calibration?

RQ3 What are the trade-offs between calibration and overall classification accuracy when incorporating AUM and mixup strategies in SSL?

RELATED WORK

Imran et al. (2016) introduced Twitter as a Lifeline, a dataset consisting of human-annotated tweets from 19 crises, supporting supervised machine learning for disaster-related classification tasks. Their work demonstrates the utility of supervised traditional machine learning (ML) models, such as Naive Bayes, Support Vector Machines (SVM), and Random Forest models, in categorizing tweets into relevant information types, such as casualties or infrastructure damage. Alam et al. (2021) applied both traditional ML (e.g., SVM) and deep learning models (e.g., BERT and RoBERTa) to classify disaster tweets across multiple categories in a supervised setting.

Semi-supervised learning has been pivotal in addressing the challenges of limited availability of labeled data during disaster-related emergencies and in response. H. Li et al. (2021) combined self-training with BERT and CNN

models and showed that the semi-supervised BERT and CNN models have improved performance when a large amount of unlabeled data is used in the training process. In a more recent study, Zou et al. (2023) proposed a novel approach to enhance the performance of SSL models for crisis-related tweet classification. Their method addresses the bias commonly encountered in SSL models, where certain classes tend to receive disproportionately more accurate pseudo-labels, resulting in imbalanced performance across categories. To mitigate this, DeCrisisMB (Zou et al., 2023) leverages a memory bank to store pseudo-labels and performs equal sampling from each class during every training iteration.

Furthermore, advances in cross-lingual models have expanded their applicability by enabling them to operate across multiple languages. Given that social media platforms contain disaster-related posts in diverse languages, cross-lingual learning ensures that models trained on data from one language can generalize to others without requiring extensive re-training. Transformer-based models such as XLM-R and mBERT have shown promise in cross-lingual tasks, as they learn shared multilingual embeddings that facilitate transfer learning across languages (Conneau et al., 2020a). Ray Chowdhury et al. (2020) proposed a cross-lingual semi-supervised learning approach combined with manifold mixup, significantly improving multi-label classification across different languages.

In crisis scenarios, information often comes from multiple channels, including text, images, videos, and audio. Multi-modal learning leverages these diverse data types to enhance model performance by combining complementary information streams. This is crucial when individual modalities, like text-only tweets, may be incomplete or ambiguous (Imran et al., 2020). Koshy and Elango (2022) used a Transformer-based Bidirectional Attention Model to incorporate both textual and visual information from tweets, providing a more holistic approach to disaster tweet classification. Similarly, Mandal et al. (2024) showed the effectiveness of multimodal contrastive learning models on the task of classifying multimodal disaster tweets.

In comparison to these works, our work focuses on approaches that address the challenge posed by the noisy pseudo-labels generated during the self-training iterations by leveraging training dynamics, combined with mixup strategies (Berthelot et al., 2019; Park & Caragea, 2022), informed by AUM. Our study is the first to comprehensively evaluate both calibration (ECE) and classification performance (F1), which is much needed in the disaster domain.

METHODS

In this section, we first discuss several prior approaches used as baselines in our work and subsequently show how we leverage such approaches to design more accurate and calibrated models for use in the SSL setting in the disaster domain. Furthermore, we describe several supervised approaches that are used as baselines.

Prior SSL Approaches

Self-Training (ST): ST is a semi-supervised learning method where a teacher model is initially trained on a small labeled dataset and then subsequently refined using its own predictions on an unlabeled dataset (Mukherjee & Awadallah, 2020; Rasmus et al., 2015; Scudder, 1965). More specifically, the teacher model generates pseudo-labels for the unlabeled data, which are then further incorporated into the training process. The model is re-trained using both the original labeled data and these pseudo-labeled instances, and a student model is obtained. Subsequently, the student becomes the teacher, and the whole process is repeated until a pre-defined convergence condition is met.

Uncertainty Aware Self-Training (UST): UST is a variant of the traditional self-training approach (Mukherjee & Awadallah, 2020) that incorporates uncertainty estimation into the learning process. Instead of relying on just confidence to select the pseudo-labels, UST leverages Monte Carlo (MC) dropout to estimate model uncertainty by sampling multiple predictions per instance, allowing the identification of more reliable pseudo-labeled data. Further, it uses acquisition functions informed by these uncertainty estimates to prioritize the most informative unlabeled instances. By quantifying uncertainty, the model can avoid incorporating unreliable predictions into the training set, thereby reducing the risk of propagating errors and improving the robustness of the self-training process by ensuring that only high-confidence predictions contribute to model refinement.

MixMatch: MixMatch is another semi-supervised learning algorithm (Berthelot et al., 2019) which combines multiple techniques, such as data augmentation, label guessing, and consistency regularization, to create a unified framework. In this approach, the pseudo-labels are softened to reflect uncertainty. These pseudo-labels, along with the labeled data, are then subjected to extensive data augmentation to create diverse versions of the same examples, while the unlabeled and labeled examples are mixed using mixup (H. Zhang et al., 2018). This allows the model to produce consistent predictions for these augmented versions, enforcing a consistency regularization constraint.

AUM-based Self-Training (AUM-ST): AUM-ST (Sosea & Caragea, 2022) is an advanced SSL method that leverages training dynamics to improve text classification models by optimizing how they handle unlabeled data. The

Algorithm 1 AUM-based Mixup in Self-Training (AUM-ST-Mixup)

Input: Labeled data D_{lbl}, unlabeled data D_{ulbl}, #iterations T
Output: Trained model

Train base teacher model on D_{lbl}
for $i = 0$ to T **do**
  Generate pseudo-labeled dataset $\hat{y}_{ulbl}$ using the teacher model on D_{ulbl}
  Identify high/low AUM samples from the pseudo-labeled dataset: D_{ulbl_high}, D_{ulbl_low}
  $Total\_Loss \leftarrow 0$
  for $k = 0$ to $|D_{lbl}| + |D_{ulbl\_high}| + |D_{ulbl\_low}|$ batchwise **do**
    $Loss_{lbl} \leftarrow CrossEntropy(f(x_{lbl\_k}), y_{lbl\_k})$
    $Loss_{ulbl\_high} \leftarrow CrossEntropy(f(x_{ulbl\_high\_k}), \hat{y}_{ulbl\_high\_k})$
    Mixup labeled and hard-to-learn samples: $m_{lbl\_k\ \&\ ulbl\_low\_k} \leftarrow Mixup(x_{lbl\_k}, x_{ulbl\_low\_k})$
    Mixup easy-to-learn (High AUM) and hard-to-learn (Low AUM) samples: $m_{ulbl\_high\_k\ \&\ ulbl\_low\_k} \leftarrow Mixup(x_{ulbl\_high\_k}, x_{ulbl\_low\_k})$
    $Loss_{lbl\ \&\ ulbl\_low} \leftarrow CrossEntropy(f(m_{lbl\_k\ \&\ ulbl\_low\_k}), \hat{ym}_{lbl\_k\ \&\ ulbl\_low\_k})$
    $Loss_{ulbl\_high\ \&\ ulbl\_low} \leftarrow CrossEntropy(f(m_{ulbl\_high\_k\ \&\ ulbl\_low\_k}), \hat{ym}_{ulbl\_high\_k\ \&\ ulbl\_low\_k})$
    $Loss \leftarrow Loss_{lbl} + Loss_{ulbl\_high} + Loss_{lbl\ \&\ ulbl\_low} + Loss_{ulbl\_high\ \&\ ulbl\_low}$
  end for
  $Total\_Loss \leftarrow Total\_Loss + Loss$
  Update model weights
end for

teacher model is trained on weakly augmented labeled data and makes predictions on weakly augmented unlabeled data to generate pseudo-labels. AUM tracks the training dynamics of the weakly augmented pseudo-labeled examples, capturing the uncertainty associated with these examples. The low AUM examples are filtered out to ensure only high-quality pseudo-labeled examples are retained. Afterward, the student model is trained on both labeled and high-quality pseudo-labeled examples along with their corresponding strong augmentations, enforcing consistency regularization during the training.

Proposed Approaches

Confidence-based Mixup in Self-Training (Conf-ST-Mixup): The Conf-ST-Mixup approach incorporates pseudo-labeling and mixup strategies to improve semi-supervised learning by leveraging confidence difference between the logits with the highest and second highest confidence values from model predictions. Pseudo-labeled samples with high confidence difference are considered reliable and categorized as *easy-to-learn*. In contrast, those with low confidence difference are treated as *hard-to-learn* or *ambiguous*. During training, labeled data and high-confidence pseudo-labeled samples are mixed with low-confidence pseudo-labeled samples to create interpolated training samples. This encourages the model to gradually refine its decision boundaries while mitigating overconfidence and improving generalization. The losses are computed separately for labeled, unlabeled, and interpolated samples before being aggregated for model updates.

AUM-based Mixup in Self-Training (AUM-ST-Mixup): The AUM-ST-Mixup approach leverages the AUM uncertainty in pseudo-labels to guide training. It borrows the concept of mixup during self-training from (Berthelot et al., 2019) and utilizes AUM-based mixup from (Park & Caragea, 2022). However, instead of simply mixing up the samples from labeled and unlabeled data or performing mixup on the limited labeled samples, AUM-ST-Mixup first categorizes the pseudo-labeled data in *easy-to-learn* (high AUM values) and *hard-to-learn/ambiguous* (low AUM values). Subsequently, AUM-ST-Mixup combines the labeled data and easy-to-learn pseudo-labeled samples with the hard-to-learn pseudo-labeled samples with the intuition that the model will gradually become better at handling noisy labels and produce a better low-AUM prediction. The pseudocode for the AUM-ST-Mixup approach is shown in Algorithm 1. As can be seen, a teacher model is first trained and used to generate pseudo-labels for the unlabeled data. The pseudo-labeled samples are categorized into two groups: high-AUM samples D_{ulbl_high} and low-AUM samples D_{ulbl_low}. In each training iteration, the model is trained using labeled data, unlabeled data, and the two interpolated samples generated from mixing up labeled data and easy-to-learn pseudo-labeled samples, respectively, with hard-to-learn pseudo-labeled samples. The losses are calculated separately for labeled, unlabeled, and interpolated samples and combined to get an overall loss to update the model.

Supervised Baselines

We use several models as supervised baselines for the classification task. **BERT** (Devlin et al., 2019) is a transformer-based model that uses a bidirectional approach to understanding texts. The model is pre-trained using the Masked

Disaster Event	Classes	Posts	Class Distribution									
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
2017 Hurricane Irma	9	6534	429	397	88	528	626	0	1317	1113	1651	430
2017 Hurricane Harvey	9	6333	379	444	233	482	488	0	852	1976	1237	287
2018 Kerala Floods	9	5543	97	585	413	39	254	0	207	3005	669	319
2019 Hurricane Dorian	9	5284	958	758	125	561	42	0	571	691	1011	612
2018 California Wildfires	10	5113	97	330	55	258	1362	125	295	991	727	923
2017 Hurricane Maria	9	5049	154	470	498	92	211	0	999	1384	1097	189
2018 Hurricane Florence	9	4339	917	330	38	446	208	0	224	1034	445	742
2019 Cyclone Idai	10	2703	62	338	100	40	303	13	248	1308	285	56
2016 Canada Wildfires	8	1529	74	113	14	266	0	0	176	653	218	55
2016 Kaikoura Earthquake	9	1491	345	302	17	61	73	0	218	145	218	157

Table 1. Disaster events and the corresponding number of classes and posts (a.k.a., tweets) in training data per event with the class distribution

Language Model (MLM) and Next Sentence Prediction (NSP) as pretext self-supervised tasks. **BERTweet** (Nguyen et al., 2020) is a language model based on the BERT architecture that is pre-trained using a vast corpus of English tweets. The model captures the unique linguistic nuances of social media content. **RoBERTa** (Liu et al., 2019) is an optimized variant of BERT that improves performance by training the model longer, with larger batches, more data, and dynamically masking tokens during pretraining. It removes the Next Sentence Prediction (NSP) objective, showing that it isn't necessary for strong downstream performance. **CrisisTransformers-M1-Complete** (CT-M1-Complete) (Lamsal et al., 2024) is one of the ensemble of transformer-based models, fine-tuned for crisis-related social media text classification tasks. It builds on domain-specific data and uses multilingual or multi-task learning strategies to enhance robustness and generalizability across diverse crisis events. **DBERT**: D-BERT (Huang et al., 2021) is a variant of the BERT model that integrates dependency-based attention mechanisms to improve relation extraction performance. The model can effectively capture complex linguistic relationships between entities given in a text by incorporating syntactic dependency information into the attention layer and providing the model with contextual awareness. **XLM-R**: XLM-R (Conneau et al., 2020b) is a robust model pre-trained on an extensive corpus of CommonCrawl data (with over 100 languages). Trained with the multilingual MLM objective, the model performs better on various NLP tasks such as Cross-lingual NLI, NER, Cross-lingual QA, and others.

EXPERIMENTAL SETUP

In this section, we describe the dataset used to perform the experiments, as well as the evaluation metrics and important hyperparameter details.

Dataset

For this work, we utilize 10 distinct disaster events from the HumAID dataset (Alam et al., 2021). This is a meticulously developed, human-annotated collection of Twitter data, comprising over 77,000 tweets spanning 19 major natural disasters across diverse geographies and years from 2016 to 2019. The specific events included in our dataset are shown in Table 1. We selected these 10 disaster events because they contain the highest number of examples with the maximum possible class diversity, including two events with 10 classes and the remaining events with 9 classes. We do not pre-process the tweets before utilizing them in our models. Instead, we utilize the tweets in their original form without applying any filtering. The categories/labels included in the dataset are shown below, along with a short description of each label.

1. *Caution and advice*: Covers warnings, guidance, and tips related to the disaster.
2. *Sympathy and support*: Comprise tweets that express prayers and emotional support.
3. *Requests or urgent needs*: Captures reports of immediate needs for supplies such as food, water, and medical aid.
4. *Displaced people and evacuations*: Tracks the movement of people affected by the crisis.
5. *Injured or dead people*: Includes reports of casualties.

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

Table 2. Data Distribution across different settings. Note that *L* and *U* show the number of total data samples in the labeled and unlabeled set, respectively, for each of the 5, 10, 25, and 50 labels per class setting.

6. *Missing or found people*: Focuses on individuals who are unaccounted for or have been located.
7. *Infrastructure and utility damage*: Records damage to essential infrastructure.
8. *Rescue, volunteering, or donation effort*: Includes efforts related to rescue operations, volunteering, and donations.
9. *Other relevant information*: Includes important tweets that do not fit into the previous categories but are still crucial.
10. *Not humanitarian*: Captures tweets that are not relevant to disaster response.

Custom Dataset Splits

The training data for 10 individual disaster events is split into labeled and unlabeled subsets, simulating a low-resource semi-supervised learning setup. Specifically, the labeled set consists of 5, 10, 25, and 50 labeled examples per class, while the remaining data is treated as unlabeled. To account for the variability in data selection, we generate three independent labeled-unlabeled splits for each event. That is, for every event, we create three different random selections of labeled data for each of the 5, 10, 25, and 50 examples per class settings, resulting in 12 distinct labeled-unlabeled training sets. The three distinct training sets per labeled data setting ensure a more comprehensive evaluation of model performance under different labeled data constraints. However, the test and validation sets remain unchanged to ensure consistency in evaluation across different settings, as shown in 2.

Evaluation Metrics

We use two metrics to evaluate the results of our proposed approaches by comparison with the results of the baselines considered. First, we evaluate performance using the standard F1 measure, which captures the trade-off between precision and recall.

To evaluate model calibration, we use the Expected Calibration Error (ECE). The ECE metric is used to determine how much a model's predicted probabilities are representative of the actual outcomes. Well-calibrated models should be able to closely match the predicted confidence scores and the true likelihood of correctness. For example, a predicted label with 85% confidence should be correct 85% of the time. The ECE calculations involve grouping the predictions in equally spaced M bins based on confidence scores. For each bin, the model's accuracy and average confidence of the predictions in that bin are compared. Then, ECE is computed as the weighted average of the absolute differences between the accuracy and confidence across all bins. ECE is formally calculated as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{1}$$

where M is the number of confidence bins, B_m is the set of samples whose predicted confidence falls into bin m, n is the total number of samples, $\text{acc}(B_m)$ is the average accuracy of the bin m and $\text{conf}(B_m)$ is the average confidence of the bin m. The model is better calibrated if its ECE value is lower, demonstrating that the model's predicted probabilities are better aligned with actual outcomes.

Metric	F1	ECE
BERT	0.665	0.139
BERTweet	0.678	0.110
RoBERTa	0.637	0.132
CT-M1-Complete	0.664	0.123
DBERT	0.663	0.147
XLM-R	0.658	0.119

Table 3. Fully supervised results: Upper-bound performance of models trained exclusively on all labeled. The results are reported in terms of F1 and ECE values averaged over the 10 events in the dataset.

Dataset/metric	F1				ECE			
#Labeled samples used in SSL	5 *lb/cl*	10 *lb/cl*	25 *lb/cl*	50 *lb/cl*	5 *lb/cl*	10 *lb/cl*	20 *lb/cl*	50 *lb/cl*
BASE (BERTweet)	0.416	0.513	0.607	0.629	0.103	0.112	0.207	0.165
ST	0.471	**0.553**	**0.626**	**0.647**	0.360	0.292	0.258	0.234
UST	0.460	0.533	0.611	0.633	0.394	0.310	0.271	0.233
MixMatch	0.437	0.530	0.611	0.636	0.417	0.343	0.265	0.252
AUM-ST	0.405	0.497	0.572	0.604	0.178	0.240	0.193	0.172
Conf-ST-Mixup	0.440	0.538	0.608	0.638	0.418	0.324	0.269	0.243
AUM-ST-Mixup (ours)	**0.529**	0.541	0.604	0.633	**0.099**	**0.076**	**0.131**	**0.163**

Table 4. Comparison between the SSL baselines and the proposed SSL approaches (ours). Furthermore, supervised baselines using all the events data are shown at the bottom. The results are reported in terms of F1 and ECE values averaged over the 10 events in the dataset.

Hyperparameter Details

This section details the hyperparameters used in our study. For all the models, we utilize a learning rate of $5e-05$ and the Adam optimizer. The batch size varies depending on the size of the training set, with a supervised batch size of 16 for fine-tuning the base model and an unsupervised batch size of 64 for self-training on pseudo-labeled data. All the supervised models (teacher models or otherwise) are trained for 18 iterations. All the semi-supervised models utilize 12 self-training iterations per run. Training was carried out on NVIDIA A5000 GPUs. Model calibration is assessed using a uniform binning approach with four bins applied consistently across all models.

RESULTS AND DISCUSSION

The supervised baseline results are shown in Table 3, and can be seen as an upper bound for the SSL results. These results show that BERTweet outperformed other pre-trained models with an F1 score of 0.678, highlighting its effectiveness in the fully supervised setting. BERT , CT-M1-Complete, and DBERT had slightly worse performance in terms of F1, and RoBERTa was the least effective. Regarding calibration, BERTweet also achieved the lowest ECE value (0.110), indicating it had the best confidence calibration among all the supervised models. XLM-R, with an ECE of 0.119, was slightly less calibrated but still within a competitive range despite its lower performance. Given these results, we used BERTweet as the base model in our SSL experiments.

Table 4 shows the F1 and ECE results of our SSL experiments (the values are averaged over all events). Specifically, the table includes the results of prior SSL approaches (at the top), followed by the results of our proposed SSL approaches that address the challenge of noisy pseudo-labels through improved calibration. Results are presented for varying numbers of labeled examples per class. Across all methods, performance improved as the number of labeled samples increased. AUM-ST-Mixup recorded the lowest ECE values for all the labeled settings, i.e., 5, 10, 25, and 50 labels per class, respectively, showing improved model prediction reliability. Regarding our RQ2, we observe that AUM-ST-Mixup, in fact, consistently has the lowest ECE values among all semi-supervised and also better ECE values than the supervised baselines when 5 or 10 labels per class are used.

The analysis of performance across different numbers of labels per class indicates that increasing the number of labels generally improves F1 scores but not necessarily calibration (ECE). For instance, F1 for the BASE method increases from 0.416 at five labels per class to 0.629 at 50 labels per class, with similar gains across the rest of the semi-supervised methods. With respect to our RQ1, the ECE values, particular methods such as ST, UST, and MixMatch have a clear pattern of decreasing ECE values with more labeled examples, although this is not always

the case for our proposed models. Overall, these trends suggest that increasing the number of labeled samples generally enhances both classification accuracy and calibration quality.

In response to our RQ3, the ECE values across the models reveal a trade-off between the F1 score and model calibration. While models such as ST offered competitive F1 scores, their ECE values were comparatively higher, indicating overconfidence in predictions. Meanwhile, across all the methods, AUM-ST-Mixup consistently has the lowest ECE values, demonstrating its ability to train better-calibrated models at the expense of a small drop in F1 score.

CONCLUSION AND FUTURE WORK

In response to our RQ1, our results demonstrate that model calibration varies significantly across different SSL approaches. While AUM-ST-Mixup achieves the best overall calibration with the lowest ECE across all labeled data settings, ST and MixMatch exhibit higher miscalibration. By managing label noise through training dynamics and mixup strategies, AUM-ST-Mixup enhances the reliability of semi-supervised learning frameworks. Readdressing our RQ2, the SSL approach leveraging training dynamics with mixup outperforms the SSL models in terms of calibration. While our discussion highlighted the benefits of AUM-ST-Mixup in improving calibration, it is important to readdress our RQ3 to demonstrate the trade-off where methods that improve classification performance, such as ST, may lead to overconfident predictions. In contrast, AUM-based methods prioritize better-calibrated outputs, sometimes at a slight cost to F1. Thus, the choice of approach depends on whether the application prioritizes raw classification performance or well-calibrated confidence estimates, especially in high-stakes disaster response scenarios.

As for future exploration, we plan to progress this study in two directions. Firstly, we plan to develop more complex algorithms that blend semi-supervised learning techniques with model confidence scoring, self-training, advanced mixup, and active learning strategies to determine if they can further improve both the model performance and calibration. Secondly, we will explore the performance and calibration of LLMs and their integration with SSL methods in the disaster response domain. A key investigation would be exploring whether the larger LLMs offer significant performance improvements over smaller, more lightweight models when combined with SSL methods. Additionally, we will focus on refining the use of LLMs by introducing ensemble methods, such as voting-based prediction strategies, to aggregate predictions from multiple LLMs and improve both accuracy and robustness. These investigations will help determine whether LLMs can enhance SSL models in low-resource settings while maintaining reliable calibration.

REFERENCES

Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). Humaid: Human-annotated disaster incidents data from twitter. *15th International Conference on Web and Social Media (ICWSM)*.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020a, July). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020b, July). Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1321–1330.

Huang, Y., Li, Z., Deng, W., Wang, G., & Lin, Z. (2021). D-bert: Incorporating dependency-based attention into bert for relation extraction. *CAAI Transactions on Intelligence Technology*, *6*(4), 417–425.

Imran, M., Mitra, P., & Castillo, C. (2016, May). Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1638–1643).

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. *Information Processing & Management*, *57*(5), 102261. https://doi.org/https://doi.org/10.1016/j.ipm.2020.102261

Koshy, R., & Elango, S. (2022). Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing and Applications*, *35*, 1607–1627.

Kumar, A., & Sarawagi, S. (2019). Calibration of encoder decoder models for neural machine translation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 132–141.

Lamsal, R., Read, M. R., & Karunasekera, S. (2024). Crisistransformers: Pre-trained language models and sentence encoders for crisis-related social media texts. *Knowledge-Based Systems*, *296*, 111916. https://doi.org/https://doi.org/10.1016/j.knosys.2024.111916

Li, F., Jin, Y., Liu, W., Rawat, B. P. S., Cai, P., & Yu, H. (2019). Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study. *JMIR Med Inform*, *7*(3), e14830. https://doi.org/10.2196/14830

Li, H., Caragea, D., & Caragea, C. (2021). Combining self-training with deep learning for disaster tweet classification. *The 18th international conference on information systems for crisis response and management (ISCRAM 2021)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv*, *abs/1907.11692*.

Mandal, B., Khanal, S., & Caragea, D. (2024). Contrastive learning for multimodal classification of crisis related tweets. *Proceedings of the ACM on Web Conference 2024*, 4555–4564.

Mukherjee, S., & Awadallah, A. (2020). Uncertainty-aware self-training for few-shot text classification. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 21199–21212, Vol. 33). Curran Associates, Inc.

Müller, R., Kornblith, S., & Hinton, G. (2019). When does label smoothing help? *Advances in Neural Information Processing Systems*, *32*, 4694–4703.

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020, October). BERTweet: A pre-trained language model for English tweets. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 9–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Park, S. Y., & Caragea, C. (2022, May). On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5364–5374). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.368

Pleiss, G., Zhang, T., Elenberg, E., & Weinberger, K. Q. (2020). Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, *33*, 17044–17056.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., & Raiko, T. (2015). Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, *28*, 3546–3554.

Ray Chowdhury, J., Caragea, C., & Caragea, D. (2020, July). Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In S. Rijhwani, J. Liu, Y. Wang, & R. Dror (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: Student research workshop* (pp. 292–298). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-srw.39

Sarabadani, S. (2019, August). Detection of adverse drug reaction mentions in tweets using ELMo. In D. Weissenbacher & G. Gonzalez-Hernandez (Eds.), *Proceedings of the fourth social media mining for health*

applications (#smm4h) workshop & shared task (pp. 120–122). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3221

Scudder, H. (1965). Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, *11*(3), 363–371. https://doi.org/10.1109/TIT.1965.1053799

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

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *CoRR*, *abs/1611.03530*.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). Mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*.

Zou, H., Zhou, Y., Zhang, W., & Caragea, C. (2023). Decrisismb: Debiased semi-supervised learning for crisis tweet classification via memory bank. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6104–6115.