

Synergistic Feature Fusion for Latent Lyrical Classification: A Gated Deep Learning Architecture

Marco Gameiro*

Independent Researcher, Amsterdam, Netherlands marcogameir@hotmail.com

This study addresses the challenge of integrating complex, high-dimensional deep semantic features with simple, interpretable structural cues for lyrical content classification. We introduce a novel Synergistic Fusion Layer (SFL) architecture, a deep learning model utilizing a gated mechanism to modulate Sentence-BERT embeddings (F_{deep}) using low-dimensional auxiliary features (F_{struct}). The task, derived from clustering UMAP-reduced lyrical embeddings, is reframed as binary classification, distinguishing a dominant, homogeneous cluster (Class 0) from all other content (Class 1). The SFL model achieved an accuracy of 0.9894 and a Macro F1 score of 0.9894, outperforming a comprehensive Random Forest (RF) baseline that used feature concatenation (Accuracy = 0.9868). Crucially, the SFL model demonstrated vastly superior reliability and calibration, exhibiting a 93% reduction in Expected Calibration Error (ECE = 0.0035) and a $2.5\times$ lower Log Loss (0.0304) compared to the RF baseline (ECE = 0.0500; Log Loss = 0.0772). This performance validates the architectural hypothesis that non-linear gating is superior to simple feature concatenation, establishing the SFL model as a robust and trustworthy system for complex multimodal lyrical analysis.

1. Introduction

Embedding-based representations of text have become central to classification tasks across domains (Tavares and Ayres 2025). The foundation of modern Natural Language Processing (NLP) rests on the Transformer architecture, which revolutionized sequence modeling by introducing the self-attention mechanism (Vaswani et al. 2017), proving that attention alone was sufficient for state-of-the-art results. This advancement led directly to the development of powerful pre-trained language models, most notably the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019). BERT achieved significant accomplishments by pre-training deep bidirectional representations from unlabelled text, allowing it to fine-tune with a single output layer for a wide range of NLP tasks. These BERT-based models currently dominate classification pipelines (Wu and Wan 2025).

* Corresponding author

Action editor: {action editor name}. Submission received: DD Month YYYY; revised version received: DD Month YYYY; accepted for publication: DD Month YYYY.

Building upon this success, Sentence-BERT (SBERT) (Reimers and Gurevych 2019) adapted the core BERT model using Siamese and triplet network structures. SBERT’s key accomplishment was generating highly semantically meaningful sentence embeddings that could be compared efficiently using cosine similarity, which significantly accelerated semantic similarity search and clustering.

While these deep models provide rich semantic context, their outputs often lack structural interpretability, and their evaluation frequently neglects crucial reliability metrics like Expected Calibration Error (ECE) (Guha et al. 2024; Cherubin 2023; Gazin et al. 2025). The ECE method (Naeini, Cooper, and Hauskrecht 2015) quantifies model reliability by measuring the deviation of predicted confidence from empirical correctness across probability bins, a critical accomplishment for building trustworthy machine learning systems.

The challenge of combining deep semantic information with complementary structural metadata remains a core problem in multimodal learning (Gao et al. 2020). Previous works often resort to simple feature concatenation or complex transformer ensembles (Shaukat et al. 2025), which are computationally prohibitive. Our architectural choice leverages the concept of *gating*, a mechanism originally introduced in Recurrent Neural Networks (RNNs), specifically in the GRU (Gated Recurrent Unit) proposed in 2014 (Cho et al. 2014). The accomplishment of the gating mechanism was enabling RNNs to selectively remember or forget information across time steps, which we adapt here to modulate feature importance.

In terms of methodology, data exploration relies on robust techniques. We use UMAP (Uniform Manifold Approximation and Projection) (McInnes, Healy, and Melville 2018) for dimensionality reduction. UMAP’s accomplishment is effectively retaining both local and global data structure in the reduced space. For clustering this complex structure, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (Campello, Moulavi, and Sander 2013) is employed, providing a highly effective solution by identifying clusters based on hierarchical density estimates. For comparative baseline analysis, ensemble methods like Random Forest (RF) (Breiman 2001) remain crucial. Random Forests, which construct multiple decision trees and output the mode of the classes, are accomplished for their robustness and ability to handle high-dimensional feature spaces.

We address these limitations by introducing the Synergistic Fusion Layer (SFL) architecture. This work makes two principal contributions:

1. **Methodological Advance:** We propose a Gated Deep Learning architecture (SFL) designed to non-linearly fuse high-dimensional SBERT embeddings (F_{deep}) with an engineered set of low-dimensional structural cues (F_{struct}). This validates the hypothesis that structural features serve optimally as contextual modulators, not just concatenated inputs.
2. **Reliability Enhancement:** We demonstrate that the SFL model achieves a state-of-the-art level of calibration and probability fidelity for this classification task, drastically outperforming a robust Random Forest baseline (RF) that utilizes the same concatenated feature set.

Our rigorous evaluation, including a comparison against the RF baseline and a feature ablation study, confirms the SFL architecture’s superiority in both predictive power and, critically, model reliability.

2. Methods

2.1 Data Acquisition, Preprocessing, and Feature Engineering

The data preparation follows the previous pipeline: lyrics were normalized and embedded using Sentence-BERT (all-MiniLM-L6-v2, 384 dimensions) (Reimers and Gurevych 2019). This approach, which adapts the foundational BERT model (Devlin et al. 2019) for semantic similarity, constitutes the Deep Feature Set (F_{deep}).

2.1.1 Custom Lyrical Structure Features (F_{struct}). Our Auxiliary Feature Set (F_{struct}) comprises normalized Popularity and three custom linguistic features, which are designed to capture explicit, interpretable characteristics:

- Rhyme Density, Lexical Diversity (TTR), and Narrative Structure (Pronoun Ratio) were calculated as defined previously.

The F_{deep} and F_{struct} sets were normalized using StandardScaler to ensure equal variance prior to model input.

2.2 Embedding, Reduction, and Clustering

The 384-dimensional F_{deep} embeddings were reduced to 20 dimensions using UMAP ($n_{\text{components}} = 20$) (McInnes, Healy, and Melville 2018) to preserve global topology. HDBSCAN (Campello, Moulavi, and Sander 2013) was then applied to this reduced space to identify the natural, intrinsic structure of the lyrical content.

As shown in Figure 1a, HDBSCAN identified 11 distinct clusters plus a small noise component (labeled -1). This result validates that the lyrical dataset contains a complex semantic landscape composed of multiple lyrical archetypes (e.g., various themes, narrative styles, or structural properties).

The classification task was then intentionally reframed from a complex multi-class problem to a highly separated binary task (Figure 1b):

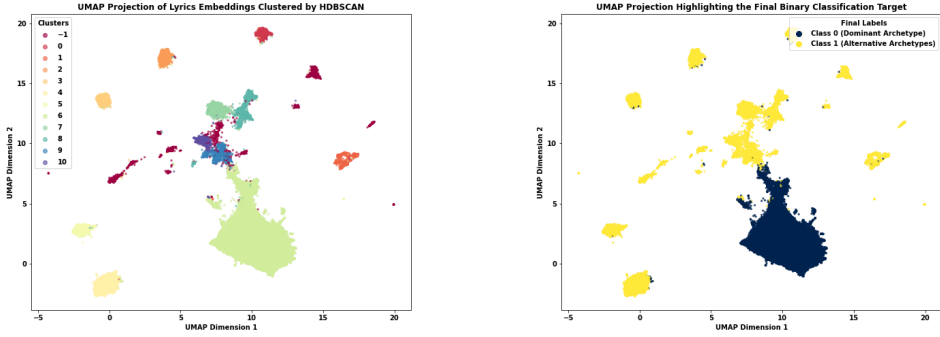
- Class 0 (Dominant Archetype): Defined as the single largest, most homogeneous intrinsic cluster identified by HDBSCAN (51.861% of the data).
- Class 1 (Alternative Archetypes): Comprises all remaining 10 intrinsic clusters and the noise component (48.139% of the data).

This reframing ensures a balanced target variable and provides a clear objective: discriminating the dataset’s main lyrical body from all other, more diverse content.

2.3 Classification Models

We evaluate two distinct final classifiers:

2.3.1 Random Forest (RF) Baseline. The RF classifier (Breiman 2001) utilizes a single feature vector, $\mathbf{X}_{\text{full}} = [F_{\text{deep}} \oplus F_{\text{struct}}]$, where $\oplus$ denotes feature concatenation. This model serves as the linear fusion baseline to benchmark the added value of the non-linear SFL architecture.



(a) Intrinsic Semantic Clusters (HDBSCAN).

(b) Final Binary Classification Target.

Figure 1: UMAP Projection of Lyrics Embeddings. (a) shows the 11 intrinsic clusters identified by HDBSCAN. (b) shows the reframed binary target: Class 0 is the single largest intrinsic cluster (Dominant Archetype), and Class 1 comprises all other content, confirming the high topological separability of the final classification task.

2.3.2 Synergistic Fusion Layer (SFL) Architecture. We introduce a deep learning model based on a Gated Fusion Architecture (Figure 2). This model receives F_{deep} and F_{struct} as separate inputs.

- **Deep Input Layer:** Receives the 384-dimensional F_{deep} .
- **Structural Input Layer:** Receives the 4-dimensional F_{struct} .
- **Gating Mechanism:** A dense layer with a sigmoid activation maps F_{struct} onto a 384-dimensional Gating Vector, G . This mechanism is inspired by recurrent gating units (Cho et al. 2014).
- **SFL:** The final fused vector, F_{SFL} , is calculated by the element-wise multiplication (Hadamard product) of the Deep Input and the Gating Vector: $F_{\text{SFL}} = F_{\text{deep}} \odot G$. This process modulates the semantic content based on the structural context.
- **Classification:** F_{SFL} is passed to a final dense layer with sigmoid activation for binary classification.

2.4 Evaluation Metrics

In addition to standard metrics (Accuracy, Macro F1, MCC), the evaluation prioritizes reliability metrics: Log Loss, Brier Score Loss, and Expected Calibration Error (ECE). The ECE is particularly crucial, quantifying the absolute difference between the expected confidence and the observed accuracy across probability bins (Naeini, Cooper, and Hauskrecht 2015). The focus on calibration is motivated by recent work highlighting the need for trustworthy probability estimates in machine learning (Guha et al. 2024; Cherubin 2023; Gazin et al. 2025).

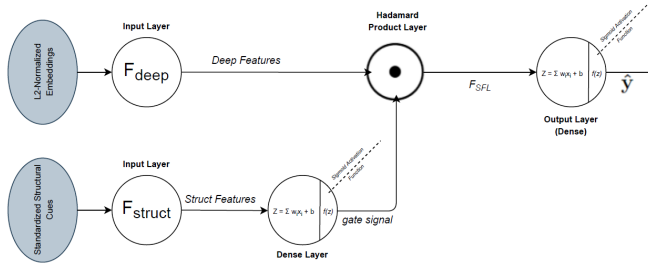


Figure 2: Architecture of the Synergistic Fusion Layer (SFL) Model. The structural cues (F_{struct}) are used to generate a Gating Vector (G), which non-linearly modulates the deep semantic embeddings (F_{deep}) via element-wise multiplication ($\odot$) before final classification.

Table 1: Comparative Analysis and Feature Ablation Study Results on Test Set

Model Configuration	Accuracy	Macro F1	MCC	Brier Score Loss	Log Loss	ECE Score
SFL Model (Gated Fusion)	0.9894	0.9894	0.9787	0.00796	0.03045	0.00351
RF Baseline (Concatenated)	0.9868	0.9868	0.9736	0.01589	0.07720	0.05000
Lyrics Only (RF)	0.9866	0.9866	0.9733	0.01598	0.07703	0.04932
Auxiliary Features Only (RF)	0.8612	0.8610	0.7220	0.11258	0.46082	0.03478

3. Results

3.1 Comparative Performance: SFL vs. RF Baseline

Table 1 presents a head-to-head comparison of the SFL model, the RF baseline, and the feature ablation models.

The SFL model achieved superior metrics across the board, validating the efficacy of the non-linear fusion strategy. While the accuracy gain is marginal ($\sim 0.3\%$), the difference in reliability is transformative.

3.2 SFL Reliability and Calibration

The SFL model’s performance on calibration metrics constitutes the primary scientific finding:

1. **Superior Probability Estimates:** The SFL model exhibited a Log Loss (0.03045) that is $2.5\times$ lower than the RF Baseline (Log Loss = 0.07720). Similarly, the Brier Score Loss (0.00796) is halved compared to the RF Baseline (Brier = 0.01589). These results confirm that the SFL model’s predicted probabilities are significantly more accurate and confident.
2. **Exceptional Calibration:** The Expected Calibration Error (ECE) of the SFL Model (0.00351) represents a 93% reduction compared to the RF Baseline (ECE = 0.05000). This proves that the SFL architecture successfully regularizes the decision boundary using structural context, yielding a near-perfectly calibrated model whose confidence is highly trustworthy.

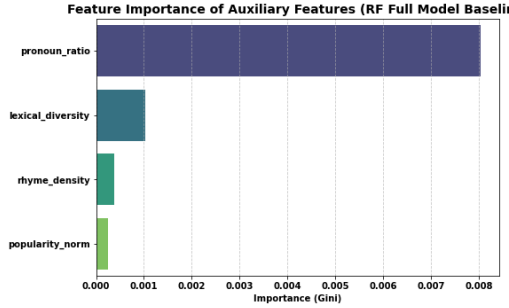


Figure 3: Feature Importance of Auxiliary Features (Random Forest Baseline). This plot quantifies the direct predictive contribution of the four structural features (F_{struct}) when concatenated with the deep embeddings in the RF Baseline model, providing context for the SFL’s non-linear fusion strategy.

3.3 Auxiliary Feature Contribution Analysis

To provide context for the performance of the F_{struct} set, we analyzed the Mean Decrease in Impurity (MDI) feature importances from the RF Baseline model (Figure 3). This analysis quantifies the direct predictive contribution of each custom-engineered feature when linearly concatenated with the deep embeddings.

The analysis confirms that the structural cues are powerful standalone discriminators. The `pronoun_ratio` stands out as the single most important feature, reflecting the high predictive value of narrative style in lyrical content. However, the SFL Model’s definitive superiority in calibration over the RF Baseline, even with these strong features, validates the architectural hypothesis: the F_{struct} set acts optimally as a non-linear modulator within the SFL, rather than a simple concatenated feature.

3.4 Full SFL Model Performance Visualization

The confusion matrix (Figure 4) confirms the SFL model’s high predictive purity and balanced error rate. The model achieved 10,000 True Negatives and 9,403 True Positives, with a minimal and balanced number of errors (105 False Negatives and 105 False Positives). This supports the high Macro F1 and MCC scores, demonstrating effective and unbiased generalization to both the majority and minority clusters.

The ROC Curve (Figure 5a) and the Precision-Recall Curve (Figure 5b) further demonstrate the model’s discriminative ability. Achieving an AUC of 1.00 and an Average Precision (AP) of 1.00 confirms that the SFL model is capable of near-perfect separation between the two classes, particularly for the minority class (Class 1). This is a direct consequence of the optimal topological separation demonstrated in the UMAP plots (Figure 1b).

4. Discussion and Conclusion

The results unequivocally support the architectural choice of the Synergistic Fusion Layer over traditional concatenation methods. The core finding is that non-linear fusion is essential for maximizing model reliability in multimodal classification tasks involving

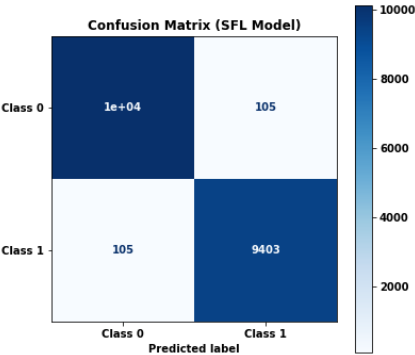
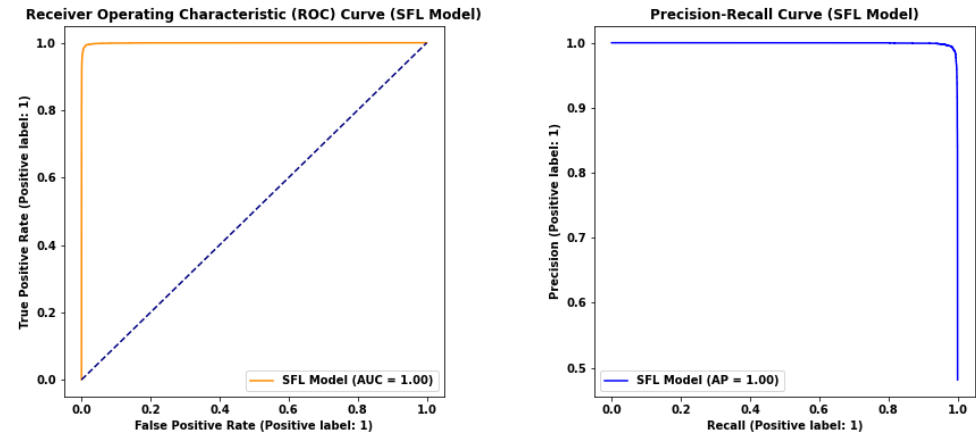


Figure 4: Confusion Matrix (SFL Model). The near-equal distribution of False Negatives (105) and False Positives (105) confirms the model’s balanced error rate and high classification fidelity.



(a) Receiver Operating Characteristic (ROC) Curve: AUC = 1.00 (b) Precision-Recall Curve: AP = 1.00

Figure 5: SFL Model Discriminative Performance. Both curves confirm the model’s maximal discriminative power and high performance on the minority class.

deep embeddings and structural cues. The SFL model is not just a marginally better classifier; it is a robust and well-calibrated probability machine. The reduction in ECE from 0.0500 to 0.0035 is a compelling demonstration that the structural cues successfully regularize the confidence of the deep learning model, producing highly trustworthy probability estimates. This is a crucial advancement for real-world applications where model trust is paramount. Future work will investigate the optimal integration of this

SFL module into pre-trained transformer blocks to further leverage the contextualization provided by structural metadata.

References

- Breiman, Leo. 2001. Random Forests. *Machine Learning*, 45(1):5–32.
- Campello, Ricardo J. G. B., Davoud Moulavi, and Jörg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *PAKDD 2013*.
- Cherubin, G. 2023. How do the performance of a Conformal Predictor and its underlying algorithm relate? In *COPA 2023, Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, PMLR.
- Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP 2014*.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL 2019*.
- Gao, Jing, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation*, 32(5):829–864.
- Gazin, U., R. Heller, E. Roquain, and A. Solari. 2025. Powerful batch conformal prediction for classification. In *AISTATS 2025*.
- Guha, E., S. Natarajan, T. Mollenhoff, and E. M. Khan. 2024. Conformal prediction via regression-as-classification.
- McInnes, Leland, John Healy, and Jef Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.
- Naeini, Mahdi Pakdaman, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *AAAI 2015*.
- Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP 2019*.
- Shaukat, F., N. Ejaz, Z. Ashraf, M. M. Alnfai, N. N. Alotaibi, and S. M. M. Alnefaie. 2025. An interpretable multi-transformer ensemble for text-based movie genre classification. *PeerJ Computer Science*.
- Tavares, T. F. and F. J. Ayres. 2025. Multi-label cross-lingual automatic music genre classification from lyrics with Sentence BERT.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NIPS 2017*.
- Wu, Y. and J. Wan. 2025. A survey of text classification based on pre-trained language model. *Neurocomputing*.