# Modeling of Artificial Intelligence-Driven Context-Aware Memotion Analysis Using Integrated Large-Scale Multimodal Corpora

Rania Aboalela
Department of Information Systems-Faculty of Computing and Information Technology,
King Abdulaziz University, Rabigh, Saudi Arabia

***Abstract*—Memotion analysis is a complex task designed to accurately identify the sentiments and feelings conveyed within internet memes. A meme is any trend, skill, or belief learned from others, rather than inherited genetically, and shared through imitation. The spread of malicious and aggressive memes has aggravated issues on social media, driving the need for this research to address and overcome such digital threats. Based on the fact that challenging memes require visual and language knowledge, recognizing offensive memes is more complex than sensing invasive texts. With the expansion of artificial intelligence (AI) and deep learning (DL) techniques, there has been an increasing interest in automatic analysis of such creative multimodal content, where the natural language processing (NLP) and computer vision (CV) research communities have usually been attentive to a single dominant modality in isolation. This study presents an Artificial Intelligence Driven Context-Aware Memotion Analysis Using Integrated Large-Scale Multimodal Corpora (AICAMA-ILSMC) model. The proposed work aims to effectively analyze and classify memes by jointly exploiting visual and textual contexts in a unified learning architecture. In the data preprocessing stage, bilateral filtering is applied for image noise removal, while textual content is processed through optical character recognition, lowercasing, punctuation, stop-word removal, tokenization, and handling slang, emojis, and abbreviations. For feature extraction, an improved Swin Transformer is utilized to extract discriminative visual features, whereas RoBERTa-Large is employed to obtain deep contextualized text embeddings. Furthermore, the multi-modal representations are combined using a cross-modal attention-based fusion strategy to enable efficient interaction between visual and embedded textual information, producing a text-dominant fused representation. Moreover, the fused features are passed into a bidirectional long short-term memory network with a self-attention mechanism for accurate memotion classification, such as Negative and Very Negative, Positive and Very Positive, and Neutral. Finally, the Nadam optimizer is integrated into the proposed model to enhance convergence and overall model performance. A series of experimental analyses is conducted, and the outcomes can be inspected under numerous aspects. The comparison study showed the improved performance of the AICAMA-ILSMC technique compared to existing approaches.**

***Keywords*—Memotion; large-scale multimodal data; artificial intelligence; feature extraction; RoBERTa-large embeddings; model optimization***

## I. Introduction

With the increase of social networking sites as a channel for users to transfer their opinions and communicate with each other, the quantity of hate online has increased in parallel. The supremacy of free, uncensored speech still leads to significant distress in virtual communities by depreciating others [1]. A prevalent type of content generating these negative contents is called memes. In general, memes comprise prevalent texts and images related to them that are aimed at triggering humor amongst the readers. A famous description of memes, now extensively utilized in the domain, defines them as "a collection of texts with shared attributes, with an exchanged main of posture, forms, and contents". Expansively, "content" implies concepts and thoughts, "form" indicates the sensory identities such as video or audio, and "posture" means the style or tone, structures for contribution, and communicative task of the meme [2]. The creative usage of text and images is gained the relevant and trending contents. While primarily utilized only for comic purposes, memes have rapidly grown as a method that can be utilized to criticize and degrade specific parts of society. They are also employed to distribute misleading information and false news. Memes function as a language of their own, with the capability to exceed principles and construct forms from the gathered individualities among people. Such distributable comic visuals are more robust and effective tools for social impact, communication, political subversion, and also expressivity [3].

A total of thirteen various forms of memes are identified, ranging from simpler text-based memes to illustrations, rage comics, and quotes. In their more identifiable form, memes generally follow particular image functions, or templates with an identifiable image and a text overlay that varies in individual instances of the meme. This multimodal element of memes causes meaning to be created from different stages. Current research works are concentrated on automated genre classification of (especially political) memes, which can be supported to enable the data gathering, a significant intention that offers an up-to-date lack of benchmark datasets for memes [4]. The multifaceted characteristics have influenced the research under the automated identification of internet memes, as separated in dual domains. The natural language processing (NLP) analysis of memes has a tendency to focus only on the textual part. In contrast, the domain of computer vision (CV) is mainly considered from a visual standpoint. The SemEval 2020 Memotion task intends to present a new process for emotion identification of internet memes. This analysis aims to respond

to the requirement for expanding attention to social media information from more hybrid viewpoints and to bridge the gap among these parts of techniques.

Memes originate on each type of social media platform and comprise images and text. It has non-verbal communications and generally indicates cultural influences increased from TV movies or programs and vibrates with an individual's opinions and emotions. Additionally, the increasing utilization of these memes needs an automatic method, which can be highly extensible when compared to manual investigation. The capability of artificial intelligence (AI) methods has significantly improved the automatic analysis of digital media information like visual, video, textual, and audio [5]. Research is achieved by implementing sentiment analysis (SA), emotion identification of memes, and automatic generation. Consequently, the major goal is to design a DL-based technique that can analyze the features and employ them for classifying the memes with respect to projected semantic data.

### A. Key Contributions of this Work

To improve the reliability of memotion analysis from multimodal meme content, this article presents an Artificial Intelligence Driven Context-Aware Memotion Analysis Using Integrated Large-Scale Multimodal Corpora (AICAMA-ILSMC) approach. The proposed AICAMA-ILSMC approach follows a comprehensive multimodal learning pipeline that consists of image and text preprocessing, deep feature extraction, attention-based fusion, hybrid classification, and parameter tuning. The empirical validation of the AICAMA-ILSMC model has been carried out on the benchmark dataset, and the results are examined under several performance measures. The key contributions of this work are given below,

- Utilize comprehensive multimodal preprocessing strategies, including bilateral filtering (BF) for image noise removal and textual normalization (optical character recognition (OCR), lowercasing, punctuation and stop-word removal, tokenization, and handling slang, emojis, and abbreviations), to increase the quality and consistency of input data for memotion analysis.
- Design a multimodal feature extraction module by integrating an improved Swin Transformer (ST) for discriminative visual representation learning and RoBERTa-Large for deep contextualized textual embeddings, allowing for effective exploitation of both visual and linguistic cues.
- Present a cross-modal attention-based fusion strategy to model fine-grained interactions between visual and textual methods and to generate text-dominant fused representations that improve contextual understanding in memotion classification.
- Develop a hybrid classification model using a bidirectional long short-term memory network with a self-attention mechanism (BiLSTM-SAM) for efficiently learning contextual dependencies for precise multi-class memotion classification.
- Integrate Nadam optimizer for hyperparameter tuning, accomplishing improved convergence and enhanced classification performance.
- Perform extensive evaluations on the benchmark Memotion dataset and exhibit the improved performance of the model across diverse performance measures.

## II. Related Works on Memotion Analysis

A summary of existing related works regarding memotion analysis is presented in this section. Jaganath et al. [6] developed a new method for SA from meme images by employing the optical character recognition (OCR) method that can be integrated with DL techniques for classifying the text images. This technique involves executing OCR on the meme images to extract the text content for the partitioning process by using textual analysis and illustration interpretations for SA. The evaluation of texts is done by employing VADER and NLP, and the following steps are obtained by inferring sentiments with respect to the identified texts. By using this innovative multi-modal method, identification of invasive content from images and text must be accomplished effectively, thus offering a nonviolent and more accountable stage in social media. Antu et al. [7] implemented wide-ranging research on the Bengali meme SA with the MemoSen database and examined the pre-trained hybrid fusion techniques and VLMs in lower-resource contexts. To overcome these difficulties, the authors present a custom multimodal sentiment method that contains a ResNet-18-based visual encoder and a Bengali-aware text processing channel with BERT tokenization and a following BiLSTM–BiGRU sequential model. A method called as attention-guided fusion has been utilized for dynamically integrating the textual and visual features that could be smoothed by employing the focal loss with balanced sampling. Sharma et al. [8] introduced the MOOD (Meme emotions dataset) that represents 6 primary emotions. Afterwards, ALFRED (emotion-Aware multi-modal fusion for emotion detection) is an innovative multimodal neural architecture that clearly models emotion-based visual indications and utilizes an effective cross-modal fusion by using the gating mechanism.

Aishy et al. [9] developed a multimodal DL architecture for categorizing memes and utilizing visual and textual data to increase the sentiment data in social media content. This research work explores multimodal and unimodal approaches through an image-based CNN framework, such as VGG19, ResNet152, YOLOv4, and MobileNetV2, integrated with text-based techniques like BiLSTM, using feature fusion methods. Hilarious. Shi et al. [10] developed MemoDetector, an innovative architecture for advanced MEU. Primarily, the authors presented a 4-stage textual improvement model that employs the significant data and intellectual abilities of multimodal large language models (MLLMs) for increasingly deducing and extracting implied and contextual data from memes. Such improved texts mainly expand the original contents of the meme and offer valuable direction for downstream classification. Later, a two-phase modal fusion approach: the primary phase is performed by shallow fusion with original meme texts and images, whereas the secondary phase intensely combines the heightened textual and visual

attributes. This hierarchical fusion allows the system to gain higher multifaceted cross-modal emotional indications.

Breiteneder et al. [11] introduced the first-of-its-kind task to expansively examine the behavior of negative meme detection techniques in the presence of imperfect modal data. Particularly, an innovative baseline technique that is learned a common representation for numerous modalities by predicting them self-reliantly. Such representations are utilized when the information is imperfect and modal. Asmawati et al. [12] intended to design an SA method for meme images corresponding to essential aspects. The investigated meme images encompass human faces. The facial features are captured by employing the main aspects, which are the mouth, eyes, and eyebrows. There are Numerous involvements like emerging a human facial SA detection technique by employing significant points, thus indicating as different graphs, and forming the meme dataset with Indonesian texts. Wahyuni et al. [13] examined a multimodal FL technique that incorporates texts and images for improving sarcasm detection. This method is well-maintained and converts characters, punctuation, and emojis into sequence features instead of eliminating them. The images have been processed through OCR to extract the text and ensure computational effectiveness without non-textual visual essentials. Word representations could be produced by applying the Word2Vec embeddings that are used, LSTM, BiLSTM, and GRU methods. Table I presents a summary of the literature on sentiment analysis applied to meme text classification.

TABLE I. OUTLINE OF EXISTING MODELS

| Authors | Objectives | Models | Datasets | Outcomes | Limitation |
|---|---|---|---|---|---|
| Jaganath et al. [6] | To enhance social media safety by implementing OCR-driven sentiment analysis combined with image recognition. | VADER, CNN, and NLP algorithms. | TamilMemes dataset. | F1 score of 90%. | The system may struggle to detect sarcasm, irony, or culturally contextual humor in memes. |
| ANTU et al. [7] | To build a system that understands the emotions expressed in Bengali memes using both pictures and text. | BERT, BiLSTM, and BiGRU | MemoSen dataset. | F1-score of 0.88. | Mixed languages and slang words may reduce accuracy. |
| Sharma et al. [8] | To explicitly model emotion-enriched visual cues for improved affective representation. | ViT and BERT models. | Meme emotions Dataset | F1-score of 82.43 %. | Domain shifts across meme datasets may reduce accuracy despite generalization efforts. |
| Aishy et al. [9] | To develop a multimodal framework for classifying humor in memes using both images and text. | CNN and BiLSTM methods. | Memotion 2.0 dataset. | Precision of 89%. | Class imbalance and minority class scarcity can limit generalization and robustness. |
| Shi et al. [10] | To implement a four-step textual enhancement module for enriching meme text with background knowledge and reasoning. | MLLMs models | MET-MEME and MOOD datasets. | Accuracy of 49.57% and 83.52%. | Generalization to new meme formats or rapidly evolving internet trends may be limited. |
| Breiteneder et al. [11] | To investigate harmful meme detection in scenarios where one or more modalities, such as text, are missing. | Multimodal Learning method. | HarMeme dataset. | F1-score of 54.25. | Increased computational complexity arises from projecting and fusing multiple modality representations. |
| Asmawati et al. [12] | Designing the sentiment analysis model for meme images depends on facial key points. | DNN model | Human face datasets. | Accuracy of 83%. | The approach may not generalize well to memes with exaggerated or stylized facial expressions. |
| Wahyuni et al. [13] | To evaluate multiple deep learning architectures for performance, generalization, and computational efficiency. | LSTM model | SemEval-2015 Task 11, SemEval-2018 Task 3, and Reddit SARC 2.0 datasets. | Accuracy of 76%. | Models may not generalize well across diverse social media platforms or meme formats. |

## III. PROPOSED SYSTEM

The proposed model introduces an effective method for classifying memotion from large-scale multimodal corpora. Fig. 1 depicts the overall structural design of the proposed AICAMA-ILSMC model.

As illustrated in Fig. 1, the process begins with a preprocessing phase, where images undergo the BF technique to remove noise, while texts undergo comprehensive steps of OCR, lowercasing, punctuation, stop-word removal, tokenization, and handling slang, emojis, and abbreviations to enhance data consistency and quality. Subsequently, the feature extraction stage is carried out, where image features are extracted using improved ST, and the embedded textual information is obtained through RoBERTa-Large. Then, a cross-modal attention fusion mechanism is incorporated to model interactions among visual and textual modalities, generating text-dominant fused representations. The fused features are fed into the BiLSTM-SAM classifier that learns to contextualize dependencies for robust motion classification. In addition, the Nadam optimizer is adopted to fine-tune model parameters, resulting in better model convergence and performance

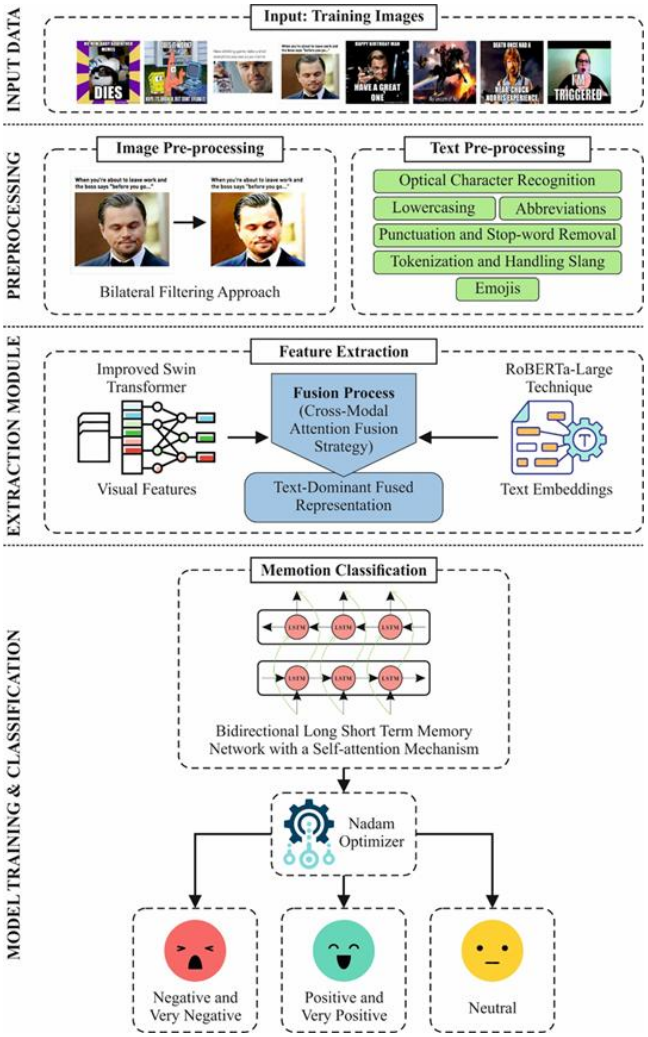



Fig. 1. Overall structural design of the proposed AICAMA-ILSMC approach.

### A. Dataset Used

The empirical validation of the proposed AICAMA-ILSMC method can be assessed by employing the Memotion Dataset 7k [14]. This dataset is a collection of around 7,000 internet meme images generated for the SemEval-2020 Memotion Analysis task, developed to support multimodal research integrating both visual and textual information.

TABLE II. DETAILS OF DATASET

| Sentiment | No. of Samples |
|---|---|
| Positive | 4160 |
| Neutral | 2201 |
| Negative | 631 |
| Total | 6992 |

TABLE III. SAMPLE TEXT

| Image Name | Text OCR | Text Corrected | Classes |
|---|---|---|---|
| image_1.jpg | LOOK THERE MY FRIEND LIGHTYEAR NOW ALL SOHALIKUT TREND PLAY THE 10 YEARS CHALLENGE AT FACEBOOK imgflip.com | LOOK THERE, MY FRIEND LIGHTYEAR NOW ALL SOHALIKUT TREND PLAY THE 10 YEARS CHALLENGE AT FACEBOOK imgflip.com | Positive |
| image_189.jpg | Whole Year Night Before Exam Sarcasmlol.com During Exam | Whole Year Night Before Exam Sarcasmlol.com During Exam | Neutral |
| image_422.jpg | Don't forget to tuck your cocks back; I mean, set your clocks back this weekend | Don't forget to tuck your cocks back; I mean, set your clocks back this weekend | Negative |
| image_169.jpeg | Photo bombing level: Mr. Bean | Photo bombing level: Mr. Bean | Positive |
| image_277.jpg | When you're trying to catch the piece of shrimp in the air at Benihana | When you're trying to catch the piece of shrimp in the air at Benihana | Neutral |
| image_266.jpg | Me: I should keep my identity a secret. Me to me: But has he heard about the tragedy of Darth Plagueis the Wise? @santa1896 | Me: I should keep my identity a secret. Me to me: but has he heard about the tragedy of Darth Plagueis the Wise? | Negative |

Every meme encompasses the image and its corresponding text (frequently extracted via OCR), allowing models to examine both modalities together. The dataset is annotated by humans for overall sentiment (positive, negative, neutral) as well as emotional and semantic classes like humor, sarcasm, offensiveness, and motivation. It is broadly utilized to train and evaluate multimodal DL techniques in sentiment and emotion analysis. The dataset has 6992 records under 3 classes, as represented in Table II. Table III demonstrates the sample text.

### B. Preprocessing Pipeline

This section defines the preprocessing pipeline implemented for preparing visual and textual inputs for effective feature learning.

*1) Visual data preprocessing*: This subsection describes the image preprocessing strategy adopted to remove noise and improve visual quality. The BF is a noise-reducing, edge-preserving, and non-linear smoothing filter that changes image pixel intensities with a weighted average of close pixels. The BF in the domain of image denoising, as represented in the mathematical form,

$$I'(i) = \frac{\sum_{j \in N(i)} w_c\,(\|j - i\|) w_s\,(|I\,(i) - I\,(j)|) I\,(j)}{\sum_{j \in N(i)} w_c\,(\|j - i\|) w_s\,(|I\,(i) - I\,(j)|)}, \quad (1)$$

Here the terms $I'(i)$ and $I(i)$ refers to the updated and original pixel values with respect to position $i = (u, v)$, correspondingly. The parameter $N(i)$ signifies the adjacent pixels of $i = (u, v)$, the parameter $w_c$ and $w_s$ describe the weighting operations. The parameter $|j - i|$ signifies the spatial distance among pixels, and the mathematical notation $|I(i) - I(j)|$ measures the color variance among them. Particularly, the BF is responsible for the spatial and color data among pixels.

The BF is used to meme images that need the pixel intensity differences for computing the weights, as represented in Eq. (2), and it is similar to BF for natural images in Eq. (1). Later, the location of the pixel value has been updated by fluctuating with the intensity gradient, as represented in Eq. (3).

$$\delta = \frac{\sum_{q \in N(p_i)} w_d(di) w_n\left(r(n_i)\right) r(n_i)}{\sum_{q \in N(p_i)} w_d(di) w_n\left(r(n_i)\right)}, \quad (2)$$

$$I_i' = I_i + \delta, \quad (3)$$

Here the parameter $I_i$ indicates the pixel intensity value at location $i$, the parameter $N(i)$ represents the adjacent of $i$, $d_i = \| i - q \|$ indicates the Euclidean distance between pixel $i$ and its adjacent pixel $q$, the mathematical term $r(i) = | I(i) - I(q) |$ refers to the intensity difference between neighbouring pixels, and $w(\cdot)$ represents the weighting function.

*2) Text data preprocessing*: This subsection outlines the text preprocessing steps used on the meme text for robust representation learning.

*a) OCR*: It is used for extracting the embedded textual content within the meme images. The OCR channel comprises recognizing text regions and learning the visual character representations by employing the pre-trained identification models, and combining recognized character-level predictions into meaningful words and sentences.

*b) Lowercasing and punctuation removal*: It converts the whole extracted text into lowercase and removes redundant symbols to ensure textual consistency.

*c) Removal of stop words*: Normally happening words like "an," "but," "a," "of," "and," "the," and more on that can be carried out the minimized semantic data are removed.

*d) Tokenization*: It is the method of treating the input text as a constant sequence and segmenting it into subword tokens or meaningful words.

*e) Handling slang, emojis, and abbreviations*: Emoji symbols, informal terms, and shortened expressions are normalized into their standard semantic representations for efficient contextual data.

### C. Multimodal Feature Extraction

This section presents the multimodal feature extraction model for learning discriminative visual and textual representations.

*1) Image feature encoder:* This subsection details the visual feature encoding process utilizing an improved ST. The ST is also known as the Shifted Window Transformer, which is an innovative visual Transformer model developed to overcome the computational complexity that can be encountered by conventional transformers with large-scale images. In contrast with the traditional global attention mechanism (AM), the ST utilizes the local window multi-head self-attention (W-MHSA) to minimize processing overhead and shifted window multi-head self-attention (SW- MHSA) to execute the data transmission. The main novelty of the ST is its hierarchical model. This is initiated by segmenting the input emotional images into patches that could be projected onto a low-dimensional subspace. Afterwards, the W-MSA mechanism captures the local attributes, whereas the SW-MSA technique improves interactions over windows, thus increasing the capability of the global model.

Since the network extends, the ST is gradually downsampling feature maps by patch integration, decreasing the spatial dimensions, while improving the feature extraction abilities, therefore increasing the computation effectiveness and image feature learning outcomes. This technique enables the model for preserving the lower computation intricacy while analyzing the large-scale meme images and is used for multimodal visual features in various semantic measures. This model is also incorporated the Squeeze-and-Excitation (SE) AM to gain the global contextual data and allocate the weights to every feature channel, thus permitting the model to concentrate on emotionally salient visual indications although minimalizing unrelated contextual redundancy. By applying the linear integration of such attributes, the fully connected (FC) layer gains deeper non-linear correlations. The global pooling layer has been utilized for preserving the main emotional feature data, and finally, it offers the last classification outcome of the memotion.

The window-based MHA is represented as

$$\text{W-MSA}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (4)$$

The channel recalculations in the SE attention mechanism can be shown below,

$$\text{z} = \sigma\big(W_2 \delta(W_1 \text{g})\big) \quad (5)$$

Here, the parameters $W_1$ and $W_2$ indicate learnable weight matrices, g represents the globally pooled feature vector, $\delta(\cdot)$ refers to the ReLU activation, and $\sigma(\cdot)$ indicates the sigmoid function that can be utilized for channel-wise features.

*2) Text feature encoder:* This subsection defines the textual feature encoding process using the RoBERTa-Large [15] model. RoBERTa-Large is a robust transformer-based masked language model (MLM) developed for greater performance in English NLP. This can improve the BERT by training under a great many different datasets, extended through the augmented hyperparameters for better accuracy in tasks, namely text classification. Later, this RoBERTa-large is used as the encoder for extracting the useful semantic representations of input texts. In order to represent each input, the vector with respect to the [CLS] location is utilized as the global representation. The RoBERTa-large embedding technique has been determined,

$$h[CLS] = RoBERTalarge(X)[CLS] \quad (6)$$

This mathematical form allows the efficient learning of emotionally dissimilar textual features, thus enhancing the robustness and generalization ability of the memotion classification architecture. Fig. 2 represents the model diagram of RoBERTa-Large.

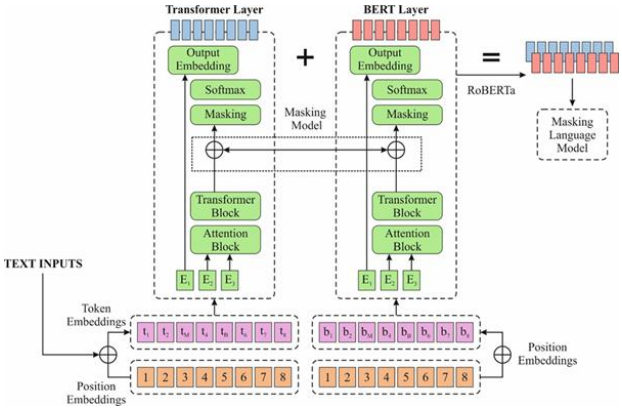



Fig. 2. Model diagram of RoBERTa-Large.

## D. Cross-Modal Attention-Based Fusion

This section presents the cross-modal attention module utilized for integrating visual and textual representations. In order to incorporate visual features, these are extracted by employing the enhanced ST and textual embeddings gained with RoBERTa-Large; a cross-modal attention fusion method has been applied. The image features could be projected to the textual data as well as arranged with the RoBERTa model for capturing the additional multimodal signs. This fusion allows the model to particularly identify useful image regions directed by textual data, while maintaining the text-dominant representation for downstream methods. The fusion process can be denoted by,

$$\mathrm{F}_{txt} = \mathrm{Attn}(\mathrm{Q} = \mathrm{T}, \mathrm{K} = \mathrm{V}, \mathrm{V} = \mathrm{V}) \quad (7)$$

Here, the parameter T represents RoBERTa-Large text embeddings, V indicates ST image features, and $\mathrm{F}_{txt}$ refers to the fused representation with respect to the textual feature space that can be employed for subsequent classification.

## E. Memotion Recognition Module

This section describes the memotion recognition method developed for classifying memes into corresponding emotional classes.

*1) BiLSTM:* It extended the traditional LSTM developed for obtaining the long-range dependency in successive data. This can be used as a secondary LSTM layer that can process the input sequences in reverse order. It could allow the network to gain both preceding and forthcoming contexts after examining data. The purpose of employing the BiLSTM [16] is demonstrated by its effectiveness in capturing the context from extended texts. The dataset has been handled by employing numerous layers, which can be developed for extracting useful features, and finally, the prediction is generated.

*a) Input layer*: The architecture takes an input layer that recognizes the sequences of textual data. The size of the input data is equivalent to the maximal length of the text sequences.

*b) Embedding layer*: The primary processing layer converts the dense vector from the input text. The embedding layer learns semantic correlations among words during training, as it is essential for text data.

*c) Bidirectional LSTM layer*: The output is determined from the embedding layer, which is handled by using the BiLSTM layer. Such a layer gains relative data from the previous and next words in the sequence. The hidden units that exist in the BiLSTM are formed, and each individual LSTM handles the sequence with respect to backward and forward directions. The BiLSTM produces an integrated vector with forward and backward contextual data for offering the comprehensive data of the sequences. The LSTM layer sequentially treats the input sequences with respective time steps and preserves a hidden state that gains data from prior time steps. The benefit of LSTMs is their capability to continue data for extended sequences, which could overcome the vanishing gradient complexity generally encountered by standard Recurrent Neural Networks (RNNs).

$$f_t = \sigma\big(W_f\,[h_{t-1}, x_t] + b_f\big) \quad (8)$$

$$i_t = \sigma(W_i\,[h_{t-1}, x_t] + b_i) \quad (9)$$

$$\widetilde{C_t} = \tanh\,(W_C \cdot [h_{t-1}, x_t] + b_c) \quad (10)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (11)$$

$$0_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (12)$$

$$h_t = 0_t * \tanh\,(C_t) \quad (13)$$

Here, the parameter $(x_\tau)$ refers to the input with respect to time step (4), $(h_{\tau-1})$ implies the hidden layer (HL) at the preceding time step, and the term $(C_\tau)$ represents the cell state with respect to the time step. The mathematical terms $(0_t)$ $(f_t)$ and $(i_\tau)$ are the output, forget, and input gates, respectively.

*d) Dropout layers*: Numerous dropout layers in the framework are utilized for preventing overfitting. Such dropout layers arbitrarily disable neuron parts during the learning procedure.

*e) Dense and output layers*: The network framework encompasses a sequence of FC layers, which can gradually improve feature extraction: a dense layer and ReLU activation. This dense layer with an individual element and sigmoid activation function produces the probability scores under classification methods. Numerous network frameworks are selected for this formation with a funnel model that can progressively learn representations, while maintaining the classification outcome by employing the ReLU activation for computational effectiveness and avoiding the vanishing gradient problem. An individual element in the last output layer integrates with the sigmoid activation to generate the probability scores for classification.

*2) Attention mechanisms*: The BiLSTM architecture is more effective in capturing long-term dependencies. During

long sequences, the BiLSTM faces the bottleneck complexity where the various contextual features can be removed. A SA technique has been employed to handle these difficulties. The SA method allows the model to concentrate on the more significant portions of the sequences, with respect to the length of the sequence. It is highly essential for identifying the patterns of emotion that could be distributed by long-range social networking content.

The attention mechanism is used to solve this difficulty by allowing the method to focus on particular areas of the sentences. This allocates various weights to every input image that depend on the significance of the context. The attention mechanism calculates the weight values for all the input elements in a sequence by employing scoring functions, which can be used on hidden states $h$, and then the softmax normalization is employed to ensure the correct weight distribution that offers precise attention to the sequence essentials.

Later, the BiLSTM layer processes the input sequences, generating concatenated hidden states. The mathematical form is $H = [h_1, h_2, \dots, h_n]$, where every term $h_i = [h_f^i, \cdot\, h_b^i]$ is integrated with the backward and forward LSTM hidden states in position $i$. The aforementioned bidirectional hidden states are obtained from all-inclusive contextual features. Considering the BiLSTM hidden states, the attention mechanism determines the attention scores,

$$e_i = W_a^T \tanh\left(W_h h_i + b_a\right) \tag{14}$$

Here, the parameter $d$ indicates the attention dimension, mathematical terms $W_a \in \mathbb{R}^{dx}1$ and $W_h \in \mathbb{R}^{dxh}$ describes learnable parameters, and $b_a$ refers to the bias term.

The attention weights have been standardized by applying the SoftMax, and the last context vector could be represented by

$$a_i = \frac{\exp(e_i)}{\Sigma_{j=1}^{n} \exp(e_j)} \tag{15}$$

$$c = \Sigma_i^{=1\mathrm{n}}\, \alpha_i h_i \tag{16}$$

This model is vital for efficiently identifying emotionally salient and task-relevant cues, which may appear at arbitrary positions within long textual sequences. Therefore, the proposed attention framework enables the model to concentrate on informative features, thus refining the robustness of memotion classification.

### *F. Optimization Strategy with Nadam*

This section outlines the optimization strategy integrated to train the proposed system effectively. Nadam [17] basically fuses Adam with the Nesterov momentum. This can be integrated with Adam's adaptive learning rates (ALR) with the "look-ahead" aspects of the Nesterov momentum (NAG) for accomplishing quicker and more consistent convergence. The main invention of NAdam is simulating the NAG's "look-ahead" update. In the NAG process, a "look-ahead" term can be primarily obtained for a temporary point,

$$\theta_{temp} = \theta_{t-1} - \beta_1 * \frac{m_{t-1}}{1 - \beta_1^{t-1}} \tag{17}$$

Later, the gradient can be determined instead of the temporary point to prevent redundant evaluation. The Nadam is gains look-ahead through accurate calculation. The look-ahead momentum estimation is given by the term $\widehat{m}_t$,

$$\widehat{m}_t^{nesterov} = \beta_1 \cdot \frac{m_t}{1 - \beta_1^t} + (1 - \beta_1) \cdot \frac{g_t}{1 - \beta_1^t} \tag{18}$$

The mathematical form calculates the momentum state with respect to the look-ahead point $\theta_{temp}$ through weighting the existing bias-corrective momentum $\widehat{m}_t^{nesterov}$ in the decay coefficient of the subsequent step $\beta_1^{t+1}$, and combining it with the existing gradient $g_t$. After that, the look-ahead momentum estimation is changed to the momentum estimation of Adam. The last parameter updates can be given by,

$$\theta_t = \theta_{t-1} - \alpha \cdot \frac{\widehat{m}_t^{nesterov}}{\sqrt{\widehat{v}_t} + \epsilon'} \tag{19}$$

Now, the 2nd moment estimation $v_t$ and the bias correction $\widehat{v}_t$ are retained as similar to the Adam method. The robustness of the model lies in these two inheritances.

The Nadam keeps Adam's benefits—ALR efficiency for sparse gradients and unrelated features—while acquiring the NAG look-ahead features. It can enable the motion identification of gradient variations, like increasing the convergence, decreasing fluctuations, and approaching the optimum.

TABLE IV. PARAMETER SETTING

| Module | Hyperparameter | Symbol | Tuned Values |
|---|---|---|---|
| Training (Common) | Optimizer | – | Nadam |
| | Learning Rate | η | 0.002 |
| | Beta1 | $\beta_1$ | 0.9 |
| | Beta2 | $\beta_2$ | 0.999 |
| | Epsilon | ε | 0.0001 |
| | Batch Size | B | 32, 64 |
| | Epochs | E | 200 |
| | Early Stopping Patience | – | 5 |
| | Gradient Clipping | – | 1.0 |
| Embedding Layer | Embedding Dimension | $d_e$ | 200 |
| | Max Sequence Length | L | 60 |
| BiLSTM Layer | Number of BiLSTM Layers | Lb | 1 |
| | Hidden Units | h | 128 |
| | Dropout Rate | – | 0.3 |
| | Recurrent Dropout | – | 0.3 |
| | Return Sequences | – | True |
| Self-Attention Layer | Attention Units | a | 64 |
| | Attention Heads (if multi-head) | H | 3 |

| | | | |
|---|---|---|---|
| | Attention Dropout | – | 0.2 |
| Fully Connected Layer | Dense Units | d | 64 |
| | Activation | – | ReLU |
| | Dropout | – | 0.4 |

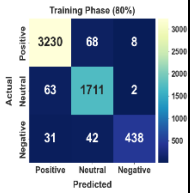
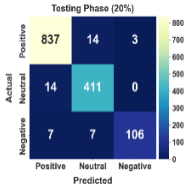
Table IV denotes the hyperparameters utilized for a DL approach integrating embedding, BiLSTM, self-attention, and fully connected layers. The model has been trained with the Nadam optimizer at a learning rate of 0.002, with batch sizes of 32 or 64 for up to 200 epochs, with early stopping after 5 epochs without expansion, and gradient clipping set to 1.0. The embedding layer maps input sequences of up to 60 tokens into 200-dimensional vectors. The BiLSTM layer has one layer with 128 hidden units, a dropout and recurrent dropout of 0.3, and returns sequences for attention. The self-attention layer utilizes 64 attention units across 3 heads with 0.2 dropout, while the final FC layer has 64 units with ReLU activation and 0.4 dropout to produce the output.

## IV. Performance Evaluation and Analysis

In this section, the empirical validation of the proposed AICAMA-ILSMC method is assessed using the Memotion Dataset 7k. Fig. 3 illustrates the sample of original and pre-processed images. Table V describes the text classifier results of the AICAMA-ILSMC system on 80:20 of TRAP/TESP. In the confusion matrices, the AICAMA-ILSMC technique proficiently classifies each class. In addition, in the text classification outcomes with 80% TRAP, the AICAMA-ILSMC system achieves $accur_y$ of 97.45%, $preci_n$ of 96.30%, $recal_l$ of 93.25%, $F1_{score}$ of 94.64% and $AUC_{Score}$ of 95.43%. At the same time, at 20% TESP, the AICAMA-ILSMC technique accomplishes $accur_y$ of 97.86%, $preci_n$ of 96.65%, $recal_l$ of 94.35%, $F1_{score}$ of 95.42% and $AUC_{Score}$ of 96.13%, respectively.

Fig. 3. Sample of original and pre-processed images.

TABLE V. Confusion Matrices and Text Classifier Results of AICAMA-ILSMC Method with 80:20

| Confusion Matrix | Classes | $Accur_y$ | $Preci_n$ | $Recal_l$ | $F1_{score}$ | $AUC_{score}$ |
|---|---|---|---|---|---|---|
|  | Positive | 96.96 | 97.17 | 97.70 | 97.44 | 96.80 |
| | Neutral | 96.87 | 93.96 | 96.34 | 95.13 | 96.73 |
| | Negative | 98.52 | 97.77 | 85.71 | 91.35 | 92.76 |
| | Average | 97.45 | 96.30 | 93.25 | 94.64 | 95.43 |
|  | Positive | 97.28 | 97.55 | 98.01 | 97.78 | 97.08 |
| | Neutral | 97.50 | 95.14 | 96.71 | 95.92 | 97.27 |
| | Negative | 98.78 | 97.25 | 88.33 | 92.58 | 94.05 |
| | Average | 97.86 | 96.65 | 94.35 | 95.42 | 96.13 |

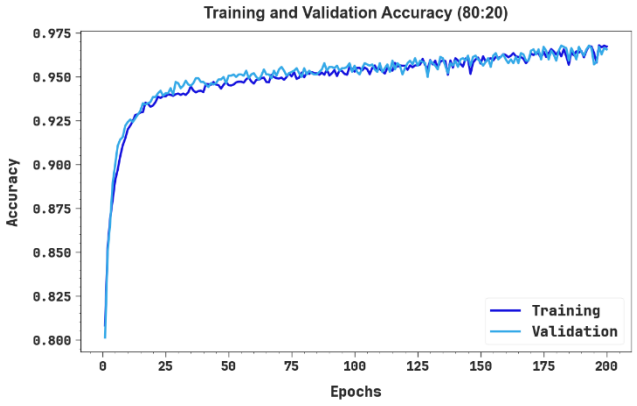



Fig. 4. $Accur_y$ curve of the AICAMA-ILSMC approach with 80:20.

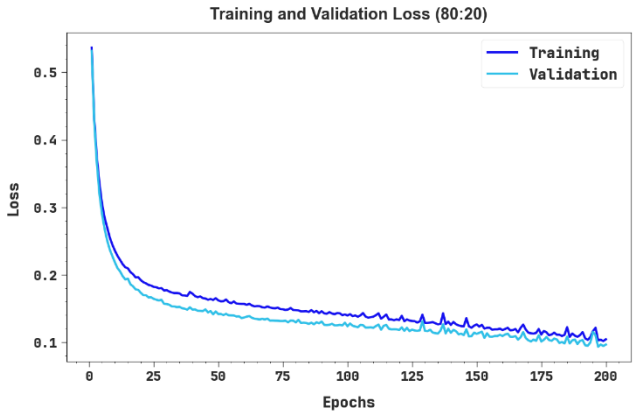



Fig. 5. Loss curve of AICAMA-ILSMC system with 80:20.

Fig. 4 portrays the training (TRAING) and validation (VALIDN) accuracy curves of the AICAMA-ILSMC framework under an 80:20 training-testing split over 200 epochs. TRAING accuracy rises gradually, which denotes efficient

learning and convergence. The VALIDN accuracy is closer to the training accuracy, with a minor and steady gap, signifying better generality and minimal overfitting.

Fig. 5 shows the TRAING and VALIDN loss curves of the AICAMA-ILSMC system with an 80:20 training–testing split over 200 epochs. Both losses reduce abruptly in the initial epochs, indicating rapid learning, and then progressively decline steadily, indicating robust convergence. The consistent overlap between training and VALIDN loss, with a small, reliable gap, implies minimal overfitting and robust generality.

Fig. 6 demonstrates the ROC curve of the AICAMA-ILSMC model. The results denote that the AICAMA-ILSMC technique achieves better ROC outcomes on every class, exhibiting significant potential in class differentiation. This regular pattern of superior ROC values over several classes specifies the efficacy of the AICAMA-ILSMC system on class prediction.

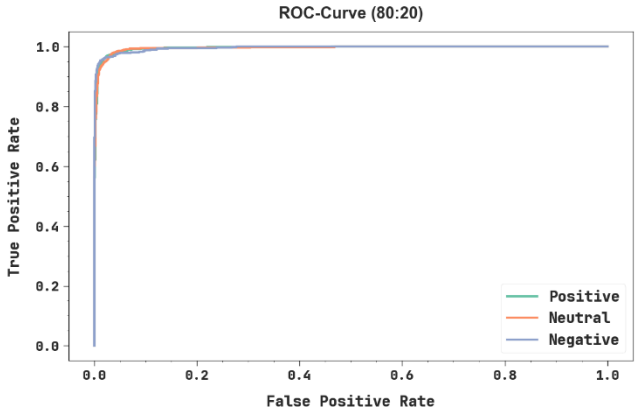


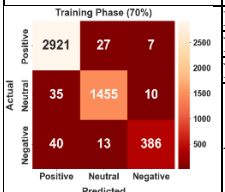
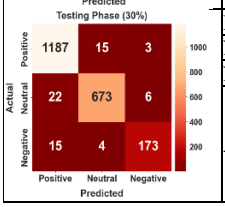

Fig. 6. ROC curve of the AICAMA-ILSMC method with 80:20.

Table VI defines the classifier outcomes of the AICAMA-ILSMC model at 70:30 of TRAP/TESP. In the confusion matrices, the AICAMA-ILSMC system accurately recognizes all classes. Additionally, in text classification with 70% TRAP, the AICAMA-ILSMC technique attains $accur_y$ of 98.20%, $preci_n$ of 96.87%, $recal_l$ of 94.59%, $F1_{score}$ of 95.67% and $AUC_{Score}$ of 96.39%. Likewise, at 30% TESP, the AICAMA-ILSMC system accomplishes $accur_y$ of 97.93%, $preci_n$ of 96.43%, $recal_l$ of 94.87%, $F1_{score}$ of 95.63% and $AUC_{Score}$ of 96.44%, correspondingly.

Fig. 7 illustrates the training and VALIDN accuracy curves of the AICAMA-ILSMC system on 70:30 training–testing split over 200 epochs. TRAING accuracy surges progressively, implying robust learning and convergence. The VALIDN accuracy closely follows the training trend with a small and steady gap, signifying robust generality and minimal overfitting.

Fig. 8 explains the TRAING and VALIDN loss curves of the AICAMA-ILSMC system by employing a 70:30 training–testing split across 200 epochs. Both losses reduce abruptly during the initial epochs, demonstrating quick learning, and then gradually decrease smoothly, displaying efficient convergence. The consistent overlap among TRAING and VALIDN loss, with only minor differences, implies lower overfitting and better generality.

TABLE VI. CONFUSION MATRICES AND TEXT CLASSIFIER RESULTS OF AICAMA-ILSMC METHOD WITH 70:30

| Confusion Matrix | Classes | $Accur_y$ | $Preci_n$ | $Recal_l$ | $F1_{score}$ | $AUC_{score}$ |
|---|---|---|---|---|---|---|
|  | Positive | 97.77 | 97.50 | 98.85 | 98.17 | 97.49 |
| | Neutral | 98.26 | 97.32 | 97.00 | 97.16 | 97.91 |
| | Negative | 98.57 | 95.78 | 87.93 | 91.69 | 93.77 |
| | Average | 98.20 | 96.87 | 94.59 | 95.67 | 96.39 |
|  | Positive | 97.38 | 96.98 | 98.51 | 97.74 | 97.18 |
| | Neutral | 97.76 | 97.25 | 96.01 | 96.63 | 97.32 |
| | Negative | 98.67 | 95.05 | 90.10 | 92.51 | 94.82 |
| | Average | 97.93 | 96.43 | 94.87 | 95.63 | 96.44 |



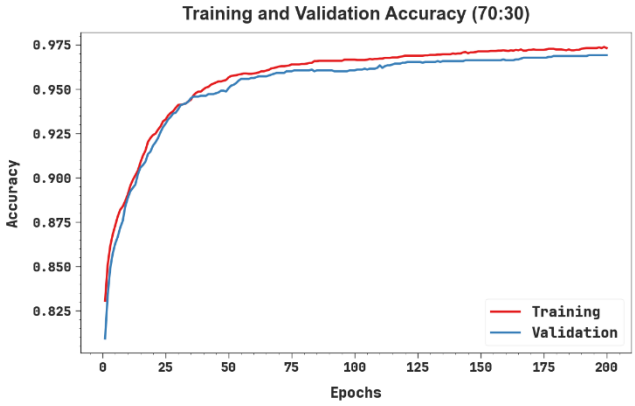



Fig. 7. $Accur_y$ curve of the AICAMA-ILSMC approach with 70:30.

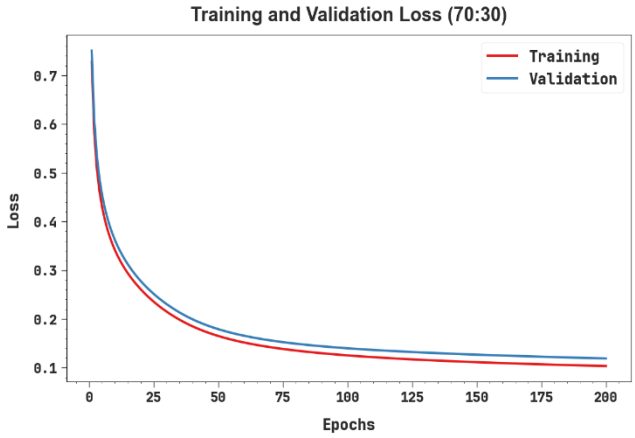



Fig. 8. Loss curve of the AICAMA-ILSMC technique with 70:30.

Fig. 9 portrays the ROC curve of the AICAMA-ILSMC method. The experimental results indicate that the AICAMA-ILSMC model achieves improved ROC values for each class, exhibiting significant potential for class discrimination. This steady pattern of enhanced ROC values over several classes shows the efficiency of the AICAMA-ILSMC system in class prediction.

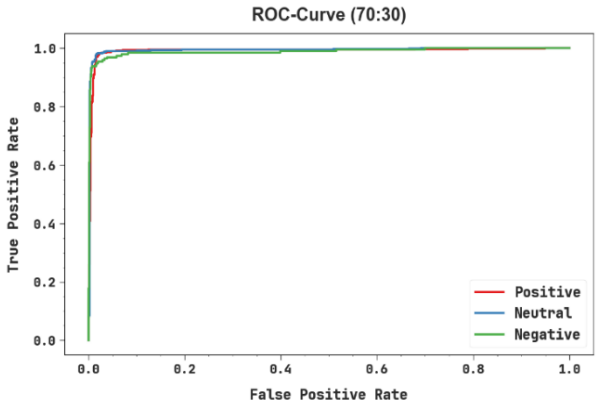



Fig. 9. ROC curve of the AICAMA-ILSMC method with 70:30.

The comparison assessment of the AICAMA-ILSMC method with existing techniques is shown in Table VII [18]. The simulation results show that the AICAMA-ILSMC method performed well.

TABLE VII. COMPARISON STUDY OF AICAMA-ILSMC METHOD WITH EXISTING MODELS

| Models | $Accur_y$ | $Preci_n$ | $Recal_l$ | $F1_{score}$ | Computational Time (sec) |
|---|---|---|---|---|---|
| BERT | 63.30 | 79.26 | 79.10 | 87.04 | 4.94 |
| ViT | 67.13 | 79.30 | 80.59 | 93.23 | 4.29 |
| MMBT | 63.37 | 81.73 | 76.66 | 80.88 | 6.29 |
| Resnet+LSTM | 40.00 | 93.40 | 79.75 | 76.49 | 2.77 |
| Kernel SVM | 88.00 | 88.07 | 75.88 | 77.98 | 5.32 |
| ViT+Word2Vec | 91.80 | 79.74 | 86.57 | 90.94 | 5.74 |
| ResNet-50+Glove | 88.60 | 89.47 | 78.80 | 88.89 | 4.47 |
| AICAMA-ILSMC | 98.20 | 96.87 | 94.59 | 95.67 | 0.83 |

Fig. 10 displays the comparative outcomes of the AICAMA-ILSMC model with various measures. The ResNet+LSTM model has lower performance with $accur_y$ of 40.00%, $preci_n$ of 93.40%, $recal_l$ of 79.75%, $F1_{score}$ of 76.49%. Moreover, the BERT, ViT, and MMBT approaches have shown slightly higher outcomes. Also, Kernel SVM, ViT+Word2Vec, and ResNet-50+GloVe methodologies have performed reasonably and achieved closer results. Finally, the AICAMA-ILSMC approach has attained the highest performance with $accur_y$ of 98.20%, $preci_n$ of 96.87%, $recal_l$ of 94.59%, $F1_{score}$ of 95.67%.

Fig. 11 represents the comparative outcomes of the AICAMA-ILSMC system, which is detailed based on computational time (CT). The results imply that the AICAMA-ILSMC method attains improved performance. Under CT, the AICAMA-ILSMC method attains a lower CT of 0.83sec, while the BERT, ViT, MMBT, ResNet+LSTM, Kernel SVM, ViT+Word2Vec, and ResNet-50+Glove methodologies attain greater CT values of 4.94sec, 4.29sec, 6.29sec, 2.77sec, 5.32sec, 5.74sec, and 4.47sec, respectively. Based on the in-depth results and discussion, the AICAMA-ILSMC model exhibits superior performance compared to others, confirming its effectiveness in memotion classification tasks.

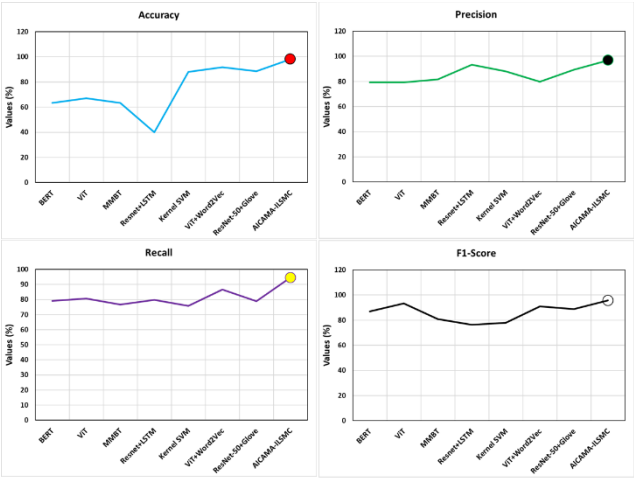



Fig. 10. Comparative analysis of the AICAMA-ILSMC method with measures.

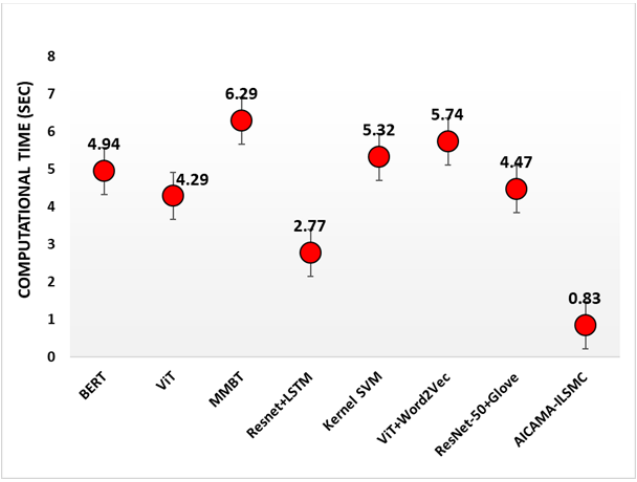



Fig. 11. Comparative result of the AICAMA-ILSMC technique with CT.

## V. CONCLUSION

In this study, the AICAMA-ILSMC approach has been designed for efficiently analyzing and classifying memes by jointly utilizing visual and textual contexts in a unified learning architecture. In the data preprocessing phase, BF is used for image noise removal. At the same time, textual content is preprocessed through OCR, lowercasing, punctuation and stop-word removal, tokenization, and handling slang, emojis, and abbreviations. For feature extraction, an improved ST is deployed to capture discriminative visual features, while RoBERTa-Large is utilized to acquire deep contextualized text embeddings. Then, the multi-modal representations are integrated using a cross-modal attention-based fusion strategy, enabling effective interaction between visual and embedded textual information, which produces a text-dominant fused representation. Furthermore, the fused features are classified using a BiLSTM-SAM for precise memotion classification. Lastly, the Nadam optimizer is incorporated in the proposed model to improve convergence stability and overall model performance. An extensive experimental analysis is conducted using the benchmark dataset, and the outcomes are examined under numerous measures. The comparison assessment

demonstrated the enhanced performance of the AICAMA-ILSMC method in accurately classifying memotion into corresponding classes, such as Negative and Very Negative, Positive and Very Positive, and Neutral, when compared to other recent approaches, achieving an accuracy of 98.20%.

## Data Availability Statement

The data that support the findings of this study are openly available in the Kaggle repository at https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k/data, reference number [14].

## References

[1] M. J. Glennon, Free speech and turbulent freedom: The dangerous allure of censorship in the digital era. Oxford University Press, 2021.

[2] T. Azar, H. M. Noori, A. R. Mahlous, A. Al-Khayyat, and I. K. Ibraheem, "Quasi-Reflection Learning Arithmetic Firefly Search Optimization with Deep Learning-based Cyberbullying Detection on Social Networking," Eng. Technol. Appl. Sci. Res., vol. 14, no. 5, pp. 17162–17169, Oct. 2024.

[3] Poon, S.T., 2022. Telling a Compelling Story: An Exploration of Cognitive Simplicity in Comic Book Design and Characterisation as Visual Communication for Political, Cultural and Social Influence. Wacana Seni Journal of Arts Discourse, 21.

[4] W. Ahmed, A. Ali, G. Hossain, M. Husain, and A. Mahmoud, "A Comparative Analysis of Machine Learning and Deep Learning Approaches to Enhanced Fake News Detection," Eng. Technol. Appl. Sci. Res., vol. 16, no. 2, pp. 33318–33322, Apr. 2026.

[5] D. M. Beskow, S. Kumar, and K. M. Carley, "The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning," Inf. Process. Manage., vol. 57, no. 2, p. 102170, 2020.

[6] Jaganath, M., Jothika, R., Keerthika, S., Pranishka, N. and Vinoharsitha, A.S., MemeCheck: Automated Meme Analysis for Identifying Offensive Text and Visuals. Proceedings Copyright, 676, p.682.

[7] ANTU, S.R.R., MUNNA, A.I., KHALIL, M.I., ABYAD, S.M., SIFAT, N.M., DHRUBO, A.F.H., QAYUM, M.A. and SAJJAD, M., 2026. Multimodal Sentiment Classification of Bengali Memes on MemoSen with Custom Fusion Models.

[8] S. Sharma, S. Ramaneswaran, M. S. Akhtar, and T. Chakraborty, "Emotion-aware multimodal fusion for meme emotion detection," IEEE Trans. Affect. Comput., vol. 15, no. 3, pp. 1800–1811, 2024.

[9] Aishy, A.T., Zanvi, S.H., Ali, M.S., Hossen, M.M., Rafi, M.S.M. and Islam, M.A., 2025. Memotion Analysis: Multimodal Fusion Techniques for Humor Classification in Memes.

[10] Y. Shi, W. Meng, Z. Guo, C. Wei, and W. Chen, 2025. Enhancing Meme Emotion Understanding with Multi-Level Modality Enhancement and Dual-Stage Modal Fusion. arXiv preprint arXiv:2511.11126.

[11] F. Breiteneder, M. Belal, M. S. Saeed, S. Masoudian, U. Naseem, K. Juhi, et al., 2026. Robust Harmful Meme Detection under Missing Modalities via Shared Representation Learning. arXiv preprint arXiv:2602.01101. https://doi.org/10.1145/3774905.3795610.

[12] E. Asmawati, A. Saikhu, and D. O. Siahaan, 2025, October. Sentiment Analysis of Meme Images Using Deep Neural Network Based on Keypoint Representation. In Informatics (Vol. 12, No. 4, p. 118). MDPI. https://doi.org/10.3390/informatics12040118.

[13] E. D. Wahyuni, T. L. M. Suryanto, and H. Arviani, "Deep Learning Multimodal Sarcasm Detection in Social Media Comments: The Role of Memes and Emojis," Journal of Artificial Intelligence and Technology, vol. 5, pp. 192–201, 2025.

[14] https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k/data

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[16] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997.

[17] S. Ruder, 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

[18] M. Velmala, S. Rajiakodi, K. Pannerselvam, and B. Sivagnanam, "Multimodal Sentiment Analysis of Online Memes: Integrating Text and Image Features for Enhanced Classification," Procedia Comput. Sci., vol. 258, pp. 355–364, 2025.