MemeCLIP: Leveraging Contrastive Language-Image Pre-training (CLIP) for Memotion Analysis

Vaishali Ganganwar, Gaurav Singh Chauhan, Jangveer Singh, Shashvat Khajuria, and Vivek Battan

Original scientific article

***Abstract*—Nowadays, memes, which commonly spread humor, ideas, or even harmful materials such as hate and propaganda, are a significant part of the Internet culture. The meme consists of an image and supporting text. Memotion Analysis, or meme Emotion Analysis, is automatic processing of memes using artificial intelligence. Unimodal solutions are now being taken over by multimodal solutions such as feature concatenation, weighted fusion, and Gated Multimodal Unit (GMU) for better Memotion Analysis. In this work, we proposed two deep learning based multimodal models for meme emotion classification. In the first model, we used ResNet and DeBERTa separately for single image-text fusion. In the second 'MemeCLIP' model an integrated CLIP-based representation with GMU employing a gated mechanism for adaptive visual and text feature fusion is used. In contrast to simple concatenation techniques, GMU demonstrates superior capability in extracting fine-grained emotional cues embedded in Memes. For the Memotion Analysis task 8 of SemEval-2020 competition, the CLIP-based model 'MemeCLIP' achieved a F1-score of 0.65, closely followed by the ResNet+DeBERTa model with a score of 0.64, compared to the SemEval baseline of 0.5118. These findings demonstrate the strength of selectively regulating modality contributions.**

***Index Terms*—Meme Classification, Memotion Analysis, CLIP, DeBerta, ResNet.**

I. Introduction

TODAY, social media is one of the most common ways of communicating and sharing thoughts through memes, i.e., a mode of media having text and images that convey the message. These help them express their emotions and opinions. Memotion Analysis [1] aims to make use of artificial intelligence and natural language processing to define and detect sentimental description of the meme. Memotion analysis takes into account both textual and visual cues for categorization, in contrast to standard sentiment analysis, which exclusively uses text as a medium. The memes are categorized based on their emotional impact and messages using machine learning and Convolutional Neural Network (CNN) based models. The memes are efficiently grouped into specific categories that aid in determining the meme's overall mood. As the memes are highly influential and engaging, leading to their rapid spread throughout the Internet, it is advised to make the network healthy for the overall audience, ranging from children to older adults. Some memes can carry harmful information which should be hidden from a specific group of users, making it a healthier platform for everyone. The automation of the meme emotion analysis is essential because of the immense volume of content generated every second. Automation can improve the accuracy of content moderation effectively classifying memes into different categories.

Participating in the call to examine the affective content contained in memes, the SemEval 2020 shared task 8 introduced the Memotion Analysis Task [1]. In this task, around 10,000 annotated memes are published with human-labeled categories including sentiment, types of emotion (sarcastic, funny, offensive, motivational), and their respective intensities. This task is not only related to labeling memes according to the polarity of sentiment (Task A), but also along some emotional and semantic dimensions—humor, sarcasm, offensiveness, and motivation—as part of Sub-Task B. Each of the four emotional tags is a binary classification task, labeling the presence or absence of that trait in a meme. The multimodal character of memes, with both text and image, makes this task very difficult and valuable as an area of research.

Memotion analysis typically involves extracting summaries of informative features from text and image, and then using an appropriate fusion mechanism to learn the emotional content. Existing approaches lack a well-balanced understanding of both modalities or fail to generalize effectively using large datasets. In this work, we proposed a pair of new bimodal architectures for Task B of Memotion Analysis to enhance precision and robustness. Our key contributions are a comparative study of two multimodal fusion approaches, both utilizing strong deep learning backbones to extract features, and the use of a Gated Multimodal Unit to learn how to wisely fuse visual and textual information.

The first method employed Residual Network (ResNet50) [2] to extract image features and Decoding-enhanced BERT with disentangled attention (DeBERTa) [3] to extract rich text embedding features. The second method employed OpenAI's Contrastive Language-Image Pre-training (CLIP) model [4], which jointly processes images and text with a single transformer-based model. Both models provide their modality-specific features to a GMU, which dynamically calculates im-

Manuscript received October 15, 2025; revised December 22, 2025. Date of publication June 1, 2026. Date of current version June 1, 2026. The associate editor prof. Damir Krstinić has been coordinating the review of this manuscript and approved it for publication.

V. Ganganwar is with the Department of Computer Engineering, Army Institute of Technology, Pune, India (e-mail: vganganwar@aitpune.edu.in).

G. S. Chauhan, J. Singh, S. Khajuria, and V. Battan are with the Army Institute of Technology, Pune, India.

Digital Object Identifier (DOI): 10.24138/jcomss-2025-0150

portance weights for each modality and produces a combined representation that retains the most important information from both inputs. The GMU outputs are then employed for multi-label binary classification over the four emotional classes. Our findings showed that both models have strong performance, with the ResNet-DeBERTa + GMU method having especially strong performance when trained on larger datasets due to DeBERTa's enhanced textual comprehension.

Aside from performance, the comparison of the CLIP-based model with the ResNet-DeBERTa model provided more information regarding the advantages and trade-offs of different feature extraction methods. Even though CLIP's pretraining on huge amounts of image-text data enables it to produce well-matched multimodal embeddings, the ResNet-DeBERTa model highlights its strength in independently extracting rich context text and high-level visual features. Both models rely solely on pre-trained representations to leverage the benefits of feature extractors, along with dynamic fusion mechanisms like GMU, to greatly improve emotion recognition in multimodal meme data. In addition, the Gated Multimodal Unit is critical in facilitating dynamic multimodal fusion. Instead of simply concatenating multimodal features or processing all modalities symmetrically, the GMU can learn to pay attention to the contributing modality with the most decision-making power for a specific meme. This attention-inspired weight mechanism not only enhanced classification performance but also offered explainability, gaining insights into what drives a meme's emotional cue more by either its image or text content.

The key contributions of the proposed work are as follows.

- We proposed two deep learning-based bimodal models for Task B of the SemEval 2020 Memotion Analysis Challenge, designed to identify the emotional aspects: humor, sarcasm, offensiveness, and motivation in memes.
- ResNet50 is used in the first model to obtain high-level image features and DeBERTa to produce dense, context-enriched text embeddings.
- The second model 'MemeCLIP' utilized OpenAI's CLIP, a transformer model that processes the image and text in parallel. CLIP generates multimodal embeddings that are inherently aligned and are then summed with the GMU for classification.
- Both models are trained and tested on the official SemEval 2020 Memotion Analysis dataset Memotion 1.0, which contains 6992 training memes and 1878 test memes.
- The performance of proposed models is compared with the existing approaches for Memotion Analysis. Comparative analysis showed that our suggested models greatly surpassed current models in F1-score, with better classification accuracy and stability in all emotional classes.

The remainder of this paper is organized as follows. Section II discusses different methodologies and approaches on Memotion Analysis. Section III describes the dataset used in our work. Section IV describes the methodology of the proposed work. Section V contains the results and discussion, followed by the final section that outlines the conclusion.

II. Related Work

Multiple studies on this subject have been carried out in the past with researchers utilizing various methodologies and approaches.

A. Uni-Model Approaches for Memotion Analysis

The study of memes—those propagated, culture-rich images with humor or sarcasm-rich text superimposed on them—has quickly become an intricate issue in sentiment analysis. Although much work has been devoted to multimodal fusion approaches, a closer examination of the history of unimodal approaches, rooted in a single modality like text or image, indicates a surprisingly robust foundation. Indeed, several seminal studies [5], [6], [7], [8] demonstrated that unimodal systems not only compete but, in certain situations, outperformed intricate multimodal hierarchies, particularly when image and text fusion is not semantically consistent.

Bonheme and Grzes [9] proposed a system named 'SESAM' for Memotion analysis. They sought to examine to what extent typical machine learning models could understand memes by aligning or fusing textual and visual information. They created a pipeline that included typical machine learning models like Logistic Regression, K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Random Forest, and Multi-Layer Perceptron (MLP). For representation of text embeddings: Universal Sentence Encoder embedding and CNN is used for representing the image in fixed length vector. These vectors are then concatenated to form a single feature vector for meme emotion classification.

They tried a variety of advanced fusion techniques—alignment-based techniques like Canonical correlation analysis (CCA) and Deep Canonical correlation analysis (DCCA) but no multimodal combinations achieved notable performance gains. In particular, image and text components of memes were not statistically related, making alignment-based techniques not useful. Even fusion-based techniques such as early-level concatenation and late fusion (ensemble voting) failed to perform well. For certain instances, even fusion models were worse than simple unimodal baselines.

Unimodal models, particularly text-only embedding-based outperformed more sophisticated counterparts consistently. The highest performance with 35% F1- score for sub-task A (sentiment classification) was achieved by applying KNN with image embeddings in their evaluation. Nevertheless, further inspection indicated that GNB with text embeddings would have been a better submission, reinforcing the argument that text-only models were more robust for meme sentiment classification. Their results provided a valuable insight: memes do not always have significant image-text alignment, and in such instances, unimodal approaches offer simplicity as well as better performance.

Keswani et al. [10] presented a set of unimodal and bimodal approaches for sentiment classification of Internet memes in SemEval-2020 Task 8. Their methods included traditional machine learning models, deep learning techniques, and transformer-based architectures. For unimodal (text-only)

approaches, they explored Naïve Bayes, BERT, and a Feed Forward Neural Network (FFNN) using Word2vec embeddings. For bimodal approaches, two models were tested: a fusion of FFNN (for text) and CNN (for image) with an SVM classifier for final prediction, and the Multimodal Bitransformer (MMBT), which integrates ResNet-152 and BERT to align image embeddings into the text space. Their experiments revealed that despite the use of sophisticated bimodal techniques, the simple text-only FFNN model with Word2vec achieved the best performance, with a macro-F1 score of 0.3546, surpassing all other methods including MMBT (0.30) and BERT (0.33). They attributed this to the dominant role of text in meme sentiment classification and the limitations of pre-trained models when applied to noisy, sarcastic, and short social media texts. Their FFNN architecture included six hidden layers and was trained with the Adam optimizer and ReLU activation. The study concluded that simpler models, when combined with effective embeddings and preprocessing, can outperform complex architectures, especially in low-resource and noisy domains.

From this finding, Singh et al. [11] created a hybrid deep learning model, text-only but making use of temporal as well as spatial features. Their model employed Bidirectional Long Short-Term Memory (BiLSTM) with a deep 2D-CNN architecture based on ResNet. The approach treated the input text as a time series, allowing the BiLSTM to learn both past and future dependencies—a strength when working with short, context-dependent meme captions. The temporal features were extracted, then translated into a grayscale image representations, based on word embeddings. These grayscale matrices were then fed into the ResNet-style 2D-CNN, making use of residual connections to avoid performance degradation characteristic of deep networks [12].

This new BiLSTM-ResNet hybrid allowed the model to utilize both sequence-aware representations (via BiLSTM) and deep pattern recognition (via CNN). The residual connections with depth facilitated deeper networks to be trained, leading to more accurate classification without affecting model stability. The model was extensively tested on the Sentiment140 corpus, and performance showed a strong positive relationship between network depth and classification accuracy. As the number of residual blocks grew and larger chunks of the dataset were utilized, the model's performance was dramatically boosted. This research not only confirmed the effectiveness of deep unimodal text models but also stressed the importance of combining temporal and spatial knowledge to construct stable sentiment classifiers.

Although these two works handled only text, Liao et al. [13] proposed image-text interaction graph neural network (ITIGNN), a multimodal model that—albeit designed for fusion—provided some information on the relative strength of unimodal routes. Their framework employed CNN for the extraction of visual features from images, such as color, texture, and shape, and Graph Neural Networks (GNN) to encode syntactic and semantic dependencies in the text. To enable the two streams to communicate, they utilized an image-text interaction layer—inferred to be based on attention—so the model could attend to the most pertinent part of each modality.

Following this, an image-text aggregation layer combined or summed features together to create an aggregated feature vector. This was then manipulated by a feed-forward attention mechanism, which attended most strongly to the most important information in each modality before the final summation layer produced the input to classification. While the system was certainly multimodal in nature, it was the strength of each single stream, that is, the textual stream based on GNNs, that accounted for its overall performance. The ITIGNN model was tested on two datasets: MVSA and Twitter26k, both Twitter-based but differing in scale and curation. The MVSA dataset had 5,129 image-text pairs initially, which were reduced to 4,511 high-quality pairs after removing conflicting labels. The Twitter26k dataset, curated from Twitter100k, had 26,951 pairs, carefully filtered with a rule-based sentiment analysis tool-VADER for text sentiment and a visual sentiment model from t4sa for images. ITIGNN, despite problems like label imbalance—where neutral sentiment outweighed—performed better than state-of-the-art methods by more than 5% in macro-F1 score on MVSA, and also fared better on Twitter26k. These performances not only asserted the model's stable management of sentiment noise but also highlighted the importance of modality-specific optimization even on a multimodal pipeline.

B. Multimodel Approaches for Memotion Analysis

With the meteoric growth of memes as a pervasive mode of online communication, their analysis has become a worthwhile pursuit in computational linguistics, computer vision, and social computing. As opposed to unimodal sentiment analysis—which considers text or image separately—multimodal meme emotion analysis combines visual and textual content to label sophisticated sentiments like humor, sarcasm, motivation, and offense. In time, researchers have come up with more and more advanced architectures to capture the visuo-lingual interplay characterizing memes, culminating in profound advances in accuracy and interpretation.

Among the early efforts in this space is MemoSYS [14], developed by Bejan. Entered in SemEval-2020 Task 8, the system experimented with four multimodal architectures, blending pre-trained VGG16 for image features with either TF-IDF or BERT for text. Fusion was done using locally connected layers, and classification was performed using either softmax or SVM classifiers. The BERT+VGG16 fusion resulted in the best performance, recording a macro-F1 of 0.3513 and accuracy of 0.4988 for sentiment classification. Preprocessing techniques involved image resizing, data augmentation, and text masking using the EAST text detector. In spite of encountering difficulties such as noisy data and label imbalance, MemoSYS established the strength of transfer learning and deep multimodal fusion in meme sentiment tasks.

The second significant contribution was made by Gupta et al. [15], as they developed a robust pipeline for SemEval-2020 Task 8. They compared BiLSTM and RoBERTa for text and AlexNet and ResNet for images. Fusion methods used Early Fusion (concatenation of features), Late Fusion (averaging predictions), and GMU. They also employed Multitask Learning (MTL) in order to share the representations of

sentiment, humor, and offense tasks. The RoBERTa+ResNet model using Early Fusion obtained the optimal performance: macro-F1 equal to 0.357 on sentiment (Task A), 0.510 on humor (Task B), and 0.306 on semantic scale (Task C). In spite of domain-transfer problems due to stylistic meme text and noisy images, this work demonstrated that plain fusion methods can be very competitive if well-tuned.

Building further, Singh et al. [16] presented a Multi-Modal Multi-Task Learning system that incorporated the idea of "Memebeddings"—shared feature vectors learned from BERT-based text and ResNet-18 image features. These embeddings were passed through a 1024-dimensional linear layer and into five task-specific classifiers. Several training paradigms were investigated: independent task training, average/weighted loss MTL, and a continual learning setup. The continual learning setting had the highest macro-F1 values: 0.2771 (Task A), 0.5077 (Task B), and 0.5069 (Task C). The strategy proved that fine-tuning vision encoders on meme-specific datasets, together with shared feature learning, greatly improves classification performance.

The paper "Little Flower" by Phan et al. [17] presented an ensemble approach employing BiLSTM with Bahdanau attention for text and VGG16 for image. Attention was used to help the model concentrate on key words in text and relevant areas in images. Two-headed multi-head attention was applied for visual features to enhance context sensitivity. The final model—cross-validated using K-fold cross-validation—obtained a weighted F1 score of 73.58%, beating ResNet- and BERT-based models. This illustrated the extent to which attention-based fusion models with thoughtful structure can uncover nuanced multimodal sentiment signals.

In the same vein, Sharma et al. [18] went a step further by adding emotion-aware features using Vision Transformers (ViT) pre-trained over AffectNet and text represented using BERT. Their approach presented GMU, which merged general and emotion-specific image features via Hadamard product-based gating. Next, Gated Cross Attention (GCA) brought text and image representations into alignment through the modulation of textual embeddings according to emotion-weighted image features. This produced a single, deeply fused representation that saw significant gains in meme emotion categorization. The architecture showcased the strength of emotion-aware visual feature extraction and cross-modal gating in boosting interpretability.

The Multi-Modal Multi-Task Transformer (MMIT) model by Bucur et al. (Blue team) [19] was a significant step forward by using frozen Sentence Transformers for text, CLIP and EfficientNetV4 for image features, and mapping them to a common 512-dimensional space. A 4-layer transformer encoder (without positional encoding) encoded cross-modal relations, and task-specific heads did classification and ordinal regression using CORAL loss and cross-entropy. In ablation studies, the addition of CLIP features greatly improved performance. text+CLIP models achieved an F1-score of 0.5567, better than image-only (0.5553) and text-only (0.5541). The model performed especially well on sentiment, sarcasm, and motivation, though it was a bit behind on humor and offense compared to text-only methods.

The BROWALLIA system [20] of Duan and Zhu continued to explore multimodal learning by mixing ResNet-50 for image and BERT/LSTM for text. It employed late fusion, boosted by Offline Gradient Blending (OGB)—a dynamic weighting mechanism that dynamically adjusts each modality's contributions according to validation performance. Focal loss was also employed to address class imbalance. This approach achieved a Macro-F1 score of 0.3649, beating all unimodal baselines. On the Memotion 2.0 leaderboard, the model was placed second for Subtasks A and C, and third for B. The findings confirmed that adaptive multimodal fusion, informed by performance-based weighting, achieves strong classification across meme subtasks.

On another front, Ahuja et al. [21] focused on transfer learning using RoBERTa for text and Xception for image features, choosing them according to size, accuracy, and parameter efficiency. The fused system applied weighted fusion to reconcile the effect of visual and textual modalities. Their model achieved high performance with 92.08% accuracy and stable precision, recall, and F1-scores at 0.92 levels. The study highlighted that parameter-efficient architecture and optimized fusion can attain top-level performance along with computational affordability.

Sharma et al. [22] introduced ALFRED, a multimodal neural model designed specifically for fine-grained emotion recognition in memes over six Ekman emotion classes. Their method pairs emotion-augmented features of an extended AffectNet dataset with a gated multimodal fusion (GMF) module and GCA mechanism for detecting subtle visual-textual associations. The authors also released MOOD, a high-quality, human-annotated meme dataset containing 10,004 memes covering a wide thematic and emotional range. The ALFRED architecture achieved an F1-score improvement of 4.94% over strong early-fusion baselines on MOOD and exhibited strong generalization on the Memotion, HarMeme, and Dank memes datasets.

Pandey and Vishwakarma [23] gave a comprehensive overview of multimodal sentiment analysis with deep learning for prominent modalities like text, image, audio, and video. Various fusion techniques, deep architectures like CNNs, RNNs, and transformers, and techniques for word embedding and feature extraction were discussed by them. Practical applications and most challenging tasks such as sarcasm detection, subjectivity tagging, and the requirements of strong multimodal datasets were addressed in the paper.

Lastly, Sharma, Kandasamy, and Vasantha [18] suggested a two-stream architecture that processed text separately through BiLSTM with GloVe embeddings and images through pre-trained Inception networks. The features were combined through dense layers and regularized by dropout and L2. A stacked BiLSTM-GRU block generated the final representation. The model performed macro-F1 of 0.325 on humor classification and 0.508 on humor quantification in SemEval-2021 Task 3, showcasing robust performance in identifying the humor content of memes.

The relation between meme stocks and social media is explored by Lee et al [24]. They have integrated NLP based techniques and econometric techniques to find the correlation

between trading volume and social media posts. Akshaya et al. [25] proposed model based on BERT and vision transformer. They tested their model on Memotion 7K corpus. For addressing correlation between text and images Zheng et al. [26] proposed global-local cross-modal interaction model. They used the MET-MEME bilingual dataset for experimentation. Jha et al. [27] proposed four different model based on variants of BERT and ResNet. There model based on DistillBERT and ResNet achieved highest accuracy compared with other proposed models.

The path of multimodal meme emotion analysis has progressed from initial feature concatenation methods to attention-guided fusion, transformer-based embeddings, and emotion-aware optimization methods. Each of these approaches has shed light on the semantic alignment between text and image, the value of domain-specific visual pretraining, and the requirement for strong training schemes such as MTL and continual learning.

III. Dataset Used

For the training and testing of our meme emotion classification model, we use the Memotion 1.0 dataset [1], which is a broad and diverse collection of 8,000 memes drawn from several social media sites, including Facebook, Instagram, and Twitter. The dataset consists of multimodal data with both textual and visual features and is annotated with sentiment as well as humor-related properties. It is split into two main subsets: a training set of 6,400 memes and a test set of 1,600 memes. The dataset can be used with three main tasks: sentiment classification, in which the memes are classified as positive, neutral, or negative; multi-label classification, which finds the occurrence of humor, sarcasm, offense, and motivation; and semantic intensity quantification, which assigns a numerical measure to the intensity of these attributes.

One of the most important features of the Memotion 1.0 dataset is that it is multi-label, i.e., a meme may belong to more than one category at once. For instance, a meme may be both funny and offending at once. This does create more complexity in the classification process. The dataset also indicates a large degree of imbalance between label distribution with 4,272 memes rated as funny, 4,358 as sarcasm, 3,424 as offensive, and merely 1,974 as motivational. This kind of imbalance requires delicate preprocessing and tuning of the model to prevent bias during classification. Table I shows the statistics of the Memotion 1.0 dataset.

TABLE I
Statistics of the Memotion 1.0 dataset

	Humour	Sarcastic	Offensive	Motivational	Total
Train	4272 (61%)	4358 (62.3%)	3424 (48.9%)	1974 (28.2%)	6992
Test	1069 (56.9%)	1090 (58%)	855 (45.5%)	493 (26.2%)	1878

The annotation types differ according to task. Sentiment classification is scored on a -1 to 1 scale, binary labels for humor categories are presence/absence, and intensity quantification scores on a 0 to 3 scale. Baseline metrics of evaluation capture the challenge of the dataset in F1-scores of 0.2176 for sentiment classification, 0.5118 for binary classification, and 0.2483 for intensity quantification, providing benchmark levels for model performance.

A. Other Existing Datasets

Although not used in this work, several other publicly available datasets are noteworthy in meme sentiment analysis research are as follows:

- Memotion 2.0: An extended version introduced in SemEval-2020, this dataset builds upon Memotion 1.0 with an emphasis on temporal shifts, more refined emotion scales, and better OCR-augmented text extraction. Its aim is to evaluate how meme sentiments evolve over time, which makes it ideal for temporal or longitudinal analysis.
- MOOD: Designed for deep emotional classification in memes, focusing primarily on fine-grained mood detection rather than coarse sentiment. It includes richer annotations but with a smaller meme set.
- AffectNet & Reddit Datasets: Often used in facial emotion or visual content classification, these datasets complement meme datasets by providing contextual visual cues and user-driven sentiment reactions, respectively.

These datasets highlight the broader landscape of meme and multimodal emotion analysis and serve as valuable benchmarks. However, for the scope and experimental setup of this work, Memotion8K was selected due to its balanced multimodal structure, task complexity, and accessibility through CodaLab.

IV. Methodology

We have proposed two different models for Memotion classification, one uses DeBERTa for the extraction of text features and ResNet50 for the extraction of image features and then merges these features using the Gated Multimodal Unit layer in a similar fashion. The second approach used the CLIP transformer for the image and text feature extraction and then merged the extracted features using Gated Multimodal Unit Layer.

The proposed models involved three steps: 1) Extracting features from images and text, 2) Utilizing a Gated Multimodal unit to combine text and image features, and 3) Employing a Fully Connected Neural Network with a Classification Layer for Memotion Classification. We used two different approaches for image and text feature extraction, with the first approach based on DeBERTa and ResNet, and the second approach based on the CLIP transformer, as illustrated in Fig 2 and 3.

A. Data Preprocessing

A training dataset containing 6992 memes and a test dataset containing 1878 memes are both publicly available for research. The text and image that make up each meme are crucial for multimodal emotion classification. To ensure

TABLE II
COUNT OF CO-OCCURRENCE ACROSS THE 4 EMOTIONS

Labels	Class	Humour		Sarcasm		Offensive		Motivational	
		0	1	0	1	0	1	0	1
Humour (not_humorous:0,humorous:1)	0	1695	0	592	1101	748	946	1097	597
	1	0	5296	873	4423	1768	3528	2954	2341
Sarcasm (not_sarcastic:0,sarcastic:1)	0	592	873	1466	0	1092	374	1090	376
	1	1101	4423	0	5525	1424	4100	2962	2563
Offensive (not_offensive:0,offensive:1)	0	748	1768	1092	1424	2517	0	1010	506
	1	946	3528	374	4100	0	4474	2042	2431
Motivational (not_motivational:0,motivational:1)	0	1097	2954	1090	2962	2010	2042	4052	0
	1	597	2341	376	2563	506	2431	0	2939

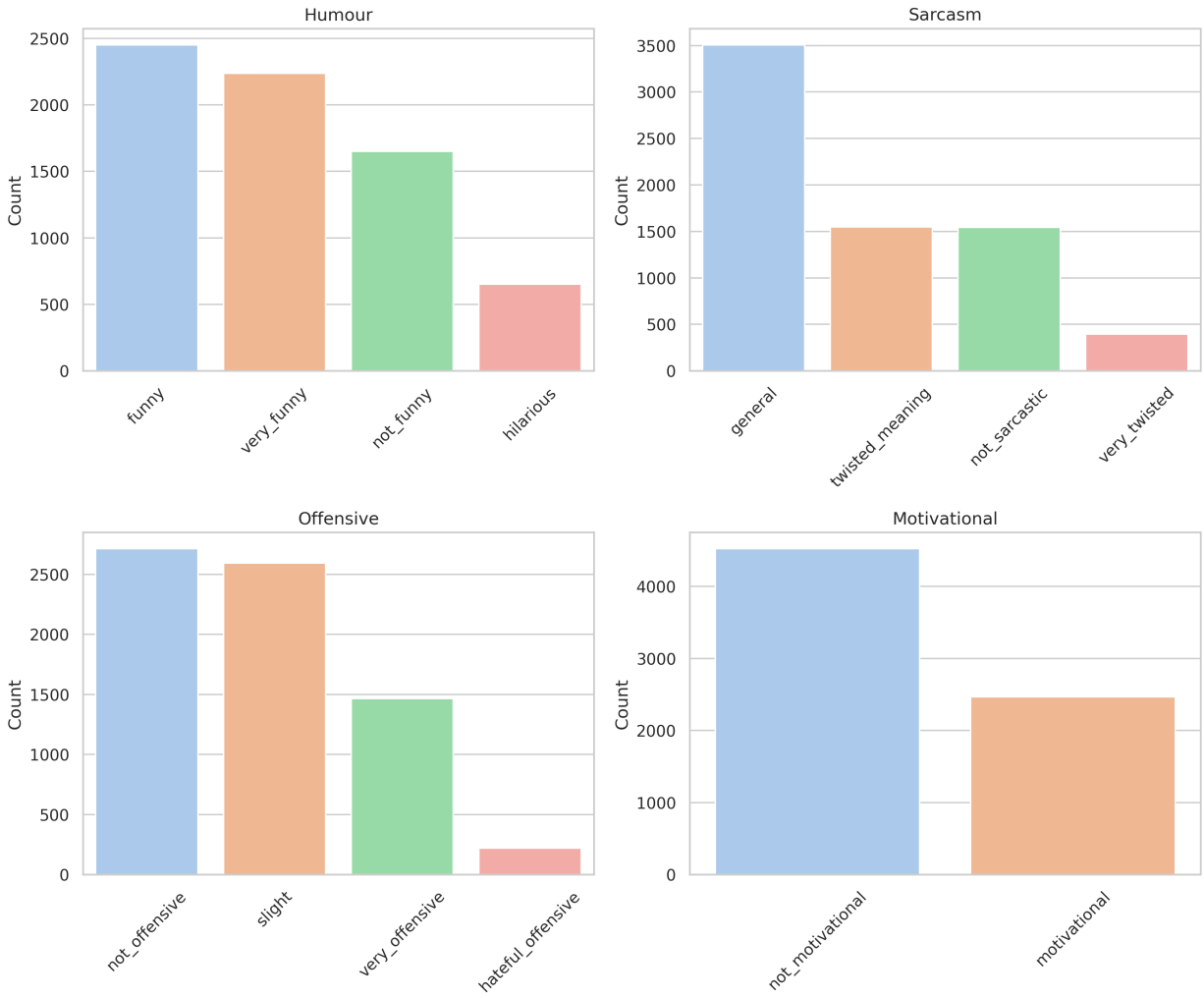

Fig. 1. Class distribution in Memotion 1.0 Dataset

objective evaluation the test set is kept separate for training and validation while the training set is split in an 80:20 ratio.

In our work, we proposed models for Sub Task B of the Sem-Eval 2020 compilation in which label encoding converts categorical annotations into binary labels. As the dataset is multi-label, the count of co-occurrence across the four emotion classes is shown in Table II. Figure 1 shows the intensity quantification score distribution for each class.In data preprocessing , Humor labels [not_funny:0, funny:1, very_funny:1 and hilarious:1] are combined to form a single binary category. memes that are humorous are given the label 1 and those that are not, given the label 0. Sarcasm [not_sarcastic:0, general:1 and twisted_meaning:1] offensiveness [not_offensive:0, slight:1, very_offensive:1 and hateful_offensive:1] and motivation [not_motivational:0 and motivational:1] are all approached similarly. These labels are stored as four-dimensional

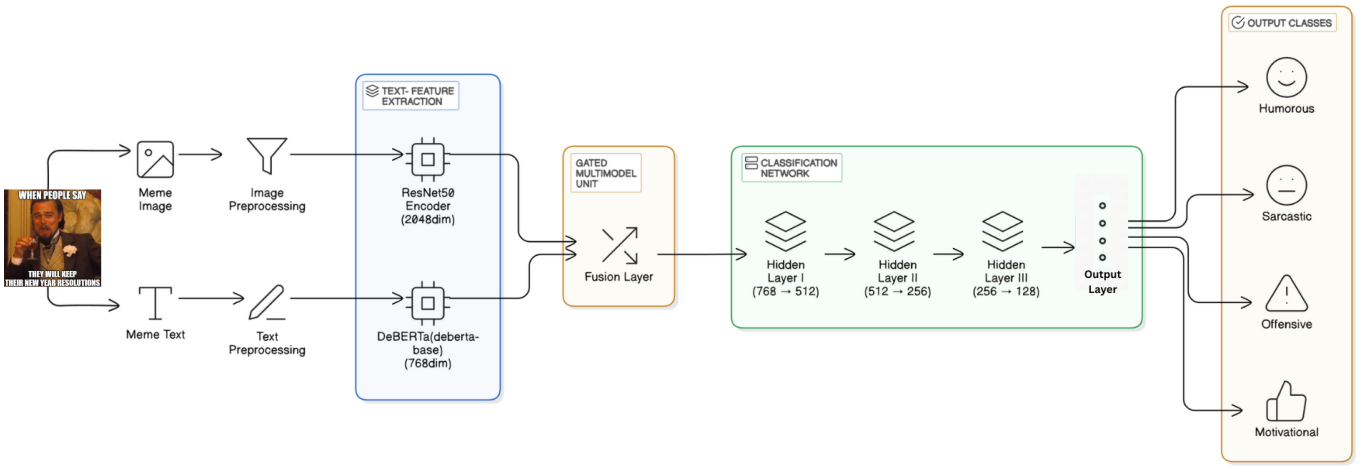

Fig. 2. DeBERTa & ResNet-based approach for Memotion Analysis

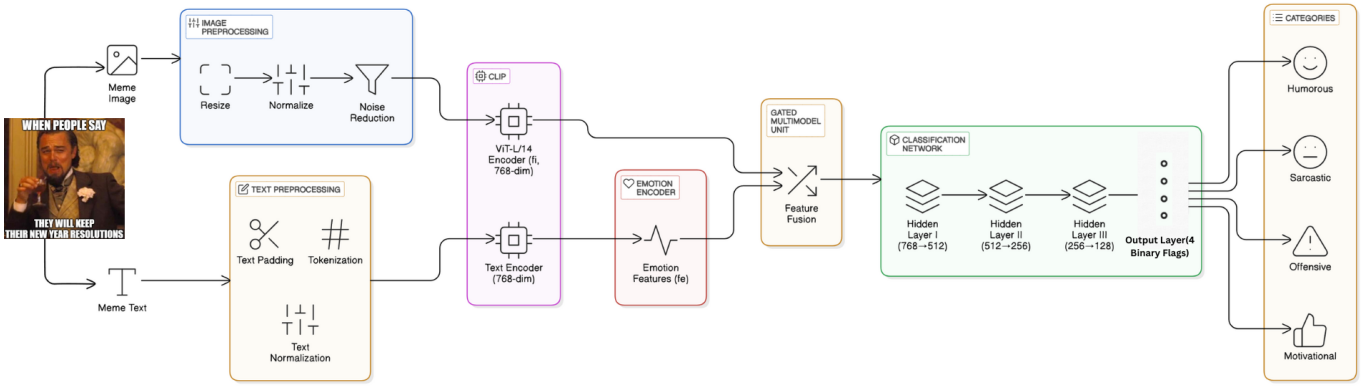

Fig. 3. MemeCLIP- CLIP-based approach for Memotion Analysis

tensors to aid in multi-label classification.

During preprocessing, images are loaded and resized, and any corrupted or unreadable ones are ignored. The text is tokenized and cleaned. The final dataset comprises image-text pairings along with the labels.

B. Image Feature Extraction

1) ResNet50 [2]: In this approach (Fig 2), ResNet50 is used for image feature extraction of the meme. It's a deep residual network pretrained on the ImageNet dataset. ResNet uses a "residual connection" in its layer to handle gradient descent and accelerate deep network convergence.The ResNet architecture utilizes a total of 50 layers which includes 48 convolution layers, one max pooling layer, and one global average pooling layer. The ResNet50 model encodes images by applying multiple convolutional layers to capture complex features of the image. The ResNet50 transformation process includes resizing images to 224×224, normalizing them, and converting them into tensors. After the image is processed through the convolutional layers, the output is directed to a fully connected layer containing 2048 output units. This layer is then transformed into a 768-dimensional feature vector, which acts as the image embedding. This 768-dimensional vector serves as the representation of the image.

2) Contrastive Language-Image Pre-training (CLIP): To extract image features OpenAI's CLIP is utilized [4]. Images are resized to 224 x 224, normalized and transformed into tensors as part of CLIP's transformation pipeline. After processing the images ((Fig 3), CLIPs Vision Transformer (ViT-L/14) encoder extracts a 768-dimensional feature vector for every image. The transformer architecture of ViT-L/14 produces rich image embeddings that can further be compared with the text embeddings to understand the underlying patterns and semantics of the memes.

C. Text Feature Extraction

1) DeBERTa: DeBERTa, a variant of the BERT model introduced by He et al. [3], is used for the text sentiment analysis of the meme. Unlike traditional models, DeBERTa incorporates disentangled attention mechanisms and enhanced

relative position encoding, which improves its ability to capture contextual dependencies in natural language. It is pretrained on a large corpus of textual data and fine-tuned for downstream tasks to achieve superior performance in natural language understanding. During text encoding, the meme text is first tokenized using the DeBERTa-large tokenizer, which segments the input into sub-word tokens suitable for the model's vocabulary. The resulting tokenized input is passed through the DeBERTa-large encoder, which outputs contextualized embeddings for each token. To obtain a fixed-size sentence representation, a pooling layer (Fig. 2) is applied to the encoder output. DeBERTa-large produces a 1024-dimensional feature vector, which is subsequently passed through a linear projection layer, called the text projection layer, to reduce its dimensionality to 768, facilitating multimodal fusion with other modalities, such as image features.

2) CLIP: To extract text features (Fig 3), OpenAI's CLIP text encoder [4], using a transformer model similar to GPT is used. To ensure consistent representation across text lengths, the text is first tokenized using the built-in tokenizer of CLIPs. A transformer encoder receives the tokenized text sequence as input, and each token undergoes self-attention calculations to enable contextualized feature extraction. The text embedding is a 768-dimensional feature vector that is the end product. Following the capture of semantic information, these embeddings are then combined with image representations for classification.

D. Gated Multimodal Unit

The GMU [28] is a neural network that is intended to integrate multiple input modalities through the learning of a gating mechanism. The gate imputes each modality (text and image in our implementation) with a soft importance weight, enabling the model to scale the contribution of each input at run-time dynamically. This proves to be valuable in Meme classification, where text vs. image importance can switch dramatically between examples.

The GMU combines two input vectors, one from text (T) and one from image (I). A gating mechanism is used to determine the importance of each modality. The gate vector z is computed as:

$$z = \sigma\left(W_z T + U_z I + b_z\right)$$

Where W_z and U_z are learnable weight matrices, b_z is a bias vector, and σ is the sigmoid function. The final fused representation h is obtained as:

$$h = z \odot \tanh\left(W_h T\right) + (1 - z) \odot \tanh\left(U_h I\right)$$

Here, W_h and U_h are additional learnable weight matrices to transform input vectors before combining them, and $\odot$ denotes element-wise multiplication. This formulation enables the model to learn an optimal blend of both modalities dynamically.

Two different transformer-based combinations are experimented within this design:

- ResNet50-DeBERTa: In this combination, The ResNet50 model provides an image feature vector of size 2048 which is proportionately reduced to 768 dimension, and the DeBERTa model provides a 768-dimensional text feature vector. These feature vectors are then fed into the GMU, which calculates weighted combinations of the features from both the image and the text.
- CLIP: In this combination, both the image transformer and text transformer are based on CLIP. GMU produces 768 size fused feature vectors.

E. Fully Connected Neural Network with Classification Layer

In both methodologies, the last component in the pipeline is a fully connected multi-layer network preceded by an output layer of 4 neurons corresponding to one out of the group of four emotions.

The fusion feature vector (dimension = 768) is forwarded to this network and it is given a sequence of linear transformations, ReLU activations, batch normalization and dropout layers. This enhances generalization, introduces non-linearity and avoids overfitting.

The model has three hidden layers of diminishing size (768 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128), and batch normalization and dropout (p=0.3) used after the first two layers. The last output layer has four neurons, each accounting for one of the four emotion categories: humour, sarcasm, offensive, and motivational. Because this is a multi-label classification problem, the model itself does not emit raw logits by not using sigmoid activation. However, the BCEWithLogitsLoss function is used, which internally enforces the use of sigmoid activation and calculates Binary Cross-Entropy loss in a numerically robust manner.

This architecture enables the model to learn sophisticated non-linear relationships within the combined multi-modal feature space while separately predicting the occurrence of several emotions and, therefore, suits meme emotion classification.

V. Results and Discussion

This section describes the hyperparameters used, evaluation metrics, and performance of both proposed models.

A. Hyperparameters

For training, a batch size of 32 was used, with the number of epochs set to 25. The learning rate was configured at 1×10^{-5} and a dropout rate of 0.3 was applied. The model employed the ReLU activation function and the Adam optimizer, while Binary Cross-Entropy with Logits was used as the loss function. For the fully connected neural network, the input dimension was set to 768, and three hidden layers were utilized. Batch normalization was applied after the first and second hidden layers to improve training stability and performance.

B. Evaluation Metrics

To measure the performance of our multimodal Meme classification models, we employed the macro-averaged F1-score. This score calculates the F1-score separately for each class of emotion (humorous, sarcastic, offensive, motivational)

and then averages the resultant scores. It is especially apt for our task, since it specifies that every class is considered equally, irrespective of class imbalance, a fact that has been acknowledged by existing Meme datasets. As a baseline, we consider the average macro F1 score reported in the paper [29] Baseline Score of 0.5118. This serves as a reference point to assess the effectiveness of our approaches.

C. Performance of Proposed Models

The graphs of training and validation loss over epochs for the ResNet-DeBERTa-based model and the MemeCLIP model are shown in Fig 4 and 5 respectively.

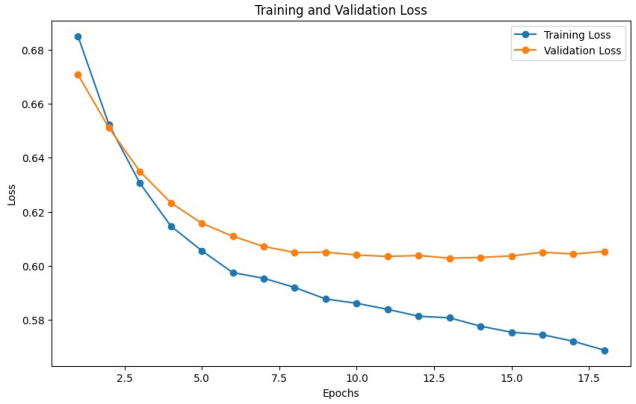

Fig. 4. Training & Validation Loss wrt epochs for ResNet-DeBERTa-based Model

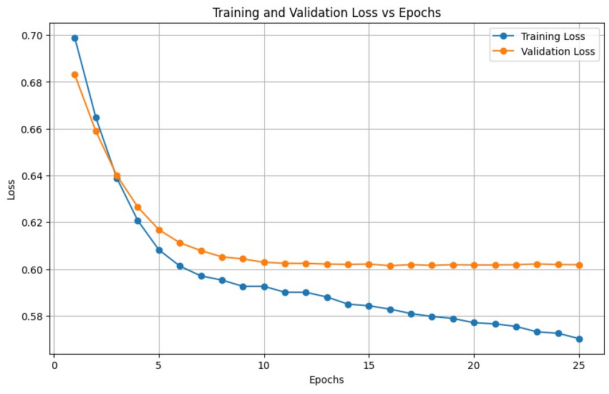

Fig. 5. Training & Validation Loss wrt epochs for MemeCLIP Model

In the DeBERTa-ResNet50 approach, text features are extracted using DeBERTa, while image features are obtained using ResNet50. These features are then fused using the Gated Multimodal Unit (GMU) and passed through a fully connected classification layer. The detailed performance of this model is shown in Table III.

The DeBERTa-ResNet50 model performed comparatively well for Humour and Sarcasm, both with F1-scores of 0.77 and 0.79, respectively. Performance takes a sharp fall for Offensive (F1-score: 0.62) and is the worst for Motivational (F1-score: 0.39), reflecting the inability to recognize motivational memes. It is notable that recall for the Motivational class is 0.42, which reflects the model's inability to identify high numbers of true motivational cases correctly.

TABLE III
PERFORMANCE OF DEBERTA-RESNET50 BASED MODEL

Metrics	Humour	Sarcasm	Offensive	Motivational
Precision	0.77	0.78	0.64	0.37
Recall	0.77	0.79	0.60	0.42
F1-score	0.77	0.79	0.62	0.39

In the MemeCLIP model, both image and text features are extracted using CLIP (ViT-L/14). These features are subsequently fused using a GMU-based fusion mechanism and then passed to a fully connected classification layer. Table IV provides the detailed performance results of the MemeCLIP model. The class-wise performance comparison of both proposed models is shown in Table V.

TABLE IV
PERFORMANCE OF MEMECLIP MODEL

Metrics	Humour	Sarcasm	Offensive	Motivational
Precision	0.78	0.78	0.63	0.39
Recall	0.80	0.81	0.65	0.36
F1-score	0.79	0.79	0.64	0.37

The MemeCLIP model demonstrated better performance across the majority of categories, especially Humour (F1-score: 0.79) and Offensive (F1-score: 0.64), than the DeBERTa-ResNet50 model. As with the earlier model, Sarcasm detection continues to be robust (F1-score: 0.79). Motivational meme detection, however, continues to trail (F1-score: 0.37), again probably because there are not many such instances in the training set.

Overall, the MemeCLIP model performed slightly better than the DeBERTa-ResNet50 model, with an average F1-score of 0.65 versus 0.64. The gains are mainly seen in the Humour and Offensive classes, where the synergistic combination of visual and textual information through CLIP and GMU is likely to have made a bigger impact.

DeBERTa is very good at processing subtle text signals, which is appropriate for detecting sarcasm and offensiveness. Thus the DeBERTa-ResNet model performance is approximately similar for the challenging Sarcasm and Offensive classes. In 'MemeCLIP' model, the simultaneous training of CLIP on image and text seems to help with increased synergy in some emotion categories. Motivational memes remain challenging due to fewer examples in the dataset.

On closer examination, we see that motivational class always received the lowest F1-score on both methods. This is because there are very few motivational samples within the dataset, which restricts the ability of the model to generalize effectively.

A more balanced and larger dataset could potentially allow our methodologies, especially those leveraging pretrained transformers, to better learn class-specific patterns. Given enough data, we believe the combination of strong feature extractors and fusion mechanisms could yield substantially better performance across all emotion classes.

TABLE V
PERFORMANCE COMPARISON OF PROPOSED MODELS

Model	Humour	Sarcasm	Offensive	Motivational	Overall
DeBERTa & ResNet50	0.77	0.79	0.62	0.39	0.64
MemeCLIP (ViT-L/14)	0.79	0.79	0.64	0.37	0.65

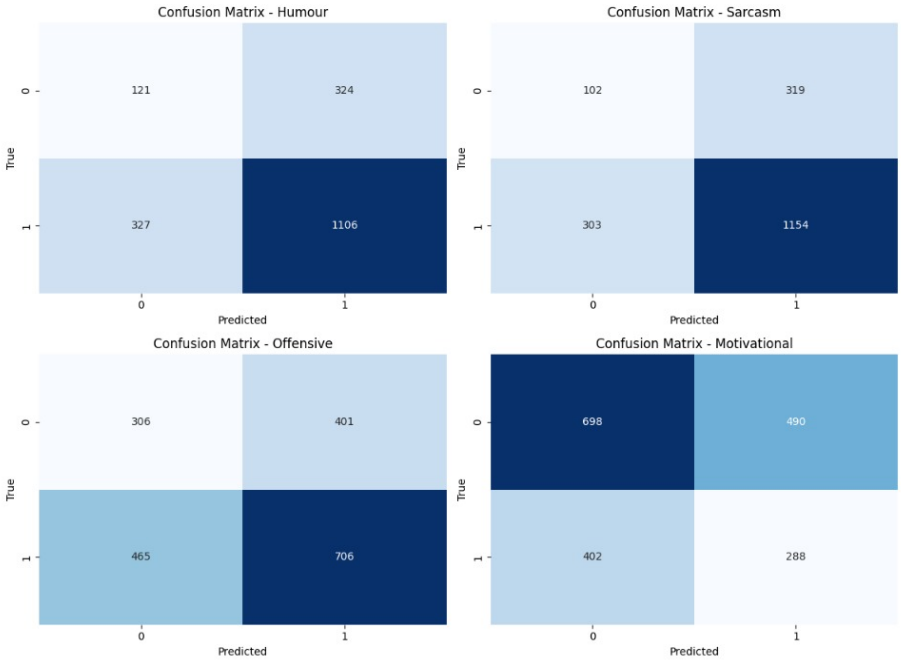

Fig. 6. Confusion Matrix for DeBERTa-ResNet50 Model

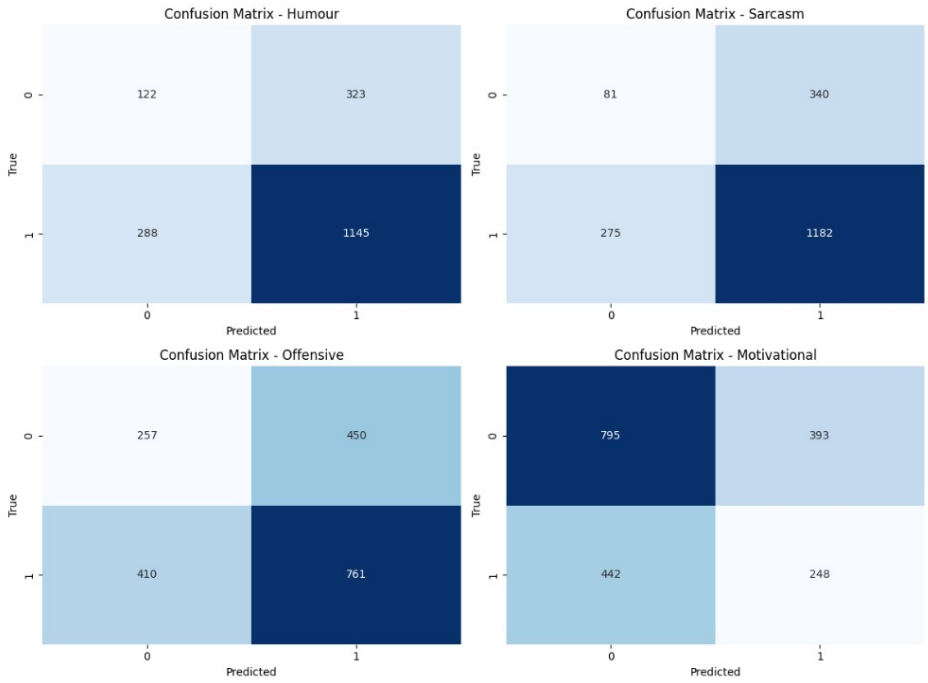

Fig. 7. Confusion Matrix for MemeCLIP Model

Fig 6 and 7 shows the confusion matrices for the four classes of the DeBERTa-ResNet50 model and the MemeCLIP model, respectively. As shown in the confusion matrices, both the models perform well in identifying true positives for the Humor and sarcasm class.However, there is some difficulty in distinguishing between non-sarcastic and non-

TABLE VI
PERFORMANCE COMPARISON OF PREVIOUS STUDIES BASED ON MODALITY AND TECHNIQUE

Study	Modality	Technique	F1-score
Baseline	-	-	0.5118
MemoSYS [14]	Text + Image	BERT + VGG	0.4519
SESAM [9]	Image	KNN	0.49
DSC IIT ISM [15]	Text + Image	BiLSTM+AlexNet	0.495
Gundapusunil [30]	Text + Image	LSTM + InceptionV3	0.5014
LT3 [16]	Text + Image	NN	0.5077
DSC IIT ISM [15]	Text + Image	RoBERTa + ResNet	0.51
DeBERTa & ResNet-based Model	Text + Image	DeBERTa + ResNet50	0.63
MemeCLIPModel	Text + Image	CLIP	0.65

humorous samples, as shown by the significant number of false positives. In the Offensive category, the model performs moderately with similar misclassifications in both classes. The Motivational category has a higher rate of misclassification, indicating challenges in distinguishing between motivational and non-motivational content. In order to put our results into context, we compared our models with the findings presented in the previous literature using the same dataset. Table VI shows the performance comparison of our models with the existing work. Both of our models performed better than the SemEval baseline and the existing work. Having access to newer transformer models and deep fusion techniques such as GMU offers tremendous value in the ability to capture sophisticated visuo-linguistic signals in memes.

VI. CONCLUSION

The main goal of our research was to investigate and compare multimodal deep learning methods for classifying memes and learning about the latest advancements in transformer-based vision and language models. Two architectures were proposed and compared: one that relied solely on OpenAI's CLIP model (ViT-L/14 and its equivalent text transformer), and another that used ResNet50 for visual representation along with DeBERTa for text sentiment classification. A Gated Multimodal Unit (GMU) was used in both models to efficiently combine the two modalities and learn dynamic feature fusion. Both models performed much better than the official SemEval Memotion 1.0 baseline for Task B. The CLIP-based model obtained slightly better F1-scores than the ResNet50–DeBERTa combination, indicating that CLIP's joint vision-language pretraining helps to provide better synergy between image and text representations.

One of the highlights of this work is the application of the CLIP architecture in the meme classification task, something that has not been done so extensively in this field previously. Our results also emphasize ongoing difficulties in the task. The Motivational class, for example, always performs worse on both architectures. This can be explained due to a variety of reasons:

1. Data imbalance: The Motivational class contains very few labeled samples in the data, causing bad representation in training.

2. Semantic subtlety: Unlike Sarcastic or Humorous memes, Motivational memes tend to use broad, context-dependent knowledge that is hard to deduce from short captions and images alone.

As a direction for future research, larger and more varied meme datasets can be used to train models with user-contextual features (viewer demographics or prior interaction history) to enable humor perception modeling. Advances can also come from the use of cross-modal attention mechanisms or prompt-based learning methods to improve the quality and interpretability of the fusion. Another area is using data augmentation or semi-supervision to resolve data imbalance and further boost performance on underrepresented classes.

REFERENCES

[1] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gamback, "Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor!" *arXiv preprint arXiv:2008.03781*, 2020.

[2] B. Koonce, "Resnet 50," in *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*. Springer, 2021, pp. 63–72.

[3] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[5] H. Xu, "Unimodal sentiment analysis," in *Multi-Modal Sentiment Analysis*. Springer, 2023, pp. 135–177.

[6] V. Ganganwar, Manvainder, M. Singh, P. Patil, and S. Joshi, "Sarcasm and humor detection in code-mixed hindi data: A survey," in *International Conference on Computing and Machine Learning*. Springer, 2024, pp. 453–469.

[7] J. He, S. Mai, and H. Hu, "A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis," *IEEE Signal Processing Letters*, vol. 28, pp. 992–996, 2021.

[8] V. Ganganwar *et al.*, "Sentiment analysis of legal emails using plutchik's wheel of emotions in quantified format," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 6, pp. 4979–4987, 2021.

[9] L. Bonheme and M. Grzes, "SESAM at SemEval-2020 task 8: Investigating the relationship between image and text in sentiment analysis of memes," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 804–816. [Online]. Available: https://aclanthology.org/2020.semeval-1.102/

[10] V. Keswani, S. Singh, S. Agarwal, and A. Modi, "IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1135–1140. [Online]. Available: https://aclanthology.org/2020.semeval-1.150/

[11] H. Singh, N. Helian, R. Adams, and Y. Sun, "Sentiment analysis using blstm-resnet on textual images," 07 2022, pp. 1–8.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] W. Liao, B. Zeng, J. Liu, P. Wei, and J. Fang, "Image-text interaction graph neural network for image-text sentiment analysis," *Applied Intelligence*, vol. 52, pp. 1–15, 08 2022.
[14] I. Bejan, "MemoSYS at SemEval-2020 task 8: Multimodal emotion analysis in memes," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1172–1178. [Online]. Available: https://aclanthology.org/2020.semeval-1.155/
[15] P. Gupta, H. Gupta, and A. Sinha, "Dsc iit-ism at semeval-2020 task 8: Bi-fusion techniques for deep meme emotion analysis," 2020. [Online]. Available: https://arxiv.org/abs/2008.00825
[16] P. Singh, N. Bauwelinck, and E. Lefever, "LT3 at SemEval-2020 task 8: Multi-modal multi-task learning for memotion analysis," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, and E. Shutova, Eds. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1155–1162. [Online]. Available: https://aclanthology.org/2020.semeval-1.153/
[17] K. N. Phan, G. Lee, H.-J. Yang, and S. hyung Kim, "Little flower at memotion 2.0 2022 : Ensemble of multi-modal model using attention mechanism in memotion analysis (short paper)," in *DE-FACTIFY@AAAI*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252015554
[18] S. Sharma, R. S, M. S. Akhtar, and T. Chakraborty, " Emotion-Aware Multimodal Fusion for Meme Emotion Detection ," *IEEE Transactions on Affective Computing*, vol. 15, no. 03, pp. 1800–1811, Jul. 2024. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TAFFC.2024.3378698
[19] A.-M. Bucur, A. Cosma, and I.-B. Iordache, "Blue at memotion 2.0 2022: You have my image, my text and my transformer," 2022. [Online]. Available: https://arxiv.org/abs/2202.07543
[20] B. Duan and Y. Zhu, "Browallia at memotion 2.0 2022: Multimodal memotion analysis with modified ogb strategies," in *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.
[21] G. Ahuja, A. Alaei, and U. Pal, "A new multimodal sentiment analysis for images containing textual information," *Multimedia Tools and Applications*, 2024. [Online]. Available: https://doi.org/10.1007/s11042-024-19999-8
[22] S. Sharma, R. S, M. S. Akhtar, and T. Chakraborty, "Emotion-aware multimodal fusion for meme emotion detection," *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 1800–1811, 2024.
[23] A. Pandey and D. K. Vishwakarma, "Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey," *Applied Soft Computing*, vol. 152, p. 111206, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623012243
[24] S. Lee, Y. Lee, J. Lee, and H. Kim, "A statistical analysis of the relationship between meme stocks and social media," *IEEE Access*, vol. 13, pp. 63 143–63 156, 2025.
[25] A. V, S. S, K. V, S. Kumaran R, and R. D. S, "Crossmemenet: A cross-modal attention framework for meme sentiment analysis using clip and bert," in *2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*, 2025, pp. 1–6.
[26] L. Zheng, H. Fei, T. Dai, Z. Peng, F. Li, H. Ma, C. Teng, and D. Ji, "Multi-granular multimodal clue fusion for meme understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, 2025, pp. 26 057–26 065.
[27] R. Jha, M. R. Panda, S. K C, and A. Dahal, "Deep learning architectures for multimodal sentiment analysis," in *2025 International Conference on Intelligent and Cloud Computing (ICoICC)*, 2025, pp. 1–6.
[28] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.
[29] C. Sharma, D. Bhageria, W. Paka, Scott, S. P Y K L, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gambäck, "SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor!" in *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*. Barcelona, Spain: Association for Computational Linguistics, Sep 2020.
[30] S. Gundapu and R. Mamidi, "Gundapusunil at semeval-2020 task 8: Multimodal memotion analysis," *arXiv preprint arXiv:2010.04470*, 2020.

Vaishali Ganganwar is working as Associate Professor in the Department of Computer Engineering, Army Institute of Technology, Pune. She completed her Ph.D. in Computer Science and Engineering from Vellore Institute of Technology, Chennai, India. She has completed her M.Tech from College of Engineering Pune. She has 20+ years of teaching experience and published several research articles in reputed international conferences and journals. Her research interests include Natural Language Processing, Sentiment Analysis, Machine Learning and Deep Learning.

Gaurav Singh Chauhan has completed B.E.(May 2025) in Computer Engineering at the Army Institute of Technology, Pune. He is currently working as Associate Consultant in ORACLE FINANCIAL SERVICES SOFTWARE LIMITED.

Jangveer Singh has completed B.E.(May 2025) in Computer Engineering at the Army Institute of Technology, Pune.

Shashvat Khajuria has completed B.E.(May 2025) in Computer Engineering at the Army Institute of Technology, Pune.

Vivek Battan has completed B.E.(May 2025) in Computer Engineering at the Army Institute of Technology, Pune.