Emotion-Based Bollywood Music Recommendation Systems: Integrating Facial Expression Recognition and Lyrics Sentiment Analysis—A Comprehensive Review

Prof. Naresh Shende[1], Amrit Gupta[2], Jainil Solanki[3], Karan Tirwa[4], Mohammed Tufail[5]

[1,2,3,4,5]*Dept. of Artificial Intelligence and Data Science, Rajiv Gandhi Institute of Technology, Mumbai, India*

***Abstract*—Emotion-based music recommendation systems are transforming the way users interact with large-scale digital music libraries, particularly in emotionally rich domains like Bollywood music. This review surveys current advances in integrating facial expression recognition (FER) and lyrics sentiment analysis, emphasizing their application in the Indian music landscape. We explore deep learning, transformer-based approaches for Hindi/English lyric emotion tagging, multimodal fusion (facial+ audio + text), system architectures, and practical challenges for real-time use. Ten pivotal research papers from 2016–2025 are analyzed comparatively, with a focus on scope, methodology, efficacy, and limitations. Our review concludes with practical directions for future research on personalized, affective music recommendation in Indian languages, stressing the need for culturally attuned models and real-time efficiency.**

***Index Terms*—Emotion Recognition, Music Recommendation, Lyrics Sentiment, Bollywood, Facial Expression, Hindi, Trans- former, Deep Learning, Multimodal, Review**

I. INTRODUCTION

Personalized music recommendation is at the forefront of digital music innovation. Traditional recommender systems rely predominantly on explicit user ratings (collaborative filtering) or acoustic features (content-based filtering). However, these methods often fail to capture the transient and highly personal nature of user intent. The user's current emotional and psychological state represents a far more natural and direct proxy for preference. In cultural domains, particularly in Bollywood and Indian music curation, emotional con- text—encompassing lyrical themes, melodic structure, and vocal style—plays an especially vital role in the user experience. The integration of advanced computer vision techniques for real-time Facial Expression Recognition (FER) and sophisticated Natural Language Processing (NLP) for lyrics sentiment analysis unlocks unprecedented levels of personalization. This paper systematically examines recent advances, datasets, and system architectures for emotion-driven music recommendation, with a particular focus on the Bollywood (Hindi/Indian regional) context. The following sections outline the historical progression of the field, survey key modalities and methodologies, present a comparative analysis of seminal works, and discuss critical challenges and future research directions.

II. LITERATURE SURVEY AND BACKGROUND

A. Historical Evolution

Early attempts at affective music recommendation relied on simple metadata and manual emotion tagging. However, the field has rapidly moved toward computational analysis. The literature evolved from content-based filtering using basic emotion tags to complex deep neural networks designed for multimodal affective computing. Pichappan's 2025 review offered a contemporary survey of emotion-induced music recommendation systems, charting the field's trajectory and highlighting the increasing dependence on hybrid and deep learning methods [1]. Emotional modeling often relies on dimensional models like Valence-Arousal-Dominance (V-A- D) or categorical models like Ekman's six basic emotions, tailored to the specific modality being processed.

B. Emotion Recognition Modalities

1) Facial Expression Recognition (FER): Initial studies focused on basic CNN architectures for real-time facial emotion detection, often utilizing datasets like FER2013 [2]. The need for high-speed, on-device recognition has driven the adoption of faster detection frameworks. The work presented in 2025 demonstrated the integration of the YOLO v11 model for FER, emphasizing its superior performance in real-time latency for live webcam inputs, allowing for rapid integration into music applications [3]. Modern systems often employ optimized models like **EfficientNet** for efficient and accurate feature extraction from the user's face.

2) Lyrics Sentiment Analysis: Sentiment analysis in music has seen a dramatic transition. Kulkarni (2023) explored the use of Bi-directional Long Short-Term Memory (BiL- STM) networks for analyzing Hindi song lyrics, providing a deep learning baseline for capturing temporal dependencies in text [10]. However, capturing the deep contextual meaning and long-range dependencies inherent in complex lyrical structures, especially in Indian languages, necessitated more powerful tools. Recent research by Shanker et al. (2023) demonstrated the superiority of Transformer-based models, such as XLNet, which significantly improved emotion tagging accuracy for Hindi lyrics by leveraging self-attention mechanisms [5]. Furthermore, Dhar et al. (2025) applied advanced Transformer techniques specifically calibrated with the Indian cultural emotion framework (Navrasa), offering results that are more culturally resonant than models based purely on Western lexical resources [4]. Modern approaches leverage models like **BigBird** or other specialized Transformers for efficient handling of long-context lyrical data.

3) Multimodal Fusion: The consensus in affective computing is that combining multiple sources of emotional data mitigates the ambiguity inherent in any single modality. Multimodal fusion, integrating audio, lyrics, and facial cues, consistently outperforms unimodal approaches. Patra et al. (2016) demonstrated an early multimodal mood classification framework for Hindi songs, combining audio features with lyric analysis using traditional machine learning (ML) [7]. More recently, the work by Bottu and Ragavan (2025) show- cased a state-of-the-art system that integrates real-time CNN- based FER with Transformer-based lyrics sentiment analysis, representing a powerful fusion technique for practical recommendation systems [8].

C. Indian and Multilingual Focus

A crucial finding in the literature is the necessity of cultural calibration. Studies focusing on Hindi/Bollywood music and utilizing models trained specifically for Indian languages show marked performance improvements over systems relying on models built on Western-centric lexicons [6]. The use of the Navrasa (Nine Emotions) framework, as opposed to simpler categorical or dimensional Western models, has proven more effective in capturing the nuanced emotional spectrum of Indian musical content [4].

III. METHODOLOGIES – SYSTEM ARCHITECTURES

Multimodal system architectures are designed to ingest asynchronous data streams (facial video, lyrical text, audio features) and fuse them at either the feature level or decision level to produce a unified emotional state prediction.

A. Conceptual System Architecture

The flow diagram in Fig. 1 illustrates a common conceptual architecture for an emotion-based music recommendation system. It details the parallel processing of data from two main streams—the static song data (lyrics) and the real-time user state (face/expression)—before converging in a central recommendation engine. The lyrics pipeline handles multilingual data (English + Hindi + Marathi), cleaning and tokenizing the text, and using an advanced Transformer-based model (e.g., BigBird) to assign one of seven emotion classes to the song. Concurrently, the webcam input undergoes face detection and a dedicated FER model (e.g., CNN or EfficientNet) detects the user's current emotion. The core of the system, the **Music Recommendation Engine**, matches the detected user emotion with the pre-tagged song emotion to curate a personalized playlist delivered through a web application.

Fig. 1. Data Flow Diagram of an Emotion-Based Music Recommendation System, illustrating parallel

processing of lyrics and facial input.

Table 1 provides a comparative summary of the methodological approaches identified in the literature, contrasting the input modalities, core computational models, key innovations, and inherent limitations.

IV. STATE-OF-THE-ART SYSTEM: PROPOSED ARCHITECTURE

Our system is designed to advance the state-of-the-art by explicitly addressing the shortcomings of unimodal or generalized systems within the Indian music domain.

The architecture integrates three primary components:

1) Data Curation: Comprehensive Bollywood song datasets spanning 1990–2025 are used, categorized by genre, era, and emotional labels.
2) Modal Analysis:
   - *Facial Expression Recognition (FER):* An EfficientNet-based pipeline is used for real-time webcam/input analysis, mapping facial features to Indian emotion taxonomies.
   - *Lyrics Sentiment Labeling:* A multilingual Transformer model (e.g., BigBird or XLNet) is employed for processing Hindi and English lyrics, specifically trained on the Navrasa/Hindi-specific sentiment analysis framework.
3) Fusion and Recommendation: A robust fusion layer combines the real-time facial emotion state with the inherent emotional metadata derived from the lyrics. This fusion assigns a robust emotion-label and mood-tag, supplemented by proxy audio features (tempo, valence) when full audio analysis is unavailable, to generate a highly personalized recommendation playlist.

This architecture aims to provide a reliable, low-latency system that integrates both the instantaneous user mood (via FER) and the contextual emotional content of the song itself (via lyrics).

V. DISCUSSION: CHALLENGES AND FUTURE SCOPE

The transition of emotion-based music recommendation from academic concept to practical application presents several critical challenges, particularly within the domain of Indian music.

A. Nuance and Diversity in Emotional Mapping

One of the most persistent issues is the inherent complexity and diversity of emotions within Indian culture. The Indian classical framework of *Navrasa* (Nine Emotions) includes complex states like *Adbhut* (Wonder) and *Karuna* (Sorrow/Compassion), which are not adequately captured by the simpler categorical models like the Ekman set typically.

TABLE I Comparative Analysis of Methodological Approaches in Emotion-Based Music Recommendation (2016–2025)

Study	Modality/ Fusion	Model(s)	Key Innovations	Limitation(s)
Patra et al. (2016)	Audio + Lyrics	Classical ML (e.g., SVM, k-NN)	Multimodal mood classification framework; focused on Indian genre distinctions.	Small dataset size; reliance on non-deep learning methods.
Velankar et al. (2021)	Lyrics	Knowledge Graph, ML	Contextual mood analysis using graph-based representations for Hindi songs.	Limited to lyrical modality; complexity of graph construction and inference.
Kulkarni (2023)	Lyrics	BiLSTM	Deep learning baseline for sentiment analysis of Hindi/English song lyrics.	Baseline model only; limited capability to capture deep, complex emotional context.
Shanker et al. (2023)	Lyrics	Transformer (XLNet)	Significant improvement in emotion recognition accuracy for Hindi lyrics via self-attention.	Unimodal focus (only lyrics); high computational requirement.
ScienceDirect (2023)	All Modalities	Machine Learning	General framework for emotion improvement; utilizes BRECVEMA mapping.	Broad scope means it's not specifically tailored to the unique Hindi/Indian context.
IJIRSET (2024)	Facial Expression	Deep CNN	Comprehensive literature review emphasizing the real-time performance of CNN-based FER	Review-based; does not address lyrical or audio components.

			systems.	
IJERT (2025)	Facial Expression	YOLO v11	Ultra-fast, real-time FER integration for direct use in live webcam music applications.	Unimodal focus; ignores the rich contextual information in the lyrics.
Dhar et al. (2025)	Lyrics	Transformer	Application of Navrasa emotion framework; advanced Hindi NLP with state-of-the-art metrics.	Limited modalities (only lyrics); lack of facial/audio fusion for comprehensive state analysis.
Bottu & Ragavan (2025)	FER + Lyrics	CNN, Transformer	End-to-end real-time fusion of facial expression and lyrics sentiment.	Possible limitations in dataset size and cultural diversity of the training data.
Pichappan (2025)	All	Survey/Review	Comprehensive analysis of trends, limitations, and future scope across 32 studies.	Secondary research only; provides no new empirical data.

used in Western systems. Future research must focus on building and validating larger, culturally-specific datasets that explicitly map Indian musical features (e.g., specific *raagas* or instrumental sounds) to these nuanced emotional states. This is essential for the system to move beyond generic mood detection to culturally relevant personalization.

B. Accuracy in Hindi/Indian Language Processing

While Transformer models have demonstrated superior performance over previous methods (like BiLSTM) in handling Hindi lyrics [5], the language still poses significant challenges. Hindi's rich morphology, code-mixing with English (Hinglish), and extensive use of figurative language and idiom require continuous retraining and cultural calibration of NLP models. Furthermore, the lack of large-scale, publicly available, manually annotated emotional datasets for regional Indian languages remains a key bottleneck. Development of transfer learning techniques that leverage resources from high- resource languages (like English) to improve emotion classification in low-resource Indian languages is a crucial area for development.

C. Real-time System Constraints and Ethics

Deploying these systems in real-time, especially on mobile or edge devices, necessitates low-latency processing. While models like YOLO v11 address the speed requirement for FER [3], the integration of multiple deep learning streams (FER + Transformer NLP) introduces significant computational over- head. Optimizing the fusion architecture to minimize latency without sacrificing prediction accuracy is vital.

Ethically, the use of facial analysis raises significant privacy and security concerns. The system must be transparent about data collection and usage. Future systems must incorporate ethical safeguards, ensuring that facial data is processed locally (on-device) and only transiently used for emotion inference, with the user retaining full control over the data pipeline.

D. Integration of Actual Audio Features

Currently, many fusion systems primarily rely on FER and lyrics sentiment, often using only proxy audio features (e.g., measured tempo or basic energy levels) when a full audio analysis is too slow or resource-intensive. However, the true emotional impact of music is intrinsically linked to its sonic characteristics (timbre, harmony, rhythm, and melody). For full model generalization and robustness, there is a critical need to integrate deeper audio analysis features using techniques like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) trained directly on raw or Mel-spectrogram audio data, as demonstrated in earlier multimodal work [7]. Future efforts should focus on creating lightweight, efficient audio feature extraction models that can run in parallel with the FER and NLP streams.

E. Dataset Expansion and Personalization Feedback Loops

The efficacy of any AI-driven recommendation system is tied directly to the quality and volume of its training data. The development of more labeled data, particularly data that links multimodal inputs (user's face, lyrics, and audio) to their explicit feedback, is essential. Furthermore, moving beyond simple static recommendations, future systems should incorporate sophisticated user feedback loops. These loops will allow the model to learn user-specific emotional responses, tuning the general emotion model to individual preferences, thereby leading to truly personalized, affective music curation, as suggested by wellness-focused frameworks [9].

VI. CONCLUSION

Emotion-based recommendation—especially when augmented with advanced Facial Expression Recognition and Transformer-enhanced Hindi lyric processing—is entering a practical phase for application in Bollywood and Indian music. This review identifies significant progress, particularly through multimodal integration and the adaptation of models to the Indian cultural context. Despite these advancements, remaining gaps exist concerning the nuance of emotional mapping, the need for robust real-time performance, and the availability of large, culturally-specific datasets. Future systems must focus on ethical data handling, low-latency fusion architectures, and deeper integration of actual audio features alongside culturally attuned NLP models to achieve truly personalized, explain- able, and affectively accurate music curation. The continued progress in this area promises to revolutionize how users experience and interact with digital music libraries.

REFERENCE

[1] Pichappan, J. (2025). "A Review of the Emotion-Induced Music Recommendation Systems." *JDIM*.

[2] "A Literature Review on Face Emotions Based Music Player," *IJIRSET*, 2024.

[3] "Music Recommendation System using YOLO v11 For Facial Expression," *IJERT*, 2025.

[4] Dhar, A., et al. (2025). "Emotion recognition from lyrical text of Hindi songs." *IJISAE*.

[5] Shanker, S., et al. (2023). "A Transformer-based Analysis of Lyrics for Improved Emotion Recognition."

[6] Velankar, A., et al. (2021). "Contextual Mood Analysis with Knowledge Graph for Hindi Songs." *arXiv*.

[7] Patra, A., et al. (2016). "Multimodal Mood Classification Framework for Hindi Songs." *Computacion y Sistemas*.

[8] Bottu, V., & Ragavan, K. (2025). "Emotion Based Music Recommendation System Integrating Facial Expression Recognition and Lyrics Sentiment Analysis." *IEEE*.

[9] "An emotion-based personalized music recommendation framework for emotion improvement," *ScienceDirect*, 2023.

[10] Kulkarni, J. (2023). "Sentiment Analysis of Hindi Song Lyrics using a BiLSTM." Thesis, NCIRL.