StructFormer: Structure-Aware Hyperbolic Representation for Coarse-to-Fine Emotion Classification in Lyrics

Yutong Hu
University of Amsterdam
yutong.hu@student.uva.nl

Reza Mohammadi
University of Amsterdam
a.mohammadi@uva.nl

Menglin Yang
HKUST (GZ)
menglin.yang@outlook.com

Abstract

Song lyrics possess a natural hierarchical structure that remains unexploited in computational models, creating a significant gap in emotion recognition systems. We introduce **StructFormer**, a novel framework that leverages the symbolic structure of lyrics and the hierarchical nature of emotions by encoding them in hyperbolic space. Our approach incorporates paragraph-line structure as an inductive bias and employs a multi-level supervision strategy with both fine-grained and coarse-grained labels. The key innovations include three theoretically-grounded components: (1) a structure-aware embedding module that fuses semantic and structural information through computationally efficient gated alignment, (2) a hyperbolic projection that captures hierarchical relationships among emotion labels with mathematical guarantees, and (3) geometric consistency losses that enforce coherence between structural segmentation and emotional representation. Extensive experimental results demonstrate that StructFormer achieves substantially improved embedding coherence while maintaining competitive classification performance across diverse emotion categories compared to state-of-the-art baselines.

1 Introduction

Emotion recognition from lyrics serves as a critical complement to acoustic features, enhancing applications from personalized music recommendation to creative AI systems (Revathy et al., 2023; Ma et al., 2021). Lyrics present unique analytical challenges due to their inherent hierarchical composition (song → paragraph → line) and emotional progression. Despite these structural properties, conventional Music Emotion Recognition (MER) models predominantly process lyrics as flat sequences, neglecting critical organizational cues (Yin et al., 2021; Zhang and McAuley, 2020). While structural modeling approaches—including hierarchical attention mechanisms (Yang et al., 2016) and graph-based encoders (Zhang and Yu, 2021)—have advanced sentiment analysis in general NLP, they remain underutilized in lyrical analysis. Similarly, emotion taxonomies naturally form hierarchical relationships (e.g., *amusement*, *pride* → *joy*) (Goel et al., 2022), yet current MER systems rarely leverage these inherent structures.

From a geometric perspective, Euclidean spaces face fundamental limitations when embedding hierarchical structures due to their fixed zero curvature. Bourgain's theorem (Bourgain, 1985) demonstrates that trees cannot be embedded into Euclidean space with arbitrarily low distortion, regardless of dimension, with complete binary trees requiring at least $\Omega(\sqrt{\log n})$ distortion. In contrast, hyperbolic spaces can embed any tree into a 2-dimensional Poincaré disk with arbitrarily low distortion $(1+\varepsilon)$ for any $\varepsilon > 0$. While recent work applies hyperbolic embeddings to label representations (Peng et al., 2023), no approach integrates structural priors from input text with suitable non-Euclidean geometry.

We introduce **StructFormer**, the first unified framework that jointly encodes paragraph-line structure and emotion hierarchies in hyperbolic space. As shown in Figure 1, our model fuses semantic and structural features via gated alignment, projects them into hyperbolic space using Möbius operations, and optimizes a multi-objective function enforcing classification accuracy and geometric consistency. Extensive Experiments show that StructFormer yields superior embedding coherence and competitive performance against strong baselines, validating our theoretical contributions. Our work makes three key contributions to emotion recognition in lyrics:

- **Structure-Aware Embedding**: A gated fusion mechanism that adaptively integrates se-

 Accepted at the KnowFM Workshop @ ACL 2025.

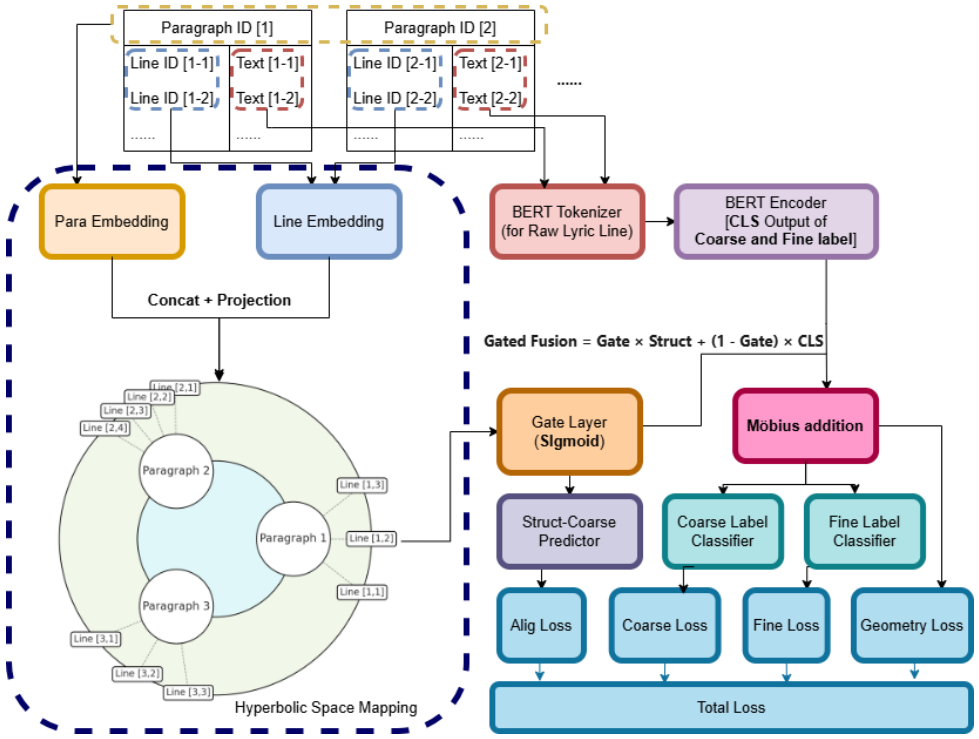

Figure 1: Overview of StructFormer. The model integrates structured embeddings (paragraph and line indices), semantic representations (BERT CLS), and non-Euclidean projection (Mobius addition) for fine- and coarse-grained emotion recognition. Gated fusion and structural losses align semantic and structural space.

mantic content with structural information (paragraph-line positions) for the first time in lyrical emotion modeling.

- **Hyperbolic Geometric Framework**: A unified projection mechanism that captures hierarchical emotion relationships with mathematical guarantees, overcoming Euclidean space limitations for tree structures.

- **Geometric Consistency Losses**: Novel loss functions that enforce coherence between structural segmentation and emotional representation through theoretically-grounded distance constraints.

2 Related Work

Music Emotion Recognition. Early MER methods relied on acoustic features with statistical classifiers (Kim et al., 2010; Yang and Chen, 2012), while recent approaches incorporate lyrics (Yin et al., 2021) or multimodal fusion (Zhang and McAuley, 2020; Huang and Yang, 2020). However, most treat lyrics as flat sequences and model emotions independently, overlooking the hierarchical relationships among emotion categories (e.g., *joy* → *pride*, *amusement*).

Structure-Aware Emotion Modeling. Structural modeling has advanced sentiment analysis via recursive networks (Socher et al., 2013), hierarchical attention (Yang et al., 2016), and tree-structured encoders (Tai et al., 2015), yet remains underexplored in lyrics despite inherent paragraph-line structure (Malheiro et al., 2021). Models like HiBERT (Zhang et al., 2019) have not been adapted to jointly leverage structure and non-Euclidean geometry for emotion representation.

Hyperbolic Emotion Embedding. Hyperbolic spaces provide exponential capacity for modeling scale-free and hierarchical structured data (Nickel and Kiela, 2017; Chami et al., 2019; Yang et al., 2022a), with applications in taxonomies (Dhingra et al., 2020), sentence embeddings (Ganea et al., 2018), recommender systems (Sun et al., 2021; Yang et al., 2022b,b; Chen et al., 2022), and knowledge graphs (Balazevic et al., 2019). Recent emotion work embeds label hierarchies (Peng et al., 2023) but omits structural input. Our model fills this gap by jointly encoding emotional and structural hierarchies within a unified hyperbolic framework.

Fine-Coarse Label Supervision. Hierarchical emotion taxonomies like Plutchik's wheel and

Ekman's basic emotions provide valuable frameworks for emotion classification (Cowen and Keltner, 2021). Recent approaches have leveraged these hierarchies through multitask learning, simultaneously predicting fine and coarse emotion labels to improve both interpretability and performance (Goel et al., 2022; Poria et al., 2018). While effective, these models typically overlook the interaction between label hierarchies and input structure, relying instead on standard Euclidean representations without explicit geometric constraints to model hierarchical relationships.

Research Gaps. Prior music emotion recognition research, despite extensive work on audio-text fusion and label hierarchies, has three key limitations (Araño et al., 2021; Schmeier et al., 2019; Raboy and Taparugssanagorn, 2024): lyrics are rarely treated as structured text despite clear paragraph-line organization; emotion taxonomies are often flattened rather than hierarchically modeled; and standard Euclidean representations fail to capture the natural hierarchical properties of both emotions and lyrical structure. Our work introduces StructFormer to address these gaps by unifying structural encoding, hierarchical supervision, and hyperbolic representations within a single coherent framework.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i, p_i, l_i, y_i^{\text{fine}}, y_i^{\text{coarse}})\}_{i=1}^{N}$ denote a dataset of N lyric lines, where x_i represents the textual content, p_i and l_i indicate structural positional information (paragraph and line indices respectively), and $y_i^{\text{fine}} \in \mathcal{Y}_{\text{fine}}$, $y_i^{\text{coarse}} \in \mathcal{Y}_{\text{coarse}}$ are hierarchical emotion labels at fine-grained and coarse-grained levels. Our objective is to learn a representation function that maps each lyric line to a vector $h_i \in \mathbb{R}^d$ that effectively captures both semantic content and structural context, supporting emotion classification in both Euclidean and hyperbolic geometric spaces.

3.2 Overall Framework

We introduce **StructFormer**, a framework that addresses the limitations of existing approaches through three key innovations: (1) a structure-aware embedding module with gated fusion, (2) a hyperbolic projection mechanism with mathematical guarantees for hierarchical relationships, and (3) geometric consistency losses that enforce coherence between structural segmentation and emotional representation. For each lyric line x_i, we extract contextual features using BERT to obtain a semantic representation h_{cls}, encode structural information through position embeddings, and adaptively fuse these signals before projecting them into hyperbolic space for classification.

3.3 Structure-Aware Embedding

Our first key innovation is a computationally efficient structure-aware embedding module that captures positional context while maintaining semantic richness. The semantic content of each lyric line is first encoded using a pretrained BERT model:

$$h_{\text{cls}} = \text{BERT}(x) \in \mathbb{R}^d. \tag{1}$$

While this captures linguistic features, it lacks structural context. To address this limitation, we encode structural information through a dedicated embedding pathway:

$$h_{\text{struct}} = \text{MLP}_{\text{struct}}([\mathbf{e}_p; \mathbf{e}_l]) \in \mathbb{R}^d, \tag{2}$$

where $\mathbf{e}_p \in \mathbb{R}^{d_s}$ and $\mathbf{e}_l \in \mathbb{R}^{d_s}$ are learnable embeddings for paragraph and line positions with dimension $d_s \ll d$ for parameter efficiency, $[\cdot;\cdot]$ denotes vector concatenation, and $\text{MLP}_{\text{struct}}$ is a multi-layer perceptron with two linear transformations and a ReLU activation:

$$\text{MLP}(x) = W_2\sigma(W_1 x + b_1) + b_2, \tag{3}$$

where $W_1 \in \mathbb{R}^{d_h \times 2d_s}$, $W_2 \in \mathbb{R}^{d \times d_h}$, $b_1 \in \mathbb{R}^{d_h}$, and $b_2 \in \mathbb{R}^d$ are learnable parameters.

To adaptively integrate these complementary signals, we employ a content-dependent gating mechanism:

$$g = \sigma(\text{MLP}_g(h_{\text{struct}})), \tag{4a}$$

$$h = g \odot h_{\text{struct}} + (1 - g) \odot h_{\text{cls}}. \tag{4b}$$

Here, σ is the sigmoid function, and $\odot$ denotes element-wise multiplication. This computationally efficient gated alignment allows the model to dynamically adjust the influence of structural information based on content, avoiding the need for more complex attention mechanisms.

3.4 Hyperbolic Projection and Classification

Our second key innovation is a hyperbolic projection mechanism that provides mathematical guarantees for representing hierarchical relationships.

A fundamental challenge in modeling hierarchical structures in Euclidean space is the exponential growth in volume required to maintain distance relationships. To address this limitation, we project the fused representation into the Poincaré ball model $\mathbb{H}^d$:

$$\tilde{h} = \exp_0(h) \in \mathbb{H}^d, \tag{5}$$

where the exponential map $\exp_0$ transforms vectors from Euclidean space to the hyperbolic manifold $\mathbb{H}^d = \{x \in \mathbb{R}^d : \|x\| < 1\}$. This guarantees exponential capacity for representing hierarchies with theoretical embedding distortion bounds.

For classification, we employ a nearest-prototype approach:

$$\hat{y}_{\text{coarse}} = \arg\min_c \; d_{\mathbb{H}}(\tilde{h}, \mu_c), \tag{6}$$

where $\mu_c \in \mathbb{H}^d$ is a learnable prototype and $d_{\mathbb{H}}$ is the Poincaré distance:

$$d_{\mathbb{H}}(u, v) = \text{arcosh}\left(1 + 2\frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)}\right). \tag{7}$$

3.5 Supervision and Loss Functions

Our third key innovation is a set of geometric consistency losses that enforce coherence between structural segmentation and emotional representation. Our training objective integrates four complementary components:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{fine}}\mathcal{L}_{\text{fine}} + \lambda_{\text{coarse}}\mathcal{L}_{\text{coarse}} \\ + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{geom}}\mathcal{L}_{\text{geom}},\end{aligned} \tag{8}$$

where the weights are set to $\lambda_{\text{fine}} = 1.0$, $\lambda_{\text{coarse}} = 0.5$, $\lambda_{\text{align}} = 0.4$, and $\lambda_{\text{geom}} = 0.05$.

For emotion classification, we employ standard cross-entropy losses:

$$\mathcal{L}_{\text{fine}} = -\sum_k \mathbb{I}(y^{\text{fine}} = k) \log \hat{p}_k^{\text{fine}}, \tag{9a}$$

$$\mathcal{L}_{\text{coarse}} = -\sum_j \mathbb{I}(y^{\text{coarse}} = j) \log \hat{p}_j^{\text{coarse}}, \tag{9b}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\hat{p}$ are softmax-normalized probabilities.

The alignment loss ensures structural representations remain semantically meaningful:

$$\mathcal{L}_{\text{align}} = \mathbb{E}\left[(1 - \cos(h, h_{\text{struct}}))^2\right]. \tag{10}$$

The curvature is set to -1.

Our novel geometric consistency loss explicitly enforces structural relationships in both spaces:

$$\mathcal{L}_{\text{intra}} = \sum_p \log\left(1 + \bar{d}_{\mathbb{H}}^{(p)}\right) - \log\left(1 + \bar{d}_{\mathbb{E}}^{(p)}\right) \tag{11a}$$

$$\mathcal{L}_{\text{inter}} = \sum_{(i,j)} \mathcal{L}_{\text{inter}}^{(i,j)}, \quad \text{where}$$

$$\mathcal{L}_{\text{inter}}^{(i,j)} = \begin{cases} \max\left(0,\ d_{\mathbb{H}}(i,j) - d_{\mathbb{E}}(i,j) + \delta\right), \\ \quad \text{if } |i-j| = 1 \\ \max\left(0,\ d_{\mathbb{E}}(i,j) - d_{\mathbb{H}}(i,j) + \delta\right) \\ \quad \text{if } |i-j|. \geq 2 \end{cases} \tag{11b}$$

Here, $\bar{d}^{(p)}$ represents average within-paragraph distances and $d(p_i, p_j)$ denotes paragraph centroid distances. This loss enforces a crucial structural-emotional coherence property: adjacent paragraphs (likely with similar emotions) should be closer in hyperbolic space than in Euclidean space, while non-adjacent paragraphs should be further apart. The complete geometric loss is:

$$\mathcal{L}_{\text{geom}} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{inter}}. \tag{12}$$

This theoretically-grounded approach ensures that the geometric properties of our representation space align with the inherent hierarchical and structural properties of both lyrical organization and emotion taxonomies.

4 Experiments

4.1 Dataset Preprocessing

We construct our dataset from the DALI corpus (Meseguer-Brocal et al., 2020), extracting lyric lines with paragraph and line indices. Emotion labels are assigned using the `monologg/bert-base-cased-goemotions` model (Park, 2022), yielding 26 fine-grained emotions mapped to 7 coarse categories and 3 sentiment polarities. To balance the dataset, we include all songs for rare emotions (¡50 examples) and sample up to 50 songs for common ones, resulting in 1,041 songs, 10,001 paragraphs, and 73,796 lyric lines with complete hierarchical annotation.

4.2 Experimental Setup

We evaluate StructFormer-Hyper against three baselines (BERT, Multi-task BERT, HAN) under controlled conditions. Models are trained with batch

Table 1: Fine-to-Coarse Label Mapping with Sentiment Polarity

Coarse Category	Fine Labels	Sentiment Polarity
Joy	amusement, pride, excitement, relief, optimism	Positive
Sadness	sadness, grief, disappointment, remorse	Negative
Anger	anger, annoyance, disapproval	Negative
Fear	fear, embarrassment, nervousness	Negative
Surprise	surprise, realization, confusion	Neutral
Love	love, gratitude, desire	Positive
Neutral	neutral, curiosity, approval	Neutral

size 16, learning rate 1×10^{-5}, and dropout 0.1 (seed=42). StructFormer-Hyper uses Riemannian Adam (Klimenko et al., 2020) for hyperbolic optimization, while baselines use standard Adam. The dataset contains 1,041 DALI songs with 20/10/70% train/validation/test splits. Structural embedding dimension is 8. Ablation studies use identical hyperparameters on 300 randomly selected songs for 10 epochs. Experiments ran on an NVIDIA RTX 4090 GPU (24GB VRAM) and Intel Xeon Gold 6430 (16 vCPUs, 120GB RAM), requiring approximately 8 hours of training.

Model Comparisons. We evaluate four models: (1) **StructFormer-Hyper** (ours) encodes paragraph-line structure with BERT embeddings via gated fusion, projects into hyperbolic space, and employs four specialized losses; (2) **BERT** (Devlin et al., 2019) uses only the [CLS] token without structural awareness; (3) **Multi-task BERT** (Goel et al., 2022) adds coarse-level classification but lacks structural modeling; and (4) **HAN** (Yang et al., 2016), adapted for lyrics, uses hierarchical attention without hyperbolic geometry.

4.3 Evaluation Metrics

Classification Performance Metrics. We employ multiple F1-score variants to comprehensively evaluate classification performance across different granularities and perspectives. The **weighted F1-score** measures overall classification effectiveness across all 26 fine-grained emotion labels, with automatic adjustment for class imbalance through sample-weighted averaging. This metric provides a holistic view of model performance while accounting for the skewed distribution inherent in emotion datasets. The **sentiment F1-score** evaluates performance at a higher abstraction level by aggregating results over three sentiment polarities (positive, negative, neutral) derived from our coarse-grained emotion groupings, enabling comparison with sentiment analysis baselines. Finally, the **macro F1-score per coarse category** provides detailed breakdowns of average performance within each of the seven coarse emotion categories (joy, sadness, anger, fear, surprise, love, neutral), offering insights into model consistency for fine-grained classification within broader emotional groups.

Representation Quality Assessment. To evaluate the geometric properties and structural coherence of learned representations, we compute the **silhouette coefficient** over the latent embeddings of lyric lines. This metric quantifies how well-separated and internally cohesive the emotional clusters are in the learned embedding space, with scores ranging from -1 (poor clustering) to +1 (well-separated clusters). The silhouette score serves as a crucial indicator of whether our hyperbolic geometric approach successfully captures meaningful emotional distinctions and hierarchical relationships, complementing classification accuracy with an assessment of representational quality that is independent of supervised labels.

4.4 Comparison and Evaluation

We evaluate StructFormer-Hyper against three strong baselines (BERT, Multi-task BERT, and HAN) on our curated dataset comprising 73,796 lyric lines across 1,041 songs, following standard evaluation protocols.

Classification Performance. Table 2 shows that all transformer-based models achieve high weighted F1 scores (≥ 0.91), with our model showing particular strength in challenging negative and neutral sentiment categories. Despite compara-

Table 2: Main experimental results with emotion polarity and coarse label breakdown. Silhouette score marked with * indicates statistical significance ($p < 0.1$) compared to other models using the Wilcoxon signed-rank test.

Model	Sentiment	Weighted F1	Sentiment F1	Macro F1 per Label	Silhouette
StructFormer-Hyper	positive		0.88	joy: 0.19, love: 0.25	
	negative	**0.96**	0.95	sadness: 0.13, anger: 0.22, fear: 0.16	**0.77***
	neutral		**0.98**	surprise: 0.28, neutral: **0.12**	
BERT	positive		**0.92**	joy: **0.25**, love: **0.28**	
	negative	**0.96**	**0.97**	sadness: **0.18**, anger: **0.25**, fear: **0.25**	0.55
	neutral		**0.98**	surprise: **0.30**, neutral: **0.12**	
Multi-task BERT	positive		0.90	joy: 0.23, love: 0.26	
	negative	**0.96**	0.96	sadness: 0.17, anger: 0.24, fear: 0.23	0.60
	neutral		**0.98**	surprise: 0.28, neutral: 0.11	
HAN	positive		0.66	joy: 0.15, love: 0.18	
	negative	0.91	0.86	sadness: 0.10, anger: 0.20, fear: 0.12	-0.09
	neutral		0.96	surprise: 0.19, neutral: 0.09	

ble overall metrics, baselines struggle with fine-grained emotion distinctions—especially for low-resource categories like *fear* and *anger*. This validates our theoretical assumption that hierarchical structure modeling is critical for capturing nuanced emotional distinctions that flat sequence representations fundamentally miss.

Representational Coherence. The significant improvement in silhouette score (0.77 vs baseline's 0.55, $p < 0.1$) confirms StructFormer-Hyper's superior embedding organization and our hyperbolic approach's advantage for emotion hierarchies. HAN's negative score (-0.09) reveals that structural information alone is insufficient without appropriate geometric constraints, demonstrating the essential synergy between structure and geometry.

Embedding Visualization. The UMAP plots in Figure 2 reveal striking differences in emotional representation quality. StructFormer-Hyper produces distinctly separated clusters with clear boundaries between semantically related emotions—especially evident in positive categories like *joy* and *love*—demonstrating its superior capacity to capture emotional hierarchies. In contrast, baselines exhibit either densely collapsed representations (BERT), diffuse boundaries (Multi-task BERT), or fragmented clusters (HAN) with significantly less discriminative structure. These visualizations validate our quantitative silhouette metrics and highlight the fundamental representational advantages of our hyperbolic geometric approach.

Loss Curve Analysis. Figure 3 shows that all models exhibit stable convergence within 20 epochs. StructFormer-Hyper reduces its composite loss steadily without abrupt plateaus, indicating effective multi-objective optimization. As loss magnitudes are not directly comparable across models, we focus on their respective convergence behaviors.

4.5 Ablation Study

To verify that our performance gains are attributed to our theoretical innovations rather than implementation details, we conducted systematic ablation experiments on 300 randomly selected songs, isolating each component's contribution.

Component Analysis. Table 3 reveals critical insights into our model's architecture. While StructFormer-NoHyper achieves marginally higher classification accuracy (0.95 F1), it suffers from significantly degraded structural coherence, empirically confirming hyperbolic geometry's essential role in capturing hierarchical relationships.
Interestingly, StructFormer-NoProj shows strong silhouette scores (0.33) but at the cost of classification performance. The severe performance degradation of StructFormer-NoCoarse (silhouette -0.13) demonstrates the critical importance of hierarchical supervision signals. These results validate our theoretical premise that effective emotional representation requires both appropriate geometry and explicit structural guidance.

Representation Quality. Figure 4 reveals how our model components affect emotion mappings. StructFormer-Hyper produces sharp, discriminative alignments while variants exhibit increasingly

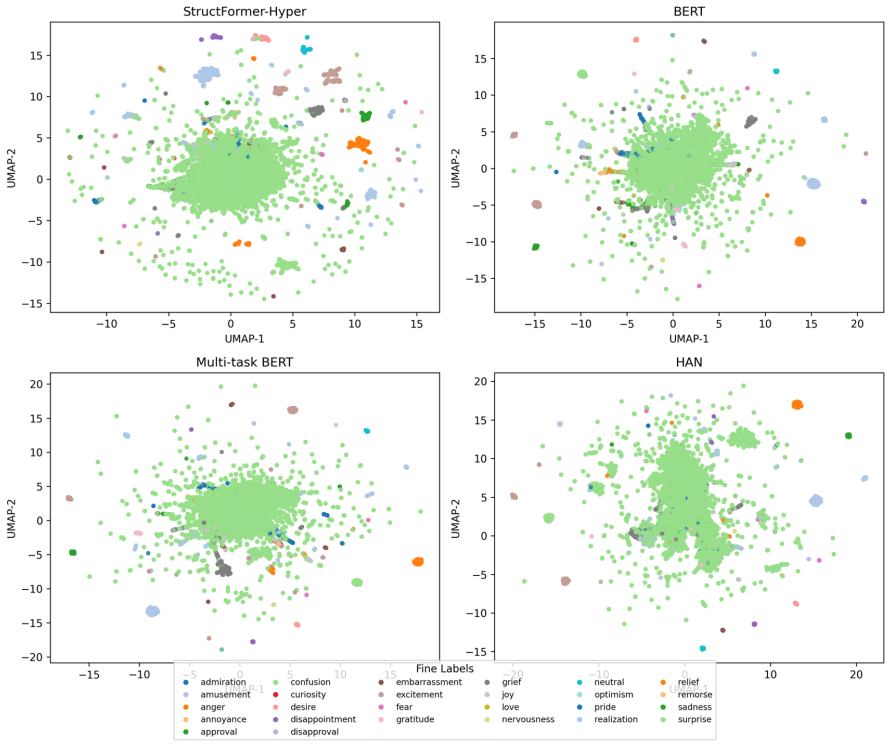

Figure 2: UMAP visualization of embedding spaces for all models, colored by fine-grained emotion labels. StructFormer-Hyper (top-left) shows more distinct clustering with clearer boundaries between emotion categories compared to baseline models, particularly for semantically related emotions.

Figure 3: Training loss trajectories over 20 epochs on a log scale. All models show stable convergence, though absolute loss values are not directly comparable due to differing loss formulations and objectives. StructFormer-Hyper exhibits consistent reduction despite optimizing multiple competing objectives.

Table 3: Ablation study results on StructFormer variants. No statistics reported as results reflect architectural variants, not independent models.

Model	Weighted-F1	Silhouette
StructFormer-Hyper	0.92	0.19
StructFormer-NoStruct	0.90	**0.32**
StructFormer-NoHyper	**0.95**	0.16
StructFormer-NoProj	0.91	0.33
StructFormer-NoCoarse	0.90	-0.13

diffuse, cross-category activations. NoHyper particularly struggles with sadness/anger distinctions, empirically confirming the necessity of both geometric and structural encoding for effective emotional representation.

5 Conclusion and Discussion

We presented **StructFormer**, the first unified architecture that successfully integrates structural hierarchy with hyperbolic geometry for emotion recognition in lyrics. By embedding both paragraph-line structure and emotional taxonomies within non-Euclidean space, our model achieves improved representational coherence while maintaining competitive classification performance. Despite its advantages, StructFormer faces limitations. The model's emphasis on structural coherence may constrain its ability to capture abrupt emotional transitions that cross structural boundaries.

Key future directions include extending StructFormer to multimodal settings with synchronized audio-lyric data, exploring temporal emotion dynamics across song progressions, and applying our hyperbolic structural framework to other domains with nested hierarchical properties. Investigating

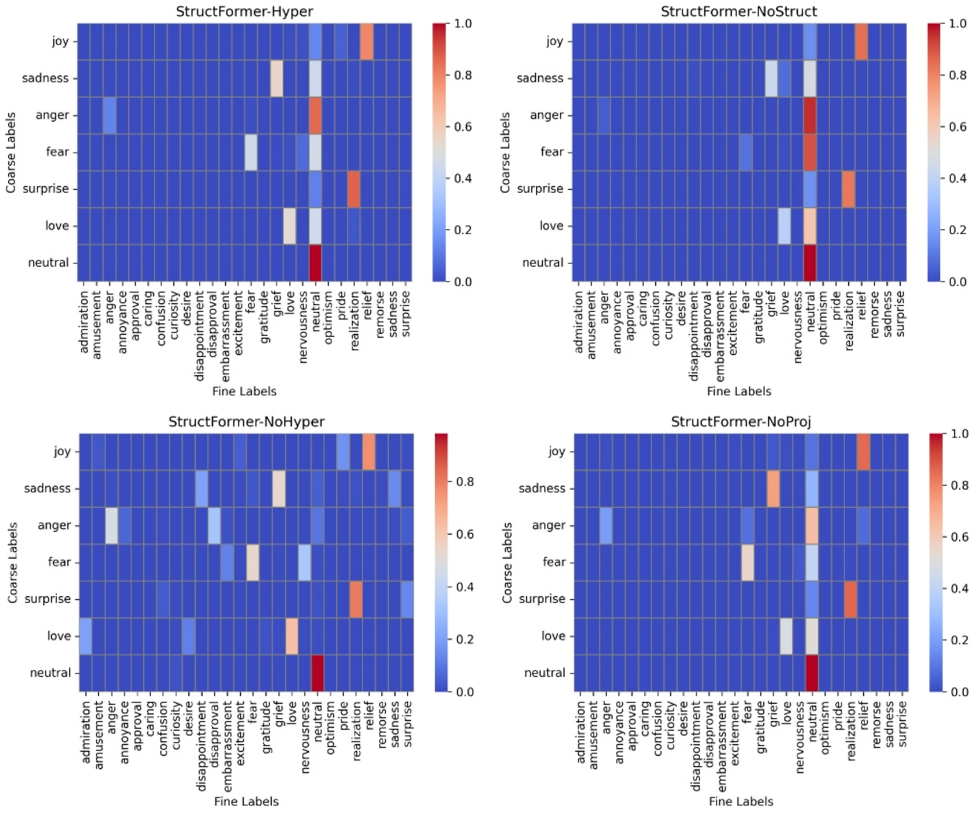

Figure 4: Heatmaps of fine-to-coarse label alignment across StructFormer variants. StructFormer-Hyper exhibits clearer hierarchical mappings, while the removal of structural or projection components leads to more diffuse alignments.

more efficient hyperbolic optimization techniques could mitigate computational costs and broaden applicability.

Acknowledgements

This paper was accepted at the KnowFM Workshop @ ACL 2025. The authors would like to thank the anonymous reviewers for their valuable feedback. We also acknowledge the open-source communities behind Hugging Face Transformers and the DALI dataset, which made this research possible. This work did not receive any specific funding. The authors declare no competing interests.

References

Keith April Araño, Carlotta Orsenigo, Mauricio Soto, and Carlo Vercellis. 2021. Multimodal sentiment and emotion recognition in hyperbolic space. *Expert Systems with Applications*, 184:115507.

Ivana Balazevic, Carl Allen, and Timothy M Hospedales. 2019. Multi-relational poincaré graph embeddings. In *NeurIPS*.

Jean Bourgain. 1985. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. *NeurIPS*, 32.

Yankai Chen, Menglin Yang, Yingxue Zhang, Mengchen Zhao, Ziqiao Meng, Jianye Hao, and Irwin King. 2022. Modeling scale-free graphs with hyperbolic geometry for knowledge-aware recommendation. In *WSDM*, pages 94–102.

Alan S Cowen and Dacher Keltner. 2021. The structure of emotions reflects both cultural and universal factors. In *Nature Human Behaviour*, volume 5, pages 909–920.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Bhuwan Dhingra, Zhengbao Zhou, Nicholas FitzGerald, Theodore Muehlbauer, and William W Cohen. 2020. Embedding text in hyperbolic spaces. In *ACL*.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*.

Kunal Goel, Diyi Yang, Charles L Isbell, and Honglak Lee. 2022. Fine-grained emotion recognition with multitask learning. In *EMNLP*.

Yu-Siang Huang and Yi-Hsuan Yang. 2020. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *ACM MM*.

Youngmoo E Kim, Erik M Schmidt, Robert Migneco, Brian G Morton, Jason Richardson, Jeffrey Scott, James A Speck, Douglas Tindal, Douglas Turnbull, and Kristjan Williams. 2010. Music emotion recognition: A state of the art review. In *ISMIR*.

Artem Klimenko, Maxim Kochurov, Vitaly Vlasov, and Mikhail Belyaev. 2020. Geoopt: Riemannian optimization in pytorch. https://github.com/geoopt/geoopt. Accessed: 2025-05-16.

Xichu Ma, Ye Wang, Min-Yen Kan, and Wee Sun Lee. 2021. Ai-lyricist: Generating music and vocabulary constrained lyrics. In *ACM MM*, pages 1002–1011.

Ricardo Malheiro, Tiago Pimentel, Emílio Rodrigues, and Rui Paiva. 2021. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *Computer Speech & Language*, volume 66.

Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2020. Creating dali, a large dataset of synchronized audio, lyrics, and notes. *Transactions of the International Society for Music Information Retrieval*, 3(1).

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*.

Jihoon Park. 2022. monologg/bert-base-cased-goemotions. https://huggingface.co/monologg/bert-base-cased-goemotions. Accessed: 2025-05-11.

Yi Peng, Ruotian Wang, Ting Liu, Maosong Sun, and Yangqiu Song. 2023. Label-aware hyperbolic embeddings for fine-grained emotion classification. In *ACL*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Prateek Vij, Navonil Majumder, and Rada Mihalcea. 2018. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *EMNLP*.

Love Jhoye Moreno Raboy and Attaphongse Taparugssanagorn. 2024. Verse1-chorus-verse2 structure: A stacked ensemble approach for enhanced music emotion recognition. *Applied Sciences*, 14(13):5761.

VR Revathy, Anitha S Pillai, and Fatemah Daneshfar. 2023. Lyemobert: Classification of lyrics' emotion and recommendation using a pre-trained model. *Procedia Computer Science*, 218:1196–1208.

Timothy Schmeier, Joeseph Chisari, Sam Garrett, and Brett Vintch. 2019. Music recommendations in hyperbolic space: an application of empirical bayes and hierarchical poincaré embeddings. In *RecSys*, pages 437–441.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Jianing Sun, Zhaoyue Cheng, Saba Zuberi, Felipe Pérez, and Maksims Volkovs. 2021. Hgcf: Hyperbolic graph convolution networks for collaborative filtering. In *WWW*, pages 593–601.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*.

Menglin Yang, Min Zhou, Zhihao Li, Jiahong Liu, Lujia Pan, Hui Xiong, and Irwin King. 2022a. Hyperbolic graph neural networks: A review of methods and applications. *arXiv preprint arXiv:2202.13852*.

Menglin Yang, Min Zhou, Jiahong Liu, Defu Lian, and Irwin King. 2022b. Hrcf: Enhancing collaborative filtering via hyperbolic geometric regularization. In *WWW*, pages 2462–2471.

Yi-Hsuan Yang and Homer H Chen. 2012. Machine recognition of music emotion: A review. In *TIST*, volume 3. ACM.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Ruiqi Yin, Yizhen Cai, and Julian McAuley. 2021. Emotionally-relevant features for classification and regression of music lyrics. In *ICASSP*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2019. Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *ACL*.

Lichao Zhang and Dong Yu. 2021. Hyperbolic graph convolutional networks for text classification. In *AAAI*.

Yasheng Zhang and Julian McAuley. 2020. Moodplay: Interactive mood-based music discovery and recommendation. In *WWW*.