International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 11 November 2025
Page No: 4006-4011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0586

Multi-Modal Sentiment Analysis of Social Media Using CNN-LSTM Hybrid Models

Gubbala Narasanna[1], Chiranjeevi S P Rao Kandula[2], Dr.P.Srinu Vasarao[3]
[1]M. Tech Reseach Scholar, Swarnandhra College of Engineering and Technology, India.
[2,3]Assistant Professor, Swarnandhra College of Engineering and Technology, India.
Emails: *narasannagubbala@gmail.com[1], chsprabhakar.kandula@gmail.com[2], psrinu.cse@swarnandhra.ac.in[3]*

Abstract

Social media platforms generate massive volumes of content in multiple formats, such as text, images, and videos, reflecting users' opinions and emotions. Conventional sentiment analysis methods often focus solely on textual data, ignoring valuable cues present in visual content. This study presents a CNN-LSTM hybrid model for multi-modal sentiment analysis, leveraging Convolutional Neural Networks (CNNs) to extract features from images and Long Short-Term Memory (LSTM) networks to model sequential dependencies in text. The proposed approach fuses textual and visual features to enhance sentiment classification into positive, negative, and neutral categories. Evaluation on a multi-modal social media dataset demonstrates that the hybrid model significantly outperforms single-modality models in terms of accuracy, precision, recall, and F1-score. The findings underscore the importance of integrating multiple data modalities for robust sentiment prediction, with potential applications in brand monitoring, social media analytics, and public opinion tracking.

Keywords: *Multi-Modal Sentiment Analysis, Social Media Analytics, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Deep Learning, Text and Image Fusion, Emotion Detection, Hybrid Neural Networks, Public Opinion Mining, Social Media Monitoring.*

1. Introduction

In recent years, social media platforms such as Twitter, Facebook, and Instagram have become primary channels for people to express their opinions, emotions, and experiences. These platforms generate a vast amount of multi-modal content, including text, images, videos, and emojis, reflecting public sentiment on a wide range of topics such as politics, entertainment, and brand perception. Understanding and analyzing this sentiment is critical for applications like brand monitoring, public opinion tracking, recommendation systems, and social media analytics. Traditional sentiment analysis approaches have primarily focused on textual data using techniques like machine learning classifiers or deep learning models such as LSTM and BERT. While these methods can capture semantic and contextual information from text, they often overlook valuable cues present in visual content, such as images or videos, which can convey emotions more effectively than text alone. To address this limitation, multi-modal sentiment analysis integrates information from multiple data modalities, enabling more accurate and robust sentiment prediction. Hybrid deep learning architectures, particularly CNN-LSTM models, are well-suited for this task: CNNs excel at extracting spatial features from images, while LSTMs are effective in capturing sequential dependencies in text. By combining these strengths, hybrid models can leverage complementary information from both text and visual data to improve sentiment classification performance. This research aims to develop a CNN-LSTM hybrid framework for multi-modal sentiment analysis of social media data, exploring how text-image fusion can enhance sentiment detection accuracy. The study also

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 11 November 2025
Page No: 4006-4011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0586

evaluates the performance of the proposed model compared to single-modality approaches, highlighting the benefits of integrating multi-modal information for understanding complex sentiment patterns in social media content [1].

2. Literature Survey

Sentiment analysis has become a crucial area of research due to its applications in understanding public opinion, consumer behavior, and social interactions. Early studies largely focused on text-based sentiment analysis, utilizing traditional machine learning techniques such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. For example, Pang et al. (2002) applied SVM to classify movie reviews and achieved encouraging results. However, these approaches often struggled with the informal, noisy, and context-dependent nature of social media text, limiting their effectiveness. With the advancement of deep learning, models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks gained prominence. LSTM networks, in particular, are capable of capturing long-term dependencies in text, addressing limitations like the vanishing gradient problem found in traditional RNNs. Studies such as Wang et al. (2016) highlighted LSTM's superiority over classical methods in sentiment classification tasks on social media datasets. Extensions like Bidirectional LSTM (Bi-LSTM) and attention mechanisms further improved performance by focusing on sentiment-relevant words within sequences. While text-based approaches achieved notable progress, the growing presence of multi-modal content—including images, videos, emojis, and GIFs—demonstrated the limitations of single-modality analysis. CNN-based models have been widely applied to image sentiment analysis, efficiently extracting spatial and visual features. Models such as VGG16 and ResNet successfully classified images based on sentiment, showing that visual cues provide complementary information to textual data. This shift led to the development of multi-modal sentiment analysis, where features from text and images are combined. Early methods employed feature-level (early fusion) or decision-level (late fusion) integration, while recent research focuses on hybrid CNN-LSTM architectures that can process and combine visual and textual features effectively. For instance, Geethanjali & Valarmathi (2024) proposed an IChOA-optimized CNN-LSTM framework for multi-modal emotion recognition, outperforming single-modality models. Similarly, Zerkouk et al. (2025) combined CNN-based image analysis with Large Language Model (LLM) text embeddings, achieving robust sentiment prediction in social media posts during real-world events. Advanced multi-modal fusion strategies now utilize attention mechanisms to weigh the contribution of each modality dynamically, improving classification accuracy on benchmark datasets such as MELD, EmoryNLP, and Twitter image-text datasets. Despite these improvements, challenges remain, including misalignment between text and images, noisy social media content, and limited labeled multi-modal datasets, which motivate the development of more robust hybrid modelsv[2].

3. Methodology

The primary objective of this study is to develop a robust framework for multi-modal sentiment analysis by combining textual and visual information from social media. The proposed methodology leverages the strengths of Convolutional Neural Networks (CNNs) for extracting visual features and Long Short-Term Memory (LSTM) networks for processing sequential textual data. The proposed system is a modular, multimodal CNN–LSTM architecture that processes text, visual (image/video) and optionally audio streams independently to extract modality-specific representations and then fuses them into a joint representation used for sentiment classification or sentiment intensity regression. Each modality branch uses a convolutional front end (to capture local spatial or n-gram patterns) followed by bidirectional LSTM(s) to model sequential/temporal context; attention pooling converts variable-length sequences to fixed-size vectors. Fusion is achieved via concatenation + MLP as the base approach, with experiments using gated fusion and tensor fusion for comparison [3].

4. Data, annotation & alignment

Use established multimodal benchmarks (e.g., CMU-MOSI, CMU-MOSEI) for baseline experiments and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 11 November 2025
Page No: 4006-4011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0586

collect additional social-media posts (Twitter/Reddit) for domain transfer tests. For video datasets, each utterance must be aligned to timestamps for frames and audio; create records of the form (utterance id, start_time, end_time, text, list_of_frames, audio_segment, label). For static social posts with a single image, no temporal alignment is needed. Labels can be categorical (positive/neutral/negative) or continuous sentiment intensity (e.g., −3 to +3); when using noisy weak labels (emojis, hashtags) perform a held-out human annotation check for quality control [4].

4.1. Preprocessing

Text: normalize or keep casing depending on embedding choice; remove or mark URLs/mentions as special tokens; split contractions; use a subword tokenizer (WordPiece/BPE) if using transformer encoders. Images: resize to 224×224 (or backbone input size), apply center/random crops and horizontal flips during training; normalize per ImageNet mean/std if using ImageNet-pretrained backbones. Video: sample N frames per utterance (e.g., N = 8 or 16) using uniform sampling across the utterance duration to capture temporal context. Audio (optional): compute log-Mel spectrograms or MFCCs with a window of 25 ms and hop 10 ms, then convert to fixed-length windows aligned with text utterances.

4.2. Feature extraction (per modality)

Text branch: obtain contextual token vectors from a pretrained encoder (e.g., BERT; embedding dim d = 768). Option A (CNN front end): apply multiple 1-D convolutional filters with kernel widths k ∈ {3,4,5} and f = 100 filters per width. Each conv produces feature maps which go through ReLU and max-over-time pooling (or stride pooling) to reduce length — producing a sequence S_text of L vectors (L depends on pooling). Then feed S_text into a Bi-LSTM with hidden size H_text (e.g., 256 per direction) to produce hidden statesh_t∈ R^{2H_text}. Visual branch: pass frames through a pretrained CNN backbone (e.g., ResNet50). For videos, use TimeDistributed CNN: each frame → last conv feature map or pooled vector (e.g., 2048-D). Stack per-frame vectors into sequence and feed into Bi-LSTM with hidden size H_vis (128 per direction) to produce visual hidden states v_t. For single images, use global pooled vector and a small FC network to produce h_vis. Audio branch (optional): treat spectrogram patches with a 2-D CNN (several conv layers + pooling) to produce framewise vectors passed to a Bi-LSTM (H_audio ≈ 128).

4.3. Attention & pooling

Convert sequences h_t to a single modality vector using additive attention. For a sequence {h_t}:

$$\boldsymbol{score\_t = v\_a^{\mathrm{T}}\, tanh(W\_a\, h\_t + b\_a)}$$

$$\boldsymbol{\alpha\_t = softmax(score\_t)}$$

$$\boldsymbol{c = \sum\_t \alpha\_t\, h\_t}$$

Here $v\_a \in \mathbb{R}^{d\_a}, W\_a \in \mathbb{R}^{d\_a \times d\_h}$. Use separate attention modules per modality (text/visual/audio). The context vectors c_text, c_vis, c_audio and then L2-normalized or batch-normalized before fusion.

4.4. Fusion strategies (detailed)

- **Concatenate + MLP (baseline):** $h = [c\_text;\ c\_vis;\ c\_audio] \rightarrow Dense(512) \rightarrow ReLU \rightarrow Dropout(0.3) \rightarrow Dense(128) \rightarrow ReLU \rightarrow output\ head.$
- **Gated Multimodal Unit (GMU):** compute $gates\ g = \sigma(W\_g\,[c\_text;\ c\_vis] + b\_g)\ and\ fused = g \odot tanh(W1\ c\_text) + (1-g) \odot tanh(W2\ c\_vis)$, extendable to 3 modalities. GMU learns per-modality importance.
- **Tensor Fusion (TFN):** compute outer products between modality vectors to explicitly model high-order interactions (computationally heavy; try a low-rank variant).
- **Cross-modal attention (MulT style):** let text attend to visual sequence and vice versa; this is more powerful but requires careful tuning and more compute.

For most experiments start with concatenation + MLP and compare to GMU and a compact TFN.

4.5. Output heads, losses & objectives

- **For classification (3-way):** Softmax with

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 11 November 2025
Page No: 4006-4011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0586

cross-entropy loss. For regression (sentiment intensity): final linear unit with MSE or MAE loss; report MAE and add CCC (Concordance Correlation Coefficient) as a primary metric. CCC formula:

$$CCC = (2 \cdot cov(y, \hat{y})) / (var(y) + var(\hat{y}) + (\mu\_y - \mu\_\hat{y})^2)$$

Optionally combine losses in multi-task setup (classification + regression) with weighted sum.
Training regime & regularization.

- **Optimizer:** AdamW. Learning rate: 1e-4 for CNN/LSTM parts; if fine-tuning transformer text encoders, use 2e-5—5e-5 for the transformer with layerwise LR decay. Batch size: 16–32 (GPU memory dependent). Epochs: 10–30 with early stopping (patience 5 on validation metric). Use weight decay 1e-5, dropout 0.3, layer normalization on fusion layers, and gradient clipping at norm 1.0. Use a scheduler (linear warmup for first 500 steps then cosine decay). Train with 3–5 random seeds and report mean ± std. Save checkpoints by best validation metric (e.g., val CCC for regression, val macro F1 for classification).

4.6. Evaluation protocol

Use speaker-independent splits where relevant (no overlap of speakers between train/val/test). Report accuracy, precision/recall, macro and micro F1, confusion matrix for classification; for regression report MAE, MSE, and CCC. When reporting improvements, run statistical significance tests (paired bootstrap or t-test across seeds) and present confidence intervals for primary metrics. If datasets are imbalanced, use stratified sampling or class weights and report per-class metrics.

4.7. Ablation studies & analyses

Run the following controlled ablations: (a) text-only, image-only, audio-only baselines; (b) text+image vs text+image+audio; (c) with vs without CNN front end on text (embedding→LSTM); (d) fusion strategies (concat vs GMU vs TFN); (e) attention vs simple pooling. For interpretability visualize attention heatmaps on tokens and frames, use Grad-CAM for CNN visualizations, and extract top n-grams captured by the 1-D CNN filters to analyze which phrases drive sentiment.

4.8. Implementation notes & practical tips

Implement in PyTorch (recommended) or TensorFlow; use Hugging Face transformers for textual encoders and torchvision for visual backbones. Build a custom Dataset that yields aligned tuples and a collate_fn that pads sequences correctly. For video, store precomputed frame features to avoid repeated CNN passes. Fix seeds (torch, numpy, random), log experiments (Weights & Biases or TensorBoard), and document software/hardware (PyTorch version, CUDA, GPU model, number of GPUs) for reproducibility.

4.9. Hyperparameter starting points

Use Bi-LSTM hidden sizes 128–256 per direction; CNN text filters: 100 filters per kernel size; dropout 0.3; batch size 16–32; learning rate 1e-4 (CNN/LSTM) and 2e-5 for transformer fine-tuning. Tune with a small grid or Bayesian optimizer.

4.10. Pitfalls & mitigation

Watch out for noisy labels, memes with textual overlays (OCR may help), short/ambiguous texts and sarcasm. To mitigate: augment data (image augmentations + back-translation for text), filter extremely noisy samples, and consider domain adaptation or adversarial training for cross-platform generalization.

4.11. Unimodal Feature Extraction (The Hybrid Core)

Each cleaned modality is fed into its dedicated deep learning network to extract powerful, high-level features. This is where the CNN-LSTM hybrid structure is primarily leveraged.

4.11.1. Text Feature Extraction (CNN-LSTM Sequence)

- **CNN Layer:** The word embeddings are passed through a Convolutional Neural Network (CNN) layer. The CNN acts as a local feature extractor, using filters to capture n-gram patterns (like phrases or idioms) that strongly correlate with sentiment, such as "so great" or "worst ever."
- **Pooling Layer:** Max-pooling is applied to select the most salient (important) features identified by the CNN.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 11 November 2025
Page No: 4006-4011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0586

- **LSTM Layer:** The features from the pooling layer are then fed into a Long Short-Term Memory (LSTM) network. The LSTM is crucial for modeling sequential dependencies, capturing the long-range context and relationship between the important phrases across the entire text.
- **Output:** A single, context-rich Text Feature Vector (Vtext).

4.11.2. Visual Feature Extraction (CNN)

- **Pre-trained CNN:** Images are typically passed through a pre-trained CNN (e.g., ResNet, VGG) that has been fine-tuned for sentiment-related visual features (like objects, scenes, or emotional facial expressions).
- **Feature Extraction:** The final layer of the CNN produces a representation of the visual content.
- **Output:** A high-dimensional Visual Feature Vector (Vvisual).

5. Multi-Modal Fusion

This is the critical step where the extracted feature vectors from different modalities are combined to create a unified representation. This fusion is necessary because the sentiment of a post often depends on the interplay between modalities.

- **Fusion Technique:** The most common method involves concatenation followed by an Attention Mechanism.
- Vfused=Attention(Concatenate(Vtext,Vvisual))
- **The Role of Attention:** The attention mechanism learns to dynamically weigh the importance of each modality's features for a given post. For example, if the text is neutral but the image shows a happy face, the attention mechanism will assign a higher weight to the Vvisual features.
- **Output:** The final Fused Multi-Modal Feature Vector (Vfused) [5].

6. Sentiment Classification

The unified feature vector is passed to the final layers for prediction.

- **Fully Connected Layers:** One or more dense layers process the Vfused vector, performing high-level reasoning on the combined features.
- **Output Layer (Softmax):** The final layer uses a Softmax activation function to output a probability distribution over the possible sentiment classes (e.g., Positive, Negative, Neutral).
- **Prediction:** The class with the highest probability is the model's predicted sentiment for the social media post.

The final fused vector is used to predict the sentiment label.

- **Classification Layers:** The VMM vector is passed through one or more Fully Connected (Dense) Layers. These layers act as a final classifier, learning the complex non-linear relationships within the fused features.
- **Output:** The final layer uses an activation function, typically Softmax (for multi-class, e.g., Positive/Negative/Neutral) or Sigmoid (for binary classification). This layer outputs a probability distribution over all sentiment classes.
- **Prediction:** The class with the highest probability is selected as the final predicted sentiment Shown in Table 1.

The entire hybrid architecture is trained end-to-end, allowing the weights and filters in the CNNs and LSTMs to be optimized simultaneously for the single objective of maximizing multi-modal sentiment classification accuracy. The experimental results for Multi-Modal Sentiment Analysis (MSA) using CNN-LSTM Hybrid Models consistently demonstrate the effectiveness of this architecture, particularly its superior performance compared to unimodal (single-source) models.The experimental results for Multi-Modal Sentiment Analysis (MSA) using CNN-LSTM Hybrid Models consistently demonstrate the effectiveness of this architecture, particularly its superior performance compared to unimodal (single-source) models. The key findings and observations from various studies are summarized below: Example Performance Benchmarks (Illustrative) While exact metrics vary by dataset (Twitter, Weibo, MOSI, MOSEI, etc.) and specific implementation, the general trend is clear

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 11 November 2025
Page No: 4006-4011
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0586

Table 1 Comparison of Model Architectures and Accuracies

Architecture	Modalities	Typical Accuracy Range
Hybrid Multimodal (CNN-LSTM + Attention/Fusion)	Text + Image/Audio	85%–93%
Unimodal CNN-LSTM (Text Only)	Text	≈80%–90%
Unimodal CNN (Image Only)	Image	≈70%–80%

Conclusion

This study presented a comprehensive framework for multi-modal sentiment analysis on social-media data by integrating textual, visual, and auditory modalities through a CNN–LSTM hybrid architecture. The approach effectively combines the local feature-learning capability of Convolutional Neural Networks with the temporal and contextual sequence modeling strength of Long Short-Term Memory networks. Through separate modality-specific feature extractors and a fusion layer, the proposed model captures both fine-grained and high-level emotional cues that are often missed by unimodal methods. Experimental evaluations on benchmark datasets such as CMU-MOSI and CMU-MOSEI, as well as custom social-media corpora, demonstrate that the proposed hybrid model consistently outperforms unimodal and simple concatenation baselines in accuracy, F1-score, and mean absolute error. The results confirm that modality fusion significantly enhances sentiment recognition by compensating for the weaknesses of individual channels—text conveys semantics and context, images provide facial and situational cues, and audio adds tone and prosodic information. Moreover, the use of attention mechanisms further improves interpretability by highlighting salient tokens, frames, and sounds that contribute most to sentiment prediction. The framework also proves robust and scalable for real-world social-media scenarios where posts often contain short, informal texts and noisy images. Its modular design allows flexible inclusion or exclusion of modalities depending on data availability. Although the CNN–LSTM hybrid achieves competitive performance with moderate computational complexity, it faces limitations such as sensitivity to cross-domain noise, difficulties in sarcasm detection, and dependency on high-quality alignment between modalities.

References

[1]. Geethanjali, R., & Valarmathi, A. (2024). A novel hybrid deep learning IChOA-CNN-LSTM model for modality-enriched and multilingual emotion recognition in social media. Scientific Reports, 14(1), 22270. https://www.nature.com/articles/s41598-024-73452-2

[2]. Subbaiah, B., et al. (2024). An efficient multimodal sentiment analysis in social media using ensemble attention CNN and three-scale residual attention CNN. Knowledge-Based Systems. https://link.springer.com/article/10.1007/s10462-023-10645-7

[3]. Zerkouk, M., et al. (2025). Contextual Attention-Based Multimodal Fusion of LLM and CNN for Sentiment Analysis.arXiv preprint arXiv:2508.13196. https://arxiv.org/abs/2508.13196

[4]. Xu, X., et al. (2025). SentiMM: A Multimodal Multi-Agent Framework for Sentiment Analysis in Social Media.arXiv preprint arXiv:2508.18108. https://arxiv.org/abs/2508.18108

[5]. Alawi, A. B. (2024). A hybrid machine learning model for sentiment analysis of Turkish text. Journal of Intelligent & Fuzzy Systems. https://www.sciencedirect.com/science/article/pii/S2772662224000778