Research Paper

Benchmarking Full Fine-Tuning, LoRA, and Adapters for Emotion Detection Across Diverse Social Text Datasets

Nikhil Kumar[1], Divyendu Mishra[2]

[1]M.Tech Scholar, Dept. of CSE, Faculty of Engineering and Technology
Email: nikhil07kr@gmail.com

[2]Assistant Professor, Dept. of CSE, Faculty of Engineering and Technology
Email: divyendu01mishra@gmail.com

ABSTRACT

Emotion detection from text has seen rapid progress, yet comparing systems across studies remains difficult because datasets, label spaces, languages, and tuning protocols vary simultaneously. Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) and adapter modules are increasingly used to reduce computational cost, but their behavior under heterogeneous emotion detection conditions has not been systematically studied. This paper presents a unified benchmark comparing three tuning strategies full fine-tuning (FFT), LoRA, and adapter modules across seventeen transformer backbones and four emotion datasets: GoEmotions, SemEval-2018 E-c, ArmanEmo, and EmoMix-3L. Holding the evaluation protocol fixed, we find that ELECTRA-base-discriminator consistently performs best among standard-size backbones, LoRA remains close to FFT in most settings, and code-mixed text remains the hardest condition for every model and strategy. The benchmark provides a practical reference for selecting efficient tuning methods for emotion detection across diverse social text environments.

Keywords: Emotion Detection; Multi-Label Classification; LoRA; Adapters; Parameter-Efficient Fine-Tuning; Benchmarking; Social NLP; Multilingual NLP.
How to cite this article: Kumar N, Mishra D. Benchmarking Full Fine-Tuning, LoRA, and Adapters for Emotion Detection Across Diverse Social Text Datasets. Int J Drug Deliv Technol. 2026;16(55s): 1076-1081. DOI: 10.25258/ijddt.16.55s.106

Source of support: Nil.

Conflict of interest: None.

1. INTRODUCTION

Emotion detection from text is an important problem in natural language processing and affective computing, with applications in mental health monitoring, crisis detection, customer feedback analysis, and conversational artificial intelligence.[1] Modern systems move beyond coarse sentiment labels and attempt to recognize fine-grained affective states, including anger, joy, sadness, fear, surprise, and other nuanced emotions. However, comparing emotion detection systems across studies remains difficult because datasets, label spaces, languages, domains, metrics, and tuning protocols often vary simultaneously.

A second challenge is computational cost. Full fine-tuning of transformer models requires updating all model parameters, which becomes expensive when separate models are needed for multiple datasets, languages, or deployment settings. Parameter-efficient fine-tuning (PEFT) methods such as LoRA[2] and adapter modules[3] reduce this cost by freezing most of the pretrained backbone and updating only a small set of task-specific parameters. Although these methods are widely used, their behavior across heterogeneous emotion detection datasets remains insufficiently studied. This paper addresses both issues by benchmarking three tuning strategies FFT, LoRA, and adapters across seventeen transformer backbones and four emotion datasets covering English Reddit,[4] multilingual tweets,[5] Persian social text,[6] and Bangla-English-Hindi code-mixed social text.[7] The goal is not to introduce a new architecture, but to provide a controlled and reproducible comparison of tuning strategies across diverse real-world emotion detection settings.

The main contributions are:

- We present a systematic benchmark comparing FFT, LoRA, and adapter tuning

across four diverse social text emotion datasets.
- We evaluate seventeen transformer backbones under a unified protocol covering English, multilingual, language-specific, and social-media-oriented models.
- We provide practical findings on when LoRA can substitute FFT and where full fine-tuning remains preferable, especially for code-mixed emotion detection.

2. RELATED WORK

Emotion detection has progressed from lexicon-based approaches such as the NRC Emotion Lexicon[8] to transformer-based contextual modeling. A cross-corpus study by Bostan and Klinger[9] showed that emotion datasets differ substantially in annotation schemes, label inventories, and domains. Large-scale datasets such as GoEmotions,[4] SemEval-2018 Task 1,[5] ArmanEmo,[6] and EmoMix-3L[7] have expanded emotion detection across English, multilingual, Persian, and code-mixed settings. However, differences in dataset design and evaluation protocol still make direct comparison difficult.

Transformer models such as BERT,[10] RoBERTa,[11] XLM-RoBERTa,[12] and DeBERTa[13] have become strong baselines for text classification. Domain- and language-specific models such as BERTweet and ParsBERT[14] further improve performance in social-media and Persian settings. At the same time, PEFT methods such as adapters,[3] LoRA,[2] prefix tuning,[15] and prompt tuning[16] reduce adaptation cost by updating only a small subset of parameters. Surveys and frameworks such as Delta Tuning and AdapterHub[17,18] summarize these methods, but their comparative behavior across heterogeneous emotion detection datasets remains underexplored.

3. MATERIALS AND METHODS

3.1. Benchmark Design

The benchmark is designed to isolate the effect of tuning strategy and backbone model. Preprocessing, optimization, metrics, and data splits are kept fixed, while only the tuning strategy and pretrained backbone are varied. This controlled design ensures that observed differences reflect genuine properties of each strategy rather than differences in experimental setup.

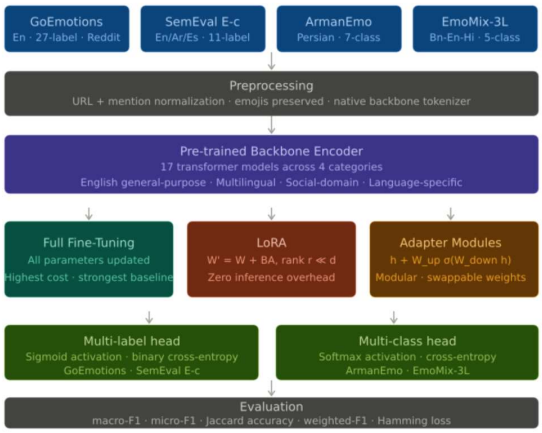

Figure 1: Overview of the benchmark pipeline used to compare full fine-tuning, LoRA, and adapter tuning across transformer backbones and emotion datasets.

3.2. Datasets

Table 1 summarizes the four datasets used in the benchmark.

Table 1: Summary of datasets used in the benchmark.

Dataset	Lang.	Domain	Task	Labels
GoEmotions	En	Reddit	Multi-label	27+N
SemEval-2018 E-c	En/Ar/Es	Tweets	Multi-label	11+N
ArmanEmo	Fa	Social/Review	Multi-class	7
EmoMix-3L	Bn-En-Hi	Social	Multi-class	5

GoEmotions[4] contains 58,000 English Reddit comments annotated with 27 emotion categories plus Neutral. It is treated as a multi-label classification task.

SemEval-2018 E-c[5] contains English, Arabic, and Spanish tweets labeled with one or more emotion categories. We use the official E-c subtask and evaluation protocol.

ArmanEmo[6] contains Persian social and review-style text across seven emotion classes. It provides a non-English evaluation condition for emotion detection.

EmoMix-3L[7] contains Bangla-English-Hindi code-mixed social text across five emotion classes. It is the most linguistically challenging dataset in the benchmark.

3.3. Backbone Models

We evaluate seventeen transformer backbones across English general-purpose, social-media-specific, multilingual, and language-specific model families. These include BERT,[10] RoBERTa,[11] DeBERTa,[13] DistilBERT, ELECTRA, BERTweet, Twitter-RoBERTa, XLM-RoBERTa,[12] mDeBERTa, ModernBERT, SONAR, ParsBERT,[14] and MuRIL.[19] English general models serve as broad baselines, social-media models test domain-matched pretraining, multilingual models test cross-lingual transfer, and language-specific models provide reference points for Persian and Indian-language settings.

3.4. Tuning Strategies

Full Fine-Tuning (FFT). In FFT, all backbone parameters and the classification head are updated jointly. This setting provides the highest adaptation flexibility and serves as the performance ceiling.

LoRA. LoRA freezes the pretrained model and injects trainable low-rank updates into selected projection layers.[2] For a pretrained weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA modifies it as:

$$W' = W + BA \quad (1)$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ are trainable matrices, with $r \ll min(d, k)$. The learned update can be merged at inference time, adding no extra latency.

Adapters. Adapter tuning inserts lightweight bottleneck layers inside transformer blocks.[3] The adapter output is:

$$Adapter(h) = h + W_{up}\sigma(W_{down}h) \quad (2)$$

Only adapter weights and the task head are trained. This makes adapters useful when a single backbone must support multiple tasks or languages.

3.5. Training and Evaluation Protocol

URLs and user mentions are normalized, whitespace is cleaned, and punctuation, emojis, and expressive repetitions are retained because they may carry emotional information. For EmoMix-3L, the code-mixed surface form is preserved without transliteration. Each model uses its native tokenizer.

Models are trained using AdamW with early stopping on validation macro-F1, and Jaccard score is used for SemEval E-c. The learning rate is 2×10^{-5} for FFT and 1×10^{-4} for PEFT heads. Batch size is 16 or 32, training runs for 5–10 epochs, weight decay is 10^{-2}, and warmup is 0.1. For multi-label tasks, sigmoid activation with binary cross-entropy is used; for multi-class tasks, softmax with categorical cross-entropy is used. Macro-F1 is the primary metric because it treats all emotion classes equally and is suitable for imbalanced datasets.

4. RESULTS AND DISCUSSION

To keep the manuscript within the journal page limit, Table 2 reports the strongest overall models and the strongest standard-size backbone across all datasets. Complete model-wise results for all seventeen backbones under FFT, LoRA, and adapter tuning can be provided as supplementary material.

Table 2: Compact summary of best-performing models across datasets and tuning strategies.

Dataset	Model	FFT F1	LoRA F1	Adapter F1
GoEmotions	SONAR†	0.941	0.913	0.887
GoEmotions	ELECTRA-base-disc.	0.628	0.602	0.583
SemEval-2018 E-c	SONAR†	0.648	0.623	0.607
SemEval-2018 E-c	ELECTRA-base-disc.	0.634	0.608	0.590
ArmanEmo	SONAR†	0.741	0.719	0.699
ArmanEmo	ELECTRA-base-disc.	0.726	0.705	0.688
EmoMix-3L	SONAR†	0.578	0.567	0.551
EmoMix-3L	ELECTRA-base-disc.	0.561	0.543	0.529

†*SONAR uses model-internal normalized scores for accuracy; F1 is standard macro-F1.*

4.1. Full Fine-Tuning Performance

Under FFT, ELECTRA-base-discriminator is the strongest standard-size backbone across all four datasets, reaching macro-F1 scores of 0.628 on GoEmotions, 0.634 on SemEval-2018 E-c, 0.726 on ArmanEmo, and 0.561 on EmoMix-3L. ELECTRA-small also performs strongly, indicating that ELECTRA's discriminative pretraining objective aligns well with the fine-grained classification demands of emotion detection.

Scaling from base to large gives modest gains for RoBERTa and DeBERTa-V3 families, while DistilBERT remains the weakest model across datasets. BERTweet performs notably well on ArmanEmo relative to GoEmotions, likely because ArmanEmo includes social-media text from Twitter and Instagram, which is closer to BERTweet's pretraining domain. SONAR achieves the highest overall F1 on ArmanEmo, showing the value of massively multilingual pretraining for non-English emotion detection.

4.2. LoRA vs. Full Fine-Tuning

LoRA remains close to FFT in most settings. For ELECTRA-base-discriminator, LoRA reaches F1 scores of 0.602, 0.608, 0.705, and 0.543 across GoEmotions, SemEval-2018 E-c, ArmanEmo, and EmoMix-3L, compared with FFT scores of 0.628, 0.634, 0.726, and 0.561. This shows that LoRA can serve as a practical substitute for FFT when computational resources are limited.

The main exception is EmoMix-3L. Code-mixed text creates a stronger distribution mismatch because of language switching, transliteration variation, vocabulary overlap, and informal spelling. In this setting, full end-to-end adaptation remains more useful than low-rank updates alone. Therefore, FFT should be preferred when the target data is highly code-mixed or distribution-shifted.

4.3. Adapter Performance

Adapters generally perform below LoRA but remain competitive. For ELECTRA-base-discriminator, adapter F1 scores are 0.583, 0.590, 0.688, and 0.529 across the four datasets. Although this is lower than LoRA, the gap is relatively small. The main advantage of adapters is modularity: separate adapters can be trained and swapped for different tasks or languages without modifying the frozen backbone. This makes adapters useful for multi-task or multi-language deployment, even when they are slightly weaker than LoRA in raw performance.

4.4. Key Findings

Four important observations emerge from the benchmark. First, ELECTRA-base-discriminator is the strongest standard-size backbone across datasets and tuning strategies. Second, LoRA is a reliable alternative to FFT in most settings, usually with a small F1 reduction. Third, adapters remain useful when modular deployment is important, although they generally perform below LoRA. Fourth, EmoMix-3L is the hardest dataset across all strategies, confirming that code-mixed emotion detection remains a challenging open problem.

4.5. Representation Analysis

The t-SNE visualizations support the quantitative findings but are omitted from the main manuscript to satisfy the page limit. Under FFT, emotion representations form clearer class clusters than under LoRA and adapter tuning. LoRA maintains moderate class separation, while adapter tuning shows greater inter-class overlap. The overlap is strongest for EmoMix-3L, further confirming that code-mixed emotion detection is the most difficult setting in the benchmark. Complete t-SNE plots can be provided as supplementary material.

5. CONCLUSION

This paper presented a unified benchmark of FFT, LoRA, and adapter tuning for emotion detection across seventeen transformer backbones and four diverse social text datasets. ELECTRA-base-discriminator emerged as the strongest standard-size backbone across datasets and tuning strategies. LoRA achieved performance close to FFT in most settings, making it a practical choice when computational efficiency is important. Adapter tuning performed slightly below LoRA but offered modular advantages for multi-task and multi-language deployment.

Code-mixed emotion detection, represented by EmoMix-3L, remained the most challenging setting across all models and strategies. Future work should explore stronger adaptation strategies for code-mixed and low-resource social text, newer PEFT methods such as IA3, and improved methods for handling label imbalance in fine-grained emotion datasets.

ACKNOWLEDGEMENT

The authors are thankful to the Faculty of Engineering and Technology (UNSIET), Veer Bahadur Singh Purvanchal University, Jaunpur, Uttar Pradesh, India, for providing academic support.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Nikhil Kumar contributed to conceptualization, methodology, experimentation, result analysis, and manuscript writing. Divyendu Mishra contributed to supervision, validation, critical review, and manuscript revision. All authors reviewed and approved the final manuscript.

References

[1] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017. doi: 10.1016/j.inffus.2017.02.003.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[3] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of Proceedings of Machine Learning Research, pages 2790–2799. PMLR, 2019.

[4] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372.

[5] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1001.

[6] Hossein Mirzaee, Javad Peymanfard, Hamid Haj Moshtaghin, and Hossein Zeinali. ArmanEmo: A Persian dataset for text-based emotion detection. *arXiv preprint arXiv:2207.11808*, 2022.

[7] Nishat Raihan, Dhiman Goswami, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. EmoMix-3L: A code-mixed dataset for bangla-english-hindi emotion detection. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation (WILDRE-7)*. European Language Resources Association, 2024.

[8] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013. doi: 10.1111/j.1467-8640.2012.00460.x.

[9] Laura Ana Maria Bostan and Roman Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119. Association for Computational Linguistics, 2018.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 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, 2020. doi: 10.18653/v1/2020.acl-main.747.

[13] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[14] Mehrdad Farahani, Mohammad Gharachorloo, Kia Jahanbakhshi, and Omid Shirali. ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53(6):3831–3847, 2021. doi: 10.1007/s11063-021-10528-4.

[15] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.353.

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.243.

[17] Ning Ding, Yujia Qin, Guang Yang, Furu Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022.

[18] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.7.

[19] Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji

Gopalan, Dilip Kumar Singh, and Partha Talukdar. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*, 2021.