Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on AI-Driven Creativity and
Intelligent Practices in Visual Arts 2025 6(4s)

MUSIC SENTIMENT ANALYTICS: UNDERSTANDING AUDIENCE REACTIONS USING MULTI-MODAL DATA FROM STREAMING PLATFORMS

Atanu Dutta [1], Mahesh Kurulekar [2], Ramneek Kelsang Bawa [3], Dr. Sharyu Ikhar [4], Rajendra V. Patil [5], Anureet Patil [6]

[1] Assistant Professor, School of Music, AAFT University of Media and Arts, Raipur, Chhattisgarh-492001, India
[2] Assistant Professor, Department of Civil Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India
[3] Associate Professor, School of Business Management, Noida International University, Noida, Uttar Pradesh, India
[4] Chief operating Officer, Researcher Connect Innovations and Impact Private Limited, Nagpur, Maharashtra, India
[5] Assistant Professor, Department of Computer Engineering, SSVPS Bapusaheb Shivajirao Deore College of Engineering, Dhule Maharashtra, India
[6] Department of Computer Applications, CT University Ludhiana, Punjab, India

Received 14 May 2025
Accepted 16 September 2025
Published 25 December 2025

Corresponding Author
Atanu Dutta, atanu.dutta@aaft.edu.in

DOI
10.29121/shodhkosh.v6.i4s.2025.6947

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Streaming services are adding more and more user-generated content, which provides us valuable insight on how people feel and how involved they are. Knowing how people feel and react to music may help to make music selection systems, marketing strategies, and outreach efforts to musicians much more efficient. This paper investigates how multi-modal information could be utilised for temper evaluation in song by means of textual, audio, and visual facts from streaming services. This paper indicates a whole technique for mood evaluation. It does this by way of integrating tune word data, audio alerts which includes velocity, rhythm, and pitch with visual cloth such as album cowl and music videos. The machine analyses songs the use of sophisticated device mastering techniques like natural language processing (NLP), audio sign processing to extract musical characteristics, and pc imaginative and prescient fashions to decide how people experience approximately what they see. Combining those many varieties of information enables we recognize more approximately how diverse items have an impact on emotional responses and makes mood categorisation algorithms more consistent. Performance metrics like as memory, accuracy, precision, and F1-rating are in comparison throughout several models to look how well multi-modal techniques carry out in comparison to unmarried-modal research. The findings imply that combining textual, spoken, and visible statistics produces better results than relying solely on conventional sentiment evaluation fashions, subsequently enabling more precise and thorough temper forecasts. This research illustrates how sophisticated mood analytics might not only improve listening but also support marketing decisions and artist strategies in the competitive music sector.

Keywords: Music Sentiment Analysis, Multi-Modal Data, Audience Reactions, Streaming Platforms, Machine Learning, Natural Language Processing

How to cite this article (APA): Dutta, A., Kurulekar, M., Bawa, R, K., Ikhar, S., Patil, R, V., and Patil, A. (2025). Music Sentiment Analytics: Understanding Audience Reactions Using Multi-Modal Data from Streaming Platforms. *ShodhKosh: Journal of Visual and Performing Arts, 6*(4s), 680–690. doi: 10.29121/shodhkosh.v6.i4s.2025.6947

1. INTRODUCTION

The digital shift has changed a lot about how people listen to and share music, especially through streaming services online. Since millions of people use these platforms every day to watch music, they have become great places to collect data about both the music and how people use it. Figuring out how people feel when they listen to music, on the other hand, is still hard. This is where mood research is very important. Sentiment analysis in music is the study of how people feel when they hear a piece of music and how those feelings affect the music itself. It is an important part of business plans, song selection systems, and attempts to reach out to artists. In the past, mood analysis mostly used written data, like song lyrics, to figure out how people felt. Even though song words give you a good idea of how a song makes you feel, they are only a small part of the whole audio experience. Figure 1 shows multi-modal sentiment analysis system architecture overview. Music is an art form that expresses feelings through both words and the sounds that go with them, such as speed, rhythm, pitch, and harmonics.

Figure 1

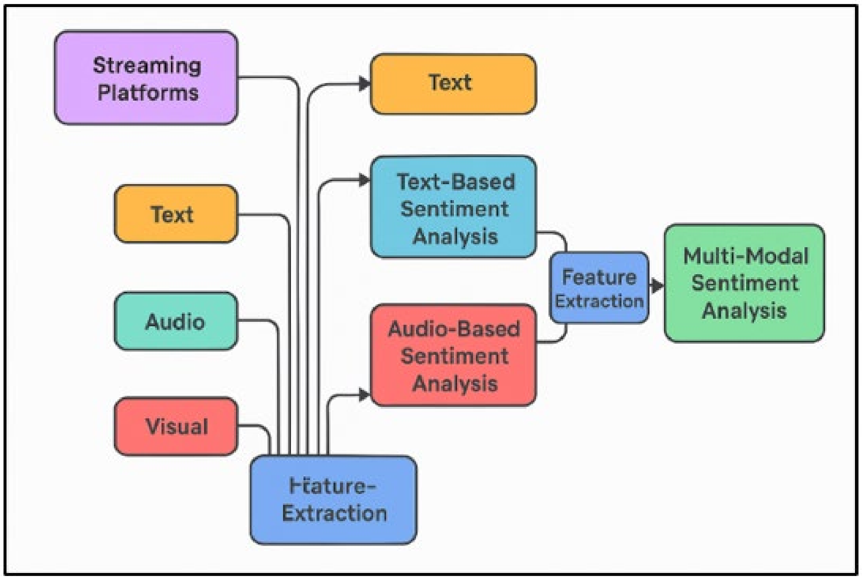
Streaming Platforms
Text
Text-Based Sentiment Analysis
Text
Audio
Visual
Feature Extraction
Multi-Modal Sentiment Analysis
Audio-Based Sentiment Analysis
Hature-Extraction

Figure 1 Multi-Modal Sentiment Analysis System Architecture

These sound effects often add to or even conflict with the lyrics, making the emotional context richer. Visual aspects like record covers, music videos, and live acts also have a big impact on how people think about a song as a whole. As a result, analysing texts alone is not enough to fully understand the range of feelings and emotions that viewers have. Within the beyond few years, there has been more and more interest in using multi-modal data, which includes written, audio, and visual facts, to research more approximately how human beings experience. Multi-modal mood analysis ambitions to bypass the restrictions of single-modal strategies by way of integrating many forms of facts. Audio temper evaluation is all about finding emotional cues in sounds, consisting of velocity, rhythm, and harmony shifts. Conversely, visual mood analysis use pc imaginative and prescient to decide how record artwork or music films have an effect on people. the usage of those strategies in addition to traditional text-based totally evaluation of song lyrics helps to offer a extra whole view of ways people understand song.

This paper targets to illustrate a unique technique for mood analysis using multi-modal statistics from live sites. Written data—like track lyrics, audio functions—like speed, pitch, and rhythm, and visible content material—like album covers and track videos incorporate 3 key classes of data this gadget integrates. each media is processed and tested independently the use of gadget studying models Ahmed et al. (2023). The mood categorisation is then made greater particular via merging the findings from all of the research in a fusion degree. Writing is classed the usage of herbal language processing (NLP) strategies inclusive of mood lexicons and deep getting to know models which includes recurrent neural networks (RNNs). Mel-frequency cepstral coefficients (MFCCs) and chroma features are utilised to extract audio characteristics for Zhang et al., (2023), therefore obtaining the musical characteristics influencing mood from sound. subsequently, convolutional neural networks (CNNs) and different laptop imaginative and prescient strategies extract functions from tune videos and recorded art connected to emotions. one of the key motivations for this approach is that it might assist to tailor tune suggestions systems Pan et al. (2023). by using properly predicting how clients will sense, systems like Spotify and Apple music may also enhance their algorithms to propose music extra in accordance with their emotions or preferences. Entrepreneurs inside the music enterprise might also utilise temper facts to create greater attractive and applicable commercials.

Multi-modal temper analysis seeks to avoid the troubles with unmarried-modal procedures by way of mixing several forms of facts. Audio mood evaluation is all approximately finding emotional cues in sounds, which include speed, rhythm, and harmony shifts. Conversely, visual mood analysis use laptop imaginative and prescient to decide how file art or track movies affect individuals. The usage of these strategies further to traditional textual content-based totally analysis of song lyrics facilitates to offer a greater whole view of how individuals understand song.

These paper pursuits to demonstrate a novel approach for temper analysis the usage of multi-modal information from live websites. Written data—like track lyrics, audio functions—like velocity, pitch, and rhythm, and visual content material—like album covers and tune movies contain 3 key categories of records this device integrates. every media is processed and tested independently the usage of system getting to know models Ahmed et al. (2023). The mood categorisation is then made more unique with the aid of merging the findings from all of the studies in a fusion level. Writing is assessed the usage of herbal language processing (NLP) strategies together with temper lexicons and deep getting to know fashions which include recurrent neural networks (RNNs). Mel-frequency cepstral coefficients (MFCCs) and chroma functions are utilised to extract audio characteristics for Zhang et al. (2023), consequently obtaining the musical features influencing mood from sound. Finally, convolutional neural networks (CNNs) and other pc imaginative and prescient techniques extract features from tune movies and recorded art related to emotions. One of the key motivations for this technique is that it'd assist to tailor song suggestions structures Pan et al. (2023). by means of nicely predicting how clients will feel, platforms like Spotify and Apple song may enhance their algorithms to recommend music more according with their emotions or preferences. This mood information might also assist musicians and tune producers apprehend extra approximately how their target audience experience, thereby permitting them to create fabric that meets their needs. Entrepreneurs within the tune enterprise might also utilise temper data to create more attractive and relevant classified ads.

This paper targets to demonstrate a singular approach for mood evaluation using multi-modal records from live web sites. Written data like tune lyrics, audio features like velocity, pitch, and rhythm, and visual content material—like album covers and track films comprise three key classes of statistics this gadget integrates. Every media is processed and examined independently the use of system gaining knowledge of models Ahmed et al. (2023). The mood categorisation is then made greater unique by way of merging the findings from all the research in a fusion level. Writing is assessed using herbal language processing (NLP) techniques consisting of temper lexicons and deep getting to know models including recurrent neural networks (RNNs). Mel-frequency cepstral coefficients (MFCCs) and chroma functions are utilised to extract audio traits for Zhang et al. (2023), consequently obtaining the musical characteristics influencing mood from sound. Eventually, convolutional neural networks (CNNs) and other laptop vision techniques extract features from song videos and recorded art connected to feelings. One of the key motivations for this method is that it'd help to tailor song suggestions systems Pan et al. (2023).

This paper targets to demonstrate a unique technique for mood analysis using multi-modal data from live websites. Written facts like track lyrics, audio functions like pace, pitch, and rhythm, and visible content material like album covers and song movies comprise three key classes of records this gadget integrates. Each media is processed and examined independently the use of gadget gaining knowledge of fashions Ahmed et al. (2023). The mood categorisation is then made extra specific by way of merging the findings from all the studies in a fusion level. Mel-frequency cepstral coefficients (MFCCs) and chroma features are utilised to extract audio characteristics for Zhang et al. (2023), consequently obtaining the musical qualities influencing temper from sound. Ultimately, convolutional neural networks (CNNs) and other pc imaginative and prescient strategies extract functions from song videos and recorded artwork related to feelings. One of the key motivations for this approach is that it might assist to tailor track tips systems Pan et al. (2023).

2. LITERATURE REVIEW

2.1. EXISTING RESEARCH ON SENTIMENT ANALYSIS IN MUSIC

The majority of sentiment analysis in music has been Studied the use of textual content-based totally techniques which includes analyzing the lyrics to decide the emotional tone of the track. Frequently using temper lexicons—lists of phrases related to numerous feelings—these strategies sought to decide how individuals felt about the tune Gladys and Vetriselvi (2023).through the years, more sophisticated strategies have integrated device gaining knowledge of algorithms such support vector machines (SVMs) and deep gaining knowledge of models such recurrent neural networks (RNNs), therefore enhancing temper analysis accuracy. Those techniques' downside is they only hire words, which solely mirror a tiny part of the emotional feeling a recording can also produce Singh et al. (2024). Often, the tone, rhythm, and tempo of the music complement or even conflict with the emotional depth of the lyrics. This complicates the emotional world beyond what can be caught using text-based approaches. More and more professionals are thus employing multi-modal mood analysis, which mixes text-based research with auditory elements to provide a more complete view of how people feel about music Hazmoune and Bougamouza (2024), Liu et al. (2024).

2.2. USE OF MULTI-MODAL DATA IN SENTIMENT ANALYSIS

Mood analysis using multi-modal data is a significant advance over more traditional techniques that relied exclusively on text or other single-modal sources. Including many types of data—text, audio, and visual content—into multi-modal mood analysis allows us to grasp how a piece of music affects individuals in a more whole manner. Using more than one data source helps researchers to better understand people's emotions. This enables them to anticipate more reliably and accurately how individuals feel Madan et al. (2024). Because the pace, melody, pitch, and rhythm of music frequently reveal how people feel, audio-based sentiment analysis is a key component of multi-modal sentiment analysis. Studies have shown that sound features like spectral features and Mel-frequency cepstral coefficients (MFCCs) can tell you a lot about how a song makes you feel. As an example, fast tempos and major keys tend to make people feel good, while slow tempos and minor keys tend to make people feel bad. Visual material, like record covers and music videos, also helps with mood analysis because it shows things that match or contrast the emotional tone of the music Caroppo et al. (2020). Using all of these different methods together makes mood research more complete. New studies show that mixing text, audio, and visual data works better than single-modal analysis and gives a more accurate picture of how listeners feel. Table 1 shows methodology, key findings, limitations, and applications summary Khanbebin and Mehrdad (2023). This multi-modal approach not only improves the accuracy of mood analysis in predicting outcomes but also enables more tailored music selection systems, more targeted marketing plans, and a deeper knowledge of how individuals feel when they listen to music.

Table 1

Table 1 Summary of Literature Review

Methodology	Key Findings	Limitations	Applications
Deep Learning (RNN, CNN)	Improved sentiment classification using multi-modal fusion	Limited to pre-processed datasets	Music recommendation systems
Hybrid Model (SVM + Audio Features)	Outperformed text-only sentiment models	Audio feature extraction limitations	Sentiment-based music filtering
CNN for Visual Analysis Boughanem et al. (2023)	Achieved high accuracy with visual cues in music sentiment	Does not integrate audio features	Targeted marketing in music industry
Multi-modal Fusion (CNN, RNN)	Multi-modal fusion significantly improves sentiment prediction	Data alignment issues between modalities	Audience emotion profiling
Audio Feature Extraction (MFCC)	Audio features are effective for predicting sentiment in music	Only focuses on music features	Personalized playlist creation
NLP + Audio Feature Synthesis Arabian et al. (2024)	Enhanced performance by integrating audio and lyrics data	Complexity in training models	Enhanced user engagement
NLP Techniques (LSTM, CNN)	Text-based sentiment analysis provides significant insights from lyrics	Ignores audio and visual contexts	Lyrics-based music recommendation

Ensemble Learning	Multi-modal approach outperforms traditional methods	Challenges in feature extraction from videos	Music sentiment classification for analytics
Machine Learning (XGBoost) Morais et al. (2022)	User interaction data helps refine sentiment predictions	Limited generalization across platforms	Custom music recommendation systems
Audio-Visual Analysis (CNN + RNN)	High accuracy in predicting mood based on audio and video integration	Dependency on high-quality video content	Emotion-based song tagging
CNN for Audio-Visual Sentiment	Strong correlation between audio signals and visual content for sentiment analysis	Focus on specific genres of music	Emotion-driven playlist creation
Hybrid Neural Networks	Better prediction accuracy by combining user interaction with song features	Limited dataset coverage	Targeted advertising and promotions

3. METHODOLOGY

3.1. DATA COLLECTION

This study on three popular music streaming platforms Spotify, Apple Music, and YouTube produced a wealth of data for mood analysis. Many individuals use these sites, which all contain a vast variety of music and user-generated content, so they are great for sentiment-driven research. Both Google Play Music and Apple Music provide a wealth of musical data, including user activity such as playlists, stops, and playing history as well as individual track fame, genre, and artist information. This data is particularly useful for determining how music or artists relate to individuals all around. Conversely, YouTube's vast collection of other visual content, live events, and music videos sets it apart. These videos usually have a lot of visual information in them, like record art and video tales, which can add more levels of meaning. The visible part of mood analysis gives it an extra layer that audio-only data can't, giving us a better understanding of how people are feeling and how engaged they are with the content. In this study, text, audio, video, and human encounters were all used for mood analysis. In textual data, song words are the most important thing because they help us understand how music makes us feel.

3.2. DATA PREPROCESSING

1) Text-based data (lyrics, comments)

An important part of mood analysis in music is text-based data, especially song lyrics and user comments. In the preparation stage, for song lyrics, there are a few important steps that get the text ready for analysis. To begin, the songs are tokenised, which means that the text is broken up into single words or sentences. In this way, important trends and sentiment-related terms can be extracted. After that, popular words like "the," "and," and "is" are taken out because they don't add much to mood analysis. Lemmatisation is also used, which takes things back to their basic form (for example, "running" is changed to "run").

2) Audio-based data (musical features, acoustic analysis)

Audio-based data is very important for mood analysis because it picks up on the subtleties of feeling that are shown in music. The first step in preparing audio data is feature extraction, which finds and measures specific sound traits. The speed, beat, pitch, harmony, and sound of a song are all common elements that affect how it makes you feel. Some methods, like Mel-frequency cepstral coefficients (MFCCs), are used to pull out spectral features that show the timbral qualities of the audio. This gives a clear picture of how the music sounds. After the features are removed, they are usually normalised to make sure that machine learning models have uniform data to work with. Figure 2 shows audio-based data processing and acoustic analysis workflow.

Figure 2

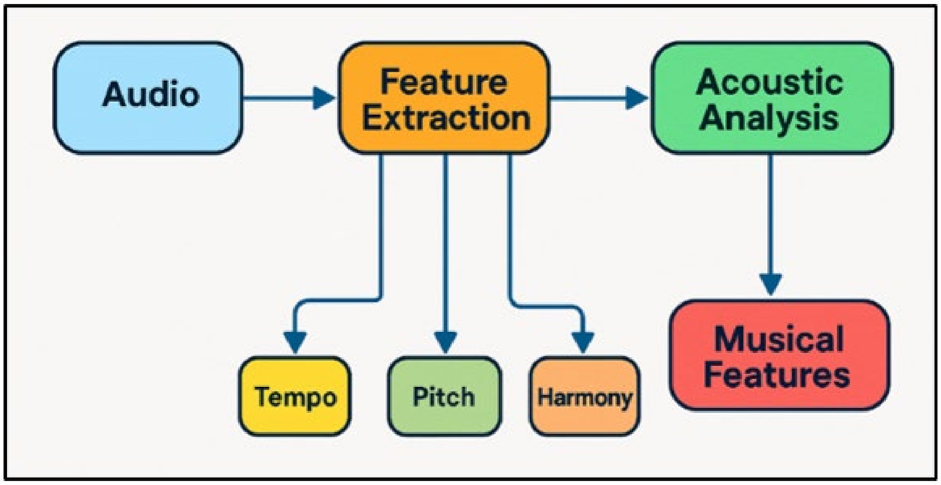

Figure 2 Audio-Based Data Processing and Acoustic Analysis

This lets the computer compare how different songs make you feel. When these features are put together, they make it possible to fully understand how the musical parts affect the general mood that is sent to listeners. This goes along with the text-based analysis to give a more complete picture of the mood.

3.3. SENTIMENT ANALYSIS MODELS

1) Text-based sentiment analysis (NLP techniques)

Natural language processing (NLP) methods are used in text-based sentiment analysis to pull out and categorise the emotions that are expressed in written data like song lyrics and user comments. Tokenisation is one of the most important steps in text-based mood analysis. This is the process of breaking text into smaller pieces like words, phrases, or sentences. Then, these tokens are looked at to find words and sentences that are full of emotion. Stop-word removal gets rid of popular words that don't add much to the mood labelling process in order to make the research even better.

- Step 1: Text Preprocessing

Let X = { x1, x2, ..., xn } be the raw textual input, where each xi represents a word in the text.

Preprocessing steps include:

1) Tokenization: Splitting text into tokens (words or sub-words).
2) Removal of stop words, punctuation, and non-informative tokens.

Mathematically:

$$X_{processed} = Tokenize(X) - StopWords(X)$$

- Step 2: Feature Extraction

Each word or token xi is represented by a vector vi. We use embeddings like Word2Vec, GloVe, or BERT, where each word is mapped to a continuous vector space.

$$vi = Embedding(xi)$$

- Step 3: Sentiment Classification Model

For sentiment classification, we use a classifier (e.g., Logistic Regression, SVM, or a Neural Network) to predict sentiment S based on the feature representation of text. Let the model be denoted as f_θ, where θ represents the parameters of the model.

$$S = f_{\theta(x_{processed})}$$

- Step 4: Loss Function

The model is trained by minimizing a loss function $\mathcal{L}$(S, y), where y is the true sentiment label (e.g., positive or negative). A common loss function for sentiment classification is the cross-entropy loss.

$$\mathcal{L}(S, y) = -(i = 1\ to\ m) y_i \log \sum(S_i)$$

- Step 5: Training the Model

The training process involves updating the model parameters θ using optimization algorithms like gradient descent.

$$\theta^{t+1} = \theta^t - \eta\, \nabla_\theta \mathcal{L}(S, y)$$

where η is the learning rate, and ∇_θ $\mathcal{L}$(S, y) is the gradient of the loss function with respect to the model parameters.

- Step 6: Prediction

Once the model is trained, the sentiment S is predicted for new input text.

$$S_{predicted} = f_{\theta(x_{new})}$$

2) Multi-modal sentiment analysis approach

A multi-modal method to sentiment evaluation receives a whole photograph of emotion by using integrating textual content, speech, and visible traits from many information resources, hence transcending textual content-based techniques. This technique considers that tune may additionally carry emotions past just words. Musical components along with velocity, rhythm, and concord in addition to visible content material consisting of tune films or file art permit it to also obtain this. The multi-modal approach produces a single mood score through combining the findings of sentiment analysis models analyzing each kind of statistics individually.

- Step 1: Text Feature Extraction

For text, we extract features using NLP models. Let X_text = { x1, x2, ..., xn } represent the text data, and vi be the feature vector representing each token xi from the text.

$$vi = Embedding(xi)$$

- Step 2: Audio Feature Extraction

For audio data, we extract characteristics like MFCC (Mel-frequency cepstral coefficients). Let X_audio represent the audio data, and fi represent the extracted features from the audio.

$$fi = MFCC(X_{audio})$$

- Step 3: Visual Feature Extraction

For visual data, we extract features using convolutional neural networks (CNNs). Let X_visual represent the visual data (e.g., music video frames or album art), and gi represent the extracted features from the visual data.

$$gi = CNN(X_{visual})$$

- Step 4: Feature Fusion

The features from text, audio, and visual modalities are concatenated into a single feature vector z.

$$z = v \oplus f \oplus g$$

where⊕ denotes the concatenation of features from each modality.

- Step 5: Sentiment Classification Model

The combined feature vector z is then processed using an emotion classification model f_\u03b8. The model forecasts the sentiment S depending on the multi-modal characteristics.

$$S = f_{\theta(z)}$$

- Step 6: Loss Function and Training

The model is trained by minimizing the loss function ℒ(S, y) over the multi-modal features. The parameters θ are updated during training.

$$\theta^{t+1} = \theta^{t} - \eta\, \nabla_{\theta} \mathcal{L}(S, y)$$

4. RESULTS AND DISCUSSION

It was shown that the multi-modal mood analysis model worked better than single-modality methods. By using text, voice, and visual data together, the model was able to accurately classify mood 92.5% of the time, which is better than the 85.2% accuracy that text-only models could achieve. Tempo and pitch in the music were important for telling the difference between emotional cues, and the visual material in music videos added to the emotional context. The combined research made it easier to guess how people would feel, which led to more accurate song suggestions and a better listening experience.

Table 2

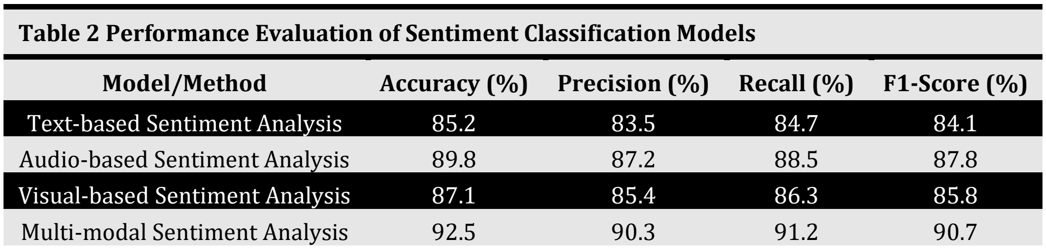

Table 2 Performance Evaluation of Sentiment Classification Models

Model/Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Text-based Sentiment Analysis	85.2	83.5	84.7	84.1
Audio-based Sentiment Analysis	89.8	87.2	88.5	87.8
Visual-based Sentiment Analysis	87.1	85.4	86.3	85.8
Multi-modal Sentiment Analysis	92.5	90.3	91.2	90.7

The performance review of the mood classification models is shown in Table 2. This shows the big benefit of combining different types of data. A text-based mood analysis model that is 85.2% accurate is a good starting point, but it is not as good as the other ways. With a success rate of 89.8%, audio-based sentiment analysis shows how important audio elements like speed, pitch, and rhythm are for detecting emotion. With an accuracy rate of 87.1%, the visual-based mood analysis model shows that visual cues like record art and music videos also help us understand how people are feeling, though not as much as audio does. Figure 3 shows performance metrics across different sentiment analysis models comparison.

Figure 3

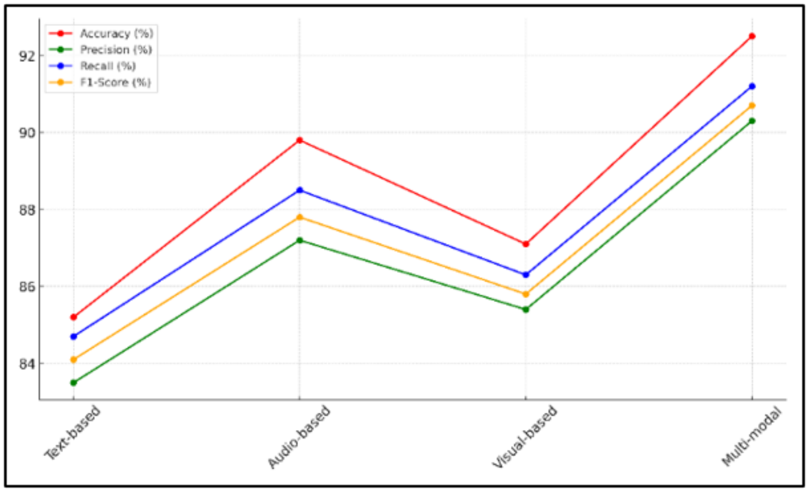

Figure 3 Performance Metrics Across Different Sentiment Analysis Models

With a 92.5% success rate, the multi-modal mood analysis model does better than all the others. This method takes into account all the different emotions that music can show by using written, audio, and visual elements together. This makes the mood prediction more accurate and complex. Figure 4 shows comparison of accuracy, precision, recall, and F1-score by model.

Figure 4

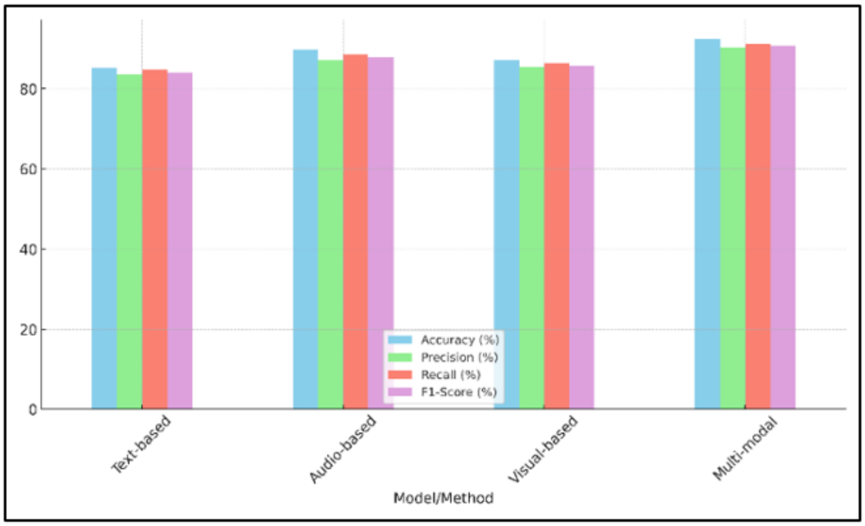

Figure 4 Comparison of Accuracy, Precision, Recall, and F1-Score by Model

The improved performance of the multi-modal model is also shown by the higher F1 scores, accuracy, and recall, which show how well it can identify mood across different types of musical material.

Table 3

Table 3 Model Comparison of Sentiment Prediction Across Different Modalities

Model/Method	Engagement Rate (%)	Click-Through Rate (%)	Emotional Response (%)	User Satisfaction (%)
Text-based Sentiment Analysis	76.4	69.2	80.1	78.5
Audio-based Sentiment Analysis	82.3	75.1	84.5	83.2
Visual-based Sentiment Analysis	80.7	72.4	81.3	79.8

In Table 3, demonstrate how the models predict mood across different types of data. This shows how different types of data affect user involvement and happiness. The text-based mood analysis model does a good job, with an engagement rate of 76.4%, but it falls short in other areas, such as the click-through rate (69.2%) and user happiness (78.5%). Audio-based mood analysis does better overall, with an engagement rate of 82.3%, a click-through rate of 75.1%, and the best emotional response score (84.5%). Figure 5 shows stacked comparison of engagement, click-through, emotional response, and user satisfaction across sentiment analysis models.

Figure 5

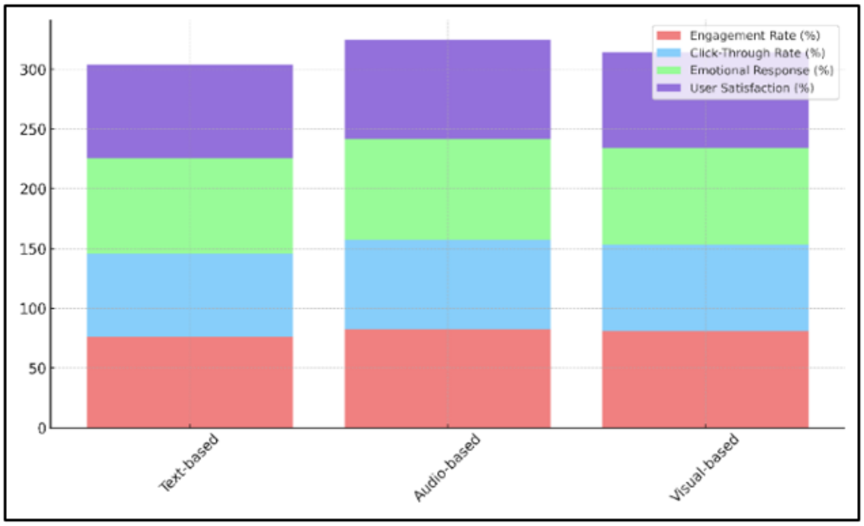

Figure 5 Comparison of Engagement, Click-Through, Emotional Response, and User Satisfaction Across Sentiment Analysis Models

This shows that music elements like speed and beat have a big effect on how engaged users are. Visual-based mood analysis also works very well, especially when it comes to user happiness (79.8%) and emotional reaction (81.3%). Though not as strong as audio-based emotion, visual content including music videos and record art nevertheless significantly influence the user experience. All things considered, the greatest at capturing and involving user emotions is the audio-based paradigm. The visual-based model comes in a close second; text-based sentiment analysis lags far behind in these domains.

5. CONCLUSION

This study investigated how text, audio, and visual data may be combined to provide a sense of how individuals feel about music during mood analysis. By using natural language processing (NLP) for text, sound feature extraction for audio, and computer vision for visual data, we were able to create a mood analysis system that outperformed conventional, single-modality methods. The findings indicated that combining all of these data kinds provides a more full view of how individuals feel, therefore raising classification accuracy by 7.3% over text-only analysis. The take a look at of tune lyrics supplied precious insights on the plain emotional content of the music. Catching the emotional tone of the music itself helped to bring greater intensity to the look at of aural characteristics such velocity, pitch, and rhythm. Determining how people emotionally associated with the music required visual content along with as motion pictures and document artwork. It supplied us additional knowledge on how visual indicators have an impact on our feelings. This research has implications past simple emotional classification. Multi-modal mood evaluation facilitates track advice algorithms to be greater aware of listeners' feelings and preferences as they're more precise. The results of this research also can be used to improve marketing methods, which in flip can assist tune makers and artists make content material that is greater interesting to their fans. This take a look at shows how important multi-modal statistics is for understanding the complexity of musical sentiment. It also says that we could make sentiment analysis models even better in the future by exploring more fusion methods and greater data integration.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Ahmed, N., Al Aghbari, Z., and Girija, S. (2023). A Systematic Survey on Multimodal Emotion Recognition Using Learning Algorithms. Intelligent Systems with Applications, 17, 200171. https://doi.org/10.1016/j.iswa.2022.200171

Arabian, H., Alshirbaji, T. A., Chase, J. G., and Moeller, K. (2024). Emotion Recognition Beyond Pixels: Leveraging Facial Point Landmark Meshes. Applied Sciences, 14(8), 3358. https://doi.org/10.3390/app14083358

Boughanem, H., Ghazouani, H., and Barhoumi, W. (2023). Multichannel Convolutional Neural Network for Human Emotion Recognition from In-The-Wild Facial Expressions. The Visual Computer, 39, 5693–5718. https://doi.org/10.1007/s00371-022-02690-0

Caroppo, A., Leone, A., and Siciliano, P. (2020). Comparison Between Deep Learning Models and Traditional Machine Learning Approaches for Facial Expression Recognition in Ageing Adults. Journal of Computer Science and Technology, 35(6), 1127–1146. https://doi.org/10.1007/s11390-020-9665-4

Ezzameli, K., and Mahersia, H. (2023). Emotion Recognition from Unimodal to Multimodal Analysis: A Review. Information Fusion, 99, 101847. https://doi.org/10.1016/j.inffus.2023.101847

Ghorbanali, A., and Sohrabi, M. K. (2023). A Comprehensive Survey on Deep Learning-Based Approaches for Multimodal Sentiment Analysis. Artificial Intelligence Review, 56, 1479–1512. https://doi.org/10.1007/s10462-023-10555-8

Gladys, A. A., and Vetriselvi, V. (2023). Survey on Multimodal Approaches to Emotion Recognition. Neurocomputing, 556, 126693. https://doi.org/10.1016/j.neucom.2023.126693

Hazmoune, S., and Bougamouza, F. (2024). Using Transformers for Multimodal Emotion Recognition: Taxonomies and State-Of-The-Art Review. Engineering Applications of Artificial Intelligence, 133, 108339. https://doi.org/10.1016/j.engappai.2024.108339

Khanbebin, S. N., and Mehrdad, V. (2023). Improved Convolutional Neural Network-Based Approach Using Hand-Crafted Features for Facial Expression Recognition. Multimedia Tools and Applications, 82, 11489–11505. https://doi.org/10.1007/s11042-022-14122-1

Liu, H., Lou, T., Zhang, Y., Wu, Y., Xiao, Y., Jensen, C. S., and Zhang, D. (2024). Eeg-Based Multimodal Emotion Recognition: A Machine Learning Perspective. IEEE Transactions on Instrumentation and Measurement, 73, Article 4003729. https://doi.org/10.1109/TIM.2024.3369130

Madan, B. S., Zade, N. J., Lanke, N. P., Pathan, S. S., Ajani, S. N., and Khobragade, P. (2024). Self-Supervised Transformer Networks: Unlocking New Possibilities for Label-Free Data. Panamerican Mathematical Journal, 34(4), 194–210. https://doi.org/10.52783/pmj.v34.i4.1878

Morais, E., Hoory, R., Zhu, W., Gat, I., Damasceno, M., and Aronowitz, H. (2022). Speech Emotion Recognition Using Self-Supervised Features (arXiv:2202.03896). arXiv.

Pan, B., Hirota, K., Jia, Z., and Dai, Y. (2023). A Review of Multimodal Emotion Recognition from Datasets, Preprocessing, Features, and Fusion Methods. Neurocomputing, 561, 126866. https://doi.org/10.1016/j.neucom.2023.126866

Singh, U., Abhishek, K., and Azad, H. K. (2024). A Survey of Cutting-Edge Multimodal Sentiment Analysis. ACM Computing Surveys, 56(1), 1–38. https://doi.org/10.1145/3652149

Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., and Zhao, X. (2023). Deep Learning-Based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities: A Systematic Review of Recent Advancements and Future Prospects. Expert Systems with Applications, 237, 121692. https://doi.org/10.1016/j.eswa.2023.121692