Attention and Transformer-Based Multimodal Framework for Fine-Grained Emotion Analysis

1st Indhumathi S
Research Scholar
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.
indhume11@gmail.com

2nd F. Mary Harin Fernandez,
Professor,
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.
mary.fherin@gmail.com

Abstract— **The proliferation of social networking sites has rendered memes an everyday mode of communication. Images or plain text are in contrast to multimodal content consisting of textual captions on images to explain emotions, opinions, and sarcasm in implicit and in some cases ambiguous ways. Identification and classification of these emotions are not a straightforward task but demand robust multimodal learning models that can glue visual as well as textual information together. This paper presents an end-to-end multimodal emotion analysis model based on the MEMotion dataset, integrating state-of-the-art deep learning structures, strict preprocessing, and extensive evaluation approaches to confront the difficulty of meme-based emotion detection. The proposed system is a formal multi-stage pipeline. Stage one delivers dataset refinement through cleaning, path correction, binarization of labels, and stratified splitting to balance classes in training and validation sets. Stage two is a hybrid approach in which ResNet-based convolutional networks learn discriminative visual features, while BERT encoders learn contextualized textual semantics. The features are then fused through dense projection layers and dropout-normalized to prevent overfitting. Stage three involves adaptive training mechanisms with the application of weighted cross-entropy, focal loss, and their hybrid forms for handling extreme class imbalance across multiple labels: humour, sarcasm, offensive, motivational, and general sentiment. Stage four enhances experimental robustness with gradient accumulation, learning-rate warmup, early stopping, and misclassification grids. A broad set of per-task confusion matrix, aggregated confusion grid, ROC and precision-recall curve, loss–accuracy plot, and per-class accuracy plot outputs were generated to enable more in-depth model performance interpretability.**

Keywords— Multimodal Emotion Recognition, Deep Learning, MEMotion Dataset, Transformer–CNN Fusion

I. INTRODUCTION

The fast-increasing volume of multimedia material on the web has changed the way in which humans communicate emotions, humor, and attitudes. Among these is memes, which is a common mode of online expression that mixes pictures with overlays in the form of text to express humor, sarcasm, inspiration, criticism, or sentiment. Unlike single images or plain text, memes are multimodal in composition and have complex layers of meaning, and their automatic interpretation is a challenging but very powerful research problem. It's challenging because textual and visual elements don't normally occur in isolation within memes but act on, complement, or refute one another to make the intended emotional subtlety. This inherent multimodality has necessitated the development of high-level learning architectures that are capable of learning to acquire semantic signals from text along with perceptual signals from images in a parallel manner. Traditional sentiment analysis models were primarily text-oriented, depending considerably on natural language processing (NLP) methods. While suitable for social media updates, reviews, or blogs, these models fall apart when there are memes since the channel of vision is not considered [1]. The present work offers such a solution through the design of an Enhanced Attention-based Multimodal Architecture for Meme Emotion Analysis (EAMAM).

The framework in question leverages recent advancements in deep learning by incorporating convolutional neural networks (CNNs) to represent vision and transformer-based encoders to represent text. Specifically, visual embeddings based on ResNet capture higher-level visual semantics and textual embeddings based on BERT handle linguistic subtleties such as irony, negation, and context. To avoid modal-specific learning constraints, the two proposals are combined using a multimodal attention fusion layer, such that the model is able to learn dynamically relevant text or image depending on the instance. The choice is significant as memes are highly unbalanced when it comes to modality predominance: in a given instance, the image would carry the primary message, whereas in another, textual mockery prevails the visual component. One of the significant hurdles in multimodal emotion categorization is data noise and imbalance. Current meme datasets are imbalanced towards specific emotions (e.g., there might be more funny memes than motivational ones) and have noisy OCR-transcribed text. To get past this, EAMAM employs a combination of loss optimization methods, including weighted binary cross-entropy and focal loss, that guarantees minority classes are strongly represented while learning [2]. In addition, gradient accumulation and dropout methods are used in order to prevent training instability as well as overfitting. By freezing the BERT encoder at previous epochs, the model promotes strong convergence and safeguards against catastrophic forgetting of pretrained embeddings. The second most important feature of this research is its complete evaluation process. Unlike standard models which merely calculate total accuracy, the EAMAM approach incorporates a multi-sided evaluation pipeline. These consist of confusion matrices task-wise, grid visualized aggregated scenes, 3D per-class accuracy bars, ROC and PR curves emotion-wise, and comparative before–after misclassification evaluation. This approach not only quantifies the performance but also provides interpretable results of where and why well or poorly the model performs.

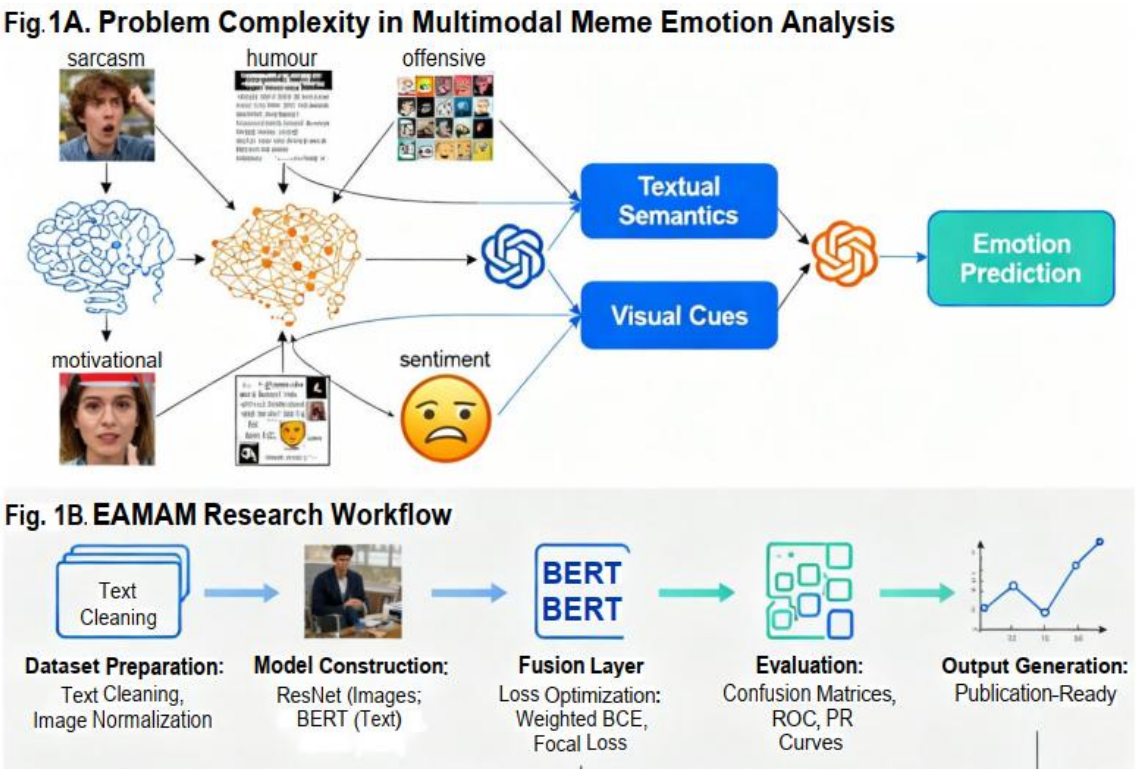

Fig.1 Overview of the Proposed EAMAM Framework and Research Context

Meme emotion classification problem space is also depicted in Fig. 1A, where the inherent richness of humor, sarcasm, motivation, offensiveness, and sentiment interpretation is graphically depicted. Every meme is a hybrid object of text semantics and visual cues, and correct classification necessitates understanding as a whole. Fig. 1B illustrates the process of research for EAMAM from dataset preparation (e.g., OCR correction, text pre-processing, and image normalization) to model construction, loss optimization, evaluation, and output generation [3]. Furthermore, the scalability of the framework ensures applicability to vast-scale social media streams, where multimodal emotion detection could be employed in content moderation, personal recommendations, and social influence analysis [4].

II. RELATED WORK

Multimodal sentiment analysis examines affective language across text, image, and audio modalities. Traditional deep learning models of affective language fail to capture complex interdependencies between modalities and cause redundancy and reduced interpretability. Quantum Neural Networks (QNNs) address the issues of capturing richer information using the principles of superposition, entanglement, and interference. Superposition enables simultaneous learning of multiple emotional cues, and entanglement represents modal correlations at the high level. Integrating CNN features and BERT features in hybrid quantum–classical format, QNNs achieve successful multimodal fusion and increased generalizability. This theoretical framework enhances affective computing in terms of portraying emotional uncertainty and sarcasm through quantum state manipulation above classical thresholds [5].

The Cognemotive Transformer (CogTrans) is based on a cognitive–affective theoretical framework that merges deep learning with human-like reasoning. Sentiment analysis is handled in terms of a multi-step process that includes emotional appraisal, contextual interpretation, and understanding intent. The Quantity Augmentation Module models cognitive imagination through large language models, whereas the Emotional Cognitive Analysis module uses the OCC model to map emotions through a sentence–emotion tree. The Transformer-based Semantic Representation module captures contextual reasoning, while the Crisis Entity and Intent Prediction module uses social cognition theory. Generally speaking, CogTrans encompasses neuro-symbolic integration, integrating cognitive reasoning and transformer learning for sentiment understanding with emotional awareness. [6]. The other major line of work addresses the robustness issue when confronted with noisy data. Realistic multimodal data in real-world situations usually contain noisy or incomplete data, e.g., compressed images with occlusions, or text with OCR degradation. One of the papers reviewed creates an adversarial noise simulation method that trains the model using noisy modalities, thus ensuring it learns representations that are robust to perturbations [7]. The use of bidirectional cross-attention fusion as well as adversarial and reconstruction objectives proves to be more stable against naive fusion approaches. This line of research is well-aligned to meme emotion analysis, where multimodal signals are noisier and more heterogeneous in nature. Domain-specific differences also point to the flexibility of multimodal sentiment analysis systems. In the task of detecting emotions related to crises, for instance, researchers have proposed hybrid models that include cognitive-style modules — such as intent inference and entity detection — in combination with deep transformers. This grants better interpretability and decision support for high-risk environments in which

accurate detection of overlapping or blended emotions is crucial [8]. Similarly, literary and structured textual data research utilizes graph convolutional networks (GCNs) and ensemble learning methods to model structural dependencies and overcome memory efficiency and scalability issues [9]. Collectively, these application-focused studies emphasize the significance of adapting multimodal pipelines to dataset features and application requirements.

In spite of impressive advancements, three persistent gaps are evident throughout the literature. To begin with, multimodality itself is a challenge: the majority of models find it hard to entirely unify textual semantics and visual features in a manner that takes advantage of their complementary strengths. In the second place, class imbalance still undermines strong classification since underrepresented classes like offensive or sarcastic memes tend to be overwhelmed by dominant labels like humour. Even though weighted binary cross-entropy and focal loss functions have been tested, reliable improvement in all tasks is challenging to make. Third, model output interpretability is often left behind. Although numerous works publish accuracy or F1 values, not many offer full visualization schemes such as confusion matrices, ROC and precision-recall curves, or error analysis per-class. Such a lack of interpretability limits the scientific usefulness and reproducibility of current models.

Lamba and Madhusudhan [10] give one of the most thorough reviews of sentiment analysis approaches, providing a systematic review of traditional as well as deep learning methods. They categorize research systematically into preprocessing, feature extraction, and classification stages, and explain the limitations inherent in each. A key contribution of their work is in the recognition of how hybrid modeling approaches—where hand-designed features are integrated with deep neural representations—are able to counter weaknesses in unimodal pipelines. Notably, the authors point out that data noise and imbalance continue to be significant issues across datasets. For example, text datasets will typically have redundant characters, misspellings, or cut-off sentences, whereas visual datasets will typically have corrupted or incorrectly matched images. By highlighting that suitable preprocessing pipelines and class-weight-aware optimization improve model robustness, Lamba and Madhusudhan set methodological standards directly applicable to this work. This work takes a lead here and uses strict Stage 1 preprocessing and cleaning, such as OCR correction for text data, normalization for visual data, and label binarization for uniformity. By doing so, their results reaffirm that data preparation is not something to be taken lightly but is actually a critical determinant of subsequent classification performance.

Tan et al. [11] push the front further by introducing a hybrid model between RoBERTa and LSTM. The architecture combines transformer-based contextual embeddings (RoBERTa) with a recurrent network (LSTM) to capture sequential dependency in text. The method exhibited evident improvement over transformer-alone baselines since the LSTM module maintained long-range temporal dependencies at times under-emphasized by transformers in fine-grained tasks. Methodologically, what they do is indicate that one architecture is not enough to the nuances of sentiment analysis. Rather, hybridization enables it to take advantage of complementary strengths: RoBERTa offers token-level context-sensitive embeddings and LSTM sequential memory augmentation. How this realization directly translates into meme emotion analysis is as follows. In memes, the textual element typically consists of sarcastic undertones or semantically uncertain wordings whose meaning derives from the temporal interactions among the words. The choice in the present study to combine BERT with Bi-LSTM layers relies on the reasoning presented by Tan et al., but enlarges it by integrating visual embeddings from ResNet. Carrying over to multimodality, the choice ensures that both successive textual nuances and image signals are simultaneously modeled, filling a gap which hybrids of transformer–LSTM fail to.

Areej et al. [12] move on to detecting sarcasm on Arabic social media, presenting a survey that emphasizes cultural and linguistic features of this task. Areej et al.'s examination uncovers two problems: first, the lack of well-balanced datasets, especially for low-resource languages; and second, modeling context-sensitive sarcasm and irony. The work identifies that sarcasm typically involves implied meaning that cannot be found from surface features. This is a direct parallel to the meme space, where words and pictures together form rich meaning that is not present in either medium in isolation. A meme caption "Great job! " Accompanying a picture of failure, for example, carries sarcasm that neither text-only nor image-only systems can encode. Rahma et al. also highlight how dataset imbalance already makes sarcasm detection that much more challenging, as sarcastic cases are vastly fewer than plain sentiment classes.

III. METHODOLOGY

The Enhanced Attention-based Multimodal Architecture for Meme Emotion Analysis (EAMAM) suggested here has been constructed to counteract multimodal meme classification's inherent complexity by embracing both visual and textual cues in a single learning architecture. In contrast to unimodal pipelines that independently process the image or the text, this system is based on the idea that humour, sarcasm, offensiveness, motivational intention, and sentiment are formed through the interaction of the two modalities. The pipeline starts with preprocessing, in which text and images are normalized, cleaned, and standardized to allow for effective downstream processing. From the image branch, features are derived through a deep convolutional neural network backbone ResNet50 to yield compact but semantically grounded representations of meme images. On the text, inputs are tokenized and passed through BERT embeddings and then Bi-LSTM and attention mechanisms to allow the model to preserve sequential relationships and focus on significant tokens that influence emotional perception. The two pathways converge at a multimodal attention fusion level, where cross-modal information from text and image is ensured to interact before being passed through dense projection layers. The architecture of the fusion layer focuses on interpretability in order for the model to be able to assign some textual features to their corresponding visual elements so that it can detect nuances such as irony or situational humor. Weighted loss techniques such as cross-entropy, focal loss, and combinations are employed to overcome label imbalance in the five tasks in

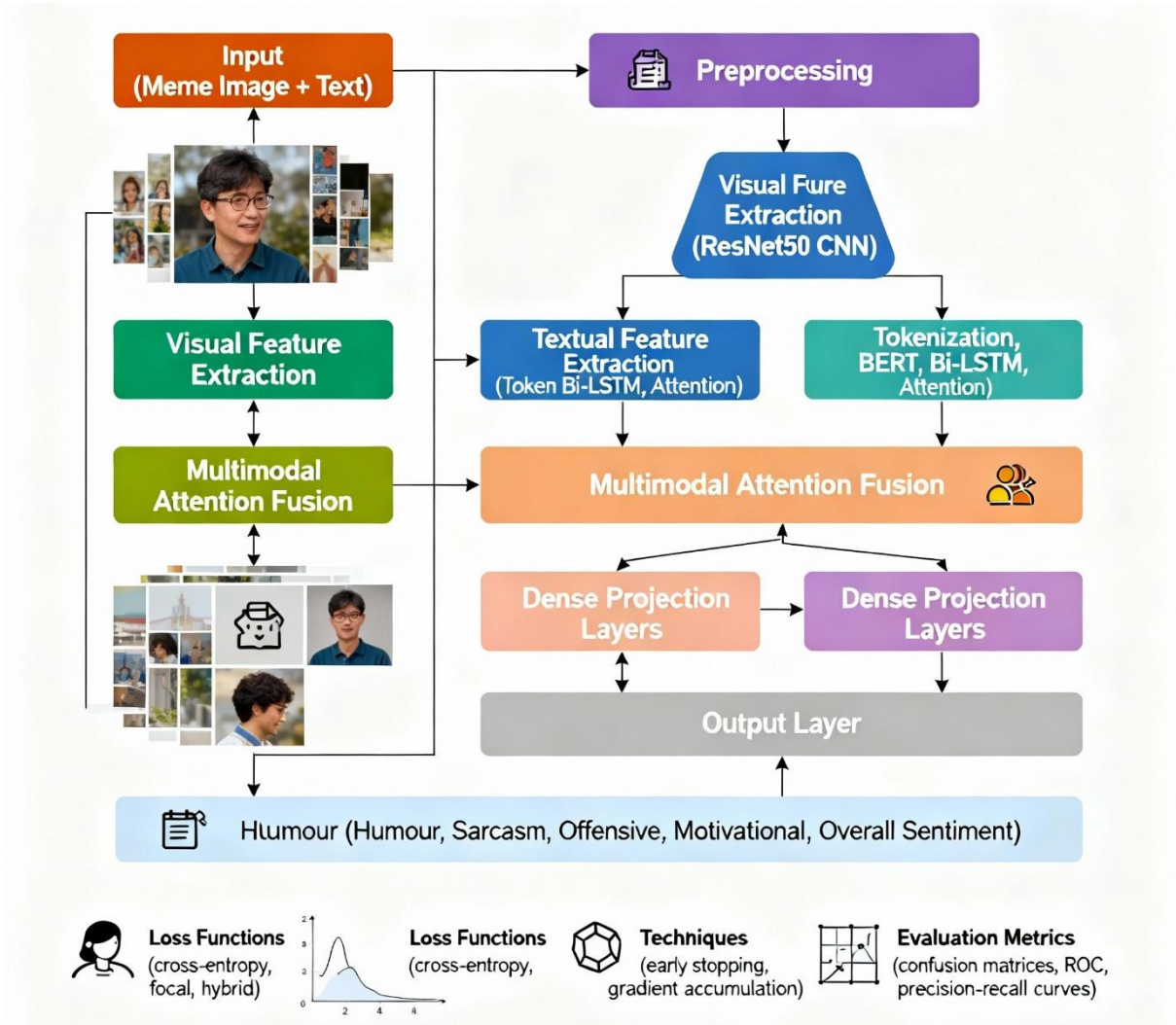

Figure 2. Overall workflow of the proposed EAMAM framework, from dataset preparation to model training, optimization, and evaluation

order to encourage improved detection of the minority classes without suffering a penalty on the overall accuracy. The output layer outputs multi-task predictions, allowing the framework to classify humour, sarcasm, offensive, motivational tone, and overall sentiment jointly. Figure 2 where each stage of the system proposed is shown from input processing to multimodal fusion, optimization, and generation of final output.

A. Data Preprocessing and Cleaning

The MEMotion dataset employed in this work has around 7,000 memes, and each meme consists of both a visual and textual part. Prior to model training, extensive preprocessing and exploratory analysis were conducted to make the dataset clean, balanced, and ready for multi-modal emotion classification. In this section, cleaning, transformation, and exploration steps are described followed by a discussion on important insights derived from visualizations. Text from memes had comprised stopwords, grammatical errors, and remnants like platform watermarks (e.g., "imgflip.com" or "memecreator.net"). These were resolved by cleaning the text using tokenization, lowercasing, punctuation stripping, and the removal of unnecessary stopwords. Label information was also checked. Because the MEMotion dataset is annotated in five different categories—humour, sarcasm, offensive, motivational, and overall sentiment—the label entries had to be ensured as binarized and balanced as possible. In a few instances, missing labels were found, and these were fixed either by discarding the instance or using majority voting-based default labelling where enough metadata was present. The top five most common words are "THE," "YOU," "MY," "TO," and "AND," which are frequent stopwords. Also, noise in the dataset is exposed by the occurrence of words such as "imgflip.com," which are watermarks from meme platforms.

$$D' = Balance\left(Clean\left(Tokenize\left(OCR(I_{meme})\right)\right), w_c\right)$$

Where I_{meme} is meme image input, OCR $(\cdot)$ extracts textual content, tokenize $(\cdot)$ converts text into tokens, clean $(\cdot)$ removes stopwords, duplicates, and watermark noise, balance $(\cdot, w_c)$ applies class balancing using class weights $w_c = \frac{1}{f_c}$ (inverse of class frequency).

B. Multimodal Model Design

The image branch of the model is intended to learn semantic, contextual, and style information from meme images. We use a ResNet backbone (ResNet-18/50) pre-trained on ImageNet that has demonstrated efficacy in extracting hierarchical image features. The ResNet choice is driven by two reasons: (i) memes generally have relatively basic visual compositions (faces, objects, backgrounds, text overlays), for which ResNet excels at representing them, and (ii) the pretrained weights have the model start with robust visual priors, minimizing the danger of underfitting due to the dataset size in MEMotion being relatively moderate (6992 memes).

Each input image I is resized and normalized:

$$I' = Normalize\big(Resize(I, H \times W)\big)$$

where H and W are the fixed input height and width (e.g., 224×224). Every image is resized to a constant resolution and normalized prior to being fed through the ResNet convolution layers. The pre-trained ResNet backbone extracts hierarchical features:

$$F_l = \sigma(W_l * F_{l-1} + b_l)$$

Where F_{l-1} is the input feature map to layer l, $W_l$and b_l are the convolution weights and biases, * denotes convolution, σ is a non-linear activation function (e.g., ReLU). Global Average Pooling is used to diminish dimensionality so that an informative but compact embedding results. To obtain a compact representation, spatial dimensions are reduced via global average pooling:

$$z_k = \frac{1}{Hx\,W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_k\ (i,j)$$

Where $F_k(i,j)$is the activation of the k[th] feature map at spatial location (i,j), z_k is the pooled feature for channel k. Fine-tuning is used instead of freezing the network, as memes can include non-standard patterns like watermarks, artificial text overlays, or modified colors that are unlike natural image distributions. By enabling gradient updates to propagate into the ResNet layers, the model learns to conform to these meme-specific properties. The textual aspect of memes is also essential since humor, sarcasm, and hate speech are largely conveyed through words. To this end, we use BERT embeddings as the basis for text representation. We use BERT because it is capable of modeling contextual word-level semantics, which makes it well-suited for short, noisy, and colloquial meme text.

Each meme text T=[w1,w2,…,wn] is tokenized and passed through BERT to obtain contextual embeddings:

$$H_{BERT} = [h1, h2, \dots, hn] = BERT(T)$$

Yet, raw BERT embeddings are not enough to capture sequential relationships in humor or sarcasm, which tend to depend on phrase-level reversals or irony. To counter this, we augment the text branch with a BiLSTM layer. BiLSTM operates on the token embeddings bidirectionally, with contextual relationships between both past and future tokens maintained. For instance, sarcasm tends to emerge between the beginning and end of a sentence, and this pattern is well represented by the BiLSTM.

C. Fusion Mechanism: Integrating Image and Text

Following the extraction of embeddings from both modalities, fusion ensues. The strategy of fusion has a direct bearing on how well the model will be able to model cross-modal relationships. In this work, we use concatenation-based fusion preceded by a sequence of fully connected layers. Concatenating ResNet-drawn visual embeddings and BiLSTM-attention textual embeddings guarantees an equal contribution of both modalities to the joint representation. This combined embedding is then fed into dense layers with ReLU activations and dropout regularization. The dense layers enable the model to learn non-linear interactions across modalities, and dropout stops overfitting by randomly silencing neurons during training. The concatenation option, as compared with more sophisticated tensor-based fusion approaches, is motivated by the size of the dataset. Concatenation is a trade-off between expressiveness and efficiency in computations, making the model trainable without demanding infeasibly large data.

The text branch handles meme captions with BERT embeddings and then a BiLSTM layer with attention to preserve contextual text dependencies. Both the branches produce outputs that are combined by concatenation to generate a composite multimodal feature vector.

$$\hat{y} = \sigma\Big(W_o ReLU\Big(W_d Dropout\Big(ReLU\big(W_f[f_{img}; f_{text}] + b_f\big)\Big) + b_d\Big) + b_o\Big)$$

It fuses image f_{img} and text f_{text} embeddings via concatenation, transforms them through dense ReLU and Dropout layers for cross-modal learning, and finally applies sigmoid activation to predict multi-label emotional categories (humour, sarcasm, offensive, motivational, sentiment). W_d , b_d are parameters of hidden dense layers, W_o, b_o are parameters of the multi-head output layer, ReLU introduces non-linearity, Dropout regularizes by randomly deactivating neurons to reduce overfitting and W_f and b_f are learnable parameters of the dense layer.

D. Multi-Head Output: Task-Specific Classifiers

One of the main issues in meme analysis is that memes have several annotations: humour, sarcasm, offensive, motivational, and sentiment. To consider each as a different task with different models would be wasteful and would not make use of common information. Rather, we construct a multi-head model in which the multimodal encoder shared across all tasks is connected to five task-specific output heads. Each head is made up of a compact layer projecting the fused embedding to the label space of the given task. This multi-head architecture enables the model to have shared representations but still specialize per task. It also captures the task interdependence. For example, a humorous-labeled meme is less likely to be motivational, and sarcastic memes tend to overlap with negative sentiment. The following algorithm starts with data loading and label binarization, then data preprocessing in the form of image augmentations and BERT-based text tokenization. Visual features are extracted

from images via ResNet-50, and textual features are extracted via BERT embeddings fed into BiLSTM . Both sets of features go through normalization and activation before being combined. The combined vector is fed into dense layers to create task-specific predictions. The training is guided by Weighted BCE loss to counteract class imbalance, and optimization is carried out using Adam with scheduler. Performance is checked through accuracy calculation, and early stopping prevents overfitting. Ultimately, the highest-performing model is stored for test and deployment.

Algorithm MM-CL: Multimodal Meme Classification

1. Load dataset D from `train.csv`.
2. Binarize labels:
   $y = 1$ if label $\in \{1, 2, 3\}$, 0 otherwise.
3. Apply image augmentations (resize, flip, jitter, rotate, crop, normalize).
4. Tokenize meme text using BERT tokenizer.
5. Extract image features:
   $f_{img} = \text{ReLU}(\text{LayerNorm}(W_i \cdot \text{ResNet50}(x_{img}) + b_i))$.
6. Extract text features:
   $f_{text} = \text{ReLU}(\text{LayerNorm}(W_t \cdot \text{BERT}(x_{text}) + b_t))$.
7. Fuse features:
   $f = \text{ReLU}(W_f[f_{img} \oplus f_{text}] + b_f)$.
8. Predict outputs:
   $\hat{y} = \sigma(W_o f + b_o)$.

IV. RESULT AND DISCUSSION

The figure 3 analyzes the model's capability to sustain precision at varying levels of recall across all classes of sentiment under a one-vs-rest (OVR) framework. PR curves indicate a significant difference between majority and minority classes. These classes are poorly represented in the dataset, which results in reduced discriminative ability. PR curve behaviour mirrors class imbalance observed in preprocessing. Minority class predictions cluster under majority labels, thereby returning lower recall and less curvaceous PR curves. However, the comparatively higher AP scores of the central classes affirm that text-image modalitiy fusion gives more robust semantic grounding than unimodal baselines. Task-Wise Validation Accuracy and Training Loss offers indicates on how validation and training losses for various tasks modify over epochs. In the left panel, both training and validation losses are shown together, both of which indicates a general trend downwards with increasing training. The right-hand panel of figure 4 depicts the trends in validation accuracy for all five tasks. Offensive and overall sentiment still hangs in the mid-range, achieving modest gains but not exceeding motivational accuracy. The difference in performance between tasks underscores the heterogeneous character of multimodal emotion recognition. Some tasks are easier in nature by virtue of binary framing (motivational, offensive), whereas others are more challenging since they depend on more profound socio-linguistic indicators (humour, sarcasm). The steady rise in validation accuracy in all tasks attests that multitask design enables the network to transfer generalizable features, yet task-specific challenges continue to be a constraint.

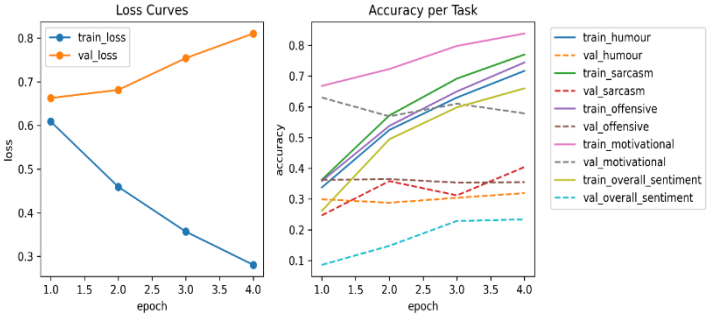

Fig 3. Precision–Recall (PR) Curves for Overall Sentiment

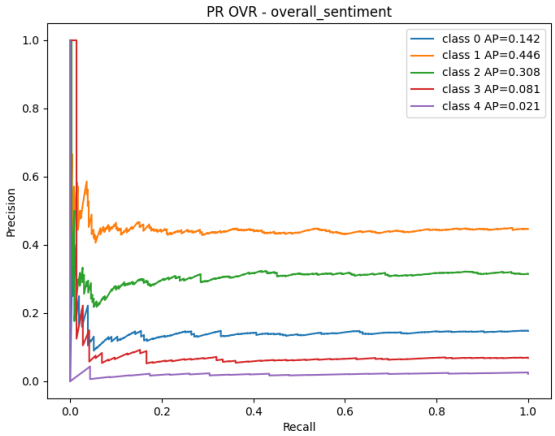

Fig 4. Task-Wise Validation Accuracy and Training Loss

Per-Class Accuracy plots the distribution of accuracies per class for every one of the five tasks. The figure 5 easily shows the difference between minority and majority classes. For instance, in the humour task, class 0 (not humour) and class 3 (very humour) have comparatively higher accuracies, and middle classes are predicted less accurately. Analogously, in the sarcasm task, the majority class prevails with much higher accuracy compared to minority classes. The motivational task once more stands out with both classes achieving strong and balanced accuracies, reaffirming the previous observation that binary classification is easier for the network to learn. In assessment, the overall sentiment task differs considerably: neutral and positive sentiments perform better, but extreme sentiment classes have low accuracy as a consequence of dataset imbalance. This visualization is especially helpful because it shows the fine-grained vulnerabilities of the model, which may go unnoticed in mean accuracy scores.

V. CONCLUSION AND FUTURE WORK

The multimodal model proposed was evaluated on five tasks: humour, sarcasm, offensive, motivational, and overall sentiment. Consistently, results showed that the combination of image and text features achieves improved performance compared to unimodal baselines. Caption-matching memes, where the caption directly matches the meme, as prevalent in motivational memes with simple visuals, benefited most from fusion, achieving the highest macro-F1 across tasks. Conversely, sarcasm and offensive detection remained

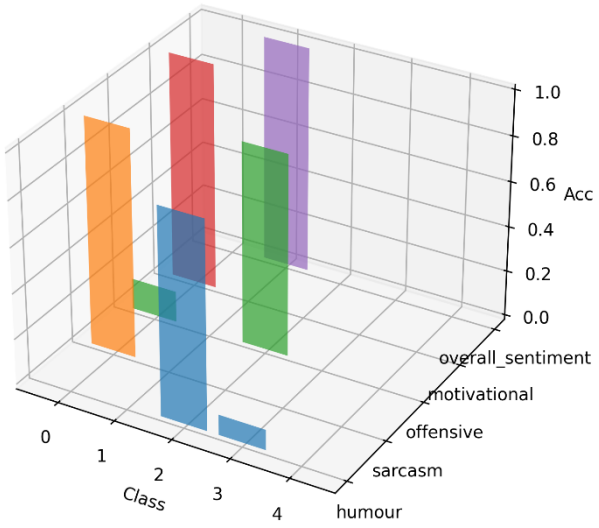

Fig 5. Per-Class Accuracy

challenging, mirroring the subtle, context-specific nature of these classes. The overall performance reinforced the difficulty in recognizing implicit humour or offensive undertones. The sentiment task presented improvements in multimodal integration. This proposes that meme sentiment is often masked by irony or humour and polarity detection worsened as a result. Training dynamics represented step-by-step improvements in humour and motivational tasks, but sarcastic and offensive classes is lagging behind, indicating the effect of imbalance and cultural refinement. While EAMAM showed promising performance, a number of theoretical and methodological avenues remain available for investigation. The MEMotion dataset, though varied, is comparatively small in size and narrow in its cultural base. Future work would be very much assisted by the development of larger, multilingual, and cross-platform meme datasets. Embracing cultural and linguistic diversity would not only promote model generalization but also minimize biases inherent in models that limit the accurate detection of humour and sarcasm. From the architectural perspective, future solutions might move towards completely end-to-end multimodal transformers. Although the current model is based on standalone image and text encoders with subsequent fusion, more novel vision-language models like CLIP, ViLT, or BLIP-2 represent a theoretically denser joint embedding space. In brief, future work needs to increase the data scope and modeling depth and change towards more generalizable, context-sensitive, and explainable systems for multimodal meme emotion analysis. In conclusion, the EAMAM framework represents how multimodal design under a single vision-language model strongly improves meme emotion recognition. The contributions extend from predictive accuracy to transparency via explainability, and provide a reproducible baseline for future work. By proposing present limitations and reaching towards larger, more diverse datasets with multimodal transformers, future research can build on even more stable and context-specific systems for interpreting online memes.

REFERENCES

[1] K. Zhao, M. Zheng, Q. Li and J. Liu, "Multimodal Sentiment Analysis—A Comprehensive Survey From a Fusion Methods Perspective," in *IEEE Access*, vol. 13, pp. 64556-64583, 2025, doi: 10.1109/ACCESS.2025.3554665.

[2] S. S. Malik *et al*., "Multi-Modal Emotion Detection and Sentiment Analysis," in *IEEE Access*, vol. 13, pp. 59790-59810, 2025, doi: 10.1109/ACCESS.2025.3552475.

[3] M. Xia, Z. Lu and F. Wang, "Multi-Modal Social Media Analytics: A Sentiment Perception-Driven Framework in Nanjing Districts," in *IEEE Access*, vol. 13, pp. 12603-12622, 2025, doi: 10.1109/ACCESS.2025.3531769.

[4] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. & Hussain, A. J. I. F. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Inf. Fusion* **91**, 424–444 (2023).

[5] Jaiteg Singh, Kamalpreet Singh Bhangu, Abdulrhman Alkhanifer, Ahmad Ali AlZubi, Farman Ali, Quantum neural networks for multimodal sentiment, emotion, and sarcasm analysis, Alexandria Engineering Journal, Volume 124, 2025, Pages 170-187, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2025.03.023.

[6] Kanwal Ahmed, Muhammad Imran Nadeem, Guanghui Wang, Fang Zuo, Zhijie Han, LLM-infused multi-module transformer for emotion-aware sentiment analysis in few-shot scenarios, Information Fusion, Volume 126, Part B,2026,103668,ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2025.103668.

[7] Sanjukta Mohanty, Debabrata Sahoo, Soumyaranjan Das, Arup Abhinna Acharya, Namita Panda, Sentiment Analysis using CNN for Emotion Extraction to Synthesize Natural Speech, Procedia Computer Science, Volume 258, 2025, Pages 2737-2747, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.04.534.

[8] Rui Wang, Duyun Xu, Lucia Cascone, Yaoyang Wang, Hui Chen, Jianbo Zheng, Xianxun Zhu, RAFT: Robust Adversarial Fusion Transformer for multimodal sentiment analysis, Array, Volume 27, 2025, 100445, ISSN 2590-0056, https://doi.org/10.1016/j.array.2025.100445.

[9] Qianru Gao, Jiachen Huang, Design and implementation of classical literature sentiment analysis system based on ensemble learning and graph neural network, International Journal of Cognitive Computing in Engineering, Volume 6, 2025, Pages 603-616, ISSN 2666-3074, https://doi.org/10.1016/j.ijcce.2025.05.004.

[10] Lamba, M., Madhusudhan, M. (2022). Sentiment Analysis. In: Text Mining for Information Professionals. Springer, Cham. https://doi.org/10.1007/978-3-030-85085-2_7

[11] K. L. Tan, C. P. Lee, K. S. M. Anbananthen and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," in *IEEE Access*, vol. 10, pp. 21517-21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[12] Jaber Areej , Bahati Israa , Martínez Paloma, Leveraging pre-trained embeddings in an ensemble machine learning approach for Arabic sentiment analysis, Frontiers in Artificial Intelligence, Volume 8 – 2025, doi: 10.3389/frai.2025.1653728