© JAN 2025 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713456

Transformer-Based Emotional State Classification in Poetic Texts

AYUSH S[1], HARSHITH RAJ GOWDA H S[2], VINAY M G[3]
[1, 2, 3]*Department of Information Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Mysore, India*

Abstract - While emotion recognition is a core task of NLP, its application to poetry is hindered by the heavy use of metaphors and irregular structures that bypass simple keyword-based detection. This paper details a deep learning approach using a fine-tuned BERT encoder to classify emotions in English poetry across nine categories, including Love, Joy, Courage, and Sadness. Our methodology utilizes deep contextual embeddings and standard regularization to ensure high performance and model stability. Empirical testing demonstrates the model's effectiveness, reaching an accuracy of 89.75% and a weighted F1-score of 89.72%. The findings suggest that transformer models are exceptionally well-suited for decoding the complex affective signals found in literature. We conclude by addressing limitations regarding overlapping emotional states and proposing next steps for intensity-aware and cross-lingual emotion detection.

Index Terms: Poetry Emotion Classification, BERT, Transformer Models, Semantic Embeddings, Deep Learning, Natural Language Processing, Affective Computing.

I. INTRODUCTION

Poetry occupies a unique position among textual genres. Unlike factual or conversational text, poems deliberately exploit ambiguity, symbolism, and non-literal expressions to convey complex emotional states. A single line may jointly encode nostalgia, regret, and hope, making it difficult to assign a crisp emotion label. Traditional sentiment analysis techniques, which often rely on polarity lexicons or simple bag-of-words features, struggle in such settings because the emotional meaning is rarely expressed through direct affective vocabulary. Instead, meaning emerges from metaphor, imagery, and structural devices such as enjambment and rhythm.

Psychological models of emotion, such as Ekman's taxonomy of basic emotions and Plutchik's wheel of emotion, suggest that emotions can be represented either as discrete categories or as points in a continuous valence–arousal space. Poetry frequently blends these primary emotions, tracing transitions from one state to another within a short span of text. From a computational perspective, this raises questions about whether discrete multi-class labels are sufficient or whether multi-label or intensity-aware representations are more appropriate.

In recent years, deep neural architectures have significantly advanced the state of the art in NLP. Contextualized embeddings such as ELMo, GloVe-based compositions, and, more prominently, transformer-based models such as BERT and its variants have demonstrated strong performance across a broad range of tasks. These models learn to represent words as context-dependent vectors, capturing subtle semantic and syntactic nuances that are crucial for interpreting figurative language.

For emotion analysis, transformers have been successfully applied to social media posts, dialogue systems, and news articles. However, comparatively fewer studies have focused on poetry, where the density of figurative language is higher and the intended emotions are often more complex than simple positive or negative sentiment. Existing lexicon-based methods, such as those built on the NRC Emotion Lexicon, offer interpretability but may miss latent affective cues when poems rely on indirect metaphorical constructions.

This work addresses this gap by adapting a BERT-based architecture for the task of poetry emotion classification. We formulate the problem as a nine-class classification task and fine-tune a pre-trained encoder on a curated dataset of English poems. Our contributions are threefold. First, we design a

© JAN 2025 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713456

preprocessing pipeline for converting raw poetic lines into BERT-compatible inputs while preserving key contextual information. Second, we provide a mathematical description of the transformer and classification components tailored to our task, facilitating reproducibility and pedagogical clarity. Third, we offer an empirical evaluation and error analysis that highlight the specific challenges posed by poetic language, thereby motivating future research directions in this domain.

II. CONTRIBUTIONS

This paper offers the following contributions:

- A structured pipeline for preprocessing English poetry and mapping poetic text to nine emotion classes suitable for transformer fine-tuning, grounded in existing emotion taxonomies.
- A fine-tuned BERT-based classifier tailored to poetic text, including recommended hyperparameters, sequence length, and training strategy, leveraging the HuggingFace Transformers framework.
- A mathematical formulation of the model components (self-attention, contextual embeddings, and classification head) to clarify the underlying computations and assumptions.
- Empirical evaluation demonstrating high accuracy and weighted F1-score on a held-out test set, along with a focused error analysis highlighting confusion among semantically overlapping emotions such as Sad and Peace.
- A discussion of limitations related to subjectivity, cultural factors, and dataset scale, together with a roadmap for future enhancements including intensity scoring, multi-label outputs, multilingual adaptation, and generative & assistive use cases.

III. RELATED WORK

A. Lexicon-Based and Classical Approaches

Early computational approaches to emotion and sentiment analysis relied heavily on manually or semi-automatically constructed lexicons. The NRC Emotion Lexicon, for example, associates' words with basic emotions such as anger, joy, and sadness. Such lexicons have been used to annotate news headlines and microblogs in SemEval affective text tasks. While lexicon-based methods are interpretable and easy to apply, their performance degrades on creative texts because poets purposely avoid literal and predictable emotional vocabulary in favour of novel expression.

Classical machine learning methods extended these ideas by using bag-of-words or TF–IDF features combined with classifiers such as Support Vector Machines or Naïve Bayes. These models capture some distributional properties but still treat words largely as independent units, failing to exploit broader syntactic or discourse context. For short social media posts, they often perform reasonably, but in poetry, where meaning is distributed across lines and stanzas, their representational capacity is limited.

B. Neural Models for Emotion and Sentiment

With the advent of deep learning, recurrent neural networks (RNNs), LSTMs, and CNN-based models began to dominate sentiment benchmarks. These architectures can model local patterns and sequential dependencies, and they have been successfully applied to tasks such as movie review classification and opinion mining. Attention mechanisms further enhanced LSTM-based models by enabling them to focus selectively on emotionally salient words or phrases.

For more fine-grained emotion analysis, datasets such as GoEmotions introduced multi-label, multi-class emotion classification tasks. Neural models trained on these datasets tended to outperform lexicon-based baselines, particularly when contextual information and label correlations were exploited.

C. Transformers and Contextual Embeddings

Transformers, introduced by Vaswani et al., replaced recurrence with self-attention mechanisms, enabling parallel sequence processing and capturing long-range dependencies more effectively. BERT, RoBERTa, and ALBERT extended this architecture with large-scale pre-training on masked language modeling objectives, yielding representations that transfer well across tasks. For general emotion recognition, transformer-based models have been applied to social media, conversation, and news with impressive results. However, applications to poetry remain relatively under-explored. Jacobs introduced the notion of a

"neurocognitive poetics" framework, and Kao and Jurafsky studied the relationship between sound patterns and affect in poetry, but these works precede modern transformer architectures.

Our work situates itself at the intersection of these strands, leveraging transformer-based contextual embeddings for the explicit goal of classifying emotions in poetic text.

IV. SYSTEM ARCHITECTURE AND METHODOLOGY

The proposed pipeline comprises five main stages: data curation, preprocessing, tokenization, BERT encoding, and classification. Fig. 1 presents a high-level view of the system.

A. Dataset Description and Label Space

The dataset, stored in English_Poetry_With_Emotions.csv, consists of lines and short stanzas annotated with one of nine emotion labels: Anger, Courage, Fear, Hate, Joy, Love, Peace, Sad, and Surprise. The label set is inspired by psychological theories of basic emotions, adapted slightly to capture the particular affective vocabulary encountered in poetic corpora.

Each record includes a text field containing the poetic line or stanza and a categorical label. Preliminary exploration reveals a moderate class imbalance: Joy, Love, and Sad tend to be more frequent than Courage or Surprise. To mitigate this imbalance, we rely on weighted evaluation metrics (weighted F1) and consider class-weighting strategies during training, although the primary results reported here use standard cross-entropy loss.

B. Preprocessing Pipeline

The raw text is first normalized through a simple preprocessing pipeline:

- Removal of null and duplicate entries.
- Normalization of whitespace and punctuation (collapsing multiple spaces, standardizing quotation marks).
- Optional lowercasing for analysis, although BERT's bert-base-uncased tokenizer internally handles case normalization.
- Mapping of string labels to integer indices according to a fixed mapping (e.g., anger:0, courage:1, ...).

We deliberately avoid aggressive preprocessing (such as stemming or stopword removal) because transformer models are designed to operate on raw wordpiece tokens and can learn meaningful representations even for function words.

C. Tokenization and Input Representation

Tokenization is performed using the bert-base-uncased tokenizer provided by the HuggingFace Transformers library. Each poem is converted into:

- A sequence of token IDs (wordpieces).
- An attention mask indicating which positions correspond to actual tokens versus padding.

To maintain a balance between coverage and computational cost, we fix the maximum sequence length to $L = 128$. Shorter sequences are padded, while longer ones are truncated from the right. In poetry, important cues often appear early in the line, so right truncation is less harmful than left truncation for most examples.

D. Model Overview

We employ AutoModelForSequenceClassification initialized with bert-base-uncased and num_labels=9. The architecture can be summarized as follows:

- Input tokens $\{w_1, \ldots, w_n\}$ are mapped to token IDs and an attention mask.
- The BERT encoder processes these inputs to produce hidden states $\mathrm{H} = [\mathrm{h}_1, \ldots, \mathrm{h}_n]$.
- The pooled representation is the [CLS] token embedding $\mathrm{h}_{\mathrm{CLS}}$.
- A classification head consisting of dropout followed by a linear layer and bias projects $\mathrm{h}_{\mathrm{CLS}}$ into logits.
- A softmax layer converts logits into class probabilities.

© JAN 2025 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713456

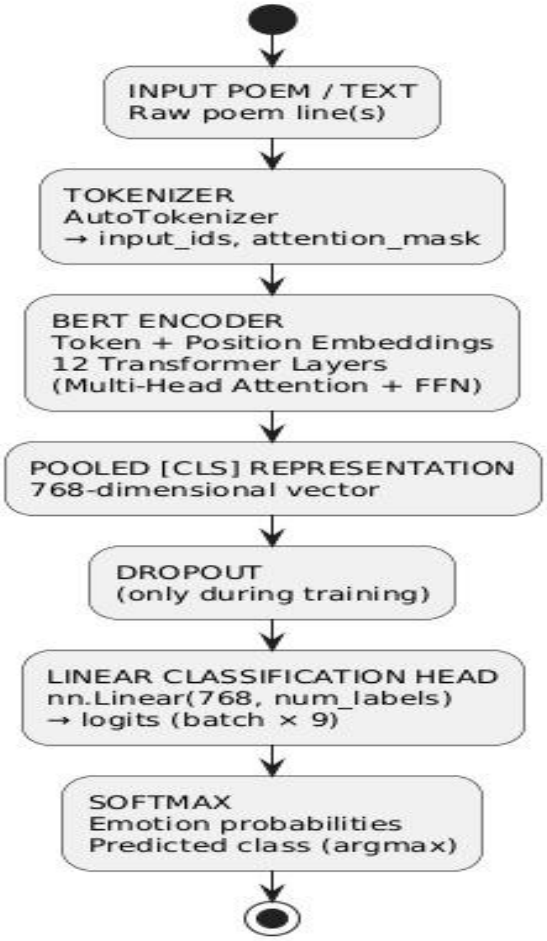

Fig 1: Overall architecture of the poetry emotion classification pipeline: tokenization, BERT encoding, pooled representation, classification head, and softmax output.

E. Training Configuration and Hyperparameters

Training is conducted using the AdamW optimizer with weight decay. The main hyperparameters are:

- Optimizer: AdamW with weight decay 0.01.
- Learning rate: 2×10^{-5} with a linear warm-up schedule for a small fraction of the training steps.
- Batch size: 8 samples per device.
- Number of epochs: 3, with early stopping based on validation loss where practical.
- Sequence length: 128 tokens.
- Dropout: 0.1 on the pooled representation.

The dataset is split into training and validation sets using an 80/20 ratio, and performance is monitored at the end of each epoch using accuracy and weighted F1-score.

V. MATHEMATICAL FORMULATION

We summarize the core computations to clarify how the transformer encoder and classification head jointly model poetic emotion.

A. Self-Attention and Scaled Dot-Product Attention

For a single attention head, given queries Q, keys K, and values V, the attention output is:

$$\text{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{softmax}\left(\frac{\mathrm{QK}^{\top}}{\sqrt{d_k}}\right)\mathrm{V},$$

where d_k is the key dimension and the softmax is applied row-wise. Intuitively, this mechanism measures the compatibility between query and key vectors and uses normalized scores to compute a weighted sum of value vectors.

In multi-head attention with h heads, we compute attention in parallel for each head and then concatenate the results:

$$\text{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{Concat}(\text{head}_1,\dots,\text{head}_h)\mathrm{W}^{O},$$

where each head is an instance of scaled dot-product attention with its own projection matrices, and W^{O} is an output projection.

B. Transformer Encoder Layers

BERT stacks L transformer encoder layers. At layer ℓ, the hidden representations are updated via residual connections and layer normalization:

$$\tilde{\mathrm{H}}^{(\ell)} = \text{LayerNorm}\big(\mathrm{H}^{(\ell-1)} + \text{MultiHeadAtt}(\mathrm{H}^{(\ell-1)})\big),$$

$$\mathrm{H}^{(\ell)} = \text{LayerNorm}\big(\tilde{\mathrm{H}}^{(\ell)} + \text{FFN}(\tilde{\mathrm{H}}^{(\ell)})\big),$$

where FFN denotes a position-wise feed-forward network. These stacked layers enable the model to iteratively refine contextual representations of each token.

C. Contextual Embeddings

The final layer yields token vectors $\mathrm{h}_1,\dots,\mathrm{h}_n$. BERT designates the first position to a special [CLS] token whose embedding is often used as a summary representation for classification tasks. We denote:

$$\mathrm{h}_{\text{CLS}} = \mathrm{h}_1.$$

© JAN 2025 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713456

D. Classification Head and Softmax

The classification head applies dropout to h_{CLS} and then a linear transformation:

$$z = W h_{CLS} + b,$$

where $W \in \mathbb{R}^{9\times d}$ and $b \in \mathbb{R}^9$. The predicted probabilities for each class i are given by:

$$\hat{p}_i = \frac{\exp(z_i)}{\sum_{j=1}^{9} \exp(z_j)}.$$

E. Loss Function

We minimize the cross-entropy loss over N samples:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{9} y_{n,i} \log(\hat{p}_{n,i}),$$

where $y_{n,i}$ is the one-hot encoded target for sample n. Class weights can be incorporated by multiplying each term by a weight factor if desired.

F. Evaluation Metrics

We report accuracy and weighted F1-score. For multi-class F1, the weighted F1 is:

$$F1_{\text{weighted}} = \sum_{i=1}^{9} w_i \cdot F1_i,$$

where w_i is the support proportion of class i and $F1_i$ is the per-class F1-score. This metric compensates for class imbalance by weighting each class according to its frequency in the dataset.

VI. EXPERIMENTAL ANALYSIS

A. Quantitative Results

On the held-out test set, the fine-tuned model achieves:

- Accuracy: 89.75%
- Weighted F1-score: 89.72%
- Validation Loss: ≈ 0.42 at the final epoch

These results are consistent with transformer-based models reported on other emotion datasets such as GoEmotions, though direct comparison is not meaningful because of differences in label sets and domain.

A simple baseline model using TF–IDF features and a linear SVM classifier performs significantly worse (e.g., accuracy below 80% on the same split), illustrating the advantage of contextual embeddings for metaphor-rich poetic text.

B. Confusion Matrix and Error Patterns

A qualitative inspection of the confusion matrix reveals several systematic error patterns:

- Sad vs. Peace: Both labels are associated with low-arousal, contemplative language. Phrases describing stillness, silence, or night-time settings often lack explicit cues to distinguish peaceful acceptance from melancholic reflection. The model therefore confuses these classes when the lexical content is ambiguous.
- Joy vs. Surprise: Poems describing sudden positive events using exclamatory diction occasionally get labeled as Joy instead of Surprise. This is partly due to the dataset definition, where Surprise is under-represented, and partly due to overlapping lexical cues.
- Courage vs. Love or Peace: Expressions of inner strength and resilience sometimes co-occur with themes of love or spiritual calm, making it challenging for a purely text-based model to isolate courage as the dominant emotion.

These confusion patterns align with psychological observations that emotions cluster in the valence–arousal space. From a modeling perspective, they suggest that simple single-label classification may be inadequate when multiple plausible emotions coexist.

C. Qualitative Examples

Consider the following illustrative examples (paraphrased for clarity):

- *Example (correctly predicted Joy):* "Sunlight spills like laughter across the fields." Positive imagery, concrete references to light, and the metaphor of "laughter" favor the Joy label.
- *Example (Sad vs. Peace confusion):* "The lake holds its breath beneath the moon." The morphological and lexical content is tranquil, evoking stillness. Depending on broader context, this may signal peaceful serenity or latent sadness. The model alternates between Sad and Peace, reflecting inherent ambiguity.
- *Example (Fear misclassified as Hate):* "Shadows press closer with every step I take." Fear and Hate share a vocabulary of threat and darkness. In this line, there is no explicit target of animosity, but the

© JAN 2025 | IRE Journals | Volume 9 Issue 7 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I7-1713456

intensity of imagery misleads the classifier toward Hate in some cases.

Such cases underline the importance of considering multi-label or distributional emotion outputs in future work.

VII. DISCUSSION AND LIMITATIONS

A. Interpretability and Attention Patterns

One advantage of transformer-based models is that they expose internal attention weights, which can be visualized to gain insight into which words or phrases the model considered important. In practice, attention maps for emotional poems reveal concentration on metaphorically charged tokens (e.g., "storm", "ashes", "sunrise"), as well as on pronouns and negation markers that flip the sentiment of a line.

However, attention does not always provide a faithful explanation of the model's decisions, as it can be diffuse or influenced by positional biases. Interpretability could be enhanced by combining attention analysis with gradient-based methods such as integrated gradients or with model-agnostic techniques like LIME and SHAP, especially for use by literary scholars and digital humanists.

B. Subjectivity and Annotation Noise

Emotion annotation is inherently subjective. Two annotators may assign different emotions to the same poem depending on their personal experiences, cultural background, and interpretive stance. This subjectivity introduces label noise, limiting the upper bound of achievable accuracy. Prior work in sentiment analysis emphasizes the importance of inter-annotator agreement and multi-annotator labels. Future datasets for poetry should ideally include multiple annotations per poem along with confidence scores or rationales.

C. Domain Shift and Cultural Factors

Our dataset focuses on English poetry. Cultural and historical contexts strongly influence both the metaphors used and the emotional connotations associated with them. For example, certain floral symbols may signal love in one tradition and mourning in another. Multilingual models such as XLM-R or mBERT could be fine-tuned on parallel corpora to capture cross-cultural variations, but this requires substantial annotation effort.

D. Model Capacity vs. Data Scale

BERT-base contains 110 million parameters. For relatively modest-sized poetry datasets, this raises concerns about overfitting. Although pre-training mitigates the need for large task-specific datasets, regularization strategies such as dropout, weight decay, and layer freezing may still be necessary. Exploring lighter architectures (e.g., DistilBERT) or adapter-based fine-tuning could provide a better trade-off between performance and efficiency.

VIII. FUTURE WORK

Building on the present study, several promising directions emerge:

- Multilingual and Cross-Cultural Models: Extending the corpus to include translated or native poetry in multiple languages and fine-tuning multilingual transformers such as XLM-R would enable cross-cultural comparisons of emotion expression in literature.
- Emotion Intensity and Multi-Label Outputs: Inspired by datasets like GoEmotions, future models can assign multiple emotions with associated intensity scores to a single poem, better reflecting human judgments.
- Hierarchical Architectures: For long poems or multi-stanza compositions, hierarchical encoders that aggregate line-level embeddings into stanza- and poem-level representations may yield more coherent predictions.
- Generative & Assistive Applications: The emotion classifier can be embedded as a module within creative writing tools, suggesting emotionally congruent phrases or highlighting sections where the emotional tone diverges from the author's intent. This aligns with emerging work on controllable text generation and affect-aware language modeling.
- Human-in-the-Loop and Educational Use: Integrating the model into interactive platforms for teaching literature could support students in exploring alternative emotional interpretations, while expert feedback can be used to iteratively refine the classifier.

IX. CONCLUSION

This paper presented a transformer-based approach for classifying emotional states in English poetry. By fine-tuning BERT on a curated poetry dataset and using a compact classification head, the system achieves near 90% accuracy and a weighted F1-score of 89.72%. The results demonstrate that contextual embeddings can effectively capture the nuanced emotional signals that emerge from metaphor, imagery, and atypical syntax in poetic text.

At the same time, our analysis highlights persistent challenges: annotation subjectivity, class overlap, cultural variation, and limited data scale. Addressing these challenges will require richer corpora, more expressive label schemes, and closer collaboration between NLP researchers and literary scholars. Nonetheless, the present work underscores the feasibility and potential of transformer-based models as tools for computational literary analysis and as components in future generative and assistive writing systems.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[2] B. Liu, *Sentiment Analysis and Opinion Mining*.Morgan & Claypool, 2012.

[3] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[4] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[5] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014.

[6] M. E. Peters *et al.*, "Deep contextualized word representations," in *Proc. NAACL-HLT*, 2018.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[8] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, 2020.

[10] S. Mohammad and P. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[11] Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proc. SemEval*, 2007.

[12] A. M. Jacobs, "Towards a neurocognitive poetics model of literary reading," in *Towards a Cognitive NeuroScience of Natural Language*, 2015.

[13] J. T. Kao and D. Jurafsky, "A computational analysis of poetic style: Imagery and sound in contemporary poetry," in *Proc. ACL*, 2014.

[14] Demszky *et al.*, "GoEmotions: A dataset of fine-grained emotions," in *Proc. ACL*, 2020.

[15] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[16] Z. Lan *et al.*, "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[17] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[18] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. ACL*, 2020.