SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I6P123

Volume 12 Issue 6, 293-303, June 2025

Original Article

Enhancing Music Emotion Recognition with LSTM: Evaluating Various Embedding Techniques

Affreen Ara[1], Rekha V[2]

[1,2]*Department of Computer Science and Engineering, Christ University, Karnataka, India.*

[1]*Corresponding Author : affreen.ara@res.christuniversity.in*

Received: 12 April 2025 Revised: 13 May 2025 Accepted: 14 June 2025 Published: 27 June 2025

Abstract *- The study investigates the application of Long Short-Term Memory (LSTM) networks for emotion classification in music lyrics. It focuses on the comparative effectiveness of various word embedding techniques. It evaluates the performance of static embeddings (GloVe, Word2Vec, FastText) versus contextual embeddings (BERT, Distil BERT) across three datasets: MER Lyrics, Mood Lyrics, and Combined Lyrics. Additionally, the study examines the role of stylistic and content-based features in enhancing classification accuracy. The results demonstrate that contextual embeddings considerably outperform static embeddings, achieving accuracy rates of up to 98% compared to 60% for static approaches. Moreover, combining multiple lyric datasets leads to improved model generalization. The findings show the potential of transformer-based models for advancing music emotion recognition. Future research will focus on optimizing large embedding models using techniques such as pruning, quantization, and distillation to enhance computational efficiency.*

Keywords *- LSTM, Embedding, BERT, Emotion classification and Emotion.*

1. Introduction

Online music platforms are the primary medium for music discovery and consumption; users increasingly engage with music on an emotional level. This engagement drives demand for personalized recommendations and has spurred a growing interest in Music Emotion Recognition (MER). MER is to discern emotions-such as anger, fear, sadness, joy, and surprise from music lyric text. Music Lyrics are an important resource for emotion classification because they have emotion embedded in them through word choice, syntactic structure, and stylistic patterns. Therefore, analyzing lyrics not only enables the classification of emotional categories but also provides insights into how a listener's emotional state is affected by music.

Word embedding methods form the basis of converting raw text into numerical form, maintaining semantic meaning. Static embeddings like GloVe, Word2Vec, and FastText build dense vectors from co-occurrence statistics and morphological attributes. GloVe, for instance, learns word representations by studying co-occurrence ratios and effectively capturing semantic relationships such as similarity and analogy. FastText builds on this by using words as bags of character n-grams. It can process rare word vocabulary and capture morphological subtleties in lyric text [1]. Static embeddings are context-free and do not capture word meanings across semantic contexts. Conversely, transformer models such as BERT produce contextual embeddings. An important characteristic of this model is the presence of self-attention mechanisms in bidirectional architecture [2, 3]. The BERT model transforms input tokens by means of several transformer layers that identify nuanced semantic relationships and distant dependencies. This is beneficial for emotion recognition tasks using lyrics [2, 3]. LSTM networks have proven efficient in processing chronological text data for emotion recognition, with multimodal residual LSTM architectures demonstrating [4] better performance. They effectively handle chronological text processing while capturing sensitivity and emotional patterns in music text.

Pre-trained BERT models optimized on emotion datasets improve emotion nuances expressed in lyric text. [2, 5]. Static embeddings provide a rudimentary sense of word meaning, but new research reveals that transformer-based embeddings excel at extracting emotional nuances in lyrics. Methods such as t-SNE validate the enhanced clustering of emotional categories. However, incorporating stylistic and semantic attributes can make the model more complex, sometimes lowering accuracy because of noise. This research identifies a notable gap regarding the limited comparison of static and contextual word embedding methods for emotion detection in music lyrics. This is reflected in the absence of an overarching framework that addresses Hyperparameter tuning and changes in datasets [2, 4]. Previous research has established the effectiveness of transformer-based architectures like BERT and Distil BERT for tasks in emotion

classification. Test their performance compared to static embeddings such as GloVe, Word2Vec, and FastText on many datasets is required. In addition, the investigation of how the model performance is affected by various stylistic, content-based attributes and other features has not yet been explored. This research proposes that extra features might sometimes reduce performance because they add more complexity [2, 4].

The key contributions of this work include: (1) Systematic evaluation and comparison of five different embedding techniques (GloVe, Word2Vec, FastText, BERT, and Distil BERT) within a unified LSTM-based framework for music emotion recognition. (2) Extensive testing across three distinct music lyric datasets (MER Lyrics, Mood Lyrics, and Combined Lyrics) to ensure robustness and generalizability of findings. (3) Establishment of clear performance benchmarks showing the significant accuracy gap between static embeddings (maximum 60%) and transformer-based embeddings (up to 98%).(4) Investigation of the impact of additional stylistic and lexical attributes on classification performance, revealing both benefits and potential drawbacks of feature complexity. (5)Empirical evidence of transformer-based models' superior ability to capture nuanced emotional content in lyrics compared to traditional static embeddings. (6) Analyse computational efficiency trade-offs and practical considerations for deploying different embedding techniques in real-world MER systems.

The uniqueness of this work lies in its comprehensive evaluation of static and transformer-based embedding techniques using an LSTM model across three distinct music lyric datasets, including MER, Mood Lyrics, and a combined dataset. The study achieves an accuracy of 98%, which surpasses many previous results, and it shows the potential of transformer-based models for music emotion recognition [2, 4]. This research explores how additional features can improve or hinder classification performance, providing insights for effective feature integration. A systematic comparison is conducted, and the result indicates good accuracy and a significant advancement in music emotion recognition using natural language processing techniques. The paper is organized as follows: the literature review discusses previous work in emotion recognition using LSTM models, MER, and Mood Lyrics dataset, followed by the methodology section detailing the experimental setup, feature extraction methods, and LSTM model architecture. The results and discussion are presented next, followed by a conclusion and future research directions.

2. Literature Review

Emotion classification from music lyrics has become a vital research area driven by applications in personalized recommendation systems, affective computing, and music therapy. Traditional machine learning techniques such as Support Vector Machines (SVM), Naive Bayes classifiers, and lexicon-based approaches were limited in capturing the word semantics, figurative language, and subjective interpretation of lyrics. To overcome these challenges, deep learning models-especially Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), and more recently, Transformer-based architectures-have emerged as powerful alternatives for modelling sequential and contextual data in textual emotion recognition.

This literature review encompasses recent research contributions in Music Emotion Recognition (MER), with a particular emphasis on studies involving emotion-based lyric analysis. The review includes papers based on applying deep learning techniques such as Long Short-Term Memory (LSTM) networks and Transformer-based models. Limited research uses music lyrics as a key modality for MER, leveraging emotion-tagged lyric datasets to infer emotional content. These works often utilize emotion-specific datasets such as the MER Lyrics and Mood Lyrics datasets, which provide valuable benchmarks for emotion classification tasks.

Transformer architectures, especially those using self-attention mechanisms, have shown superior performance in emotion classification tasks. Previously employed an XLNet model with a bidirectional LSTM for classifying emotions in music lyrics, achieving an accuracy of 84.3%, surpassing conventional deep learning models. Similarly, Loreto et al. [6] study propose a hybrid approach that combines synchronized lyrical content with vocal features to improve emotion classification accuracy. The study uses word embeddings (e.g., FastText) and architectures like Bidirectional LSTM with attention mechanisms. Previously designs transformer architecture for music emotion recognition with attention mechanisms to enhance feature extraction. The transformer architecture with BERT embeddings achieves 83.5% accuracy on a Mood lyrics dataset.

Despite the dominance of transformer models, LSTM-based architectures continue to be used for their efficiency and effectiveness in sequence modelling. Cong Jin et al. [7] propose a sentiment analysis model based on Bi-DLSTM using Beijing Opera lyrics. A Bi-LSTM network with dilated recurrent skip connections (Bi-DLSTM) is used along with an attention mechanism. This attention mechanism ensures that important words are in the text sequence.

Previously work integrates TF-IDF and Word2Vec features within CNN and LSTM frameworks, implementing improved attention mechanisms to better capture emotionally salient words. The CNN-LSTM model achieves an accuracy of 85.7%. Jia et al. [8] propose a hybrid LTM-GRU model emphasising regularization techniques to mitigate overfitting, achieving 72.51% accuracy.

Pyrovolakis et al. [9] utilize the audio and lyrics of a musical track separately and together for emotion classification. The study also evaluates the performance of different embeddings (GloVe, BERT, and Word2Vec).

The BERT achieves 69.11% accuracy, and traditional embeddings like Word2Vec and GloVe achieve 41.66% and 53.33%. Delbouys et al. [10] aim to enhance music mood detection by integrating audio and lyrics. The study uses deep learning techniques to predict emotion dimensions of valence and arousal.

Previously introduced MoodNet, which combines lyrics and audio data and attained 87.2% accuracy. Shaday et al. [11] evaluate how different word embedding methods, such as Word2Vec, GloVe, and FastText, affect the emotion classification accuracy of Bi-LSTM.

The Mood Lyrics dataset achieves an accuracy of 62%. Agarwal et al. [12] implemented XLNet with bidirectional LSTM and achieved 84.3% accuracy.

R. Guru et al. [13] work compares embedding techniques such as Word2Vec and FastText with LSTM architectures. The study achieves an F1-score of 0.88 and confirms that pre-trained embeddings significantly improve the semantic representation of lyrics.

2.1. Research Gap

Although transformer models and ensemble approaches have shown promising results, several gaps remain. First, there is a lack of systematic comparison between Transformer-based methods and optimized LSTM architectures using classical embeddings like GloVe or FastText. Most studies prioritize state-of-the-art models but overlook simpler models that may offer lower computational costs with competitive accuracy. Second, dataset limitations are prevalent.

Many models are trained and evaluated on small, pre-annotated datasets like Mood Lyrics, limiting generalizability and reproducibility. Few studies address how their models perform on larger, noisier, or sparsely annotated datasets, which are more realistic in practical applications.

3. Methodology for Emotion Classification using LSTM Models

Figure 1 illustrates the methodology for classifying music lyrics using an LSTM (Long Short-Term Memory) model. The process begins with the input lyrics undergoing text preprocessing, including tokenization, lowercasing, and removing stop words or special characters. The pre-processed lyric text is then transformed into a numerical form, such as word embeddings (e.g., Word2Vec, GloVe or BERT) to capture semantic meaning.

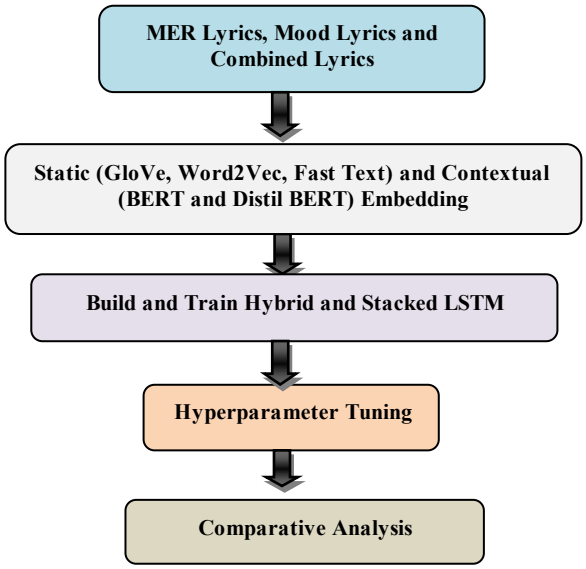

Fig. 1 Methodology for emotion classification using LSTM models

The LSTM model is trained on output numeric representation, using its ability to learn long-term dependencies in sequential text. Parameter Tuning helps improve model accuracy and robustness for emotion classification. The model outputs are Russell Quadrants (Q1, Q2, Q3 and Q4) for the input lyrics. Lastly, classification accuracy and other metrics are evaluated for MER, Mood, and Combined Lyric datasets, and the results of contextual and static embedding are compared.

3.1. Emotion Dataset

The lyrics (English) are taken from the Moody Lyrics [14] and Ricardo Malheiro [15] datasets. The Ricardo Malheiro et al. (2017) dataset contains 771 emotion-labeled lyrics taken from the AllMusic platform. MER dataset songs are labeled with four Russell Quadrants: Q1, Q2, Q3 and Q4. The Mood Lyrics is a sentiment-annotated dataset. Songs are labeled with four Russell Quadrants of Russell's 2D model with output classes as sad, relaxed, angry and happy. MER and Mood Lyrics are combined, which resulted in a new dataset comprising 2680 music lyrics with four uniform Russell Quadrants (Q1, Q2, Q3 and Q4).

3.2. Lyric Preprocessing

The input lyrics are fed into the emotion classification model, followed by preprocessing steps. The lyrics text is cleaned using methods such as punctuation, special characters, stop words, and tokenization. Lemmatization or stemming can also reduce words to their base forms.

3.3. Feature Extraction Techniques

3.3.1. Traditional Text Representation

The tokenized lyric text is POS-tagged. POS tagging refers to the identification of words related to parts of speech, i.e., nouns, verbs, adjectives, or adverbs. POS tagging aids in

syntactic structure understanding of the lyrics. TF-IDF and Bag of Words (BOW) are basic text representation techniques that provide different insights for emotion classification. The TF-IDF method weights words according to their prevalence in a document against the whole corpus, highlighting unique and informative terms. BOW represents text as word-frequency-based vectors that identify rudimentary patterns independent of grammar and word order. Also, length-based features, including word, character, or sentence count, not count and BE verbs count, contain structural and stylistic information that can augment these representations.

3.3.2. Word Embedding Techniques

Glove [5] is a word embedding that learns word representation through the examination of words' co-occurrence statistics within a corpus. It captures semantic word relationships and applies to applications like word similarity, analogy, and text classification. Glove embeddings are generally dense vectors capturing the meaning of a word in a high-dimensional space. The word embedding scheme associates words into high-dimensional vectors**.** In contrast to one-hot encoding, where words are encoded as sparse vectors with binary values, GloVe encodes dense vector form that encode the semantic meaning of words depending on the context where they are used. GloVe seeks to preserve the ratios of co-occurrence probabilities. For example, if the word "ice" often occurs with "cold" and rarely with "steam," the ratio of the two probabilities should be evident in the distance between their corresponding embeddings. After training, each word in the dictionary is encoded in form of a fixed dense vector**.**

Word2Vec captures semantic relationships between words. Word2Vec is trained on the song lyric text using the skip-gram model with negative sampling. The embedding dimension is set to 300 to ensure uniformity across all models. A series of word vectors are created from each lyric and fed into the LSTM model.

FastText [16] is a sophisticated variant of the Word2Vec method, which takes subword information into account. In contrast to Word2Vec, In FastText, words are represented as character n-gram bags, dividing them into subparts like "class," "lass," and "as if" for the word "classification." Fast Text's capacity to embrace morphological structure within words makes it especially efficient for dealing with rare or Out-Of-Vocabulary (OOV) words. A training corpus word might not have a vector, but inferring its vector from its sub-word units is still possible. FastText can train big data sets while still keeping up with the capacity to capture syntactic and meaningful word relationships. Sub-word-level representation in FastText in sentiment classification is useful when words are segmented into sub-words. It retains the emotional meaning of the text for misspelt, unseen, or rare words. The model's ability to retain word-level and sub-word-level information gives the meaning of text for misspelt, unseen, or rare words. The model's ability to retain word-level and sub-word-level information allows it to extract strong features, improving emotion classification performance.

3.3.3. Transformer-Based Embeddings

Bidirectional Encoder [9] Representations from Transformers (BERT) is a two-way representation that allows BERT to learn how language works and how words depend on each other. Using self-attention to parse input text in parallel, BERT manages long-range dependencies while utilizing the Transformer design. BERT leverages a transformer-based architecture that captures bidirectional contextual representations of text.

For emotion classification, BERT's pretrained models are fine-tuned on emotion-specific datasets, enabling a deep contextual understanding of linguistic nuances. The model generates contextual embeddings by processing input tokens through multiple transformer layers, which learn complex semantic relationships and capture emotional subtleties across different sentence structures.

Distil BERT [3] is a compressed version of the original BERT model, designed to reduce computational complexity while maintaining high performance. With 40% fewer parameters and 60% better performance, it keeps 97% of BERT's language understanding potential. Distil BERT reduces complexity by using 6 transformer layers (compared to BERT's 12) and removing components like token-type embeddings and the pooler.

Distil BERT maintains much of the accuracy of the original BERT model while being more competent, which makes it appropriate for real-time applications and deployment in production environments with few computational resources. The compact size and quick processing of the model facilitate quicker inference and less memory consumption. Distil BERT offers a computationally lightweight substitute that retains vital semantic representation for emotion classification. It is appropriate for resource-limited settings with little performance loss.

3.4. LSTM Model Framework for Emotion Classification

3.4.1. Feature Integration

For the hybrid feature integration approach, the model combines numeric features and word embeddings such as GloVe, BERT, etc. Word embeddings can infer the semantic meaning of words. Numeric features and stylistic features are used for feature fusion. Contextual embeddings are directly used for BERT-based models. GloVe and Word2Vec embeddings contribute dense representations of words, focusing on their semantic relationships. This feature fusion enables the model to leverage the strengths of each feature type for enhanced emotion classification.

3.4.2. Model Architecture

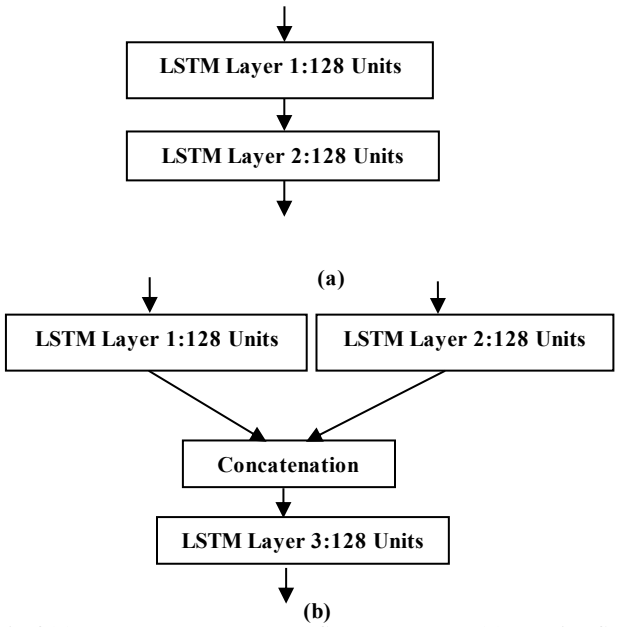

Fig. 2(a) Fully connected stacked LSTM Model, and (b) Hybrid LSTM model.

The Long Short-Term Memory (LSTM) network is designed to handle sequential data such as lyrics text. The word embeddings are passed through an embedding layer and processed by an LSTM. Figure 2(a) illustrates a neural network model with two fully connected LSTM layers placed on top of each other, each containing 128 units. Stacking LSTM layers enabled the model to learn a hierarchical representation of sequential data, with the first layer capturing lower-level features and the second layer building on these to identify more complex patterns. This model uses BERT, FastText and Distil BERT embedding as feature representation.

Figure 2(b) illustrates a hybrid LSTM network. The architecture integrates multiple feature types by processing diverse inputs through separate LSTM layers for word embeddings of (GloVe, Word2Vec), numeric features (POS tags, length features), and text-based representations (TF-IDF, Bag of Words). An attention mechanism is also applied to weigh the most relevant features, which are then concatenated and passed through dense layers with softmax activation for emotion classification. This approach enables the simultaneous processing of semantic and syntactic information, allowing the hybrid model to capture complex linguistic nuances across different feature representations.

3.4.3. Fine-Tuning and Model Optimization

Hyperparameter tuning is applied to optimize the model's performance. Parameters like LSTM units, dropout rate, batch size, and epochs are fine-tuned to achieve an optimal balance between generalization and performance.

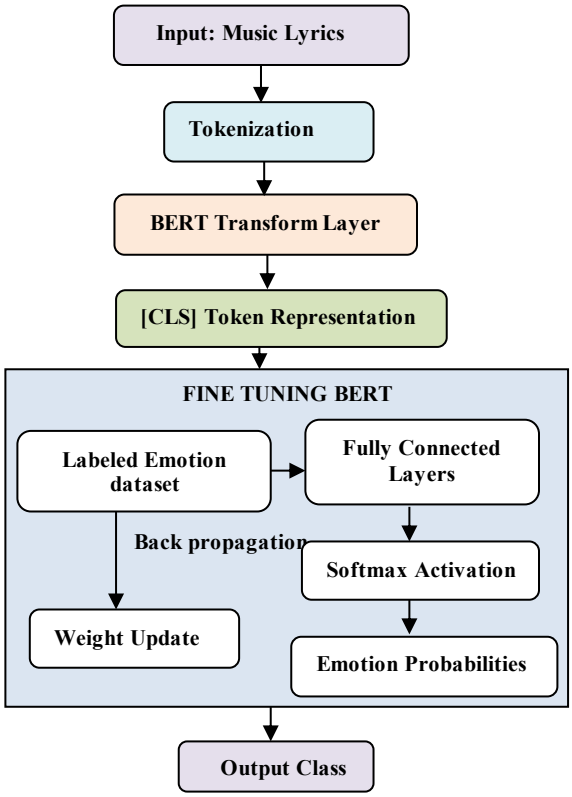

Fig. 3 Fine tuning BERT for emotion classification

Figure 3 shows an illustration for fine-tuning BERT for emotion classification. The proposed approach adapts the pretrained BERT model for emotion classification using a fine-tuning strategy. The input lyrics are tokenized, with BERT's transformer layers creating contextualized embeddings that capture linguistic nuances.

The [CLS] token's representation serves as an inclusive feature vector. Using softmax activation, a completely connected neural network is attached to the [CLS] token form to map high-dimensional embeddings to discrete emotion probabilities. The output layer gives the emotion class with the highest probability.

The emotion model is trained on labelled dataset (MER lyric, Mood and Combined Lyrics) through backpropagation, which adjusts pretrained BERT weights to study task-specific representation features. It also preserves the rich contextual knowledge from initial pretraining. This approach influences BERT's linguistic understanding to transform textual representations into emotion classifications. During training, regularization techniques like dropout are used to stop overfitting by arbitrarily turning off neurons. The classification layer applies the Binary Cross-Entropy with Logits (BCEWithLogits) loss function for BERT models and Categorical Cross Entropy for the Word2Vec/ Glove model. The Sigmoid activation function combines binary cross-

entropy, making it suitable for multi-label emotion classification. For models using GloVe or Word2Vec embeddings, categorical cross-entropy is the loss function for four class classification problems.

3.5. Accuracy Metric

Accuracy is the primary performance evaluation metric for emotion classification. Accuracy is computed as the proportion of correctly predicted emotion values to the total number of predictions:

Accuracy = (Number of Correct Predictions) / (Total Number of Predictions) × 100%

This metric provides a simple measure of the model's complete performance across different embedding techniques and datasets.

4. Result and Discussion

The study evaluated the performance of three datasets—Mood Lyrics [14], MER Lyrics [15], and Combined Lyrics—using different embeddings such as BERT, Distil BERT, FastText, Word2Vec and GloVe. The goal is to analyze how these embeddings perform across varying epochs and identify the best-performing configurations. It uses Random Search, Optuna, and manual tuning for hyperparameter optimization. Experiments are conducted using the stacked LSTM model for (Word2Vec and Glove) embedding, and for the Hybrid LSTM Model (BERT, Distillation BERT and FastText) embedding are used with the following key parameters tuned for optimization:

1. Learning Rate (LR): A fixed learning rate of 0.01 to 0.001 is used across all experiments to ensure consistent training dynamics.
2. Epochs: The number of training epochs varies between 10 and 40, with higher epochs generally improving accuracy for most embeddings.
3. Batch Size: Batch sizes 32, 64,100 and 300 are tested to balance training efficiency and model performance.
4. LSTM Layers: The model architecture includes two LSTM layers to determine sequential dependencies in the data.
5. Hidden Layer Size: Each LSTM layer has 128 hidden units, providing sufficient capacity for learning complex patterns.

4.1. Experimental Results: Embedding Performance on Three Lyrics Datasets: MER, Mood and Combined Lyrics

Table 1. MER lyrics performance is for different embeddings (BERT, GloVe, FastText, Distil BERT)

Embedding	LR	Epoch	Batch Size	Num Layers	Hidden Layers	Accuracy (%)
BERT	0.001	20	32	2	128	87.03
BERT	0.001	30	64	2	128	97.54
BERT	0.001	30	64	2	128	94.81
GLOVE	0.2	-	300	-	-	52.00
FASTTEXT	0.001	30	64	2	128	74.06
Distil BERT	0.001	20	64	2	128	87.03
Distil BERT	0.001	30	64	2	128	94.81

Table 2. Mood lyrics performance for different embeddings (BERT GloVe, FastText, Distil BERT)

Embedding	LR	Epoch	Batch Size	Num Layers	Hidden Layers	Accuracy (%)
BERT	0.001	20	64	2	128	89.83
BERT	0.001	30	64	2	128	93.86
BERT	0.001	30	150	2	128	96.83
BERT	0.001	10	64	2	128	67.00
BERT	0.001	20	100	2	128	97.63
FastText	0.001	20	32	2	128	87.61
FastText	0.001	30	64	2	128	97.11
GLOVE	0.2		300	-	-	60.00
Distil BERT	0.001	10	32	2	128	66.08
Distil BERT	0.001	10	100	2	128	66.75
Distil BERT	0.001	20	100	2	128	97.63

Table 3. Combined lyrics performance for different embedding (BERT, FastText, Distil BERT)

Embedding	LR	Epoch	Batch Size	Num Layers	Hidden Layers	Accuracy (%)
Distil BERT	0.001	20	64	2	128	91.56
Distil BERT	0.001	30	64	2	128	98.39
FastText	0.001	40	32	2	128	56.83
FastText	0.001	20	100	2	128	58.81
BERT	0.001	30	64	2	128	98.17

Table 1 shows MER Lyrics dataset performance using different embeddings. The BERT achieves high accuracy, with performance peaking at higher epochs. The best accuracy is 97.54% at epoch 30. Table 2 shows the performance of the Mood Lyrics dataset using different embeddings. The BERT embedding performs exceptionally well, with an accuracy of 97.63% at Epoch 20. BERT performs exceptionally well, with accuracy improving as epochs increase. Table 3 shows the Combined Lyrics Dataset performance using different embeddings. The Distil BERT gives the best accuracy of 98.39% at epoch 30. It outperforms other embeddings, with BERT achieving strong results (98.17% at Epoch 20). BERT consistently delivered high accuracy across all datasets, achieving 97.63% on Mood Lyrics and 97.54% on MER Lyrics.Distil BERT outperformed other embeddings on the Combined Lyrics dataset, achieving % highest accuracy of 98.39% at Epoch 30. FastText shows poor performance on the Combined Lyrics dataset, with an accuracy of only 56.83% at Epoch 40.

4.2. Embedding Performance Analysis and Training Dynamics

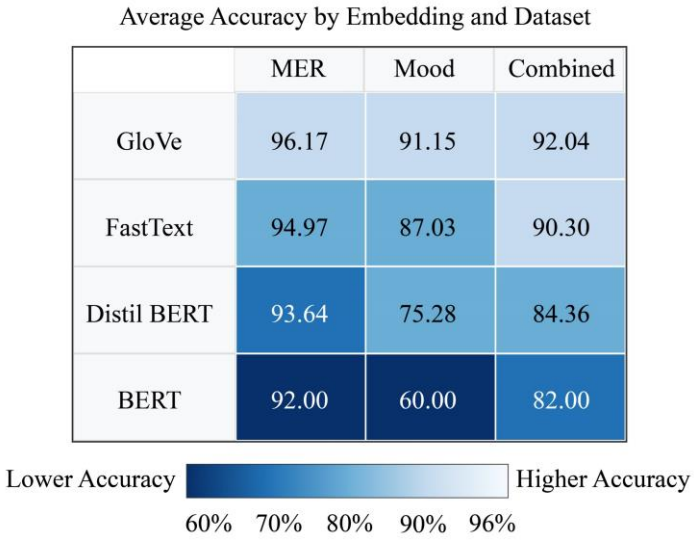

Fig. 4 Average classification accuracy for all datasets

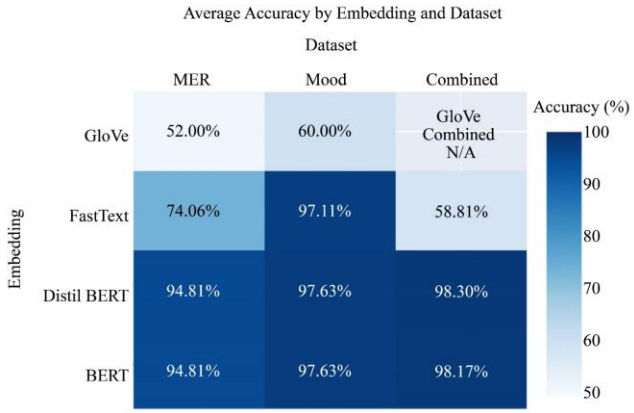

Fig. 5 Classification accuracy of all datasets and embeddings for different Epochs

4.2.1. Average Classification Accuracy for each Embedding across the Mood, MER, and Combined Lyrics Datasets

Figure 4 shows the average classification accuracy for each embedding (Glove, Fast Text, BERT and Distil BERT) across the Mood Lyrics, MER Lyrics and Combined Lyrics Datasets. This study calculates the average accuracy for each embedding across the Mood Lyrics, MER Lyrics, and Combined Lyrics datasets. The results provide a comprehensive overview of how each embedding performs on average, helping to identify the most reliable and effective embedding for lyrical analysis tasks.

- BERT achieves the highest average accuracy of 96.17%. It consistently performed well across all datasets, making it the most reliable embedding.
- Distil BERT achieves an average accuracy of 94.97%, slightly outperforming BERT. It delivers exceptional performance on the Combined Lyrics dataset, contributing to its high average.
- FastText achieves the lowest average accuracy of 56.83%. It performed poorly on the Combined Lyrics dataset, resulting in a low overall average.

Numerous experiments were conducted utilizing Glove and Word2Vec embeddings, yielding overall accuracy rates ranging from 53% to 60% across various feature combinations for the hybrid LSTM model. In contrast, Distil BERT and BERT emerged as the highest-performing embeddings, achieving average accuracies exceeding 94%...

4.2.2. Epoch-Wise Performance Dynamics

Figure 5 shows the effect of epochs on the performance of different datasets. BERT(E20) means BERT at 20 epochs; it has an accuracy of 89.39%. The analysis shows that increasing the number of epochs generally leads to improved accuracy for most models, particularly for BERT and Distil BERT. This trend indicates that these models benefit from additional training time, allowing them to learn more multifaceted patterns in the data. However, the performance gains are not uniform across all embeddings. For instance, FastText and GLOVE do not show significant improvements with increased epochs, suggesting that these models may have reached their performance ceiling early in the training process.

4.2.3. Performance Analysis of Different Embeddings

GLOVE / Word2Vec Performance

The performance of GLOVE and Word2Vec embeddings ranges between 54% and 60%. The result indicates that traditional word embeddings do not retain the contextual nuances of lyrics compared to advanced transformer models. For hybrid model configurations that include additional features, the best accuracy of Word2Vec Glove remains between 54% and 60%. The result is the same even when experimenting with a large batch size of 100 to 300 and using 30 to 50 epochs. Experiments were also conducted with GRU, Bi-LSTM, and RNN models using GloVe and Word2Vec embeddings, but the accuracy remained consistent across all architectures. Results show that GloVe provides some semantic understanding but lacks the depth of contextual representation that modern transformer-based models offer. The result highlights the limitations of static embeddings in tasks requiring a nuanced understanding of song lyrics.

BERT Performance

BERT embeddings show good results for accuracies with up to 98.39%. The results show that BERT is highly effective in understanding the context and semantics of lyrics, likely due to its transformer architecture and attention mechanisms. For hybrid model configurations with additional features, BERT's performance decreases to around 51%. Adding features to the BERT model may introduce noise or complexity that diminishes the model's ability to generalize effectively. The drop in performance could be due to overfitting.

Distil BERT Performance

Distil BERT also shows strong performance, with accuracies peaking at 97.63%. Distil BERT is a pre-compiled version of BERT; it retains much of the original's performance while being more efficient in speed and resource usage.

The results indicate that Distil BERT is a viable alternative to BERT, especially in scenarios where computational resources are restricted. The Distil BERT performs consistently for different datasets and epochs, suggesting robustness in various contexts.

Batch Size and Learning Rate

Batch size and learning rate help in improving the accuracy of classification. For example, the highest accuracy for BERT is achieved with a learning rate of 0.001 and a batch size of 64.

4.2.4. Inconsistency in Datasets

The results indicate Inconsistency in performance for different datasets. For instance, the Mood Lyrics dataset consistently yields higher accuracies than the MER Lyrics dataset. Higher accuracy could be due to differences in the complexity and structure of the lyrics in each dataset and the inherent characteristics of the English language.

4.2.5. t-SNE Visualization

Figure 6 visualizes the embedding spaces for different techniques using t-SNE dimensionality reduction. Transformer-based models (BERT and Distil BERT) show superior separation of emotion categories compared to traditional embedding techniques. This visualization corresponds with quantitative performance metrics, explaining why transformer models achieve higher classification accuracy. The figure indicates that:

- There is a clear separation of emotion categories in BERT and Distil BERT, which indicates that the four emotion categories form distinct clusters with minimal overlap.
- There is a partial overlap in FastText embeddings. A significant overlap in GloVe embeddings is present in visualization due to fewer clear boundaries between emotion categories.
- This indicates that emotion categories are somewhat separated but have more boundary cases.

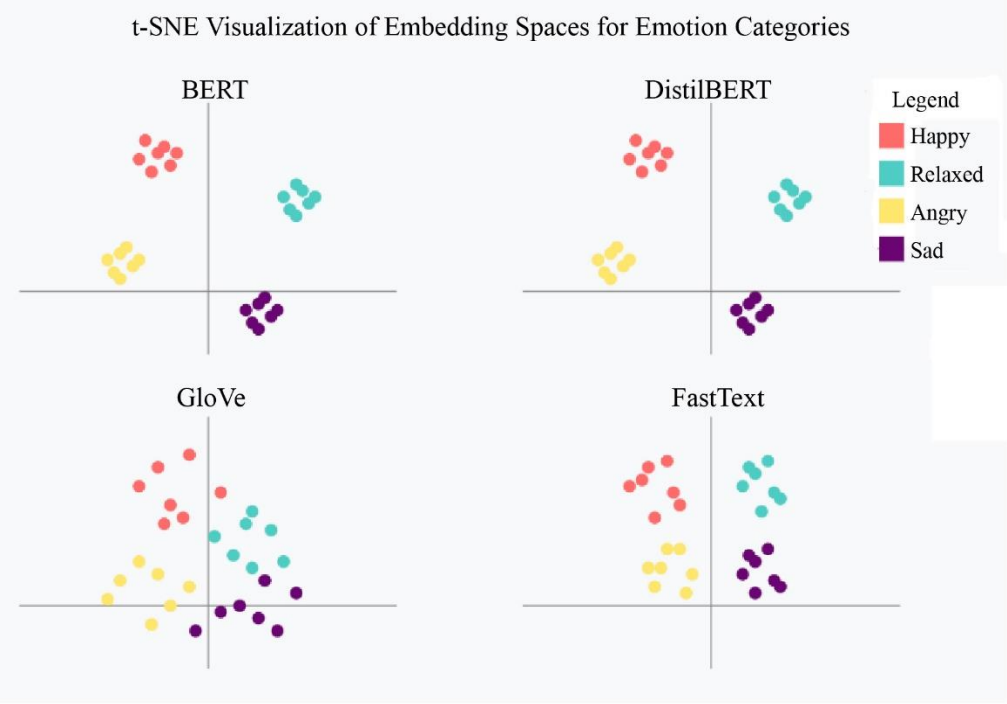

Fig. 6 t-SNE visualization of embedding space for all emotion categories/classes

(For the Emotion Categories, Happy is Red, Relaxed is Blue, Angry is Yellow, and Sad is Purple. Russell emotion quadrants are Q1 (Happy), Q2 (Angry), Q3 (Sad) and .Q4 (Calm /Relaxed))

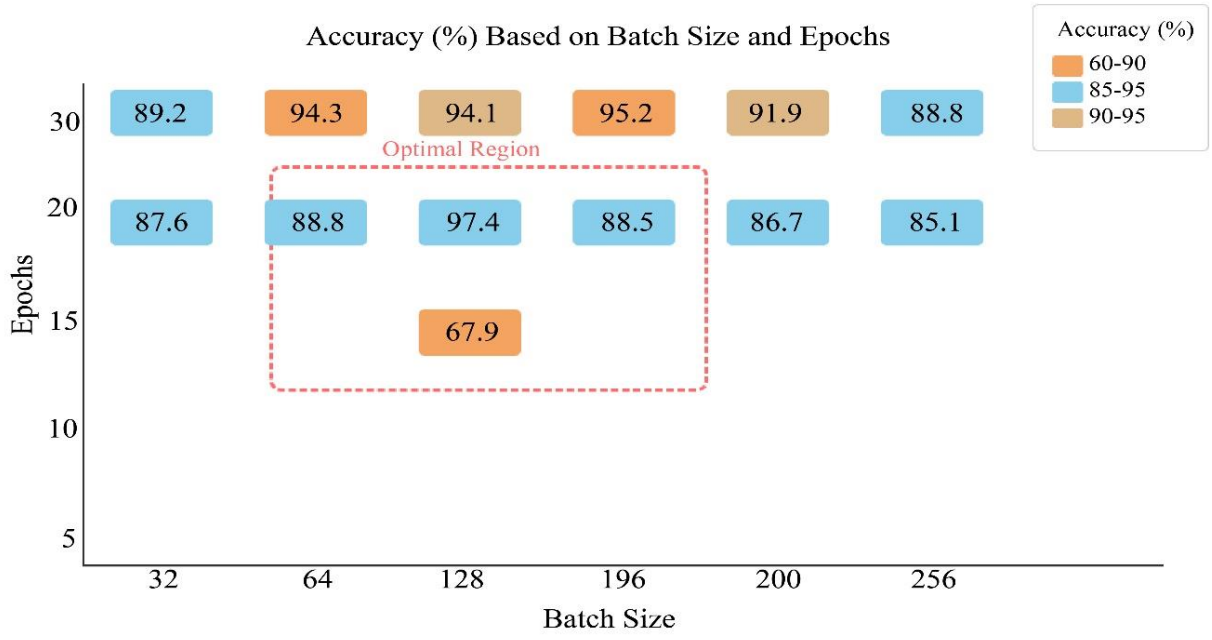

Fig. 7 Hyperparameter sensitivity analysis for BERT performance

4.2.6. Hyperparameter Sensitivity Analysis

The hyperparameter sensitivity visualization shows how batch size and epoch count affect BERT's performance. Figure 7 shows a Hyperparameter sensitivity analysis for BERT embeddings, revealing that optimal performance occurs with batch sizes between 100-150 and epochs from 20-30. The visualization demonstrates why our highest accuracy (97.63%) was achieved with a batch size of 100 and 20 epochs. Very small batch sizes (32) or large ones (256) show reduced performance, even with increased training epochs. The figure indicates the following:

- Optimal performance region: The region highlighted in the middle-top area (batch sizes 100–150, epochs 20–30) indicates the optimal performance.
- Performance patterns: Increasing epochs generally improves performance, but only up to a certain point.
- Batch size impact: Moderate batch sizes (64–150) tend to perform better than very small or large ones.

4.3. Comparison to Existing Work

This study uses a dataset of 771 songs from the MER and Mood Lyrics datasets to evaluate LSTM-based models with different embeddings. The BERT model achieves an accuracy of 97.54% on the MER dataset and 97.63% on the Mood Lyrics dataset. The Distil BERT model achieves 98.39% accuracy on the combined dataset. The results for the Mood Lyrics dataset surpass those of Shanker et al. (2023) [17], who achieved 93.8% accuracy with a transformer model. The improved accuracy is attributed to the use of transformer-based embeddings such as BERT and Distil BERT. The use of optimized hyperparameters: a batch size of 64, 20-30 raining epochs, two LSTM layers, and 128 hidden units contributed to improved accuracy.

The BERT implementation reaches 97.63%, reflecting a 3.83% improvement due to better training configurations, including larger batch sizes (64 to 100) and optimal epoch settings (20-30). The results show that Distil BERT outperforms regular BERT in music emotion recognition, achieving an accuracy of 98.39% on the combined dataset. This shows that the result significantly improved for the reported accuracy of 94.2% reported in Previously study. The combined datasets with Distil BERT show high accuracy. This suggests the transformative potential of multi-task learning and joint training in advancing emotion recognition capabilities.

The findings of Konstantinos et al. [9] show that transformer-based embeddings perform better than conventional methods, such as Word2Vec and Glove, in emotion classification. Their result indicates that BERT delivers 69.11% accuracy under the model setup. Conventional embeddings such as Word2Vec (41.66%) and GloVe (53.33%) performed much worse. Word2Vec and Glove also yielded less than 60% accuracy compared to their work. Nonetheless, BERT and DistilBERT obtained nearly 98% accuracy as a breakthrough compared to their results. The performance gap could be due to the fact that this work employed a better-optimized architecture. The use of diverse training data and hyperparameter optimization is an advantage in this research. Contrary to Konstantinos et al. [9], experimented with BERT within a different configuration, this research proves BERT's dominance over Glove and Word2Vec embeddings with better accuracy and heat map plots.

In contrast to Yinan Zhou's (2022) thesis [18], which reports an accuracy of 85.2% on the MER dataset using an RNN-based architecture-the findings reveal an impressive enhancement of 12.34%. The improvement in accuracy is due to transformer-based embeddings like BERT and Distil BERT, along with optimized hyperparameters: a batch size of 64, training for 20-30 epochs, two LSTM layers, and 128 hidden units. This study [18] also reports BERT achieving an F1-score between 47.48% and 53.31% (51.13% without text preprocessing); it outperforms with 97.54% accuracy on the MER dataset and 97.63% on the Mood Lyrics dataset using BERT alone. When integrating additional features like TF-IDF, BOW, length features such as (word count, sentence count, not count, BE verb count), and POS tags, the accuracy reduces to 30-50%. Combining extra features can interfere with BERT's ability to detect emotional nuances in lyrics, highlighting the difficulties of integrating features for emotion classification. Devlin et al. (2019) [19] highlight BERT's pretraining benefits for fine-tuning emotion-specific datasets. Previously Studies like further confirm BERT's strong performance in music emotion recognition, demonstrating its suitability for emotion-related tasks.

The study demonstrates that transformer-based models, particularly BERT and Distil BERT, are highly effective for emotion recognition. Hyperparameter tuning and model configuration play important roles in enhancing the performance of the model. Transformers without additional features performed better than transformers with additional features in emotion classification for BERT models, which identify difficulties in integrating features.

5. Conclusion

The performance of conventional and transformer-based embedding methods for emotion classification is assessed in this study using an LSTM model. The study compares how GloVe, Word2Vec, FastText, BERT, and Distil BERT perform on three music lyric datasets: Music Emotion Recognition (MER), Mood Lyrics, and a Combined Lyric dataset. The result shows a stark difference in performance between contextual and static embeddings. The static embeddings achieve the highest accuracy of 60%, and pre-trained BERT embeddings reach 98% accuracy. This highlights the effect of contextual embeddings in improving

LSTM performance, especially on emotion datasets. The combined dataset methodology is novel, as most current research examines the MER and Mood Lyrics datasets in isolation. The findings show that transformer-based models work best with combined datasets. This suggests they hold promise for enhancing music emotion classification. The findings also highlight that meticulous hyperparameter tuning and well-planned model design influence emotion model performance. The research also demonstrates that performance embedding is influenced by a number of factors, including model structure, corpus quality, and embedding selection. Future research will maximize the computational efficiency of BERT and other large models by investigating methods like model pruning, quantization, or knowledge distillation to minimize memory and processing time.

References

[1] Xin Rong, "Word2vec Parameter Learning Explained," *arXiv*, pp. 1-21, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[2] M. Alfa Riza, and Novrido Charibaldi, "Emotion Detection in Twitter Social Media Using Long Short-Term Memory (LSTM) and Fast Text," *International Journal of Artificial Intelligence and Robotics*, vol. 3, no. 1, pp. 15-26, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Zixiao Zhu, and Kezhi Mao, "Knowledge-Based BERT Word Embedding Fine-Tuning for Emotion Recognition," *Neurocomputing*, vol. 552, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jiaxin Ma et al., "Emotion Recognition Using Multimodal Residual LSTM Network," *Proceedings of the 27th ACM International Conference on Multimedia*, Nice France, pp. 176-183, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532-1543, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] Loreto Parisi et al., "Exploiting Synchronized Lyrics and Vocal Features for Music Emotion Detection," *arXiv*, pp. 1-8, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Cong Jin et al., "Attention-Based Bi-DLSTM for Sentiment Analysis of Beijing Opera Lyrics," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Xiaoguang Jia, "Music Emotion Classification Method Based on Deep Learning and Improved Attention Mechanism," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Konstantinos Pyrovolakis, Paraskevi Tzouveli, and Giorgos Stamou, "Multi-Modal Song Mood Detection with Deep Learning," *Sensors*, vol. 22, no. 3, pp. 1-23, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Rémi Delbouys et al., "Music Mood Detection Based on Audio and Lyrics with Deep Neural Net," *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Elangel Neilea Shaday, Ventje Jeremias Lewi Engel, and Hery Heryanto, "Application of the Bidirectional Long Short-Term Memory Method with Comparison of Word2Vec, GloVe, and FastText for Emotion Classification in Song Lyrics," *Procedia Computer Science*, vol. 254, pp. 137-146, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Anshul Wadhawan, and Akshita Aggarwal, "Towards Emotion Recognition in Hindi-English Code-Mixed Data: A Transformer Based Approach," *arXiv*, pp. 1-8, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri, "Transformer-Based Approach towards Music Emotion Recognition from Lyrics," *Proceedings, Part II 43rd European Conference on IR Research, Advances in Information Retrieval*, pp. 167-75, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Erion Çano, and Maurizio Morisio, "MoodyLyrics: A Sentiment Annotated Lyrics Dataset," *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, Hong Kong, pp. 118-124, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ricardo Malheiro et al., "Emotionally-Relevant Features for Classification and Regression of Music Lyrics," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 240-254, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Matthew V. Mahoney, "Fast Text Compression with Neural Networks," *Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2000)*, pp. 1-5, 2000. [Google Scholar] [Publisher Link]

[17] R. Guru Ravi Shanker, "*Emotion Unmasked: A Transformer-Based Analysis of Lyrics for Improved Emotion Recognitionm*," Thesis, International Institute of Information Technology, pp. 1-51, 2023. [Google Scholar] [Publisher Link]

[18] Yinan Zhou, *Music Emotion Recognition on Lyrics Using Natural Language Processing*, McGill University Libraries, 2022. [Google Scholar] [Publisher Link]

[19] Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, vol. 1, pp. 4171-4186, 2019. [CrossRef] [Google Scholar] [Publisher Link]