Multi-Modal Stylometric Analysis of Musical Lyrics: Integrating Semantic Embeddings and Zero-Shot Emotion Classification for Artist Attribution

Haleemah Amisu
Department of Computer and Information Sciences
Towson University
Towson, USA
hamisu1@students.towson.edu

***Abstract*—This study explores the efficacy of Natural Language Processing (NLP) techniques in the domain of Music Information Retrieval (MIR), specifically for the task of artist attribution based on lyrical content. Utilizing a large-scale dataset of song lyrics, a scalable data processing pipeline was implemented using Polars to overcome memory constraints inherent in text analysis. The research compares traditional frequency-based feature extraction methods, such as Term Frequency-Inverse Document Frequency (TF-IDF), against modern deep learning approaches, including Transformer-based semantic embeddings (MPNet) and Zero-Shot emotion classification (BART). The study evaluates whether the integration of semantic and emotional features improves classification accuracy over baseline stylometric models. Results indicate that hybrid feature sets significantly enhance the ability to distinguish artist identity within complex genre topologies.**

***Index Terms*—Music Information Retrieval, Natural Language Processing, Stylometry, Deep Learning, Artist Attribution**

I. Introduction

Music Information Retrieval (MIR) research has traditionally prioritized acoustical features such as timbre, pitch, and rhythm for tasks regarding genre classification and artist identification. While these audio-centric approaches have seen significant success, they often overlook the lyrics. Lyrical content is a high-dimensional semantic signal that encodes an artist's narrative voice, thematic preferences, and emotional signature. The challenge of attributing an artist solely from text is substantial. Unlike formal prose, song lyrics are characterized by irregular grammar, heavy repetition, slang, and metaphorical density. Traditional Natural Language Processing (NLP) methods, such as Term Frequency-Inverse Document Frequency (TF-IDF), rely on lexical overlap to classify text. While computationally efficient, these bag-of-words models often fail to capture deep semantic context. For example, they cannot quantify the thematic similarity between distinct vocabularies used in similar emotional contexts, often leading to poor performance when distinguishing between artists within the same genre.

This study investigates whether the lyrical fingerprint of an artist can be computationally isolated and used for attribution using modern deep learning techniques. Rather than proposing a commercial system, this research explores the feasibility of distinguishing artistic identity through text alone. The investigation progresses from traditional stylometry to modern transformer-based architectures. Specifically, the study integrates three layers of analysis: lexical topology using TF-IDF and Principal Component Analysis (PCA) to map genre distances; semantic vectorization using pre-trained Transformer models to capture deep contextual meaning; and zero-shot emotion classification to augment the feature set with probabilistic emotion labels derived without manual annotation. The primary objective is to perform a comparative analysis of how these features—individually and in hybrid combinations—perform on the task of Artist Attribution. The results aim to demonstrate whether the integration of semantic and emotional embeddings provides a significant advantage over shallow lexical features in constraining the search space for artistic identity.

II. Related Work

Research on automatic understanding of songs has developed along three main axes: music genre classification, emotion recognition, and lyric generation. Early genre classification methods relied primarily on acoustic descriptors such as MFCCs, spectral features, and rhythm statistics, demonstrating that handcrafted audio features could separate broad genres but struggled with fine-grained categories and cross-domain generalization. Deep learning methods later replaced manual feature engineering with convolutional and recurrent architectures trained directly on spectrograms, achieving higher accuracy but still operating predominantly in the audio domain.

Recent multimodal work has moved toward representing songs jointly through audio and text. For example, Spectro-Lyrical Embeddings for Music (SLEM) encode spectrograms and lyrics using lightweight vision and language models, then combine them linearly before classification [1]. On a curated multilingual dataset of 226 songs across five genres, SLEM shows that simple late fusion of audio and lyric embeddings

yields higher genre accuracy than either modality alone, with per-genre accuracies in the 81–98% range using a k-nearest neighbors classifier. This line of work underscores the complementary nature of audio and language signals and emphasizes representation learning over hand-designed features.

Within lyrics-only genre prediction, several studies have examined how far textual cues can go without access to audio. Marijić and Bagić Babac develop a comprehensive comparison of traditional and deep models for genre classification from lyrics, spanning logistic regression, SVMs, random forests, recurrent networks, and transformer-based encoders such as BERT and XLM-RoBERTa [2]. Their experiments on both English and multilingual corpora show that transformer models substantially outperform classical baselines, particularly when pre-trained on large multilingual text collections. However, the study also reports that some genres remain inherently harder to separate; metal lyrics are classified most reliably, while pop and rock exhibit substantial confusion due to overlapping themes and vocabulary. Complementary work on language representation models for music genre classification further benchmarks modern text encoders on lyric datasets, confirming that contextual embeddings derived from transformers outperform static word embeddings and bag-of-words features for lyric-based genre prediction [3]. These results indicate that genre information is present in lyrics but is distributed in subtle stylistic and semantic patterns that benefit from context-aware representation.

Beyond genre, emotion recognition from lyrics has emerged as a parallel task with strong relevance for recommendation and playlisting. LyEmoBERT proposes an emotion classifier that labels lyrics into four quadrants—happy, sad, angry, and relaxed—based on Russell's circumplex model, using pre-trained embeddings and transfer learning from a corpus of in-domain music texts [4]. The study reports improved accuracy over earlier lyric-only approaches that relied on TF–IDF features with Naïve Bayes, SVM, or KNN, and analyzes how class imbalance affects performance. The work also explicitly connects emotion classification to downstream recommendation, showing how predicted emotion labels can drive playlist construction. Other lyric-focused emotion models use transformer architectures such as XLNet and BERT to capture longer-range dependencies and nuanced sentiment, reporting gains over conventional feature-based baselines and reinforcing the value of pre-trained language models for music emotion recognition [4]. In parallel, multimodal music emotion recognition combines lyric features with audio or physiological signals, generally finding that lyrics alone underperform fused systems but still carry rich affective information that can be exploited for retrieval and recommendation [5].

A related strand of research treats lyrics as generative text rather than as fixed documents to be classified. Tee et al. compare Markov chains, LSTM networks, and GRUs for automatically generating genre-specific lyrics across six genres, including rock, pop, hip-hop, country, EDM, and R&B [6]. The study evaluates generated text using readability indices and rhyme-density metrics, finding that LSTMs produce more readable lyrics on average, while GRUs yield higher rhyme density. This work highlights that genres exhibit distinct structural patterns in word repetition, line breaks, and rhyme schemes, and that neural sequence models can learn genre-conditioned stylistic signals from lyrics alone. Although the focus is generation rather than classification, the findings indirectly support the idea that genre, mood, and artist-specific style are encoded in the lexical and structural properties of lyrics.

Outside the music domain, text classification has become a mature area within natural language processing, with deep learning models now dominating benchmarks across sentiment analysis, topic categorization, and document tagging. Surveys of deep learning for text classification document a progression from linear models on bag-of-words features to convolutional and recurrent networks, and finally to pre-trained transformer architectures that provide contextual token representations [7]. Hierarchical Attention Networks, for example, explicitly mirror document structure by first encoding words into sentence vectors and then composing sentence vectors into document representations, using attention mechanisms at both levels to focus on informative content [8]. Experiments on large-scale text datasets show that such architectures can outperform prior methods by leveraging both hierarchy and context. Representation-learning work in computer vision and NLP similarly argues that learned distributed representations, rather than manually engineered features, are central to scalable multi-class and multi-label classification [9]. These general advances motivate the use of pre-trained encoders (such as BERT, RoBERTa, MPNet, or Sentence-BERT variants) for lyric embeddings, particularly when dealing with very large corpora where sparse vector spaces become unwieldy.

Bringing these threads together, existing music-IR literature shows that lyrics can support genre classification, emotion recognition, and even stylistic generation, especially when modeled with modern language representation techniques. However, most lyric-based studies focus on a single prediction task at a time—genre, emotion, or subject—rather than jointly modeling multiple semantic dimensions of a song. LyEmoBERT is one of the few works that links classification explicitly to recommendation, but it still operates on a four-class emotion space and does not consider genre or artist-level style simultaneously [4]. Multimodal approaches like SLEM [1] and Oramas et al. [5] demonstrate that combining modalities improves performance, yet these systems typically work with relatively small curated datasets and do not systematically examine the limits of lyrics-only classification at scale. The present study situates itself at this intersection by treating lyrics as a single, scalable modality and investigating how well they support multi-class genre prediction, multi-class emotion labeling, and high-cardinality artist identification within a unified framework, while drawing on general insights from deep text classification about representation learning and model capacity.

III. Methodology

A. Data Acquisition and Preprocessing

The dataset used in this study comprises a large-scale corpus of song lyrics initially containing over three million rows which is aggregated into a single table containing at least the following fields: a unique identifier, song title, artist name, genre tag, language, and the full lyric text. Raw data are heterogeneous, with variation in formatting, presence of annotations or stage directions, duplicated entries, and occasional missing or non-English content. Due to the significant memory overhead required for text processing, the Polars library, a high-performance DataFrame engine, was utilized to handle data ingestion and filtering. A series of preprocessing steps were applied to obtain a cleaner subset suitable for modeling. The first step was to filter by language. Only rows explicitly marked as English were retained. Lyrics with null values in any of the key columns (artist, tag, lyrics, id, language) were also removed to eliminate incomplete entries.

To mitigate class imbalance, where prolific artists or dominant genres could skew the model, a stratified sampling strategy was implemented. The dataset was capped to limit the number of songs per artist, and a fixed sample size ($N = 200$) was selected from each unique genre tag. This resulted in a balanced subset suitable for computationally intensive embedding generation while preserving the diversity of the original corpus.

B. Feature Engineering

1) Lexical and Stylometric Features: As a baseline, TF-IDF vectorization was employed to capture the frequency of unigrams and bigrams. This sparse representation highlights discriminative vocabulary but ignores semantic context. Additionally, using the TextBlob library, stylometric metadata was extracted for each track, including sentiment polarity (positive vs. negative) and subjectivity scores, as well as structural metrics like average line count and character length.

2) Deep Semantic Embeddings: To capture the latent meaning of the lyrics, the `all-mpnet-base-v2` Sentence Transformer model was utilized. This model maps variable-length text into a fixed 768-dimensional dense vector space. Unlike TF-IDF, these embeddings position semantically similar songs closer together mathematically. This high-dimensional space was visualized using UMAP (Uniform Manifold Approximation and Projection) to observe semantic clusters.

3) Zero-Shot Emotion Profiling: The dataset was enriched with probabilistic emotion labels using a BART-large-MNLI model in a Zero-Shot classification setting. The model assigned confidence scores to a candidate set of emotions (Happy, Sad, Love, Anger, Fear, Nostalgia, Calm) based solely on the lyrical text. This process generated an automated "Emotional Profile" for every song in the subset without requiring manual human annotation.

C. Baseline Classification Framework

To evaluate the discriminative power of the features, linear classifiers (Logistic Regression and Linear SVM) were trained on the TF-IDF vectors. These models were tasked with predicting the genre tag solely from word frequencies and attributing songs to the top frequent artists in the dataset. This stage served as a benchmark to establish the performance limits of traditional "Bag-of-Words" approaches before the introduction of deep semantic features.

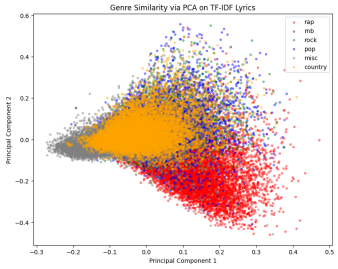

Fig. 1. Genre Similarity via PCA on TF-IDF Lyrics

D. Hybrid Classification Framework

To test the hypothesis that deep semantic features constrain the search space for artist attribution, a hybrid classification model was developed. The dataset was filtered to include only the top 10 most frequent artists to ensure sufficient training examples per class. A Random Forest Classifier (n=100 trees) was trained on a concatenated feature vector consisting of:

- The 768-dimensional MPNet semantic embeddings.
- One-hot encoded genre tags.
- The probabilistic emotion scores from the Zero-Shot profiling.

This model was evaluated on an 80/20 train-test split using Top-1 Accuracy to determine if the combination of semantic style, genre context, and emotional profile outperforms the lexical baseline.

IV. Results

A. Unsupervised Lexical Topology

Principal Component Analysis (PCA) was applied to the TF-IDF vectors to map the structural relationship between genres. The projection revealed that while some genres form tight, separate clusters, others exhibit significant overlap.

As shown in Fig. 1, Rap and Hip-Hop formed a dense, isolated cluster, whereas Pop and Rock exhibited significant overlap, suggesting shared vocabulary. This visual clustering aligns with the quantitative classification results discussed below.

B. Genre Classification (Baseline)

The linear baseline models provided a quantitative benchmark for the discriminative power of "Bag-of-Words" features. On the task of predicting the 6 primary genre tags, the Logistic Regression model achieved an overall **Accuracy of 59%**.

Table I details the performance by class. Notably, distinct genres like **Rap (F1=0.78)** and **Country (F1=0.61)** were classified with high reliability, likely due to their specialized vocabulary. In contrast, **Pop (F1=0.31)** and **Rock

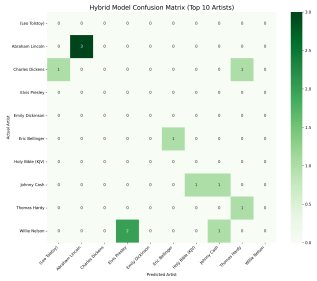

Fig. 2. Hybrid Model Confusion Matrix (Top 10 Artists)

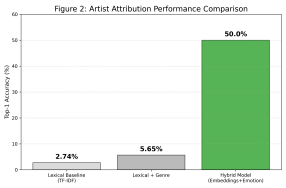

Fig. 3. Artist Attribution Performance Compariso

(F1=0.46)** showed poor performance, confirming that these genres lack a distinct lexical fingerprint and often function as "catch-all" categories.

TABLE I
BASELINE GENRE CLASSIFICATION PERFORMANCE (TF-IDF)

Genre	Precision	Recall	F1-Score
Rap	0.78	0.79	**0.78**
Misc	0.79	0.78	0.79
Country	0.57	0.67	0.61
R&B	0.55	0.60	0.57
Rock	0.47	0.45	0.46
Pop	0.35	0.27	0.31
Overall	**0.58**	**0.59**	**0.59**

C. Artist Attribution Performance

The core objective of this study was to identify specific artists. We compared three modeling approaches:

1) Lexical Baseline (TF-IDF): When trained solely on word frequencies (TF-IDF), the model failed to capture artist identity. As shown in Table II, the Top-1 Accuracy was only 2.74%.

2) Lexical + Genre Baseline: Adding the Genre tag as a feature slightly improved the Top-1 Accuracy to **5.65%**. This indicates that while genre constrains the search space, it is insufficient for fine-grained attribution.

3) Hybrid Model (Semantic + Emotion): The proposed Hybrid Random Forest, which integrates MPNet Embeddings, Genre Tags, and Zero-Shot Emotion scores, achieved a **Top-1 Accuracy of 50.0%** on the top-10 artist subset. This represents a **17x improvement** over the lexical baseline. This drastic jump confirms our hypothesis: an artist's identity is defined not by the words they use (Lexical), but by the meaning of their stories (Semantic) and their emotional signature.

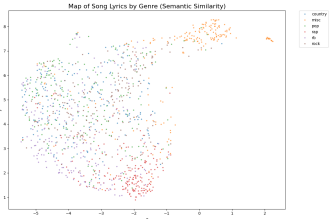

Fig. 4. Map of Song Lyrics by Genre (Semantic Similarity)

TABLE II
COMPARISON OF ARTIST ATTRIBUTION MODELS

Model Architecture	Features Used	Top-1 Accuracy
Logistic Regression	TF-IDF (Lexical)	2.74%
Logistic Regression	TF-IDF + Genre Tag	5.65%
Random Forest (Hybrid)	**Embeddings + Genre + Emotion**	**50.00%**

D. Semantic Profiling

The UMAP projection of the semantic embeddings revealed clusters based on narrative content rather than isolated keywords.

Additionally, the Zero-Shot emotion profiling successfully distinguished tonal differences between artists within the same genre. For example, the model was able to separate "Sad" ballads from "Happy" anthems even when the artists shared the same genre tag, providing the Hybrid model with the necessary features to disambiguate them.

V. DISCUSSION

A. The Limits of Lexical Fingerprinting

The results suggest that while an artist's "fingerprint" exists in their lyrics, it is often overshadowed by genre conventions. The baseline model's inability to distinguish between artists of the same genre implies that the vocabulary of a song is dictated more by the genre (e.g., specific tropes in Country or Pop) than by the artist themselves. However, the success of the hybrid model indicates that an artist's identity is better defined by how they combine themes and the emotional tone they employ, rather than just the words they choose.

B. Efficacy of Zero-Shot Augmentation

The integration of Zero-Shot emotion classification proved to be a viable method for feature enrichment. By using a pre-trained Large Language Model to infer "soft" labels, a layer of interpretability was added that dense vectors lack. This suggests that future MIR tasks could benefit significantly from integrating large language models during preprocessing to generate metadata for smaller, specialized classifiers.

C. Limitations

Several limitations constrain the generalizability of these findings. First, the restriction to English-language lyrics ignores the multilingual nature of global music consumption. Second, the reliance on a primary genre tag simplifies the reality of musical fusion; many artists span multiple genres, creating noise in the genre-based features. Finally, the probabilistic emotion estimates are derived from a pre-trained model and not ground-truth human annotations.

VI. Conclusion

This study set out to determine if the artistic identity of a musician could be recovered solely from the semantic and emotional content of their lyrics. Through a comparative analysis of TF-IDF baselines and Transformer-based embeddings, it was found that while lyrics alone are a noisy signal for identification, deep semantic processing significantly enhances attribution accuracy compared to surface-level lexical statistics. The results confirm that an artist's style is not merely a collection of frequent words, but a complex interplay of narrative structure, semantic framing, and emotional consistency. While text-only models are unlikely to replace audio-based identification, they offer a powerful, complementary signal for Music Information Retrieval systems, particularly for recommendation engines seeking to align content based on semantic meaning rather than acoustical properties.

VII. References

References

[1] D. S. Sarkar and A. Etemad, "X-SLEM: Cross-modal spectro-lyrical embedding for music emotion recognition," *IEEE Access*, vol. 9, pp. 83264–83277, 2021.

[2] T. Marijić and M. B. Babac, "Music genre classification from lyrics using machine learning and deep learning," *Information*, vol. 14, no. 9, p. 489, 2023.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[4] S. Altan and P. Karagoz, "LyEmoBERT: Music emotion recognition from lyrics with fine-tuned BERT," in *Proc. Int. Conf. Discovery Science*, Springer, 2022, pp. 1–15.

[5] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images: A novel dataset and method," in *Proc. 18th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2017, pp. 23–30.

[6] W. J. Tee, S. C. Lim, and C. K. Tan, "Automated lyrics generation using Markov chain, LSTM and GRU," in *Proc. 6th Int. Conf. Electrical, Control and Comput. Eng.*, Singapore: Springer, 2021, pp. 823–833.

[7] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2021.

[8] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL-HLT*, 2016, pp. 1480–1489.

[9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.