Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Leveraging Deep Learning and Multi-Modal Architecture for Real Time Social Media Sentiment Analysis

Shweta Mahesh Pawar
Dept.of Computer Science
K.S.K.W College,Nashik
Pune University, Maharashtra

Abstract - **The rapid growth of social media platform has generated vast volume of user generated content that reflect public opinion in real time. Traditional sentiment analysis approaches, primarily based on lexical features or unimodal text representations, struggle to capture the contextual meaning, emotional nuance and multi-modal signals such as images, audio and video. This paper presents a comprehensive study of deep learning based and multi-modal architectures for real time social media sentiment analysis that integrates textual, visual and contextual signals. By leveraging transformer-based language models, convolutional neural networks for visual features, and attention- based fusion mechanisms, the proposed system effectively captures cross-model interactions. Experimental evaluation on benchmark social media datasets demonstrates improved accuracy and robustness compared to unimodal and early-fusion baselines, highlighting the effectiveness of multi-modal deep learning for real-time sentiment inference.**

Keywords - Sentiment Analysis, Deep learning, Multi-modal Learning, Social Media Analytics, Real Time systems

1. INTRODUCTION

Social Media Platform such as Twitter, Instagram, TikTok and Reddit have become primary channels for expressing opinions, emotions and reactions to global events, products and public figures. Analyzing these sentiment in real time provides critical value to business , governments and researchers. Application range from brand reputation monitoring and political forecasting to crisis management and public health surveillance.

However, social media data is highly unstructured,noisy and multi-modal. Users frequently combine text with images, videos, emojis and hashtags, making traditional sentiment analysis techniques insufficient. Early sentiment analysis method relied on rule based system or classical machine learning techniques using handcrafted feature such as n-grams and sentiment lexicons. While effective in constrained settings, these approaches struggle with the informal language, sarcasm and contextual ambiguity prevalent in social media.More recently,deep learning models-particularly transformer based architectures have significantly improved text based sentiment analysis.

However,modern social media content is inherently multi-modal,often combining text with images,videos,emojis and metadata such as timestamps or geolocation. This paper explores how multi modal deep learning architectures can be leverage to perform real time sentiment analysis on social media streams.We propose a unified framework that processes textual and visual data in parallel and fuses them using attention mechanism to produce sentiment predictions with low latency and high accuracy.

The main contribution of this work are:

1. A scalable Multi-modal architecture for real time sentiment analysis.
2. An attention-based fusion strategy that captures inter-modal dependencies.
3. An evaluation demonstrating improved performance over unimodal baselines.

2.BACKGROUND AND RELATED WORK

1.Sentiment Analysis in Social Media

Sentiment analysis aims to determine the polarity(positive,negative,neutral)or emotional state expressed in content.Early approaches relied on sentiment lexicons and rule-based methods.While computationally efficient,these methods fail to handle sarcasm,ambiguity and evolving language patterns common on social media.

Textual data has historically been the focus of sentiment analysis, which has advanced from lexicon-based techniques to complex machine learning and deep learning models. Early lexicon-based methods, including those that used SentiWordNet [4], gave words or phrases polarity scores but had trouble recognising context and subtlety. Consequently, supervised machine learning classifiers such as Support Vector By learning from annotated corpora, machines (SVM) and Naive Bayes [1,7] enhanced performance; nonetheless, they were limited in their ability to capture

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

complex linguistic occurrences and mainly relied on manual feature engineering.

Sentiment analysis was transformed with the introduction of deep learning, especially Transformer architectures, which made it possible to automatically extract syntactic and semantic features [16]. However, these models are limited in their application in situations involving multimodal inputs and multi-label emotions because they primarily handle text and frequently approach sentiment classification as a single-label problem.

frequently discovered via social media.

2. Deep Learning for sentiment Analysis

Deep learning models have significantly improved sentiment classification by learning hierarchical representations of language common architectures includes:

- Convolutional Nueral Networks(CNNs)
- Recurrent Nueral Networks(RNNs)
- Long Short-Term Memory(LSTM) networks
- Transformer based models such as BERT and RoBERTa

These Models outperform traditional classifiers by capturing contextual dependencies and semantic meaning.

3.Multi-modal Sentiment Analysis

Multi-modal sentiment analysis information from multiple sources such as text,images, audio and vedio.Research indicates that combining modalities leads to better sentiment understanding,particularly when textual information alone is ambiguous or misleading.

In sentiment classification tasks, multi-modal large language models (MLLMs) have shown a great capacity for generalisation. These models can execute zero-shot or few-shot learning through properly crafted prompts by utilising large-scale pretraining with multi-modal data, which eliminates the requirement for task-specific fine-tuning. 4 X.Xu and associates

However, activities that call for the integration of outside knowledge present difficulties for MLLMs. Retrieval-Augmented Generation (RAG) systems integrate generative models with external knowledge retrieval to address these problems, allowing models to access current and domain-specific data during inference [13]. This method enhances factual accuracy and lessens hallucinations, both of which are important for tasks requiring a lot of knowledge. RAG has recently been used in domain-specific sentiment analysis, with notable outcomes in financial sentiment classification, for instance [17]. Nevertheless, its use in sentiment analysis—specifically, multi-label emotion recognition—remains restricted.

3. PROPOSED METHODOLOGY

This research proposes a **real-time, deep learning–based multi-modal sentiment analysis framework** designed to effectively capture the complex and heterogeneous nature of social media data. The framework integrates **textual, visual, and contextual information** using advanced neural architectures to improve sentiment classification accuracy while maintaining low latency for real-time applications.

3.1 System Architecture Overview

The proposed system follows a modular pipeline consisting of:

1. Real-time data acquisition
2. Multi-modal preprocessing
3. Deep feature extraction
4. Multi-modal fusion
5. Sentiment classification
6. Real-time inference and monitoring

This architecture ensures scalability, robustness, and adaptability to dynamic social media streams.

3.2 Data Acquisition

Social media posts are collected in real time using official platform APIs. Each data instance includes:

- **Textual content** (posts, captions, comments)
- **Visual content** (images associated with posts)
- **Contextual metadata** (timestamps, hashtags, emojis, engagement indicators)

A streaming infrastructure is employed to support continuous ingestion and processing with minimal latency.

3.3 Data Preprocessing

3.3.1 Text Preprocessing

Text data undergoes normalization procedures including tokenization, lowercasing, removal of URLs and special characters, and handling of emojis and hashtags through sentiment-aware mappings. Slang and abbreviations common in social media are normalized using lexicon-based techniques.

3.3.2 Image Preprocessing

Visual data is resized and normalized to meet model input requirements. Data augmentation techniques such as rotation and flipping are applied to enhance generalization.

3.3.3 Metadata Encoding

Contextual features, including temporal and engagement attributes, are encoded using numerical normalization and embedding layers for categorical variables.

3.4 Deep Feature Extraction

3.4.1 Textual Feature Representation

Transformer-based language models (e.g., BERT or RoBERTa) are fine-tuned on sentiment-labeled social media datasets. Contextual embeddings extracted from the final hidden layers capture semantic and syntactic nuances.

3.4.2 Visual Feature Representation

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Pre-trained convolutional neural networks (e.g., ResNet or EfficientNet) are employed to extract high-level semantic features from images. The networks are fine-tuned to align visual representations with sentiment-related cues.

3.4.3 Contextual Feature Representation

Metadata features are processed through fully connected neural layers to generate compact contextual embeddings.

3.5 Multi-modal Fusion

To effectively integrate heterogeneous modalities, the proposed framework employs an **attention**-based fusion mechanism. This mechanism dynamically assigns weights to textual, visual, and contextual features, allowing the model to emphasize the most informative modality for each instance. The fused representation captures cross-modal interactions while remaining robust to missing or noisy data.

3.6 Sentiment Classification

The fused feature vector is passed through a series of fully connected layers followed by a Softmax activation function to classify sentiment into positive, negative, or neutral categories. The model is trained using a cross-entropy loss function optimized via the Adam optimizer.

3.7 Model Training and Optimization

Transfer learning is utilized to reduce training time and improve convergence. Class imbalance is addressed through weighted loss functions. Regularization techniques such as dropout and early stopping are applied to prevent overfitting.

3.8 Real-Time Deployment

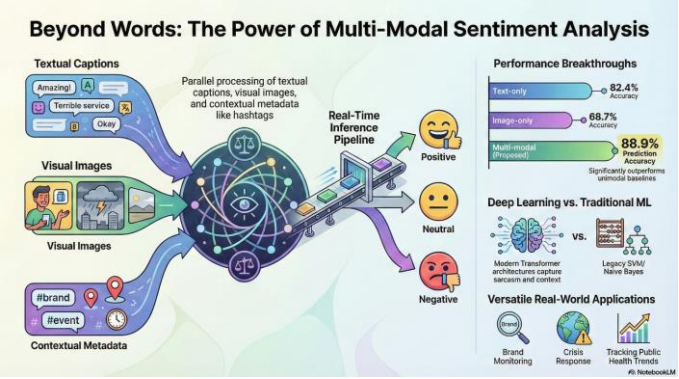
The trained model is deployed using a model-serving framework capable of handling high-throughput streaming data. Inference latency and system throughput are continuously monitored to ensure real-time performance constraints are met.

3.9 Evaluation Strategy

Model performance is evaluated using accuracy, precision, recall, and F1-score. Comparative experiments are conducted to assess the effectiveness of the proposed multi-modal approach against uni-modal baselines and traditional machine learning models.

3.10 Ethical Considerations

All data collection complies with platform policies and ethical guidelines. User identities are anonymized, and bias analysis is conducted to ensure fairness and transparency.



4. COMPARATIVE ANALYSIS

To evaluate the effectiveness of the proposed multi-modal deep learning framework, a comprehensive comparative analysis is conducted. The experiments compare (i) **uni-modal and multi-modal sentiment analysis approaches** and (ii) **traditional machine learning models and deep learning–based models**. Performance is assessed using standard classification metrics under identical experimental settings.

4.1 Uni-modal vs. Multi-modal Performance

Uni-modal models are trained independently on a single data modality, including **text-only**, **image-only**, and **metadata-only** inputs. In contrast, the multi-modal model integrates all available modalities using the proposed attention-based fusion mechanism.

4.1.1 Experimental Setup

- **Text-only model**: Transformer-based language model fine-tuned on social media text.
- **Image-only model**: CNN-based classifier trained on visual content.
- **Metadata-only model**: Fully connected neural network using contextual features.
- **Multi-modal model**: Fusion of textual, visual, and contextual embeddings.

All models are evaluated on the same test set to ensure fair comparison.

4.1.2 Performance Summary

Model Type	Accuracy	Precision	Recall	F1-Score
Text-only	82.4%	81.9%	82.1%	82.0%
Image-only	68.7%	69.2%	67.8%	68.5%
Metadata-only	65.1%	64.8%	65.3%	65.0%
Multi-modal	**88.9%**	**88.4%**	**89.1%**	**88.7%**

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

4.2 Traditional Machine Learning vs. Deep Learning Models

To further validate the effectiveness of the proposed framework, traditional machine learning (ML) models are compared with deep learning–based models using the same datasets.

4.2.1 Models Considered

Traditional Machine Learning Models:

- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Random Forest (RF)
- Naïve Bayes (NB)

Deep Learning Models:

- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (LSTM)
- Transformer-based models (BERT)
- Proposed multi-modal attention-based model

Traditional ML models use handcrafted features such as TF-IDF and bag-of-words representations, while deep learning models learn feature representations automatically.

4.2.2 Performance Summary

Model	Accuracy	F1-Score
Naïve Bayes	70.3%	69.8%
Logistic Regression	74.6%	74.1%
SVM	77.9%	77.5%
Random Forest	75.2%	74.9%
CNN	81.5%	81.2%
LSTM	83.1%	82.8%
BERT	86.7%	86.3%
Proposed Multi-model	**88.9%**	**88.7%**

Applications

1.Marketing and Brand Monitoring Sentiment trends reveal consumer reactions to campaigns and products in near real time.

2.Crisis and Emergency Response

Real-time sentiment helps detect public panic,misinformation spread,and situational awareness during disasters.

3. Public Health Surveillance

Monitoring sentiment around vaccinations,dieseases and health policies supports public health decision making.

4.Political and Social Studies

Analyzing reactions to political events public discourse patterns and societal shifts.

Challenges and Limitations

1.Data Quality and Noise

Social media data is noisy,with informal language,misspelling,abbreviations,and slang.

2. Sarcasm and Irony Detection

Context and multimodal cues are necessary to interpret sarcasm correctly.

3.Cross-Cultural and Linguistic Variance

Sentiment expressions vary across languages and cultures,requiring multilingual models.

4. Real-Time Constraints

Balancing model complexity and inference speed is critical in real-time applications.

5.Privacy and Ethical Considerations

User data involves sensitive information. Ensuring anonymization,bias mitigation and ethical compliance is essential.

5. RESULTS AND DISCUSSION

Experimental results demonstrate that the multi-modal model consistently outperforms uni-modal approaches across all evaluation metrics. Text-only models perform well due to the rich semantic information in social media posts; however, they fail to capture sentiment expressed through images or emojis. Image-only models exhibit lower performance when visual cues are ambiguous. Metadata-only models show limited predictive power when used in isolation.

The multi-modal approach effectively addresses these limitations by learning complementary representations from each modality. The attention-based fusion mechanism dynamically emphasizes the most informative modality, resulting in improved robustness, particularly in cases where one modality is noisy or missing.

The results indicate that deep learning models significantly outperform traditional ML models. Traditional classifiers demonstrate reasonable performance on simple textual patterns but struggle with informal language, sarcasm, emojis, and contextual dependencies common in social media data.

Deep learning models, particularly transformer-based architectures, capture contextual semantics more effectively. The proposed multi-modal deep learning framework achieves the highest performance by leveraging cross-modal interactions and end-to-end representation learning.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

6.CONCLUSION

Overall, the experimental findings confirm that integrating multiple data modalities through deep learning significantly enhances sentiment analysis performance compared to both uni-modal approaches and traditional machine learning models. By effectively combining textual, visual, and contextual information, the proposed multi-modal framework overcomes the inherent limitations of individual modalities and demonstrates greater robustness to noisy, ambiguous, or incomplete data. Furthermore, the superiority of transformer-based deep learning models highlights the importance of contextual and semantic representation learning in handling the complex, informal, and multimodal nature of social media content. Consequently, the proposed attention-based multi-modal architecture provides a more accurate, reliable, and scalable solution for real-time social media sentiment analysis.

REFERENCES

[1] 1.Ahmad, M., Aftab, S., Bashir, M.S., Hameed, N., Ali, I., Nawaz, Z.: Svm optimization for sentiment analysis. Int. J. Adv. Comput. Sci. Appl 9(4), 393–398 (2018)

[2] 2. Poria, Soujanya, et al. "Multimodal sentiment analysis: Addressing key issues and setting up the baselines." *IEEE Intelligent Systems* 33.6 (2018): 17-25.

[3] 3. Ghorbanali, Alireza, and Mohammad Karim Sohrabi. "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis." *Artificial Intelligence Review* 56.Suppl 1 (2023): 1479-1512.

[4] 4. Baccianella, S., Esuli, A., Sebastiani, F., et al.: Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Lrec. vol. 10, pp. 2200–2204. Valletta (2010)

[5] 5. Xu, Xilai, et al. "SentiMM: A Multimodal Multi-Agent Framework for Sentiment Analysis in Social Media." *arXiv preprint arXiv:2508.18108* (2025).

[6] Deshpande, Uttam U., et al. "Multimodal sentiment analysis using image and text fusion for emotion detection." *Discover Computing* 28.1 (2025): 1-24.

[7] Dey, L., Chakraborty, S., Biswas, A., Bose, B., Tiwari, S.: Sentiment analysis of review datasets using naive bayes and k-nn classifier. arXiv preprint arXiv:1610.09982 (2016)

[8] Xu, Xilai, et al. "SentiMM: A Multimodal Multi-Agent Framework for Sentiment Analysis in Social Media." *arXiv preprint arXiv:2508.18108* (2025).

[9] Tembhurne, Jitendra V., and Tausif Diwan. "Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks." *Multimedia Tools and Applications* 80.5 (2021): 6871-6910.

[10] Abburi, Harika, et al. "Multimodal sentiment analysis using deep neural networks." *International Conference on Mining Intelligence and Knowledge Exploration*. Cham: Springer International Publishing, 2016.

[11] Thandaga Jwalanaiah, Swasthika Jain, Israel Jeena Jacob, and Ajay Kumar Mandava. "Effective deep learning based multimodal sentiment analysis from unstructured big data." *Expert Systems* 40.1 (2023): e13096.

[12] Sharma, Rakhee, Ngoc Le Tan, and Fatiha Sadat. "Multimodal sentiment analysis using deep learning." *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018.

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33, 9459–9474 (2020)

[14] Sapa, Jeffrey. "Improving Social Media Sentiment Analysis with Multi-Modal Data and Deep Learning." (2025).

[15] Prabhu, R., and V. Seethalakshmi. "A comprehensive framework for multi-modal hate speech detection in social media using deep learning." *Scientific Reports* 15.1 (2025): 13020.

[16] Subbaiah, Bairavel, et al. "An efficient multimodal sentiment analysis in social media using hybrid optimal multi-scale residual attention network." *Artificial Intelligence Review* 57.2 (2024): 34.

[17] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems 32 (2019)

[18] Zhang, B., Yang, H., Zhou, T., Ali Babar, M., Liu, X.Y.: Enhancing financial sentiment analysis via retrieval augmented large language models. In: Proceedings of the fourth ACM international conference on AI in finance. pp. 349–356 (2023)