Emotion Classification in Creative Text: Domain Adaptation and Fine-Tuning Strategies for RoBERTa on Poetry and Song Lyrics

Lia Berman **Wilson Sun** **Zeke Delaughter** **Kee Chee Pheng**
University of Arizona
{liaberman,wilsonsun,zedelaughter,cheephengkee}@arizona.edu

Abstract

Emotion classification in creative text, such as poetry and song lyrics, presents a significant challenge due to the reliance on metaphor, narrative, and mixed emotional states. This paper details the implementation and evaluation of a complete text-based emotion classification system built on the RoBERTa-base transformer model. We investigate the effectiveness of four distinct fine-tuning strategies for domain adaptation across three diverse datasets (tweets, poems, and songs). Our results demonstrate that full fine-tuning yields the best performance in poetic text (Poem $F1 \approx 0.60$), confirming the need for comprehensive domain adaptation when shifting from general social media language to complex creative writing. The primary bottleneck to further improvement is identified as the severe class imbalance and small size of the evaluation datasets.

1 Introduction

1.1 Motivation

Why Text-Based Mood Classification? Emotion in creative text is complex. Lyrics and poems often contain multiple overlapping emotions and rely heavily on metaphor, imagery, narrative framing, and figurative language. Because of this, surface-level audio cues such as tempo, loudness, or genre frequently misrepresent the underlying emotional meaning.

For example, Mr. Brightside is widely perceived as upbeat due to its fast tempo and energetic production, yet the lyrics describe jealousy, insecurity, and emotional distress. Audio-based systems would mislabel it, while a text-based model analyzing the actual words can correctly identify its emotional content.

This distinction matters for real-world applications. Users searching for "happy songs" often receive fast or energetic tracks whose lyrics are negative. Text-based emotion classification enables more accurate retrieval of songs that genuinely express positive affect, which is important for personalized recommendations, mood-aware interfaces, and wellness-oriented applications.

Context Matters in Emotion Analysis In "From Sentiment to Emotion: Understanding the Nuances of Emotional Expression in Text" (Mohammad, 2021), the author emphasizes that emotional expression in text is subtle, layered, and highly context-dependent. It cannot be captured through shallow cues such as individual keywords, sentiment markers, or audio characteristics. Effective emotion analysis must consider lexical choices, contextual relationships, and broader narrative structure.

This directly supports our approach: because creative works like poems and lyrics often blend conflicting emotions or shift tone across lines, a transformer-based NLP model which encodes deep

Preprint. Under review.

contextual and semantic relationships is better suited for emotion detection than systems relying on audio signals or keyword spotting. This motivates our use of RoBERTa and contextual fine-tuning strategies for accurate mood classification in creative domains.

These challenges motivate our use of a two-stage domain adaptation pipeline and four fine-tuning strategies to evaluate how different levels of parameter updating affect representation learning in creative text.

1.2 Approach and contribution

We investigate a transfer-learning pipeline designed to progressively adapt a pretrained RoBERTa model from general emotion data to the figurative and structurally complex domain of creative writing. Our study examines not only whether transfer is possible, but also which adaptation strategies are most effective given limited data and substantial class imbalance. We further analyze how dataset properties interact with fine-tuning depth to influence model variance and stability.

Our study makes four contributions. First, we present a two-stage domain-adaptation process that transfers emotional priors from large-scale social media data to poetry and lyrics. Second, we provide a systematic comparison of four parameter-update strategies (classifier-head-only training, partial unfreezing, full fine-tuning, and LoRA) evaluated across both poetic and lyrical text. Third, we conduct a detailed variance analysis demonstrating how dataset size, label distribution, and representational depth jointly shape model behavior. Finally, we identify intra-song emotional heterogeneity as a key limitation of single-label lyric classification, motivating segment-level modeling as a promising direction for future research.

1.3 Datasets

Our experiments use three datasets that together form a progressive transfer-learning pipeline from general emotional language to highly figurative creative writing.

Tweet Emotion Dataset The first stage of training uses a large-scale social media emotion dataset containing six basic categories: *anger, fear, joy, love, sadness*, and *surprise*. Tweets offer diverse but generally literal emotional expressions, making this dataset effective for initializing emotional priors in the model. Because of its size and label clarity, it serves as the foundation for Stage 1 pretraining.

PERC Poem Emotion Corpus The second dataset is a curated corpus of poems written between the nineteenth and twenty-first centuries, annotated with nine fine-grained emotional labels. This collection represents the primary domain-adaptation target due to its extensive use of metaphor, irregular narrative structure, and stylistic distance from modern text corpora. Exposure to poetic language allows the model to acquire representations relevant to figurative emotional expression that are unlikely to be learned during generic pretraining.

500 Songs Emotion Dataset Finally, we evaluate cross-domain generalization on a collection of 500 English-language songs annotated by multiple raters. The dataset is intentionally small and displays severe label imbalance: positive emotions such as "joy" and "love" dominate, while categories like "anger" and "fear" are sparse. This imbalance produces volatile validation metrics and reflects the challenges of real-world lyric annotation. We use this dataset solely for downstream evaluation to assess how well creative-domain adaptation transfers to musical lyrics, which differ from poetry in both structure and emotional pacing.

2 Methodology

2.1 Training Pipeline

We adopt a two-stage training pipeline designed to gradually expose the model to increasingly figurative text.

Stage 1 We fine-tuned RoBERTa on the Tweet Emotion dataset to learn general patterns associated with basic emotional categories. This step ensures that the model develops stable emotional representations before being introduced to more ambiguous or metaphorical expressions.

Stage 2 We utilized the PERC poem corpus to further train the model. This stage emphasizes learning stylistic and semantic properties unique to creative writing, such as indirect emotional signaling and non-literal imagery. A standard classification head consisting of a dropout layer followed by a linear projection and softmax activation is used to generate label predictions.

2.2 Fine-Tuning Strategies

To understand how parameter-update depth influences performance in creative-text emotion classification, we compare four fine-tuning strategies commonly used in transfer learning.

Cls Head Train only the final classification head while keeping all transformer layers frozen; this approach is parameter-efficient but limits the model's ability to adapt to domain-specific linguistic features.

Last N Layers Unfreeze the final two transformer layers in addition to the classification head, allowing limited structural adaptation while controlling the risk of overfitting on small datasets.

Full Fine-tuning Update all model parameters. Although this method offers the greatest flexibility, it may be unstable on data-scarce domains, particularly when emotional categories are highly imbalanced.

LoRA Low-Rank Adaptation method inserts trainable matrices into the attention layers while keeping the original parameters frozen. LoRA updates fewer than 1% of total parameters yet often provides substantial performance gains by directly modifying attention subspaces most relevant to emotional semantics.

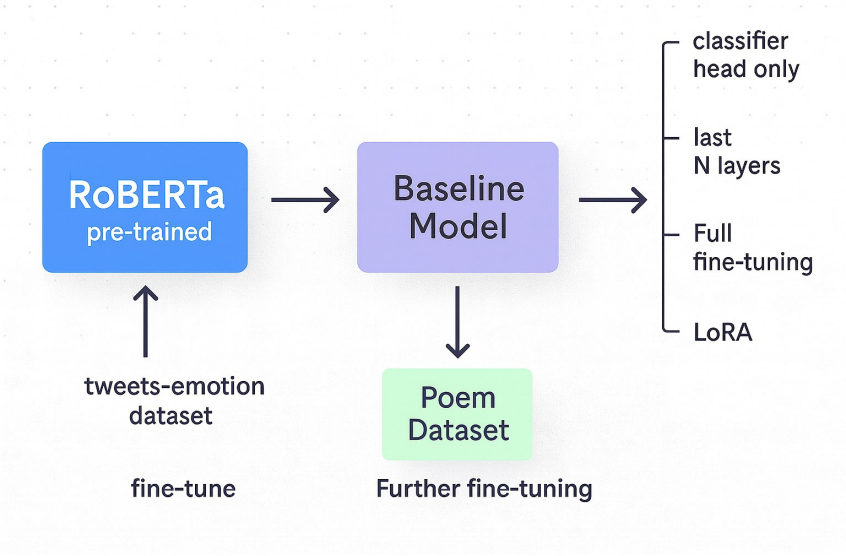

Figure 1: Two-stage domain adaptation pipeline and four fine-tuning strategies.

3 Experiments

3.1 Metrics

To evaluate emotion classification under substantial class imbalance, we adopt the Weighted F1 score as the primary metric. This choice reflects the need to account for minority emotional categories

that would otherwise be overshadowed by majority-class predictions. Robustness is assessed by examining variations across random seeds, the stability of validation loss throughout training, and the structure of per-class confusion matrices. Together, these indicators allow us to understand not only overall performance but also the consistency and interpretability of model behavior.

3.2 Baselines

We compare our methods against two baselines. The first is a simple majority-class predictor, which reflects the effect of extreme class imbalance within the creative-text datasets. The second baseline is a pretrained RoBERTa model without any domain-specific fine-tuning, representing the performance ceiling of using general-purpose language understanding for figurative text. Both baselines perform poorly on poetry and lyrics, confirming that conventional representations do not sufficiently capture the metaphorical or emotionally layered structure of creative writing.

3.3 Results

Table 1: Weighted F1 Scores on Poetry and Song Lyrics

Method	**Poem F1**	**Song F1**
Classifier Head Only	≈ 0.45	≈ 0.13
Last 2 Layers	≈ 0.44	≈ 0.14
LoRA	≈ 0.57	≈ 0.17
Full Fine-Tuning	$\approx \mathbf{0.60}$	$\approx \mathbf{0.18}$

Full fine-tuning achieves the strongest performance, particularly on poetry, where deep layers appear to benefit from extensive adaptation to figurative semantics. LoRA, despite updating less than 1% of model parameters, delivers performance nearly comparable to full fine-tuning. This suggests that creative-domain adaptation can be effectively captured by adjusting attention subspaces rather than modifying the full parameter space. In contrast, training only the classification head or the top transformer layers yields limited gains, likely because modifying shallow layers introduces noise on such a small dataset.

Song lyrics exhibit substantially lower and more volatile scores across all training configurations. This degradation partially stems from label inconsistency and the inherent structural heterogeneity of lyrics, where emotional tone often shifts across verses and choruses, making single-label annotation an insufficient representation of the underlying emotional landscape.

4 Analysis

4.1 Validation Instability

Although the training loss curves display smooth convergence across all models, validation F1 fluctuates sharply—often varying by more than 0.10 between consecutive epochs. This instability arises from several factors. The validation set is small relative to the model capacity, increasing sensitivity to sampling noise. Emotional labels in creative text are inherently variable, particularly for metaphor-rich or ambiguous passages, which compounds annotation inconsistency. Moreover, the severe class imbalance amplifies variance, since rare classes disproportionately influence the weighted F1 metric when misclassified.

4.2 Emotional Confusion Patterns

The confusion matrices further illuminate recurring classification challenges. The most prominent error involves "Joy" and "Love," which are consistently misclassified as each other across baseline and fine-tuned models. This aligns with their overlapping lexical cues and shared positive sentiment markers. Negative emotions display similar overlap; for instance, "Fear" is frequently predicted as "Sadness," especially in narratives that express vulnerability or emotional withdrawal. Additionally, "Anger" remains significantly under-detected, reflecting its low frequency in the dataset as well as its subtle expression in poetic and lyrical contexts. These findings align with observations from prior

work indicating that figurative emotional signals are often intertwined, making sharp categorical distinctions difficult without domain-specific representations.

4.3 Limitations of Single-Label Supervision

The structural mismatch between song composition and single-label supervision imposes additional performance constraints. Songs commonly alternate emotional tone across sections. In one representative case, the model correctly identifies an upbeat emotional shift in the chorus when the lyrics are segmented, yet the global single-label annotation forces the entire piece to be classified as *sadness*. This discrepancy illustrates that current annotation granularity restricts model performance and masks genuine understanding of emotional dynamics.

5 Future Work

5.1 Hierarchical Segment-Based Modeling

Given the multi-layered emotional structure of songs, future research should incorporate hierarchical models that process lyrics at the segment level. By dividing texts into structural components such as verses, choruses, and bridges, models can first predict segment-level emotional distributions and subsequently integrate them into a holistic profile. Such a design would mitigate label inconsistency, better capture temporal emotional shifts, and provide more interpretable predictions for downstream applications.

5.2 Domain-Specific Pretraining

Our findings indicate that pretrained RoBERTa embeddings are insufficiently representative of figurative language. Narrative, poetic, and lyrical texts rely heavily on metaphor, imagery, and symbolic expressions that are underrepresented in general corpora. Large-scale domain-specific pretraining or intermediate training on creative-writing corpora may significantly enhance metaphor sensitivity and emotional nuance. This direction aligns with growing evidence that continued pretraining on targeted domains can substantially boost performance in tasks requiring specialized linguistic competence.

5.3 Multimodal and Prosodic Extensions

Although this study focuses exclusively on textual features, song lyrics are inherently multimodal. Incorporating audio cues such as pitch contour, tempo, and vocal affect could help disambiguate subtle emotional states. Additionally, structural features such as rhyme density, meter, and prosodic rhythm may offer complementary signals aligned with both emotional intensity and thematic progression. Multimodal fusion therefore presents a promising path toward more comprehensive creative-text understanding.

6 Conclusion

This work investigates the challenges of emotion classification in poetry and song lyrics, two domains characterized by figurative language, emotional layering, and structural variability. Full fine-tuning of RoBERTa yields the strongest quantitative results, while LoRA demonstrates an appealing balance between efficiency and performance. Despite these improvements, substantial volatility and persistent confusion between semantically adjacent emotions highlight the limitations of small datasets and single-label annotation schemes.

More broadly, the results underscore the need for methodological innovations beyond straightforward fine-tuning. Segment-level modeling, domain-specific pretraining, and multimodal integration represent promising avenues toward robust emotional understanding in creative text. As emotion-rich content plays an increasingly central role in applications ranging from recommendation systems to mental-wellness tools, advancing models capable of capturing nuanced human affect remains an essential goal for future NLP research.

References

[1] Y. Liu, M. Ott, N. Goyal, J. Du, M. Li, V. Ko, A. Mahoney, L. Zettlemoyer, V. Stoyanov, X. Ade, M. Kettl, S. Ramea, and P. Singh. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Lu, W. Wen, H. Wang, and L. Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv:2106.09685*, 2021.

[3] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018.

[4] S. Kant and R. Gupta. PERC: Poem Emotion Recognition Corpus. *Mendeley Data, V1*, 2019.

[5] H. Inan, A. O. Muis, and L. Bing. Song Lyrics Emotion Dataset (500 Song Dataset). 2020.

[6] A. Vaswani, N. Shazeer, N. Parmar, L. Uszkoreit, J. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Github repository: https://github.com/zlu3s/csc396_project