Journal of Information Systems Engineering and Management
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Adversarial Multimodal Sentiment Analysis with Cross-Modal and Cross-Domain Alignment

Vani Golagana[1], Prof. S. Viziananda Row[2], Prof. P. Srinivasa Rao[3]
[1,2,3] *Department of Computer Science & Systems Engineering, AUCE, Andhra University, Visakhapatnam, AP, India.*
[1] *Corresponding author: Email: vani.srr22@gmail.com*

ARTICLE INFO

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Multimodal sentiment analysis faces significant challenges due to data scarcity and domain shift, which hinder model generalization across different datasets. To address these issues, we propose a Multimodal Domain Adaptation framework that combines advanced feature extraction techniques, attention-based fusion, and adversarial domain adaptation to enhance sentiment classification. Our approach builds on Domain-Adversarial Neural Networks (DANN) to facilitate knowledge transfer from a labeled source dataset to an unlabeled target dataset, reducing domain discrepancies while preserving sentiment prediction accuracy. Unlike existing methods that focus primarily on single-modality adaptation or basic feature alignment, our framework performs comprehensive cross-modal and cross-domain feature alignment to improve generalization. Specifically, we extract high-quality feature embeddings for both text and image modalities using state-of-the-art deep learning models. To bridge modal gaps, we integrate an attention-based fusion mechanism that prioritizes the most informative modality, ensuring optimal feature integration. Additionally, we employ a Gradient Reversal Layer (GRL) and a domain discriminator to reduce domain loss, enabling the model to learn domain-invariant representations. Our experimental results demonstrate that Adversarial Multimodal Sentiment Adaptation with Cross-Modal and Cross-Domain Alignment (AMSA-CMCDA) significantly enhance performance in sentiment classification across datasets. By effectively addressing both modality mismatch and domain shift, our approach proves its effectiveness in real-world sentiment analysis applications.

Keywords: Multimodal Sentiment Analysis, Domain Adaptation, Feature Extraction, Attention-based Fusion, Domain-Adversarial Neural Networks (DANN), Cross-Modal Alignment, Cross-Domain Alignment, Gradient Reversal Layer (GRL), Domain-Invariant Representations, Sentiment Classification.

INTRODUCTION

Multimodal Sentiment Analysis (MSA) has emerged as a pivotal research area within artificial intelligence, aiming to integrate information from multiple modalities—such as text, images, and audio—to enhance sentiment classification accuracy [1]. Unlike unimodal sentiment analysis, MSA leverages the complementary strengths of each modality, providing a deeper understanding of human emotions and opinions. However, despite its potential, MSA faces significant challenges, particularly the scarcity of labeled multimodal datasets and the impact of domain shift. One of the primary challenges in MSA is the lack of large-scale annotated multimodal datasets. Annotating data across multiple modalities is a complex, labor-intensive process that requires domain expertise to ensure accuracy. Consequently, the limited availability of high-quality labeled data restricts the performance and generalization of MSA models. Moreover, MSA models often suffer from domain adaptation issues, where models trained on one dataset fail to generalize effectively across different datasets due to variations in linguistic style, image content, or cultural context. Another key challenge arises from missing or incomplete modalities during inference. Some datasets may lack textual descriptions for images, while others may have textual information but no corresponding visual data. To address this issue, Golagana et al. [2] introduced an attention-based fusion approach, which assigns sentiment-aware attention weights to text and image features before fusing them for classification. By leveraging attention-based fusion, this method enhances robustness, ensuring accurate sentiment prediction even when one modality is incomplete or missing. A fundamental difficulty in sentiment prediction across different domains is the intrinsic variation in data distributions. Traditional supervised learning models assume that training and testing data share the same distribution and follow the independent and identically distributed (i.i.d.) condition. However, in real-world scenarios, this assumption often fails, leading to domain shift and suboptimal model performance.

Adversarial discriminative domain adaptation methods have emerged as effective solutions to mitigate domain shift. These approaches aim to bridge the gap between the source domain (where labeled data is available) and the target domain (where labeled data is scarce or unavailable) by ensuring that the learned feature representations are well-aligned [3]. By reducing distributional discrepancies, domain adaptation techniques enable models to generalize across different domains without requiring extensive labeled data for every new dataset.

Domain Adaptation (DA) is a specialized approach within Transfer Learning (TL) [4] that has gained increasing importance in sentiment analysis. While early DA methods primarily focused on learning domain-invariant features, recent research highlights the need to preserve both domain-specific and invariant features to improve adaptability. Fully unsupervised domain adaptation remains a complex challenge, requiring models that can automatically extract meaningful representations even when no labeled data is available in the target domain [5].

A straightforward but ineffective approach to cross-domain sentiment analysis is to apply a model trained on one domain directly to another without adaptation. This overlooks domain-specific differences in sentiment expressions, often leading to poor performance. To overcome this, researchers have explored various domain adaptation techniques. One approach is feature extraction-based adaptation, which identifies transferable features across domains. Another is pivot-based learning, where common linguistic or visual features serve as bridges between domains [6], [7].

Existing domain adaptation methods aim to minimize distributional differences between source and target domains while preserving domain-specific discriminative structures. They achieve this through techniques like Maximum Mean Discrepancy (MMD) and discriminative distance measures [8]. In practice, the accuracy of a supervised learning model does not always transfer well to datasets that were not part of the training. This occurs because the assumption that source and target domain data share the same distribution is often violated, which arises from differences in data distributions between source and target domains. Early research on Domain Adaptation (DA) was primarily focused on traditional machine learning techniques. However, with the rise of Deep Learning (DL), the focus has shifted toward DA in DL-based methods. The introduction of Generative Adversarial Networks (GANs), Attention Mechanisms, and Transformer-based architectures has significantly advanced DA techniques in deep learning [9]. Domain Adaptation (DA) aims to address the performance degradation of Deep Neural Networks (DNNs) when applied to different datasets (source vs. target domains). This survey explores the significance of DA in practical applications, real-world scenarios, and industrial domains. While numerous DA methods have been developed for specific data domains, no comprehensive review covers datasets, methods, applications, and challenges across all domains.To fill this gap, the paper presents a comprehensive taxonomy of DA approaches across five major areas: computer vision, natural language processing, speech, time-series, and multimodal data. It delves into the available datasets, key challenges, and practical industrial use cases, demonstrating how DA can significantly improve model performance and adaptability. It examines available datasets, challenges, and industrial use cases, highlighting how DA enhances model efficiency. The survey also outlines future research directions, focusing on evolving DA methods, datasets, and industrial implementations [10].

Multimodal DA introduces additional complexity by requiring alignment across both domains and modalities. While prior works have explored multimodal adaptation for tasks like Visual Question Answering (VQA) [11], [12], [13], challenges remain in effectively fusing multimodal data under domain shift. To address these challenges, this work proposes a domain-adapted multimodal sentiment classification framework that combines advanced feature extraction, attention-based fusion, and adversarial domain adaptation.

Our primary objective is to predict sentiment labels for target data by identifying domain-invariant features using a domain discriminator. The learning process jointly optimizes classification and domain losses to ensure effective knowledge transfer from the source to the target domain. In this work, we present a domain-adapted multimodal sentiment classification framework that effectively tackles the challenge of domain shift. Our approach leverages feature representation learning, attention-based fusion, and adversarial domain adaptation to enhance model generalization across different domains.

This paper is structured as follows: Section II reviews related work on multimodal sentiment analysis and domain adaptation, emphasizing prior approaches to handling modality mismatch and domain shift. Section III details our proposed Multimodal Domain Adaptation framework, including feature extraction, attention-based fusion, and adversarial domain adaptation techniques. Section IV presents experimental results, evaluating the effectiveness of our approach in improving sentiment classification across datasets. Finally, Section V concludes the paper with key findings and future research directions, particularly in enhancing cross-modal and cross-domain adaptation for real-world sentiment analysis applications.

II. RELATED WORK

Domain Adaptation (DA) aims to enhance model generalization across different domains, particularly in sentiment analysis, where data distributions vary significantly. Domain adaptation for text sentiment analysis focuses on minimizing discrepancies between the source and target domains to improve sentiment classification accuracy. Tzeng et al. [14] introduced a method to learn domain-invariant features using an adaptation layer and domain confusion loss, minimizing differences between source and target domains with metrics such as Maximum Mean Discrepancy (MMD). Similarly, Peddinti et al. [15] proposed a two-step domain adaptation framework for Twitter sentiment analysis, involving feature alignment techniques like MMD and adversarial training to improve generalization. Unlabeled Twitter data was leveraged through semi-supervised learning and self-training. To enhance domain adaptability in deep learning models, Du et al. [16] integrated adversarial learning with BERT, making it both domain-aware through post-training tasks and domain-invariant via adversarial adaptation. This approach led to improved sentiment classification even when labeled target data was unavailable.

In sentiment analysis, image-based domain adaptation methods tackle variations in image distributions across domains. Hassan et al. [17] proposed an approach that combines generative models with self-ensembling. A modified CycleGAN was employed to generate diverse source-domain variations, aligning them with the target domain to enhance domain-invariant learning. Additionally, a teacher-student self-ensembling framework was introduced to stabilize training, yielding superior results on benchmark datasets.

Multimodal domain adaptation aims to align text and image modalities while mitigating domain shifts. Meegahapola et al. [18] explored distribution shifts in multimodal mobile sensing and introduced M3BAT, a multi-branch adaptation method employing Domain Adversarial Neural Networks (DANN) to address domain discrepancies. While their work focused on multimodal

sensor data, similar strategies are applicable to sentiment analysis.
A deep residual deformable subdomain adaptation framework was introduced by Liang et al. [19] for cross-domain fault diagnosis. A residual network was enhanced with a deformable convolution module to improve feature extraction and transferability. Unlike traditional DANN-based methods, adversarial training was replaced with local maximum mean discrepancy (LMMD), leading to more efficient distribution alignment. Although primarily designed for industrial applications, its methodology provides insights into multimodal DA techniques.

For sentiment analysis across multimodal domains, the importance of graph-based adaptation methods has also been explored. Liang et al. [20] introduced SSAGCN (Semi-Supervised Subdomain Adaptation Graph Convolutional Network), integrating Graph Convolutional Networks (GCN) with Semi-Supervised Adaptation (SSA) to improve feature extraction and reduce domain discrepancies. This method demonstrated high accuracy and efficiency in transferring knowledge between domains, making it relevant for MSA.

A robust unsupervised domain adaptation framework was proposed by Ganin et al.[21] utilizing a domain classifier and a gradient reversal layer (GRL) to ensure that deep learning models learn both discriminative and domain-invariant features. By reversing gradients during backpropagation, the method effectively minimized domain shifts, improving performance in unseen target domains.The framework's adaptability to various feed-forward architectures made it a scalable solution for leveraging unlabeled target domain data while maintaining high classification performance. This review highlights key domain adaptation techniques across text, image, and multimodal sentiment analysis. While prior research has tackled individual modalities using MMD, adversarial training, and self-ensembling, multimodal adaptation remains a challenging frontier. Our work builds specifically on adversarial training using Domain-Adversarial Neural Networks (DANN), applying this approach to both text and image modalities. By doing so, we aim to enhance cross-domain generalization in multimodal sentiment analysis.

III. METHODOLOGY

This approach builds on Domain-Adversarial Neural Networks (DANN) to learn domain-invariant representations, reducing both cross-modal and cross-domain discrepancies. It focuses on transferring knowledge from a labeled source domain to an unlabeled target domain, where both domains include multiple modalities such as text and images. The framework involves extracting meaningful features, fusing multimodal data effectively, and applying adversarial training to align the domains. The overall workflow is illustrated in Figure 1, and a detailed step-by-step process is provided in Algorithm 1.

A. Feature Extraction

The feature extraction process transforms raw textual and visual data into meaningful representations for sentiment classification. For text, we utilize a transformer-based model such as Bidirectional Encoder Representations from Transformers (BERT), which tokenizes input text, generates contextualized embeddings from the last hidden layer, and applies a linear projection to obtain compact feature representations.

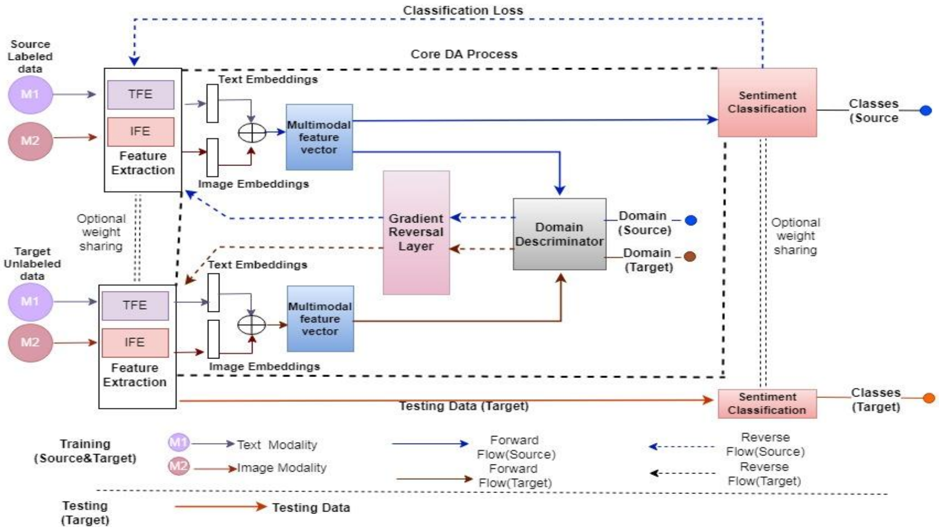


Fig. 1: Proposed Workflow of AMSA-CMCDA for Sentiment Label Prediction

For images, a deep convolutional neural network (CNN), specifically ResNet50, is employed to extract hierarchical visual features. The images are passed through ResNet50's pre-trained layers, with the final classification layer removed, and feature vectors are obtained from the global average pooling (GAP) layer [22]. To ensure effective alignment between textual and visual representations, the extracted feature vectors are further transformed using a fully connected (FC) layer. By leveraging BERT for text and ResNet50 for images, our framework ensures that both semantic and visual information is effectively captured, enabling robust multimodal sentiment analysis. The processed features are then integrated through an attention-based fusion mechanism for optimal multimodal sentiment analysis.

B. Attention-Based Fusion

Fusion plays a critical role in multimodal domain adaptation, requiring the integration of complex, high-level interactions between different modalities to extract meaningful features for recognition and retrieval tasks. In multimodal sentiment analysis, effective integration of textual and visual features is crucial for improving classification performance.

Traditional fusion methods often treat both modalities equally, which may not always be optimal. To address this, we employ weight-based attention fusion [2], which dynamically assigns importance to each modality based on its contribution to sentiment prediction. Attention weights are computed by evaluating the significance of each modality's feature representation. The modality contributing more to sentiment prediction is assigned a higher weight, ensuring that the fusion process prioritizes the most informative features. This adaptive weighting improves robustness, particularly in cases where one modality is more expressive than the other. The final fused representation is computed as follows:

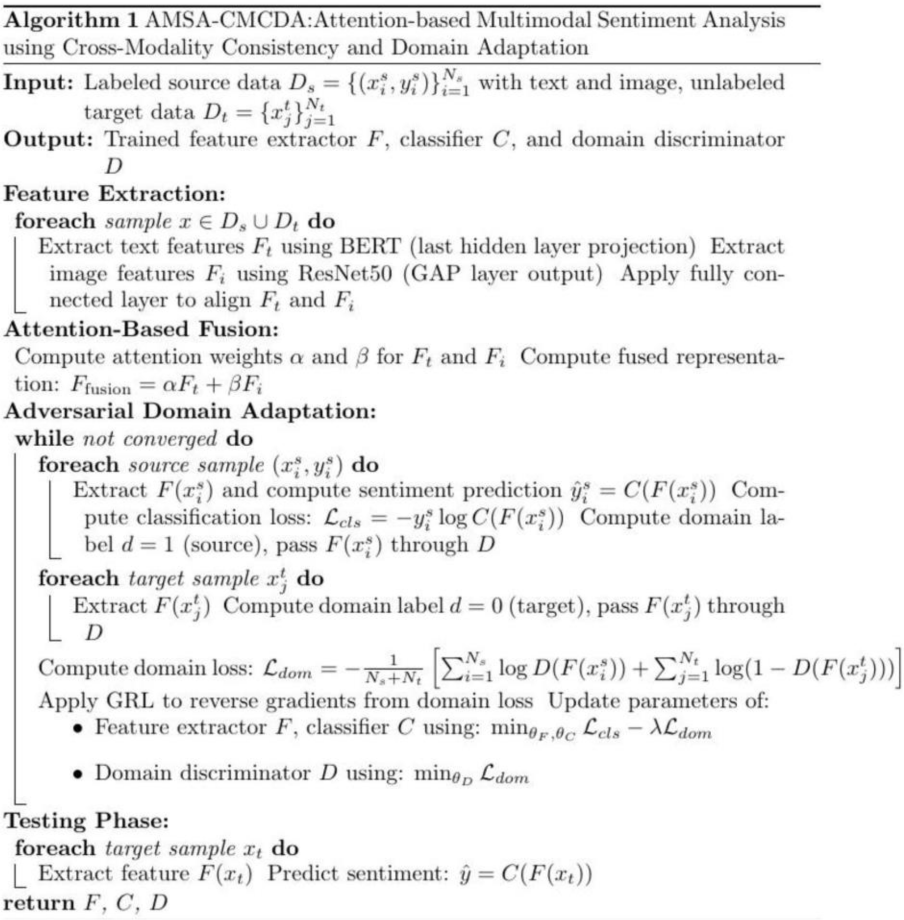
$$F_{\text{fusion}} = \alpha F_t + \beta F_i \qquad (1)$$

where F_t and F_i represent text and image features, respectively. The attention weights α and β dynamically adjust to control the contributions of text and image modalities based on their relevance to sentiment prediction.

These fused representations are then passed to the adversarial domain adaptation framework for further processing.

C. Adversarial Domain Adaptation Framework

The input data undergoes feature extraction and fusion, integrating both text and image modalities. The fused features are then used for classification and domain discrimination. The adversarial training framework, based on Domain-Adversarial Neural Networks (DANN), aims to learn domain-invariant representations by introducing an adversarial loss



Algorithm 1 AMSA-CMCDA:Attention-based Multimodal Sentiment Analysis using Cross-Modality Consistency and Domain Adaptation

```
Input: Labeled source data $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ with text and image, unlabeled
       target data $D_t = \{x_j^t\}_{j=1}^{N_t}$
Output: Trained feature extractor $F$, classifier $C$, and domain discriminator
        $D$
Feature Extraction:
 foreach sample $x \in D_s \cup D_t$ do
     Extract text features $F_t$ using BERT (last hidden layer projection)  Extract
     image features $F_i$ using ResNet50 (GAP layer output)  Apply fully connected layer to align $F_t$ and $F_i$
Attention-Based Fusion:
 Compute attention weights $\alpha$ and $\beta$ for $F_t$ and $F_i$  Compute fused representation: $F_{\text{fusion}} = \alpha F_t + \beta F_i$
Adversarial Domain Adaptation:
 while not converged do
     foreach source sample $(x_i^s, y_i^s)$ do
         Extract $F(x_i^s)$ and compute sentiment prediction $\hat{y}_i^s = C(F(x_i^s))$  Compute classification loss: $\mathcal{L}_{cls} = -y_i^s \log C(F(x_i^s))$  Compute domain label $d = 1$ (source), pass $F(x_i^s)$ through $D$
     foreach target sample $x_j^t$ do
         Extract $F(x_j^t)$  Compute domain label $d = 0$ (target), pass $F(x_j^t)$ through $D$
     Compute domain loss: $\mathcal{L}_{dom} = -\frac{1}{N_s+N_t}\left[\sum_{i=1}^{N_s} \log D(F(x_i^s)) + \sum_{j=1}^{N_t} \log(1 - D(F(x_j^t)))\right]$
     Apply GRL to reverse gradients from domain loss  Update parameters of:
         • Feature extractor $F$, classifier $C$ using: $\min_{\theta_F,\theta_C} \mathcal{L}_{cls} - \lambda\mathcal{L}_{dom}$
         • Domain discriminator $D$ using: $\min_{\theta_D} \mathcal{L}_{dom}$
Testing Phase:
 foreach target sample $x_t$ do
     Extract feature $F(x_t)$  Predict sentiment: $\hat{y} = C(F(x_t))$
return $F$, $C$, $D$
```

This loss encourages the model to extract features that are indistinguishable across source and target domains. The framework consists of three key components:

- **Domain Discriminator (D)**: Determines whether a feature comes from the source or target domain, helping the model learn shared representations.
- **Gradient Reversal Layer (GRL)**: Enables adversarial learning by flipping the gradient during backpropagation, making it harder for the domain discriminator to differentiate between domains.

To enforce domain invariance, we introduce the Gradient Reversal Layer (GRL), which inverts the gradient during backpropagation. The domain classification loss is defined as:

$$L_d = -\mathbb{E}[y_d \log D + (1 - y_d)\log(1 - D)] \qquad (2)$$

where y_d represents the true domain label (1 for source, 0 for target).

During backpropagation, GRL modifies the gradient update for the feature extractor as:

$$\frac{\partial L_d}{\partial F} = -\lambda \frac{\partial L_d}{\partial D} \qquad (3)$$

where λ is a hyperparameter that controls the strength of the adversarial loss. This forces the feature extractor to learn domain-invariant representations, improving generalization to the target domain.

D. Adversarial Objective for Domain-Invariant Features

The adversarial objective balances task-specific learning and domain adaptation by incorporating two opposing losses. The task-specific loss (e.g., classification loss) minimizes errors on the labeled source domain, ensuring effective learning. Conversely, the domain loss maximizes confusion between source and target domains, encouraging domain-independent feature learning. The total loss function is defined as:

$$L_{\text{total}} = L_{\text{task}} - \lambda L_{\text{domain}} \qquad (4)$$

where λ controls the trade-off between task-specific learning and domain adaptation.

The categorical cross-entropy loss is used for both sentiment classification and domain discrimination, effectively guiding the model to minimize prediction errors.

1) Sentiment Classification Loss: For labeled source data, the classifier $\boldsymbol{C}$ minimizes the cross-entropy loss:

$$\mathcal{L}_{cls} = -\frac{1}{N_s}\sum_{i=1}^{N_s} y_i^s \log C(F(x_i^s)) \qquad (5)$$

where, N_s is the number of labeled source samples, x_i^s is the i-th source sample, y_i^s is the corresponding ground truth sentiment label, $F(x_i^s)$ represents the extracted feature of x_i^s, and $C(F(x_i^s))$ is the predicted sentiment probability distribution by the classifier C.

2) Domain Discrimination Loss: The domain discriminator $\boldsymbol{D}$ classifies features as source ($\boldsymbol{d} = \mathbf{1}$) or target ($\boldsymbol{d} = \mathbf{0}$), using binary cross-entropy:

$$\mathcal{L}_{dom} = -\frac{1}{N_s + N_t}\sum_{i=1}^{N_s} \log D(F(x_i^s)) - \sum_{j=1}^{N_t} \log(1 - D(F(x_j^t))) \qquad (6)$$

where N_t is the number of unlabeled target samples.

3) Adversarial Training with Gradient Reversal Layer (GRL): To achieve domain invariance, we formulate an adversarial training objective where the feature extractor $\boldsymbol{F}$ is trained to minimize classification loss while maximizing domain confusion. This is achieved using a Gradient Reversal Layer (GRL) with the following optimization objectives:

$$\theta_F^*, \theta_C^* = \underset{\theta_F, \theta_C}{\operatorname{argmin}} \mathcal{L}_{cls} - \lambda \mathcal{L}_{dom} \qquad (7)$$

$$\theta_D^* = \underset{\theta_D}{\operatorname{argmin}} \mathcal{L}_{dom} \qquad (8)$$

where, θ_F, θ_C, and θ_D are the parameters of the feature extractor, classifier, and domain discriminator respectively.

$\mathcal{L}_{cls}$ is the classification loss for sentiment prediction, $\mathcal{L}_{dom}$ is the domain classification loss and λ is a trade-off parameter that controls the adversarial effect.

E. Testing Strategy

During inference, the trained feature extractor F and classifier C are used to predict sentiment labels for target data. The learned domain-invariant representations ensure adaptation to the target domain, despite the absence of labeled target data.

The final classification decision is obtained as:

$$\hat{y} = C(F(x_t)) \qquad (9)$$

where x_t is a target sample, and $\hat{y}$ is the predicted sentiment label.
By iteratively optimizing these objectives, our model learns domain-invariant feature representations, ensuring effective adaptation across different domains. Since the feature space is shared between source and target data, the classifier generalizes well to the target domain using only the source-labeled supervision.

IV. RESULTS

In this section, we present the results of our experiments, highlighting the effectiveness of the proposed domain adaptation approach for multimodal sentiment analysis. We begin with a description of the datasets used for training and evaluation, followed by the experimentation process.

A. DataSets

In this study, we utilize two multimodal datasets: MVSA-S (Multi-View Sentiment Analysis-Single) and the Memotion dataset, both consisting of image-text pairs annotated with sentiment labels. The MVSA-S dataset contains 4,869 image-text pairs collected from Twitter, each labeled for sentiment by a single annotator. To ensure consistency and comparability with prior studies, we focus on pairs where both the text and image exhibit the same sentiment polarity. Using the filtering and preprocessing strategies outlined in [23], [24], we refine the dataset to retain only sentiment-consistent samples. The Memotion dataset originally [25] comprises 6,992 meme images, annotated with both sentiment polarity (positive, negative, neutral) and emotion-related categories, including humor, sarcasm, offensiveness, and motivation. We apply similar preprocessing and filtering techniques to obtain a balanced and representative subset from this dataset for our experiments.

B. Experimentation

To evaluate the effectiveness of proposed adversarial domain adaptation approach, we conducted experiments using the MVSA-S dataset as the labeled source domain and the Memotion dataset as the unlabeled target domain. The model incorporates BERT for textual feature extraction and ResNet50 for visual feature extraction, followed by a fusion strategy. The fused features from both domains are processed through a domain discriminator to reduce domain shift, while a classifier, trained on the source domain, handles sentiment prediction. A Gradient Reversal Layer (GRL) is used to encourage domain-invariant learning.

To systematically evaluate the impact of domain adaptation on multimodal sentiment classification, we designed three experimental configurations using the MVSA-S dataset as the labeled source domain and the Memotion dataset as the unlabeled target domain. The configurations and corresponding results are summarized in Table I.

- Source (Labeled) Only: In this baseline setting, the model was trained solely on the MVSA-S dataset without incorporating any domain adaptation strategy. We evaluated the performance using three input modalities: text-only (BERT+CNN), image-only (ResNet50+CNN), and a fused text+image representation (BERT+ResNet50+CNN). The text-only model achieved 54% accuracy, while the image-only model attained 49%, indicating that textual features contributed more significantly to sentiment prediction. The fused modality without adaptation resulted in a slightly higher accuracy of 58%, showing modest benefit from multimodal fusion alone.

- Source (Labeled) + Target (Unlabeled) with Domain Adaptation (DANN): To address domain shift, we introduced Domain-Adversarial Neural Networks (DANN) to align source and target domain distributions. When DANN was applied to the text-only modality, performance improved to 56%. For the image-only modality, DANN yielded a slight increase to 53%, demonstrating its potential even in unimodal settings.

- Source + Target with Domain Adaptation using Fused Modality (AMSA-CMCDA): In the final and most effective setup, we applied our proposed Adversarial Multimodal Sentiment Adaptation with Cross-Modal and Cross-Domain Alignment (AMSA-CMCDA) method. This involved fusing both text and image features and applying DANN to the combined representation. This approach achieved the highest accuracy of 66.36% on the Memotion dataset, clearly demonstrating the effectiveness of combining multimodal features with adversarial domain adaptation.

Table I. Performance evaluation of domain adaptation strategies across modalities

Input Domain	**Input Modality**	**Methodology**	**MVSA-S->Memotion (Accuracy %)**
Source (Labeled) Only	Text	BERT+CNN	54%
	Image	ResNet50+CNN	49%
	Text+Image(Fused)	BERT+ResNet50+CNN	58%
Source (Labeled) + Target (Unlabeled)	Text	DANN	56%
	Image	DANN	53%
	Text+Image(Fused)	**AMSA-CMCDA**	**66.36%**

We trained the model over 20 epochs and observed the behavior of classification and domain losses throughout the training process. During the initial epochs, particularly around Epoch 3, there was a brief spike in total loss, indicating some early instability. Despite this, the classification loss started at a relatively high value and gradually decreased as training continued. Both classification and domain losses consistently declined over subsequent epochs, eventually stabilizing at lower values by the final epoch. This steady convergence suggests that the model was learning effectively and becoming increasingly stable over time.

Table II presents examples from the Memotion dataset, showcasing actual sentiment labels alongside predictions made by our model. These results demonstrate the effectiveness of adversarial domain adaptation in transferring knowledge from the source domain to the target domain. The domain discriminator successfully reduced discrepancies between source and target feature distributions, leading to improved generalization in sentiment classification.

Image	**Caption**	**Actual**	**Predicted**
LOOK THERE MY FRIEND NOW ALL SOHAL IKUT TREND PLAY THE 10YEARS CHALLENGE AT FACEBOOK	Look there my friend Lightyear, now all Sohalikut trend play the 10 years challenge at Facebook imgflip.com	Positive	Positive

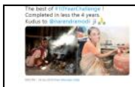
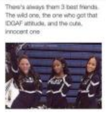
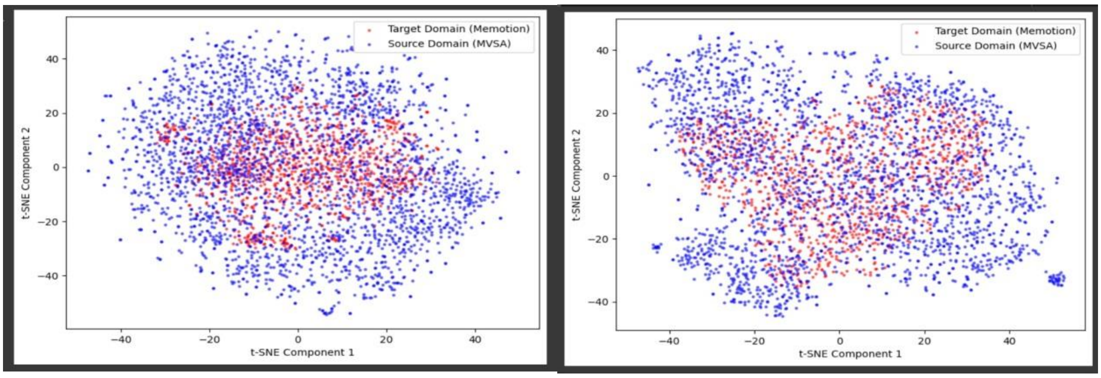
	The best of #10YearChallenge! Completed in less than 4 years. Kudos to @narendramodi ji 8:05 PM - 16 Jan 2019 from Mumbai, India	Positive	Positive
	I don't know why this dude is named Mr. Beans because he doesn't look like he got none	Neutral	Neutral
	Obama: Don't discuss Titanic with Joe. DiCaprio: Why? Obama: He's still upset. He thinks you could've fit on that door—and I don't disagree	Neutral	Neutral
	The 'distracted boyfriend' meme presidential version imgflip.com	Negative	Negative
	Gizmodo @Gizmodo Charlie Chaplin invented the "distracted boyfriend" meme back in 1922	Neutral	Negative
	WE "LOST" hn cells when they invade your body. i am not gonna copy myself. Actually uses thet cells to copy itself and kills thet-cell it copied off of. Doctor Evil laugh	Neutral	Positive
	WE "LOST" THE E-MAILS. memegenerator.net We "LOST" The E-mails. | Dr. Evil quote	Negative	Negative
	There's always them 3 best friends. The wild one, the one who got that IDGAF attitude, and the cute innocent one	Positive	Positive
	Wants to ban violent video games, but doesn't try to ban violent wars in the Middle East quickmeme.com	Positive	Negative

Table II. Examples from the Memotion dataset with actual and predicted sentiment labels

We trained the model over 20 epochs using a batch size of 100, the Adam optimizer, and a learning rate of 0.0001. During training, we monitored both the sentiment classification loss and the domain classification loss to ensure stable convergence and effective domain adaptation. During the initial epochs, particularly around Epoch 3, there was a brief spike in total loss, indicating some early instability. Despite this, the classification loss started at a relatively high value and gradually decreased as training continued. Both classification and domain losses consistently declined over subsequent epochs, eventually stabilizing at lower values by the final epoch. This steady convergence suggests that the model was learning effectively and becoming increasingly stable over time.

To further validate this observation, we visualized the feature alignment between the source and target domains using t-SNE projections. Figure 3 presents the t-SNE (t-Distributed Stochastic Neighbor Embedding) plot of the fused feature representations before and after training in a multimodal sentiment analysis framework. The plot before training, as shown in Figure 3(a), shows a clear separation between the source domain (MVSA-S) and the target domain (Memotion), indicating a significant domain shift due to differences in data distributions. In contrast, the plot after training, as shown in Figure 3(b), demonstrates improved alignment between the two domains, suggesting that domain adaptation techniques, such as adversarial training, have successfully reduced the domain gap.



(a) t-SNE Plot Before Training (b) t-SNE Plot After Training

Fig. 3: Visualization of t-SNE Plot of Feature Distributions: Pre-Training vs. Post-Training with Adversarial Learning

The red and blue points represent data samples from MVSA-S and Memotion, respectively, with increased overlap in the after-training plot, highlighting effective feature adaptation. This improved alignment in the fused feature space enhances the model's generalization ability for sentiment analysis across diverse datasets.

V. CONCLUSION

This study demonstrates the effectiveness of Adversarial Multimodal Sentiment Analysis with Cross-Modal and Cross-Domain Alignment (AMSA-CMCDA) in improving sentiment classification across datasets with significant domain shift. By leveraging advanced feature extraction, attention-based fusion, and adversarial adaptation, our approach ensures robust cross-modal and cross-domain alignment. The integration of Domain-Adversarial Neural Networks (DANN) and the Gradient Reversal Layer (GRL) facilitates knowledge transfer from a labeled source dataset to an unlabeled target dataset, enabling the learning of domain-invariant features. Experimental results demonstrate that AMSA-CMCDA significantly improves sentiment prediction, emphasizing its importance in real-world applications, particularly when labeled target data is scarce. Future work can focus on enhancing generalization across diverse domains like social media, reviews, and news, as well as expanding to multi-source adaptation and fine-grained sentiment analysis to further improve robustness of the proposed model.

REFERENCES

[1] William C Sleeman IV, Rishabh Kapoor, and Preetam Ghosh. Multimodal classification: Current landscape, taxonomy and future directions. ACM Computing Surveys, 55(7):1–31, 2022.

[2] Vani Golagana, S Vizianandа Row, and P Srinivasa Rao. Adaptive multimodal sentiment analysis: Improving fusion accuracy with dynamic attention for missing modality. Journal of Electrical Systems, 20, 2024.

[3] Huanjie Wang, Xiwei Bai, Jie Tan, and Jiechao Yang. Deep prototypical networks based domain adaptation for fault diagnosis. Journal of Intelligent Manufacturing, 33(4):973–983, 2022.

[4] Pedro Marcelino. Transfer learning from pre-trained models. Towards data science, 10(330):23, 2018.

[5] Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuan-Jing Huang. Cross-domain sentiment classification with target domain specific information. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2505–2513, 2018.

[6] Yftah Ziser and Roi Reichart. Pivot based language modeling for improved neural domain adaptation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1241–1251, 2018.

[7] Yftah Ziser and Roi Reichart. Task refinement learning for improved accuracy and stability of unsupervised domain adaptation. In proceedings of the
57th annual meeting of the Association for Computational Linguistics, pages 5895–5906, 2019.

[8] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. IEEE Transactions on Neural Networks and Learning Systems, 34(1):264–277, 2021.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[10] Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Domain adaptation: challenges, methods, datasets, and applications. IEEE access, 11:6973–7020, 2023.

[11] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In Computer Vision–
ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, pages 451–466. Springer,2016.

[12] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In Proceedings of the 26th ACM international conference on Multimedia, pages 429–437, 2018.

[13] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057. PMLR, 2015.

[14] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474, 2014.

[15] Viswa Mani Kiran Peddinti and Prakriti Chintalapoodi. Domain adaptation in sentiment analysis of twitter. Analyzing Microtext, 11:05, 2011.

[16] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware bert for cross-domain sentiment analysis. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics, pages 4019–4028, 2020.

[17] Eman T Hassan, Xin Chen, and David Crandall. Unsupervised domain adaptation using generative models and self-ensembling. arXiv preprint arXiv:1812.00479, 2018.

[18] Lakmal Meegahapola, Hamza Hassoune, and Daniel Gatica-Perez. M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8(2):1–30, 2024.

[19] Pengfei Liang, Bin Wang, Guoqian Jiang, Na Li, and Lijie Zhang. Unsupervised fault diagnosis of wind turbine bearing via a deep residual deformable convolution network based on subdomain adaptation under time-varying speeds. Engineering Applications of Artificial Intelligence, 118:105656, 2023.

[20] Pengfei Liang, Leitao Xu, Hanqin Shuai, Xiaoming Yuan, Bin Wang, and Lijie Zhang. Semisupervised subdomain adaptation graph convolutional network for fault transfer diagnosis of rotating machinery under time-varying speeds. IEEE/ASME Transactions on Mechatronics, 29(1):730–741, 2023.

[21] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, pages 1180–1189. PMLR, 2015.

[22] Vani Golagana, S Viziananda Row, and P Srinivasa Rao. Multimodal feature fusion for image retrieval using deep learning. Journal of Data Acquisition and Processing, 39(1):823–839, 2024.

[23] Nan Xu and Wenji Mao. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 2399–2402, 2017.

[24] Nan Xu, Wenji Mao, and Guandan Chen. A co-memory network for multimodal sentiment analysis. In The 41st international ACM SIGIR conference on research & development in information retrieval, pages 929–932, 2018.

[25] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! arXiv preprint arXiv:2008.03781, 2020.