Neural Networks 162 (2023) 162–174

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2022 Special Issue

Lifelong Text-Audio Sentiment Analysis learning

Yuting Lin [a], Peng Ji [b], Xiuyi Chen [c], Zhongshi He [a,*]

[a] College of Computer Science, Chongqing University, Chongqing, China
[b] Department of Computing, Hong Kong Polytechnic University, Hongkong, China
[c] Baidu Inc., Beijing, China

ARTICLE INFO

Article history:
Available online 17 February 2023

Keywords:
Text-audio sentiment analysis
Lifelong machine learning
Cross-modality learning
Multi-task learning

ABSTRACT

Sentiment analysis refers to the mining of textual context, which is conducted with the aim of identifying and extracting subjective opinions in textual materials. However, most existing methods neglect other important modalities, e.g., the audio modality, which can provide intrinsic complementary knowledge for sentiment analysis. Furthermore, much work on sentiment analysis cannot continuously learn new sentiment analysis tasks or discover potential correlations among distinct modalities. To address these concerns, we propose a novel Lifelong Text-Audio Sentiment Analysis (LTASA) model to continuously learn text-audio sentiment analysis tasks, which effectively explores intrinsic semantic relationships from both intra-modality and inter-modality perspectives. More specifically, a modality-specific knowledge dictionary is developed for each modality to obtain shared intra-modality representations among various text-audio sentiment analysis tasks. Additionally, based on information dependence between text and audio knowledge dictionaries, a complementarity-aware subspace is developed to capture the latent nonlinear inter-modality complementary knowledge. To sequentially learn text-audio sentiment analysis tasks, a new online multi-task optimization pipeline is designed. Finally, we verify our model on three common datasets to show its superiority. Compared with some baseline representative methods, the capability of the LTASA model is significantly boosted in terms of five measurement indicators.

1. Introduction

Sentiment analysis, which is commonly referred to as opinion mining (Birjali, Kasri, & Beni-Hssane, 2021; Cambria, Das, Bandyopadhyay, & Feraco, 2017; Du, Liu, Peng, & Jin, 2022), has recently attracted interest and mostly focuses on the examination of people's sentiment towards specific entities. As sentiment analysis technology has rapidly developed, it significantly improves the performance of sentiment analysis tasks and promotes the cross-integration of sentiment analysis with other fields. In addition, sentiment analysis (Peng, Lu, Pan, & Liu, 2021; Xue, Zhang, Niu, & Wu, 2022; Yang, Xu, & Gao, 2020) derives a series of new tasks such as dialogue sentiment analysis and multi-modality sentiment analysis. To date, it has been widely used in numerous real-world scenarios, such as facial expression recognition (Ouzar, Bousefsaf, Djeldjli, & Maaoui, 2022), voiceprint recognition (Li, Li, Xiong, Chen, & Li, 2021), and human–computer emotional interaction (Chowdary, Nguyen, & Hemanth, 2021).

Generally, most sentiment analysis methods (Bibi et al., 2022; Phan, Nguyen, & Hwang, 2022) consider only one modality, e.g., the text modality, to explore emotional expressions during communication by using words, phrases or their semantic associations. Nevertheless, it is not easy to obtain the correct emotional state with a single modality, since it ignores the complementary information of other important modalities, e.g., audio modality, and cannot accurately determine the emotional expression in some exceptional cases. Taking a real-world example to illustrate the importance of audio modality, there are significant differences in emotional expression for the same utterance "I'm so happy the plane is late" expressed in different tones. Different tones of the word "happy" also have a different influence on the emotion classification results, while the word "happy" in the text modality has only one emotion result (i.e., happiness). Thus, it is necessary to consider the audio modality with complementary information to improve the performance of sentiment analysis instead of determining the emotions in text modalities via the content of the text.

To this end, some representative studies (Peng et al., 2021; Yang et al., 2020; Zadeh, Liang, Poria, Vij et al., 2018) focus on combining the text and audio modalities to obtain informative emotional features and improve the performance of sentiment

* Corresponding author.
E-mail addresses: yutinglin@cqu.edu.cn (Y. Lin), peng01.ji@connect.polyu.hk (P. Ji), chenxiuyi01@baidu.com (X. Chen), zshe@cqu.edu.cn (Z. He).

https://doi.org/10.1016/j.neunet.2023.02.008
0893-6080/

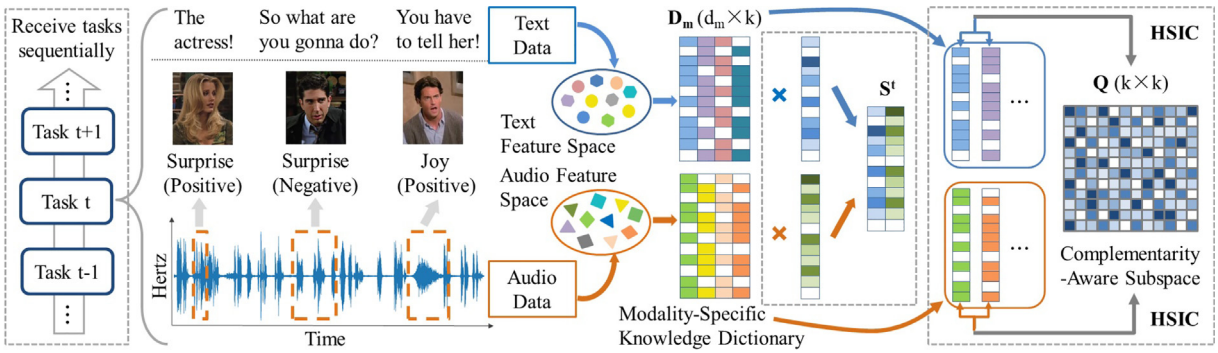

Fig. 1. Graphical illustration of the proposed LTASA model to consecutively learn new text-audio sentiment analysis tasks online. It includes a modality-specific knowledge dictionary for each (i.e., text or audio) modality to obtain intra-modality information shared among various tasks and a complementarity-aware subspace to capture inter-modality inherent complementary knowledge among different modalities.

analysis systems. For example, Yang et al. (2020) leveraged audio information to assist the text modality in dynamically assigning weights to words. Zadeh, Liang, Poria, Vij et al. (2018) proposed a multi-attention loop network that exploited multi-attention blocks to explore interactions across different modalities. In general, these existing multi-modality sentiment analysis methods (Peng et al., 2021; Yang et al., 2020; Zadeh, Liang, Poria, Vij et al., 2018) unreasonably assume that all sentiment analysis tasks are fixed in advance and cannot change over time. However, in real-world dynamic scenarios where sentiment analysis tasks consecutively arrive, the performance of these methods will significantly degrade (i.e., catastrophic forgetting on old tasks). A straightforward solution is to store all training data and retrain the sentiment analysis model. However, it consumes ample storage space and high computational costs, which make them impractical in real-world applications.

Another trivial solution is to adopt the existing single-modality lifelong learning approaches (Isele, Rostami, & Eaton, 2016; Rannen, Aljundi, Blaschko, & Tuytelaars, 2017; Ruvolo & Eaton, 2013) to learn continuous text-audio tasks by integrating both text and audio modalities into a high-dimensional vector. Unfortunately, it increases the computational costs by handling high-dimensional features and ignores the large distribution gap (Dong, Cong, Sun, Fang, & Ding, 2021; Dong, Cong, Sun, Zhong, & Xu, 2020) among different modalities that may lead to significant performance degradation. For example, text features contain semantic and grammatical information of sentences, while acoustic features include rhythmic, spectral, and cepstral information from audio signals (Alías, Socoró, & Sevillano, 2016). In addition, this method neglects inter-modality intrinsic complementary information, since different modalities are independent and equal to one another in single-model lifelong learning by sharing one common knowledge base. To tackle this issue, a learning model rLM^{2}L for multi-task multi-view representations under a lifelong learning setting was proposed in Sun, Cong, Li, and Fu (2018), which employed a group of modality-specific libraries to accumulate relevant information within intra-modality. However, this method could not settle the question of complementarity across different modalities well for text-audio sentiment analysis, since it neglected the nonlinear dependence embedding various sensors or feature representations. Hence, our chief aim is to mine the inherent nonlinear complementary knowledge between text and audio modalities in the lifelong learning environment.

To address these challenges, we propose a new Lifelong Text-Audio Sentiment Analysis (LTASA) learning model, as shown in Fig. 1, which is configured with the ability to continuously learn online text-audio sentiment analysis tasks. The LTASA model can effectively explore the shared features among various tasks in the intra-modality respect and simultaneously extract potential nonlinear complementary relationships across distinct modalities in the inter-modality respect. Specifically, concerning correlations within intra-modality, a modality-specific knowledge dictionary is established separately for text and audio modalities to capture shared experiences across various tasks and semantic knowledge acquired from new tasks is continuously integrated into this shared embedding space. For inter-modality relationships, a complementarity-aware subspace is constructed to explore the nonlinear complementary knowledge shared among the text and audio modalities by incorporating the Hilbert–Schmidt independence criterion (HSIC) (Gretton, Bousquet, Smola, & Schölkopf, 2005) on distinct modality-specific knowledge dictionaries. As a result of the non-convexity and NP-hard characteristics in our model optimization process, an Alternating Direction Method of Multipliers (ADMM) (Boyd, Parikh, & Chu, 2011) policy is utilized to optimize our LTASA model when new text-audio sentiment analysis tasks continuously arrive. Finally, to verify the validity of the LTASA model, synthetic experiments are conducted on three public datasets, and the numerical results show that the LTASA model is robust and has significant performance improvements compared to the state-of-the-art approaches.

The major contributions in this paper are summarized as follows:

- A novel LTASA model is proposed to continuously learn text-audio sentiment analysis tasks under lifelong learning settings. To the best of our knowledge, this paper is an earlier exploration of lifelong text-audio sentiment analysis learning in the multimodal emotion analysis field.
- A complementarity-aware subspace is designed to explore nonlinear complementary knowledge shared across text and audio modalities by considering HSIC on different modality-specific knowledge dictionaries.
- We propose a novel online multi-task majorization strategy to process the LTASA model. Extensive experiments on several benchmark datasets illustrate its efficiency compared to the state-of-the-art approaches.

2. Related work

In this part, some typical works on multimodal sentiment analysis, lifelong machine learning and their combinations are introduced.

2.1. Multimodal sentiment analysis

In the field of sentiment analysis (Cambria et al., 2017; Poria, Cambria, Bajpai, & Hussain, 2017), multimodal information, e.g., text and audio modalities, is crucial for improving the performance of sentiment analysis. Generally, multimodal fusion

strategies for sentiment analysis are divided into three main categories: feature-level fusion, decision-level fusion, and hybrid fusion.

Feature-level fusion simply concatenates features extracted from different modalities. For example, the authors of Hazarika et al. (2018), Majumder et al. (2019) and Poria, Chaturvedi, Cambria, and Hussain (2016) performed feature-level fusion of multiple modalities for sentiment analysis. Recently, some works have proposed reinforcement learning-based multimodal sentiment analysis. Zhang, Li, Wang, Cambria, and Li (2021) revised the outcomes of reinforcement learning module recognition by first combining reinforcement learning with domain knowledge acquired from sentiment pairs.

Decision-level fusion combines predictions from various modalities to make a final decision. Wu and Liang (2010) integrated the prediction results based on prosodic features and semantic labels in a weighted manner. Zhao, Cao, Lin, Yu, and Cao (2021) proposed a decision fusion model whose results depended on modalities with higher reliability. However, feature-level and decision-level fusion methods neglect the heterogeneous distribution gap across different modalities.

Therefore, we focus on the **hybrid fusion** strategy in this paper, e.g., some scalable multimodal fusion methods, multi-task learning for multimodal sentiment analysis, and recent multimodal resources. For some **scalable multimodal fusion methods**, Cambria, Howard, Hsu, and Hussain (2013) proposed a sentic blending, where the output of each module was mapped into a multidimensional vector space AffectiveSpace; then, the information of all modalities was fused into a dynamic multidimensional flow. Zadeh, Chen, Poria, Cambria, and Morency (2017) proposed a tensor-based fusion method to outer-product the features of the three modalities to explore the correlations among different modalities. Zadeh, Liang, Poria, Cambria and Morency (2018) designed a graph fusion network to fuse information of different modalities, and each fusion was treated as a vertex of the graph. Gkoumas, Li, Yu, and Song (2021) employed quantum theory to further model the correlation and inseparability across different modalities. For **multi-task learning for multimodal sentiment analysis,** Yang et al. (2022) learned a representation of the text modality in the first stage and fused it with other modalities in the second stage, where the fused information was used with each modality for multi-task learning. Concerning **recent multimodal resources,** Stappen et al. (2021) designed a toolbox to extract time-series features of continuous signals, which helped transform the emotional dimension into discrete emotional categories.

However, these existing sentiment analysis models assume that all training data are fixed in advance and cannot consecutively learn new sentiment analysis tasks, which is impractical in real-world applications.

2.2. Lifelong learning

Lifelong learning (Chen, Ma, & Liu, 2018; Dong, Wang et al., 2022; Gai, Chen, & Wang, 2021), which is also known as continuous learning, is an appealing paradigm of machine learning that incrementally learns new tasks by accumulating the experiences gained from past tasks. To solve the catastrophic forgetting problem, some continuous learning methods have been proposed, e.g., replay-based methods, regularization-based methods, and reusable knowledge-based methods.

Replay-based methods: Rebuffi, Kolesnikov, Sperl, and Lampert (2017) designed a classic replay-based incremental learning approach that saved some of the most representative samples for each previous task and trained the saved data with the new task. Lopez-Paz and Ranzato (2017) proposed a new gradient memory segment algorithm, the idea of which was to update the parameters of the new task without affecting the parameters of the previous tasks. However, replay-based incremental learning approaches require additional data storage space, while regularization-based methods tackle this problem by adding a regularization term to the loss function.

Regularization-based methods: Li and Hoiem (2017) applied a fine-tuning method to train a new model on a new task and utilized a knowledge distillation technique to distill knowledge from the old model to the new model without using previous data. Kirkpatrick et al. (2017) introduced a regularization term for the weights to control the direction of weight optimization with less variation to the more important parameters of previous tasks.

Reusable knowledge-based methods: Ruvolo and Eaton (2013) proposed an effective lifelong learning model (ELLA) for multi-task learning, which could continuously learn a series of tasks by constructing a shared knowledge base to deliver effective learning experiences. Irfan et al. (2021) focused on the adaptability of users, tasks, context, and environment in lifelong learning and long-range interaction in various domains. For multi-view lifelong learning, Zhang, Fu, Liu, Liu, and Cao (2015) proposed a decision function to combine lifelong learning and multi-view learning in a shared potential space for the first time. In practical applications, Dong, Cong, Sun and Zhang (2022) established a modal-specific knowledge library and a cross-modal invariant space to probe the common knowledge of different tasks and complementary knowledge among visual-tactile modalities, respectively.

To achieve single-modality lifelong sentiment analysis learning, Chen et al. (2018) proposed a lifelong learning multi-domain emotion classification method and introduced a punishment term to efficaciously employ the previously learned knowledge obtained in past tasks. Xia, Jiang, and He (2017) tackled the problem of emotion analysis of massive social media by employing lifelong learning settings with remote supervision. However, these previous lifelong sentiment analysis methods only consider single-modality sentiment prediction and cannot be directly applied to lifelong multi-modality sentiment analysis tasks.

From this discussion, this work is the first exploration of lifelong text-audio sentiment analysis learning. Moreover, existing lifelong multi-modality learning methods cannot effectively capture complementary knowledge among different modalities for sentiment analysis because they neglect the nonlinear dependence among feature representations of different modalities introduced by significant distribution divergence.

3. The proposed model

3.1. Problem definition

Following traditional multimodal lifelong learning methods (Ruvolo & Eaton, 2013; Sun et al., 2018), we set the fair experimental settings of task-incremental learning for lifelong sentiment analysis and conduct comparison experiments with baseline methods under identical settings to verify the effectiveness of our model.

Our proposed LTASA model is conceived with the aim of learning continuous text-audio sentiment analysis tasks. Denote $\mathcal{W} = \{\mathcal{W}^t\}_{t=1}^{T}$ as a sequence of consecutive learning tasks, where T is the total number of tasks. For the tth sentiment analysis task $\mathcal{W}^t = \{f_m^t, \mathbf{X}_m^t\}_{m=1}^{M}$, where M is the number of modalities, f_m^t is the linear mapping of the mth modality in the tth task, $\mathbf{X}_m^t = [\mathbf{x}_{1m}^t, \mathbf{x}_{2m}^t, \ldots, \mathbf{x}_{n_t m}^t] \in \mathbb{R}^{d_m \times n_t}$ and $\mathbf{Y}^t = [y_1^t, y_2^t, \ldots, y_{n_t}^t] \in \mathbb{R}^{n_t}$ are n_t pairs of samples and corresponding labels, respectively, and d_m is the characteristic dimension of the mth modality. Intuitively, let

$M = 2$ due to the text and audio modalities. To learn the linear mapping f_m^t, i.e., $f_m^t : \mathbf{X}_m^t \rightarrow \mathbf{Y}^t$, the function f_m^t is formulated as $\mathbf{Y}^t = f_m^t(\mathbf{X}_m^t; \boldsymbol{\theta}_m^t) = \langle \mathbf{X}_m^t, \boldsymbol{\theta}_m^t \rangle$, where $\boldsymbol{\theta}_m^t \in \mathbb{R}^{d_m}$ is the task classifier. When the samples from a new text-audio sentiment analysis task (such as the tth task $\mathcal{W}^t$) come, the LTASA model is designed to predict the new task and $t-1$ previously learned tasks without the need for revisiting the training data in the previous tasks. In this work, our model has no prior knowledge about the distribution of each sentiment analysis task, the order of arriving tasks, and the total number of tasks.

3.2. Lifelong single-modality sentiment analysis

Most existing lifelong single-modality sentiment analysis models (Chen et al., 2018; Xia et al., 2017) only learn text-modality sentiment analysis tasks in a consecutive manner. They ignore other important modalities, e.g., the audio modality, which can provide complementary knowledge to improve sentiment analysis performance. Generally, these single-modality lifelong learning models (Ruvolo & Eaton, 2013) assume that the task classifier $\boldsymbol{\theta}_m^t$ of the text modality in all tasks shares a communal knowledge dictionary $\mathbf{D} \in \mathbb{R}^{d_m \times k}$, where parameter k represents the number of bases in $\mathbf{D}$. Thus, $\boldsymbol{\theta}_m^t$ can also be expressed as $\boldsymbol{\theta}_m^t = \mathbf{D}\mathbf{s}_m^t$, where $\mathbf{s}_m^t$ indicates the sparse coding in the modality-specific knowledge dictionary $\mathbf{D}$. The sparse coding $\mathbf{s}_m^t$ utilizes only a small number of atoms to represent the latent characterizations of the tth task in the mth modality. Finally, the objective of lifelong single-modality sentiment analysis is formally written as follows:

$$\min_{\mathbf{D}} \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}_m^t} \left\{ \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}(f_m^t(\mathbf{x}_{im}^t, \mathbf{D}\mathbf{s}_m^t), y_i^t) + \lambda_1 \|\mathbf{s}_m^t\|_1 \right\} + \mu_1 \|\mathbf{D}\|_F^2, \tag{1}$$

where the symbol $\mathcal{L}(\cdot)$ denotes the classification loss function. $\lambda_1 > 0$ and $\mu_1 > 0$ are the trade-off coefficients that control the sparsity of $\mathbf{s}_m^t$ and the importance of $\mathbf{D}$. $\|\cdot\|_1$ and $\|\cdot\|_F$ represent the ℓ_1 norm and Frobenius norm, respectively.

Eq. (1) shows that the shared knowledge dictionary $\mathbf{D}$ in the single-modality lifelong learning method can continuously accumulate task-shared knowledge for text-modality sentiment analysis tasks. However, it is infeasible to directly extend Eq. (1) to lifelong text-audio sentiment analysis learning by concatenating the features of multiple modalities into a high-dimensional vector, since it neglects the complementary information among different modalities. Moreover, the shared knowledge dictionary $\mathbf{D}$ cannot capture intra-modality task-shared knowledge due to the significant distribution gap among different modalities.

3.3. Lifelong text-audio sentiment analysis

To handle the aforementioned problems, we propose an LTASA model, as shown in Fig. 1. It can explore shared knowledge of different tasks from an intra-modality aspect. It also learns complementary information from an inter-modality aspect to improve the performance of lifelong multi-modality sentiment analysis learning.

Intra-modality correlations. Define the finite modality set for the sentiment analysis as $\mathcal{M} = \{1, \ldots, M\}$ and let $M = 2$ due to the text and audio modalities. In addition, for the tth task, the ith sample data of its mth modality is denoted as $\mathbf{x}_{im}^t \in \mathbb{R}^{d_m}$, where d_m is the feature dimension. To explore the relationship of different tasks within each modality, a modality-specific knowledge dictionary $\mathbf{D_m}$ is defined for the mth modality to accumulate task-shared knowledge in the intra-modality aspect. Specifically, the classifier $\{\boldsymbol{\theta}_m^t\}_{t=1}^T$ for the mth modality under the tth task is decomposed into $\mathbf{D}_m$ and $\mathbf{s}_m^t$ via a matrix factorization mechanism, and θ_m^t is concretely written as follows:

$$\boldsymbol{\theta}_m^t = \mathbf{D}_m \mathbf{s}_m^t, \qquad \forall m \in \mathcal{M} = \{1, \ldots, M\}, \tag{2}$$

where $\mathbf{D}_m \in \mathbb{R}^{d_m \times k}$ denotes the mth modality knowledge dictionary, and k is the number of atoms in knowledge dictionary $\mathbf{D}_m$. Moreover, $\mathbf{s}_m^t \in \mathbb{R}^k$ is the characteristic representation of the corresponding task for the mth modality in the tth task. Obviously, the constructed modality-specific knowledge dictionaries $\{\mathbf{D}_1, \ldots, \mathbf{D}_m\}$ can explore the intrinsic task-shared relationships among continuous text-audio tasks within intra-modality. Therefore, Eq. (1) can be reformulated as follows:

$$\min_{\mathbf{D}_m} \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}_m^t} \left\{ \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}(f_m^t(\mathbf{x}_{im}^t, \mathbf{D}_m\mathbf{s}_m^t), y_i^t) + \lambda_1 \|\mathbf{s}_m^t\|_1 \right\} + \mu_1 \|\mathbf{D}_m\|_F^2. \tag{3}$$

Inter-modality relationships. Although Eq. (3) can explore the task-shared relationship among different text-audio sentiment analysis tasks within intra-modality, it cannot fully exploit the complementary information in inter-modality, which significantly decreases the performance of lifelong text-audio sentiment analysis learning. Considering this problem, we focus on learning a complementarity-aware subspace to explore the nonlinear inter-modality complementary information by exploring inter-dependent relations among different modalities. As introduced in Li, Zhang, Hu, Zhu, and Wang (2019), HSIC can effectively measure the distribution dependency among different modalities. As a result, HSIC (Gretton et al., 2005) is utilized to learn a complementarity-aware subspace $\mathbf{Q} \in \mathbb{R}^{k \times k}$ by maximizing its semantic dependence with different modality-specific knowledge dictionaries $\{\mathbf{D}_m\}_{m=1}^M$.

Furthermore, the large heterogeneity of different modality-specific dictionaries makes it difficult for traditional lifelong multi-modality methods (Dong, Cong et al., 2022; Li, Chandrasekaran, & Huan, 2017; Sun et al., 2018) to explore the modality complementary information. To tackle this problem, we follow Cao, Zhang, Fu, Liu, and Zhang (2015) and Wang et al. (2019) to utilize an inner product kernel and a linear kernel, which are used to better explore the high-level nonlinear important responses and low-level linear knowledge of complementary information among different modalities. Specifically, the inner product kernel $\mathbf{K}_m^1 = \mathbf{D}_m^\top \mathbf{D}_m$ is used for the mth modality-specific knowledge dictionary $\mathbf{D}_m$ and the linear kernel $\mathbf{K}_m^2 = \mathbf{Q} - \mathbf{A}$ is used for the complementarity-aware subspace, where $\mathbf{Q} \geq 0$ and $\mathbf{A}$ is a diagonal matrix whose ith diagonal entry is defined as $\sum_j \mathbf{Q}_{ij}$. Thus, to explore the inter-modality nonlinear complementary information by constructing a complementarity-aware subspace, we have the following constraint:

$$\begin{aligned} &\max_{\mathbf{Q}} \sum_{m=1}^{M} \mathrm{HSIC}(\mathbf{D}_m, \mathbf{Q}) = \max_{\mathbf{Q}} \sum_{m=1}^{M} tr(\mathbf{K}_m^1 \mathbf{H} \mathbf{K}_m^2 \mathbf{H}) \\ &= \max_{\mathbf{Q}} \sum_{m=1}^{M} tr(\mathbf{D}_m^\top \mathbf{D}_m \mathbf{H}(\mathbf{Q} - \mathbf{A})\mathbf{H}) = -\max_{\mathbf{Q}} \sum_{m=1}^{M} tr(\mathbf{D}_m \mathbf{H} \mathbf{P} (\mathbf{D}_m \mathbf{H})^\top) \\ &= \min_{\mathbf{Q}} \sum_{m=1}^{M} tr(\mathbf{D}_m \mathbf{H} \mathbf{P} (\mathbf{D}_m \mathbf{H})^\top), \end{aligned} \tag{4}$$

$\mathbf{H} = \mathbf{I} - k^{-1}\mathbf{1}\mathbf{1}^\top$ is the centring matrix that centres the complete space, and $\mathbf{P} = \mathbf{A} - \mathbf{Q}$ is considered the Laplacian matrix of $\mathbf{Q}$. Furthermore, to make the constructed similarity points lie within the union of complementarity-aware subspace $\mathbf{Q}$, we constrain $\mathbf{Q}_i^\top \mathbf{1} = 1$, where $\mathbf{Q}_i$ is the ith column of $\mathbf{Q}$. These constraints

also encourage the proposed model to better capture complementary information among different modalities while guaranteeing stability. Therefore, the following optimization objective is obtained:

$$\min_{\mathbf{Q}} \sum_{m=1}^{M} tr(\mathbf{D}_m\mathbf{H}\mathbf{P}(\mathbf{D}_m\mathbf{H})^\top)$$
$$\text{s.t.} \quad \mathbf{Q}_i^\top \mathbf{1} = 1, \ \mathbf{Q}_i \geq 0, \ \forall i = 1, \ldots, k. \tag{5}$$

Before providing a solution for the complementary-aware subspace $\mathbf{Q}$ proposed in Eq. (4), its potential implications are investigated. When the modality-specific knowledge dictionary $\mathbf{D}_m$ is centred, multiplying $\mathbf{D}_m$ by the centring matrix $\mathbf{H}$ does not make any change, i.e., $\mathbf{D}_m = \mathbf{D}_m\mathbf{H}$. Therefore, Eq. (5) can be reformulated as the following function:

$$\begin{aligned}
&\min_{\forall i, \ \mathbf{Q}_i^\top \mathbf{1}=1, \ \mathbf{Q}_i \geq 0} \sum_{m=1}^{M} tr(\mathbf{D}_m\mathbf{H}\mathbf{P}(\mathbf{D}_m\mathbf{H})^\top) \\
&= \min_{\forall i, \ \mathbf{Q}_i^\top \mathbf{1}=1, \ \mathbf{Q}_i \geq 0} \sum_{m=1}^{M} tr(\mathbf{D}_m\mathbf{P}\mathbf{D}_m^\top) \\
&= \min_{\forall i, \ \mathbf{Q}_i^\top \mathbf{1}=1, \ \mathbf{Q}_i \geq 0} \sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{D}_m^i - \mathbf{D}_m^j\|_2^2 \mathbf{Q}_{\mathbf{ij}},
\end{aligned} \tag{6}$$

where $\mathbf{D}_m^i$ and $\mathbf{D}_m^j$ are the ith column and jth column of the knowledge dictionary $\mathbf{D}_m$, respectively. Intuitively, if knowledge dictionaries from different modalities have a smaller measure distance (i.e., larger similarity), the complementarity-aware subspace $\mathbf{Q}$ will have a larger complementarity value, which implies that more inter-modality shared complementarity information will be captured, and vice versa.

However, simply optimizing Eq. (6) may result in a trivial solution such that the learned complementarity-aware subspace $\mathbf{Q}$ is an identity matrix. Furthermore, the complementarity-aware subspace $\mathbf{Q}$ in Eq. (6) cannot address large information corruptions such as outliers and noise; thus, the recovered complementarity-aware subspace may not be robust. To prevent the trivial solution, the Frobenius norm is applied to $\mathbf{Q}$. In addition, to address the outliers, the ℓ_2 distance between $\mathbf{D}_m^i$ and $\mathbf{D}_m^j$ is replaced with the ℓ_1 distance. Consequently, Eq. (6) can be reformulated as follows:

$$\begin{aligned}
\min_{\mathbf{Q}} \ &\beta \sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{D}_m^i - \mathbf{D}_m^j\|_1 \mathbf{Q}_{ij} + \lambda_2 \|\mathbf{Q}\|_F^2, \\
&\text{s.t.} \quad \mathbf{Q}_i^\top \mathbf{1} = 1, \ \mathbf{Q}_i \geq 0, \ \forall i = 1, \ldots, k,
\end{aligned} \tag{7}$$

where β and λ_2 are nonnegative trade-off parameters.

Furthermore, due to different contributions of modalities to lifelong text-audio sentiment analysis learning, the constraint $\|\cdot\|_{1,\infty}$ is utilized for task characterizations $\{\mathbf{s}_1^t, \ldots, \mathbf{s}_M^t\}$ across different modalities. Then, the objective formulation of the proposed LTASA model is written as follows:

$$\begin{aligned}
\min_{\mathbf{D}_m, \ \mathbf{Q}} \ &\frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{S}^t} \Big\{ \frac{1}{n_t}\sum_{i=1}^{n_t} \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}(f_m^t(\mathbf{x}_{im}^t, \mathbf{D}_m\mathbf{s}_m^t), y_i^t) \\
&+ \lambda_1 \|\mathbf{S}^t\|_{1,\infty} \Big\} \\
&+ \beta \sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{D}_m^i - \mathbf{D}_m^j\|_1 \mathbf{Q}_{ij} + \lambda_2 \|\mathbf{Q}\|_F^2, \\
&\text{s.t.} \quad \mathbf{Q}_i^\top \mathbf{1} = 1, \ \mathbf{Q}_i \geq 0, \ \forall i = 1, \ldots, k,
\end{aligned} \tag{8}$$

$\mathbf{S}^t = [\mathbf{s}_1^t, \ldots, \mathbf{s}_M^t] \in \mathbb{R}^{k \times M}$ is the potential representation matrix from all sentiment analysis modalities, which can measure the contribution of each modality to the tth sentiment analysis task.

4. Model optimization

The detailed optimization procedure of the LTASA model is illustrated in this chapter. Since the objective function is a nonconvex function over the variables $\{\mathbf{S}^t, \mathbf{D}_m, \mathbf{Q}\}$ in Eq. (8), we follow the ADMM algorithm (Boyd et al., 2011) to split the objective into several subproblems. This method optimizes only one variable during each iteration while fixing other variables. As introduced in Boyd et al. (2011), this optimization strategy has been proven to have a stable convergence and achieve the optimal performance for nonconvex optimization. Therefore, the objective function is decomposed into small subproblems, which are solved by the ADMM algorithm by alternately optimizing each variable.

4.1. Taylor expansion

Before calculating the optimization of subproblems, we adopt the Taylor expansion (Ruvolo & Eaton, 2013) to approximate the first term in Eq. (8). Define the first term in Eq. (8) as $\mathcal{L}_m^t(\mathbf{D}_m\mathbf{s}_{\mathbf{m}}^t)$ and approximate it by the second-order Taylor expansion of $\mathcal{L}_m^t(\mathbf{D}_m\mathbf{s}_m^t)$ near $\mathbf{D}_m\mathbf{s}_m^t = \boldsymbol{\theta}_m^t$. The specific expansion is shown in Appendix. Therefore, we obtain $\boldsymbol{\theta}_m^t$, and the Hessian matrix $\mathbf{H}_m^t$ is defined as follows:

$$\begin{aligned}
\boldsymbol{\theta}_m^t &= \arg\min_{\mathbf{D}_m\mathbf{s}_m^t} \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}(f_m^t(\mathbf{x}_{im}^t, \mathbf{D}_m\mathbf{s}_m^t), y_i^t), \\
\mathbf{H}_m^t &= \nabla^2_{\mathbf{D}_m\mathbf{s}_m^t, \mathbf{D}_m\mathbf{s}_m^t} \mathcal{L}(\mathbf{D}_m\mathbf{s}_m^t)|_{\mathbf{D}_m\mathbf{s}_m^t = \boldsymbol{\theta}_m^t}.
\end{aligned} \tag{9}$$

For the second-order Taylor expansion of $\mathcal{L}(\mathbf{D}_m\mathbf{s}_m^t)$ in Appendix, its constant and linear terms are ignored in the optimization process, and it is then employed to substitute for the first term of Eq. (8), which results in the following optimization function:

$$\begin{aligned}
\min_{\mathbf{D}_m, \ \mathbf{Q}} \ &\frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{S}^t} \Big\{ \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\theta}_m^t - \mathbf{D}_m\mathbf{s}_m^t\|_{\mathbf{H}_m^t}^2 + \lambda_1 \|\mathbf{S}^t\|_{1,\infty} \Big\} \\
&+ \beta \sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{D}_m^i - \mathbf{D}_m^j\|_1 \mathbf{Q}_{ij} + \lambda_2 \|\mathbf{Q}\|_F^2, \\
&\text{s.t.} \quad \mathbf{Q}_i^\top \mathbf{1} = 1, \ \mathbf{Q}_i \geq 0, \ \forall i = 1, \ldots, k.
\end{aligned} \tag{10}$$

Furthermore, to make the objective function Eq. (10) more concise, inspired by Kodirov, Xiang, Fu, and Gong (2016), the second term in Eq. (10) is simplified as follows:

$$\sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=1}^{k} \|\mathbf{D}_m^i - \mathbf{D}_m^j\|_1 \mathbf{Q}_{ij} = \sum_{m=1}^{M} \|\mathbf{D}_m\mathbf{B}_{\mathbf{Q}}\|_1, \tag{11}$$

where $\mathbf{B}_{\mathbf{Q}} = \mathbf{U}\boldsymbol{\Sigma}^{\frac{1}{2}}$, and the multipliers $\mathbf{U}$ and $\boldsymbol{\Sigma}$ can be calculated by the eigenvalue decomposition of the Laplace matrix $\mathbf{P} = \mathbf{A} - \mathbf{Q} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top$. To make this subproblem easier to resolve, we assume an auxiliary variable $\mathbf{G}_m$ instead of $\mathbf{D}_m\mathbf{B}_{\mathbf{Q}}$ as an additional constraint (i.e., $\mathbf{G}_m = \mathbf{D}_m\mathbf{B}_{\mathbf{Q}}$). Therefore, Eq. (10) can be expressed as an augmented Lagrangian objective:

$$\begin{aligned}
\min_{\mathbf{D}_m, \ \mathbf{G}_m, \ \mathbf{Q}, \ \mathbf{Z}} \ &\frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{S}^t} \Big\{ \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\theta}_m^t - \mathbf{D}_m\mathbf{s}_m^t\|_{\mathbf{H}_m^t}^2 + \lambda_1 \|\mathbf{S}^t\|_{1,\infty} \Big\} \\
&+ \beta \sum_{m=1}^{M} \|\mathbf{G}_m\|_1 + \lambda_2 \|\mathbf{Q}\|_F^2 + \Phi(\mathbf{Z}, \mathbf{G}_m - \mathbf{D}_m\mathbf{B}_{\mathbf{Q}}), \\
&\text{s.t.} \quad \mathbf{Q}_i^\top \mathbf{1} = 1, \ \mathbf{Q}_i \geq 0, \ \forall i = 1, \ldots, k,
\end{aligned} \tag{12}$$

where $\Phi(\mathbf{Z}, \mathbf{C}) = \langle \mathbf{Z}, \mathbf{C} \rangle + \frac{\mu}{2}\|\mathbf{C}\|_F^2$, $\mathbf{Z}$ is the Lagrangian multiplier, $\langle \cdot, \cdot \rangle$ denotes the inner product of the matrix, and μ is the quadratic penalty term coefficient.

Based on the above discussion, the optimization function of our model in Eq. (12) has five variables: $\{\mathbf{S}^t, \mathbf{G}_m, \mathbf{Q}, \mathbf{D}_m, \mathbf{Z}\}$, which are optimized using the ADMM algorithm. The optimization details are presented in the following subsections.

4.2. Updating $\mathbf{S}^t$ while fixing other variables

While fixing the variables $\{\mathbf{G}_m, \mathbf{Q}, \mathbf{D}_m, \mathbf{Z}\}$, the optimization function over the optimization variable $\mathbf{S}^t$ is as follows:

$$\min_{\mathbf{S}^t} \frac{1}{M}\sum_{m=1}^{M}\|\boldsymbol{\theta}_m^t - \mathbf{D}_m\mathbf{s}_m^t\|_{\mathbf{H}_m^t}^2 + \lambda_1\|\mathbf{S}^t\|_{1,\infty}. \tag{13}$$

The variable $\mathbf{S}^t$ is solved by using the proximal alternating linearized minimization (PALM) (Pock & Sabach, 2016) strategy, which can be written as follows:

$$\begin{aligned}\mathbf{S}^t &= \arg\min_{\mathbf{S}} \mathfrak{g}(\mathbf{S}) + r(\mathbf{S}),\\ \mathfrak{g}(\mathbf{S}) &= \frac{1}{M}\sum_{m=1}^{M}\|\sqrt{\mathbf{H}_m^t}\boldsymbol{\theta}_m^t - \sqrt{\mathbf{H}_m^t}\mathbf{D}_m\mathbf{s}_m^t\|_F^2.\end{aligned} \tag{14}$$

We multiply $\boldsymbol{\theta}_m^t$ and $\mathbf{D}_m$ by $\sqrt{\mathbf{H}_m^t}$ so that $\mathfrak{g}(\mathbf{S})$ is a convex function. $r(\mathbf{S}) = \lambda_1\|\mathbf{S}\|_{1,\infty}$ represents the regularization term. Since the convex function $\mathfrak{g}(\mathbf{S})$ at the previous approximate solution $\mathbf{S}_{l-1}$ can be regularized by the quadratic approximation, the approximate solution $\mathbf{S}_l$ at the current optimization step is expressed as the following objective:

$$\mathbf{S}_l = \arg\min_{\mathbf{S}} \mathfrak{g}(\mathbf{S}_{l-1}) + \langle\nabla_{\mathbf{S}}\mathfrak{g}(\mathbf{S}_{l-1}), \mathbf{S}-\mathbf{S}_{l-1}\rangle + \frac{\xi_l}{2}\|\mathbf{S}-\mathbf{S}_{l-1}\|_F^2, \tag{15}$$

where $\nabla_{\mathbf{S}}\mathfrak{g}(\mathbf{S}_{l-1})$ denotes the first-order gradient of $\mathfrak{g}(\mathbf{S}_{l-1})$. According to the backtracking rule (Beck & Teboulle, 2009), the step size parameter ξ_l can be appropriately determined. Furthermore, the irrelevant constant terms are eliminated, and the regularization term $r(\mathbf{S})$ is added to Eq. (15) to obtain the optimization solution with respect to the variable $\mathbf{S}$:

$$\mathbf{S}_l = \arg\min_{\mathbf{S}} \frac{\xi_l}{2}\|\mathbf{S} - (\mathbf{S}_{l-1} - \frac{1}{\xi_l}\nabla_{\mathbf{S}}\mathfrak{g}(\mathbf{S}_{l-1}))\|_F^2 + \lambda_1\|\mathbf{S}\|_{1,\infty}. \tag{16}$$

Inspired by Liu and Ye (2010), Eq. (16) can be solved by updating each row of $\mathbf{S}$ separately. As the optimization steps described in **Algorithm 1**, when Eq. (16) satisfies the condition of convergence, the solution for $\mathbf{S}^t$ is obtained.

Algorithm 1 Updating $\mathbf{S}^t$ via PALM Strategy (Pock & Sabach, 2016).

Input: $\{\boldsymbol{\theta}_m^t, \mathbf{H}_m^t, \mathbf{D}_m\}_{m=1}^M$, $\lambda_1 > 0$ and MAX-ITER;
Output: $\mathbf{S}^t$;
1: **Initialize**: $\mathbf{S}_0 \in \mathbb{R}^{k\times M}$, $\xi_0 > 0$;
2: **for** $l = 1, \cdots,$ MAX-ITER **do**
3: Settle $\mathbf{S}_l$ by Eq. (16);
4: Update ξ_l by backtracking relu (Beck & Teboulle, 2009);
5: **if** satisfy the convergence condition **then**
6: $\mathbf{S}^t = \mathbf{S}_l$;
7: break;
8: **end if**
9: **end for**
10: **return** $\mathbf{S}^t$;

4.3. Updating $\mathbf{G}_m$ while fixing other variables

The formula to update variable $\mathbf{G}_m$ can be written as:

$$\min_{\mathbf{G}_m} \beta\sum_{m=1}^{M}\|\mathbf{G}_m\|_1 + \Phi(\mathbf{Z}, \mathbf{G}_m - \mathbf{D}_m\mathbf{B}_{\mathbf{Q}}). \tag{17}$$

In this subsection, a soft threshold operator is employed to acquire the solution $\mathbf{G}_m$:

$$\mathbf{G}_m = \text{sign}(\mathbf{D}_m\mathbf{B}_{\mathbf{Q}} - \frac{\mathbf{Z}}{\mu})\max(|\mathbf{D}_m\mathbf{B}_{\mathbf{Q}} - \frac{\mathbf{Z}}{\mu}| - \frac{\beta}{\mu}). \tag{18}$$

4.4. Updating $\mathbf{Q}$ while fixing other variables

To optimize variable $\mathbf{Q}$, the update formula of the complementarity-aware subspace $\mathbf{Q}$ is expressed as follows:

$$\begin{aligned}\min_{\mathbf{Q}}\ &\beta\sum_{m=1}^{M}\sum_{i=1}^{k}\sum_{j=1}^{k}\|\mathbf{D}_m^i - \mathbf{D}_m^j\|_1\mathbf{Q}_{ij} + \lambda_2\|\mathbf{Q}\|_F^2,\\ &\text{s.t.}\quad \mathbf{Q}_i^\top\mathbf{1} = 1,\ \mathbf{Q}_i \geq 0,\ \forall i = 1, \ldots, k.\end{aligned} \tag{19}$$

The optimization of Eq. (19) can be split into a group of smaller independent subproblems:

$$\begin{aligned}&\mathbf{Q}_i = \arg\min_{\mathbf{Q}}\ \|\mathbf{Q}_i + \mathbf{d}_i^{\mathbf{D}_m}\|_2^2,\\ &\text{s.t.}\quad \forall i,\ \mathbf{Q}_i^\top\mathbf{1} = 1,\ \mathbf{Q}_i \geq 0,\ \forall i = 1, \ldots, k,\end{aligned} \tag{20}$$

where $\mathbf{d}_i^{\mathbf{D}_m} \in \mathbb{R}^k$ is a vector, and its jth element can be represented as $\mathbf{d}_{ij}^{\mathbf{D}_m} = \frac{\beta\sum_{m=1}^{M}\|\mathbf{D}_m^i - \mathbf{D}_m^j\|_1}{2\lambda_2}$. Inspired by Nie, Wang, and Huang (2014), a closed-form solution for each $\mathbf{Q}_i$ is obtained as follows:

$$\mathbf{Q}_i = \left(\frac{1 + \sum_{j=1}^{n}\tilde{\mathbf{d}}_{ij}^{\mathbf{D}_m}}{n}\mathbf{1} - \mathbf{d}_i^{\mathbf{D}_m}\right)_+, \tag{21}$$

where $\tilde{\mathbf{d}}_{ij}^{\mathbf{D}_m}$ is an ascending variant of $\mathbf{d}_{ij}^{\mathbf{D}_m}$. Operator $(\mathbf{u})_+$ indicates that all negative elements of $\mathbf{u}$ become zero, leaving the other elements unchanged. The parameter $n = \{1, 2, \ldots, k\}$ is used to dominate the number of nearest neighbours of $\mathbf{D}_m^i$, which may have the opportunity to link to the $\mathbf{D}_m^i$. In addition, the solution of $\mathbf{Q}$ in Eq. (21) is usually unbalanced, and we adopt $\mathbf{Q} = \frac{\mathbf{Q}+\mathbf{Q}^T}{2}$ to achieve balance. In addition, motivated by Nie et al. (2014), the parameter λ_2 can be obtained using the following formula:

$$\lambda_2 = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{n}{2}\mathbf{d}_{i,n+1}^{\mathbf{D}_m} - \frac{1}{2}\sum_{j=1}^{n}\mathbf{d}_{ij}^{\mathbf{D}_m}\right). \tag{22}$$

4.5. Updating $\mathbf{D}_m$ while fixing other variables

The optimization function of $\mathbf{D}_m$ can be expressed as follows:

$$\min_{\mathbf{D}_m} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\sum_{m=1}^{M}\|\boldsymbol{\theta}_m^t - \mathbf{D}_m\mathbf{s}_m^t\|_{\mathbf{H}_m^t}^2 + \Phi(\mathbf{Z}, \mathbf{G}_m - \mathbf{D}_m\mathbf{B}_{\mathbf{Q}}). \tag{23}$$

To update the modality-specific dictionary $\mathbf{D}_m$, the gradient of Eq. (23) is set to zero, and $\mathbf{D}_m$ is calculated. Specifically, the column-wise vectorization of $\mathbf{D}_m$ is updated by the formula $\text{vec}(\mathbf{D}_m) = (\mathbf{R}_m)^{-1}\mathbf{V}_m$, where $\text{vec}(\cdot)$ refers to the operation of column-vectorization, and $\mathbf{R}_m \in \mathbb{R}^{(k\mathbf{D}_m)\times(k\mathbf{D}_m)}$ and $\mathbf{V}_m \in \mathbb{R}^{k\mathbf{D}_m}$ are the statistical records of the mth modality, which are formulated as follows:

$$\begin{aligned}\mathbf{R}_m &= \frac{1}{T}\sum_{t=1}^{T}\left\{(\mathbf{s}_m^t(\mathbf{s}_m^t)^\top)\otimes\mathbf{H}_m^t + \mu(\mathbf{B}_{\mathbf{Q}}\mathbf{B}_{\mathbf{Q}}^\top)\otimes\mathbf{I}_k\right\},\\ \mathbf{V}_m &= \frac{1}{T}\sum_{t=1}^{T}\left\{\mathbf{s}_m^t\otimes(\mathbf{H}_m^t\boldsymbol{\theta}_m^t) + \text{vec}(\mu\mathbf{G}_m\mathbf{B}_{\mathbf{Q}}^\top + \mathbf{Z}\mathbf{B}_{\mathbf{Q}}^\top)\right\},\end{aligned} \tag{24}$$

where $\otimes$ indicates the Kronecker product, and $\mathbf{I}_k \in \mathbb{R}^{k\times k}$ is an identity matrix. Moreover, $\mathbf{R}_m$ and $\mathbf{V}_m$ consecutively increase with the continuous arrival of the new text-audio tasks. In addition, the entire optimization pipeline of the LTASA model is revealed in **Algorithm 2**.

4.6. Updating **Z** while fixing other variables

Furthermore, the multiplier **Z** must also be updated, which can be accomplished simply via the following objective:

$$\mathbf{Z} = \mathbf{Z} + \rho(\mathbf{G}_m - \mathbf{D}_m\mathbf{B}_\mathbf{Q}), \tag{25}$$

where ρ controls the update pace of **Z**.

Algorithm 2 The Optimization Pipeline of Our LTASA Model.

Input: A series of text-audio learning tasks $\{\{\mathbf{X}_m^t\}_{m=1}^M, \mathbf{Y}^t\}_{t=1}^T$, $\mu > 0$, $\mathbf{R}_m = \mathbf{0}_{k\mathbf{D}_m, k\mathbf{D}_m}$, $\mathbf{V}_m = \mathbf{0}_{k\mathbf{D}_m, 1}$;
Output: $\{\mathbf{D}_m\}_{m=1}^M$, $\{\mathbf{S}^t\}_{t=1}^T$;

Initialize: $\{\mathbf{D}_m\}_{m=1}^M$;
for $t = 1, \cdots, T$ **do**
  The t-th new task $\{\mathbf{X}_m^t, \mathbf{Y}^t\}$ arrives;
  for $m = 1, \cdots, M$ **do**
    Compute $\{\boldsymbol{\theta}_m^t, \mathbf{H}_m^t\}$ via Eq. (9);
    Compute $\mathbf{S}^t$ via **Algorithm 1**;
    Update $\mathbf{G}_m$ via Eq. (18);
    Update $\mathbf{Q}$ by optimizing Eq. (20);
    Update $\mathbf{R}_m$: $\mathbf{R}_m \leftarrow \mathbf{R}_m + \left((\mathbf{s}_m^t(\mathbf{s}_m^t)^\top) \otimes \mathbf{H}_m^t + \mu(\mathbf{B}_\mathbf{Q}\mathbf{B}_\mathbf{Q}^\top) \otimes \mathbf{I}_k\right)$;
    Update $\mathbf{V}_m$: $\mathbf{V}_m \leftarrow \mathbf{V}_m + \left(\mathbf{s}_m^t \otimes (\mathbf{H}_m^t\boldsymbol{\theta}_m^t) + vec(\mu\mathbf{G}_m\mathbf{B}_\mathbf{Q}^\top + \mathbf{Z}\mathbf{B}_\mathbf{Q}^\top)\right)$;
    Update $\mathbf{D}_m$: $\mathbf{D}_m \leftarrow (\frac{1}{t}\mathbf{R}_m)^{-1}(\frac{1}{t}\mathbf{V}_m)$;
    Update $\mathbf{Z}$ via Eq. (25);
  end for
end for
return $\{\mathbf{D}_m\}_{m=1}^M$, $\{\mathbf{S}^t\}_{t=1}^T$;

4.7. Complexity analysis

As shown in **Algorithm 2**, the computing complexity of optimizing the LTASA model is mainly composed of four subproblems, the details of which are presented as follows:

- $\{\boldsymbol{\theta}_m^t, \mathbf{H}_m^t\}$ are calculated first by running a single task learner (Ruvolo & Eaton, 2013), which costs $O(\Sigma_{m=1}^M \delta(d_m, n_t))$, where $\delta(\cdot)$ is the complexity of single task learner (Ruvolo & Eaton, 2013), d_m is the dimension, and n_t is the number of data samples.
- The computational complexity of optimizing $\mathbf{S}^t$ is $O(\Sigma_{m=1}^M((d_m)^3 + k(d_m)^2 + kd_m) + k\log k + M\log M)$, where $\boldsymbol{\theta}_m^t$ and $\mathbf{D}_m$ multiplied by $\sqrt{\mathbf{H}_m^t}$ consume $O(\Sigma_{m=1}^M(d_m)^3 + k(d_m)^2)$, the gradient computation of $\mathbf{S}^t$ costs $O(\Sigma_{m=1}^M kd_m)$, and the computational complexity of optimizing Eq. (16) is $O(k\log k + M\log M)$.
- The computational complexity of updating the variables $\mathbf{G}_m$, $\mathbf{Z}$ and $\mathbf{Q}$ totally consumes $O(\Sigma_{m=1}^M(kd_m + d_mk^2))$.
- $\mathbf{D}_m$ is updated via $\text{vec}(\mathbf{D}_m) = (\mathbf{R}_m)^{-1}\mathbf{V}_m$, whose computational complexity is $O(\Sigma_{m=1}^M k^2(d_m)^3)$.

Based on the above description and discussion, the overall computation complexity of the LTASA model is $O(\Sigma_{m=1}^M(\delta(d_m, n_t) + (d_m)^3 + k(d_m)^2 + kd_m + k^2(d_m)^3))$, when a new text-audio sentiment analysis task comes. In addition, there are much fewer bases k in the knowledge dictionary than the feature dimension d_m of each modality and number of samples n_t. Therefore, our proposed LTASA model has high computational efficiency in real-world lifelong text-audio sentiment analysis tasks.

5. Experiments

The experimental setups, implementation details, and comparative evaluation of model performance are presented in this chapter.

5.1. Datasets and evaluation metrics

Two benchmark sentiment analysis datasets are adopted to validate the validity of the proposed model, and one representative public dataset is utilized to illustrate the generalization of our model.

IEMOCAP (Busso et al., 2008) is one of the most widely used datasets in dialogue emotion recognition and was collected by the SAIL Laboratory at the University of Southern California, including 12 h of multi-modality audio-visual data. It has five sessions in total; in each session, one is assigned to one male and one female for dialogue. The content of the dialogue is divided into two parts: a fixed script and free play in a given theme and scenario. This dataset contains 151 conversations with a total of 7433 sentences, where six emotions are marked: neutral, happiness, sadness, anger, frustrated, excited, and non-neutral emotions account for 77%. For each text-audio sentiment analysis learning task, we only consider the distinction between a certain emotion category and all remaining categories by employing the method of logistic regression. Moreover, the emotion categories that any two tasks learn to distinguish are not the same. Hence, the total number of text-audio sentiment analysis tasks in the random task order of this dataset is 6.

MELD (Poria et al., 2018) is also a common dialogue emotion recognition dataset, which contains 1433 conversations with 13,708 accumulated sentences. This dataset is composed of 7 types of emotions: neutral, happiness, surprise, sadness, anger, disgust, and fear. Similar to the IEMOCAP dataset, each task classifies a category and the remaining categories, and the categories distinguished by any two tasks are different. Similarly, the total number of text-audio sentiment analysis tasks for this dataset is 7, and the task order is randomized.

CIFAR-10 (Krizhevsky, Hinton, et al., 2009) is a dataset to recognize real-world objects, which was collected and sorted by Hinton's students Alex Krizhevsky et al. It includes 60,000 RGB colour images of size 32×32, where 50,000 images serve as the training dataset, and 10,000 images are considered the testing dataset to verify the generalization of our model to other lifelong learning tasks. Specifically, the dataset contains ten classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, each of which can be considered a learning task. In each task, two categories (i.e., positive and negative modes) are set with an identical number of samples. Thus, in the CIFAR-10 dataset, the number of learning tasks is 10, and these classification tasks are randomly ordered.

Implementation Details: For the IEMOCAP and MELD datasets, we utilize the pre-trained "bert-base-uncased" (Devlin, Chang, Lee, & Toutanova, 2018) and "facebook/wav2vec2-base-960h" (Baevski, Zhou, Mohamed, & Auli, 2020) models to extract the "[CLS]" representations and mean pooling representations as text features and audio features, respectively. For the CIFAR-10 dataset, we utilize Resnet-18 (He, Zhang, Ren, & Sun, 2016) to extract the 512-D features of RGB images and the 256-D features of grayscale images. As introduced in the traditional multimodal method (Gao, Lian, Wang, & Sun, 2020), RGB and grayscale images are considered different modalities, and there is a large distribution gap between these two modalities. Thus, in this paper, we follow the multimodal data processing strategy proposed in Gao et al. (2020) to obtain the RGB and gray modalities of the CIFAR-10 benchmark dataset. For each dataset, we randomly select 10%, 20%, and 30% of the samples from each task as training sets for three different experiments and treat the corresponding remaining samples as their respective test sets. All learning tasks are continuously forwarded to our proposed LTASA model without any human prior. The evaluation results in this paper are the average of 3 random runs.

Table 1
Performance comparisons between the proposed model and other baseline approaches in terms of 5 indicators (mean ± standard deviation) (%) on the IEMOCAP dataset, where 10%, 20% and 30% represent the training set.

Metric	Train	MTFL (Gong et al., 2012)	MFM (Lu et al., 2017)	ELLA (Ruvolo & Eaton, 2013)	lslMTMV (Li et al., 2017)	rLM^{2}L (Sun et al., 2018)	L^2HMT (Liu et al., 2019)	LVTL (Dong, Cong et al., 2022)	MSCNN (Cai et al., 2016)	Ours
Auc	10%	79.21 ± 0.53	80.16 ± 0.22	82.87 ± 0.51	84.40 ± 0.17	83.15 ± 0.63	83.94 ± 0.58	84.67 ± 0.83	81.33 ± 0.64	**86.24 ± 0.54**
	20%	78.85 ± 0.27	79.06 ± 0.74	82.13 ± 0.38	83.36 ± 0.23	83.47 ± 0.74	84.17 ± 1.02	84.83 ± 0.59	82.53 ± 0.17	**86.94 ± 0.41**
	30%	78.06 ± 0.31	78.95 ± 0.14	80.67 ± 0.42	82.49 ± 0.72	81.96 ± 0.52	81.62 ± 0.65	82.03 ± 0.38	81.19 ± 0.84	**83.76 ± 0.38**
Macro-F1	10%	61.85 ± 0.23	62.28 ± 0.19	63.17 ± 0.58	64.79 ± 0.37	64.33 ± 0.62	64.01 ± 0.63	64.25 ± 0.47	63.85 ± 0.39	**66.79 ± 0.59**
	20%	62.53 ± 0.81	61.19 ± 0.41	64.01 ± 0.52	62.53 ± 0.93	64.28 ± 0.73	64.73 ± 0.58	65.02 ± 0.57	64.73 ± 0.53	**70.50 ± 0.32**
	30%	59.83 ± 0.42	60.14 ± 0.72	61.75 ± 0.48	62.36 ± 0.59	62.84 ± 0.82	63.05 ± 0.74	63.87 ± 0.68	63.36 ± 0.26	**66.67 ± 0.45**
Micro-F1	10%	61.27 ± 0.31	61.85 ± 0.46	63.83 ± 0.48	64.57 ± 0.18	64.83 ± 0.63	65.02 ± 0.91	65.31 ± 0.36	63.96 ± 0.31	**66.57 ± 0.83**
	20%	60.18 ± 0.51	59.26 ± 0.23	61.83 ± 0.83	60.45 ± 0.39	62.07 ± 0.62	62.83 ± 0.59	63.18 ± 0.71	61.29 ± 0.18	**70.34 ± 0.51**
	30%	61.03 ± 0.48	59.73 ± 0.46	58.04 ± 0.66	59.87 ± 0.72	61.13 ± 0.49	60.55 ± 0.82	62.29 ± 0.73	63.05 ± 0.53	**66.29 ± 0.87**
Acc	10%	76.15 ± 0.37	77.84 ± 0.62	79.74 ± 0.57	81.15 ± 0.85	81.03 ± 0.69	80.74 ± 0.80	81.33 ± 0.62	80.74 ± 0.43	**82.30 ± 0.63**
	20%	75.83 ± 0.16	76.74 ± 0.28	78.63 ± 0.29	78.30 ± 0.37	78.92 ± 0.81	79.16 ± 0.47	81.14 ± 0.74	79.65 ± 0.28	**83.70 ± 0.49**
	30%	77.16 ± 0.41	78.06 ± 0.15	79.15 ± 0.44	78.98 ± 0.84	79.13 ± 0.71	77.52 ± 0.63	80.03 ± 0.68	79.85 ± 0.95	**81.50 ± 0.51**
Recall	10%	63.38 ± 0.72	64.48 ± 0.29	66.52 ± 0.72	68.41 ± 0.84	67.33 ± 0.92	68.02 ± 0.49	67.84 ± 0.83	65.77 ± 0.28	**68.77 ± 0.26**
	20%	57.95 ± 0.28	59.18 ± 0.71	62.59 ± 0.82	63.87 ± 0.71	64.48 ± 0.73	64.92 ± 0.36	66.14 ± 0.83	66.85 ± 0.56	**73.34 ± 0.45**
	30%	59.72 ± 0.25	60.17 ± 0.30	61.83 ± 0.33	61.70 ± 0.19	62.11 ± 0.60	62.64 ± 0.82	64.11 ± 0.53	62.08 ± 0.27	**67.67 ± 0.27**

evaluation Metrics: To effectively illustrate the capability of the LTASA model, five general evaluation indicators are employed: Accuracy (Acc) (Sokolova, Japkowicz, & Szpakowicz, 2006), Area Under Curve (Auc) (Ling, Huang, & Zhang, 2003), Macro-F1 (Opitz & Burst, 2019), Micro-F1 (Opitz & Burst, 2019) and Recall (Sokolova et al., 2006).

5.2. Baseline comparison methods

The effectiveness of our LTASA model is demonstrated by comparing it with some of the most advanced learning methods, which are mainly divided into four categories: a multi-task learning model, a multi-view multi-task learning model, five lifelong learning approaches, and a conventional sentiment analysis learning model.

- Multi-Task Learning Model: **MTFL** (Gong, Ye, & Zhang, 2012) mainly explores the common feature set among related tasks, where the characteristics of distinct modalities for each task are concatenated into a high-dimensional vector to consider the robustness to anomalous tasks. However, instead of lifelong learning settings, it assumes that all training data are simultaneously available.
- Multi-View Multi-Task Method: **MFM** (Lu, He, Shao, Cao, & Yu, 2017) addresses the issues of multi-task and multi-view by learning task-specific features and exploring the information shared by distinct views in various tasks, but it also simultaneously accesses all training data.
- Lifelong Learning Methods: The **ELLA** (Ruvolo & Eaton, 2013) algorithm combines the characteristics of distinct modalities into a high-dimensional space and exploits it to explore the sparse shared base of all learning tasks. **lslMTMV** (Li et al., 2017) maps multiple views into a shared potential space and learns the decision function among multiple views by transferring knowledge from the learned tasks to new arriving tasks. **rLM^{2}L** (Sun et al., 2018), which has strong robustness, can seek out views specific to a task from continuous multi-view tasks. **L^{2}HMT** (Liu, Sun, & Fang, 2019) continuously learns the intrinsic representations among heterogeneous modalities by constructing an effective online dictionary and constantly updates the learning classifier for each multi-modality task. **LVTL** (Dong, Cong et al., 2022) constructs a modality invariant space based on sparse constraints to acquire the inherent relevances within intra-modality.
- Sentiment Analysis Learning Method: **MSCNN** (Cai, Fan, Feris, & Vasconcelos, 2016) is a multi-scale convolutional neural network that we use to learn text-audio sentiment analysis tasks. It assumes that all training data can be accessed at once.

5.3. Comparison experiments

By presenting the experimental results of our LTASA model and some comparison algorithms, as shown in Tables 1–3, we prove the efficiency and generalization of our model to address lifelong text-audio sentiment analysis tasks. Compared with the multi-task learning model MTFL (Gong et al., 2012), our model can mine the complementary knowledge shared between text and audio modalities, significantly improving the evaluation performance of lifelong multi-modality sentiment analysis tasks. Moreover, the proposed model performs better than the multi-view multi-task method MFM (Lu et al., 2017) by a large margin, since our model can consecutively learn new text-audio sentiment analysis tasks by integrating learning experiences into a modality-specific shared knowledge dictionary. Compared with lifelong learning methods (Dong, Cong et al., 2022; Li et al., 2017; Liu et al., 2019; Ruvolo & Eaton, 2013; Sun et al., 2018), the proposed model significantly outperforms by approximately 2% in five indicators on the benchmark datasets, which fully confirms the validity of our LTASA model. The comparison results in Tables 1, 2, and 3 show that our model establishes a modality-specific knowledge dictionary for each modality to learn the shared experiences of different tasks from an intra-modality aspect and explores complementary knowledge between text and audio modalities from an inter-modality aspect. The comparison experiments between our model and the convolutional neural network-based sentiment analysis learning model (i.e., MSCNN Cai et al., 2016) also illustrate that our method has continuous learning capability to consecutively learn new text-audio sentiment analysis tasks.

5.4. Comparison experiments of computational time

This section compares the efficiency of the LTASA model with other baseline approaches in terms of running time, as introduced in Table 4. Although our model has comparable computational performance (i.e., running time) with MTFL (Gong et al., 2012) and lslMTMV (Li et al., 2017), it significantly outperforms these two comparison methods in terms of 5 evaluation indicators on benchmark datasets. Furthermore, the proposed model has a shorter computational time than other baseline comparison methods, demonstrating our model's efficiency and real-time ability. Compared with the convolutional neural network-based sentiment analysis learning model, i.e., MSCNN (Cai et al., 2016), our model has better evaluation performance and computational

Table 2
Performance comparisons between the proposed model and other baseline approaches in terms of 5 indicators (mean ± standard deviation) (%) on the MELD dataset, where 10%, 20% and 30% represent the training set.

Metric	Train	MTFL (Gong et al., 2012)	MFM (Lu et al., 2017)	ELLA (Ruvolo & Eaton, 2013)	lslMTMV (Li et al., 2017)	rLM^2L (Sun et al., 2018)	L^2HMT (Liu et al., 2019)	LVTL (Dong, Cong et al., 2022)	MSCNN (Cai et al., 2016)	Ours
Auc	10%	74.16 ± 0.52	75.18 ± 0.63	77.85 ± 0.41	77.27 ± 0.29	78.64 ± 0.16	78.07 ± 0.42	78.94 ± 0.32	78.16 ± 0.33	**80.98 ± 0.52**
	20%	76.19 ± 0.52	75.86 ± 0.24	78.14 ± 0.33	77.58 ± 0.53	77.91 ± 0.36	78.17 ± 0.26	78.61 ± 0.26	77.63 ± 0.28	**80.15 ± 0.31**
	30%	71.85 ± 0.63	72.05 ± 0.73	74.18 ± 0.62	75.02 ± 0.15	75.83 ± 0.39	76.22 ± 0.40	75.85 ± 0.37	75.73 ± 0.73	**77.71 ± 0.28**
Macro-F1	10%	55.17 ± 0.28	55.84 ± 0.25	57.13 ± 0.62	57.55 ± 0.85	58.13 ± 0.73	58.01 ± 0.62	58.92 ± 0.34	58.16 ± 0.73	**61.28 ± 0.62**
	20%	54.18 ± 0.73	53.86 ± 0.63	55.73 ± 0.63	55.11 ± 0.35	56.83 ± 0.31	56.17 ± 0.21	57.06 ± 0.53	57.18 ± 0.62	**60.63 ± 0.53**
	30%	51.17 ± 0.74	52.06 ± 0.42	53.73 ± 0.28	52.74 ± 0.62	53.16 ± 0.84	54.58 ± 0.62	54.91 ± 0.70	54.85 ± 0.75	**56.74 ± 0.73**
Micro-F1	10%	53.31 ± 0.29	53.88 ± 0.62	55.19 ± 0.39	56.53 ± 0.73	57.14 ± 0.73	56.38 ± 0.42	57.74 ± 0.51	57.85 ± 0.51	**60.86 ± 0.62**
	20%	54.09 ± 0.83	53.75 ± 0.41	55.19 ± 0.38	54.46 ± 0.71	55.73 ± 0.83	56.11 ± 0.73	56.72 ± 0.83	55.08 ± 0.26	**60.22 ± 0.51**
	30%	50.85 ± 0.39	51.71 ± 0.42	52.83 ± 0.74	52.04 ± 0.83	53.16 ± 0.73	54.33 ± 0.81	53.69 ± 0.18	53.48 ± 0.64	**56.63 ± 0.52**
Acc	10%	72.84 ± 0.63	73.27 ± 0.51	75.85 ± 0.62	76.19 ± 0.72	76.83 ± 0.42	75.42 ± 0.48	76.49 ± 0.18	76.89 ± 0.41	**78.79 ± 0.47**
	20%	73.84 ± 0.18	74.25 ± 0.53	75.27 ± 0.64	75.68 ± 0.47	74.19 ± 0.73	75.94 ± 0.59	76.17 ± 0.40	75.37 ± 0.88	**78.83 ± 0.84**
	30%	71.95 ± 0.64	71.34 ± 0.63	72.26 ± 0.19	73.23 ± 0.34	73.39 ± 0.65	74.21 ± 0.30	73.93 ± 0.58	73.04 ± 0.29	**76.60 ± 0.41**
Recall	10%	58.83 ± 0.75	59.03 ± 0.48	61.73 ± 0.82	60.68 ± 0.37	61.15 ± 0.31	61.58 ± 0.92	62.04 ± 0.62	61.85 ± 0.73	**65.01 ± 0.19**
	20%	55.62 ± 0.31	56.06 ± 0.42	57.73 ± 0.19	58.25 ± 0.27	58.67 ± 0.33	56.69 ± 0.51	60.17 ± 0.24	59.16 ± 0.26	**62.54 ± 0.41**
	30%	55.03 ± 0.53	55.86 ± 0.74	57.92 ± 0.28	57.70 ± 0.42	58.25 ± 0.33	57.33 ± 0.21	58.73 ± 0.48	58.03 ± 0.22	**60.35 ± 0.64**

Table 3
Performance comparisons between the proposed model and other baseline approaches in terms of 5 indicators (mean ± standard deviation) (%) on the CIFAR-10 dataset, where 10%, 20% and 30% represent the training set.

Metric	Train	MTFL (Gong et al., 2012)	MFM (Lu et al., 2017)	ELLA (Ruvolo & Eaton, 2013)	lslMTMV (Li et al., 2017)	rLM^2L (Sun et al., 2018)	L^2HMT (Liu et al., 2019)	LVTL (Dong, Cong et al., 2022)	MSCNN (Cai et al., 2016)	Ours
Auc	10%	98.52 ± 0.17	97.95 ± 0.42	98.94 ± 0.25	99.84 ± 0.33	99.48 ± 0.21	99.65 ± 0.31	**99.89 ± 0.17**	99.84 ± 0.17	99.68 ± 0.27
	20%	97.85 ± 0.33	97.28 ± 0.25	99.31 ± 0.42	99.93 ± 0.19	99.84 ± 0.25	**99.97 ± 0.14**	98.85 ± 0.35	99.65 ± 0.19	99.59 ± 0.23
	30%	97.63 ± 0.16	97.84 ± 0.62	99.73 ± 0.18	99.96 ± 0.42	98.94 ± 0.51	98.45 ± 0.26	99.75 ± 0.13	99.48 ± 0.73	**99.99 ± 0.31**
Macro-F1	10%	90.84 ± 0.34	92.13 ± 0.25	93.79 ± 0.43	94.09 ± 0.39	95.17 ± 0.35	95.84 ± 0.27	96.85 ± 0.70	96.33 ± 0.27	**99.14 ± 0.45**
	20%	93.18 ± 0.36	94.01 ± 0.37	94.48 ± 0.75	93.91 ± 0.18	95.10 ± 0.36	95.84 ± 0.75	96.43 ± 0.51	95.19 ± 0.42	**99.01 ± 0.20**
	30%	94.28 ± 0.53	93.87 ± 0.83	96.73 ± 0.42	95.15 ± 0.64	95.74 ± 0.31	97.01 ± 0.66	96.74 ± 0.45	97.32 ± 0.25	**99.13 ± 0.28**
Micro-F1	10%	90.04 ± 0.42	90.85 ± 0.42	92.18 ± 0.44	93.42 ± 0.52	94.63 ± 0.74	95.19 ± 0.28	96.87 ± 0.64	94.85 ± 0.23	**99.14 ± 0.35**
	20%	91.74 ± 0.28	92.19 ± 0.37	95.17 ± 0.62	93.35 ± 0.15	94.73 ± 0.28	94.96 ± 0.43	95.68 ± 0.73	95.75 ± 0.41	**99.04 ± 0.16**
	30%	92.33 ± 0.53	92.85 ± 0.19	95.38 ± 0.73	94.63 ± 0.41	95.16 ± 0.28	96.41 ± 0.30	95.74 ± 0.71	96.94 ± 0.27	**99.13 ± 0.53**
Acc	10%	94.06 ± 0.41	94.83 ± 0.72	97.14 ± 0.84	96.75 ± 0.53	95.82 ± 0.55	97.65 ± 0.28	97.84 ± 0.35	97.04 ± 0.43	**99.52 ± 0.61**
	20%	93.45 ± 0.28	93.88 ± 0.24	96.41 ± 0.64	96.73 ± 0.29	97.31 ± 0.38	97.64 ± 0.22	98.15 ± 0.41	97.05 ± 0.33	**99.45 ± 0.83**
	30%	94.18 ± 0.53	93.36 ± 0.15	96.13 ± 0.52	97.39 ± 0.64	97.75 ± 0.62	98.13 ± 0.24	97.59 ± 0.28	96.88 ± 0.67	**99.52 ± 0.82**
Recall	10%	87.95 ± 0.39	89.04 ± 0.18	91.43 ± 0.58	89.89 ± 0.38	92.74 ± 0.63	93.38 ± 0.64	95.12 ± 0.52	94.84 ± 0.16	**98.58 ± 0.71**
	20%	88.25 ± 0.26	88.96 ± 0.73	90.13 ± 0.55	88.94 ± 0.45	92.14 ± 0.35	90.87 ± 0.26	93.36 ± 0.48	95.17 ± 0.25	**98.29 ± 0.73**
	30%	87.18 ± 0.53	89.17 ± 0.22	91.62 ± 0.32	91.13 ± 0.13	92.16 ± 0.28	93.34 ± 0.18	94.19 ± 0.52	94.86 ± 0.31	**98.39 ± 0.52**

Table 4
Comparison of running time between the proposed model and other baseline approaches, where s and h denote the second and hour, respectively.

Datasets	MTFL (Gong et al., 2012)	MFM (Lu et al., 2017)	ELLA (Ruvolo & Eaton, 2013)	lslMTMV (Li et al., 2017)	rLM^2L (Sun et al., 2018)	L^2HMT (Liu et al., 2019)	LVTL (Dong, Cong et al., 2022)	MSCNN (Cai et al., 2016)	Ours
IEMOCAP	1.79 ± 0.34 s	20.03 ± 1.22 s	18.74 ± 1.14 s	3.16 ± 0.97 s	6.83 ± 1.31 s	9.85 ± 0.95 s	6.50 ± 1.13 s	8.14 ± 0.13 h	6.16 ± 0.86 s
MELD	2.03 ± 1.14 s	19.17 ± 0.79 s	18.19 ± 0.24 s	3.91 ± 1.14 s	6.04 ± 0.85 s	9.17 ± 0.95 s	8.55 ± 0.73 s	7.86 ± 0.17 h	5.63 ± 0.18 s
CIFAR-10	2.96 ± 0.86 s	31.64 ± 0.57 s	29.38 ± 1.31 s	10.89 ± 1.03 s	18.17 ± 1.04 s	24.65 ± 0.86 s	13.63 ± 0.77 s	10.35 ± 0.09 h	12.04 ± 0.74 s
Average	2.26 ± 0.78 s	23.61 ± 0.86 s	22.10 ± 0.90 s	5.99 ± 1.05 s	10.35 ± 1.07 s	14.56 ± 0.92 s	9.56 ± 0.88 s	8.78 ± 0.13 h	7.94 ± 0.59 s

time because it can construct a modality-specific knowledge dictionary for each modality to learn the shared experiences of different tasks from the intra-modality aspect, and it can capture complementary knowledge between text and audio modalities from the inter-modality aspect. Considering the computational time, our model can be better applied to address lifelong text-audio sentiment analysis tasks in real-world applications.

5.5. Investigation of the number of learned tasks

This section investigates the effect of the number of learning tasks on the function of the LTASA model with the continuous arrival of new text-audio sentiment analysis tasks. As Fig. 2 displays, when there are more text-audio sentiment analysis tasks, the performance of our model on public benchmark datasets converges to a steady level on five evaluation metrics. Although the evaluation curves for the CIFAR-10 dataset fluctuate in the learning process, they tend to converge in the final phase. Complicated semantic relationships across different tasks in each modality may cause such fluctuation. However, the proposed modality-specific knowledge dictionary in this paper can efficiently capture each modality's knowledge and experiences under lifelong text-audio sentiment analysis learning, which effectively tackles such fluctuations and encourages our proposed model to converge.

5.6. Analysis of the size of the modality-specific dictionary

As Fig. 3 displays, we consider the impact of the size parameter k of the modality-specific knowledge dictionary on the capability of the LTASA model. Obviously, when k is set as 7, our model converges to a stable performance on the IEMOCAP,

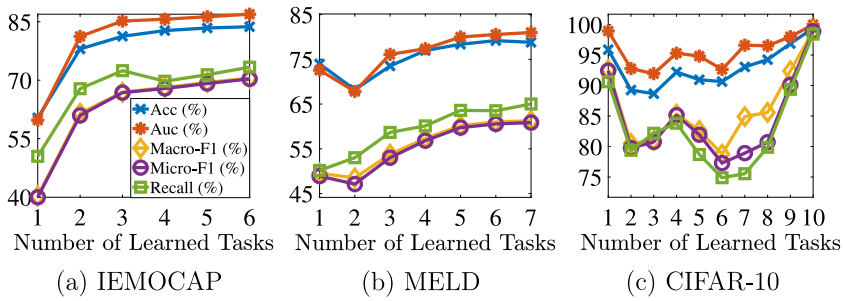

Fig. 2. Impact of the number of learned text-audio tasks on the (a) IEMOCAP, (b) MELD, and (c) CIFAR-10 datasets.

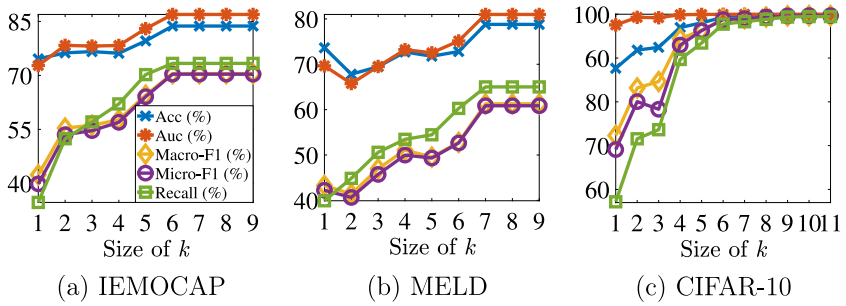

Fig. 3. Influence of parameter k of the modality-specific knowledge dictionary size on the (a) IEMOCAP, (b) MELD, and (c) CIFAR-10 datasets..

Table 5
Comparison of different choices of kernels for K_m^1 and K_m^2 when sampling 20% samples as training set, where I, L, and G represent the inner product kernel, linear kernel, and Gaussian kernel, respectively.

$K_m^1 + K_m^2$	Metric	I + I	L + L	G + G	I + G	L + G	Ours (I + L)
IEMOCAP	Auc	86.43 ± 0.27	84.37 ± 0.61	85.26 ± 0.40	85.82 ± 0.29	86.04 ± 0.25	**86.94 ± 0.41**
	Micro-F1	69.61 ± 0.16	68.75 ± 0.31	69.04 ± 0.28	68.29 ± 0.33	68.27 ± 0.51	**70.34 ± 0.51**
	Acc	83.04 ± 0.25	82.14 ± 0.24	82.74 ± 0.22	82.73 ± 0.30	83.14 ± 0.14	**83.70 ± 0.49**
MELD	Auc	79.41 ± 0.28	78.36 ± 0.25	78.71 ± 0.38	79.02 ± 0.24	78.85 ± 0.33	**80.15 ± 0.31**
	Micro-F1	58.73 ± 0.36	58.02 ± 0.15	58.35 ± 0.17	59.42 ± 0.22	59.11 ± 0.28	**60.22 ± 0.51**
	Acc	78.12 ± 0.33	77.65 ± 0.26	78.31 ± 0.48	77.95 ± 0.14	78.03 ± 0.32	**78.83 ± 0.84**

MELD, and CIFAR-10 datasets. When k increases at the beginning, the model can learn more shared specific knowledge for different modalities (i.e., text and audio modalities) and make the text and audio modalities more distinctive. However, when k exceeds the corresponding optimal value, the modality-specific knowledge dictionary introduces more modality information while introducing much noise, which heavily degrades the performance of the proposed model. More importantly, Fig. 3 shows that the LTASA model can capture potential complementary knowledge and simultaneously ignore the irrelevant representations in intra-modality.

5.7. Investigation of the kernel K_m^1 and K_m^2

We compare the effects of different choices of K_m^1 and K_m^2 kernels, as shown in Table 5. Specifically, we conduct experiments using different combinations of the Gaussian kernel (denoted as G), inner product kernel (denoted as I), and linear kernel (denoted as L). Compared with other integration methods, our model achieves the best experimental results on the IEMOCAP and MELD datasets in terms of three metrics, which validates the effectiveness of our model. This result also shows that this cross-integration of the inner product kernel and the linear kernel can better explore the complementary modality information from different Hilbert spaces.

5.8. Analysis of the hyperparameters λ_1 and λ_2

In this paper, λ_1 and λ_2 denote the regularization coefficients of the sparse matrix in intra-modality and the complementarity-aware subspace across modalities, respectively. In this section, we change the values of λ_1 and λ_2 to observe their experimental effects on three public benchmark datasets while fixing other coefficients, e.g., k and β. We set the range of hyperparameters λ_1 and λ_2 as $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ in our experiments. According to the three-dimensional bar charts in Figs. 4, 5, 6, 7, and 8, we find that the optimal values of λ_1 and λ_2 have little difference for different datasets. For example, on the IEMOCAP dataset, the evaluation results are best when λ_1 and λ_2 are both set to 10^{-3}, while on the MELD dataset, the evaluation results are relatively high when λ_1 and λ_2 are both set to 10^{-4}. For the CIFAR-10 dataset, the performance of our LTASA model improves when both λ_1 and λ_2 are set in the range of $\{10^{-4}, 10^{-3}, 10^{-2}\}$. Under these hyperparameter settings, our model can stably and efficiently learn the shared knowledge in intra-modality and the complementary knowledge across modalities. However, when both λ_1 and λ_2 are large, Auc and Acc are relatively low on all three datasets. Thus, in this case, our model is relatively weak in learning a shared knowledge dictionary within each modality and a semantically consistent space across different modalities.

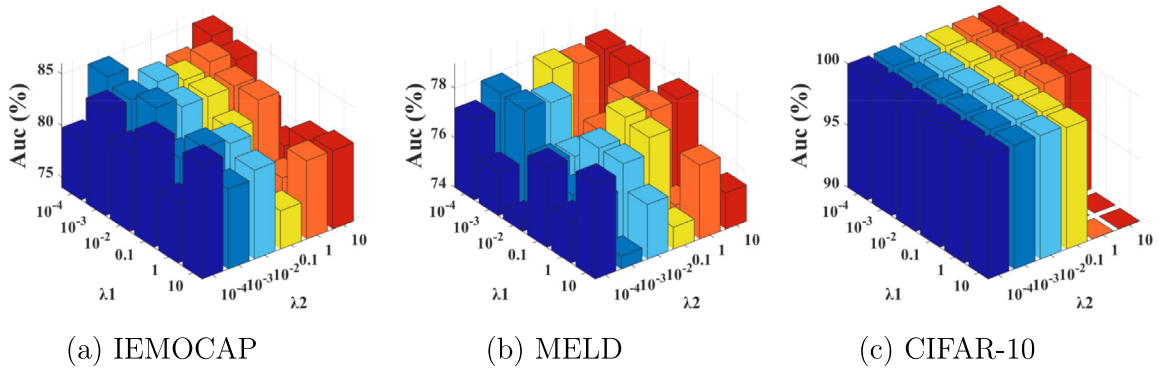

Fig. 4. Investigation of λ_1 and λ_2 on three datasets in terms of the **Auc** evaluation metric.

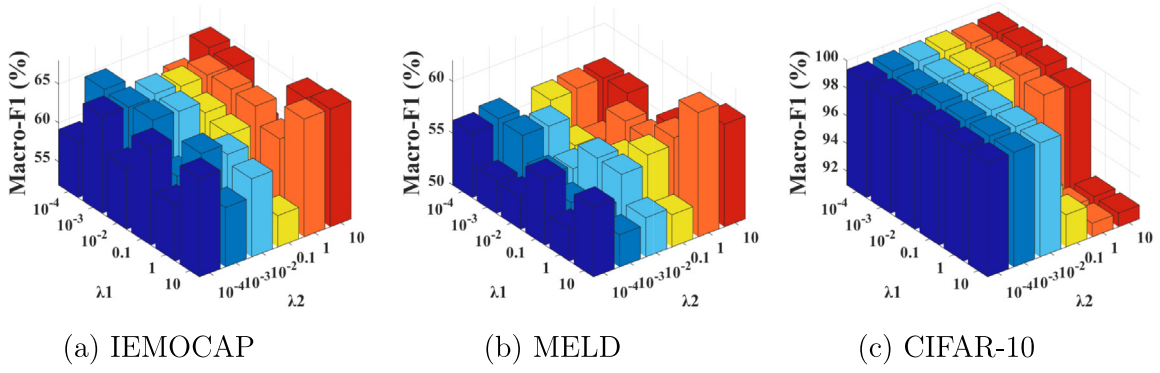

Fig. 5. Investigation of λ_1 and λ_2 on three datasets in terms of the **Macro-F1** evaluation metric.

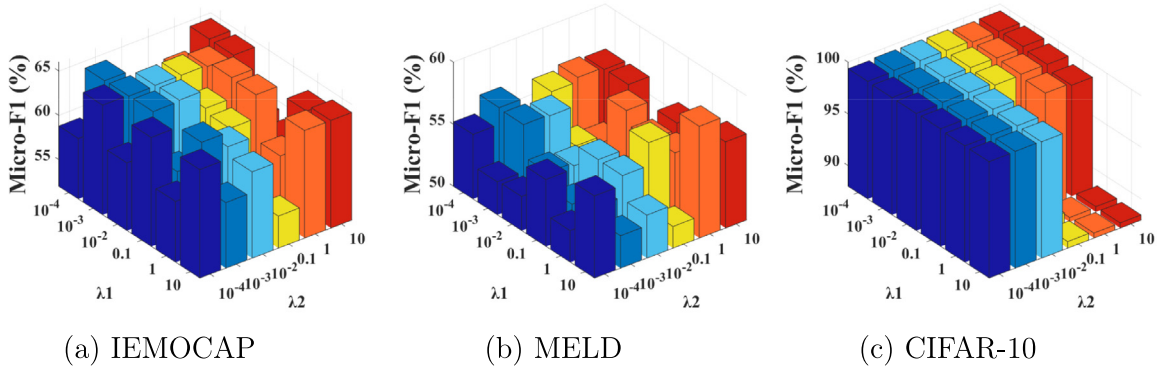

Fig. 6. Investigation of λ_1 and λ_2 on three datasets in terms of the **Micro-F1** evaluation metric.

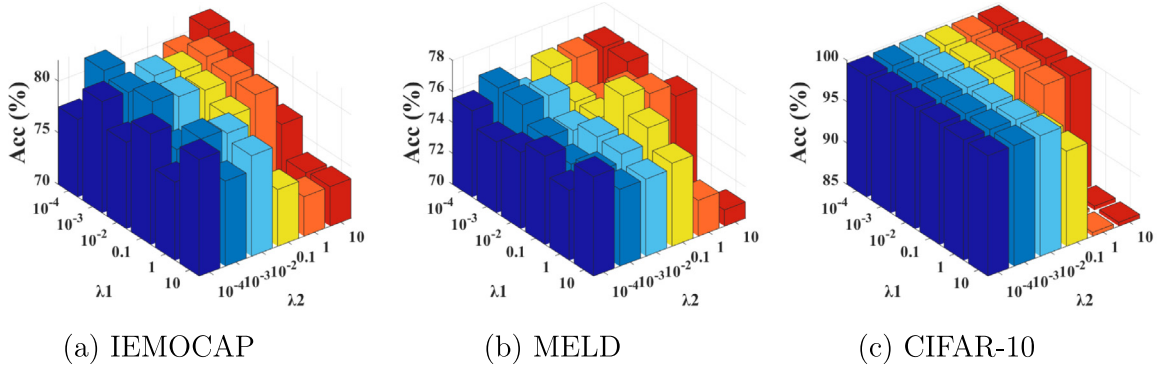

Fig. 7. Investigation of λ_1 and λ_2 on three datasets in terms of the **Acc** evaluation metric.

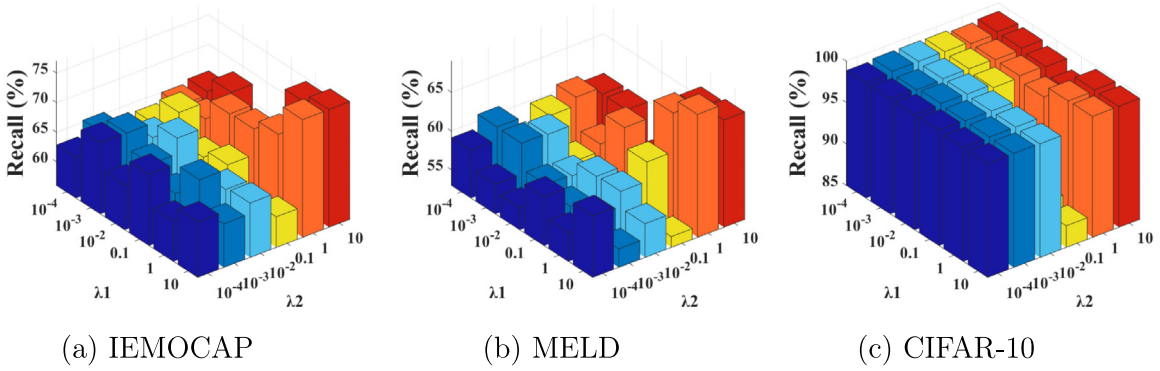

Fig. 8. Investigation of λ_1 and λ_2 on three datasets in terms of the **Recall** evaluation metric.

6. Conclusion and future work

In this work, a novel LTASA model has been proposed to continuously learn text-audio sentiment analysis tasks. Specifically, we design a modality-specific knowledge dictionary for each modality, which captures common experiences within intra-modality from a series of text-audio sentiment analysis tasks. Meanwhile, to explore complementary knowledge between text and audio modalities, we construct a complementarity-aware subspace to explore nonlinear inter-modality complementary information by incorporating a modality semantic consistency space under the HSIC. This complementarity-aware subspace encourages the semantic information of the text and audio modalities to better interact with each other and consequently improves the performance of our lifelong sentiment analysis model. The effectiveness of our model is verified on two commonly used dialogue sentiment analysis datasets and a general image classification dataset. Compared with some baseline representative methods, the capability of the LTASA model is significantly boosted in terms of five measurement indicators. In summary, the proposed LTASA learning model can explore shared knowledge among different tasks within each modality and complementary knowledge between text and audio modalities, which can be successfully applied to speech emotion recognition, dialogue emotion analysis, etc. In future research, the proposed model can be extended to address lifelong text-audio sentiment analysis tasks with missing modality information in real-world applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix. Taylor expansion

Assuming that the first term in Eq. (8) is defined as $\mathcal{L}_m^t(\mathbf{D}_m\mathbf{s}_m^t)$, the Taylor expansion is as follows:

$$\begin{aligned}\mathcal{L}_m^t(\mathbf{D}_m\mathbf{s}_m^t) &= \frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}(f_m^t(\mathbf{x}_{im}^t, \mathbf{D}_m\mathbf{s}_m^t), y_i^t) \\ &= \mathcal{L}_m^t(\boldsymbol{\theta}_m^t) + \langle\nabla_{\boldsymbol{\theta}_m^t}\mathcal{L}(\mathbf{D}_m\mathbf{s}_m^t), \mathbf{D}_m\mathbf{s}_m^t - \boldsymbol{\theta}_m^t\rangle \\ &\quad + \frac{1}{2}\|\mathbf{D}_m\mathbf{s}_m^t - \boldsymbol{\theta}_m^t\|_{\mathbf{H}_m^t}^2,\end{aligned} \tag{A.1}$$

where $\nabla_{\boldsymbol{\theta}_m^t}\mathcal{L}(\mathbf{D}_m\mathbf{s}_m^t)$ is the first-order gradient about $\boldsymbol{\theta}_m^t$, and $\mathbf{H}_m^t$ is the Hessian matrix of $\mathcal{L}(\mathbf{D}_m\mathbf{s}_m^t)$ surrounding around $\boldsymbol{\theta}_m^t$.

References

Alías, F., Socoró, J. C., & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences*, *6*(5), 143.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In *2009 IEEE international conference on acoustics, speech and signal processing* (pp. 693–696). IEEE.

Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., et al. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters*, *158*, 80–86.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, Article 107134.

Boyd, S., Parikh, N., & Chu, E. (2011). *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359.

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision* (pp. 354–370). Springer.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis* (pp. 1–10). Springer.

Cambria, E., Howard, N., Hsu, J., & Hussain, A. (2013). Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In *2013 IEEE symposium on computational intelligence for human-like intelligence* (pp. 108–117). IEEE.

Cao, X., Zhang, C., Fu, H., Liu, S., & Zhang, H. (2015). Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–594).

Chen, Z., Ma, N., & Liu, B. (2018). Lifelong learning for sentiment classification. arXiv preprint arXiv:1801.02808.

Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, 1–18.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, J., Cong, Y., Sun, G., Fang, Z., & Ding, Z. (2021). Where and how to transfer: knowledge aggregation-induced transferability perception for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dong, J., Cong, Y., Sun, G., & Zhang, T. (2022). Lifelong robotic visual-tactile perception learning. *Pattern Recognition*, *121*, Article 108176.

Dong, J., Cong, Y., Sun, G., Zhong, B., & Xu, X. (2020). What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4023–4032).

Dong, J., Wang, L., Fang, Z., Sun, G., Xu, S., Wang, X., et al. (2022). Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10164–10173).

Du, Y., Liu, Y., Peng, Z., & Jin, X. (2022). Gated attention fusion network for multimodal sentiment classification. *Knowledge-Based Systems*, *240*, Article 108107.

Gai, S., Chen, Z., & Wang, D. (2021). Multi-modal meta continual learning. In *2021 international joint conference on neural networks* (pp. 1–8). IEEE.

Gao, Q., Lian, H., Wang, Q., & Sun, G. (2020). Cross-modal subspace clustering via deep canonical correlation analysis. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 34* (pp. 3938–3945).

Gkoumas, D., Li, Q., Yu, Y., & Song, D. (2021). An entanglement-driven fusion neural network for video sentiment analysis. In *Proceedings of the thirti-eth international joint conference on artificial intelligence* (pp. 1736–1742). International Joint Conferences on Artificial Intelligence Organization.

Gong, P., Ye, J., & Zhang, C. (2012). Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 895–903).

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory* (pp. 63–77). Springer.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., & Zimmermann, R. (2018). Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. association for computational linguistics. North American chapter. Meeting, Vol. 2018* (p. 2122). NIH Public Access.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Irfan, B., Ramachandran, A., Spaulding, S., Kalkan, S., Parisi, G. I., & Gunes, H. (2021). Lifelong learning and personalization in long-term human-robot interaction (LEAP-HRI). In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 724–727).

Isele, D., Rostami, M., & Eaton, E. (2016). Using task features for zero-shot knowledge transfer in lifelong learning. In *Ijcai, Vol. 16* (pp. 1620–1626).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Person re-identification by unsupervised ℓ_1 graph learning. In *European conference on computer vision* (pp. 178–195). Springer.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Li, X., Chandrasekaran, S. N., & Huan, J. (2017). Lifelong multi-task multi-view learning using latent spaces. In *2017 IEEE international conference on big data (Big Data)* (pp. 37–46). IEEE.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 2935–2947.

Li, Z., Li, Y., Xiong, W., Chen, M., & Li, Y. (2021). Research on voiceprint recognition technology based on deep neural network. In *Proceedings of the 2021 international conference on bioinformatics and intelligent computing* (pp. 412–417).

Li, R., Zhang, C., Hu, Q., Zhu, P., & Wang, Z. (2019). Flexible multi-view representation learning for subspace clustering. In *IJCAI* (pp. 2916–2922).

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: a better measure than accuracy in comparing learning algorithms. In *Conference of the canadian society for computational studies of intelligence* (pp. 329–341). Springer.

Liu, H., Sun, F., & Fang, B. (2019). Lifelong learning for heterogeneous multi-modal tasks. In *2019 international conference on robotics and automation* (pp. 6158–6164). IEEE.

Liu, J., & Ye, J. (2010). Efficient L1/Lq norm regularization. arXiv preprint arXiv:1009.4766.

Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, *30*.

Lu, C.-T., He, L., Shao, W., Cao, B., & Yu, P. S. (2017). Multilinear factorization machines for multi-task multi-view learning. In *Proceedings of the tenth ACM international conference on web search and data mining* (pp. 701–709).

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 6818–6825).

Nie, F., Wang, X., & Huang, H. (2014). Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 977–986).

Opitz, J., & Burst, S. (2019). Macro f1 and macro f1. arXiv Preprint arXiv:1911.03347.

Ouzar, Y., Bousefsaf, F., Djeldjli, D., & Maaoui, C. (2022). Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2460–2469).

Peng, Z., Lu, Y., Pan, S., & Liu, Y. (2021). Efficient speech emotion recognition using multi-scale CNN and attention. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing* (pp. 3020–3024). IEEE.

Phan, H. T., Nguyen, N. T., & Hwang, D. (2022). Convolutional attention neural network over graph structures for improving the performance of aspect-level sentiment analysis. *Information Sciences*, *589*, 416–439.

Pock, T., & Sabach, S. (2016). Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, *9*(4), 1756–1787.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining* (pp. 439–448). IEEE.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508.

Rannen, A., Aljundi, R., Blaschko, M. B., & Tuytelaars, T. (2017). Encoder based lifelong learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 1320–1328).

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). Icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2001–2010).

Ruvolo, P., & Eaton, E. (2013). ELLA: An efficient lifelong learning algorithm. In *International conference on machine learning* (pp. 507–515). PMLR.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015–1021). Springer.

Stappen, L., Schumann, L., Sertolli, B., Baird, A., Weigell, B., Cambria, E., et al. (2021). Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox. In *Proceedings of the 2nd on multimodal sentiment analysis challenge* (pp. 75–82).

Sun, G., Cong, Y., Li, J., & Fu, Y. (2018). Robust lifelong multi-task multi-view representation learning. In *2018 IEEE international conference on big knowledge* (pp. 91–98). IEEE.

Wang, X., Lei, Z., Guo, X., Zhang, C., Shi, H., & Li, S. Z. (2019). Multi-view subspace clustering with intactness-aware similarity. *Pattern Recognition*, *88*, 50–63.

Wu, C.-H., & Liang, W.-B. (2010). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, *2*(1), 10–21.

Xia, R., Jiang, J., & He, H. (2017). Distantly supervised lifelong learning for large-scale social media sentiment analysis. *IEEE Transactions on Affective Computing*, *8*(4), 480–491.

Xue, X., Zhang, C., Niu, Z., & Wu, X. (2022). Multi-level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*.

Yang, B., Wu, L., Zhu, J., Shao, B., Lin, X., & Liu, T.-Y. (2022). Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yang, K., Xu, H., & Gao, K. (2020). Cm-bert: Cross-modal bert for text-audio sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 521–528).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 2236–2246).

Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L.-P. (2018). Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 32*.

Zhang, C., Fu, H., Liu, S., Liu, G., & Cao, X. (2015). Low-rank tensor constrained multiview subspace clustering. In *Proceedings of the IEEE international conference on computer vision* (pp. 1582–1590).

Zhang, K., Li, Y., Wang, J., Cambria, E., & Li, X. (2021). Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1034–1047.

Zhao, Y., Cao, X., Lin, J., Yu, D., & Cao, X. (2021). Multimodal affective states recognition based on multiscale cnns and biologically inspired decision fusion model. *IEEE Transactions on Affective Computing*.