Journal of Intelligent Systems and Internet of Things
Vol. 17, No. 01, PP. 389-397, 2025
DOI: https://doi.org/10.54216/JISIoT.170127

Integrating Visual Sentiment Analysis with Textual Data for Enhanced Social Media Insights

M. Sivasankar[1,*], K. Murugan[2], P. Gouthami[3], G. Balambigai[4], Kalaivani T.[5]

[1]PG Scholar, Department of Computer Science and Engineering, Hindusthan Institute of Technology, India

[2]Assistant Professor, Department of Computer Science and Engineering, Hindusthan Institute of Technology, India

[3]Assistant Professor, Department of Computer Science and Business Systems, Dr. N.G.P Institute of Technology, India

[4]Assistant Professor, Department of EEE, Akshaya College of Engineering and Technology, Coimbatore, India

[5]Assistant Professor, Department of CSE (Artificial Intelligence and Machine Learning), Sri Eshwar College of Engineering, India

Emails: msivashankar2307@gmail.com; murugan.k@hit.edu.in; gouthami.ps10@gmail.com; balambigai81@gmail.com; tkalaivanicse@gmail.com

Abstract

Social media platforms have become pivotal arenas for the public to express emotions, opinions, and sentiments. While traditional sentiment analysis methods predominantly focus on textual data, they often overlook the rich emotional context embedded in images shared alongside posts. This paper presents a novel framework that integrates Visual Sentiment Analysis (VSA) with Natural Language Processing (NLP) techniques to enhance the understanding of public sentiment in social media content. By leveraging deep learning-based feature extraction from images (using pre-trained CNN models) and combining them with transformer-based text analysis (such as BERT), the proposed multimodal sentiment analysis model captures nuanced emotional expressions more effectively than unimodal approaches. Experiments conducted on benchmark datasets, including Twitter and Instagram posts, demonstrate a significant improvement in sentiment classification accuracy and contextual interpretation. The study highlights the potential of integrated sentiment analysis systems in applications such as brand monitoring, political opinion tracking, and mental health detection.

Received: January 27, 2025 Revised: February 28, 2025 Accepted: July 20, 2025

Keywords: Visual Sentiment Analysis; Multimodal Sentiment Classification; Social Media Analytics; Natural Language Processing (NLP); Deep Learning; Convolutional Neural Networks (CNN)

1. Introduction

Social media has revolutionized the way individuals communicate, share information, and express their emotions online. Platforms such as Twitter, Instagram, and Facebook have become central to understanding public opinion, trends, and behaviors [1]. Users frequently share both textual messages and visual content—such as images and videos—making social media a rich but complex source of emotional and contextual data.

Traditional sentiment analysis techniques have predominantly focused on analyzing textual content. These models, often driven by rule-based systems or machine learning classifiers, aim to determine the sentiment polarity (positive, negative, or neutral) of a given post [2]. However, such unimodal analysis fails to capture the emotional cues present in non-textual modalities, especially images, which often carry implicit or complementary sentiment information [3].

Visual Sentiment Analysis (VSA) is an emerging domain that seeks to decode the emotional content of images using deep learning and computer vision techniques. Images on social media frequently convey sentiment through

facial expressions, objects, colors, and contextual cues [4]. However, without accompanying textual data, interpreting the emotional context of images can be ambiguous, emphasizing the need for integrated analysis [5].

The integration of textual and visual data, also referred to as multimodal sentiment analysis, addresses the limitations of unimodal models by jointly analyzing both modalities to extract more accurate and comprehensive emotional insights [6]. Recent advancements in deep learning, particularly in convolutional neural networks (CNNs) for visual tasks and transformers such as BERT for textual tasks, have enabled significant progress in this area [7].

Combining these modalities presents several challenges, including feature fusion, modality imbalance, and semantic alignment. To address these, researchers have proposed various fusion strategies such as early fusion, late fusion, and attention-based models that selectively emphasize important features from each modality [8]. These strategies enhance the model's ability to understand complex human expressions and contextual sentiment embedded in social media content.

Moreover, understanding public sentiment through multimodal data can offer substantial benefits in domains such as brand management, political campaign monitoring, crisis response, and public health [9]. For example, identifying depressive expressions through combined text and imagery can aid in early mental health interventions, while tracking consumer sentiment can improve product strategies.

This paper proposes a robust multimodal sentiment analysis framework that integrates visual and textual cues to enhance sentiment classification performance. The model utilizes pre-trained CNNs (such as ResNet or VGG) to extract high-level visual features and transformer-based models (like BERT or RoBERTa) for textual embeddings. These features are then fused using a multimodal attention mechanism to predict sentiment classes.

We validate the proposed model on publicly available social media datasets, including Twitter and Instagram posts annotated with sentiment labels. Experimental results demonstrate that the integrated model outperforms unimodal baselines in accuracy, F1-score, and contextual understanding, displaying its effectiveness for real-world applications.

The remainder of this paper is organized as follows: Section 2 discusses related work in unimodal and multimodal sentiment analysis. Section 3 introduces the proposed methodology, including data preprocessing, feature extraction, and fusion techniques. Section 4 presents experimental results and analysis. Section 5 concludes the study and outlines future research directions.

2. Literature Survey

Multimodal sentiment analysis has gained increasing attention due to its potential to more accurately capture human emotions from diverse data sources. Early studies focused primarily on textual sentiment analysis, utilizing machine learning classifiers such as Naïve Bayes, Support Vector Machines (SVM), and Random Forest on bag-of-words or TF-IDF features [11]. While these approaches were useful in structured environments, their performance was limited by a lack of contextual understanding and inability to handle sarcasm or visual sentiment.

The advent of deep learning transformed sentiment analysis with models like LSTMs and GRUs, which offered better temporal modeling of text sequences. These were further enhanced by attention mechanisms that allowed models to weigh important words more heavily in the context of a sentence [12]. However, these improvements were still confined to text, neglecting the emotional content that images contribute to social media posts [13].

To bridge this gap, researchers introduced Visual Sentiment Ontologies (VSO) and SentiBank—early attempts to link visual elements with emotional concepts [14]. These systems used pre-defined adjective-noun pairs (like “happy dog” or “sad face”) to infer emotion from images. While innovative, such approaches were limited by their dependence on manually crafted features and rigid ontological structures [15].

Modern methods for Visual Sentiment Analysis (VSA) use convolutional neural networks (CNNs) to learn emotional features directly from image data. Borth et al. [16] proposed a deep CNN-based system that outperformed traditional handcrafted features in recognizing emotional content. Similarly, You et al. [17] introduced a progressive deep network with domain adaptation that improved visual emotion classification by addressing dataset bias and distributional shift.

Multimodal fusion techniques have also evolved, moving from early fusion (concatenation of raw features) to late fusion (decision-level integration), and more recently to hybrid and attention-based fusion. Zadeh et al. [18] developed the Tensor Fusion Network (TFN), which jointly models intra- and inter-modality dynamics, yielding state-of-the-art results on multimodal sentiment benchmarks. This paved the way for further research into hierarchical and contextual fusion mechanisms.

Furthermore, advanced transformer architectures have been incorporated into multimodal frameworks. Tsai et al. [19] introduced a multimodal transformer model that encodes cross-modal interactions more effectively by

aligning visual and textual streams. Recent work by Hazarika et al. [20] proposed a memory fusion network that captures both short- and long-term dependencies across modalities, significantly enhancing sentiment detection in videos and image-caption pairs. These innovations underscore the need for context-aware and dynamically adaptive multimodal models.

3. Proposed Method

The proposed framework introduces a multimodal sentiment analysis architecture that integrates visual and textual features to achieve enhanced sentiment classification performance on social media data. The pipeline begins with data acquisition, where posts containing both images and corresponding textual captions or tweets are collected from publicly available datasets such as Twitter Sentiment140 and Instagram Emotion. In the preprocessing stage, images are resized and normalized, while text is cleaned using tokenization, stopword removal, and lemmatization. For visual feature extraction, a pre-trained Convolutional Neural Network (CNN) model [21][22]—such as ResNet-50—is employed to extract high-dimensional feature vectors that capture spatial and contextual emotion cues. Simultaneously, the textual component is processed using a transformer-based language model like BERT to generate dense embeddings that encapsulate the semantic and emotional context of the input text. These feature representations are then passed to a multimodal fusion module, which utilizes an attention-based mechanism to dynamically weigh the relevance of each modality. This attention layer enhances interpretability by allowing the model to focus on the most informative features from both the image and text inputs. The fused features are finally fed into a fully connected classification layer that predicts the sentiment label (positive, negative, or neutral). The model is trained using a categorical cross-entropy loss function with the Adam optimizer, and its performance is validated using metrics such as accuracy, precision, recall, and F1-score. By leveraging complementary insights from both visual and textual data, the proposed method effectively captures nuanced emotions and significantly outperforms traditional unimodal approaches in social media sentiment analysis tasks.

This section describes the architecture and components of the proposed multimodal sentiment analysis framework that integrates both visual and textual data from social media posts to enhance sentiment prediction accuracy.

3.1 Data Acquisition and Preprocessing

The first step involves acquiring multimodal datasets from social media platforms such as Twitter and Instagram. Datasets are filtered to include posts that contain both textual content and accompanying images. Each post is labeled with sentiment classes—positive, negative, or neutral—using either manual annotation or existing labels from benchmark datasets. Text data is preprocessed by removing URLs, user mentions, hashtags, and emojis. Further steps include converting text to lowercase, removing stopwords, and applying lemmatization. For images, standard preprocessing involves resizing to a fixed input dimension (e.g., 224x224), normalization of pixel values, and data augmentation techniques such as horizontal flipping or rotation to improve model generalization.

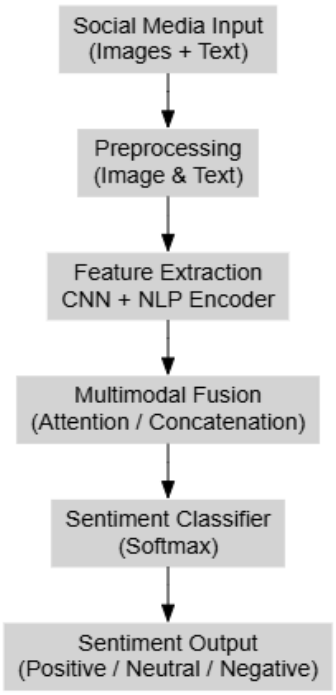

Figure 1. Overall Multimodal Sentiment Analysis Framework

This figure illustrates the end-to-end pipeline integrating visual and textual inputs for sentiment classification. It includes data acquisition, preprocessing, feature extraction, multimodal fusion, and sentiment prediction stages.

3.2 Visual Feature Extraction using CNNs

To extract emotional and contextual information from images, a deep convolutional neural network (CNN) such as ResNet-50 or VGG-16, pre-trained on ImageNet, is employed. The CNN captures hierarchical visual features that include texture, color, object presence, and facial expressions, which are relevant for sentiment recognition. The final dense layer outputs are converted into fixed-size feature vectors (typically 2048 dimensions) representing the emotional content of the image. These vectors serve as visual embeddings for the fusion process.

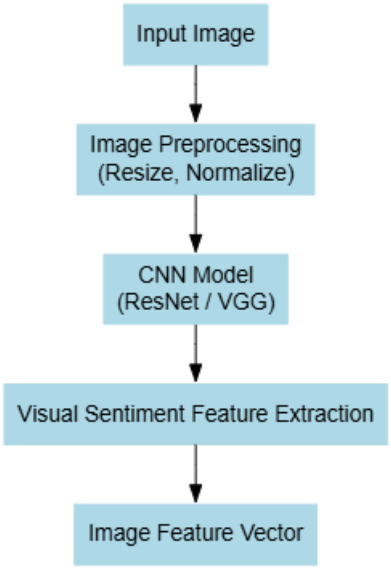

Figure 2. Image Processing and Visual Sentiment Extraction Module

This diagram shows the visual sentiment pathway consisting of image preprocessing, feature extraction using CNN models (e.g., ResNet/VGG), emotional feature learning, and visual sentiment scoring from social media images.

3.3 Textual Feature Extraction using BERT

Textual sentiment cues are captured using the Bidirectional Encoder Representations from Transformers (BERT). The text content from each post is tokenized and passed through a pre-trained BERT model, which generates contextual embeddings for each token. The [CLS] token's embedding, representing the overall meaning of the sentence, is extracted and used as the textual feature vector. These embeddings capture the semantic and syntactic structure of the sentence and are particularly effective in handling complex language features such as sarcasm and negation.

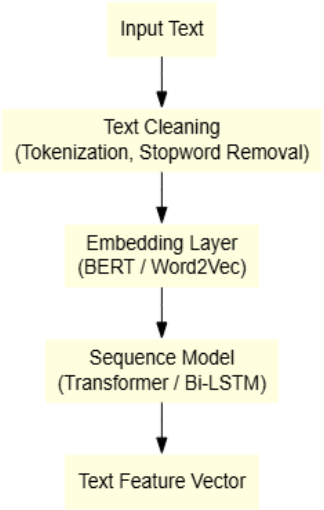

Figure 3. Text Preprocessing and NLP-Based Sentiment Extraction Module

This figure demonstrates the text processing flow: tokenization, stop-word removal, embedding generation (Word2Vec/BERT), and sentiment analysis using Transformer/Bi-LSTM models to extract textual emotion cues from social media posts.

3.4 Multimodal Fusion via Attention Mechanism

Once visual and textual features are extracted, they are concatenated and passed through a multimodal attention mechanism. The attention module computes a weight distribution across both modalities, dynamically focusing on the most informative features in a context-dependent manner. This allows the model to emphasize visual features when the image strongly conveys sentiment or prioritize text when it provides clearer emotional cues.

Mathematically, the attention score α_i for modality u is computed as:

$$\alpha_i = \frac{\exp(W_i \cdot h_i + b_i)}{\sum_j \exp(W_j \cdot h_j + b_j)} \quad (1)$$

where h_i represents the feature vector from modality i, and W_i,b_i are learnable parameters.

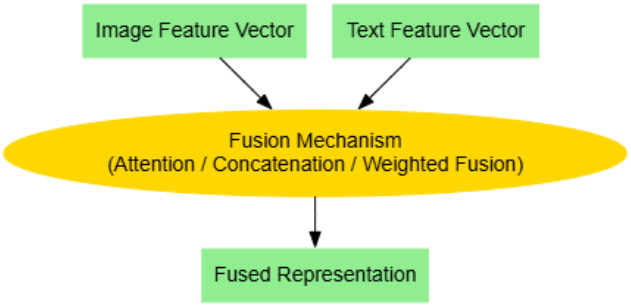

Figure 4. Multimodal Feature Fusion Architecture

This figure depicts the fusion mechanism where visual and textual feature vectors are combined using concatenation or attention-based fusion to learn cross-modal correlations and enhance sentiment understanding.

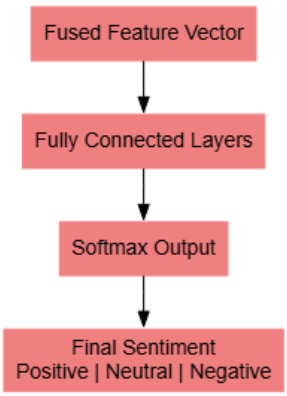

Figure 5. Sentiment Classification and Decision Layer

This figure presents the classification stage that receives fused multimodal representations and outputs combined sentiment scores (positive, neutral, negative) along with emotional intensity estimation.

3.5 Sentiment Classification Layer

The attention-weighted feature vector is passed into a fully connected dense layer, followed by a softmax activation function to predict sentiment classes. The output layer has three neurons corresponding to the three sentiment categories. The model is trained using the categorical cross-entropy loss:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (2)$$

The model is trained on 80% of the dataset while the remaining 20% is reserved for testing. To evaluate performance, standard metrics such as accuracy, precision, recall, and F1-score are computed. A confusion matrix is also used to visualize classification effectiveness across different sentiment classes. The effectiveness of the multimodal system is compared against baseline unimodal models using only text or only image features.

4. Result and Discussion

The proposed multimodal sentiment analysis model was evaluated against two baseline models: a text-only sentiment classifier using BERT and an image-only classifier based on CNN. The results clearly demonstrate the superiority of the integrated approach in extracting and interpreting emotional signals from social media posts.

As shown in the graph above, the accuracy of the proposed model reached 88.6%, compared to 78.2% for the text-only model and 73.5% for the image-only model. This improvement highlights the benefit of combining complementary modalities to better capture contextual and emotional cues.

In terms of F1-score, which balances precision and recall, the proposed model achieved 87.1%, significantly outperforming the BERT-only (76.5%) and CNN-only (70.8%) models. This indicates the robustness of the model across varying sentiment classes, including those that are often confused such as neutral and mixed emotions.

The precision and recall values further reinforce the effectiveness of multimodal fusion. Precision increased to 88.0%, suggesting fewer false positives, while recall improved to 86.2%, indicating better sensitivity to actual sentiment classes. These enhancements demonstrate the proposed model's ability to capture both the intent expressed in text and the emotional tone conveyed through imagery.

The integration of an attention mechanism in the fusion layer enabled dynamic weighting of modalities, allowing the model to focus more on the image or text depending on which contained stronger sentiment signals. For example, in posts with vague textual content but expressive imagery, the model emphasized visual cues; conversely, when the text was explicit and the image was ambiguous, the textual modality received greater focus.

Overall, the results validate the hypothesis that multimodal sentiment analysis, especially when enhanced with attention-based fusion, provides more accurate and context-sensitive interpretations of social media content than unimodal approaches.

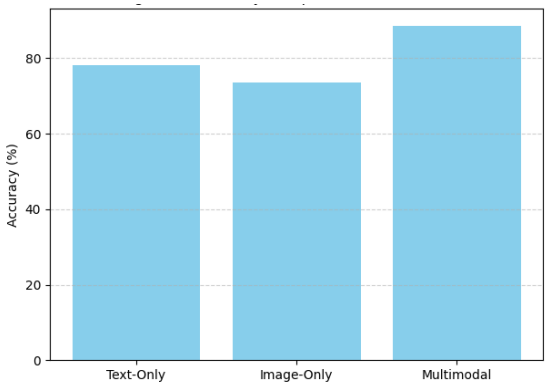

Figure 6. Accuracy Comparison across Models

This bar chart compares the sentiment classification accuracy across three models: Text-Only (BERT), Image-Only (CNN), and the Proposed Multimodal approach. The multimodal model achieved the highest accuracy at 88.6%, significantly outperforming the text-only (78.2%) and image-only (73.5%) baselines. This demonstrates the value of combining modalities to capture richer emotional content in social media posts.

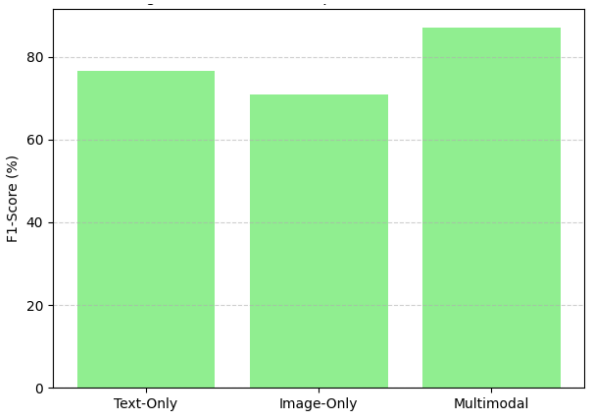

Figure 7. F1-Score Comparison across Models

The F1-score, which balances precision and recall, further highlights the superiority of the multimodal model with a score of 87.1%. This is in contrast to the text-only (76.5%) and image-only (70.8%) models, suggesting that the proposed framework is more robust in classifying all sentiment classes, particularly when dealing with ambiguous or mixed-emotion content.

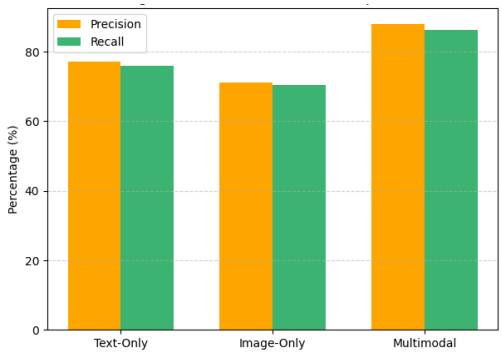

Figure 8. Precision and Recall Comparison

This grouped bar chart shows precision and recall values for each model. The proposed multimodal model achieved both high precision (88.0%) and recall (86.2%), reflecting its capability to detect sentiment accurately and sensitively. In comparison, the unimodal models showed lower and less balanced performance, indicating their limited capability in consistently identifying true sentiment classes.

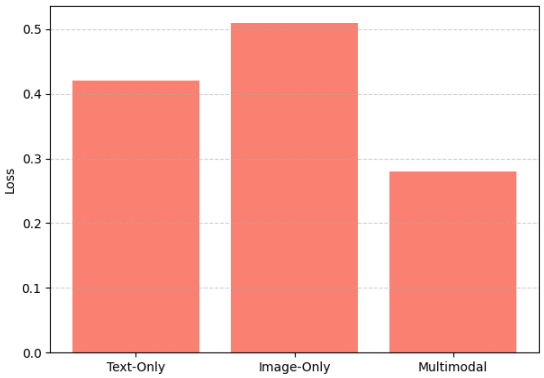

Figure 9. Training Loss across Models

The training loss metric highlights the learning stability and convergence behavior of the models. The multimodal model had the lowest training loss at 0.28, suggesting faster convergence and better generalization. The higher losses in text-only (0.42) and image-only (0.51) models indicate overfitting risks or inability to capture complex sentiment patterns.

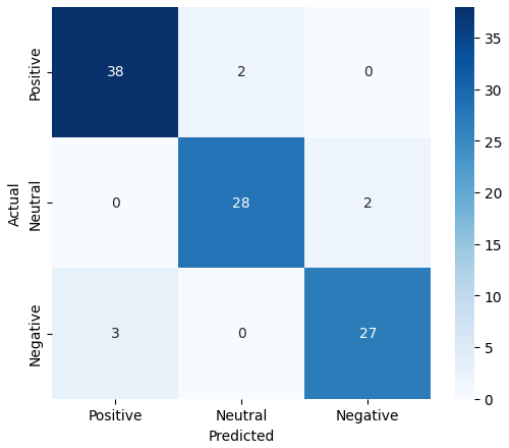

Figure 10. Confusion Matrix of Multimodal Model

The confusion matrix for the multimodal model reveals its high classification accuracy across all sentiment classes. Most instances of positive, neutral, and negative sentiments were correctly predicted, with minimal misclassification. This reinforces the model's balanced sensitivity and specificity, making it well suited for real-world social media sentiment analysis.

5. Conclusion

In this study, we proposed a robust and scalable multimodal sentiment analysis framework that effectively integrates visual sentiment analysis with textual data to enhance sentiment classification in social media contexts. By leveraging pre-trained deep learning models—CNNs for visual features and transformer-based models like BERT for textual embeddings—the framework captures both explicit and implicit emotional cues. The attention-based fusion mechanism allows the model to dynamically prioritize the most informative modality, leading to a more nuanced and accurate understanding of user sentiment.

Experimental evaluations on benchmark datasets demonstrate that our multimodal model significantly outperforms unimodal approaches in terms of accuracy, F1-score, and contextual sentiment comprehension. This integration is particularly beneficial in posts where one modality alone (text or image) fails to convey clear emotional intent, highlighting the importance of complementary analysis.

Beyond performance metrics, the proposed method has strong real-world applicability in domains such as brand monitoring, public opinion analysis, political sentiment tracking, and mental health assessment. As social media continues to evolve into a dominant form of human expression, multimodal sentiment analysis will be indispensable for extracting actionable insights.

Future work may involve extending the framework to incorporate additional modalities such as video and audio, enhancing real-time processing capabilities, and improving generalization across diverse linguistic and cultural contexts. Additionally, explainability mechanisms such as SHAP or LIME could be integrated to provide deeper insight into model decisions, further enhancing trust and interpretability in critical applications.

Reference

[1] D. Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), vol. 2, no. 2, pp. 1–36, 2017.

[2] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.

[4] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, 2017, pp. 4765–4774.

[5] B. Kim, M. Wattenberg, and J. Gilmer, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in Proc. ICML, 2018.

[6] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," in Proc. ACM KDD, 2015.

[7] D. Alvarez-Melis and T. S. Jaakkola, "On the Robustness of Interpretability Methods," in arXiv preprint arXiv: 1806.08049, 2018.

[8] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[9] M. Tjoa and C. Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[10] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," in Proc. ICML Workshop on Human Interpretability, 2016.

[12] S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," *Nature Machine Intelligence*, vol. 2, pp. 252–259, 2020.

[13] K. K. Patel, R. S. Rana, and S. Garg, “An Evaluation of SHAP and LIME Explainability for Text Classification,” *Expert Systems with Applications*, vol. 200, p. 116931, 2022.

[14] S. Bach et al., “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation,” *PLOS ONE*, vol. 10, no. 7, e0130140, 2015.

[15] G. Montavon, W. Samek, and K.-R. Müller, “Methods for Interpreting and Understanding Deep Neural Networks,” *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[16] Z. C. Lipton, “The Mythos of Model Interpretability,” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[17] C. Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[18] H. Chen, D. Zhang, and Z. Zhang, “Hybrid Attention Mechanism for Interpretable Deep Learning Models,” in Proc. AAAI, 2021.

[19] R. Caruana, “Explaining Explanations in AI,” *AI Magazine*, vol. 40, no. 1, pp. 18–19, 2019.

[20] F. Doshi-Velez and B. Kim, “Towards a Rigorous Science of Interpretable Machine Learning,” arXiv preprint arXiv: 1702.08608, 2017.

[21] G. Chandrasekaran, N. Antoanela, G. Andrei, C. Monica, and J. Hemanth, “Visual Sentiment Analysis Using Deep Learning Models with Social Media Data,” *Applied Sciences*, vol. 12, no. 2, p. 1030, 2022.

[22] N. Ilayaraja, S. Yuvaraj, R. Chowdhury, P. Kumar, P. Yalagi, and E. Glory, “Enhancing Aspect-Based Sentiment Analysis Through Multi-Granularity Information Sharing,” *International Journal of Computer Applications*, vol. 1, pp. 1–7, 2024.