A Hybrid LSTM-CNN Approach for Multimodal Sentiment Analysis: Combining Text and Image Features

Zannirah Muhammed Sammani
College of Engineering and Technology, Arab Academy for Science Technology and Maritime Transport, Cairo, Egypt

Mohammed Abo Rizka
Professor and Dean of Faculty of Computers, and Information, Arab Academy for Science & Technology, Cairo, Egypt

ABSTRACT

An efficient deep learning framework is proposed for sentiment analysis that leverages both textual and visual modalities. The architecture integrates Long Short-Term Memory (LSTM) networks for capturing sequential dependencies in textual data with Convolutional Neural Networks (CNNs) for analyzing visual content. This multimodal fusion enhances sentiment classification accuracy. The model is assessed on two benchmark datasets—Memes and MVSA—and its performance is compared to traditional machine learning models such as Support Vector Machines and Logistic Regression, as well as the transformer-based VisualBERT. Although VisualBERT achieves slightly higher accuracy (83.18% on Memes and 81.29% on MVSA), the proposed approach delivers comparable results (77.70% and 80.42%, respectively) while maintaining a much lower computational footprint. This balance between performance and efficiency highlights the model's practical value for applications where computational resources are limited or real-time analysis is required.

General Terms

Machine Learning, Deep Learning, Natural Language Processing, Computer Vision, Multimodal Analysis, Sentiment Classification, Algorithms

Keywords

Convolutional Neural Networks (CNNs), Deep Learning Hybrid Models, Long Short- Term Memory (LSTM), Multimodal Sentiment Analysis

1. INTRODUCTION

The widespread exchange of updates on major social networking platforms has become a central part of modern digital communication, enabling users to share information, opinions, and emotions globally [1]. Sentiment analysis, a key area within natural language processing (NLP), aims to identify and categorize the emotional tone behind user-generated content—spanning both textual and visual modalities [1], [2]. This domain focuses on analyzing large-scale content from platforms such as Twitter, Facebook, Instagram, and Flickr [3].

Traditional sentiment analysis has primarily relied on textual data. However, recent advancements highlight the importance of multimodal approaches, where the fusion of textual and visual features significantly enhances classification performance compared to single-modality models [4]. Given the complex and subjective nature of emotions—ranging from positive, negative, and neutral to more nuanced states like joy or sarcasm—leveraging diverse data sources is essential for building accurate and robust sentiment analysis systems [5].

Historically, sentiment analysis employed statistical and machine learning techniques that depended heavily on handcrafted features, limiting their adaptability and scalability. In contrast, modern deep learning methods—particularly neural network architectures—have demonstrated superior performance by learning high-level features automatically [6, 7].

The present study proposes a lightweight hybrid deep learning model that combines Long Short-Term Memory (LSTM) networks for textual analysis with Convolutional Neural Networks (CNNs) for image-based sentiment understanding [8]. The goal is to fuse temporal and spatial information effectively to improve sentiment classification accuracy. the model is evaluated on two widely used benchmark datasets: MVSA and Memes. It is compared against traditional machine learning models—such as Support Vector Machines (SVM) and Logistic Regression—as well as the transformer-based VisualBERT model.

While VisualBERT achieves the highest accuracy (83.18% on Memes and 81.29% on MVSA), hybrid model offers a competitive alternative (77.70% and 80.42%, respectively), with significantly lower computational requirements. This makes it well-suited for real-time or resource-constrained environments. The results demonstrate the limitations of unimodal approaches and reinforce the effectiveness of multimodal fusion in sentiment analysis.

Overall, this work contributes both theoretically and practically to the field by developing and evaluating a hybrid LSTM-CNN framework that can enhance the interpretation of sentiment in social media content. VisualBERT is also assessed to provide a deeper comparative understanding of current multimodal architectures.

2. RELATED WORK

2.1 Sentiment Analysis for Text

Sentiment analysis is key to understanding public and customer opinions, using three main approaches: sentiment lexicons, machine learning, and deep learning [9]. Supervised ML models like Linear SVM and Logistic Regression perform well in classification tasks [10] but struggle with domain transferability and require manual annotation. Deep learning addresses these issues by learning complex features automatically through neural networks such as RNNs and CNNs. Attention mechanisms enhance emotional cue detection, while LSTM networks excel in capturing long-term dependencies in text, making them highly effective for sentiment classification [11].

2.2 Sentiment Analysis for Images

Visual sentiment analysis studies emotional responses elicited by visual cues, posing unique challenges due to the subjective nature of emotions [12]. Deep learning, especially CNNs, has revolutionized computer vision tasks by improving feature extraction and computational efficiency through convolutional, pooling, and normalization layers [13]. CNNs are widely applied to scene understanding, object recognition, and image-based sentiment prediction [14].

Several CNN-based approaches have demonstrated significant improvements in representing visual sentiment. For example, DeepSentiBank uses adjective-noun pairs for emotion classification, while fine-tuned CNNs trained on large-scale datasets achieve superior emotion prediction performance. Architectures such as PCNN effectively leverage noisy web data for sentiment tasks [15]. Despite these advances, challenges such as sentiment ambiguity and category overlap remain. To mitigate these issues, hybrid models combining CNNs with RNNs have been proposed to capture multi-level features [16]. CNNs generally require fewer parameters than fully connected networks, resulting in efficient training while maintaining high accuracy [17].

2.3 Multimodal Text and Image Sentiment Analysis

Multimodal Sentiment Analysis (MSA) involves the combination of information from various sources—such as text and images—to assess emotions and sentiments. It is widely used in applications like personalized advertising, opinion analysis, emotion-aware recommendation engines, and human-computer interaction [18]. The rapid increase in user-generated content on social media has positioned MSA as an important area of research [19].

MSA presents several challenges, including the creation of effective multimodal representations, alignment of modalities both in time and meaning, and reliable fusion of diverse data types. These tasks are further complicated by issues such as asynchronous inputs, varying data quality, and modality-specific noise [20]. To tackle these problems, deep learning models—such as CNNs, LSTMs, and transformers enhanced with attention mechanisms—have been widely adopted.

Two prominent difficulties in MSA are the semantic gap, referring to the disconnect between raw input features (like pixels) and abstract sentiment concepts (like emotions), and data fusion, which involves integrating complementary features from different modalities effectively [21]. As multimodal systems become more sophisticated, ensuring interpretability and explainability is essential for understanding the individual impact of each modality, thereby enabling more transparent and dependable predictions [22].

2.4 Multimodal Fusion for Hybrid LSTM-CNN Models

Recent research in multimodal fusion for emotion recognition has explored various strategies to combine complementary data from different modalities. Fusion methods are generally classified as early fusion (feature-level), late fusion (decision-level), or hybrid fusion. Hybrid models that combine LSTM and CNN architectures are particularly promising for text and image sentiment analysis, as they exploit LSTMs' sequential modeling strength for text and CNNs' spatial feature extraction for images [9].

Several studies propose fusion mechanisms that concatenate LSTM-derived textual features with CNN-extracted visual features, which are then fed into dense or attention layers for sentiment prediction [23]. More advanced methods employ gating mechanisms or attention-based fusion, allowing the model to dynamically weight and integrate features based on their relevance to sentiment interpretation. For example, Gated Multimodal Units (GMUs) utilize gating units to control the contribution of each modality via trainable parameters [24].

In contrast, transformer-based models like VisualBERT directly integrate text and image inputs within a unified architecture by embedding image features as special tokens alongside tokenized text. Self-attention layers in these models learn cross-modal dependencies effectively [25, 26]. Although powerful, transformer-based models typically require extensive computational resources and large-scale pretraining datasets [27].

The study proposes a lightweight hybrid LSTM-CNN architecture as a practical and interpretable alternative to resource-intensive transformer models. While it does not surpass large-scale models like VisualBERT in accuracy, it offers a favorable balance between performance, computational efficiency, and deployment ease. This makes it particularly suitable for real-world applications such as content moderation, sentiment monitoring, and user feedback analysis on resource-constrained platforms including mobile and embedded devices [28].

3. METHODOLOGIES

3.1 Dataset

The datasets used in this study are sourced from Kaggle [29] and include internet memes as well as the MVSA dataset, both containing images paired with corresponding textual content [30]. Prior research has examined image color palettes and the emotional tone of associated text [31]. features are extracted to classify memes into three sentiment categories: positive, negative, and neutral. Negative memes typically express emotions such as sadness, anger, or disgust, while positive memes convey happiness or surprise. Neutral memes exhibit minimal emotional expression [30].

3.2 Adversarial Robustness and Overfitting

To address challenges related to overfitting and robustness, this study examines training on small-scale benchmark datasets. A key limitation observed is that although models may achieve high accuracy on standard test sets, their performance often degrades when exposed to adversarial examples. Achieving robust generalization usual requires substantially more training data than typical procedures provide [32]. Additionally, training with large batch sizes can lead to a generalization gap, where the model performs well on training data but poorly on unseen samples. The random walk on a random landscape framework is employed to describe the stochastic evolution of model parameters during early training phases [33].

3.3 Text Preprocessing and Feature Extraction

Text preprocessing transforms raw textual data into analyzable forms, a critical step in natural language processing (NLP) [34]. The following techniques are applied:

3.3.1 Tokenization: Text is segmented into tokens using punctuation and non-alphabetic characters as delimiters.

3.3.2 Stop-word Filtering: Commonly occurring words and tokens based on predefined length constraints are removed.

3.3.3 Stemming: Words are reduced to their root forms using algorithms such as Porter, Lovin's, and the Snowball stemmer, which implements 41 rule-based transformations [35].

3.3.4 Noise Removal: Eliminates punctuation, Twitter symbols, and HTML tags [36].

The cleaned text (TPT) is then used to extract features via the Term Frequency–Inverse Document Frequency (TF-IDF) method:

3.3.5 Term Frequency (TF): Measures how frequently a word appears in a document.

Inverse Document Frequency (IDF): Applies a logarithmic transformation to assess term rarity across the corpus [37].

3.4 Image Processing

Image classification and segmentation use descriptors like texture, color, edge maps, HOG, and GIST [38]. In a related study, twelve 64×64 sub-images were extracted from brain scans and labeled for Invasive Ductal Carcinoma detection [39, 40]. For meme analysis, OCR is used to extract embedded text from images [41].

3.5 Multimodal Deep Learning Architecture

To combine heterogeneous data sources for sentiment classification, multimodal deep learning framework ís proposed integrating textual and visual information via a hybrid LSTM-CNN architecture enhanced by gated fusion and cross-modal attention mechanisms.

3.5.1 Text Branch:

A Long Short-Term Memory LSTM recurrent network processes preprocessed text inputs, capturing temporal and contextual relationships within token sequences to generate a fixed-size semantic vector [42]. Text is tokenized, embedded into 100-dimensional vectors, and passed through a 128-unit LSTM, followed by batch normalization and dropout 0.5 to enhance generalization.

3.5.2 Image Branch:

A Convolutional Neural Network CNN extracts spatial features from resized meme images, capturing texture, color, and sentiment cues relevant to classification [43]. Images pass through two convolutional layers (with 32 and 64 filters), each followed by max pooling, then flattening and dropout, before a dense layer project them into a 128-dimensional space.

3.5.3 Fusion Mechanism:

A gated fusion module combines modalities by learning the importance of each. Inspired by Gated Multimodal Units (GMUs) [24], it uses trainable gating parameters to weight visual and textual features. Both modalities are projected into a shared semantic space. Global pooling is applied to text outputs, and image features are reshaped. These are concatenated and passed through a sigmoid-activated gating layer to dynamically balance their contributions.

3.5.4 Cross-Modal Attention:

A multi-head bidirectional attention layer [43] enables image and text features to attend to one another. This two-way interaction highlights sentiment-relevant cues and models fine-grained inter-modal relationships. Residual connections further enhance these features with cross-modal context.

3.5.5 Training Enhancements:

The model is trained for 10 epochs with a batch size of 16 using the Adam optimizer and categorical cross-entropy loss, with validation on a separate test set. To enhance robustness and avoid overfitting—especially on small meme-based datasets—techniques such as early stopping, learning rate scheduling, dropout, L2 regularization, and architectural constraints are applied [44]. Is there any spelling error in this section

3.6 Transformer-Based

An alternative model variant integrating transformer-based components:

3.6.1 Text Tokenization: The BERT tokenizer preprocesses text inputs.

3.6.2 Image Features: Pre-trained ResNet-50 with frozen convolutional layers extracts 1024-dimensional visual features.

3.6.3 Cross-Modal Transformer: These features are embedded as tokens within a VisualBERT architecture, enabling self-attention mechanisms to learn semantic relationships across visual and textual modalities for sentiment classification [45].

While VisualBERT demonstrates strong performance, its reliance on extensive pretraining and significant computational resources makes it less suitable for real-time applications [26]. In contrast, the hybrid LSTM-CNN model offers a balanced trade-off between accuracy and efficiency. This makes it particularly useful for resource-constrained environments, such as mobile applications, content moderation systems, or embedded sentiment tracking tools, where low latency and computational efficiency are critical.

including both architectures allows to evaluate trade-offs between model complexity and performance in real-world applications.

4. TECHNIQUES

4.1 Traditional Machine Learning Models

Support Vector Machine (SVM) is a linear classifier that minimizes errors while maximizing the margin between classes, making it a maximum margin classifier. It projects data into a higher-dimensional space to identify the optimal separating hyperplane, ensuring clear class separation [46].

Logistic Regression is a commonly used binary classification model that estimates class probabilities using the logistic function. It is valued for its simplicity and interpretability but may struggle with complex language features like sarcasm or idioms due to its linear assumptions [47].

4.2 Deep Learning Models

4.2.1 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are an advanced type of recurrent neural network (RNN) specifically developed to model long-term dependencies in sequential data, such as natural language or time series. To use LSTMs effectively, the input text must undergo preprocessing steps like cleaning, tokenization, and word embedding to transform it into numerical vectors suitable for the model. LSTMs are particularly effective at maintaining information over long sequences while addressing the vanishing gradient issue commonly found in standard RNNs [48, 49].

4.2.2 Convolutional Neural Networks (CNNs)

CNN have become dominant in the field of computer vision due to their capacity to extract complex spatial features. A typical CNN architecture includes an input layer, multiple convolutional and pooling layers, normalization, and one or more fully connected layers. CNNs are highly effective for structured data prediction tasks and are well-optimized for operations involving matrices and vectors [50, 51].

4.3 Hybrid Models

Hybrid models that integrate CNNs and LSTMs leverage the strengths of both architectures. CNNs are efficient for identifying local patterns in text, while LSTMs capture sequential dependencies. This synergy is particularly useful for sentiment classification and depression detection in social media posts [52].

In the proposed CNN–LSTM architecture, CNN layers first extract spatial features, followed by MaxPooling and a Flatten layer to reshape outputs for LSTM input. The LSTM then processes temporal relationships across sequences. To prevent overfitting—a common deep learning challenge—Dropout layers are used to randomly disable neurons during training. The final classification is handled by a fully connected (FC) layer [49, 53, 54].

VisualBERT is a transformer-based model that combines BERT (for text) and Faster R-CNN (for images). It treats object proposals as pseudo-tokens and integrates them with textual input into a unified transformer pipeline. Pre-training is conducted on image-caption datasets using masked language modeling and text-image matching tasks. VisualBERT excels at identifying nuanced sentiment and offensive content in memes by jointly processing both modalities [55, 56].

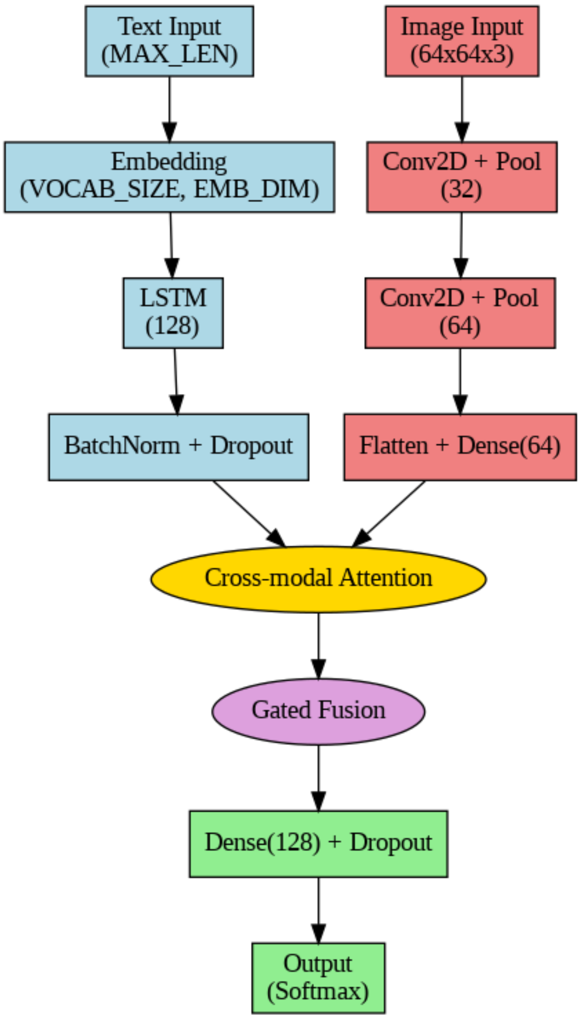

Figure 1 CNN–LSTM Architecture

4.4 Evaluation Metrics

To assess model performance, several metrics are employed:

Accuracy: Measures the ratio of correctly predicted instances to the total instances [57].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+TN} \quad (1)$$

$$Weighted\text{- } Accuracy = \frac{1}{N}\sum_{i=1}^{N} Wi \cdot \frac{Tp_i+TN_i}{TP_i+TN_i+FP_I+TN_I} \quad (2)$$

Weighted Accuracy: Adjusts for class imbalances by assigning weights Wi to each class:

Precision: Proportion of true positive predictions among all positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Recall (Sensitivity): Measures how well the model identifies actual positive instances.

$$\text{Recall} = \frac{TP}{TP+FP} \quad (4)$$

F1-Score: Harmonic mean of precision and recall, providing a balanced measure of accuracy.

$$\text{F1-Score} = 2 \cdot \frac{precision_i \cdot Recall_i}{precision_i \cdot Recall_i} \quad (5)$$

Weighted F1-Score: Accounts for class imbalance by weighting individual F1-scores:

$$\text{Weighted-F1-score} = \frac{1}{N}\sum_{i=1}^{N} Wi \cdot 2 \cdot \frac{precision_i \cdot Recall_i}{precision_i \cdot Recall_i} \quad (6)$$

5. EXPERIMENT & EVALUATION AND RESULTS

This section outlines the datasets used and the experimental configuration employed for model training and evaluation.

5.1 Datasets

This study utilized two benchmark multimodal datasets: the Memes dataset and the MVSA dataset. Both comprise paired image and text data annotated for sentiment classification into three categories—positive, negative, and neutral. However, each dataset exhibits class imbalance, which was appropriately addressed during model training.

5.1.1 Memes Dataset

The Memes dataset comprises approximately 6,992 valid samples, each consisting of a paired image and its corresponding textual caption. Initially, missing text entries (NaN values) were imputed using corrupted or truncated image files were excluded. After preprocessing, a one-to-one alignment was ensured between the text and image components to maintain modality consistency.

5.1.2 MVSA Dataset

The MVSA dataset includes approximately 20,000 samples in total, each consisting of a paired image and its corresponding text. While the dataset contains three sentiment classes (positive, negative, and neutral), the distribution across these categories is imbalanced. This class imbalance was addressed during training using weighted loss functions.

5.1.3 Tools and Libraries

The experiments were implemented using Python and various deep learning and NLP libraries, including:

Pandas for data handling, NLTK for text preprocessing, TensorFlow and Keras for LSTM-based models, PyTorch and TorchVision for CNN and VisualBERT models, Scikit-learn for traditional machine learning classifiers transformers by hugging face for implementing visualbert

5.2 Experimental Setup

All datasets were split into training and testing sets using an 80:20 ratio. This split was consistently applied across all experiments involving text-only, image-only, and multimodal text image inputs. Model performance was evaluated using standard classification metrics including accuracy, precision, recall, and F1-score.

5.3 EVALUATION CONFIGURATIONS AND RESULTS

Table 1 Traditional Machine Learning Models

Aspect Feature	SVM	Logistic Regression
Model Type	Classical ML	Classical ML
Input Modalities	Text + Image	Text + Image
Text Representation	TF-IDF (OCR + Text)	TF-IDF (OCR + Text)
Image Representation	Flattened Pixels	Flattened Pixels
Fusion Strategy	Manual Feature Fusion	Manual Feature Fusion
Regularization	Standardization	Standardization
Data Augmentation	None	None
Memes Accuracy	66.82%	70.93%
MVSA Accuracy	61.45%	70.91%

Table 2 VisualBERT Variants Comparison

Aspect Feature	VisualBERT (Text+Image)	VisualBERT (Text Only)	VisualBERT (Image Only)
Model Type	Transformer (Multimodal)	Transformer (Text)	Transformer (Image)
Input Modalities	Text + Image	Text only	Image only
Text Representation	BERT embeddings	BERT embeddings	Dummy input
Image Representation	ResNet-50 (frozen)	None	ResNet-50 (frozen)
Fusion Strategy	Early Fusion (Token-Level)	None	Dummy Fusion
Classifier Head	Linear (768 → 3)	Linear (768 → 3)	Linear (768 → 3)
Regularization	Dropout 0.5	Dropout 0.3	Dropout 0.4
Data Augmentation	None	None	None
Optimizer	AdamW (LR = 3e-5)	AdamW (LR = 3e-5)	AdamW (LR = 3e-5)
Epochs	10	10	10
Early Stopping	Patience = 3	Patience = 3	Patience = 3
Memes Accuracy	83.18%	78.52%	60.48%
MVSA Accuracy	81.29%	81.12%	0.6048

Table 3 Deep Learning Models (LSTM/CNN)

Aspect Feature	Top	In-between	Bottom
Model Type	Hybrid DL	RNN-based	CNN-based
Input Modalities	Text + Image	Text only	Image only
Text Representation	LSTM Embeddings	TF-IDF	None
Image Representation	CNN (custom)	None	ResNet-50
Fusion Strategy	Feature Concatenation	N/A	N/A
Classifier Head	Attention + FC	Dense Layers	FC Layer
Regularization	Dropout 0.5	Dropout 0.5	Dropout 0.4
Data Augmentation	None	None	Horizontal Flip
Optimizer	ADAM	ADAM	ADAMW
Epochs	10	10	10
Early Stopping	Patience = 3	Patience = 3	Patience = 3
Memes Accuracy	77.70%	76.97%	53.38%
MVSA Accuracy	80.42%	78.39%	47.50%

Table 4 Classification Report

DATASET	Model	INPuT TYPE	ACCURACY	F1-NEG	F1-NEU	F1-POS	MACRO F1
MEMES	LSTM +CNN	Text+Image	0.78	0.63	0.83	0.71	0.73
MEMES	LSTM	Text only	0.77	0.54	0.84	0.65	0.68
MEMES	CNN	Image only	0.53	0.17	0.69	0.24	0.37

MEMES	Visual BERT	Text Image	0.83	0.70	0.87	0.81	0.79
MEMES	Visual BERT	Text only	0.79	0.65	0.83	0.75	0.75
MEMES	Visual BERT	Image only	0.67	0.00	0.80	0.00	0.27
MVSA	LSTM +CNN	Text. Image	0.80	0.62	0.65	0.90	0.72
MVSA	LSTM	Text. only	0.78	0..52	0.67	0.88	0.69
MVSA	CNN	Image only	0.47	0.06	0.39	0.59	0.35
MVSA	Visual BERT	Text Image	0.81	0.63	0.70	0.90	0.75
MVSA	Visual BERT	Text only	0.81	0.63	0.69	0.91	0.74
MVSA	Visual BERT	Image only	0.60	0.00	0.00	0.75	0.25

6. CONCLUSION AND FUTURE WORK

This study presented a comprehensive evaluation of sentiment classification using text, image, and multimodal inputs. A lightweight hybrid model combining LSTM for text and CNN for image features was proposed and benchmarked against traditional classifiers and the transformer-based Visualbert model.

While VisualBERT achieved the highest accuracy on both the Memes (83.18%) and MVSA (81.29%) datasets, the proposed hybrid LSTM-CNN model delivered competitive performance77.70% and 80.42%, respectively at significantly lower computational cost. This makes it a practical option for deployment in real-time or resource-constrained environments, such as mobile applications or content moderation platforms.

Key observations include the relatively strong performance of text-only models, particularly on the MVSA dataset, indicating that textual features often carry the bulk of sentiment-related information. In contrast, image-only models performed poorly, highlighting the limited standalone utility of visual cues for sentiment analysis. Neutral sentiment classification also remains a challenge, primarily due to its subtle and ambiguous nature.

Future work will focus on improving multimodal alignment using advanced attention mechanisms and exploring more powerful vision-language models, such as CLIP, ALBEF, and BLIP-2. Enhancing datasets to better represent nuanced emotional expressions will also be a priority, with the goal of improving model generalizability and robustness across diverse social media contexts.

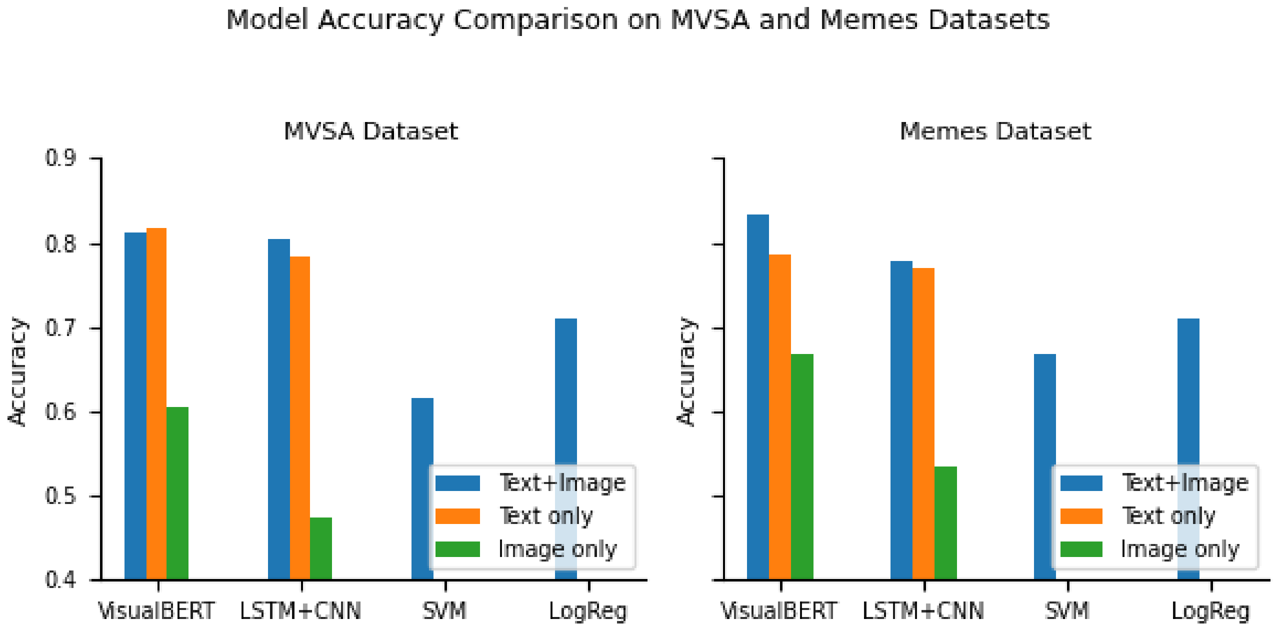

Figure 2 Illustration showcasing the accuracy of models that utilize text, images, and a combination of both text and images

7. ACKNOWLEDGMENTS

The author, Zannirah Sammani Muhammad, extends sincere gratitude to Prof. Mohammed Abo Rizka and her parents for their invaluable support and guidance throughout this research.

7.1 Declaration

The author, confirm that the manuscript is original, not sponsored, has not been published or accepted elsewhere in English, or any language and is not under current review.

8. REFERENCES

[1] Xu, C., Cetintas, S., Lee, K.-C., and Li, L.-J. 2014. Visual sentiment prediction with deep convolutional neural networks. arXiv:1411.5731v1.

[2] Qiu, K., Zhang, Y., Zhao, J., Zhang, S., Wang, Q., and Chen, F. 2024. A multimodal sentiment analysis approach based on a joint chained interactive attention mechanism. Electronics, 13(1), 1922.

[3] Al-Tameemi, I. K. S., Feizi-Derakhshi, M.-R., Pashazadeh, S., and Asadpour, M. 2024. A comprehensive review of visual–textual sentiment analysis from social media networks. Journal of Computational Social Science, 7(3), 2767–2838.

[4] Sánchez Villegas, D., Preoțiuc-Pietro, D., and Aletras, N. 2024. Improving multimodal classification of social media posts by leveraging image-text auxiliary tasks. arXiv:2309.07794v2.

[5] Dao, P. Q., Roantree, M., Nguyen-Tat, T. B., and Ngo, V. M. 2024. Exploring multimodal sentiment analysis models: A comprehensive survey. Preprints.

[6] Liu, B. 2012. Sentiment analysis and opinion mining. Morgan & Claypool Publishers.

[7] Jiang, T., Wang, J., Liu, Z., and Ling, Y. 2020. Fusion-extraction network for multimodal sentiment analysis. In Advances in Knowledge Discovery and Data Mining, Vol. 12085, 785–797. Springer.

[8] Dang, N. C., Moreno-García, M. N., and De la Prieta, F. 2020. Sentiment analysis based on deep learning: A comparative study. Data, 5(2), 35.

[9] Li, H., Lu, Y., and Zhu, H. 2024. Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism. Electronics, 13, 2069.

[10] Aliman, G. B., Nivera, T. F. S., Olazo, J. C. A., Ramos, D. J. P., Sanchez, C. D. B., Amado, T. M., Arago, N. M., Jorda Jr., R. L., Virrey, G. C., and Valenzuela, I. C. 2022. Sentiment analysis using logistic regression.

[11] Qixuan, Y. 2024. Three-class text sentiment analysis based on LSTM. Preprint submitted to Computer, Zhongnan University of Economics and Law.

[12] 15. You, Q., Jin, H., and Luo, J. 2017. Visual sentiment analysis by attending on local image regions. In Proceedings of the AAAI Conference on Artificial Intelligence. Retrieved from www.aaai.org.

[13] 16. You, Q., Luo, J., Jin, H., and Yang, J. 2016. Building a Large-Scale Dataset for Image Emotion Recognition: The Fine Print and the Benchmark. Proceedings of the AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence. Retrieved from www.aaai.org.

[14] 12. You, Q. and Luo, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the AAAI Conference on Artificial Intelligence.

[15] 13. Yang, J., She, D., and Sun, M. 2017. Joint image emotion classification and distribution learning via deep convolutional neural network. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).

[16] 14. Dongyu, S., Yang, J., Cheng, M.-M., Lai, Y., Rosin, P., and Liang, W. 2020. WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection. IEEE Transactions on Multimedia, 22(5), 1358–1371

[17] 17. Chen, T., Borth, D., Darrell, T., and Chang, S.-F. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. Proceedings of the IEEE International Conference on Multimedia and Expo (ICME).

[18] 18. Jiang, T., Wang, J., Liu, Z., and Ling, Y. 2020. Fusion-Extraction Network for Multimodal Sentiment Analysis. In H. W. Lauw et al. (Eds.), Proceedings of the PAKDD 2020 (Vol. 12085, pp. 785–797). Springer Nature. https://doi.org/10.1007/978-3-030-47436-2_59

[19] Dao, P. Q., Roantree, M., Nguyen-Tat, T. B., and Ngo, V. M. 2024. Exploring Multimodal Sentiment Analysis Models: A Comprehensive Survey. Preprints. https://doi.org/10.20944/preprints202408.0127.v1

[20] Luo, X. Y., Liu, J., Lin, P., and Fan, Y. 2021. Multimodal sentiment analysis based on deep learning: Recent progress. In Proceedings of The International Conference on Electronic Business (ICEB'21), Vol. 21, 293–303. Nanjing, China, December 3–7, 2021.

[21] Hakimov, S., Cheema, G. S., and Ewerth, R. 2025. Processing multimodal information: Challenges and solutions for multimodal sentiment analysis and hate speech detection. In I. Marenzi et al. (Eds.), Event Analytics across Languages and Communities, 71–94. Springer. https://doi.org/10.1007/978-3-031-64451-1_4

[22] Li, H., Lu, Y., and Zhu, H. 2024. Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism. Electronics, 13(11), 2069. https://doi.org/10.3390/electronics13112069

[23] Su, J., Liang, J., Zhu, J., and Li, Y. 2024. HCAM-CL: A novel method integrating a hierarchical cross-attention mechanism with CNN-LSTM for hierarchical image classification. Symmetry, 16(9), 1231. https://doi.org/10.3390/sym16091231

[24] Arevalo, J., Montes-y-Gómez, M., Solorio, T., and González, F. A. 2017. Gated Multimodal Units for Information Fusion. arXiv:1702.01992v1 [stat.ML]. https://arxiv.org/abs/1702.01992

[25] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557v1 [cs.CV]. https://arxiv.org/abs/1908.03557

[26] Shan, F., Liu, M., Zhang, M., and Wang, Z. 2024. Fake News Detection Based on Cross-Modal Message Aggregation and Gated Fusion Network. Computers, Materials & Continua, 2024, Article 10.32604/cmc.2024.053937. https://doi.org/10.32604/cmc.2024.053937

[27] Fields, C., and Kennington, C. 2023. Exploring transformers as compact, data-efficient language models. Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), 521–531.

[28] Wei, L., Wang, Z., Xu, J., Shi, Y., Wang, Q., Shi, L., Tao, Y., and Gao, Y. 2023. A lightweight sentiment analysis framework for a micro-intelligent terminal. Sensors, 23(2), 741. https://doi.org/10.3390/s23020741

[29] Pareek, P., Sharma, N., Ghosh, A., and Nagarohith, K. 2022. Sentiment analysis for Amazon product reviews using logistic regression model. Journal of Development Economics and Management Research Studies, 09(11), 29–42. https://doi.org/10.53422/09(11), 29-42

[30] Suryawanshi, S., Chakravarthi, B. R., Arcan, M., and Buitelaar, P. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 32–41. Language Resources and Evaluation Conference (LREC 2020), Marseille, May 11–16, 2020. European Language

Resources Association (ELRA).

[31] Barnes, K., Juhász, P., Nagy, M., and Molontay, R. 2024. Topicality boosts popularity: A comparative analysis of NYT articles and Reddit memes. Social Network Analysis and Mining, 14, 119. https://doi.org/10.1007/s13278-024-01272-3

[32] Schmidt, L., Talwar, K., Santurkar, S., and Tsipras, D. 2018. Adversarially robust generalization requires more data. Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

[33] Hoffer, E., Hubara, I., and Soudry, D. 2017. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[34] Aleqabie, H. J., Sfoq, M. S., Albeer, R. A., and Abd, E. H. 2024. Review of text mining techniques: Trends and applications in various domains. International Journal of Computer Science and Management, 5(1), 1–9. https://doi.org/10.52866/ijcsm.2024.05.01.009

[35] Kalra, V., and Aggarwal, R. 2018. Importance of text data preprocessing & implementation in RapidMiner. In Proceedings of the First International Conference on Information Technology and Knowledge Management, Vol. 14, 71–75. https://doi.org/10.15439/2018KM46

[36] Vidyashree, K. P., and Rajendra, A. B. 2023. An improvised sentiment analysis model on Twitter data using stochastic gradient descent (SGD) optimization algorithm in stochastic gate neural network (SGNN). SN Computer Science, 4, 190. https://doi.org/10.1007/s42979-022-01607-x

[37] Liu, C., Sheng, Y., Wei, Z., and Yang, Y. Q. 2018. Research of text classification based on improved TF-IDF algorithm. In Proceedings of the International Conference of Intelligent Robotic and Control Engineering. College of Information Science & Engineering, Ocean University of China.

[38] Valente, J., António, J., Mora, C., and Jardim, S. 2023. Developments in image processing using deep learning and reinforcement learning. Journal of Imaging, 9(10), 207. https://doi.org/10.3390/jimaging9100207

[39] Tachibana, Y., Obata, T., Kershaw, J., Sakaki, H., Urushihata, T., Omatsu, T., Kishimoto, R., and Higashi, T. 2019. The utility of applying various image preprocessing strategies to reduce the ambiguity in deep learning-based clinical image diagnosis. Magnetic Resonance in Medical Sciences, 19, 92–98. https://doi.org/10.2463/mrms.mp.2019-0021

[40] Murcia-Gómez, D., Rojas-Valenzuela, I., and Valenzuela, O. 2022. Impact of image preprocessing methods and deep learning models for classifying histopathological breast cancer images. Applied Sciences, 12(22), 11375. https://doi.org/10.3390/app122211375

[41] Barnes, K., Juhász, P., Nagy, M., and Molontay, R. 2024. Topicality boosts popularity: A comparative analysis of NYT articles and Reddit memes. Social Network Analysis and Mining, 14, 119. https://doi.org/10.1007/s13278-024-01272-3

[42] Guo, R., Wei, J., Sun, L., Yu, B., Chang, G., Liu, D., Zhang, S., Yao, Z., Xu, M., and Bu, L. 2024. A survey on advancements in image-text multimodal models: From general techniques to biomedical implementations. arXiv preprint arXiv:2309.15857.

[43] Jiang, M., and Ji, S. 2022. Cross-modality gated attention Jian fusion for multimodal sentiment analysis. arXiv

preprint arXiv:2208.11893.

[44] Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., and Liu, T. 2021. Understanding and Improving Early Stopping for Learning with Noisy Labels. arXiv preprint arXiv:2106.15853.

[45] Ren, J. 2024. Multimodal Sentiment Analysis Based on BERT and ResNet. School of Information and Engineering, Zhongnan University of Economics and Law. arXiv preprint arXiv:2412.03625v1

[46] Majumder, S., Aich, A., and Das, S. 2021. Sentiment analysis of people during the lockdown period of COVID-19 using SVM and logistic regression analysis.

[47] Henderi, & Siddique, Q. (2024). Comparative analysis of sentiment classification techniques on Flipkart product reviews: A study using logistic regression, SVC, random forest, and gradient boosting. *Journal of Data Mining and Decision Making*, 1(1), 4. https://doi.org/10.47738/jdmdc.v1i1.4

[48] Chiny, M., Chihab, M., Chihab, Y., & Bencharef, O. (2021). LSTM, VADER, and TF-IDF based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*, 12(7).

[49] Ur Rehman, A., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications*, 78, 26597–26613. https://doi.org/10.1007/s11042-019-07788-7

[50] Meena, G., Mohbe, K. K., & Kumar, S. (2023). Sentiment analysis on images using convolutional neural networks-based Inception-V3 transfer learning approach. *International Journal of Information Management Data Insights*,3, 100174.

[51] Bart, M. P., Savino, N. J., Regmi, P., Cohen, L., Safavi, H., Shaw, H. C., Lohani, S., Searles, T. A., Kirby, B. T., Lee, H., and Glasser, R. T. 2022. Deep learning for enhanced free-space optical communications. arXiv. https://arxiv.org/abs/2208.07712

[52] Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi- directional LSTM. *Multimedia Tools and Applications*, 81, 23649–23685. https://doi.org/10.1007/s11042-022-12648-y

[53] Vaydande, R. 2022. Retinal Fundus Image Classification using LSTM - Convolution Neural Network. MSc Research Project, Data Analytics. National College of Ireland, School of Computing. Supervisor: Vladimir Milosavljevic.

[54] Li, T., Hua, M., and Wu, X. 2020. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5). IEEE Access, Special Section on Feature Representation and Learning Methods with Applications in Large-Scale Biological Sequence Analysis. https://doi.org/10.1109/ACCESS.2020.2971348

[55] Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. 2019. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

[56] Bandyopadhyay, D., Hasanuzzaman, M., and Ekbal, A. 2024. Seeing through VisualBERT: A causal adventure on memetic landscapes. arXiv preprint arXiv:2410.13488.

[57] Yang, H., Zhao, Y., Wu, Y., aWang, S., Zheng, T., Zhang, H., Ma, Z., Che, W., and Qin, B. 2024. Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey. arXiv:2406.08068v2 [cs.CL]. https://arxiv.org/abs/2406.080