DAMM: Dynamic Modality Aware Weighted Embeddings Fusion for Multimodal Meme Detection

A Preprint

Mohsin Imam *
University of Delhi
New Delhi, India
mohsingpu@gmail.com

Utathya Aich *
Machine Learning Engineer
CNH Industrial ITC, India
Department of Information Technology
Jadavpur University
Kolkata ,India
us4decaich@gmail.com

Ram Sarkar
Department of Computer Science and Engineering
Jadavpur University
Kolkata ,India
ramjucse@gmail.com

January 25, 2025

Warning: Some content in this paper may be offensive to some readers; reader discretion is advised.

Abstract

The exponential growth of social media platforms and digital communication forums, along with the volume of data being shared, has accelerated the dissemination of information in various forms. However, this surge has also provided masses with routes to spread harmful content, targeting movements, communities, or individuals. Memes, as one of the distinctive form of multimodal content, integrate visual and textual elements to convey nuanced messages, often with a blend of humour and satire to convey complex ideas. Despite their creative potential, memes are often exploited as a medium for advancing hatespeech. These hateful memes frequently target specific groups, attributes, religions, or ideologies, creating social division and animosity. This highlights the dire requirement for robust content moderation systems to ensure that online spaces remain safe, inclusive, and respectful for all. This has alleviated multiple research efforts focusing on capturing hatespeech in online spaces. In this work, we propose `DAMM` (Dynamic Modality-Agnostic Weighted Embedding Fusion for Multimodal Meme Detection), a novel deep learning architecture designed for multimodal analysis. `DAMM` employs multiple multi-modal early fusion approach across weighted image-image and image-text modalities to leverage the strengths of the diverse components of memes, both visual and textual. This is achieved through two sub-modules within `DAMM`: the DeepVisionMixer (DVM) and the CrossEmbeddingMixer (CEM), inputting embeddings generated from CLIP, CNN-based EfficientNet-B3, and a RoBERTa-based text encoder (TweetEval), effectively capturing and analyzing critical features necessary for understanding hate speech in memes. Extensive experiments were conducted on four established datasets — MAMI, Multioff, Memotion 3, and Misogynistic MEME (MIME), demonstrating superior performance as evidenced by comprehensive performance evaluations. Additionally, we

*Equal contribution.

performed modality importance analysis on sample data and conducted ablation studies to validate the optimality of DAMM 's architectural modules.

Keywords Meme categorization · Hateful meme · Multimodal data · Information fusion · Deep learning
Need to discuss

1 Introduction

The growing digital dependence of humans has led to the exchange of everyday information primarily through online platforms, particularly through social media such as Reddit, Twitter, Instagram, etc. Meme is one such information exchange medium that has gained significant prominence in the digital age, functioning as a powerful medium of communication on the internet. Memes are often composed of images with text superimposed, complementing the visual content. They are often used for jokes, social commentary, healthy humor, and creative expression. However, they are increasingly being used to mock individuals or groups and attack specific ideologies, which are conventionally categorized as "hateful memes." Hateful content generated through such memes, containing discriminatory, racist, offensive, or violent messages, leverages their virality to spread hate to a large audience in a much shorter time, reinforcing communal bias and extremist ideology, provoking actions. There are multiple subcategories inside hate speech or cyberbullying spread via hateful memes, such as misogyny, racism, religious persecution, ethnic vilification, political intolerance, etc. Previous research describe hate speech as a hostile and harmful form of speech directed at an individual or a social group, often targeting aspects of their inherent or fundamental characteristics [1]. Content that promotes hostility or violence against individuals or groups is often based on specific characteristics such as ethnicity, race, religion, disability, gender, age, or veteran status. This includes assaults targeting individuals based on their race, identity, gender, character, disability, or other distinguishing attributes, often with the intent to demean or harm a certain group or ideology [2]. Thus, it is important to filter such content from the Internet and maintain a hate-free environment for users.

Hate speech has existed for a long time, initially in textual form, appearing in written materials, public speeches, and broadcasts [3].However, in the exponentially growing digital era, it has transitioned to visual content, complemented by text, referred to as vision-language hate speech [4]. There are multiple subcategories inside hate speech or cyberbullying spread via hateful memes, such as misogyny, racism, religious persecution, ethnic vilification, political intolerance, etc. Previous research describe hate speech as a hostile and harmful form of speech directed at an individual or a social group, often targeting aspects of their inherent or fundamental characteristics [1]. Content that promotes hostility or violence against individuals or groups is often based on specific characteristics such as ethnicity, race, religion, disability, gender, age, or veteran status. This includes assaults targeting individuals based on their race, identity, gender, character, disability, or other distinguishing attributes, often with the intent to demean or harm a certain group or ideology [2]. Thus, it is important to filter such content from the Internet and maintain a hate-free environment for users.

Figure 1: Interplay of text and image in classifying memes, where the textual content influences the overall meaning of a picture, shifts its interpretation from non-hateful to hateful.

Multimodal problems involve tasks that require the integration of data from multiple modalities or sources, such as text, images, audio, or video, to enhance understanding and make more accurate predictions by leveraging all available information. In contrast to multimodal problems, unimodal tasks typically involve a single type of data, such as text,images,etc. Multimodal problems demand models that can process and

High Correlation | Moderate Correlation | Low Correlation

Figure 2: Text-Image Correlation in memes highlights the relevance of influencing and perceiving the categorization of a meme in a particular class.

combine information from diverse formats. These tasks can be particularly challenging because it is often more difficult to fuse information from modalities that vary in scale and are independent of one another. The interplay of text and image in classifying meme is very crucial.In Figure 1 (left), an image of an old man resting his face on his hand with a thoughtful expression, typically conveys a neutral or positive interpretation. The addition of Hateful text in Figure 1 (in the middle) drastically transforms the meaning of the image. The phrase *"When they asks why you dropped the muslim kid off the school to give the parents back their bomb"* reframe the same visual as symbol of hatred, making it Hateful. Whereas in the Figure 1 (right), the addition of the phrase *"Should I buy another Lamborghini, or will the neighbors start talking again?"* augment a positive or neutral narrative as a whole making the situation leaning towards non-hateful category.

Modalities Correlation. Detecting multimodal hateful memes poses a challenge due to the inherent complexities associated with the interplay of two or more modalities. In certain cases, the textual and visual elements are strongly correlated, which can aid in the detection and classification of the meme as either hateful or non-hateful. As illustrated in figure 2 (left), the image of a man with high correlation to sexualizing actions is paired with the caption *"DAT ASS."* In this instance, the alignment between the visual and textual contents clearly conveys hateful intent, making the classification process more straightforward due to the explicit visual nature of the sexual undertones, inclining it towards the hateful category. However, it is important to note that this does not unequivocally classify the meme as hateful. The lack of explicit association with a specific entity, race, or group positions the meme in a negative context but not necessarily in the hateful category. While such content might offend certain groups of social media users, others could perceive it as humorous, highlighting the subjective nature of such memes. In contrast, in many cases, there is minimal correlation between the modalities, making it challenging even for humans to categorize the content into a specific class. An example of this is depicted in Figure 2 (see middle), where an image of a contemplative sun bear resting quietly includes a caption conveying bear's personal confession with a humorous, self-deprecating tone is paired with *"I'm Such a 9gagger That I Don't Have Enough Courage to Call a Hooker Because I'm Afraid of Getting Rejected."* Like aforementioned nature of subjective interpretation, it is important to note that there is no explicit racism evident in this particular meme. The humor stems from the confession's ironic and exaggerated insecurity, without targeting any specific group. However, in certain contexts, subtle associations with racial undertones could arise, particularly when interpreted in connection with stereotypes particularly about black individuals, making it placing at liminal position in contrastive categories of hate detection systems. In multiple cases, the counterpart visual element diverges from the ominous and threatening tone of the text, creating ambiguity and making the intent unclear. An example of such a case is shown in Figure 2 (right), which presents a sparse representation: a woman sitting in fear, surrounded by spiders, coupled with a caption that refers to spiders as terrorists. While this combination creates an atmosphere of fear as visible by the state of the lady, it does not establish a direct correlation that allows the instance to be definitively classified into either category (i.e., negative or positive). These examples when represented as sophisticated feature embeddings, when represented in a joint feature space, results in sparse representation which complicates the classification process. Both the text and image interaction are needed to create meaning to determine whether a message is perceived as hateful or non-hateful. As a result, advanced representation models are needed to effectively handle and integrate these diverse modalities to effectively capture and integrate multimodal information to classify the meme [5].

Over the last decade, the concept of multimodal data fusion has been widely applied in the deep learning domain, addressing various problem areas across multiple modalities [6]. In the audio-video domain, tasks

include audio summarization, video summarization, and event detection in videos. For image-text integration, applications encompass image captioning, visual question answering (VQA), visual entailment, visual reasoning, and image-text retrieval. Lastly, in the video-text domain, where frames of videos are typically processed, tasks involve video-text retrieval, action recognition in videos, and generating textual summaries of video content. This has been achieved using models like Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and attention mechanisms. Different forms of categorization have been proposed by researchers for different types of data fusion [7].

Hierarchical feature fusion leverages the ability of deep neural networks (DNNs) to learn hierarchical representations, allowing multimodal features to be fused at different abstraction levels rather than directly combining raw data. This approach integrates features from various levels of the network, utilizing their combined strengths to improve model performance more effectively. Some other forms of data fusion involve combining data through multiple operations such as element-wise summation, multiplication, concatenation and cross-product of different modality embeddings [8]. Decision-level fusion is a straightforward approach, where cross-modal information is combined at the final or penultimate layers of decoders or classifiers. While easy to implement, it offers limited flexibility and interpretability of multimodal interactions [9]. Hee et al. [4] emphasize the importance of developing hate speech detection models that not only combine representations from different modalities but also effectively capture the contextual subtleties of multimodal inputs.

One significant challenge is the identification and interpretation of implicit hate speech, where the harmful intent is subtly embedded in seemingly neutral language or actions [10, 11]. This complexity arises from the nuanced nature of human communication, where hate speech can be delivered indirectly or through coded language, making it difficult to detect. Furthermore, obtaining data from various platforms such as Reddit, YouTube and 4chan complicates the process due to differences in content formats and platform specific contexts, leading to difficulties in standardization and interpretation [12]. Moreover, the uneven distribution of hate speech across dataset presents a challenge in training and inferencing models effectively, limiting the model's ability to generalize [13].

To address the research gaps of implicit hate speech, data complexity, model generalizability, and adaptability, we introduce Dynamic Modality Agnoistic Weighted Embeddings Fusion for Multimodal Meme Detection (`DAMM`), a deep learning-based architecture designed to tackle these challenges. `DAMM` employs a multiple-feature fusion design, which integrates a prior distribution loss function powered by multi-head attention mechanism for both binary-class and multiclass classification of memes into hateful and non-hateful categories. To effectively weigh the fused-modality representations of both text and image features, we incorporate a squeeze block that is integrated with the fused embeddings, which is specifically designed to enhance the model's ability to understand and classify complex, multi-modal content in a nuanced and context-aware manner.

In brief, following are the main contributions of our paper:

- We introduce dynamic weighted embeddings leveraging an early fusion approach that integrates intra-modality and inter-modality fusion for rich contextualization of spatial and linguistic features for robust meme classification in a multimodal setting.
- We employ a Squeeze-and-Excitation (SE) block to dynamically adjust the intra-modality visual weighting for feature representations generated by CLIP and EfficientNetB3, allowing the creation of a fused representation, where the contribution of each modality is adaptively optimized. This fused representation is further combined with text features from TweetEval, ensuring that the most relevant features from both visual and textual modalities are emphasized. The SE block assigns higher weights to the modality that contributes more significantly, maintaining most relevant features from each modality are emphasized in the final representation.
- We perform a thorough evaluation on a range of hatespeech genres including misogyny, political offensiveness, standard hatespeech, enabled by analysis of four datasets, namely MAMI, MultiOFF, Memotion 3 and Misogynistic-MEME (MIME), seminal in multimodal hatespeech research and demonstrate the effectiveness of `DAMM` by performing detailed ablation analysis.

2 Related Work

Due to the high usage of memes in spreading hate and persuading the public against specific ideologies or political movements, a significant number of researchers have focused on detecting hateful content on social media. In this section, we formally discuss previous research pertaining to multimodal hate meme detection.

Although advancements have previously been made, machine learning has significantly enhanced hate speech detection across languages by employing supervised models like FastText, SVM, Multinomial Naïve Bayes, and Logistic Regression. These models classify hate speech in datasets from platforms like Twitter, YouTube, and Wikipedia, addressing diverse categories such as hateful, offensive, and clean content with improved accuracy [14, 15, 16, 17]. Since, memes consist of both visual and textual elements, making a multimodal framework essential for their detection. Such frameworks integrate image analysis with language processing to evaluate the visual and textual aspects of memes simultaneously. To find a proper fusion strategy, researchers have explored various fusion methods, primarily categorized into early fusion and late fusion approaches [18].

Nguyen et al. [19] addressed the challenge of multimodality in hate meme classification on Memotion 2.0 dataset, by leveraging advanced attention mechanisms such as multi-hop attention and stacked attention to produce comprehensive aggregated features. To tackle the issue of data imbalance, they employed an auto augmentation method. For feature extraction, they utilized EfficientNet-V2 for images and for text they used RoBERTa and LSTM . The authors explored three types of fusion models: concatenation, multi-hop attention, and stacked attention networks. Additionally, they implemented a reinforcement learning-based augmentation technique that dynamically generates optimal augmentation strategies. To ensure that the visual modality focuses solely on the image content, they used the EAST (Efficient and Accurate Scene Text Detector) module to detect the the text and remove it from the images before extracting visual features. For the sentiment detection task, their dual-modality fusion approach with multi-hop attention achieved a Weighted F1 score of 0.5316.

Deng et al., in [20], proposed MuAL, a lightweight model that addressed limitations often encountered with contrastive learning-based models like CLIP or BLIP, as well as large transformer-based models like VisualBERT. MuAL employed Cross-Modal Attention (CAM) for cross-modal information integration, enabling it to capture richer semantics between modalities. The authors introduced a difference loss function to reduce discrepancies in image and text feature representations by imposing a constraint, thereby enhancing the model's robustness. Their approach outperformed existing methods. The authors also focused on testing their approach by freezing the underlying pre-trained models, highlighting MuAL's suitability for transfer learning. However, the study did not discuss the potential noise that could have arisen when integrating features from different modalities using CAM.

In the work [21], researchers proposed a targeted approach for detecting hateful memes with a focus on religious sentiments. They curated a dataset comprising over 2,000 meme images paired with their respective captions. For feature extraction, they utilized a ResNeXt-152-based Mask R-CNN for images and BERT for encoding text features, for the generation of a comprehensive feature representation. These features were combined using an early fusion technique. Additionally, the solution was fine-tuned with VisualBERT. To diminish biases in the collected dataset, the authors extended it by incorporating the Facebook Hateful Memes dataset. However, since the study mainly focused on religious hateful memes, the authors did not explore expanding their approach across diverse culture, heritage, or geographical contexts.

The study conducted by Wang et al. [22] focused on a comparative analysis of UNITER (UNiversal Image-TExt Representation) [23], a unified stream approach that processes image and text pairs together in a single stream, and the processing of independent image and text features of misogynistic memes using CLIP. The test aimed to evaluate how strongly or weakly the image and text were related. The authors used XGBoost as a final classifier, which gave the best results when features were extracted with CLIP. The study also examined the effect of domain shift, where both approaches performed worse compared to cases with no domain shift. However, CLIP performed better than UNITER in domain shift scenarios. From this analysis, they proposed the PBR (Pretrained, Boosting, Rule-based adjustments) method for misogyny detection, aided by manual rule-based adjustments. The proposed approach involved training XGBoost on image features extracted by CLIP and fine-tuning UNITER and BERT for image-text and text-only tasks using the MAMI dataset. Subsequently, the XGBoost predictions were refined using outputs from UNITER and BERT along with logical inference, achieving the highest macro F1 score of 0.834, placing first in SemEval-2022 Task 5.

Roy et al. [24] created a framework called MMFFHS (Multi-Modal Feature Fusion for Hate Speech Detection on Social Media) to detect hate speech using both text and images. This method was tested on a large dataset with 150,000 examples, divided into six groups: homophobic, religious, other hate, racist, sexist, and non-hate. For text, the framework used an LSTM model, and for images, it used ResNet50. The features from both text and images were combined and passed through a dense layer to classify the data. This research aims to detect hate speech on social media platforms like Twitter, Facebook, and YouTube, using both text and images. Three setups were tested: using only text and image, respectively and using both (multimodal). The LSTM model alone achieved an accuracy of 0.69, and combining text and image data gave better results.

Results showed that combining text and image data worked best, with precision of 0.72, recall of 0.68, an F1 score of 0.69, and accuracy of 0.70.

Arya et al. [25] proposed a contrastive learning framework combined with prompt engineering techniques to identify hate speech in viral meme content on social networking platforms. The authors leveraged CLIP to study the model's ability to correlate image and text features associated with memes. They set a threshold of 20% for classifying memes as non-hateful, acknowledging that the model may fail to relate image and text features even if the meme is inherently hateful.To enhance classification accuracy, they developed custom prompts for "good memes" and "hateful memes," which were used to classify memes. For memes exceeding the 20% threshold on matching the image features and associated text features in a shared latent space, the authors further calculated the cosine similarity between image and caption embeddings against predefined "good meme" and "hateful meme" templates. If both the image and caption showed stronger similarity to the "hateful meme" description, the meme was classified as hateful; otherwise, it was categorized as "non-hateful.

In [26], to analyze the intricate relationship between meme images and their embedded captions, a triplet relation model was proposed to improve the connection between visual regions of memes and their associated captions, addressing the complex reasoning required for hate meme analysis. The model comprised an encoder-decoder module to generate captions for the image, an OCR extraction module, and an object detection module powered by R-CNN to extract semantic regions from the meme. The unified text and image modalities were then fed into a triplet relation network within a transformer block, leveraging self-attention mechanisms. Experiments were conducted with both one-stream and two-stream approaches: in the one-stream approach, textual embeddings were combined before processing, whereas in the two-stream approach, the two textual embeddings were processed separately. Another study by Yang et al. [27] explored the fusion of meme visual and associated textual features using a combination of 1D CNN, max pooling, and MLP. The study demonstrated the comparative effectiveness of various fusion approaches, including attention-based fusion, gated fusion, and bilinear fusion techniques.

The proposed architecture leverages a multi-level fusion strategy across multiple hierarchical levels of image representation from dual visual models, followed by the integration of both image and text modalities enabled by a RoBERTa-based encoder. We utilize the Squeeze-and-Excitation (SE) mechanism to enhance feature representations of the image modality through the synergistic use of EfficientNetB3 and CLIP models for improving visual understanding of meme images. For the textual modality, we employ the TweetEval model, known for its robust performance in sentiment analysis tasks. This dual-modality approach facilitates rich, contextualized representations, which are subsequently fused to improve the model's overall performance. A comprehensive description of the methodology is provided in the subsequent section.

3 Methodology

To address the complex relationships inherent in meme hate detection, we propose `DAMM`, *D*ynamic modality *A*ware weighted embeddings fusion for *M*ultimodal *M*eme detection, as shown in Figure 3, a three-stage framework leveraging an early fusion strategy. This approach integrates both intra-modality and cross-modality interactions, specifically focusing on the interplay between image and text modalities. The framework is designed to capture and represent the nuanced features associated with each modality more effectively, thereby enhancing the detection capabilities for hate speech in multimodal contents. In the following sections, we discuss each module and step of the framework in detail, highlighting their rationale and relevance in constructing `DAMM`.

Problem Formalization. Given a meme X, let the meme image be represented as X^i, where $i \in \{1, 2, \ldots, n\}$, with n being the total number of samples in our dataset, and the associated caption as X^t. Each meme X comprises two modalities: $m \in M$, where $M = \{\text{Image}, \text{Text}\}$. The objective is to classify each meme X into one of the predefined classes $\{C_1, C_2, .., C_k\}$, where k represents category of memes primarily `non-hateful` or `normal` memes and `hateful` memes. Hateful memes include content that is objectifying, derogatory, racist, or contains other forms of harmful or offensive messaging.

3.1 Dynamic modality Aware weighted embeddings fusion for Multimodal Meme detection (DAMM)

Our framework, referred to as `DAMM`, as shown in Figure 3 designed to address the binary and multiclass problem of distinguishing between hateful and non-hateful content. The key innovation of `DAMM` lies in its novel weighted embedding fusion mechanism, which leverages embeddings from multiple modalities. The fusion process is enabled by two modules proposed within `DAMM`: DeepVisionMixer (DVM) and CrossEmbeddingMixer

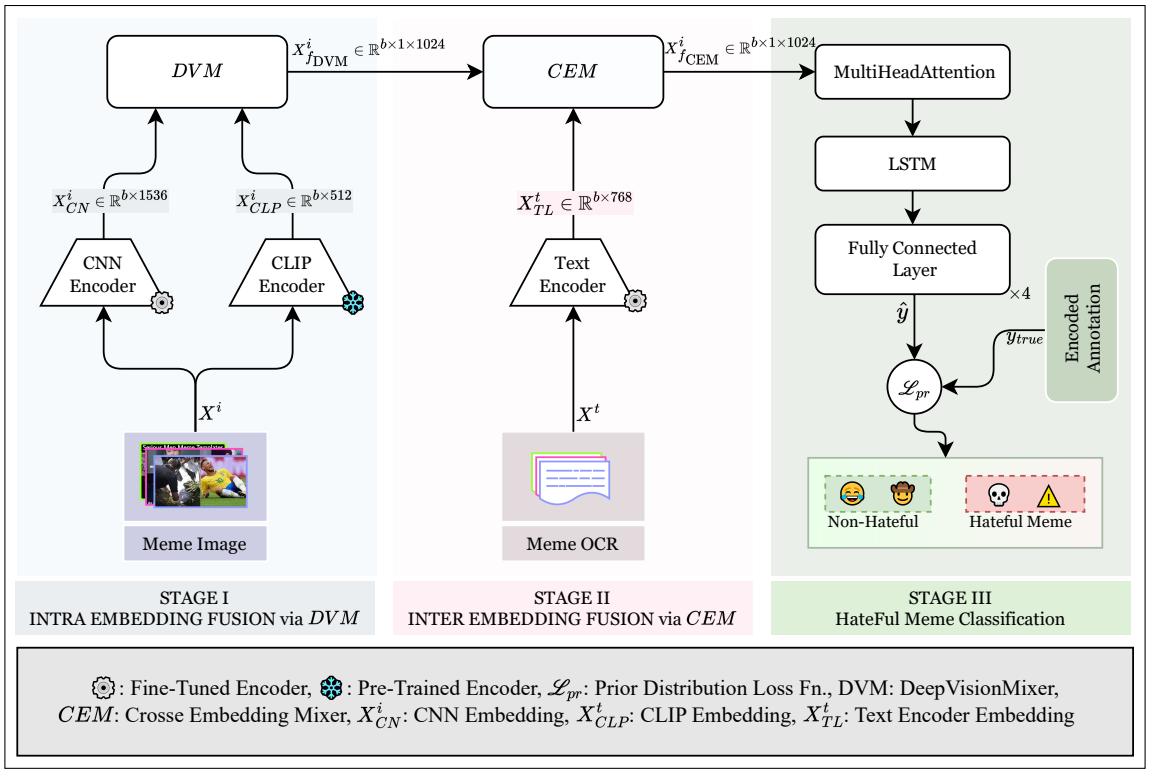


Figure 3: The overall architecture of DAMM is used for multimodal meme classification. Stage I comprises generating embeddings from CNN (EB3) and CLIP, which are used as inputs in the DVM block for generating detailed visual features. In stage II, the CEM takes the meme text embedding and the DVM's output as input, generating a cross-modal representation of the whole meme. This is then passed to the stage III of DAMM, aided by Multi-Head Attention and a Feed Forward Neural Network, for final classification.

(CEM). The DVM module plays an important role in processing and understanding visual features, ensuring that the extracted embeddings are both contextually aware and rich in semantic details. These embeddings are generated by two distinctive visual-spatial feature generation models. Meanwhile, the CEM module is dedicated to combine embeddings across modalities, enabling interaction between diverse data streams such as text and image features. Together, these modules ensure that embeddings from each modality are not only concatenated but integrated in a way that focuses on their contextual and complementary contributions which is important in multimodal tasks emphasizing on mutual fusion (refer to the section 4.4). This advanced approach addresses the limitations of traditional fusion techniques, such as simple concatenation, by generating weighted, contextualized, and robust embedding representations. Through extensive experimentation, we have demonstrated that the weighted embedding fusion mechanism employed by DAMM consistently outperforms existing approaches, providing a powerful and effective solution for tackling challenging multimodal tasks like meme classification.

3.1.1 Weighted Embedding Fusion

The weighted embedding fusion process is performed via squeezing the modalities and simple concatenation. In squeezed approach, the feature representation vector is passed through a dense layer for performing linear squeeze operation, which outputs a scalar value of $\mathbb{R}^{1\times1}$. This scalar value, derived from the respective embeddings, is then concatenated ($\oplus$) along the first axis. Squeezing, intrinsically captures a global summary of the respective embeddings, which is crucial for assigning weight contributions to the embeddings. This weight assignment is achieved through a dot product ($\odot$) operation between the scalar output of the squeezed approach and the feature representation from the simple concatenation operation known as unsqueezed concatenation. Unsqueezed concatenation (or normal concatenation), preserves the embeddings in their dimensions by combining them without any transformation or compression, to enable compatibility for the dot product operation, the resulting feature representation is reshaped appropriately. These comprehensive

embeddings, enables the fusion process to leverage both the global summary and the detailed feature representations effectively.

Let us consider two embeddings from two modalities generated from respective modality specific encoder, denoted as I_1 and I_2, where $I_1 \in \mathbb{R}^{1\times a}$ and a representing the embedding dimension of I_1. Similarly, $I_2 \in \mathbb{R}^{1\times b}$, with b representing the embedding dimension of I_2 such that $a >= b$ or $b >= a$. The final weighted embeddings is implicitly determined through the dot product of the embeddings, i.e., the squeezed and unsqueezed concatenated embeddings.

The squeeze function, denoted by $\mathbf{F_{sq}}$, is applied to feature representation of both modalities (I_1, I_2) as follows:

$$i_1 = \mathbf{F_{sq}}(I_1), i_2 = \mathbf{F_{sq}}(I_2) \tag{1}$$

where $i_1 \in \mathbb{R}^{1\times 1}, i_2 \in \mathbb{R}^{1\times 1}$. These squeezed embeddings are scalars derived from the input embeddings. The squeeze function, $\mathbf{F_{sq}}$, has a learnable parameter W, where $W \in \mathbb{R}^{n\times 1}$, and n represents the size of the feature input for a particular modality. The squeeze function is defined as $\mathbf{F_{sq}} = W \cdot I$, where I represents embedding from some given modality.

Following the application of the squeeze operation on both modalities, the concatenation of these embeddings, i_1 and i_2, facilitated by the concatenation function $\mathbf{F_c}$ along the first axis, combines the squeezed embeddings as:

$$i_{f_{sq}} = \mathbf{F_c}(i_1, i_2), \text{where } \mathbf{F_c}(i_1, i_2) = i_1 \oplus i_2 \tag{2}$$

Here, the concatenated embedding $i_{f_{sq}} \in \mathbb{R}^{1\times 2}$ represents the fused squeezed representation of the two modalities.

Since the calculation of the weighted embedding requires the unsqueezed (or normal) concatenated embedding, it is obtained using $\mathbf{F_c}$, which takes as input the original embeddings derived from the respective modality encoders. The unsqueezed concatenation is defined as:

$$i_f = \mathbf{F_c}(I_1, I_2), \quad \text{where } \mathbf{F_c}(I_1, I_2) = I_1 \oplus I_2 \tag{3}$$

Now, after obtaining both the squeezed concatenation $i_{f_{sq}}$ and the unsqueezed concatenation i_f, dot product operation is performed after reshaping i_f to avoid dimensionality issues. This operation integrates the information from both modalities by leveraging their respective representations. The dot product is defined as:

$$i_w = \mathbf{F_w}(i_{sq}, i_f), \text{where } \mathbf{F_w}(i_{f_{sq}}, i_f) = i_{f_{sq}} \cdot i_f \tag{4}$$

where i_w represents the weighted embedding, encapsulating the fused information from both modalities. This step ensures that both the compressed scalar embeddings ($i_{f_{sq}}$) and the full-dimensional concatenated (I_f) embeddings contribute to the final fused representation.

The produced weighted embedding leverages scalar weights generated via squeezed operation, $\mathbf{F_c}$, which adjusts the importance of each modality based on its context, ensuring relevance for the specific modality. This is not only limited to cross-modality; even when two feature representations are obtained via different encoders but from same modality or same data, the it allows for weighing the importance of different features based on data from the same modality (i.e., intra-modality). Moreover, the scalar weights offer clear interpretability, focusing on the contribution of each modality, improving the semantic understanding when dealing with cross-modality and intra-modality tasks. Additionally, this weighted embedding function, $\mathbf{F_w}$, can be extended to multiple modalities as well. Thus, this modality-agnostic embedding generation makes it more scalable to other multimodal tasks, making it more versatile compared to existing methods. This novel weighted embedding fusion approach in the DVM and CEM modules leads to a richer, more nuanced, and contextualized representation of multimodal data, leading to superior performance.

3.1.2 DeepVisionMixer — DVM

DeepVisionMixer is the first stage of `DAMM`, which extracts embeddings from both a CNN-based model and CLIP. Based upon [28], where the proposed fusion method learns relationships between the modalities'

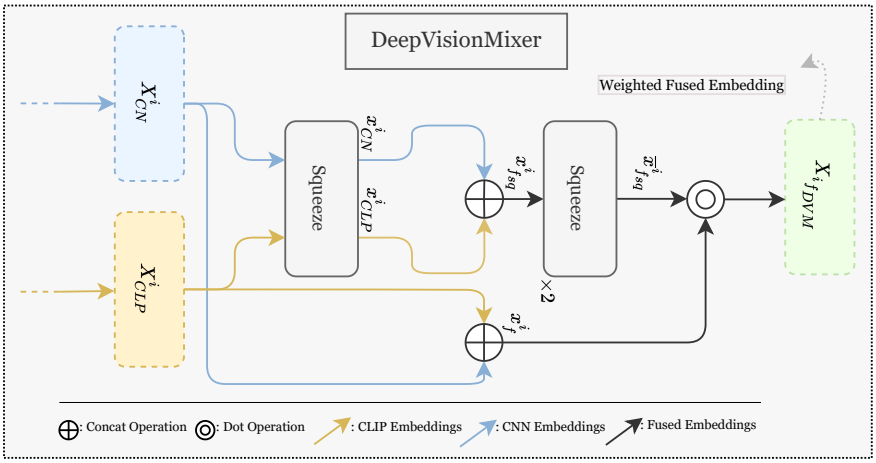


Figure 4: An illustration of the DeepVisionMixer, which takes as input the feature representation from the CNN and CLIP encoder.

dependence for textual and visual modalities of memes, we adopt a similar approach. They used a squeezed and excitation block for cross-modality interaction, which we have projected in the DVM block, but for the same modality i.e., intra-modality — the visual modality from the meme image (see figure 4). Since combining text with image modalities for memes is a well-established fusion based method for addressing meme classification, the rationale for using the same modality and exploiting their relationship via weighted embeddings is to examine the impact of different visual feature extraction models when combined with the text modality. Features from the image data are extracted using two feature representation models: CNN and CLIP. CNN focuses on extracting deep visual features, which are crucial for capturing fine-grained spatial details in images. In contrast, CLIP, built on a contrastive learning framework, excels in aligning visual features with semantic concepts due to its multimodal pre-training on vast image-text pairs. CLIP's strength in encoding semantically rich, text-aligned representations enhances the overall feature quality, enabling robust performance in tasks requiring both low-level detail and high-level abstraction. Thus DVM block serves a critical component in `DAMM` framework designed to effectively integrate visual features produced through a dual channel (i.e., CNN and CLIP). This enables `DAMM` to facilitate a comprehensive understanding of visual data enabling enhanced meme image understanding.

EfficientNet-B3 (EB3) — EfficientNet [29] is a hierarchical family of CNNs that focuses on computational efficiency while enhancing performance. The EfficientNet-B3 model, a variant within this family, is designed to achieve a high level of accuracy while maintaining computational efficiency. The core building component of EfficientNet-B3 is based on compound scaling principle. The scaling factors are determined by a compound coefficient. Unlike traditional CNN architectures that scale depth, width, or resolution independently, EfficientNet scales these three dimensions in a balanced manner to achieve better performance with fewer resources. The compound coefficient determines the optimal scaling of all three dimensions, improving the network's accuracy and efficiency. Each version (EfficientNet-B0 to EfficientNet-B7) applies a different compound scaling factor, leading to differences in performance and computational cost. The transition from EfficientNetB0 to EfficientNet-B3 increases the depth, width, and input resolution, resulting in better performance.

The core of EfficientNet-B3's architecture is based on Mobile Inverted Bottleneck Convolutions (MBConv), which leverage depthwise separable convolutions. This enables a reduction in computational complexity while maintaining or improving performance compared to other existing convolutional architectures. EfficientNet-B3 uses squeeze-and-excitation (SE) blocks to further enhance feature representation by refining channel-wise feature responses.

CLIP — CLIP [30], which stands for Contrastive Learning-Image Pretraining, is a framework introduced by OpenAI based on the contrastive learning methodology to learn image and text representations in a shared embedding space. The focus is on training a single model that can associate natural language descriptions with corresponding images and effectively differentiate non-similar images. This capability can be utilized for multiple problems involving image and text modalities, making it a state-of-the-art choice for many multimodal tasks. The essence of contrastive learning lies in mapping both images and their corresponding

textual descriptions into a joint embedding space. The training objective, aided by contrastive loss, persuades the model to contrast positive pairs (correct pairs of images and their descriptions) with negative pairs (incorrect combinations of images and text). CLIP's training goal is to maximize the similarity between positive pairs while diminishing the similarity of incorrect (or negative) pairs.

$$\mathcal{L} = -\log\left(\frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)}\right) \tag{5}$$

In equation 5, the variable $\mathbf{v}_i$ denotes the image embedding for the i-th image, while $\mathbf{t}_i$ represents the corresponding text embedding. The function $\text{sim}(\mathbf{v}_i, \mathbf{t}_j)$ computes the cosine similarity between the image embedding $\mathbf{v}_i$ and the text embedding $\mathbf{t}_j$. The parameter τ is a temperature term that adjusts the sharpness of the softmax function.

The training of CLIP on a vast dataset (~400 million image-text pairs) curated from a variety of publicly available sources on the Internet allows it to generalize to many unseen tasks, such as image classification and object detection, using only textual prompts eliminating task-specific retraining. In addressing our problem, we leverage CLIP directly for feature extraction, as the majority of the visual content in meme images aligns well with the data on which CLIP has been pre-trained. While fine-tuning would undoubtedly improve feature enrichment, we opted not to, due to computational constraints. We chose to freeze CLIP's weights to preserve its generalization ability and avoid overfitting. Fine-tuning CLIP on a small dataset could compromise its ability to maintain the broad, however, high-level semantic features needed for grasping the contextual and conceptual aspects of memes.

We leverage `CLIP-ViT-B-32` for image feature extraction in the intra-modal fusion of the DVM block, we normalize image embeddings using the L2 norm along the last dimension of the extracted embeddings, preserving the normalized feature dimensions for appropriate further operations, ensuring that the embeddings lie on a unit hyper-sphere. This process confirms numerical stability and scale invariance, which are important for the balanced alignment. This normalization is numerically expressed in the following equation:

$$\mathbf{v}_{\text{norm}} = \frac{\mathbf{v}}{||\mathbf{v}||_2} = \frac{\mathbf{v}}{\sqrt{\sum_{i=1}^{m} v_i^2}} \tag{6}$$

Mathematically, the operations performed in the DVM block are as follows:

Let X denotes an arbitrary sample from the dataset, with X^i representing the meme image and X^t the corresponding OCR caption from the meme image. The feature vector obtained from the CNN block is (i.e., from EfficientNetB3) represented as X^i_{CN}, where $X^i_{CN} \in \mathbb{R}^{1\times 1536}$ and the feature vector obtained from CLIP is represented as X^i_{CLP} where $X^i_{CLP} \in \mathbb{R}^{1\times 512}$. The first operation involves a squeeze operation, which is essential for enabling the weighted embedding, and is defined as:

$$x^i_{CN} = \mathbf{F_{sq}}(X^i_{CN}), \quad x^i_{CLP} = \mathbf{F_{sq}}(X^i_{CLP}) \tag{7}$$

The $\mathbf{F_{sq}}$ represents the squeeze function, which produces the global summaries of the input feature representations obtained from EfficientNetB3, denoted as x^i_{CN}, and from CLIP, denoted as x^i_{CLP} (see equation 1). These summarized feature representation are then concatenated represented as $x^i_{f_{sq}}$, which then undergo two dense neural network layer neural network layers:

$$x^i_{f_{sq}} = \mathbf{F_c}(x^i_{CN}, x^i_{CLP}) \tag{8}$$

The $\mathbf{F_c}$ represents the concatenation function as described in section 3.1.1. This obtained concatenated summarized features contains representation from both image features obtained from respective models which is fed into $\mathbf{F_{ac}}$ activation, containing a dual series of dense layer, where first dense layer is powered by rectified linear unit, ReLU ($\mathcal{R}$) activation function, and the latter one is supported by Sigmoid activation function, $\mathcal{S}$, as follows:

$$\bar{x}^i_{f_{sq}} = \mathbf{F_{ac}}(x^i_{f_{sq}}), \text{where } \bar{x}^i_{f_{sq}} \in \mathbb{R}^{1\times 2} \tag{9}$$

$$\mathbf{F_{ac}}(x^i_{f_{sq}}) = \mathcal{R}(W_{\mathcal{R}}(\mathcal{S}(W_{\mathcal{S}} \cdot x^i_{f_{sq}}))) \tag{10}$$

In equation 10, $W_{\mathcal{S}}$ and $W_{\mathcal{R}}$, refer learnable weight matrices for enabling appropriate shape for performing the intra-modality weighted fusion. The purpose of the passing summarized feature representation ($x^i_{f_{sq}}$) through the ReLU activation function ($\mathcal{R}$) and subsequently through Sigmoid activation ($\mathcal{S}$) is to refine and normalize the squeezed modal weights for effective fusion, enabling the generation of intra-modality representations essential for classification. $\mathcal{R}$ introduces non-linearity by emphasizing positive inter-dependencies and suppressing insignificant or negative values, ensuring robust feature selection. This step enhances sparsity and highlights dominant modal features. $\mathcal{S}$ maps the refined values into the range (0, 1), converting them into normalized probabilistic weights that emphasize the importance of the features obtained from their respective models.

To obtain the normal concatenated fusion from the initially extracted features from the respective encoders (X^i_{CN}, X^i_{CLP}), these features are combined using the concatenation function $\mathbf{F_c}$ (refer to equation 2). The concatenated output is then reshaped into dimensions of $2 \times (\mathcal{N}/2)$, where $\mathcal{N}$ represents the fused embedding weight in the latent space. This process allows the generation of the weighted fusion between $\bar{x}^i_{f_{sq}}$ and x^i_f, producing the final output of the DVM block, X_{DVM}, as follows:

$$x^i_f = \mathbf{F_c}(X^i_{CN}, X^i_{CLP}) = X^i_{CN} \oplus X^i_{CLP}, \quad \text{where } x^i_f \in \mathbb{R}^{1\times 2048} \tag{11}$$

$$x^i_f \in \mathbb{R}^{1\times 2048} \rightarrow \mathcal{RESHAPE} \rightarrow x^i_f \in \mathbb{R}^{1\times 2\times 1024} \tag{12}$$

$$X^i_{DVM} = \mathbf{F_w}(\bar{x}^i_{f_{sq}}, x^i_f), \quad X^i_{DVM} \in \mathbb{R}^{1\times 1\times 1024} \tag{13}$$

The X^i_{DVM} representing the fused intra-modality based on the weighted embedding mechanism serves as one of the inputs for the second stage of `DAMM` i.e., CrossEmbeddingMixer (CEM) as discussed in the following section.

3.1.3 CrossEmbeddingMixer — CEM

The CEM module is the second stage of `DAMM`. This stage is akin to the typical multimodal classification framework, which generally involves two different modalities, typically text and image, for most meme-focused problems. As visible in figure 5, the CEM block is build upon DVM block with multiple variations. In multimodal settings, visual content, independently, often falls short of providing sufficient clarity for categorizing data into a specific category. Textual data can often independently aid classification due to its explicit nature, such as the presence of words that directly indicate detrimental categories, including abusive language or racist remarks. In such cases only, the significance of other modalities may be undermined. CEM enables inter-modal fusion, where the visual modality encompasses the already fused intra-modal available from the DVM block, and the textual available from each dataset, generally extracted by different methods.

TweetEval [31] is an unified benchmark designed to study Twitter data in the NLP landscape, introduced by the Cardiff NLP group. It consists of seven main tasks that capture essential aspects of natural language evident in most tweets, including sentiment analysis, emotion recognition, irony detection, stance prediction, and emoji detection. Notably, the evaluations for benchmarking are performed independently but are encompassed within the same TweetEval framework. While primarily focused on classification, TweetEval's structure is highly versatile, with potential applications extending to multi-label and multimodal tasks. We have employed the fine-tuned `RoBERTa-base` model utilized in TweetEval's evaluation made available from hugging face[2]. This model, trained on approximately 58 million Twitter tweets, was further fine-tuned specifically for sentiment analysis under the TweetEval benchmark. In our experiments, we have utilized TweetEval for feature extraction due to its relevance to hate speech detection tasks. TweetEval includes datasets and models fine-tuned for textual analysis, particularly in domains like hatespeech, making it well-suited for extracting features from the OCR-generated captions of memes. This ensures the linguistic nuances of hatespeech are effectively captured, thereby cooperating with a multimodal focus of our study.

The operation in the CEM blocks begins by projecting the text and fused (intra-modal) image embeddings into scalar values to produce global summaries. The text feature, represented by X^i_{TL}, is extracted using a

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

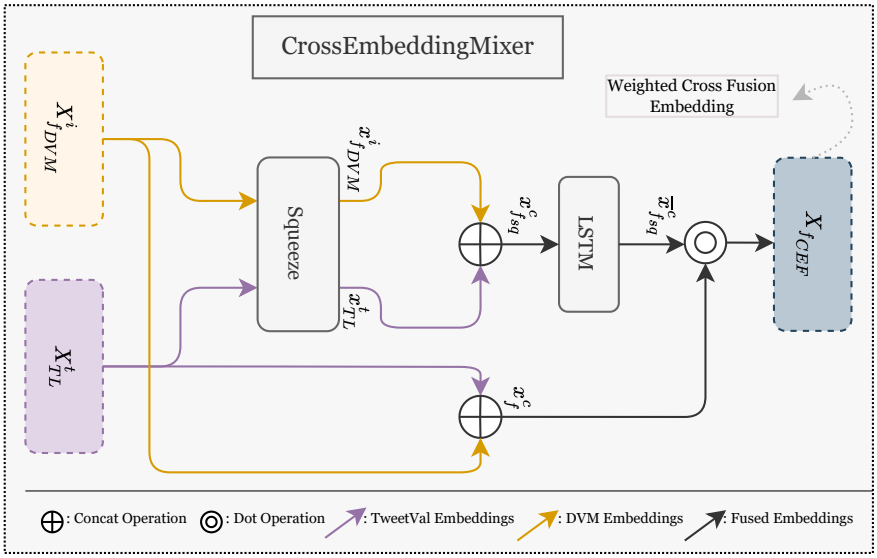


Figure 5: An overview of the CrossEmbeddingMixer, which takes as input the intra-modal fused embedding from the DVM block and the TweetVal-generated representation derived from meme-embedded text, producing a cross-modal representation of the meme image.

fine-tuned TweetEval model on the respective datasets. This process is performed similarly to the approach used in the DVM block, as follows:

$$x_{DVM}^{i} = \mathbf{F_{sq}}(X_{DVM}^{i}), \quad X_{TL}^{t} = \mathbf{F_{sq}}(X_{TL}^{t}), \quad \text{where } X_{DVM}^{i} \in \mathbb{R}^{1\times1\times1024},\ x_{TL}^{t} \in \mathbb{R}^{1\times768} \tag{14}$$

Here, the squeezed $x_{DVM}^{i} \in \mathbb{R}^{1\times1}$ represents the compressed (fused) image modality, while $x_{TL}^{t} \in \mathbb{R}^{1\times1}$ denotes the suppressed text modality.

$$x_{f}^{c} = \mathbf{F_{c}}(X_{DVM}^{i}, X_{TL}^{t}) \tag{15}$$

$$x_{f_{sq}}^{c} = \mathbf{F_{c}}(x_{DVM}^{i}, x_{TL}^{t}) \tag{16}$$

In equation 15 and equation 16, the concatenation operation (as shown in equation 2) of inter-modalities takes place, producing the concatenated form of squeezed image-text features, i.e., $x_{f_{sq}}^{c} \in \mathbb{R}^{1\times2}$. Additionally, the base form of concatenated image-text features is obtained in a similar way, i.e., $x_{f}^{c} \in \mathbb{R}^{1\times1792}$.

However, the CEM block differs from the DVM block with the introduction of an LSTM block (shown in figure 5) through which the squeezed inter-fused feature vector is passed.

LSTM — LSTM were introduced to overcome the limitations of RNNs, particularly the well-known issue of vanishing gradients during backpropagation in RNNs. This issue arises when gradients diminish exponentially, hindering the network's ability to capture long-range dependencies in sequential data. To address this, LSTMs are designed with memory cells and gating mechanisms. LSTMs excel at learning long-range dependencies in sequential data, particularly in tasks like language modeling, speech recognition, and time-series analysis. However, LSTMs are also known to enhance performance when combined with different neural networks, even when the problem is not explicitly sequential [32]. The main components of an LSTM—the forget, input, and output gates—allow the network to regulate the flow of information, selectively retaining long-term dependencies while discarding irrelevant details. This ability to store and retrieve information selectively over extended sequences makes LSTMs highly effective for a variety of machine learning and deep learning tasks, which we leverage in the CEM block as well.

Although the presence of dual activation functions enhances the summarized representations of the fused features (see sub-section 3.1.2), some disadvantages exist, such as it may suppress valuable information containing important context, which could be crucial for correctly weighing the unsqueezed embeddings. As

a result, because both values of the squeezed feature comes from different modalities in the CEM block, we pass the squeezed features through an LSTM to leverage sequential dependencies within the fused squeezed features ($x^c_{f_{sq}}$). This is beneficial for capturing nuanced relationships in the multimodal embedding space, where specific patterns may emerge sequentially, enriching the inter-fusion modality, which builds upon intra-fusion. Additionally, passing squeezed embeddings through LSTMs allows dynamic recalibration of the learned features. As LSTMs update their hidden states, they can assign weights to features based on their importance, enhancing the relevance of key features from both modalities. We used an LSTM with 8 units, allowing the model to capture compact, yet significant, temporal relationships relevant to the size of the squeezed feature. In equation 17, the $\bar{x}^c_{f_{sq}}$ represents the temporal pattern of the squeezed feature learned iteratively during the training phase.

$$\bar{x}^c_{f_{sq}} = \mathcal{LSTM}(x^c_{f_{sq}}), \quad \text{where } \bar{x}^c_{f_{sq}} \in \mathbb{R}^{1\times 2\times 8} \tag{17}$$

Now, akin to DVM block we proceed with producing unsqueezed concatenation of the embeddings produced from TweetVal and DVM block as:

$$x^c_f = \mathbf{F_c}(X^i_{DVM}, X^t_{TL}) = X^i_{DVM} \oplus X^t_{TL}, \quad \text{where } x^c_f \in \mathbb{R}^{1\times 1792} \tag{18}$$

$$x^c_f \in \mathbb{R}^{1\times 1792} \rightarrow \mathcal{RESHAPE} \rightarrow x^c_f \in \mathbb{R}^{1\times 2\times 896} \tag{19}$$

$$X^c_{CEM} = \mathbf{F_w}(\bar{x}^c_{f_{sq}}, x^c_f), \quad \text{where } X^c_{CEM} \in \mathbb{R}^{1\times 1\times 896} \tag{20}$$

The output of equation 20, X^c_{CEM}, represents the cumulative intra- and cross-modal representation, which is then passed into the classification stage, marking the progression to stage III of `DAMM`.

3.1.4 Classification Stage

The final stage of the `DAMM` leverages the synergistic output from the previous stages. The classification component of `DAMM` is primarily supported by a Multihead attention mechanism followed by a fully-connected layer (see in figure 3). This is pivotal in translating the fused multimodal representations into task-specific outputs, effectively addressing the challenges posed by both binary and multiclass classification problems. Multihead attention mechanisms allow the modeling of complex inter dependencies among the fused image-text feature representations. By utilizing multiple attention heads, the model can focus on diverse aspects of the fused data simultaneously. Each head (consisting of q, k, and v, produced from input with learnable matrices W_q, W_k, and W_v) operates with a key dimension of 64, ensuring a rich and detailed capture of relationships. Subsequently, the contextually enriched information is passed through an LSTM layer with 256 units and then through a series of dense layers, ultimately narrowed down to the number of labels we aim to predict. The final output is wrapped with a softmax activation function, categorizing the memes into their respective classes.

Loss Function — Sparse Categorical Cross-Entropy (SCCE) is a loss function that addresses the memory inefficiency and computational complexity issues of Categorical Cross-Entropy (CCE) when dealing with large output spaces. While CCE is defined as $\text{CCE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$ and works well for one-hot encoded labels, it can be inefficient for tasks with multi classes. SCCE simplifies this by using integer labels directly, with the equation $\text{SCCE} = -\log(\hat{y}_i)$. We have implemented the SCCE with prior distribution to incorporate prior knowledge about the data distribution into the loss function. This method adjusts the predicted logits, $\hat{y}$, by adding a scaled log-prior distribution, where α is a tunable temperature parameter. The prior distribution is defined as $\text{prior} = [p_1, p_2, \ldots, p_C]$, where C is the number of classes. The log-prior is computed as $\log(p_i + \epsilon)$, where ϵ is a small constant to prevent $\log(0)$. The modified loss function is expressed as:

$$\mathcal{L}_{\text{PSCCE}} = \mathcal{L}_{\text{SCCE}}\left(\hat{y} + \alpha \log(\text{prior} + \epsilon)\right) \tag{21}$$

Here, $\hat{y}$ is the predicted logits (without softmax), α is the scaling factor, and prior represents the prior probabilities of each class. This approach helps in incorporating class distribution knowledge directly into the training process, improving model performance when prior knowledge is available.

4 Results and Discussion

In this section, we formally analyze the efficacy of DAMM in comparison to existing approaches on the same datasets using some standard evaluation metrics. The focus during evaluation is on approaches that emphasize modality fusions with respect to context elevation for accurate identification, to assess the impact of same-modality and different-modality fusion within a network.

4.1 Implementation Setup

For implementing DAMM, we used Google Colab and Kaggle, utilizing a T4 GPU with ∼13GB of RAM and ∼112GB of storage. We used a variable number of epochs depending on the dataset, with 30 epochs being the most frequent. During training, we saved the models based on the lowest validation loss and highest F1 score, using the ModelCheckpoint available in Keras. However, on the testing data, the trained models were evaluated using the models saved with the lowest validation loss. The evaluation metrics after every epoch were calculated using the scikit-learn library. All experiments were conducted with a batch size of 256 data samples. The Adam optimizer was used during training with a learning rate of 0.0001. The embedding extraction process was performed statically, indicating that embedding extraction and fusion were carried out in separate environments.

4.2 Evaluation Metrics

1. Accuracy (ACC):
Accuracy measures the proportion of correctly classified samples out of the total number of samples and provides an overall evaluation of the model's performance across all categories, calculated as $\text{Accuracy} = \frac{T_p+T_n}{T_p+T_n+F_p+F_n}$, where T_p represents True Positives, T_n represents True Negatives, F_p represents False Positives, and F_n represents False Negatives.

2. Precision (P):
Precision evaluates the proportion of samples predicted as belonging to a certain category that are actually correct, focusing on reducing false positives, and is given by $\text{Precision} = \frac{T_p}{T_p+F_p}$, where T_p and F_p are True Positives and False Positives, respectively.

3. F1-Score: The F1-Score, which is the harmonic mean of Precision and Recall, balances the trade-off between these two metrics and is defined as $\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$, where Precision is the proportion of correctly predicted positive cases to all predicted positive cases, and Recall is the proportion of correctly predicted positive cases to all actual positive cases.

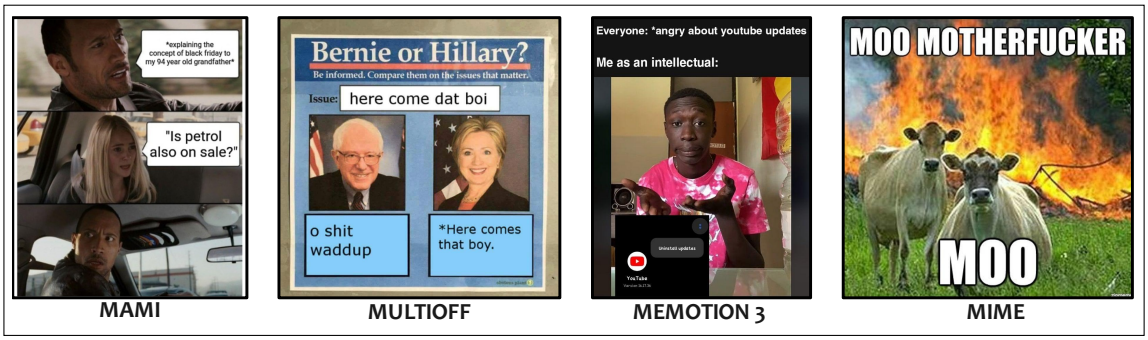
F1-Score(Macro) (F1-M) : Macro average computes the metric for each category individually and then takes an unweighted average, ensuring equal consideration for all categories, defined as $\text{Macro Average Metric} = \frac{1}{N}\sum_{i=1}^{N} \text{Metric for Meme Class } i$, where N is the total number of meme categories, and Metric for Meme Class i is the performance metric (e.g., Precision, Recall, F1-Score) for the i-th category.

F1-Score (Weighted) (F1-W): The weighted average calculates the metric by considering the proportion of samples in each category, thereby reflecting real-world distributions, and is given by $\text{Weighted Average Metric} = \sum_{i=1}^{N} w_i \cdot \text{Metric for Meme Class } i$, where $w_i = \frac{\text{Number of Instances in Meme Class } i}{\text{Total Instances}}$, N is the total number of categories, and Metric for Meme Class i is the performance metric for the i-th category.

4.3 Datasets

To thoroughly evaluate DAMM, we conduct experiments using four standard datasets, each designed to address different aspects of hateful content, including misogyny, offensiveness, abusiveness, and violence. Detailed descriptions of these datasets are provided below. Examples of hateful memes (see figure 6a) and non-hateful memes (see figure 6b) from all four datasets are illustrated in figure 6. It is important to note that the presence of derogatory language alone does not inherently classify a meme as hateful. A meme is categorized as hateful only if it specifically targets or mocks a particular group, gender, or race (refer to section 1). We

(a) Samples of Hateful memes



(b) Sample of Non-hateful memes

Figure 6: Sample of hateful and non-hateful memes for each dataset.

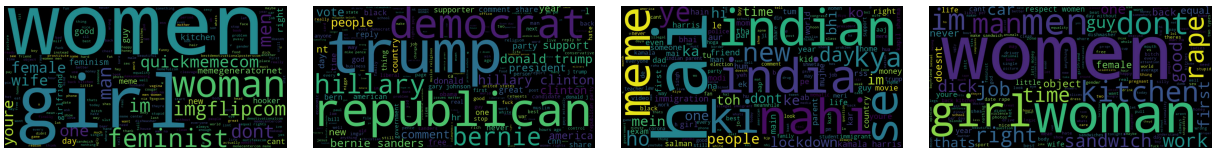
utilize a Word Cloud visualization, as shown in figure 7, to highlight the distribution of harmful words across the four datasets. This representation provides an intuitive and impactful way to identify frequently occurring harmful terms, with their dominance in the word cloud corresponding to their frequency in the datasets. A class-wise distribution of memes for four datasets is shown in figure 8. About 40% of the memes are neutral in the Memotion 3 dataset. An equal distribution class is observed on both MAMI and MIME datasets. MultiOFF constitutes a total of about 40% of offensive memes.

MAMI: This dataset originates from the SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) challenge [33], specifically Sub-task A, which focuses on binary classification of memes as misogynistic or non-misogynistic. Developed to address the growing prevalence of misogyny in online spaces, it consists of 15,000 memes sourced from platforms like Twitter, Reddit, 9GAG, and Imgur. Of these, 11,000 memes are annotated by human evaluators, ensuring a balanced distribution of misogynistic and non-misogynistic content for supervised learning tasks. The annotations are performed through a rigorous crowd-sourcing process, ensuring the reliability and relevance of the labels for detecting misogynistic content. This dataset provides a challenging benchmark for evaluating multimodal learning methods in identifying misogyny. We preserve the original data distribution, with 9,000 samples for training, 1,000 for validation, and 1,000 for testing. Following the approach of Grasso et al. [34], where they test the efficacy of their proposed approach on the validation dataset due to the unavailability of the test dataset, we also test and report the performance of `DAMM` in a similar data distribution. Table 1 summarizes the four datasets used in the study, including their respective splits for training, validation, and testing:

Table 1: Splitting of train, validation and test sets of the datasets used here

Dataset	Train	Validation	Test	Total
MAMI[33]	9000	1000	1000	11000
MultiOFF[35]	445	149	149	743
Memotion 3[36]	7000	1500	1500	10000
MIME[37]	512	128	160	800

MultiOFF: The MultiOFF dataset [35] is created to detect offensive content in memes by combining text and visual information, addressing the lack of multimodal datasets for such analysis. It is built using memes from the 2016 U.S. presidential election, sourced from platforms like Reddit, Facebook, Twitter, and Instagram. Preprocessing involves cleaning text captions, removing irrelevant metadata, and validating image URLs. The memes are annotated as offensive or non-offensive by a diverse group of annotators, following clear guidelines that consider personal attacks, homophobic or racial abuse, attacks on minorities, and sarcasm. The dataset includes 743 annotated memes, balanced for offensive and non-offensive categories, and split into training, validation, and test sets. From the dataset, 445 samples are used for training, whereas 149 samples are used for each validation and testing. MultiOFF is a valuable resource for studying the complex interplay of text and images in memes, supporting the development of multimodal machine learning models and addressing challenges in analyzing implicit offensive content.



(a) MAMI (b) MultiOFF (c) Memotion 3 (d) MIME

Figure 7: Word clouds representing the most frequent terms in four datasets: MAMI, MultiOFF, Memotion 3 and MIME

Memotion 3: The Memotion 3 dataset is a novel resource designed for sentiment and emotion analysis in memes, released as a shared task in AAAI '23 [36]. It focuses on multilingual memes, particularly Hindi-English (Hinglish) memes. The dataset consists of 10,000 annotated memes, each featuring a combination of visual and textual components, making it a multimodal dataset. The Google Vision API was used to extract text from memes, while the meme images were scraped from platforms such as Reddit and Google Images. Multiple tasks are given to solve using the dataset. Our model aims to solve Task A which is dedicated to an overall sentiment analysis of memes based on the categories: 'negative,' 'neutral,' and 'positive.' Other tasks in this dataset focus on multi-level classification of 'emotions', including 'motivational', 'sarcastic', and 'humorous' content. The dataset is divided into 7000 samples for training, whereas validation and testing contain 1000 samples each.

Misogynistic-MEME (MIME): The Misogynistic-MEME dataset [37] focuses on facilitating the detection of misogynistic contents in online memes, addressing the growing issue of cybersexism, which includes abusive remarks, body shaming, stereotypical comments, etc. It should be noted that in the literature, this dataset is referred to as Misogynistic-MEME, which we have referred to as MIME throughout this paper for simplicity. This multimodal collection comprises 800 memes, with an equal balance of 400 misogynistic and 400 non-misogynistic memes. These memes were collected from popular social media platforms, such as Facebook, Twitter, Instagram, and Reddit, as well as from websites dedicated to meme creation and collection. To

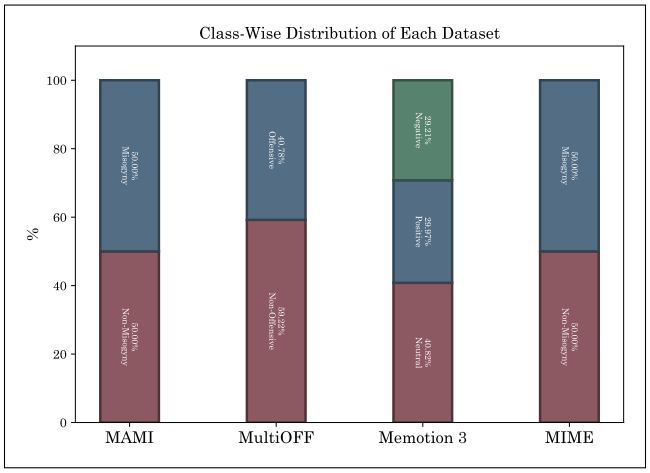


Figure 8: Class-wise distribution for dataset. **order maintain and remove dataset term from image**

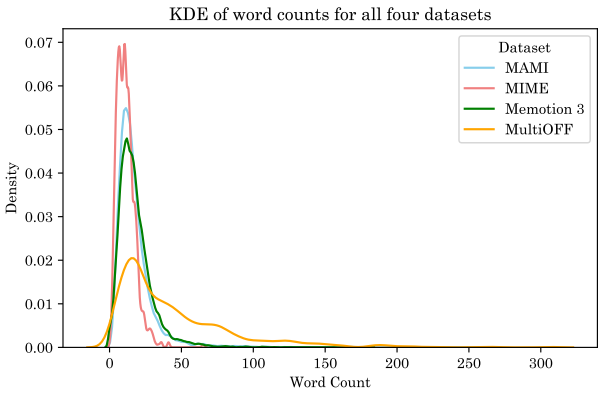


Figure 9: KDE plot of all four datasets showing the distribution of text lengths associated with memes.

ensure diverse and authentic representations, memes were curated using specific misogyny-related keywords, including themes like body shaming, stereotyping, objectification, and violence. 160 samples are kept for testing DAMM's performance while 512 samples and 128 samples are kept for training and validating DAMM. Each meme is annotated with several key attributes, including a unique identifier, the manually transcribed text from the image, and binary labels related to the presence of misogynistic content, aggressiveness, and irony, as assessed by both domain experts and crowd-sourced annotators.

Figure 9 presents a KDE (Kernel Density Estimation) plot for all four datasets. This visualization offers a comparison of textual characteristics across the datasets, providing insights into variations in text lengths and their overall distribution. of the extracted captions from meme images. The density plots show that most caption lengths average between 30 and 40 words, with a density ranging from 55% to 70% for the Memotion to MIME dataset. In contrast, for the MultiOFF dataset, a notable number of captions have lengths ranging from 50 to 100 words, which causes the curve to be more spread out compared to the other three datasets. It should be noted that slight discrepancies may exist between the actual number of words in the embedded captions on images and the actual extracted captions, depending on the OCR methods used for each dataset.

4.4 Significance of Modality Fusion

In the context of meme classification, for a meme to be classified as hateful or harmful, as discussed above (refer section 1) it is often the two-sided cooperation between the visual content and the accompanying text that completes the necessary semantic context for classification into a specific category. As shown in figure 1, the presence or absence of textual content can influence the perceived meaning of an image, leading to a misinterpretation that contradicts its actual classification. Textual features from images are typically extracted using Optical Character Recognition (OCR) solutions such as Google API, Tesseract, or Amazon Textract. These methods extracts texts compared to processing embedded text in meme images through visual encoders, which treat textual content as visual features. This approach often misses crucial linguistic patterns, resulting in features that lack the depth and precision necessary for accurate classification. However, relying solely on a single modality, either the image or the text can often lead to misclassification by failing to capture the full context or nuanced meaning of memes. This underscores the importance of integrating both modalities — or all available sources of information to achieve a more accurate and holistic understanding of the data. We conducted a brief analysis on an image from MAMI dataset, summarized in figure 10 to evaluate the effectiveness of the DAMM model in distinguishing between classes by emphasizing the combined features of both modalities, as opposed to classifying based on standalone modalities. For this analysis, we selected a misogynistic image sample from the MIME dataset, which contains a high degree of misogynistic content and is labeled in the misogyny class (labeled as 1). We start by analyzing the visual content of the meme by generating attention map via Grad-CAM[38], which highlights the visual features of the image, showing that it focuses on multiples facial entities.

The image modality attention is primarily driven by passing the image features into EfficientNet-B3, which we fine-tune, working within the DVM block to handle visual data. As shown in Figure 10, the image modality alone does not consider the textual caption in the prediction; instead, it processes the text as part of the visual content, which might offer some assistance in classification, but may not be always helpful

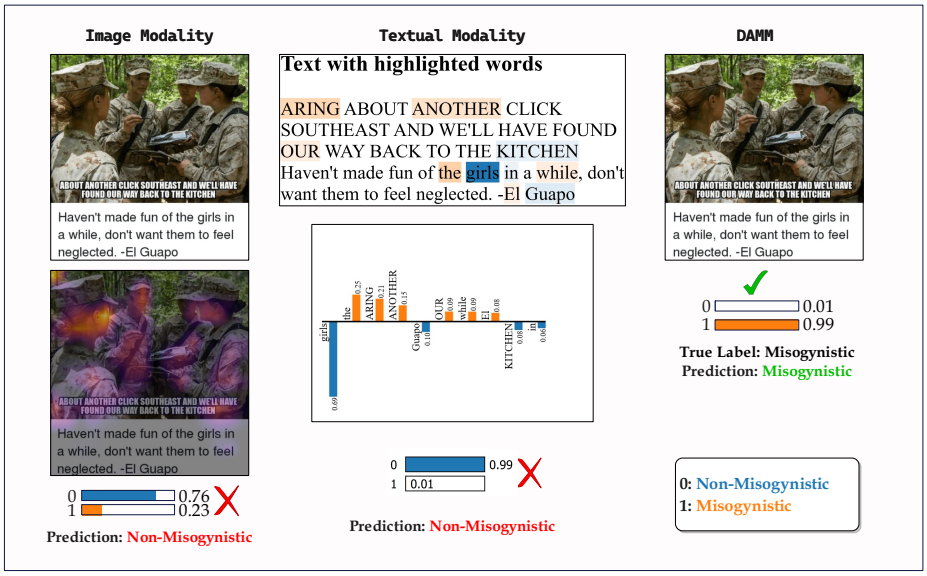


Figure 10: Significance of individual modalities and DAMM

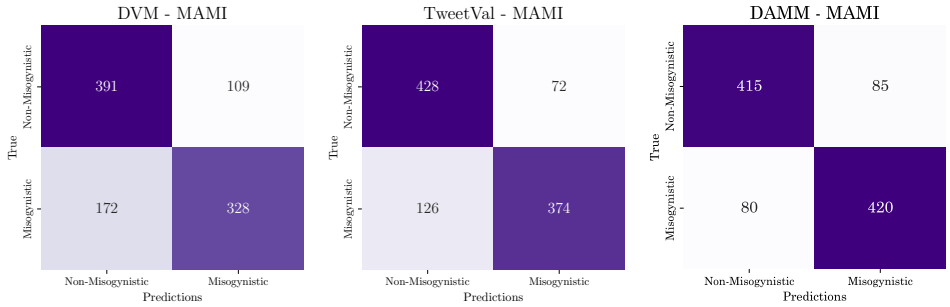


Figure 11: Comparison of confusion matrices shows DAMM's superior performance over DVM and TweetVal in capturing hateful content

in influencing the outcome. This leads to a misclassification in which the model assigns the image a high likelihood of being non-misogynistic, despite the presence of misogynistic content, due to the lack of full contextual understanding from the text and the absence of any hateful/negative or misogynistic signs that might persuade the visual classifier to place it in the label 1 (misogynistic) category.

Moving forward, we focus on studying the text content's role in understanding misogyny in the sample, utilizing LIME [39] (an Explainable AI (XAI)) framework that uses LIME values to determine feature importance in a given context). This is facilitated by passing the OCR output through TweetVal, which is primarily responsible for handling textual modality in the CEM block of DAMM. By applying LIME to the meme caption, we can visualize the focus on each token of the text via dual shades (orange and blue) representing non-misogynistic and misogynistic content, respectively. The opacity of the shades highlights the importance of each word in determining the final classification. In this case, the meme was incorrectly classified as non-misogynistic due to incomplete understanding or insufficient focus on critical features. For example, words like *"girls"* and *"kitchen"* influenced the classification towards the non-misogynistic label (0). However, words like *"fun"* and *"feel"* were overlooked in conjunction with *"kitchen,"* causing the model to subtly miss the misogynistic undertones of the entire content. The absence of critical features is effectively addressed by the DAMM block, as demonstrated in Figure 10 (in the right). In this case, the content is accurately classified as misogynistic with high confidence, leveraging the complementary structure across

both modalities. Some samples may perform well using standalone models such as DVM, TweetVal, or other encoder-based models. In certain cases, an enhanced architecture or standalone modality based model may outperform the DAMM. This can be attributed to the limited interpretability of such models. We observed that DAMM tends to deliver superior performance compared to standalone models. This is evident in Figure 11, where the confusion matrix for two standalone models (left – DVM, right – TweetVal) is shown alongside the DAMM model (right).

The correctly classified samples, represented by the diagonal alignment of respective correct prediction classes, are strongly highlighted, underscoring the robust performance of DAMM in identifying misogynistic content. DAMM outperforms in finding misogynistic memes more correctly when compared to the value findings from individual modalities, images and text, respectively.

4.5 Analysis of Results

In this section, we present a comprehensive evaluation of DAMM's performance across the selected datasets. The evaluation is further categorized based on its efficacy in classifying hatefulness, which encompasses specific subcategories such as misogyny, politically offensive content, and generalized hate speech (not associated with a particular category). A detailed summary of DAMM's performance metrics—including accuracy, precision, and F1 scores is provided in Table 2.

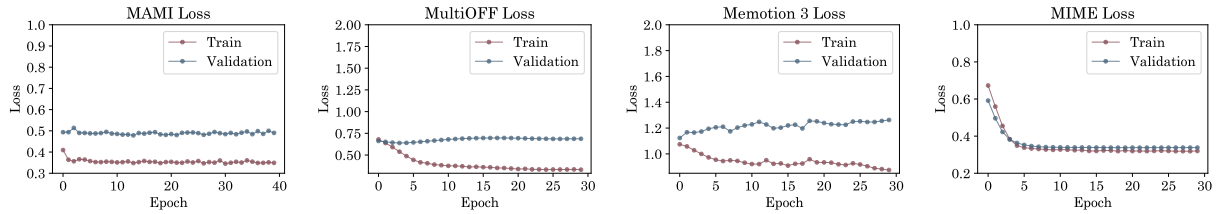
Table 2: Performance of DAMM across datasets.

Dataset	Accuracy	Precision	F1
MAMI[33]	0.835	0.835	0.835
MultiOFF[35]	0.664	0.664	0.663
Memotion 3[36]	0.367	0.364	0.363
MIME[37]	0.925	0.925	0.925

DAMM *on Misogyny Detection.* The earliest dataset in our comparison is the MAMI dataset, which is a part of SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification [33]. In this task, the Macro-F1 score is utilized as the official evaluation metric for assessing the performance of models. DAMM leverages the synergistic pattern contextualization of memes for MAMI misogynistic meme samples, resulting in significantly improved outcomes compared to most existing methods, achieving an F1 score of 0.835, accuracy of 0.8349, and precision of 0.8350. Regarding the loss, as illustrated in Figure 12a, the loss visualization for the MAMI dataset demonstrates a relatively consistent yet narrow gap between training and validation losses across 30 epochs. Despite minor fluctuations in validation loss, it remains stable, indicating that the model generalizes effectively to unseen data with minimal overfitting and is stabilizing well.
Through the ModelCheckpoint mechanism, we saved the best-performing model, the DAMM framework achieved superior empirical results during the testing phase of the experimentation. As shown in Figure 13a, the model correctly classified 415 non-offensive instances (class 0) and 420 offensive instances (class 1). However, it misclassifies 85 non-offensive instances as offensive and 80 offensive instances as non-offensive. Over the course of training across 30 epochs, the validation F1 score reached its maximum at the 14th epoch, attaining a peak value of 0.8350. This represents an improvement of 0.187 from the F1 score of 0.8197 observed at the first epoch. This performance surpasses that of several proposed fine-tuned and pre-trained architectures, highlighting the zero-shot capabilities of DAMM. Furthermore, it underscores the effectiveness of the cross-dual-level fusion approach used in our approach. Table 3 summarizes all the model performances on the MAMI Dataset.

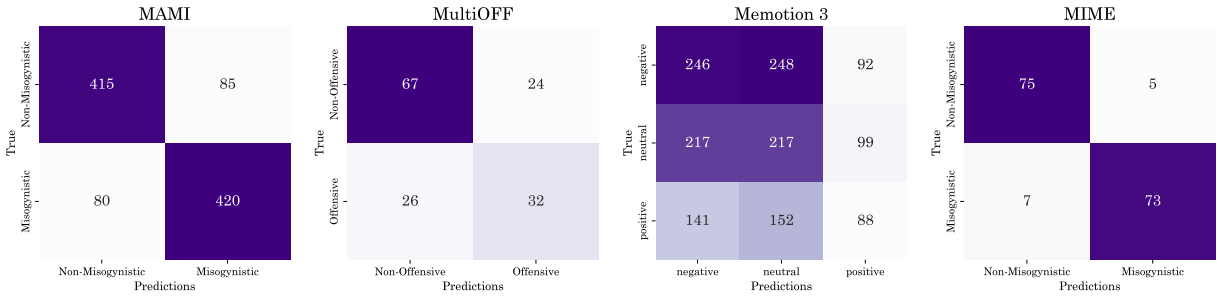
DAMM outperforms the method proposed by Hakimov et al. [40] by a margin of 0.101, which utilized a standard CLIP encoder for image and text processing in their multimodal study of misogynistic memes, achieving an F1 score of 0.734 on the same dataset. Cao et al. [41] proposed a modularize network architecture for hateful meme detection, utilizing LoRA modules fine-tuned from large language models (LLMs) alongside a module composer to enhance task-specific reasoning in low-resource scenarios. Their approach achieved a maximum accuracy of 0.611 across all experiments, where as, DAMM demonstrates a better accuracy by an increase of 0.224, reinforcing efficacy of our method. This highlights that, even without the presence of reasoning-based modules within deep neural networks and advanced tuning techniques, the choice and implementation of fusion strategies play a pivotal role in accurately identifying misogynistic cues in memes.

A recently proposed MISTRA framework [53], which we surpasses by a significant increase of 0.10 in F1 score, compared to MISTRA's 0.735. MISTRA utilizes a multimodal approach that utilizes CLIP and DistilBERT



(a) MAMI (b) MultiOFF (c) Memotion 3 (d) MIME

Figure 12: Training and Validation loss curves for four datasets: MAMI, MultiOFF, Memotion 3 and MIME **Figure c , image title should be Memotion 3**



(a) MAMI (b) MultiOFF (c) Memotion 3 (d) MIME

Figure 13: Confusion matrices illustrating DAMM performance across four datasets: MAMI, MultiOFF, Memotion 3, MIME.

for image and text feature representation, respectively, combined with a Variational Autoencoder to reduce high dimensionality into a latent space, resulting in a modest 1.5%. improvement. However, this approach does not surpass the effectiveness of inherently capturing the same modality in diverse ways, as demonstrated by DAMM, which is comparatively a simpler model architecture while yielding superior results. Additionally, MISTRA utilizes dual fusion of text modalities, with text representations dynamically extracted from meme images via BLIP and parallely by DistilBERT. These are then fused with image embeddings at the same layer using triple fusion. In contrast, our method of weighted embedding fusion provides the model with a more nuanced understanding of the relative contributions of each modality. This enables DAMM to better capture modality-specific features and achieve enhanced performance.

Another substantiation of the effectiveness of integrating weighted embedding fusion during concatenation for intra-modality and inter-modality fusion is demonstrated by a recent approach, [52], that employs multiple vision-transformer (Swin, ConvNeXt, and ViT) and text-transformer based models (BERT, ALBERT, and XLM-R) for detecting misogynistic memes. This approach performs standard early concatenation of embeddings before passing them through a single perceptron layer, achieving an F1 score of 0.728, showing 0.107 lower than that achieved by DAMM. These findings emphasize the pivotal role of advanced fusion techniques in effectively capturing the contextual interplay between textual and visual modalities, particularly in the heterogeneous and complex representations inherent in meme images.

Evaluating* DAMM *on Memotion 3 For the Memotion 3 dataset, we focus solely on Task A, which involves understanding the overall sentiment of memes, without delving into specific aspects or degree of hate (such as slight misogyny or high politically offensive content). We achieve the highest weighted F1 score of 0.363, surpassing the top scorer of the Memotion 3 challenge, NUAA-QMUL-AIIT[28], which has a F1 score of 0.344 by 0.019 on the test dataset. Table 4 summarizes all results on the Task A, DAMM outperforms NYCU_TWO [54] by 0.021, CUFE by 0.026, CSECU-DSG by 0.03, and the Baseline [55] established by the challenge organizers by 0.031. This demonstrates the impact of weighted embedding fusion in tackling multimodal hate speech, in an entwined manner, covering both image-wise and image-text fusion.

In the same challenge, wentaorub[56], the runner-up of the challenge, used an early concatenation of embeddings obtained from a fine-tuned CLIP model. This approach achieved a test F1-score of 0.3288, using CLIP independently for image and text embedding extraction and fusing them under a Multihead attention

Table 3: Summary of the performances of different approaches proposed on the MAMI dataset. DAMM scores are highlighted in gray, and the second-highest score is italicized.

Approach	Year	Model	Accuracy	Precision	F1-Score (M)
Oscar-Large [42, 43]	2022	Multimodal Pre-training	0.696	-	0.689
Uniter-Large [44, 43]	2022	Multimodal Pre-training	0.692	-	0.684
VisualBERT-Large [45, 43]	2022	Multimodal Pre-training	0.692	-	0.68
ERNIER-Vil-Large [46, 43]	2022	Multimodal Pre-training	0.715		0.707
DD-TIG [43]	2022	ERNIER-Vil-large + Word Masking + Image Captioning	0.794	-	0.793
SRCB [22]	2022	$CLIP_{Image}$ + XGBoost	-	-	0.776
RIT Boston [47]	2022	$CLIP_{Image+Text}$ + Semi-supervised Learning	-	-	0.778
ASRtrans [48]	2022	MMBT + VisualBERT	-	-	0.761
Poirot [49]	2022	ResNet + Sentence-BERT + Graph NN			0.759
AMS_ADAN [50]	2022	ResNet-18 + BERT	-	-	0.746
MISTRA [51]	2024	DistilBERT + CLIP + VAE + BLIP	-	0.773	0.715
VisualBERT COCO [45, 34]	2024	Multimodal Pre-training	-	-	0.742
KERMIT [34]	2024	Knowledge Graphs + ConceptNet + ConcatBERT	-	-	*0.834*
V-LTCS [52]	2024	BERT + ViT	0.666	0.614	0.728
DAMM (Ours)	2025	EB3 + CLIP + TweetVal	0.834	0.835	**0.835**

framework. This highlights that even when using robust vision-language models like CLIP and training them on in-domain specific datasets, our approach outperforms it by 0.0341 through the efficient framework utilizing a frozen CLIP encoder based on the effective fusion mechanism used.

Throughout the training, we observe spikes in loss, with both increases and decreases, while training the model (see Figure 12c). On this dataset, we also surpass Ignacio et al. [57]'s approach by 0.0148 on F1 metric, where they use a two-stage method utilizing large language models and U-Net Encapsulated Transformers (UET). The first stage helped in generating BLIP-2 captions, which in our case is done directly by the DVM block. In the second stage, they used GPT-4 integration and KeyBERT for key-phrase extraction to understand the essential textual content. In contrast, we achieved this with the CEM block, making our approach effective. This eliminates the need for large language models for identifying hate, making it easier to use DAMM. We emphasize on generalizing capabilities of DAMM in detecting negative hate speech to capture and identify negative content on the Internet and social media, which can then be used to narrowly classified based on external domain-specific tuned model.

Table 4: Summary of the Weighted F1 scores of different approaches proposed on the Memotion 3 dataset. DAMM scores are highlighted in gray, and second highest score is italicized.

Model	Year	Approach	F1-Score
NYCU-TWO [54]	2023	SwinTransformer + CLIP	0.342
CUFE [55]	2023	Hinglish-DistilBERT + ResNet18	0.338
Baseline [55]	2023	Hinglish-BERT + ViT	0.339
wentaorub [56]	2023	$CLIP_{Image}$ + $CLIP_{Text}$ + OSCAR	0.329
NUAA-QMUL-AIIT [28]	2023	SEFusion: RoBERTa + CLIP-ViT	0.344
CSECU-DSG [55]	2023	—	0.333
Doc HMT E3 [57]	2024	GPT4 + BLIP-2 + KeyBERT + U-Net	*0.349*
Doc HMT E10 [57]	2024	GPT4 + BLIP-2 + KeyBERT + U-Net	0.324
DAMM (Ours)	2025	EB3 + CLIP + TweetVal	**0.363**

DAMM on Identifying Politically Offensive Content. For MultiOFF, the baseline is based on the official experiments from the dataset release [35], which focuses on politically offensive meme identification, using the weighted F1 score as the official metric and also serves as the baseline for this dataset. The traditional deep

learning approach uses to handle modalities independently including variants of LSTM (stacked, biLSTM) and a deep CNN network, VGG16. Their experiments achieve the highest F1 score of 0.54 when using CNN for the textual modality, and 0.50 in a multimodal setting [35]. In contrast, our approach surpassed this baseline by achieving a score of 0.663. This represents a 0.163 absolute improvement over the multimodal approach of the baseline. As can be seen in Table 5, `DAMM` consistently outperforms all other multimodal and unimodal approaches from the baseline experiments, including Logistic Regression, Naive Bayes, DNN, and Stacked LSTM [58]. As shown in Figure 13b, for MultiOFF, `DAMM` correctly classified 67 non-offensive instances (represented by class 0) and 32 offensive instances (represented by class 1). However, it misclassified 24 non-offensive instances as offensive and 26 offensive instances as non-offensive. This highlights that while `DAMM` achieves comparable results to the latest approaches, significant improvements are still needed in handling class imbalances and borderline cases. Over the course of 30 epochs in Figure 12b), the training loss decreased from 0.6796 to 0.3311, while the validation loss showed a more modest decrease suggesting potential marginal overfitting, indicating that the model effectively learned from the training data.

Grasso et. al. proposed KERMIT [34], a framework that utilizes external knowledge to improve the classification of harmful memes, and uses ConcatBERT which internally uses BERT and ResNet-152 to process text and visual information in multimodal setting. KERMIT builds a knowledge-enriched information network with relevant external knowledge sourced from ConceptNet. The reliance on external knowledge gives KERMIT an advantage over `DAMM` in handling edge cases, as memes are increasingly dynamic and obscure. However, `DAMM` outperforms KERMIT on the MultiOFF dataset with a margin of approximately 0.012, as KERMIT achieved an F1 score of 0.651 on the test dataset. With respect to KERMIT, `DAMM`, without any external connections, benefits from a more comprehensive modality perspective by leveraging pretrained data with efficient, learnable blocks (CEM, DVM) that capture deep contextual knowledge.

We also outperform the novel disentanglement-based framework, DisMultiHate, proposed by Lee et al. [59], which focuses on simultaneously extracting relevant entities from both image and text modalities, processing them through respective visual and textual encoders, and translating them into a shared latent space. Our approach achieves a 0.203 higher from their 0.646 F1 score. This improvement underscores the efficacy of integrating modality information holistically into the model, rather than isolating relevancy (disentanglement of entities) as the primary focus. Although prioritizing relevance might appear intuitively beneficial, models that leverage a comprehensively fused representation of contextual information from both modalities demonstrates superior capability in achieving precise classification as evident by `DAMM`.

To evaluate the generalizability and robustness of our approach, we tested it on a smaller and less extensively explored dataset, Misogynistic-MEME (or MIME) [37], which has not been a primary focus in prior research on misogyny detection in social media. As a baseline, we referred to the authors' prior work on benchmarking the MIME dataset for sexism detection [60]. Their experiments separately evaluated unimodal performance, achieving an F1 score of 0.757 using the textual modality. For multimodal analysis, their early fusion approach, employing a decision tree classifier, resulted in a lower F1 score of 0.693, which they attributed to the dominance of visual features over textual ones. This outcome highlights the necessity for optimized feature representation techniques when utilizing early fusion, such as weighted feature extraction or enhanced processing strategies. In comparison our early fusion approach, `DAMM`, leverages dual-channel modality-specific feature extraction, achieved a significantly higher F1 score of 0.9250. Furthermore, while their late fusion method achieved 0.758, which `DAMM` outperforms it by 0.197.

We acknowledge that the dataset used for this experiment is relatively smaller compared to other experimental setups. To assess overfitting, as observed in Figure 12d, both the training and validation losses decreased significantly during the initial four epochs. After this point, the losses plateaued for both training and validation, showing only a marginal decline thereafter. This indicates that the model reached its optimal performance. With the least loss recorded and used to save the model `DAMM` performance results in high number of true positives (75) and true negatives (73) (refer Figure 13d), indicating that the model accurately classified both misogynistic and non-misogynistic memes. The relatively low false positives (7) and false negatives (5) suggest that the model's misclassifications were minimal, with only a small fraction of non-offensive memes being mistakenly labeled as offensive and vice versa. This balanced distribution highlights the model's precision in distinguishing between the two classes, underscoring its efficiency in the task of misogynistic meme classification.

4.6 Error Analysis

For the purpose of analyzing the factors causing errors in the images, we performed error analysis, which `DAMM` faced in meme classification when faced with specific types of meme images lacking critical properties

Table 5: Summary of the performances of different approaches proposed on the MultiOff dataset. DAMM scores are highlighted in gray, and second highest score is italicized.

Approach	Year	Model	Precision	F1-Score (W)
Suryawanshi et al. [35]	2020	CNN_{Text}	0.390	0.540
Suryawanshi et al. [35]	2020	VGG16	0.410	0.240
Suryawanshi et al. [35]	2020	Stacked LSTM + VGG16	0.400	0.500
Suryawanshi et al. [35]	2020	BiLSTM + VGG16	0.400	0.410
Suryawanshi et al. [35]	2020	CNN_{Text} + VGG16	0.380	0.480
DisMultiHate [59]	2021	BERT + Faster R-CNN w/ MultiHeadAttention	0.645	0.646
DisMultiHate $_{\text{w/o Disentangle}}$ [59]	2021	BERT + Faster R-CNN w/ MultiHeadAttention	0.614	0.608
MHA-Meme [61]	2021	LSTM + VGG-19 w/ MultiHopAttention	-	0.591
MemeFier [62]	2023	$CLIP_{Image}$ + $CLIP_{Text}$	-	0.625
PromptHate [63, 34]	2023	VilBERT + RoBERTa	-	0.420
Bates et al. [64]	2023	CLIP w/ Meme Templating	-	0.619
Grasso et al. [34]	2024	CLIP Multimodal Pretraining	-	0.617
Grasso et al. [34]	2024	OSCAR Multimodal Pretraining	-	0.606
Ernie-Vil [34, 46]	2024	Cross-modal + Two-stream Transformers + Object Detection	0.540	0.531
KERMIT [34]	2024	Knowledge Graphs + ConceptNet + ConcatBERT	-	*0.651*
DAMM (Ours)	2025	EB3 + CLIP + TweetVal	0.650	**0.663**

Table 6: Summary of the performances of different approaches proposed on the MIME dataset. DAMM scores are highlighted in gray, and the second-highest score is italicized.

Approach	Year	Model	Accuracy	Precision	F1-Score (M)
Fersini et al. [65]	2019	Late Multimodal Fusion	-	-	0.758
V-LTCS [52]	2024	ALBERT + ViT	0.779	0.792	0.777
V-LTCS [52]	2024	BERT + ViT	0.792	0.797	*0.792*
DAMM (Ours)	2025	EB3 + CLIP + TweetVal	0.925	0.925	**0.925**

important for correct classification. Often, images containing large areas of blank content, visuals with filler plain colors or irrelevant patterns, low-resolution content, and multilingual content are a few of the features responsible for errors. Blank images with text overlays relying entirely on the textual component to convey meaning often use idiomatic expressions, sarcasm, or cultural references. A subset of sample images from our chosen datasets is visualized in Figure 14 with the possible reasons for misclassification.

***Low Resolution Meme Images*:** Misclassification happens when the model struggles without visual cues to support the text, the model also misinterpret the context or tone in meme having hateful content. Low-resolution images, as shown in Figure 14(b), exemplify how they complicate classification by degrading the clarity of both visual and textual information. Pixelation and compressed visual details make it harder for the model to identify fine-grained features such as object edges or text readability. For instance, a meme intended to convey humour through subtle expressions or wordplay may be misclassified if the image quality hampers the accurate recognition of these elements. Similarly, text in low-resolution memes can become distorted or illegible, causing OCR systems to extract incomplete or inaccurate information, as shown in Figure 14(b).

***Multilinguality in Meme Captions*:** The inclusion of multilingual text in memes introduces an additional layer of complexity, as memes often blend languages or incorporate regional scripts alongside stylized fonts. For example, a meme featuring text in both English and a regional language like Telugu can confuse a monolingual encoder model, which lacks robust multilingual capabilities. In such cases, OCR may fail to recognize non-English text or mix up language-specific semantics, leading to misclassification. Additionally, the limited interpretability of multilingual features by monolingual text encoder models may prevent the generation of rich embeddings for accurate classification. Furthermore, certain words or phrases can carry

Table 7: Results of the ablation study for Memotion 3 and MAMI datasets.

Ablation	Variation	Memotion 3		MAMI	
		Accuracy	F1-Score (W)	Accuracy	F1-Score (W)
Ablation 1	Added dual LSTM network in the CEM block	0.355	0.351	0.833	0.833
Ablation 2	Disabled return sequence of LSTM layer in CEM Block	0.330	0.329	0.825	0.825
Ablation 3	Discarded X^t_{TL} from $x^c_{f_{sq}}$, i.e., removed text embeddings in the squeezed feature concatenation in CEM	0.353	0.348	0.828	0.829
Ablation 4	Discarded X^t_{TL} from x^c_f, i.e., removed text embeddings in the unsqueezed feature concatenation in CEM	0.330	0.326	0.838	0.838
Ablation 5	Discarded $X^i_{f_{DVM}}$ from $x^c_{f_{sq}}$, i.e., removed squeezed fused visual embeddings when producing cross-modal weighted embedding in CEM	0.324	0.317	0.833	0.833
Ablation 6	Discarded $X^i_{f_{DVM}}$ from x^c_f, i.e., removed unsqueezed fused visual embeddings when producing the cross-modal weighted embedding in CEM	0.359	0.354	0.815	0.817
Ablation 7	**DAMM (Ours)**	0.367	0.363	0.835	0.835

different cultural meanings depending on the language which may be wrongly interpreting the contextuality of the meme, further complicating the interpretation process.

***Cluttered Visual Entities in Memes*:** Some images in the dataset contain cluttered visuals, presenting an additional challenge. These images often feature dense visual elements combined with overlapping or poorly aligned text, resulting in disturbance for the correct focus on key elements' association with the textual component responsible for correct classification. When multiple characters, objects, or complex backgrounds compete with the text for attention, the model may misidentify key features, leading to errors in interpreting the meme's intent.

***Excessive Text on Meme Image*:** DAMM misclassifies memes with excessive text, as shown in Figure 14(d), by overwhelming the system with multiple, often disjointed textual elements, resulting in superficial embeddings. The model finds it more difficult to pinpoint the exact hateful content. DAMM struggles to determine which parts of the text are relevant for classification. It has been observed that meme images with excessive text often use varied fonts, sizes, and alignments, further complicating OCR (Optical Character Recognition) tasks. This variability makes it difficult for the model to identify harmful content, leading to misclassification.

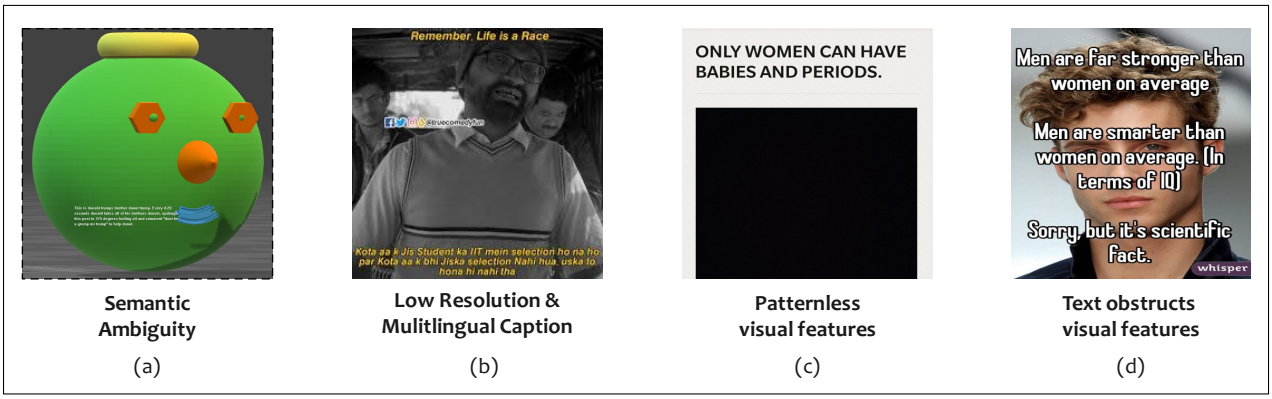


Figure 14: Misclassified meme samples, with potential reasons for misclassification mentioned in texts

4.7 Ablative Experiments

Since DAMM consists of multiple blocks superimposed in a nested manner, we performed an ablation study to assess the effectiveness of each block. Table 7 summarizes the F1 and accuracy scores obtained from

permutations of different configurations of blocks within the DAMM structure on the Memotion 3 dataset for multi-class classification and MAMI dataset for binary classification task. The ablation experiments provide critical insights into the contribution of individual components and mechanisms to overall performance. By systematically excluding or modifying specific features of DAMM, the results highlight their influence on test accuracy (ACC) and weighted F1-score (F1-W), which are the primary focus.

LSTM impact in CEM Block. Next, we studied the impact of the LSTM network in the CEM block (refer sub-subsection 3.1.3), aiming to understand whether it aids in the sequential learning of interdependent features from the DVM block (refer sub-subsection 3.1.2) representing the visual features. To investigate this, we extended the network by adding an additional LSTM block, resulting in a total of two stacked LSTMs in the CEM block. Replacing a single LSTM with two stacked LSTMs slightly diminished the test accuracy to 0.3553, with a similar drop in the F1 score to 0.3511, indicating a performance degradation of around 0.0103 compared to DAMM. A similar trend is visible, with a drop of 0.0021 in the F1 score of the MAMI dataset. This suggests that while a deeper network may capture additional temporal patterns in the feature space, it also introduces the risk of overfitting, as evidenced by the modest degradation in test scores. Therefore, careful tuning and experimentation are required to avoid redundancy or instability introduced by LSTMs in enhancing cross-modality representation.

Skipping Textual Modality. For stage two of DAMM, we aimed to understand the impact on performance when excluding the textual embeddings in both unsqueezed and squeezed concatenations, to determine whether performance significantly changes. For the first experiment, in the Memotion 3 dataset, discarding the unsqueezed concatenation (X^t_{TL} from x_{c_f}) reduced the test accuracy to 0.3534 (a drop of 0.0139) and the F1 score to 0.3481 (a drop of 0.0153). For the same configuration, DAMM observed a drop of 0.0065 in accuracy and 0.0060 in the F1 score, respectively in the MAMI dataset. In the second experiment with squeezed text feature skip, that is, when excluding text-average pooling (x^t_{TL} from $x^c_{f_{sq}}$), it resulted in a more moderate decline, with accuracy remaining the same at 0.3672, while F1 dipped to 0.3536. For MAMI, the elevation in F1 shifted to 0.8380 , which can be attributed to intrinsic dataset features, underlining the expendability of textual features in this case. These results show that textual features often provide explicit semantic cues, such as offensive language or specific terms, which are critical for meme classification. Their exclusion limits the model's ability to capture such explicit patterns, leaving classification reliant solely on visual features. This highlights the importance of modality fusion to capture the complementary interplay between embeddings.

Skipping Visual Modality. We then conducted the same experiment, focusing on assessing the importance of including the visual modality in both squeezed and unsqueezed versions. When the unsqueezed visual features were skipped in the CEM block (i.e., removing $X^i_{f_{DVM}}$ from x^c_f), the test accuracy dropped to 0.3594 and F1 to 0.3539 in the Memotion 3 dataset. This shows that image representation is crucial, offering a richer representation by aggregating embeddings from different visual feature extractors (e.g., EfficientNetB3 and CLIP). Without these features, the model loses its ability to discern detailed visual patterns, such as tone and subtle symbolism, resulting in performance degradation. This highlights the importance of feature-level complementarity in visual modalities for enhancing meme classification. Similarly, when we discard the squeezed combination of image features (i.e., excluding $x^i_{f_{DVM}}$ from $x^c_{f_{sq}}$), the test accuracy dropped to 0.3379 and F1-W to 0.3391, emphasizing the significance of suppressed image features. This also affects the output of the CEM block, as the squeezed image feature is used to weight the unsqueezed fusion. The remaining ablation experiments are conducted to understand the individual importance of visual features from EfficientNetB3 and CLIP, as well as the impact of LSTM in the classification stage, all of which are summarized in Table 7.

4.8 Conclusion and Future Work

In this work, we explored the efficacy of combining dual embedding channels from a single modality with cross-modality integration to effectively capture the nuances of hate meme categorization utilizing an early fusion . This is critical in the data- and analytics-driven world we are moving towards, aiming to create an inclusive digital environment for the future. We propose DAMM, a three-stage framework designed using squeeze-and-excitation techniques to generate weighted embeddings for both intra- and inter-modality through an attention mechanism. Our approach demonstrates superior performance across four hate meme datasets of varying genres. Our observation reveals that text plays a crucial role in shaping the overall perception of a meme. However, we believe that substantial further work is needed, and we outline several future directions based on this preliminary approach involving dual-stream embeddings. First, we aim to extend this methodology to multi-label settings, where final labels incorporate varying degrees of hate speech, such as

misogyny or offensive categories, allowing fine-grained hatespeech classification. Building on the dual visual feature extraction employed by `DAMM`, we also intend to focus on a multilingual text extractor for enhancing meme's embedded text feature representation, encompassing diverse dialects, tones, and languages. This could improve performance in varied cultural contexts. In the future, we plan to extend the evaluation of `DAMM` on more diversified and varied datasets of different multi-modalities to better assess its performance across different use cases. A potential approach in future may involves separating the modalities by removing textual content from images during reprocessing. This allows the visual modality to focus on non-textual features while the textual modality processes extracted text independently, potentially improving feature specialization, reduced noise, and improve overall performance.

References

[1] Raphael Cohen-Almagor. Freedom of expression v. social responsibility: Holocaust denial in canada. *Journal of Mass Media Ethics*, 28(1):42–56, 2013.

[2] Richard Delgado and Jean Stefancic. Images of the outsider in american law and culture: Can free expression remedy systemic social ills. *Cornell L. Rev.*, 77:1258, 1991.

[3] Laura Beth Nielsen. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social issues*, 58(2):265–280, 2002.

[4] Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Lee. Recent advances in online hate speech moderation: Multimodality and the role of large models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419, 2024.

[5] Fan Wu, Guolian Chen, Junkuo Cao, Yuhan Yan, and Zhongneng Li. Multimodal hateful meme classification based on transfer learning and a cross-mask mechanism. *Electronics*, 13(14):2780, 2024.

[6] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36, 2024.

[7] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021.

[8] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[9] Mehmet Aygün, Yusuf Hüseyin Şahin, and Gözde Ünal. Multi modal convolutional neural networks for brain tumor segmentation. *arXiv preprint arXiv:1809.06191*, 2018.

[10] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

[11] Nicolas Benjamin Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013. Association for Computational Linguistics, 2023.

[12] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. " you know what to do" proactive detection of youtube videos targeted by coordinated hate attacks. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21, 2019.

[13] Rui Cao and Roy Ka-Wei Lee. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, 2020.

[14] Hao Chen, Susan McKeever, and Sarah Jane Delany. Abusive text detection using neural networks. 2017.

[15] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, 2018.

[16] Rohit Pawar, Yash Agrawal, Akshay Joshi, Ranadheer Gorrepati, and Rajeev R Raje. Cyberbullying detection system with multiple server configurations. In *2018 IEEE International Conference on Electro/Information Technology (EIT)*, pages 0090–0095. IEEE, 2018.

[17] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17, 2011.

[18] Faseela Abdullakutty and Usman Naseem. Decoding memes: A comprehensive analysis of late and early fusion models for explainable meme analysis. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1681–1689, 2024.

[19] Thanh Tin Nguyen, Nhat Truong Pham, Ngoc Duy Nguyen, Hai Nguyen, Long H Nguyen, and Yong-Guk Kim. Hcilab at memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities (short paper). In *DE-FACTIFY@ AAAI*, 2022.

[20] Yang Deng, Yonghong Li, Sidong Xian, Laquan Li, and Haiyang Qiu. Mual: Enhancing multimodal sentiment analysis with cross-modal attention and difference loss. *International Journal of Multimedia Information Retrieval*, 13(3):31, 2024.

[21] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Amanullah Yasin, Gautam Srivastava, Dawid Połap, Thippa Reddy Gadekallu, and Zunera Jalil. Multimodal religiously hateful social media memes classification based on textual and image data. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(8):1–19, 2024.

[22] Jing Zhang and Yujin Wang. Srcb at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, 2022.

[23] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[24] Pradeep Kumar Roy. Mmffhs: Multi-modal feature fusion for hate speech detection on social media. *IEEE Transactions on Big Data*, 2024.

[25] Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 2024.

[26] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE, 2021.

[27] Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18, 2019.

[28] Xiaoyu Guo, Jing Ma, and Arkaitz Zubiaga. Nuaa-qmul-aiit at memotion 3: Multi-modal fusion with squeeze-and-excitation for internet meme emotion analysis. *arXiv preprint arXiv:2302.08326*, 2023.

[29] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

[32] Amir Farzad, Hoda Mashayekhi, and Hamid Hassanpour. A comparative performance analysis of different activation functions in lstm networks for classification. *Neural Computing and Applications*, 31:2507–2521, 2019.

[33] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, 2022.

[34] Biagio Grasso, Valerio La Gatta, Vincenzo Moscato, and Giancarlo Sperlì. Kermit: Knowledge-empowered model in harmful meme detection. *Information Fusion*, 106:102269, 2024.

[35] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41, 2020.

[36] Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes. *arXiv preprint arXiv:2303.09892*, 2023.

[37] Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in brief*, 44:108526, 2022.

[38] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[40] Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. *arXiv preprint arXiv:2204.06299*, 2022.

[41] Rui Cao, Roy Ka-Wei Lee, and Jing Jiang. Modularized networks for few-shot hateful meme detection. In *Proceedings of the ACM on Web Conference 2024*, pages 4575–4584, 2024.

[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[43] Ziming Zhou, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. Dd-tig at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 563–570, 2022.

[44] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[45] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[46] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3208–3216, 2021.

[47] Lei Chen and Hou Wei Chou. Rit boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from clip model and data-centric ai principle. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 636–641, 2022.

[48] Ailneni Rakshitha Rao and Arjun Rao. Asrtrans at semeval-2022 task 5: Transformer-based models for meme classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 597–604, 2022.

[49] Harshvardhan Srivastava. Poirot at semeval-2022 task 5: Leveraging graph network for misogynistic meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 793–801, 2022.

[50] Da Li, Ming Yi, and Yukai He. Ams_adrn at semeval-2022 task 5: A suitable image-text multimodal joint modeling method for multi-task misogyny identification. *arXiv preprint arXiv:2202.09099*, 2022.

[51] N Jindal, PK Kumaresan, R Ponnusamy, S Thavareesan, S Rajiakodi, and BR Chakravarthi. Mistra: Misogyny detection through text–image fusion and representation analysis, natural language processing journal 7 (2024) 100073. *URL: https://www. sciencedirect. com/science/article/pii S*.

[52] Sneha Chinivar, MS Roopa, JS Arunalatha, and KR Venugopal. V-ltcs: Backbone exploration for multimodal misogynous meme detection. *Natural Language Processing Journal*, 9:100109, 2024.

[53] Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073, 2024.

[54] Yu-Chien Tang, Kuang-Da Wang, Ting-Yun Ou, and Wen-Chih Peng. Nycu-two at memotion 3: Good foundation, good teacher, then you have good meme analysis. *arXiv preprint arXiv:2302.06078*, 2023.

[55] Shreyash Mishra, S Suryavardan, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. Overview of memotion 3: Sentiment and emotion analysis of codemixed hinglish memes. *arXiv preprint arXiv:2309.06517*, 2023.

[56] Wentao Yu and Dorothea Kolossa. wentaorub at memotion 3: Ensemble learning for multi-modal meme classification. *Proceedings of De-Factify*, 2, 2021.

[57] Marvin John Ignacio, Thanh Tin Nguyen, Hulin Jin, and Yong-guk Kim. Meme analysis using llm-based contextual information and u-net encapsulated transformer. *IEEE Access*, 2024.

[58] Susmita Ray. A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE, 2019.

[59] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5138–5147, 2021.

[60] Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE, 2019.

[61] Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. Exercise? i thought you said'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 513–524, 2021.

[62] Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 586–591, 2023.

[63] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156*, 2023.

[64] Luke Bates, Peter Ebert Christensen, Preslav Nakov, and Iryna Gurevych. A template is all you meme. *arXiv preprint arXiv:2311.06649*, 2023.

[65] Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231, 2019.