YNU-HPCC at SemEval-2026 Task 9: Hybrid Augmentation and Regularization Strategies for Multilingual Polarization Type Classification

Di Bao, **Jin Wang** and **Xuejie Zhang**
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: baodi@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper introduces a system based on fine-tuned pretrained language models, which is constructed for SemEval 2026 Task 9: Multilingual Polarization Type Classification. The task aims to perform multi-label polarization classification on texts covering 22 languages, identifying five types of polarization: political, racial/ethnic, religious, gender/sexual, and others. The main challenges of the task lie in handling uneven data distribution across languages, extreme class imbalance, and the complexity of cross-lingual semantic understanding. To address these challenges, a training framework integrating hybrid augmentation and multi-strategy regularization is proposed. Based on XLM-RoBERTa-large, the framework combines feature-space Mixup augmentation, an asymmetric loss function, adversarial training, and exponential moving average. Multi-label decisions are made through dynamic threshold optimization. Experimental results show that the proposed method achieves a macro-F1 score of 0.48 on the validation set, effectively improving classification performance and generalization capability in multilingual and imbalanced scenarios.

1 Introduction

With the proliferation of social media, online polarization has become increasingly severe, posing a significant threat to the health and safety of cyberspace. SemEval-2026 Task 9 introduces, for the first time, a task on multilingual, multicultural, and multi-event online polarization detection(Naseem et al., 2026a). This paper focuses on Subtask 2: multi-label classification of five polarization types. The task faces three major challenges: imbalanced data distribution across 22 languages, extreme class imbalance, and substantial differences in cross-lingual and cross-cultural expressions.

Traditional multilingual classification methods generally rely on fine tuning multilingual pre-trained models but often overlook the aforementioned data imbalance issues. Early attempts used focal loss(Lin et al., 2017) or weighted cross entropy to mitigate class imbalance, yet their effectiveness remains limited in extremely imbalanced scenarios. While adversarial training(Goodfellow et al., 2015) and model ensemble(Breiman, 1996) can enhance robustness, they fail to fully exploit the potential of data augmentation. Mixup technology(Zhang et al., 2018), though successful in computer vision, remains underexplored in natural language processing, especially when applied in feature space rather than input space.

To address the aforementioned deficiencies, this paper proposes an integrated solution. The main contributions are four-fold: (1) Feature-space Mixup is adapted to multilingual text classification, forming a complementary enhancement with adversarial training; (2) A tuned asymmetric loss function is designed, incorporating differentiated positive-negative sample focusing and probability clipping to optimize gradient updates; (3) A dynamic language-balanced sampling strategy based on the 75th percentile is proposed to alleviate data scarcity in low-resource languages individually; (4) Multi-sample dropout, EMA, and layer-wise learning rate decay are integrated to construct a stable and efficient training framework. Experimental results show that the proposed method achieves a macro-F1 score of 0.48 on the validation set, outperforming the baseline.

The remainder of this paper is organized as follows. Section 2 details the proposed methodology; Section 3 presents the experimental results and analysis; and Section 4 concludes the paper with future directions.

2 Related work

The development of multilingual pretrained models has laid the foundation for cross-lingual natural lan-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 699–705

guage processing tasks. Early multilingual models such as mBERT(Devlin et al., 2019) learned shared cross-lingual representations through joint training on multiple languages(Lample and Conneau, 2019). XLM-RoBERTa(Conneau et al., 2020) was further improved on this basis. It was trained on CommonCrawl data in 100 languages based on the RoBERTa architecture and achieved leading performance in various cross-lingual tasks through a shared subword vocabulary for cross-lingual representation alignment. Subsequently proposed mDeBERTa-v3 (He et al., 2021) introduced a disentangled attention mechanism and an enhanced masked decoder, demonstrating excellent performance in multilingual understanding tasks(Xue et al., 2021). However, these general models still require task-specific optimization when directly applied to fine-grained, culturally sensitive tasks such as polarization detection.In particular, polarized discourse often relies on specific cultural backgrounds and social contexts, which imposes higher demands on the cross-cultural understanding capabilities of multilingual models.

Handling class imbalance is a classical problem in machine learning. Common approaches include techniques at the data level and the algorithmic level. At the data level, SMOTE(Chawla et al., 2002) alleviates imbalance by synthesizing minority class samples, but it is difficult to apply directly to discrete text data(He et al., 2008). At the algorithmic level, cost-sensitive learning addresses imbalance by adjusting the weights of different classes in the loss function. Focal loss(Lin et al., 2017) reduces the weight of easy-to-classify samples through a modulating factor, but its symmetric design is limited in multi-label scenarios and cannot distinguish the different importance of positive and negative samples(Cui et al., 2019). Asymmetric loss(Ben-Baruch et al., 2021), through differentiated focusing mechanisms and probability clipping for positive and negative samples, has shown outstanding performance in long-tailed visual recognition, providing important inspiration for this work. However, existing research mostly focuses on single-language or visual domains. How to effectively adapt these techniques to multilingual text classification tasks still requires in-depth exploration.

Data augmentation is an effective means to improve model generalization. Mixup(Zhang et al., 2018) enhances data distribution by linearly interpolating between sample pairs, improving model robustness near decision boundaries(Yun et al., 2019). Subsequent research extended it to the feature space(Chen et al., 2020), avoiding semantic disruption in the text space. Adversarial training(Goodfellow et al., 2015) enhances model robustness against adversarial examples by adding small perturbations to the input(Miyato et al., 2019). Although PGD attacks(Madry et al., 2019) are stronger, they incur high computational costs, while FGM(Miyato et al., 2021) achieves a better balance between efficiency and effectiveness. Exponential moving average smooths parameter updates to reduce training oscillations and improve model stability(Tarvainen and Valpola, 2018). In multilingual scenarios, regularization techniques are particularly important because models need to handle the dual complexity arising from both language and cultural differences. However, existing studies have seldom systematically explored the synergistic mechanisms of these techniques in multilingual multi-label classification.

3 System Overview

An overview of the proposed multilingual polarization detection framework is presented in Figure 1. The framework consists of six main stages: data loading and balancing, preprocessing and splitting, tokenization and encoding, model architecture, advanced training strategies, and inference and output generation.

3.1 Dynamic Language-Balanced Sampling

To address the disparity in data volume across the 22 languages, a dynamic sampling strategy based on percentiles is proposed. First, the sample counts for each language are statistically analyzed, represented as $L = l_1, l_2, ..., l_{22}$, and the 75th percentile $T = P_{75}(L)$ is selected as the target sampling size. For low-resource languages with a sample count $l_i < T$, the oversampling factor and the remainderare calculated as follows.

$$r = T/l_i \tag{1}$$

$$m = T \bmod l_i \tag{2}$$

The sample size is then increased to the target T through repeated sampling and random supplementation, specifically formulated as follows.

$$D_{i'} = \text{concat}(D_i \times r, \text{sample}(D_i, m)) \tag{3}$$

For languages with ample data, the original data are kept unchanged. This strategy ensures that all

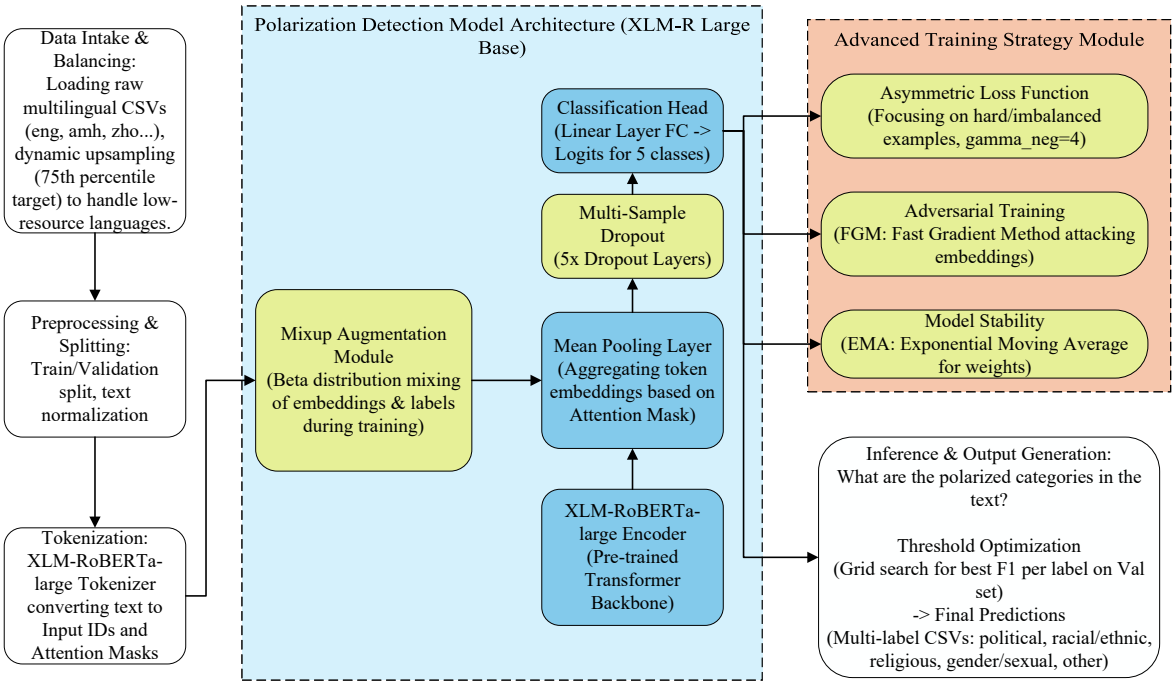

Figure 1: Overall architecture of the proposed hybrid augmentation training framework for multilingual polarization detection

languages receive adequate exposure during training while avoiding the damaging effects of destructive downsampling on high-resource languages. Experiments indicate that, compared to simple oversampling or undersampling, this strategy leads to significant improvements for low-resource languages while maintaining the performance for high-resource languages.

3.2 Feature-Space Mixup Augmentation

Traditional Mixup is applied in the input space, potentially disrupting the semantic integrity of the text. In this work, Mixup is performed in the feature space. Given the feature representations $E \in \mathbb{R}^{B\times d}$ and the labels $Y \in \mathbb{R}^{B\times 5}$ of a batch of samples, a mixing coefficient λ is randomly sampled from Beta(α,α), where α=0.4 controls the mixing intensity. A random permutation is generated to obtain an index set I. The mixed features and labels are then calculated as follows.

$$E_{\text{mix}} = \lambda E + (1-\lambda)E_I \tag{4}$$

$$Y_{\text{mix}} = \lambda Y + (1-\lambda)Y_I \tag{5}$$

The resulting soft labels provide richer supervisory signals, encouraging the model to learn smoother decision boundaries between categories. The advantage of feature-space Mixup is that it avoids direct interpolation of discrete text sequences. Instead, a linear combination is performed in the continuous feature space, thereby preserving semantic continuity. Furthermore, Mixup enhances generalization by expanding the training distribution, while adversarial training improves robustness through optimization under worst-case perturbations.

3.3 Asymmetric Loss Function

The standard binary cross-entropy loss treats all samples equally, which in extremely imbalanced scenarios may bias the model toward the majority class. An asymmetric loss function is adopted in this paper, with two key improvements introduced. First, different focusing parameters are applied to positive and negative samples, where $\gamma_{\text{pos}} = 1\ and\ \gamma_{\text{neg}} = 4$, thereby strengthening the focus on negative samples (non-polarized content). This design is motivated by the characteristic of polarization detection tasks, in which negative samples far outnumber positive ones, and it prevents the model from neglecting minority classes. Second, the predicted probabilities for negative samples are clipped as follows.

$$p_{\text{neg}} = \min(p+\delta, 1) \tag{6}$$

where $\delta = 0.05$, to prevent overconfidence on negative samples. This probability clipping mechanism ensures that a certain gradient signal is retained even when the model exhibits high confidence in negative predictions, thus avoiding training stagnation. The loss function is formulated as follows.

$$\mathcal{L}_{\text{ASL}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i(1-p_i)^{\gamma_{\text{pos}}}\log(p_i) + (1-y_i)p_i^{\gamma_{\text{neg}}}\log(1-p_i)\Big] \tag{7}$$

Asymmetric loss is inherently compatible with feature-space Mixup, since the soft labels generated by Mixup can be used directly as optimization targets for the loss function without binarization.

3.4 Adversarial Training Enhancement

Adversarial training improves model robustness by optimizing the model under worst-case perturbations. In this paper, the Fast Gradient Method (FGM) is adopted, and perturbations are added in the word embedding space as follows.

$$\theta_{\text{emb}}{}' = \theta_{\text{emb}} + \epsilon \cdot \frac{\nabla_{\theta_{\text{emb}}}\mathcal{L}}{||\nabla_{\theta_{\text{emb}}}\mathcal{L}||} \tag{8}$$

where $\epsilon = 1.0$ controls the perturbation strength. Adversarial training is performed after the normal forward pass. The adversarial loss $\mathcal{L}_{\text{adv}}$ is computed and its gradient is accumulated. The original parameters are then restored for updating. This process enables the model to learn representations that are invariant to semantic perturbations, which is particularly beneficial for handling common issues in social media texts such as spelling errors, abbreviations, and informal expressions. Compared to iterative attack methods such as PGD, FGM achieves a better balance between computational efficiency and effectiveness, making it suitable for large-scale multilingual training scenarios.

3.5 Multi-Technique Integrated Training

Multiple techniques are integrated into the training pipeline. For each batch, the following operations are performed sequentially: a mixup-augmented forward pass, normal loss computation, an adversarial training attack and recovery, and gradient accumulation update. The AdamW optimizer is used with a weight decay of 10^{-2}, effectively decoupling weight decay from learning rate adjustments. A linear warm-up schedule is applied for the first 10% training steps, followed by linear decay. The warm-up phase helps stabilize the early stage of training and prevents gradient explosions. Layer-wise learning rate decay is employed to assign different learning rates to the embedding layer, intermediate layers, and the classification layer, formulated as follows.

$$\eta_l = \eta_0 \times \phi^{L-l} \tag{9}$$

where $\phi = 0.95$,L is the total number of layers, and l is the layer index. By assigning higher learning rates to higher layers, they can adapt more quickly to the target task, while the general linguistic knowledge in lower layers is preserved. Exponential moving average is used to maintain a smoothed version of the model parameters.

$${\theta_t}^{\text{EMA}} = \beta {\theta_{t-1}}^{\text{EMA}} + (1-\beta)\theta_t \tag{10}$$

where decay rate $\beta = 0.999$.EMA reduces training oscillation by smoothing parameter updates. EMA parameters are applied during validation and model saving to improve stability. Multi-sample dropout further improves prediction stability by averaging predictions from multiple Dropout masks during inference, approximating model ensemble.

3.6 Dynamic Threshold Optimization

The decision threshold in multi-label classification significantly affects performance. A fixed threshold of 0.5 is often suboptimal, especially when the imbalance across categories varies. A fine-grained threshold search based on the validation set is proposed. For each polarization category i, an optimal threshold is independently searched as follows.

$$t_i^* = \arg\max_{t\in[0.2,0.8]} \text{F1}_i(t) \tag{11}$$

The search range is set to [0.2, 0.8] with a step size of 0.01, and the F1 score for each category is maximized on the validation set. Independent search allows each category to be assigned a personalized threshold according to its data distribution and classification difficulty. Experimental results demonstrate that dynamic threshold optimization yields significant performance gains over a fixed threshold, particularly for categories with the least data.

4 Experiment

4.1 Dataset and Evaluation Metrics

Experiments are conducted on the official SemEval-2026 Task 9 dataset(Naseem et al., 2026b). which

Model	Macro F1
bert-base-multilingual-cased	0.36
microsoft/mdeberta-v3-base	0.41
xlm-roberta-large	0.43

Table 1: Model experiment results

covers 22 languages, including Amharic, Arabic, Bengali, Burmese, Chinese, English, German, Hausa, Hindi, Italian, Khmer, Nepali, Odia, Persian, Polish, Punjabi, Russian, Spanish, Swahili, Telugu, Turkish, and Urdu. The data are collected from various social media platforms and online forums, spanning diverse cultural backgrounds and social events. For each language, the dataset is partitioned into training, validation, and test sets. The task is formulated as multi-label classification across five polarization categories: political, racial/ethnic, religious, gender/sexual, and other. Dataset statistics reveal significant discrepancies in the number of samples across languages and severe class imbalance across polarization labels. Macro-averaged F1 score is adopted as the primary evaluation metric, The macro-averaged F1 score is computed as follows.

$$\text{Macro} - \text{F1} = \frac{1}{C}\sum_{i=1}^{C}\frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \qquad (12)$$

where $C = 5$ denotes the number of categories, and P_i and R_i represent the precision and recall for the i-th class, respectively. Micro-averaged F1 score and per-class F1 scores are also reported as supplementary metrics to provide a more comprehensive performance analysis.

4.2 Comparative Experiments

First, model selection was performed. Due to the limited number of encoder models supporting all 22 languages involved in this task, three models that cover these languages were compared under identical parameter settings. XLM-RoBERTa-large achieved the best performance, as shown in Table 1. Therefore, it was selected as the final baseline model.

Subsequently, a series of ablation experiments were conducted to validate the effectiveness of each technical component, with results presented in Table 2. All experiments were performed on the same training and validation sets to ensure fair comparison. The baseline model was the standard fine-tuned XLM-RoBERTa-large, trained with binary cross-entropy loss and a fixed threshold of 0.5.

Configuration	Macro F1
baseline	0.43
+ Focal loss (=0.25)	0.35
+ Focal loss (=0.75)	0.46
+ Weighted layer pooling	0.39
+ FGM adversarial training	0.44
+ FGM + Focal loss	0.46
+ R-Drop regularization	0.38
+ SCL contrastive learning	0.32
+ PGD adversarial training	0.37
+ FGM + LLRD	0.45
+ FGM + Asymmetric loss	0.48
5-Fold Cross Validation	0.46
Full method (Ours)	**0.48**

Table 2: Ablation experiment results

Experimental results indicate that: (1) Asymmetric loss is more suitable for extreme imbalance scenarios than focal loss, particularly when negative samples far outnumber positive samples; (2) FGM adversarial training outperforms PGD in both efficiency and effectiveness, making it appropriate for large-scale training; (3) Feature-space Mixup and adversarial training complement each other effectively: Mixup expands the training distribution while adversarial training enhances robustness.

5 Conclusion

In this paper, a multilingual polarization detection framework is proposed for SemEval-2026 Task 9, integrating dynamic balanced sampling, hybrid augmentation training, and multi-strategy regularization. The proposed framework effectively addresses core challenges, including imbalanced multilingual data distribution, extreme class imbalance, and insufficient model generalization. Experimental results show that a macro-F1 score of 0.48 is achieved on the validation set, which represents a significant improvement over the baseline model. However, due to cultural differences in interpreting polarized discourse, performance in certain languages remains suboptimal. Future work will focus on incorporating fine-grained cultural knowledge to improve context-specific understanding and on exploring the transferability of this framework to broader content security tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification.

Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24:123–140.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. 2020. Semi-supervised learning under class distribution mismatch. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:3569–3576.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186. Association for Computational Linguistics.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007. IEEE.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards deep learning models resistant to adversarial attacks.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2021. Adversarial training methods for semi-supervised text classification.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization.

Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031. IEEE.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization.