Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 2109-2122
Journal homepage: http://iieta.org/journals/isi

An Adaptive Cross-Modal Transfer Learning Framework with Metaheuristic Optimization for Multimodal Sentiment Analysis

Dasari Lakshminarayana Reddy[1*], Chigarapalle Shoba Bindu[2]

[1] Department of Computer Science and Engineering, Anantha Lakshmi Institute of Technology and Sciences, Anantapuramu 515721, India
[2] Department of Computer Science and Engineering, JNTUA CEA, Anantapuramu 515002, India

Corresponding Author Email: lakshmi1217phd@gmail.com

https://doi.org/10.18280/isi.310629

Received: 16 February 2026
Revised: 17 April 2026
Accepted: 30 April 2026
Available online: 30 June 2026

Keywords:
Multimodal Sentiment Analysis, Cross-Modal Transfer Learning, deep learning, BERT-based representation learning, Long Short-Term Memory networks, metaheuristic optimization, multimodal fusion

ABSTRACT

Multimodal Sentiment Analysis (MSA) aims to identify emotional states by jointly analyzing heterogeneous information sources, including textual, acoustic, and visual signals. Although recent deep learning approaches have improved sentiment recognition performance, many existing methods still suffer from limited cross-modal interaction, weak domain adaptability, and high computational requirements when dealing with heterogeneous data distributions. This study proposes an adaptive Cross-Modal Transfer Learning (CTL)framework with metaheuristic optimization for MSA. The proposed framework first learns modality-specific representations using Bidirectional Encoder Representations from Transformers (BERT) for textual features and Long Short-Term Memory (LSTM) networks for acoustic and visual features. A multimodal latent representation learning strategy is then developed to project different modalities into a shared feature space, where domain alignment is achieved through transfer learning mechanisms. To improve model efficiency and predictive performance, a Pied Kingfisher Optimizer Algorithm (PKOA) is introduced to optimize key hyperparameters of the BERT and LSTM models. The proposed framework was evaluated on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Experimental results demonstrate that the optimized framework achieves classification accuracies of 97.89%, 98.19%, and 97.03% for the business, family, and people topics, respectively. The results indicate that adaptive cross-modal transfer and optimization-based parameter tuning can improve multimodal sentiment classification under complex data conditions. Further validation on larger and more diverse datasets is required to evaluate its generalization capability.

1. INTRODUCTION

Topic modeling is an effective method of detecting and isolating latent thematic patterns of large datasets, mostly textual in nature [1]. It assists one in identifying trends, associations and major topics of large textual data, and is therefore useful in organizing, summarizing and accessing information efficiently [2]. But in the real world, the data tends to exist in various forms such as audio, video and text and more advanced approaches are needed where a single modal does not convey the richness of the data information giving weaker coherence [3]. Multimodal topic modeling can fulfill this requirement by modeling a variety of data modalities and enabling the extraction of more comprehensive topics that a single modality would otherwise miss [4]. This multimodal approach by utilizing several sources enhances better understanding of the context, better topic coherence and a more comprehensive representation of thematic structures [5].

On the same note, Sentiment Analysis (SA) is a popular technique of identifying emotions, opinions, and subjective declarations of information [6]. SA has traditionally been used on text, like online reviews or social media posts and news items to determine whether an emotion is positive, negative or neutral [7]. Nonetheless, text alone can be restrictive because emotional expressions are usually affected by the tone of voice, facial expression and other contextual factors, which are not written, resulting in the uncertainty of perceiving the sentiment [8]. Multimodal Sentiment Analysis (MSA) analysis overcomes these drawbacks by adding audio, video, and text, which allows interpreting the sentiment more precisely and holistically [9]. This multimodal technique combines the analysis of speech intonations with facial expressions and linguistic characteristics to gain a deeper insight on the emotions in a variety of contexts, which include interviews, social interactions, and multimedia content [10].

In general, the topic modeling and SA have a high level of interrelation because the two methods are complementary in deriving meaningful information about data [11]. Topics modeling gives us the background information, which will enable SA to understand emotions within particular topics [12]. An example is that a discussion on the environmental policies could have both positive and negative feelings depending on the angle one views it. Conversely, SA complements topic modeling by identifying sentimental

aspects of topics in a topic model, thus aiding in narrowing down topic classification as well as discerning the emotional coloring of various topics [13]. Such synergy can be of special importance to opinion mining, customer feedback analysis, and media content understanding applications [14].

Most of the research has concentrated on topic modelling approach derived from Linear Discriminant Analysis (LDA). A Joint Sentiment/topic Model (JST) model [15] is a probabilistic method designed to identify thematic structures in textual data while simultaneously determining document-level sentiment. Despite its capabilities, traditional LDA-based models depend heavily on rich word distributions to infer meaningful topics which limits their effectiveness for short texts and leads to reduced topic coherence and poor scalability which extended to multimodal data sources.

To address these limitations, Coherence Improved Partitioned LDA (CIPLDA) [16] with an Alternating Minimization Algorithm (AMA) was proposed for topic-aware SA. In this framework, CIPLDA was applied for multimodal topic modelling. Long Short-Term Memory (LSTM) were applied to process acoustic and visual features, while Bidirectional Encoder Representations from Transformers (BERT) was applied for handling the textual data. Features extracted from multiple modalities were fused using a multimodal embedding approach to generate a coherent representation. This combined representation maintains modality-specific contextual cues thereby substantially improving the sentiment classification accuracy.

The CIPLDA-MSA despite enhancing SA, has a number of limitations. The fusion approach mainly integrates features on a representation level that might not reflect richer semantic interactions between modalities, resulting in poorer performance in scenarios where one of the modalities is noisy, or incomplete. Moreover, the learning process in the various modalities is still in isolation whereby knowledge acquired in one modality will not dynamically affect or improve the others. The model is also shown to be reliant on particular datasets and is usually unable to stay accurate when used in other areas, meaning that it is not as adaptable in the real-world context and promotes less transparency in decision-making.

Therefore, Cross-Modal Transfer Learning (CTL) is developed to be effectively utilized in MSA in heterogeneous domains in this paper through the transfer of knowledge of labelled source domains to unlabelled target domains. Multimodal Latent Dirichlet Allocation (MLDA) first learns latent topic distributions on audio, text and video data in the target domain to produce semantically consistent multimodal representations. These encodings, together with labeled source-domain inputs, are then computed with modality-specific encoders, with LSTM networks used to extract audio and visual features, and BERT used to learn textual features. Multimodal fusion and domain adaptation are performed by projecting the extracted features to a common modality- and domain-invariant latent space. With the homo-instance samples, the hetero-instance samples will be segregated in this common space to maintain semantic consistency and discriminative properties. In addition, total variance loss has been added to ensure the consistency of features across the modalities whereas Kullback-Leibler (KL)-divergence-based domain alignment reduces the difference in the distribution between the source and target domains, which improves domain adaptation and generalization performance in MSA tasks.

Although the proposed CTL model demonstrates good results in cross-modal transfer with MSA, one of its weaknesses is poor tuning of BERT and LSTM parameters. Indeed, poorly parameterized BERT and LSTM segments might not capture subtle aspects of multimodal data, restricting their performance and efficiency. This non-optimality also leads to increased computational complexity and lower accuracy.

Optimized CTL (OCTL) is proposed to reduce the complexity of calculations and improve accuracy performance of MSA. The hyperparameter of LSTM and BERT in CTL-MSA is fine-tuned based on Pied Kingfisher Optimizer Algorithm (PKOA) in this model. Hyper-parameter consists of learning rate, optimizer, dropout rate, weight decay, epoch, batch size and momentum. The drive behind this PKOA is the intelligent behavior of wild pied kingfishers especially how they find their prey, hover, dive and catch fish, which involves the exploration, hovering, and exploitation phases. The pied kingfisher in this behavior-based policy systematically inspects the large water surfaces to see the possible prey, keeps a unitary hover to monitor motions, and then swiftly and precisely dives to capture the fish, which adjusts dynamically as wind and water alterations happen. Based on this PKOA mechanism, the optimal hyperparameters of the LSTM and BERT models are selected efficiently for model training for MSA. By applying PKOA, hyperparameters are fine-tuned effectively, resulting in enhanced prediction accuracy and reduced computational complexity. Based on the PKOA mechanism, optimal hyperparameters of LSTM and BERT are selected efficiently to model training of MSA that enhances the prediction accuracy and reduces the computational complexity in MSA.

The remaining article continues with the following format: Section II encloses previous studies. Section III describes the proposed OCTL-MSA and Section IV assesses its efficiency. Section V concludes the study.

2. LITERATURE SURVEY

Deng et al. [17] developed a MSA framework using a Cross-Modal Multihead Attention Mechanism (CMMAM). Initially, the raw data was preprocessed and converted into word vector representations with Positional Encoding (PE) to capture semantic and sequential information. The encoded inputs were then processed through a transformer model employing CMMAM to learn relationships among modalities. Finally, the extracted features were passed through a feedforward neural network and classification layer to generate the emotional prediction.

Huang et al. [18] presented a Transformer-based Multimodal Binding Learning (TMBL) model for MSA. In this model, binding learning model was applied to capture modality-specific and modality-invariant features while enabling cross-modality interactions. Fine-grained convolution modules were integrated into the feedforward and attention layers of the transformer architecture to enhance feature interaction. In addition, Classification Token (CLS) and PE feature vectors were employed to represent modality-invariant and specific features.

Cai et al. [19] developed a MSA that combines hierarchical feature integration with multi-task learning strategies. A Unimodal Feature Extraction Network (UFEN) was developed to determine more expressive representations from individual modalities. Subsequently, a multi-task fusion network

(MTFN) was employed to strengthen cross-modal integration and inter-model relationships. The potential dependencies among features were explored through attention-based mechanism and Transformer architectures. A MSA built on Unimodal Label generation and Modality Decomposition (ULMD) was introduced by Zhu et al. [20]. By using a multi-task learning, this approach breaks down the MSA problem into three unimodal tasks and one multimodal task. A modality representation separator was used to decompose each modality's features into shared (modality-invariant) components and unique (modality-specific) components.

Sun et al. [21] developed Mutual Information-based Disentangled Representation Learning (MIDRL) for MSA. The proportions of modality-invariant, -specific and -complementary information were quantitatively adjusted for feature extraction. In feature fusion, the amount of information retained by each modality in the fused representation are evaluated. The mutual information was used to estimate each type of information content. Wang et al. [22] introduced Non-Uniform Attention Module (NUAM) integrated with LSTM for MSA. In this method, NUAN module emphasizes textual representation while integrating acoustic and visual modalities to improve emotion recognition. A cluster-based loss function was incorporated to enhance feature representation and capture data structure. This model captures temporal dynamics by embedding NUAN within LSTM's time steps leads to improved SA by refining the model's understanding of multi-modal data.

He et al. [23] developed a Text-derived Multi-phase Representation with Multi-level Spatial-Memory data integration (T-MRMSM) model. Three modules like Two-Step Cross-attention Fusion (2Steps-AT), Enhanced Memory (EM) and Scale-Global Information Extraction (S-GIE) was integrated to capture multi-level and multi-scale features. This model captures long-range spatio-temporal features for intra and inter-modal fusion reducing reliance on text while enhancing the other two modalities. Wang et al. [24] introduced a Contextual Interaction-based Multi-modal Emotion (CIME) analysis with Enhanced Semantic Information. CIME was employed a text-centric cross-modal attention mechanism to refine semantic representations, while simultaneously leveraging a Graph Convolutional Network (GCN) to model contextual dialog information by capturing both intraspeaker and interspeaker relationships.

Yu et al. [25] presented a Multi-Level Perceptual Cross-modal Fusion (MPCF) framework for MSA. The framework employs a three-level perceptual architecture for pairwise modality fusion, collaborative multimodal perception, and deep affective interaction. It integrates specialized modules to extract latent emotional features while reducing redundant information within and across modalities. Additionally, the Affective Information Integration Module enhances cross-modal feature representation and improves fine-grained emotional perception.

Gao et al. [26] introduced a Triple-query Cross-modal Hierarchical Adaptive Fusion Network (TQ-CHAFN) for MSA. The model employs a Triple-Query Grouping strategy to preserve modality-centric interactions and separate semantic information from modality-specific noise.

A Hierarchical Adaptive Attention Fusion (HAAF) mechanism dynamically generates attention weights for accurate cross-modal alignment. The hierarchical fusion architecture balances local feature representation with global semantic consistency.

Table 1. Comparison of strength and limitations of proposed and existing models

References	Methods	Strengths	Limitations
[17]	CMMAM	Captures inter-modal dependencies using cross-modal attention	Difficulty in preserving fine-grained emotional dependencies across heterogeneous modalities.
[18]	TMBL	Learns modality-specific and invariant representations effectively	Complex feature binding may reduce interpretability of multimodal representations.
[19]	UFEN + MTFN	Enhances hierarchical feature integration and multimodal interaction	Multi-task learning may introduce feature redundancy between modalities.
[20]	ULMD-based MSA	Separates shared and unique modality representations	Negative transfer between shared and modality-specific tasks affects classification consistency
[21]	MIDRL	Improves disentangled multimodal feature learning	Limited scalability and reduced effectiveness on larger multimodal datasets.
[22]	NUAM-LSTM	Captures temporal multimodal dependencies effectively	Difficulty in adapting to unseen contextual and conversational variations
[23]	T-MRMSM	Extracts long-range spatio-temporal features	Dependence on dataset-specific spatio-temporal patterns limits generalization.
[24]	CIME	Enhances contextual semantic understanding	Performance degradation when textual information is noisy, incomplete, or ambiguous
[25]	MPCF	Improves affective interaction and redundant feature reduction	Redundant affective interactions may reduce fusion efficiency across modalities.
[26]	TQ-CHAFN	Achieves accurate cross-modal alignment and semantic consistency	Hierarchical fusion may introduce imbalance between local and global semantic representations.
Proposed model	OCTL-MSA	Effectively integrates multimodal semantic representations with adaptive hyperparameter optimization for improved sentiment classification accuracy and computational efficiency	Performance may be affected under noisy, incomplete, or cross-domain multimodal conditions requiring further multilingual and real-world validation.

Note: CMMAM = Cross-Modal Multihead Attention Mechanism; TMBL = Transformer-based Multimodal Binding Learning; UFEN = Unimodal Feature Extraction Network; MTFN = multi-task fusion network; ULMD = Unimodal Label generation and Modality Decomposition; MSA = Multimodal Sentiment Analysis; MIDRL = Mutual Information-based Disentangled Representation Learning; NUAM = Non-Uniform Attention Module; LSTM = Long Short-Term Memory; T-MRMSM = Text-derived Multi-phase Representation with Multi-level Spatial-Memory; CIME = Contextual Interaction-based Multi-modal Emotion; MPCF = Multi-Level Perceptual Cross-modal Fusion; TQ-CHAFN = Triple-query Cross-modal Hierarchical Adaptive Fusion Network; OCTL = Optimized Cross-Modal Transfer Learning.

Recent MSA research has focused on transformer-based architectures, cross-modal attention mechanisms, and multimodal fusion strategies for effective sentiment representation learning. Existing methods such as CMMAM [17], TMBL [18], CIME [24], MPCF [25] and TQ-CHAFN [26] improve inter-modal interaction and contextual feature fusion, while ULMD [20] and MIDRL [21] enhance modality-specific and invariant feature learning. However, these approaches often suffer from high computational complexity, limited generalization capability, and sensitivity to noisy multimodal information. Therefore, the proposed OCTL-MSA framework integrates MLDA, LSTM, BERT and PKOA-based optimization to improve sentiment classification accuracy and computational efficiency. Table 1 depicts the comparison of strength and limitations of proposed and existing models.

3. PROPOSED METHODOLOGY

The OCTL-MSA model is designed in this department. The overall workflow in the proposed approach is depicted in Figure 1.

3.1 Pre-processing

The preprocessing step is aimed at making the audio, text, and video modalities of the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset ready to allow the successful analysis of multimodal emotions. At audio mode, they apply noise removal as well as cepstral mean subtraction normalization to enhance the quality of speech and consistency in utterances. After pre-processing, 13-dimensional Mel Frequency Cepstral Coefficients (MFCC) features are extracted as they effectively represent vocal emotions such as anger, happiness, and sadness.

The audio data signal is divided into fixed-size segments of 200 ms and sequences less than 100 frames are zero-padded to ensure the same input dimensions when training the model. For the textual modality, tokenization and stop-word removal are performed on the transcript data to eliminate irrelevant words and divide the sentences into meaningful tokens. In addition, lowercasing is applied to convert all words into a uniform format, while lemmatization is utilized to reduce words to their root forms and improve semantic consistency. Subsequently, BERT-based word embedding representations are generated to capture semantic and contextual emotional information from the conversational text. In the video modality, frame extraction and face detection are utilized to identify expressive facial regions from conversational recordings. The extracted frames are resized to 224 × 224 pixels using spatial aspect normalization and processed to obtain facial expression features and motion-based emotional cues. The audio and text modality feature dimension is 768 where as video modality size is 2048. zero-padding strategy is used for making uniform dimensions for all samples. Furthermore, cross model fusion dropout is applied during training to reduce overfitting and improve multimodal generalization performance. These pre-processed multimodal features are fed into DL models for emotion and sentiment classification within the business, family and people thematic categories.

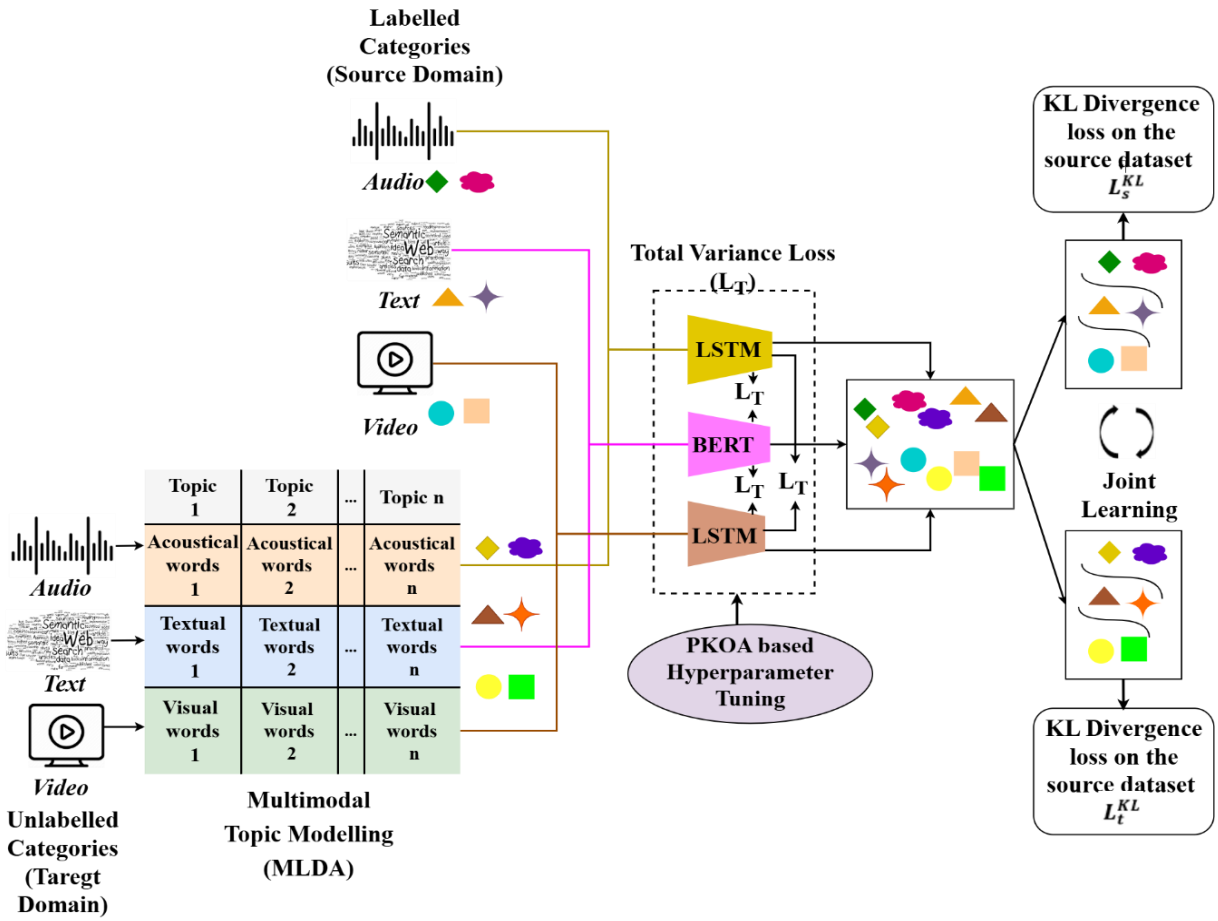

Figure 1. Overall flow of the OCTL-MSA approach
Note: MSA = Multimodal Sentiment Analysis; OCTL = Optimized Cross-Modal Transfer Learning.

3.2 Multimodal topic learning by Coherence Improved Partitioned Linear Discriminant Analysis

In this study, LDA is expanded to include three modalities by performing inference over n-tuples instead of pairs, ensuring each modality is conditionally independent based on the topic. For a tuple $(w_i, f_i, f_i', \dots)$ in document d, the conditional probability is computed across topics using modality-specific distributions. To enhance semantic representations, textual features are further enriched using BERT, while acoustic and visual features are temporally modelled through LSTM networks before entering the multimodal topic inference process.

CIPLDA [16] is based on the AMA which makes use of low-rank approximations of topic-document probability matrices to refine the topic-word and feature association. It is a reliable means to enhance topic recovery and document representation, which lead to an increased topic coherence and enhanced scalability of multimodal data. The representation-level fusion of CIPLDA-MSA, though, is better at SA, but it lacks deep cross-modal interactions and fails when there are noisy or missing modalities. It has isolated cross modal learning, is susceptible to high computational complexity, can poorly adapt to new domains and are less accurate.

3.3 Cross-Modal Transfer Learning

The proposed CTL framework facilitates effective cross-domain MSA based on the transfer of annotated multimodal knowledge on labelled source datasets into the unlabelled target datasets. In the first step, the unlabeled target-domain multimodal data, such as audio, text, and video streams are modeled by the MLDA module to extract latent topic distributions and produce semantically decodable multimodal data. In this topic modeling procedure, the underlying semantic relationships across various modalities are captured even with no sentiment labels. Simultaneously, labelled source-domain multimodal data are utilized to guide the transfer learning process.

Subsequently, the extracted modality-specific representations from both labelled source data and unlabelled target data are fed into dedicated feature encoders, where LSTM networks are employed for acoustic and visual feature learning, while BERT is utilized for contextual textual representation learning. The learned embeddings are then projected into a shared modality- and domain-invariant latent space for multimodal fusion and domain adaptation. Within this shared space, homo-instance samples are compactly clustered, whereas hetero-instance samples are separated to improve discriminative representation learning.

Furthermore, the total variance loss $\mathcal{L}_T$ is introduced to preserve feature consistency and compactness across modalities during representation learning. In addition, KL-divergence-based alignment minimizes the distribution discrepancy between source and target domains through source-domain loss $\mathcal{L}_s^{KL}$ and target-domain loss $\mathcal{L}_t^{KL}$.

Since the hyper-parameters of LSTM and BERT are not optimized properly, the model suffers from increased computational complexity, slower convergence, suboptimal feature learning, and reduced predictive accuracy across heterogeneous domains. To solve this, the proposed framework employs the PKOA-based hyperparameter optimization mechanism to optimally tune the parameters of LSTM and BERT. This reduces computational complexity while improving convergence speed and predictive accuracy.

3.4 Hyperparameter optimization of Long Short-Term Memory and Bidirectional Encoder Representations from Transformers using Pied Kingfisher Optimizer Algorithm

This section explains how to use the PKOA model to optimize the hyperparameters of LSTM and BERT that improves the model's accuracy and reduces the computational complexity. PKOA is most appropriate algorithm to fine tune parameters of DL models to MSA due to its ability to dynamically isolate the most informative modality and to optimally tune cross-modal fusion weights, avoiding redundant information. It is efficient to balance between complicated feature spaces of text, audio and visual representations. The overall example of PKOA based hyperparameter optimization of LSTM and BERT is as shown below.

The pied kingfisher's flight patterns and smart feeding habits served as design inspiration for the PKOA, a new optimization approach that is bio-inspired. An analysis of the pied kingfisher's hunting habits led to the development of this algorithm, which identifies three critical phases: exploration (hovering and perching), exploitation (diving), and commensalism (local escape). The PKOA works very fast in identifying the optimal strategy in complex optimization problems whilst maintaining diversity.

The proposed model applies PKOA to automatically find the most optimal hyperparameter combination of both LSTM and BERT models. First, a series of candidate hyperparameter sets are randomly chosen with each candidate having a learning rate, optimizers type, dropout rate, weight decay, epoch size, batch size and momentum. In the process of optimization, PKOA refines these candidate solutions with the help of exploration, exploitation, and commensalism steps to maximize classification accuracy and to minimize computation complexity. The optimal hyperparameter set obtained through PKOA is used for training the LSTM and BERT models in MSA.

3.4.1 Initialization

PKOA, similar to other population-based swarm intelligence algorithms, begins the optimization process by randomly generating initial candidate solutions within the predefined search space. In the proposed framework, each candidate solution represents a hyperparameter configuration for LSTM and BERT models, including learning rate, optimizer, dropout rate, weight decay, epoch size, batch size, and momentum. These candidate solutions serve as the starting points for searching optimal hyperparameter combinations. The process for generating the initial population is defined as:

$$X_{i,j} = LB + (UB - LB) \times R \tag{1}$$

In Eq. (1), $X_{i,j}$ is the i^{th} pied kingfisher individual in the j-th dimension, where $1 \leq i \leq N$ and $1 \leq i \leq D$. N is the total population (search space) size i.e., total number of candidates hyperparameter configuration for LSTM/BERT. D indicates dimensionality of the optimization problem (the considered hyperparameters D = 7). UB and LB is the upper and lower bounds of the search space, R represents the random number in the interval [0, 1], where i, j are positive integers of [1, N]. Practically, this initialization stage as shown in Pseudocode 1 generates diverse hyperparameter combinations for LSTM and

BERT models, enabling PKOA to efficiently explore different regions of the search space during optimization.

Pseudocode 1: Initialization stage

Input: Search ranges of LSTM and BERT hyperparameters (learning rate, optimizer, dropout rate, weight decay, epoch size, batch size, and momentum)
Output: Initial population of candidate solutions
Define population size *N* and dimensionality *D*
1. Specify UB and LB for each hyperparameter
2. for $i = 1$ to *N* do
3. for $j = 1$ to *D* do
4. Generate random value $R \in [0, 1]$
5. Initialize candidate solution $(X_{i,j})$ as in Eq. (1)
6. end for
7. end for
8. Return initialized hyperparameter population

3.4.2 Exploration stage

This stage is motivated by hovering and perch-based hunting strategies of the pied kingfisher. Filed studies show that these birds alternate between stationary hovering attacks and dives initiated from resting locations. Based on these natural behaviours, the PKO algorithm updates the positions of search agents by emulating the bird's adaptive foraging patterns. For hyperparameter optimization, this stage explores diverse regions of the search space by continuously modifying parameter values. This adjustment is defined in Eq. (2),

$$X_i(t+1) = X_i(t) + \alpha \times T \times \left(X_j(t) - X_i(t)\right) \quad i, j = 1,2,\dots,N \& j \neq i \tag{2}$$

In Eq. (2), $X_i(t+1)$ indicates the eventual position of the bird in the following iteration $1 \leq i \neq j \leq N$, and $X_i(t)$ resembles its current position. α represents the randomly sampled value drawn from a normal distribution while *T* parameter varies according to the selected strategy either by perching or by hovering. PKOA exploits the bird's natural perching and hovering behavior to efficiently explore different regions of the hyperparameter search space, varying parameter values to identify optimal configurations. *T* is tailored to every strategy to ensure best performance under different scenarios. The perching strategy parameter *T* is calculated using Eq. (3):

$$T = \left(Exp(1) - Exp\left(\frac{t-1}{M}\right)^{1/BF}\right) \times cos(2\pi R) \tag{3}$$

where, *BF* represents the constant value of 8. *M* represents the upper limit of iterations. The pied King fisher hovers by rapidly flapping its wings. In this hovering method, *T* is calculated by Eq. (4):

$$T = BR \times \left(\frac{t^{\frac{1}{BF}}}{M}\right) \tag{4}$$

$$BR = R \times \left(\frac{fit(j)}{fit(i)}\right) \tag{5}$$

In Eq. (5), *BR* represents the beating rate. *BR* represents the rate at which the pied kingfisher flaps its wings to maintain stability and control during the hovering phase, with the value being influenced by the fitness of the current and neighboring solutions. *fit* denotes the fitness function. It is the classification accuracy of LSTM and BERT models training. Perching strategy is usually applied in situations where the search space needs to be explored larger to prevent the premature convergence and finding a variety of candidate solutions. Conversely, the hovering method is applied when the searching algorithm nears in good areas of search, allowing narrow searches over smaller areas and consistent swarming towards good solutions.

In practice the exploration stage in Pseudocode 2, in PKOA sends the candidate solutions to random parts of the hyperparameter space, constantly updating them using new parameter values. In this procedure, learning rate, dropout rate, type of optimizers, batch size and momentum of LSTM and BERT models are tested in different combinations. This wide search methodology helps to avoid premature convergence and enhance the likelihood of finding the best hyperparameter settings globally.

Pseudocode 2: Exploration stage

Input: Current candidate solutions X_i
Output: Updated exploratory candidate solutions
1. for each candidate solution X_i do
2. Select neighboring solution X_j randomly
3. Generate random exploration factor α
4. If current fitness improvement is low or global exploration is required then
5. // Perching strategy for wider search
6. Compute *T* using Eq. (3)
7. else
8. // Hovering strategy for local refinement
9. Compute *T* using Eq. (4)
10. Compute beating rate (BR) using Eq. (5)
11. end if
12. Update candidate position using Eq. (2)
13. end for
14. Return updated candidate solutions

3.4.3 Exploitation stage

This stage is inspired by the perching and hovering hunting strategies of the PKO. Studies in natural habitats show that these birds alternate between launching attacks from stationary perches and engaging prey while hovering in midair. Accordingly, PKO algorithm updates the position of search agents by mimicking these adaptive foraging behaviours. For LSTM and BERT models, this stage corresponds to intensively fine-tuning hyperparameter configurations around promising search regions to obtain improved classification performance. This updated rule is delineated as in Eq. (6) to Eq. (9)

$$X_i(t+1) = X_i(t) + H_A \times \vartheta \times \alpha \times \left(b - X_{best}(t)\right) \tag{6}$$

$$H_A = R \times \left(\frac{fit(j)}{gf_b}\right) \tag{7}$$

$$\vartheta = Exp\left(-\frac{t}{M}\right)^2 \tag{8}$$

$$b = X_i(t) + \vartheta^2 \times R \times X_{best}(t) \tag{9}$$

In Eqs. (6)-(9), gf_b denotes the global optimal value, *HA*

denotes the hunting ability, ϑ represents the adaptive control factor that controls the exploitation intensity, α represents the random exploration coefficient that controls the stochastic movement of candidate solutions during optimization and b defines the uppermost fitness value attained through total iterations.

In practice, the exploitation stage in Pseudocode 3 focuses on refining the best-performing hyperparameter combinations identified during the exploration stage. Instead of conducting a global search across the entire solution space, PKOA performs local adjustments around the optimal candidate solutions to further improve convergence stability and classification accuracy. This process enables effective optimization of the LSTM and BERT hyperparameters while reducing unnecessary computational costs.

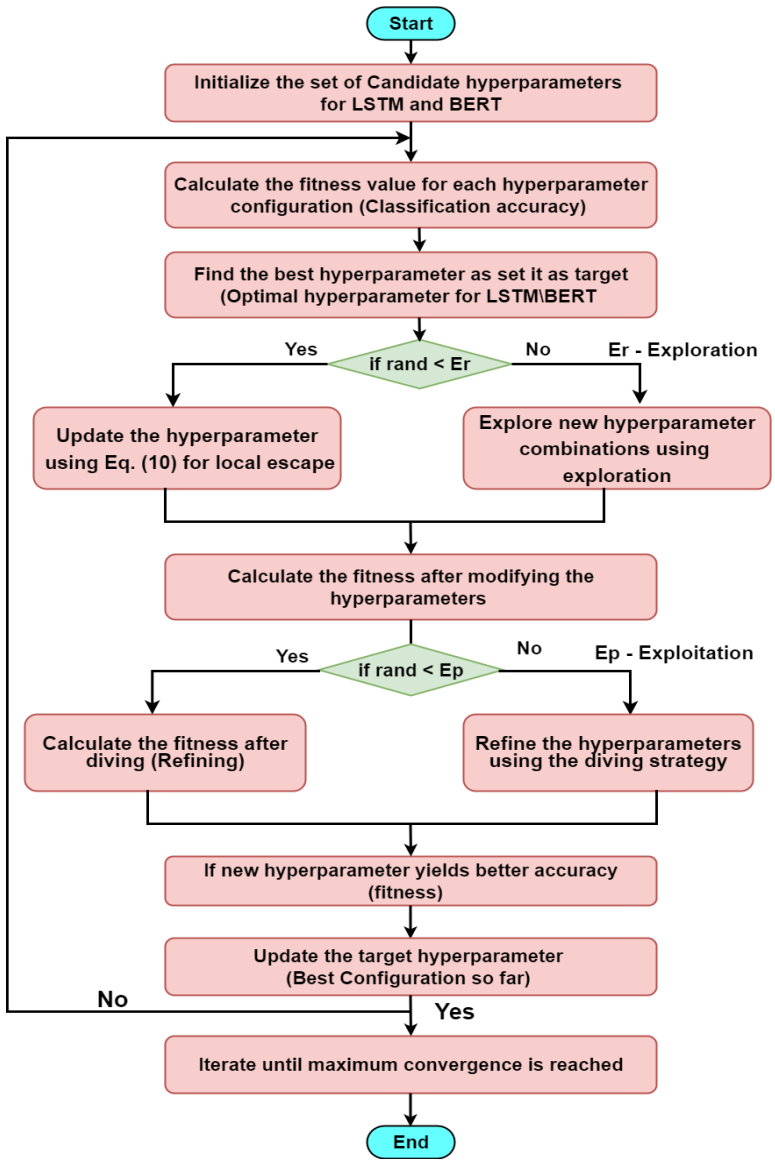

Figure 2. Flow diagram of Pied Kingfisher Optimizer Algorithm (PKOA) based hyperparameter optimization

Pseudocode 3: Exploitation stage

Input: Exploratory candidate solutions X_i
Output: Refined candidate solutions
1. for each candidate solution X_i do
2. Compute hunting ability (HA) using Eq. (7)
3. Compute decay factor ϑ using Eq. (8)
4. Determine best candidate position b using Eq. (9)
5. Update candidate position using Eq. (6)
6. Evaluate fitness using classification accuracy
7. end for
8. Return refined candidate solutions

3.4.4 Commensalism stage

Several species of otter share symbiotic relationships with pied kingfishers. Both species gain from this interaction without harming each other. This mutually beneficial interaction is mathematically represented as follows.

$$X_i(t+1) = \begin{cases} X_m(t) + \alpha|X_i(t) - X_n(t)|, if\ r > (1-PE)(a) \\ X_i(t), otherwise(b) \end{cases} \quad (10)$$

$$PE = PE_{max} - (PE_{max} - PE_{min}) \times \left(\frac{t}{M}\right) \quad (11)$$

where, X_m and X_n are the locations of two randomly selected individuals from the candidates. The pied kingfisher's hunting proficiency, represented by PE and determined by the variables PE_{max} and PE_{min}, is always assigned 0.5 and 0. Practically, the commensalism stage in Pseudocode 4 enables

candidate solutions to interact with neighboring solutions to preserve population diversity and dynamically control local search around the best LSTM and BERT hyperparameter configurations, thereby preventing premature convergence and improving optimization stability.

Pseudocode 4: Commensalism stage

Input: Refined candidate solutions X_i
Output: Diversity-preserved candidate solutions
1. for each candidate solution X_i do
2. Randomly select X_m and X_n
3. Compute hunting proficiency PE using Eq. (11)
4. Generate random value r
5. Update candidate position using Eq. (10)
6. Evaluate fitness using classification accuracy
7. end for

Return updated candidate solutions

Algorithm 1. PKOA-based hyperparameter optimization for MSA

Input: Set of hyperparameters for LSTM and BERT, maximum iterations *M*, population size *N*, dimensionality *D*
Output: Optimal hyperparameters of LSTM and BERT in CTL

1. Begin
2. Initialize the population of candidate hyperparameters by Eq. (1);
3. Compute the fitness function (classification accuracy) for each candidate by Eq. (5);
4. Discover the best solution and set it as the target;
5. while ($t \leq M$)
6. Apply exploration stage (perching and hovering) by Eq. (2)–Eq. (4);
7. Evaluate fitness of updated solutions by Eq. (5);
8. if (exploration successful)
9. Update positions using exploitation stage (diving) by Eq. (6)–(9);
10. Else Apply commensalism stage (local escape) by Eq. (10)–(11);
11. end if
12. Compute the new fitness values;
13. Update best solution (optimal hyperparameters) if improved;
14. $t = t + 1$;
15. end while
16. Acquire the final ideal hyperparameter set for LSTM and BERT
17. End

The hyperparameter optimization of LSTM and BERT is especially the focus of PKOA since both have a large and non-linear search space, whose parameters exhibit a complex interdependence. Existing standard optimization algorithms might have an early convergence point and failure to search around local optima in a problem of multimodal feature learning. Conversely, PKOA offers an efficient tradeoff between global exploration and local exploitations, allowing the results to be easily obtained in the form of optimal hyperparameter settings. Exploration phase enhances diversity in a search, exploitation phase narrows down the promising candidate solutions, and commensalism phase averts stagnation on local optima. The aforementioned properties attract PKOA to be very useful in enhancing convergence consistency, minimizing the complexity in calculations and boosting the performance of multimodal sentiment recognition. Figure 2 presents the flow work of PKOA based hyperparameter optimization in Algorithm 1.

Therefore, the suggested model of the OCTL-MSA enables the combination of the CTL and PKOA-based hyperparameter optimization to enhance the multimodal sentiment classification with less computational complexity and resource consumption.

4. RESULT AND DISCUSSION

4.1 Dataset details (multimodal corpus)

The experimental uses an IEMOCAP data [27]. It is a properly established dataset developed by the University of California (USC) to aid research in affective computing and interactive systems. It includes approximately 12 hours of audio and video (synchronized) content such as video recordings, verbal dialogue, facial expression, body movement data and transcripts of speech obtained during rehearsed and spontaneous exchanges. The series will be divided into five recording sessions with 10 trained performers with an equal number of male and female participants. All verbal blocks are coded with discrete emotion values such as happiness, sadness, anger, surprise, excitement, frustration, fear, disgust, neutrality and other, as well as continuous emotion scales: dominance, arousal and valence.

The dataset of the IEMOCAP is especially applicable to MSA since it features a rich expression of emotions in the context of natural conversations. In this study, the sample distribution across categories is divided into 70% for training and 30% for testing. The statistical distribution of sentiment categories is presented in Table 2 to ensure balanced experimental evaluation across all multimodal sentiment classes. The data instances are additionally organized into three thematic groups, namely business, family, and people.

Table 2. Statistical summary of the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus

Emotions	Number of Samples	Training (70%)	Testing (30%)
Happy	595	408	187
Excited	1041	739	302
Sad	1084	750	334
Anger	1103	764	339
Frustration	1849	1298	551
Surprised	107	74	33
Neutral	1753	1260	493
Others	2507	1734	773
Total	10039	7027	3012

4.2 Experimental setup and performances metrics

In this section, the proposed OCTL-MSA model is compared with existing approaches using the gathered dataset. The experiments were conducted on a system equipped with an Intel® Core™ i7 processor, 16 GB RAM, and an NVIDIA GTX-1060 GPU with 6 GB dedicated memory, running on the Windows 10 operating system. The implementation was carried out using Python 3.10 and TensorFlow 2.13.0 with supporting libraries such as NumPy, SciPy, and scikit-learn. GPU-accelerated batch processing was utilized during training with a batch size of 32 to improve computational efficiency.

The performance metrics like accuracy, precision, recall, and F1-score are employed to evaluate the classification performance of multimodal sentiment classes. The dataset contains eight classes which are reformulated into a binary classification task by designating one class as positive and remaining seven as negative for metric evaluation.

Accuracy: It shows the classification ability of the model tested based on the percentage of sentiment classes correctly detected. Eq. (12) is computed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where, True Positive (TP) is a sample labeled as positive that is correctly labeled as positive by the model. False Positive (FP) is when the model classifies a negative as positive. False Negative (FN) is when a positive sample is classified as negative. True Negative (TN): When the model predicts a negative class and it is indeed negative.

Precision: It indicates the correctness of a positive prediction that may have been made by the model by calculating the percentage of all positive samples predicted correctly by the model. The calculation is given in Eq. (13).

$$Pre = \frac{TP}{TP + FP} \tag{13}$$

Recall\Sensitivity: It denotes the function of the model in detecting the TP instances in the dataset. It is calculated as per Eq. (14).

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

F1-score: It represents the harmonic mean of Precision and Recall, providing a balanced evaluation of classification performance. It is defined as in Eq. (15).

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{15}$$

4.3 Performance evaluation of Multimodal Sentiment Analysis

From the compiled dataset, a test subset containing 3012 samples was used to evaluate the model's performance. Topic modelling was performed on the test data, which identified three dominant topics: business, family, and people. For the business topic, 1783 samples were selected from the test set. Among these, the proposed OCTL-MSA model correctly classified 1,676 samples, while 107 samples were misclassified. The confusion matrix for CTL-MSA model on business topic are given in Figure 3. For the family topic, 581 samples were considered, of which 550 were correctly classified and 31 were incorrectly classified.

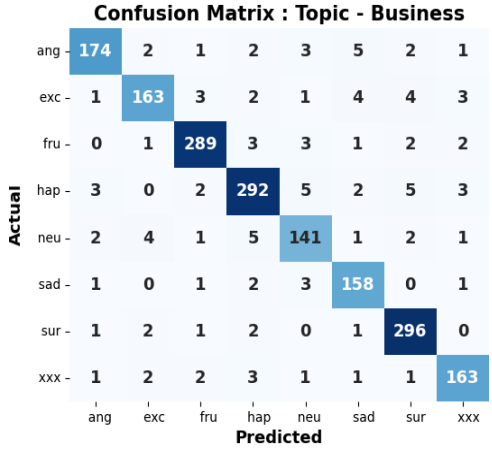

Figure 3. Confusion matrix of OCTL-MSA model: business topic (test dataset)
Note: MSA = Multimodal Sentiment Analysis; OCTL = Optimized Cross-Modal Transfer Learning.

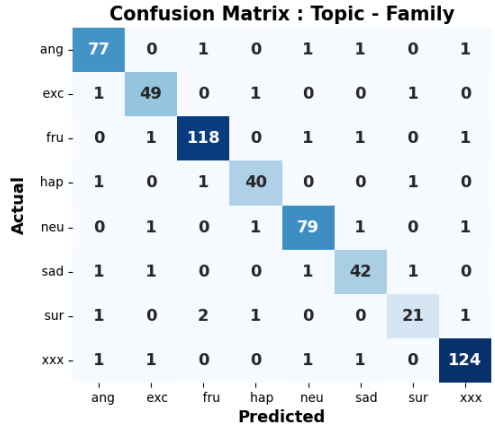

Figure 4. Confusion matrix of OCTL-MSA model for topic family (test dataset)
Note: MSA = Multimodal Sentiment Analysis; OCTL = Optimized Cross-Modal Transfer Learning.

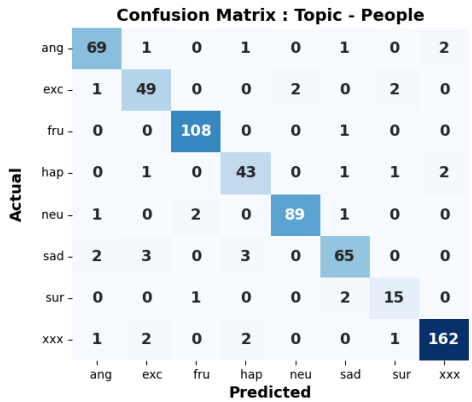

Figure 5. Confusion matrix of OCTL-MSA model: People topic (test dataset)
Note: MSA = Multimodal Sentiment Analysis; OCTL = Optimized Cross-Modal Transfer Learning.

The confusion matrix for the family topic is presented in Figure 4. Regarding the people topic, 637 samples were included in the test set. Out of these, 600 samples were accurately classified, while 37 samples were assigned to incorrect categories. Figure 5 presents the confusion matrix for people topic for proposed OCTL-MSA framework. Furthermore, the newly introduced OCTL-MSA framework is evaluated against multiple previously published techniques, namely CMMAM-MSA [17], MIDRL-MSA [21], MPCF-MSA [25], TQ-CHAFN-MSA [26], CIPLDA-MSA [16] and CTL-MSA.

All models are evaluated according to standard performance metrics, as detailed in Section 4.2. The parameter setting for PKOA is provided in Table 3. Table 4 lists the identical hyperparameter settings for proposed model and existing models when applying on the datasets given in section 4.1. In this work, seven common hyperparameters, namely learning rate, optimizer, dropout rate, weight decay, epochs, batch size, and momentum, are jointly tuned for both LSTM and BERT within a unified optimization framework. The optimized values, including a learning rate of 0.001, dropout rate of 0.4, Adam optimizer, batch size of 32, and 120 epochs, provide stable convergence, reduced overfitting, and improved generalization, thereby enhancing the overall accuracy of the OCTL model.

Table 3. Parameter settings for Pied Kingfisher Optimizer Algorithm (PKOA)

Parameters	Range
Population size (N)	30
Dimensionality (D)	7
Number of iterations	500
PE_{max}; PE_{min}	0.5; 0
Random Number	[0, 1]

Table 4. List of optimal hyperparameters for existing and proposed Optimized Cross-Modal Transfer Learning (OCTL) model

Parameters	Search Range	Optimal Range
Learning rate	0.00001–0.1	0.0001
Optimizer	[SGD, Adam, RMSprop, Adagrad]	Adam
Dropout rate	0.3-0.7	0.4
Weight decay	0-0.01	0.0005
Epoch	100-150	120
Batch size	8-64	32
Momentum	0.5-0.99	0.9

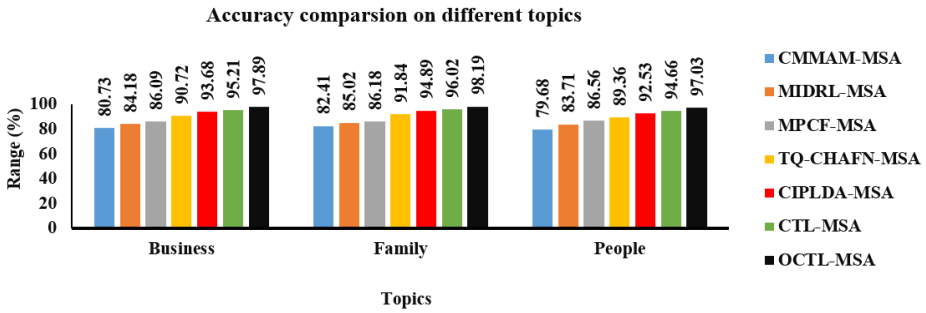

Figure 6. Accuracy evaluation on three topics for Multimodal Sentiment Analysis (MSA)

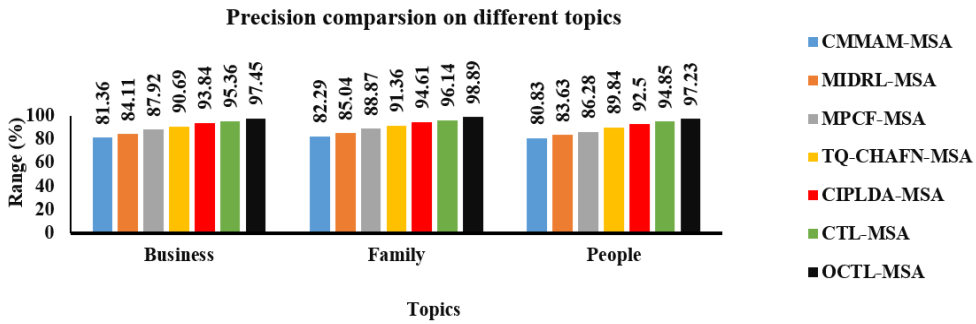

Figure 7. Precision evaluation on three topics for Multimodal Sentiment Analysis (MSA)

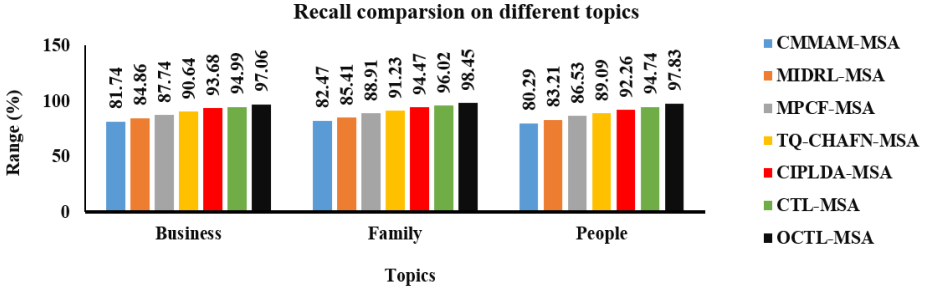

Figure 8. Recall evaluation on three topics for Multimodal Sentiment Analysis (MSA)

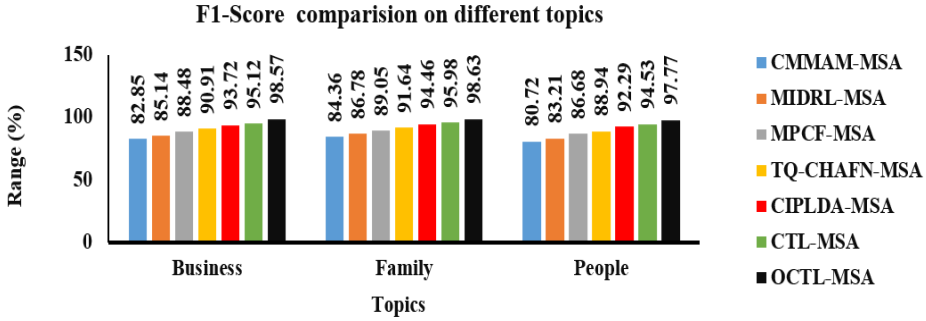

Figure 9. F1-score evaluation on three topics for Multimodal Sentiment Analysis (MSA)

Table 5. Performance comparison of with, without preprocessing and hyperparameter optimization

Metrics (%)	CIPLDA-MSA			CTL-MSA			OCTL-MSA		
	Business	Family	People	Business	Family	People	Business	Family	People
Without Pre-processing									
Accuracy (%)	88.45	89.32	87.21	90.12	91.87	89.54	89.76	90.45	88.92
Precision (%)	88.12	88.95	86.88	89.85	91.52	89.21	89.45	90.12	88.70
Recall (%)	88.55	89.10	87.34	90.05	91.73	89.47	89.80	90.50	88.95
F1-score (%)	88.33	89.02	87.11	89.95	91.62	89.34	89.62	90.31	88.82
With Pre-processing									
Accuracy (%)	93.68	94.89	92.53	95.21	96.02	94.66	97.89	98.19	97.03
Precision (%)	93.84	94.61	92.5	95.36	96.14	94.85	97.45	98.89	97.23
Recall (%)	93.68	94.47	92.26	94.99	96.02	94.74	97.06	98.45	97.83
F1-score (%)	93.72	94.46	92.29	95.12	95.98	94.53	98.57	98.63	97.77
With Hyperparameter Optimization (Random Search)									
Accuracy (%)	92.10	92.45	92.30	92.55	92.20	92.40	92.25	92.50	92.35
Precision (%)	92.00	92.30	92.15	92.50	92.10	92.35	92.20	92.45	92.25
Recall (%)	92.05	92.40	92.25	92.55	92.15	92.38	92.25	92.48	92.30
F1-score (%)	92.03	92.35	92.20	92.53	92.13	92.36	92.23	92.47	92.28
With Hyperparameter Optimization (PKOA)									
Accuracy (%)	93.68	94.89	92.53	95.21	96.02	94.66	97.89	98.19	97.03
Precision (%)	93.84	94.61	92.5	95.36	96.14	94.85	97.45	98.89	97.23
Recall (%)	93.68	94.47	92.26	94.99	96.02	94.74	97.06	98.45	97.83
F1-score (%)	93.72	94.46	92.29	95.12	95.98	94.53	98.57	98.63	97.77

Note: MSA = Multimodal Sentiment Analysis; CIPLDA = Coherence Improved Partitioned Linear Discriminant Analysis; CTL = Cross-Modal Transfer Learning; OCTL = Optimized Cross-Modal Transfer Learning; PKOA = Pied Kingfisher Optimizer Algorithm.

Figure 6 presents an accuracy comparison for MSA on different topics. The results indicate that the proposed OCTL-MSA model outperforms CMMAM-MSA [17], MIDRL-MSA, MPCF-MSA, TQ-CHAFN-MSA, CIPLDA-MSA and CTL-MSA by 21.26%, 16.29%, 13.69%, 7.91%, 4.49%, and 2.81%, respectively, on the business topic. Similarly, on the family topic, OCTL-MSA achieves accuracy improvements of 19.15%, 15.47%, 13.91%, 6.91%, 3.48%, and 2.26% over the same baseline models. For the people topic, the accuracy gains are 21.78%, 15.91%, 12.09%, 8.59%, 4.86%, and 2.50%, respectively.

Figure 7 illustrates the precision comparison for MSA on three topics. The precision of the OCTL-MSA model exceeds that of CMMAM-MSA [17], MIDRL-MSA, MPCF-MSA, TQ-CHAFN-MSA, CIPLDA-MSA and CTL-MSA by 19.77%, 15.86%, 10.85%, 7.46%, 3.85%, and 2.19%, respectively, on the business topic. On the family topic, OCTL-MSA shows precision improvements of 20.17%, 16.29%, 11.26%, 8.24%, 4.52%, and 2.86% compared to the same methods. Likewise, for the people topic, the proposed model achieves precision gains of 20.28%, 16.26%, 12.67%, 8.23%, 5.12%, and 2.51% over the corresponding baseline approaches.

Figure 8 presents the recall comparison for MSA on three topics. The proposed OCTL-MSA model achieves higher recall values than CMMAM-MSA [17], MIDRL-MSA, MPCF-MSA, TQ-CHAFN-MSA, CIPLDA-MSA and CTL-MSA by 18.74%, 14.38%, 10.62%, 7.08%, 3.61%, and 2.18%, respectively, on the business topic. For the family topic, the recall improvements are 16.61%, 12.81%, 8.20%, 5.31%, and 1.64% over the corresponding baseline models, while on the people topic, the proposed model shows recall gains of 17.37%, 13.41%, 10.07%, 5.64%, and 2.69%.

Figure 9 illustrates the F1-score comparison for MSA on three topic categories. On the business topic, OCTL-MSA outperforms CMMAM-MSA [17], MIDRL-MSA, MPCF-MSA, TQ-CHAFN-MSA, CIPLDA-MSA and CTL-MSA by 18.97%, 15.77%, 11.41%, 8.43%, 5.17%, and 3.63%, respectively. Similarly, for the family topic, the F1-score improvements are 16.91%, 13.65%, 10.75%, 7.63%, 4.41%, and 2.76%. On the people topic, the proposed model achieves even larger gains of 21.14%, 17.50%, 12.79%, 9.94%, 5.94%, and 3.43% over the same baseline methods. Overall, these results demonstrate that the OCTL-MSA model significantly enhances multimodal SA performance compared with existing standard approaches.

Table 5 provides a comparative analysis of the CIPLDA-MSA, CTL-MSA and OCTL-MSA models under without pre-processing, with pre-processing, with hyperparameter optimization of random search and PKOA across the business, family and people topics. In case of without preprocessing, CTL-MSA slightly outperformed CIPLDA-MSA and OCTL-

MSA across all metrics and topics. After applying preprocessing, the performance of all models improved significantly, with OCTL-MSA achieving the best results, followed by CTL-MSA. These results illustrate the effectiveness of pre-processing on improving the classification and the effectiveness of OCTL-MSA as a more robust solution for dealing with noisy textual data.

In a similar manner, OCTL-MSA model showed significant performance improvement all three topics when tuning hyper parameter. While random search boosted all models, OCTL-MSA optimized with PKOA achieved the best results among all the models, i.e., greater than 97% in most of the assessment metrics. The results, overall, validate that the OCTL-MSA combined with PKOA yields the lowest and most stable error rate for sentiment classification under different types of contextual conditions.

It is presented in Table 6 the computational efficiency of the models all three models for the three topics. This quantifies how long the training can take, how long it takes to run the trained model (inference), and how much memory the model uses when trained on a CPU or GPU. The conventional multimodal learning framework of CIPLDA-MSA causes it to have the highest computational cost compared to the three models, in terms of training time, inference latency and memory consumption. CTL was done by combining CTL-MSA to achieve a moderate training time, a low memory consumption, and a faster inference time than CIPLDA-MSA.

Table 6. Computational efficiency performance for various topics

Models	Topics	Training Time (sec)		Inference Time (sec)		Memory Usage (MB)	
		CPU	GPU	CPU	GPU	CPU	GPU
CIPLDA-MSA	Business	6025	1510	3.20	0.96	12480	8765
	Family	6112	1536	3.24	0.99	12563	8827
	People	6198	1568	3.29	1.02	12645	8891
CTL-MSA	Business	4128	1042	2.48	0.71	9324	6487
	Family	4196	1061	2.52	0.74	9392	6541
	People	4278	1084	2.56	0.77	9463	6598
OCTL-MSA	Business	2386	602	1.73	0.46	6945	4826
	Family	2432	614	1.76	0.48	6996	4869
	People	2489	629	1.81	0.51	7058	4918

Note: MSA = Multimodal Sentiment Analysis; CIPLDA = Coherence Improved Partitioned Linear Discriminant Analysis; CTL = Cross-Modal Transfer Learning; OCTL = Optimized Cross-Modal Transfer Learning.

Table 7. Paired *t*-test analysis of four-fold cross validation accuracy of proposed OCTL-MSA with existing models

Model	*t*-Value	df	Sig. (2-Tailed)	Mean Difference (%)	Standard Deviation	95% Confidence Interval	
						Lower	Upper
			Business				
CIPLDA-MSA	6.650	3	0.0069	4.21	1.26	2.195	6.224
CTL-MSA	9.782	3	0.0022	2.68	0.54	1.808	3.551
			Family				
CIPLDA-MSA	6.264	3	0.0082	3.30	1.05	1.624	4.980
CTL-MSA	11.163	3	0.0015	2.17	0.38	1.551	2.788
			People				
CIPLDA-MSA	14.971	3	0.0006	4.5	0.60	3.543	5.456
CTL-MSA	7.788	3	0.0044	2.37	0.60	1.401	3.338

Note: MSA = Multimodal Sentiment Analysis; CIPLDA = Coherence Improved Partitioned Linear Discriminant Analysis; CTL = Cross-Modal Transfer Learning; OCTL = Optimized Cross-Modal Transfer Learning.

CTL-MSA has the advantages of CTL, which significantly reduces the computational overhead, and much faster inference speed and is more time-efficient than CIPLDA-MSA when training. The proposed OCTL-MSA model can be optimized in hyper-parameters using the PKOA method which results in the best computational performance with significant reduction in the training time, faster inference performance on CPU and GPU platforms as well as reduction in memory usage, and with high classification accuracy. Generally, the results show that good balance between classification performance and computational efficiency is achieved by using OCTL-MSA, which are suitable for MSA.

The results presented in Table 7 show the paired t-test analysis for four-fold cross validation accuracy of the proposed model OCTL-MSA to existing models on various categories. According to the statistical results, there are significant differences in the performance of the proposed model compared to the comparative models. The greater the difference between the mean, and the narrower the confidence interval, the more robust and reliable it is. The result of the analysis in overall seems to assert the efficiency of adaptation of OCTL-MSA in multimodal sentiment classification.

Table 8. Ablation study on classification accuracy of the proposed OCTL-MSA framework with modality removal (%)

Configurations	Business	Family	People
Audio only	81.42%	82.16%	80.84%
Text only	88.63%	89.41%	87.92%
Video only	79.84%	80.27%	78.93%
Audio + Text	93.18%	94.06%	92.74%
Audio + Video	89.74%	90.42%	88.96%
Text + Audio	94.02%	94.88%	93.57%
Text + Video	92.84%	93.63%	91.94%
Video + Audio	88.91%	89.73%	87.84%
Video + Text	93.06%	93.87%	92.18%
Audio + Text+ Video	97.89%	98.19%	97.03%

Note: MSA = Multimodal Sentiment Analysis; OCTL = Optimized Cross-Modal Transfer Learning.

Table 8 presents the ablation study of the proposed OCTL-MSA framework across the business, family, and people topics under different modality configurations. The results show that unimodal configurations achieved lower classification performance, whereas bimodal combinations improved sentiment classification through enhanced multimodal feature interaction. Furthermore, the complete Audio + Text + Video configuration achieved the highest accuracies for business, family, and people topics, respectively, demonstrating the effectiveness of multimodal fusion for accurate SA.

4.4 Discussion

The proposed OCTL-MSA framework achieved superior classification performance across the business, family, and people topics. The misclassifications of different conversational topics are examined to assess the domain robustness of the model. Interestingly, there was a non-linear trend in performance: the highest overall accuracy was obtained with the topic of family, then it follows the topic of business and the lowest accuracy was obtained with the topic of generic people. This behavior can be explained by a modality investigation of the features:

Family dynamics with high emotional signaling: The interactions within the topic of family are highly emotional, with explicit displays of emotion. This gives powerful, clear acoustic differences (vocal energy) and clear facial expressions that enable easy convergence of the Audio and Video modalities. Moreover, the text domain has dense relational keywords (mom, husband, love) anchoring the text embeddings.

Uniformity of structure in business: The business topic is structurally uniform because of the vocabulary and professional behavioral restrictions, which minimally affect the volatility of unexpected features.

The uncertainties of generic interpersonal (people) data: The topic of people is not so good as it is mostly made of informal or social activities with friends. As opposed to family, the emotional expressions are structurally defended and delicate (e.g. the polite compliance or passive frustration), which leads to low-amplitude audio/visual cues. At the same time, it does not have the very specific vocabulary of corporate (business) or domestic (family) sub-context, and thus is the most complicated area to cross-modally fuse.

Despite the superior performance of the proposed framework, certain limitations remain in handling noisy and incomplete multimodal data. Variations in speech quality, facial occlusions, illumination changes and incomplete textual transcripts may affect multimodal feature extraction and classification stability. Furthermore, the current study has been validated only on the IEMOCAP dataset using English conversational data. Since emotional expressions and linguistic patterns may vary across languages and application domains, the generalization capability of the proposed framework for multilingual and cross-domain SA has not been fully explored.

Therefore, future work can focus on evaluating the proposed framework using multilingual and cross-domain multimodal datasets containing diverse conversational styles, emotional expressions, and real-world interaction scenarios. In addition, adaptive modality fusion, noise-robust feature extraction, and transformer-based contextual learning techniques can be incorporated to further improve robustness, scalability, and generalization capability. Furthermore, integrating larger multimodal datasets and real-time SA frameworks may enhance the applicability of the proposed model for effective MSA.

5. CONCLUSION

In this study, a CTL framework is proposed for efficient MSA across heterogeneous domains. The CTL framework employs MLDA for extracting semantically coherent multimodal representations, while LSTM and BERT models are utilized to effectively capture acoustic, visual, and textual features. Furthermore, total variance loss and KL-divergence-based domain alignment improve feature consistency and generalization performance across source and target domains. To further enhance the performance of CTL, an OCTL framework is developed by integrating the PKOA for efficient hyperparameter tuning of LSTM and BERT models. PKOA optimizes parameters such as learning rate, dropout rate, batch size, momentum, and weight decay, thereby reducing computational complexity and significantly improving prediction accuracy and overall MSA performance. Experimental results on the IEMOCAP dataset demonstrate that the proposed OCTL-MSA framework achieves superior classification accuracies of 97.89%, 98.19% and 97.03% for the business, family and people topics, respectively.

REFERENCES

[1] Churchill, R., Singh, L. (2022). The evolution of topic modeling. ACM Computing Surveys, 54(10s): 1-35. https://doi.org/10.1145/3507900

[2] Barde, B.V., Bainwad, A.M. (2017). An overview of topic modeling methods and tools. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, pp. 745-750. https://doi.org/10.1109/iccons.2017.8250563

[3] Alghamdi, R., Alfalqi, K. (2015). A survey of topic modeling in text mining. International Journal of Advanced Computer Science and Applications, 6(1). https://doi.org/10.14569/ijacsa.2015.060121

[4] Zhang, H.K., Cai, Y., Ren, H.P., Li, Q. (2022). Multimodal topic modeling by exploring characteristics of short text social media. IEEE Transactions on Multimedia, 25: 2430-2445. https://doi.org/10.1109/tmm.2022.3147064

[5] Zheng, Y., Zhang, Y.J., Larochelle, H. (2014). Topic modeling of multimodal data: An autoregressive approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, pp. 1370-1377. https://doi.org/10.1109/CVPR.2014.178

[6] Li, H., Qian, Y., Jiang, Y.C., Liu, Y.Z., Zhou, F. (2023). A novel label-based multimodal topic model for social media analysis. Decision Support Systems, 164: 113863. https://doi.org/10.1016/j.dss.2022.113863

[7] Wankhade, M., Rao, A.C.S., Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7): 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

[8] Medhat, W., Hassan, A., Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain

Shams Engineering Journal, 5(4): 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011
[9] Mejova, Y. (2009). Sentiment analysis: An overview. University of Iowa, United States.
[10] Devika, M.D., Sunitha, C., Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. Procedia Computer Science, 87: 44-49. https://doi.org/10.1016/j.procs.2016.05.124
[11] Dahal, B., Kumar, S.A., Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. Social Network Analysis and Mining, 9(1): 24. https://doi.org/10.1007/s13278-019-0568-8
[12] Chu, K.E., Keikhosrokiani, P., Asl, M.P. (2022). A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts. Pertanika Journal of Science & Technology, 30(4): 2535-2561. https://doi.org/10.47836/pjst.30.4.14
[13] Rana, T.A., Cheah, Y.N., Letchmunan, S. (2016). Topic modeling in sentiment analysis: A systematic review. Journal of ICT Research & Applications, 10(1): 76-93. https://doi.org/10.5614/itbj.ict.res.appl.2016.10.1.6
[14] Jeong, B., Yoon, J., Lee, J.M. (2019). Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management, 48: 280-290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009
[15] Tharwat, A., Gaber, T., Ibrahim, A., Hassanien, A.E. (2017). Linear discriminant analysis: A detailed tutorial. AI Communications, 30(2): 169-190. https://doi.org/10.3233/AIC-170729
[16] Reddy, D.L., Bindu, C.S. (2025). A coherence improved multimodal topic modeling for multimodal sentiment analysis. International Journal of Intelligent Engineering & Systems, 18(9): 737-752. https://doi.org/10.22266/ijies2025.1031.48
[17] Deng, L.J., Liu, B.Y., Li, Z.H. (2024). Multimodal sentiment analysis based on a cross-modal multihead attention mechanism. Computers, Materials & Continua, 78(1): 1157-1170. https://doi.org/10.32604/cmc.2023.042150
[18] Huang, J.H., Zhou, J., Tang, Z.C., Lin, J.Y., Chen, C.Y.C. (2024). TMBL: Transformer-based multimodal binding learning model for multimodal sentiment analysis. Knowledge-Based Systems, 285: 111346. https://doi.org/10.1016/j.knosys.2023.111346
[19] Cai, Y.J., Li, X.G., Zhang, Y.Y., Li, J.S., Zhu, F.Z., Rao, L. (2025). Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Scientific Reports, 15(1): 2126. https://doi.org/10.1038/s41598-025-85859-6
[20] Zhu, L.N., Zhao, H.Y., Zhu, ZC., Zhang, C.W., Kong, X.J. (2025). Multimodal sentiment analysis with unimodal label generation and modality decomposition. Information Fusion, 116: 102787. https://doi.org/10.1016/j.inffus.2024.102787
[21] Sun, H., Niu, Z.W., Wang, H.Y., et al. (2025). Multimodal sentiment analysis with mutual information-based disentangled representation learning. IEEE Transactions on Affective Computing, 16(3): 1606-1617. https://doi.org/10.1109/taffc.2025.3529732
[22] Wang, B.Q., Dong, G., Zhao, Y.Q., Li, R.G., Yin, W.F., Lu, L.H. (2025). Tripartite interaction representation learning for multi-modal sentiment analysis. Expert Systems with Applications, 268: 126279. https://doi.org/10.1016/j.eswa.2024.126279
[23] He, X.J., Fang, Y.J., Xiang, N., et al. (2025). Text-guided multi-level interaction and multi-scale spatial-memory fusion for multimodal sentiment analysis. Neurocomputing, 626: 129532. https://doi.org/10.1016/j.neucom.2025.129532
[24] Wang, R., Guo, C.P., Shabaz, M., Rida, I., Cambria, E., Zhu, X.X. (2025). CIME: Contextual interaction-based multimodal emotion analysis with enhanced semantic information. IEEE Transactions on Computational Social Systems, 13(3): 4001-4011. https://doi.org/10.1109/TCSS.2025.3572495
[25] Yu, L., Liu, A.Z. (2026). Multi-level perception cross-modal fusion framework for multimodal sentiment analysis. International Journal of Machine Learning and Cybernetics, 17(3): 91. https://doi.org/10.1007/s13042-026-02991-z
[26] Gao, C.Y., Liu, J.H., Wang, J.C., Chen, L.Y. (2026). TQ-CHAFN: Triple-query cross-modal hierarchical adaptive fusion network for multimodal sentiment analysis. Knowledge-Based Systems, 339: 115654. https://doi.org/10.1016/j.knosys.2026.115654
[27] Busso, C., Bulut, M., Lee, C.C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4): 335-359. https://doi.org/10.1007/s10579-008-9076-6