Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

A Systematic Review of Multimodal Sentiment Analysis Based on Text-Image Fusion: Trends, Models, and Research Gaps

[1)]Mohammed Abdul Mohsen Hamidi, [2)]Alaa Yaseen Taqa, [3)]Yahya Ismail Ibrahim
[1,2,3)] University of Mosul, Iraq
Mohammed.23esp4@student.uomosul.edu.iq, alaa.taqa@uomosul.edu.iq , yahyaismail@uomosul.edu.iq

Submitted :May 31, 2025 | **Accepted** : Jun 19, 2025 | **Published** : Jun 21, 2025

Abstract: Sentiment analysis has evolved from text-based approaches to multimodal sentiment analysis (MSA), which integrates textual and visual data to enhance the accuracy of emotional understanding, especially in visually rich social media contexts. This study presents a systematic literature review (SLR) focusing on recent developments in text-image-based MSA, aiming to identify prevailing methods, fusion strategies, and major research gaps. Following the PRISMA protocol, a total of 20 key articles published between 2019 and 2024 were selected and analyzed. The results indicate that deep learning models such as LXMERT, ViLBERT, and ERNIE-ViL outperform traditional architectures, achieving accuracies above 80% on datasets like MVSA and Twitter. Attention mechanisms and advanced feature fusion techniques significantly contribute to improving both accuracy and interpretability. However, challenges remain in terms of annotation quality, semantic alignment across modalities, and real-time implementation constraints. This study contributes by mapping the state-of-the-art in multimodal sentiment analysis, highlighting underexplored research gaps, and offering directions for future work toward more adaptive and context-aware sentiment systems..

Keywords: multimodal sentiment analysis, text-image fusion, deep learning, systematic review, attention mechanism.

INTRODUCTION

To find out what people think about a product or item, it's simple to create a poll and ask a few questions about the product, but it's much easier for individuals to ignore the poll. Instead, by searching for specific keywords on social media platforms, users can access millions of posts, tweets, and comments written by others. Social media is an irresistible desire that makes people freely post their praise or criticism about everything they encounter in their daily lives. Then comes sentiment analysis, which extracts opinions from those tweets.

Sentiment analysis (SA), also known as opinion mining (OM), could be defined as the study of extracting people's opinions, feelings, attitudes, and emotions about objects such as products, services, people, events, topics, and their specifications (Rajesh & Hiwarkar, 2023). It is defined as the process of identifying the main ideas within a text (Gherkar et al., 2022). It aims to generate opinions from texts (H. D. Sharma & Goyal, 2023).The main aim of SA is to determine whether the presented text represents a positive, negative, or neutral attitude (Aftab et al., 2023), which leads to SA being considered a text classification (Choi & Lee, 2017).

SA is the systematic investigation of the pervasive human inclination to articulate valuations and emotional responses within digital discourse, particularly on social media platforms, to discern and categorize these expressions as positive, negative, or neutral concerning a specific entity or concept.

Sentiment analysis aids decision-making by understanding and extracting human intentions, feelings, and responses, and is therefore used in many practical applications in business, social communication analysis, and public policy.

SA is classified into two main types, based on the type of data used as inputs to the model: unimodal and multimodal sentiment analysis. Unimodal SA is further divided into textual, visual, and audio SA. Unimodal SA methods typically achieve lower recognition rates.

Sentiment analysis has shifted from purely text-based sentiment analysis to multimodal analysis, which incorporates diverse input sources, such as audio, images, and video, to enhance sentiment detection (Das & Singh, 2023). Figure 1 illustrates the general framework of multimodal sentiment analysis.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

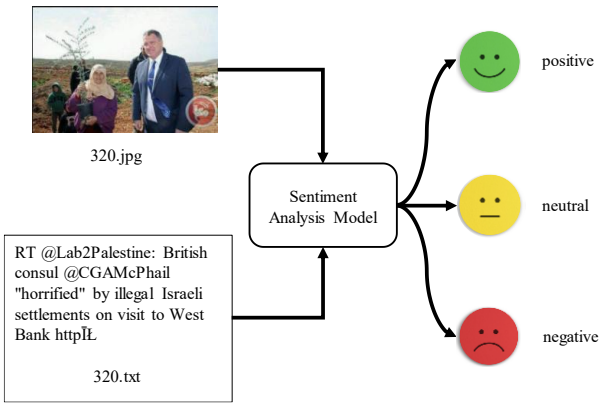

Figure 1: General framework for multimodal sentiment analysis

Some applications of SA include: healthcare, education, academic research, digital marketing, financial markets, and security applications. Table 1 below provides a summary of these applications and their benefits, with a reference to each.

Table 1 shows some of the applications and benefits of using sentiment analysis in various fields.

Ref.	application	benefits
(Raza et al., 2019)	Scientific articles	SA helps in selecting high-quality research papers by categorizing citation sentences.
(Denecke & Reichenpfader, 2023)	Healthcare	Improving healthcare services and patient care.
(K. Wang & Zhang, 2020)	Education	SA enhances teaching, supporting e-learning retention, , and enabling students to express opinions freely.
(Tan et al., 2023)	Digital marketing	SA aids in market research, assessing brand reputation, providing reliable insights on products and services, improving customer satisfaction.
(Chan & Chong, 2017)	Financial and stock markets	SA helps to identifying investment trends, and providing valuable insights for informed financial decisions.
(Deb et al., 2018)	Security	SA enhances digital safety and threat prevention.

RESEARCH OBJECTIVE

The main objective of this literature review is to review the latest studies in the field of MSA, specifically those that combine images and text for sentiment classification. The review begins by presenting the datasets used in these studies, and analyzing the current architectures and their components (feature extraction methods, fusion methods, attention mechanisms, and results). It concludes by discussing recurring problems in current models and evaluating proposed solutions to improve model performance on multimodal sentiment classification tasks.

Topic: Recent models of MSA combining images and text and the challenges they face.
Key question: How do modern image-text integration architectures improve sentiment classification accuracy? What are the challenges in achieving top performance in MSA?

METHOD

Search Strategy

These studies that were reviewed were gathered from different reputable databases: Google Scholar, Semantic Scholar, SpringerLink, Science Direct, and IEEE Xplore. For this review, studies on MSA through image and text fusion, e.g., those that were published after 2019 and onwards, were only considered. The keywords for searching were "multimodal sentiment analysis" and "sentiment analysis through image and text." The literature search was only done in English.

Inclusion and Exclusion Criteria

The study's inclusion criteria were peer-reviewed papers that appeared in research journals and focused on multimodal sentiment analysis. Guided by the Preferred Reporting Items for Systematic Reviews and Meta-

*name of corresponding author

Analyses (PRISMA) guidelines (Page et al., 2021), the present review was specific to recent intelligent model developments that were applicable to sentiment analysis. Screening was based on a review of the titles and abstracts of identified studies. Research that met the eligibility criteria were those which fulfilled the following conditions: (1) the work was on multimodal sentiment analysis; (2) it was limited to modalities combining images and text; (3) it had a clear methodology and results; and (4) it included a clear explanation of the dataset used in the experiment. Research not meeting these requirements was excluded. Specifically, we excluded articles that covered other forms of media (e.g., audio or video), posts published prior to 2019, and articles with non-relevant titles to the review topic.

Data Collection and Analysis

Studies were systematically collected based on relevant keywords. Zotero software was used to organize and sort the retrieved references, including duplicate deletion. Mendeley software was used to manage citations and references during the review process.

Screening and Data Extraction

Scientific articles related to multimodal sentiment analysis, which relies on combining textual and image data, were reviewed. Information was extracted from the studies for narrative analysis, focusing on recent trends, used AI models, and challenges of text-image integration. The extraction included collecting data related to 1) the source, including the researcher's name and year; 2) identifying information about the model (type of fusion, method of extracting textual and image features, classification method, and evaluation metrics such as accuracy and F1 score); 3) the datasets used, the number of instances, and their links, if any; 4) the strengths and weaknesses of the model; 5) conclusions and practical implications in fields such as marketing, medicine, and social networks.

RESULT

Study Selection

Before beginning the study review, a duplicate extraction process is conducted. The review process consists of three stages: title review, abstract review, and full article content review. Of the total 94,895 articles obtained, 38,990 were excluded due to duplication. A full-text selection was then conducted for 225 articles deemed feasible. Of these, 205 articles were excluded because they contained other formats such as audio or video, were reviews, lacked clear results, or were not published in reputable journals. After a thorough screening, 20 research articles were found to meet the inclusion criteria for the literature review (see Figure 2).

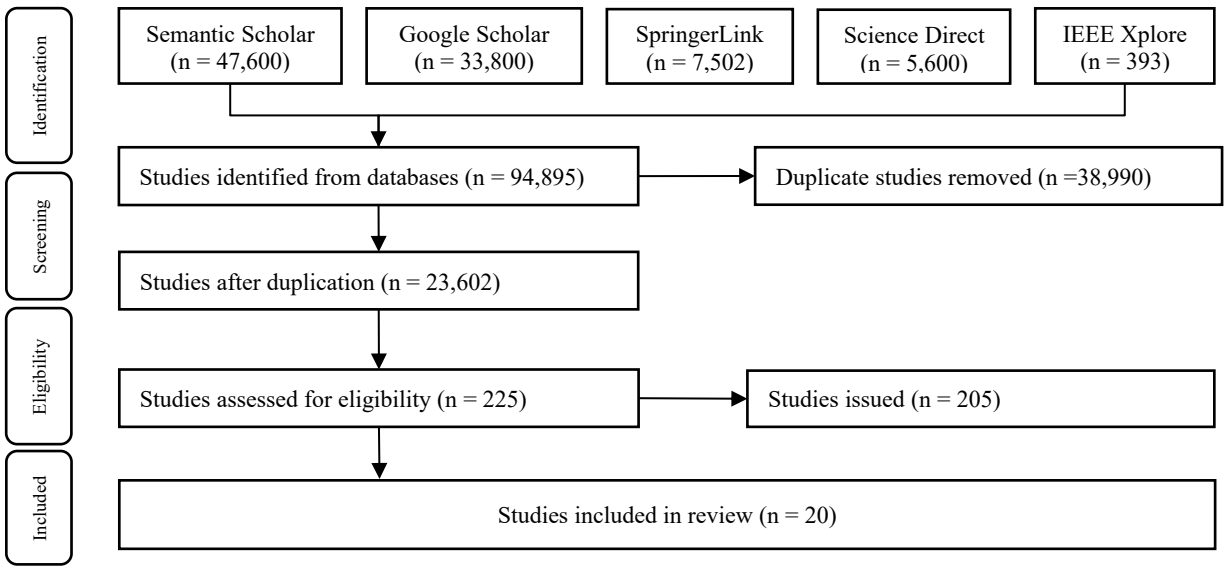

Figure 2: PRISMA Flow Diagram for Study Selection Process

Characteristics of Included Studies

The 20 selected studies were published between 2019 and 2024 in peer-reviewed journals or reputable conference proceedings. The majority of these studies utilized publicly available datasets such as Twitter, Instagram or MVSA-Single. Most employed deep learning-based architectures for feature extraction and fusion, including CNNs, RNNs, Transformers, and attention mechanisms. Regarding modality, all studies focused exclusively on combining textual and visual data. Evaluation metrics primarily included accuracy, F1-score.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Datasets

Even though many researchers use personal datasets and want to collect their own, they can still use some online datasets for free or for a fee. Some valuable datasets are available for training textual, visual, or multimodal sentiment analysis approaches. Using published datasets saves time and lets researchers compare results with others who used the same dataset. Table 2 represents some of the datasets used in multimodal sentiment analysis, respectively. Here are some of the most popular datasets in sentiment analysis

Table 2: Datasets for Multimodal Sentiment Analysis

No	Ref	Dataset Name	Modalities	Size	Structure	Outputs	Year	Link
1	(Yang et al., 2021) (K. Zhang et al., 2020) (Zhu et al., 2023) (Lei et al., 2024)	MVSA-Single	Text + Images	4,869	Social media posts with text and images	Positive, Neutral, Negative	2017	https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/
2	(Yang et al., 2021) (K. Zhang et al., 2020) (Zhu et al., 2023)	MVSA-Multiple	Text + Images	19,665	Social media posts with text and images	Positive, Neutral, Negative	2017	https://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/
3	(C. Sharma et al., 2020)	Memotion 1.0	Text + Images	7,000	Images and metadata with multi-label annotations	5-point sentiment scale	2020	https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k
4	(Ramamoorthy et al., 2022)	Memotion 2.0	Text + Images	10,000	Images and metadata with multi-label annotations	Positive, Neutral, Negative	2022	https://www.kaggle.com/datasets/weipengfei/memotion2
5	(Zhou et al., 2021)	MASAD	Text + Images	38,532	Folders for texts and images	Positive, Negative	2021	https://github.com/DrJZhou/MASAD?tab=readme-ov-file
6	(Gu et al., 2021) (X. Yin & Chen, 2023) (Liu et al., 2025)	Twitter 15	Text + Images	3,502	Multimodal tweets: text, image, target, and sentiment labels.	Positive, Neutral, Negative	2015	https://github.com/yhit98/FITE
7	(Gu et al., 2021) (X. Yin & Chen, 2023) (Liu et al., 2025)	Twitter 17	Text + Images	2,910	Multimodal tweets: text, image, target, and sentiment labels.	Positive, Neutral, Negative	2017	https://github.com/yhit98/FITE

Features Extraction:

The process of collecting relevant information from a text or image to use as input to supervise or guide subsequent processes. For feature extraction from text, deep learning delivers superior results over previous techniques on many machine learning-based tasks, enabling the development of new technologies in this field globally (W. Yin et al., 2017). Transfer learning is a powerful and effective technique for extracting valuable and important features from a wide range of different textual data types. Advanced transfer learning methodologies, such as BERT, Roberta, and GPT-3, improve the efficiency of natural language processing and reduce training time and computational resources, particularly in text analysis and generation. For image features, deep learning, specifically convolutional neural networks (CNNs), are also used to extract features from images. The advantage of CNNs lies in their ability to automatically detect image features without human intervention (Alzubaidi et al., 2021). CNNs form the basis of many pre-trained (transformer) models trained on ImageNet. Table 3 presents some models used to extract features from texts and images.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 3: Text feature extraction models

Ref	Model	Features	Description	Advantages	Disadvantages
(Huang et al., 2019)	GloVe + LSTM	Text	Uses pre-trained GloVe with LSTM to model text sequences.	Captures semantics and temporal patterns.	Sensitive to unknown words; slower training.
(Xu et al., 2021)	BiGRU	Text	Processes text forward and backward.	Captures full context; faster than LSTM.	Weaker for long texts than attention models.
(Yang et al., 2021)	BiLSTM + CNN	Text	Combines BiLSTM and CNN for sequence and local features.	Balances global and local feature extraction.	Complex model with a higher parameter count.
(Chen et al., 2024)	BERT	Text	The standard BERT model provides deep bidirectional context for text features.	High performance on many tasks.	Computationally expensive; requires large-scale pre-training.
(de Toledo & Marcacini, 2022)	DistilBERT	Text	A lighter, distilled version of BERT that produces robust contextual embeddings efficiently.	Faster and smaller than BERT.	Slightly less accurate compared to full BERT.
(Salman Al-Tameemi et al., 2023)	BERT-Base + CNN-GRU	Text	Fuses BERT embeddings with both CNN and GRU layers, capturing local and long-range patterns.	Combines local and long-term patterns.	More complex; needs careful tuning.
(Zhu et al., 2023)	BERT + BiGRU	Text	Enhances BERT's contextual embeddings with a bidirectional GRU layer for additional sequential modeling.	Richer semantic context.	Heavier model; needs more computation.
(Gong et al., 2024)	BERT + BiLSTM	Text	Augments BERT embeddings with a bidirectional LSTM layer to further capture sequential nuances.	Captures deeper meaning.	Model complexity and training time.
(Lei et al., 2024)	ALBERT + BiLSTM	Text	Uses the lightweight ALBERT model with a BiLSTM layer for efficient contextual learning.	Efficient and memory-saving.	May require careful fine-tuning on specific tasks
(Huang et al., 2019) (Xu et al., 2021)	VGG19	Image	Deep CNN with 19 layers providing robust feature extraction.	Well-known; generalizable features.	High computational cost; slower inference.
(Yang et al., 2021)	VGG-Object/Place	Image	VGG-based model tuned specifically for object and scene recognition.	Better for specific visual contexts.	Limited domain generalization.
(de Toledo & Marcacini, 2022) (H. Wang et al., 2023) (X. Yin & Chen, 2023) (Chen et al., 2024)	ResNet50	Image	A residual network with 50 layers offering a balance between depth and computational efficiency.	Balanced performance and efficiency.	Less powerful than deeper models.
(Gu et al., 2021)	ResNet-152	Image	Deep residual network with 152 layers that uses skip connections to ease training.	High accuracy; avoids vanishing gradients.	Very slow; high computational demand.
(Zhu et al., 2023)	Faster R-CNN (ResNet-101)	Image	Object detection model using a ResNet-101 backbone for region proposal and recognition tasks.	Accurate detection and localization.	Complex; resource-intensive.
(Salman Al-Tameemi et al., 2023) (Yadav & Vishwakarma, 2023)	Inception-V3	Image	CNN that uses inception modules to factorize convolutions and balance cost with accuracy.	Efficient with strong performance.	Complex architecture.
(Lei et al., 2024)	DenseNet121 + CBAM	Image	DenseNet121 enhanced with a Convolutional Block Attention Module (CBAM) to focus on key regions.	Highlights key regions; better feature reuse.	More complex and costly.
(Gong et al., 2024)	ViT + Faster R-CNN	Image	Integrates a Vision Transformer with Faster R-CNN to leverage both global attention and detection.	Superior integration of global and local features.	Complex; high resource need.
(Liu et al., 2025)	YOLOv8 + ViT	Image	Integrates YOLOv8's fast object detection with Vision Transformer feature extraction.	Fast and accurate detection.	Requires fine-tuning and resource-heavy

Multimodal Fusion Strategies:

Multimodal fusion strategies (e.g., text, images, and audio) are essential to improving the performance of models in understanding complex data. These strategies are divided into four main types. Each of the fusion methods has a merge location inside the model, which affects the computational complexity. Also, each method has its pros and

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

cons, which leads to its use accordingly. Table 4 here shows the merge location, pros and cons, application, and recalculation of each fusion method.

Table 4: Some of the differences between the fusion types.

Ref.	Fusion Type	Fusion Stage	Advantages	Disadvantages	Computational Complexity
(Boulahia et al., 2021)	Early Fusion	Before feature extraction	- Effective when patterns have uniform data types, ensuring stable properties. - Increases accuracy when pattern are combined early.	- Ineffective for heterogeneous inputs due to representation disparities. - Potential loss of source-specific information.	Medium to high (depends on data size)
(Boulahia et al., 2021)	Intermediate Fusion	After extracting some features, before full processing	- Enables interaction between features from different sources. - Reduces noise from raw features.	- Requires careful feature selection. - Needs balance between sub-models.	High (depends on the number of sub-models)
(Gadzicki et al., 2020)	Late Fusion	After processing data through independent models	- Enhances accuracy by allowing sub-models to work independently. - Easy to implement if models are well integrated.	- May lose data interactions before final decisions. - Possible incompatibility in model performance.	Low to medium
(Li et al., 2022)	Hybrid Fusion	A mix of early, intermediate, and late fusion	- Combines strengths of all fusion types - Increases model flexibility and diversity.	- Highly complex and hard to design. - Requires more training data.	Very high

Attention Mechanisms

Attention mechanisms are designed to help models focus on words in the text that are relevant to emotions and highlight important areas of the image, such as facial expressions or emotional cues. When combining text and images, attention facilitates cross-modal compatibility by enabling the model to establish meaningful associations between linguistic elements and corresponding visual features. These mechanisms enhance interpretability and performance of sentiment classification, especially in complex contexts. Table 5 shows details of some attention mechanisms.

Table 5: Details of some attention mechanisms.

Ref.	Attention mechanisms	Type	Description	Advantages	Disadvantages
(Xu et al., 2021) (Yadav & Vishwakarma, 2023) (Lei et al., 2024)	Channel attention	Visual	Focuses on the importance of different feature channels.	Improves feature discrimination; lightweight.	Ignores spatial or sequential relationships.
(Xu et al., 2021) (Yadav & Vishwakarma, 2023) (Lei et al., 2024)	Spatial Attention	Visual	Focuses on important spatial locations within visual data.	Effective for object detection and scene understanding.	Limited in capturing semantic content.
(Huang et al., 2019) (Xu et al., 2021)	Region attention	Visual	Attends to specific spatial or textual regions relevant to the context.	Effective in vision and dense-text analysis.	May overlook global context.
(Guo et al., 2021)	Layout-guided attention	Visual	Uses spatial layout to determine feature importance in images.	Captures structural relationships effectively.	May be limited for dynamic or unstructured layouts.
(Huang et al., 2019) (Yadav & Vishwakarma, 2023)	Semantic attention	Textual	Highlights important words in text sequences for better semantic learning.	Improves contextual understanding.	Dependent on model tuning for optimal performance.
(Salman Al-Tameemi et al., 2023) (Yadav & Vishwakarma, 2023)	Self-attention	General	Each element attends to every other in the same modality.	Captures global dependencies; widely used in NLP and vision.	Inefficient with long sequences unless optimized.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

(Gong et al., 2024) (Mu et al., 2025)	Multi-head attention	General	Divides attention into multiple heads to capture diverse dependencies.	Captures diverse patterns; enhances representation power.	Requires careful head balancing.
(Thuseethan et al., 2020)	Implicit attention	Multimodal	Learns feature importance implicitly rather than through explicit weighting.	No manual annotation required.	Harder to interpret and optimize.
(Zhu et al., 2023)	Adaptive Gating	Multimodal	Uses dynamic gates to control flow of attention between modalities.	Filters noise; adaptive control.	Adds architectural complexity.
(K. Zhang et al., 2020)	Fine-grained attention	Multimodal	Focuses on small-scale details within features for deeper analysis.	Improves fine feature differentiation.	Computationally heavy for dense datasets.
(Zhu et al., 2023) (Lei et al., 2024) (Liu et al., 2025)	Cross-modal attention	Multimodal	Enables attention across different modalities (e.g., text ↔ image).	Facilitates deep multimodal interaction.	Requires careful modality alignment.
(Hu & Yamamura, 2022)	Global-Local Attention	Multimodal	Balances global context with localized feature weighting.	Retains broad and detailed insights.	Needs proper calibration to avoid bias.
(Yang et al., 2021)	Memory-Guided Attention	Multimodal	Uses memory-based mechanisms to refine attention across sequences.	Enhances context retention over long inputs.	Requires additional memory storage

Multimodal sentiment analysis: state-of-art

In this regard, there has been remarkable progress in recent years using deep learning techniques and transformed models that have merged text and images more effectively. Most of the research on multimodal sentiment analysis in the last five years has focused on combining visual and textual information to enhance sentiment prediction. Different fusion techniques, such as intermediate, late, hybrid, and more complex approaches based on mutual attention and competition, have been tried to effectively fuse features from images and texts, while early fusion has been neglected due to its inability to achieve good results on heterogeneous data. The following tables provide a summary of the above studies, with table 6 presenting the models methods and it is performance, while table 7 illustrates the strengths, weaknesses.

Table 6: Presenting the models methods.

No	Ref.	Fusion type	Attention mechanism	Feature extraction method	Model architecture	Dataset	Acc.	F1	Strengths	Weaknesses
1	(Huang et al., 2019)	Intermediate + late (hybrid)	Visual (region), semantic attention	VGG19 (image) GloVe+ LSTM (text)	DMAF	Getty Image, Twitter, Flickr-w, Flickr-m	86.9 76.3 85.9 88.0	86.6 76.9 85.0 87.6	Effective multimodal fusion, attention mechanism for key regions, handles missing data.	High computational complexity, reliance on pre-trained models, risk of overfitting.
2	(Guo et al., 2021)	Late	Layout-guided attention	CNN (image) GRU (text)	LD-MAN	DMON	80.81	-	Incorporates layout for alignment, handles long text & multiple images, multimodal attention.	Complex layout modeling, dependent on image-text alignment, weak for non-layout datasets.
3	(Thuseethan et al., 2020)	Intermediate + late (hybrid)	Implicit attention	VGG-16 (image) VD-CNN (text)	Three-stream architecture (VFS + TFS + AFS)	Mix from (Flickr, Getty and Google)	82.81	82.75	Enhances intermodal relations, transfer learning improves efficiency, 10-fold validation.	Not benchmarked against public datasets, lower neutral-class performance, manual annotation needed.

*name of corresponding author

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

No	Ref.	Fusion type	Attention mechanism	Feature extraction method	Model architecture	Dataset	Acc.	F1	Strengths	Weaknesses
4	(Xu et al., 2021)	Intermediate (multi-layer)	Channel + region attention	VGG-19 (image) bi-GRU (text)	AHRM	Flickr Getty Image	87.1 87.8	87.5 88.4	Captures semantics, integrates social relations via GCN, strong benchmark performance.	Model complexity, over-smoothing in deep layers, assumes strong text-image correlation.
5	(Yang et al., 2021)	Intermediate	Memory-guided attention	VGG-Object/Place (image) BiLSTM + CNN (text)	MVAN	MVSA-Single MVSA-Multiple	72.98 72.36	72.98 72.30	Models cross-modal interactions, multi-view features improve robustness, state-of-the-art results.	High complexity, distant supervision may introduce noise.
6	(K. Zhang et al., 2020)	Intermediate	fine-grained attention	Improved VAE (image) Denoising Autoencoder (text)	CFF-ATT	MVSA-Single MVSA-Multiple	71.44 69.62	71.06 69.35	Reduces textual noise, extracts key image features, balances text-image interaction.	Assumes one image per text, removes ambiguous samples, preprocessing limitations.
7	(Gu et al., 2021)	Intermediate (multi-layer)	Multi-head attention	ResNet-152 (image) GloVe (text)	EF-Net	Twitter15 Twitter 17	73.65 67.77	67.77 65.32	Integrates MHA & capsule networks, handles multiple aspects.	High computational cost, reliance on manual annotation, risk of overfitting.
8	(Zhou et al., 2021)	Intermediate (multi-layer)	Multimodal interaction layer	ResNet-50 (image) GloVE + BiLSTM (text)	MMAE + MMAP	MASAD	95.63	95.09	Large-scale dataset, adversarial alignment of text-image features.	Handling image-text interaction complexity, unverified domain generalization.
9	(Hu & Yamamura, 2022)	Intermediate (multi-layer)	Global-local attention	CLIP (text/image) Faster RCNN	GLFN	MVSA-Single MVSA-Multiple	77.21 75.87	76.42 73.99	Combines global and local features, leverages pre-trained models for efficiency.	Relies on pre-trained models, computationally complex, struggles with unrelated image-text pairs.
10	(de Toledo & Marcacini, 2022)	Intermediate	Multi-head attention	ResNet-50(image) DistilBERT (text)	ResNet + DistilBERT	MVSA-Single MVSA-Multiple	81.31 74.61	74.54 73.49	Optimized fine-tuning for low-resource settings, reduces computational cost.	Requires paired data, struggles with complex dependencies, less effective than fully multimodal models.
11	(Zhu et al., 2023)	Intermediate (multi-layer)	Cross-Modal Attention + Adaptive Gating	Faster R-CNN (mutimodal) ResNet-18 (images) BERT + BiGRU (text)	ITIN	MVSA-Single MVSA-Multiple	75.19 73.52	74.97 73.49	Captures region-word correspondence, filters misaligned pairs, integrates visual-text context.	Less effective when text does not describe image, depends on context features if misalignment occurs.
12	(Salman Al-Tameemi et al., 2023)	Intermediate + late (hybrid)	Self-attention	Inception-V3 (image) BERT + CNN-GRU (text)	MCJF	BG BIS Twitter MVSA-Single	99.70 98.81 95.14	95.14 98.79 95.16	Handles missing modalities, integrates multimodal	High computational complexity, assumes strong image-text correlation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

No	Ref.	Fusion type	Attention mechanism	Feature extraction method	Model architecture	Dataset	Acc.	F1	Strengths	Weaknesses
							75.92	74.69	features effectively.	
13	(H. Wang et al., 2023)	Intermediate	Multi-head attention	ResNet50 (image) GloVe + BERT (text)	MLFC	MVSA-Single MVSA-Multiple	76.44 70.53	75.61 67.97	Effective multimodal fusion, robust with data augmentation, outperforms SOTA models.	High computational complexity, limited dataset diversity.
14	(Yadav & Vishwakarma, 2023)	Intermediate (multi-layer)	Channel, spatial, semantic and self-attention	Inception V3 (image) LSTM (text)	DMLANet	MVSA-Single MVSA-Multiple Flickr Getty Images	79.47 77.89 89.30 92.65	79.59 75.26 89.19 92.60	Enhances image-text correlation, extracts sentiment-rich features, uses bi-attentive visual attention.	Struggles with datasets lacking close cross-modal correlation.
15	(X. Yin & Chen, 2023)	Intermediate (multi-layer)	Multi-head attention	ResNet-50 (images), RoBERTa-BiLSTM (text/aspect)	TF-MMATI:	Twitter-15 Twitter-17	78.66 72.67	74.53 70.82	Models inconsistency and correlation separately, improves representation.	Struggles with mixed language, short text, or missing words.
16	(Lei et al., 2024)	Intermediate (multi-layer)	Channel, spatial and cross-attention	DenseNet121 (image) ALBERT + BiLSTM (text)	MCAM	MVSA-Single	86.5	85.5	Strong multimodal fusion via cross-attention, high accuracy, adapts to image/text sentiment dominance.	Sensitive to text misspellings, potential annotation errors, struggles with conflicting sentiments.
17	(Chen et al., 2024)	Intermediate + late (hybrid)	Multi-head attention	ResNet-50 (image) RoBERTa (text)	CF-MSA	MVSA-Single MVSA-Multiple	76.17 67.12	74.54 64.92	Reduces modality bias via counterfactual causal reasoning, achieving superior accuracy and F1 scores by balancing cross-modal interactions.	Computational complexity from triple-branch architecture; residual text-dependency limiting image-bias mitigation.
18	(Gong et al., 2024)	Intermediate + late (hybrid)	Multi-head and cross-modal attention	ViT + Faster R-CNN (Image) BERT + Bi-LSTM (Text)	MMJL	MVSA-Single MVSA-Multiple	76.98 75.32	76.23 75.29	Boosts multimodal sentiment analysis via multi-channel feature extraction (MCIF/MSTF) and cross-modal fusion (EMT), achieving top accuracy by balancing global-local image-text interactions.	High computational complexity and struggles with ambiguous cases due to inconsistent image-text cues.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

No	Ref.	Fusion type	Attention mechanism	Feature extraction method	Model architecture	Dataset	Acc.	F1	Strengths	Weaknesses
19	(Liu et al., 2025)	Intermediate (multi-layer)	Cross-modal attention	YOLOv8 + ViT (Image), BERT (Text)	CORSA	Twitter-15 Twitter-17	69.9 70.6	69.9 70.6	Enhances multimodal aspect-based sentiment analysis by filtering irrelevant visual data and localize condition-related regions, achieving state-of-the-art accuracy despite unmet image-text conditions.	High computational complexity; Relies on error-prone automated annotations; struggles with multiple aspects and complex cases
20	(Mu et al., 2025)	Intermediate + late (hybrid)	Multi-head and cross-modal attention	ResNet-101 (image) BERT (text)	SECIF	MVSA-Single MVSA-Multiple Weibo	79.10 73.09 90.86	78.74 72.79 90.91	Integrates text and image via smart cross-modal interaction, achieves top accuracy by enhancing text features by using semantic enhancement.	High computational complexity and resource demands; struggles with sarcasm.

DISCUSSIONS

Dataset Limitations

The review indicated that the majority of research is based on small datasets, e.g., MVSA-Single and MVSA-Multiple (Yang et al., 2021); (Zhu et al., 2023), which are comparatively small (fewer than 20,000 samples) and restricted to social media environments. This hinders the capacity of models to generalize findings to other domains like healthcare or education (K. Wang & Zhang, 2020). In addition, certain datasets like Memotion have issues of class imbalance and are based on human classification processes that might introduce training biases (Ramamoorthy et al., 2022).

Some other challenges are training models on a single dataset that is not linguistically and culturally diverse, leading to the development of inefficient models for multilingual or multicultural settings (Blodgett et al., 2020). While incorporating dataset diversity has the potential to bring about better generalizability for models, it also increases training time and computational expenses (Krizhevsky et al., 2017); (Sun et al., 2017).

In addition to this, there is a myriad of ethical issues that should be taken into perspective. One of the main issues involves the use of datasets responsibly and the protection of privacy for individuals. Another issue entails the validation of labels, which can be biased because they are based on human judgment. Finally, it is important to ensure that the model is fair, especially through keeping a balanced ratio across different classes (Bender & Friedman, 2018); (Paullada et al., 2021).

Model Architectures and Fusion Strategies

Some of the latest advancements in the area of multimodal sentiment analysis (MSA) have concentrated heavily on multi-level and hybrid architectures for enhancing the textual and visual information fusion. A prime example of the same is the MMJL model (Gong et al., 2024), through its hybrid fusion mechanism with a quite satisfactory accuracy of 76.98 on the MVSA-Single dataset but at a very high cost of computational resources. Other approaches, like MCAM (Lei et al., 2024) and GLFN in (Hu & Yamamura, 2022), utilize multi-layer intermediate fusion and multi-head and cross-modal attention mechanisms. These techniques have achieved impressive performance improvement; e.g., MCAM achieves an accuracy of 86.5. MCAM surpasses other architectures in dynamically attending to valuable sentiment indicators in several modalities and channels, thereby facilitating its higher performance scores. Other models, like CF-MSA (Chen et al., 2024) and SECIF (Mu et al., 2025), are designed based on sophisticated reasoning and semantic methods. Although these models perform adequately with accuracy scores of 76.17 and 79.10, respectively, on the MVSA-Single dataset, they exhibit high complexity.

It is worth mentioning that recent research on multimodal sentiment analysis (MSA) favors intermediate fusion methods, more specifically multi-layer fusion, as such methods allow for step-by-step interactions and alignments

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

between modalities, for instance, text and images. For example, the DMAP model (Huang et al., 2019) illustrates the efficiency of this particular approach. While, on the other hand, early fusion provides an easier and computationally lighter solution but tends to suffer from the issue of modalities with different alignments, leading to inferior performance in modeling complex correlations between modalities. Hybrid fusion methods, which blend components of early, intermediate, and late fusion, can potentially achieve better performance but tend to come with high computational complexity, lengthy training, and higher inference latency—issues that preclude their application to real-time applications such as social media surveillance.

In conclusion, attention-based architectures yield improved performance, with MCAM performing better than MVSA-Single, yet it is still a significant challenge to achieve a balance between complexity and performance.

Attention Mechanisms

Attention mechanisms are an important part of modern multimodal sentiment analysis models. They help the model focus on the most useful features from both text and images. Different types of attention, such as channel, spatial, regional, and multimodal attention, allow models like MCAM (Lei et al., 2024) to better connect information from different sources and highlight key sentiment cues.

While these mechanisms process noisy or missing data, they add computational complexity to the model architecture. Moreover, adding these mechanisms to multi-layer fusion enhances feature interaction between modalities and, consequently, boosts performance.

Feature Extraction

In text feature extraction, state-of-the-art models like BERT and DistilBERT (Chen et al., 2024) (de Toledo & Marcacini, 2022) demonstrate ability in maintaining semantic contexts but at the expense of significant computational power. In image processing, on the contrary, models like ResNet-152 and ViT display greater accuracy rates but at the expense of processing speed (Zhu et al., 2023) (Gong et al., 2024). These findings clearly indicate an urgent demand for even lighter models, e.g., ALBERT, that can effectively trade off performance and operational efficiency (Lei et al., 2024).

Research Gaps and Future Directions

In spite of increasing sophistication of multimodal sentiment models, there are many research gaps. First, most studies rely heavily on limited datasets such as MVSA and Memotion, which lack linguistic and cultural diversity, reducing generalizability. Second, fusion strategies that incorporate early-intermediate-late combinations (i.e., hybrid) remain computationally expensive and unsuitable for real-time applications. Third, the interpretability of deep multimodal models is still limited, especially in distinguishing the contribution of each modality. Future research should focus on lightweight, explainable architectures, as well as real-world multilingual datasets for robust cross-cultural applications.

CONCLUSION

Amid the rapid developments in multimodal sentiment analysis (MSA), this review highlights the tremendous potential of integrating textual and visual analytics to improve sentiment classification accuracy. The results demonstrate that modern media fusion techniques, such as hybrid and interactive fusion, along with advanced attention mechanisms (e.g., interlaced and multilayered attention), significantly improve robustness compared to traditional unimodal approaches. This progress stems from innovations in AI models like BERT (text) and ResNet/ViT (images), which enable deeper intermedia interactions. However, challenges persist, including computational complexity, limited datasets (e.g., culturally homogeneous data), and modality imbalance. Ethical concerns, such as annotation bias and lack of diversity representation, further complicate real-world deployment. Future research must prioritize lightweight models, diversified datasets, and cross-modal alignment mechanisms. By addressing these issues, MSA can revolutionize domains like healthcare (patient sentiment monitoring), education (learning experience evaluation), and digital reputation management, delivering precise, actionable insights.

This review contributes to the field by consolidating recent trends in multimodal fusion and attention mechanisms, highlighting their trade-offs, and mapping the future landscape for sentiment analysis that is both adaptive and ethically sound.

REFERENCES

Aftab, F., Bazai, S. U., Marjan, S., Baloch, L., Aslam, S., Amphawan, A., & Neo, T. K. (2023). A Comprehensive Survey on Sentiment Analysis Techniques. *International Journal of Technology*, *14*(6), 1288–1298. https://doi.org/10.14716/ijtech.v14i6.6632

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A.,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). Language (Technology) is power: A critical survey of ⇜bias" in NLP. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *c*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, *32*(6). https://doi.org/10.1007/s00138-021-01249-8

Chan, S. W. K., & Chong, M. W. C. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, *94*(August), 53–64. https://doi.org/10.1016/j.dss.2016.10.006

Chen, F., Huang, P., Ge, X., Huang, J., & Bao, Z. (2024). Multimodal Sentiment Analysis Based on Causal Reasoning. *ArXiv Preprint ArXiv:2412.07292*.

Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers*, *19*(5), 993–1012. https://doi.org/10.1007/s10796-017-9741-7

Das, R., & Singh, T. D. (2023). Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. *ACM Computing Surveys*, *55*(13). https://doi.org/10.1145/3586075

de Toledo, G. L., & Marcacini, R. M. (2022). Transfer learning with joint fine-tuning for multimodal sentiment analysis. *ArXiv Preprint ArXiv:2210.05790*.

Deb, A., Lerman, K., & Ferrara, E. (2018). Predicting cyber-events by leveraging hacker sentiment. *Information (Switzerland)*, *9*(11), 1–18. https://doi.org/10.3390/info9110280

Denecke, K., & Reichenpfader, D. (2023). Sentiment analysis of clinical narratives: A scoping review. *Journal of Biomedical Informatics*, *140*(March). https://doi.org/10.1016/j.jbi.2023.104336

Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020). Early vs late fusion in multimodal convolutional neural networks. *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*, *July 2020*. https://doi.org/10.23919/FUSION45008.2020.9190246

Gherkar, Y., Gujar, P., Gaziyani, A., & Kadu, S. (2022). *Keyword : 03029*, 1–6.

Gong, L., He, X., & Yang, J. (2024). An Image-Text Sentiment Analysis Method Using Multi-Channel Multi-Modal Joint Learning. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2371712

Gu, D., Wang, J., Cai, S., Yang, C., Song, Z., Zhao, H., Xiao, L., & Wang, H. (2021). Targeted Aspect-Based Multimodal Sentiment Analysis: An Attention Capsule Extraction and Multi-Head Fusion Network. *IEEE Access*, *9*, 157329–157336. https://doi.org/10.1109/ACCESS.2021.3126782

Guo, W., Zhang, Y., Cai, X., Meng, L., Yang, J., & Yuan, X. (2021). LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition. *IEEE Transactions on Multimedia*, *23*, 1785–1798. https://doi.org/10.1109/TMM.2020.3003648

Hu, X., & Yamamura, M. (2022). Global Local Fusion Neural Network for Multimodal Sentiment Analysis. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178453

Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, *167*, 26–37. https://doi.org/10.1016/j.knosys.2019.01.019

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lei, Y., Qu, K., Zhao, Y., Han, Q., & Wang, X. (2024). Multimodal Sentiment Analysis Based on Composite Hierarchical Fusion. *Computer Journal*, *67*(6), 2230–2245. https://doi.org/10.1093/comjnl/bxae002

Li, J., Zhang, Z., Lang, J., Jiang, Y., An, L., Zou, P., Xu, Y., Gao, S., Lin, J., Fan, C., Sun, X., & Wang, M. (2022). Hybrid Multimodal Feature Extraction, Mining and Fusion for Sentiment Analysis. In *MuSe 2022 - Proceedings of the 3rd International Multimodal Sentiment Analysis Workshop and Challenge* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3551876.3554809

Liu, X., Li, R., Ye, S., Zhang, G., & Wang, X. (2025). Multimodal Aspect-Based Sentiment Analysis under Conditional Relation. *Proceedings of the 31st International Conference on Computational Linguistics*, 313–323.

Mu, G., Chen, Y., Li, X., Dai, L., & Dai, J. (2025). Semantic enhancement and cross-modal interaction fusion for sentiment analysis in social media. *PloS One*, *20*(4), e0321011.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: an updated guideline for reporting

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Number 2, April 2025
DOI : https://doi.org/10.33395/sinkron.v9i2.14840
e-ISSN : 2541-2019
p-ISSN : 2541-044X

systematic reviews. *Bmj*, *372*.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Rajesh, A., & Hiwarkar, T. (2023). Sentiment analysis from textual data using multiple channels deep learning models. *Journal of Electrical Systems and Information Technology*, *10*(1). https://doi.org/10.1186/s43067-023-00125-x

Ramamoorthy, S., Gunti, N., Mishra, S., Suryavardan, S., Reganti, A., Patwa, P., DaS, A., Chakraborty, T., Sheth, A., & Ekbal, A. (2022). Memotion 2: Dataset on sentiment and emotion analysis of memes. *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.

Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, *10*(12), 157–165. https://doi.org/10.14569/ijacsa.2019.0101222

Salman Al-Tameemi, I. K., Feizi-Derakhshi, M. R., Pashazadeh, S., & Asadpour, M. (2023). Multi-Model Fusion Framework Using Deep Learning for Visual-Textual Sentiment Classification. *Computers, Materials and Continua*, *76*(2), 2145–2177. https://doi.org/10.32604/CMC.2023.040997

Sharma, C., Bhageria, D., Scott, W., Pykl, S., Das, A., Chakraborty, T., Pulabaigari, V., & Gamback, B. (2020). SemEval-2020 Task 8: Memotion Analysis--The Visuo-Lingual Metaphor! *ArXiv Preprint ArXiv:2008.03781*.

Sharma, H. D., & Goyal, P. (2023). *An Analysis of Sentiment : Methods , Applications ,. Ml*.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 843–852. https://doi.org/10.1109/ICCV.2017.97

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences (Switzerland)*, *13*(7). https://doi.org/10.3390/app13074550

Thuseethan, S., Janarthan, S., Rajasegarar, S., Kumari, P., & Yearwood, J. (2020). Multimodal deep learning framework for sentiment analysis from text-image web data. *Proceedings - 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2020*, 267–274. https://doi.org/10.1109/WIIAT50758.2020.00039

Wang, H., Li, X., Ren, Z., Wang, M., & Ma, C. (2023). Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion. *Sensors*, *23*(5), 1–15. https://doi.org/10.3390/s23052679

Wang, K., & Zhang, Y. (2020). Topic Sentiment Analysis in Online Learning Community from College Students. *Journal of Data and Information Science*, *5*(2), 33–61. https://doi.org/10.2478/jdis-2020-0009

Xu, J., Li, Z., Huang, F., Li, C., & Yu, P. S. (2021). Social Image Sentiment Analysis by Exploiting Multimodal Content and Heterogeneous Relations. *IEEE Transactions on Industrial Informatics*, *17*(4), 2974–2982. https://doi.org/10.1109/TII.2020.3005405

Yadav, A., & Vishwakarma, D. K. (2023). A Deep Multi-level Attentive Network for Multimodal Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(1), 1–11. https://doi.org/10.1145/3517139

Yang, X., Feng, S., Wang, D., & Zhang, Y. (2021). Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, *23*(c), 4014–4026. https://doi.org/10.1109/TMM.2020.3035277

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing*. http://arxiv.org/abs/1702.01923

Yin, X., & Chen, L. (2023). Image and Text Aspect Level Multimodal Sentiment Classification Model Using Transformer and Multilayer Attention Interaction. *International Journal of Data Warehousing and Mining*, *19*(1), 1–20. https://doi.org/10.4018/IJDWM.333854

Zhang, K., Geng, Y., Zhao, J., Liu, J., & Li, W. (2020). Sentiment analysis of social media via multimodal feature fusion. *Symmetry*, *12*(12), 1–14. https://doi.org/10.3390/sym12122010

Zhou, J., Zhao, J., Huang, J. X., Hu, Q. V., & He, L. (2021). MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, *455*, 47–58. https://doi.org/10.1016/j.neucom.2021.05.040

Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., & Qian, J. (2023). Multimodal Sentiment Analysis with Image-Text Interaction Network. *IEEE Transactions on Multimedia*, *25*, 3375–3385. https://doi.org/10.1109/TMM.2022.3160060

*name of corresponding author