

# Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for Explainable Meme Analysis

Faseela Abdullakutty  
School of Computing  
Robert Gordon University  
Aberdeen, UK  
f.abdullakutty@rgu.ac.uk

Usman Naseem  
School of Computing  
Macquarie University  
Sydney, Australia  
usman.naseem@mq.edu.au

## ABSTRACT

Memes are important because they serve as conduits for expressing emotions, opinions, and social commentary online, providing valuable insight into public sentiment, trends, and social interactions. By combining textual and visual elements, multi-modal fusion techniques enhance meme analysis, enabling the classification of offensive and sentimental memes effectively. Early and late fusion methods effectively integrate multi-modal data but face limitations. Early fusion integrates features from different modalities before classification. Late fusion combines classification outcomes from each modality after individual classification and reclassifies the combined results. This paper compares early and late fusion models in meme analysis. It showcases their efficacy in extracting meme concepts and classifying meme reasoning. Pre-trained vision encoders, including ViT and VGG-16, and language encoders such as BERT, ALBERT, and DistilBERT, were employed to extract image and text features. These features were subsequently utilized for performing both early and late fusion techniques. This paper further compares the explainability of fusion models through SHAP analysis. In comprehensive experiments, various classifiers such as XGBoost and Random Forest, along with combinations of different vision and text features across multiple sentiment scenarios, showcased the superior effectiveness of late fusion over early fusion.

## CCS CONCEPTS

• Information systems → Multimedia and multimodal retrieval.

## KEYWORDS

Multi-modal Meme analysis, fusion, explainability

### ACM Reference Format:

Faseela Abdullakutty and Usman Naseem. 2024. Decoding Memes: A Comprehensive Analysis of Late and Early Fusion Models for Explainable Meme Analysis. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3589335.3652504>



This work is licensed under a Creative Commons Attribution International 4.0 License.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0172-6/24/05  
<https://doi.org/10.1145/3589335.3652504>

## 1 INTRODUCTION

Memes are rapidly shared images, videos, or text on the internet, often with humorous or satirical intent [21]. They have become a significant part of online culture and communication, expressing emotions, opinions, and social commentary. Analyzing memes is crucial as they can convey sentiments, emotions, and intensity of emotions, providing valuable insights into public opinion, trends, and social dynamics [17]. Understanding the sentiment and emotion behind memes can help identify and address issues such as offensive content, hate speech, and spreading false information. Additionally, analyzing memes can help detect and mitigate the harmful effects of toxic memes, which can agitate arguments, disputes, and social wars. Figure. 1 shows a meme, a sample taken from [20].

Memes are analyzed by extracting and interpreting both textual and visual features as they contribute to the meaning and impact of the meme. Textual and visual features are not always correlated in meme datasets, making analysis challenging. Meme analysis requires developing techniques that address the multiple modality problem and improve the classification rate of sentiment, emotion, and intensity. As multi-modal fusion techniques integrate textual and visual modalities, they provide a comprehensive understanding of memes [24]. By combining features, decisions, or data, fusion techniques enhance sentiment analysis.

Due to the fact that memes are multi-modal data, early fusion, and late fusion have been used in multi-modal analysis to combine information from text and images. These two methods can

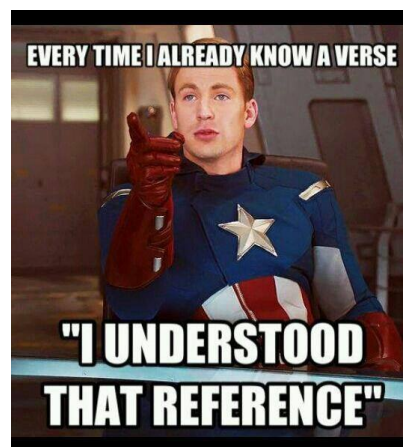


Figure 1: A meme sample from [20]

be applied to detect offensive content, hate speech, and false information in memes. Early fusion involves combining textual and visual content at the input level to provide simplicity as well as joint representation [9]. In spite of this, it may result in a loss of modality-specific information and an increase in noise sensitivity. In contrast, late fusion processes each modality independently and combine the outputs at a later stage, preserving modality-specific information and allowing greater flexibility [7]. It may, however, be more difficult to implement and may lead to inconsistencies. Depending on the task and data characteristics, both approaches may be beneficial. In early fusions, textual elements and visual images provide complementary information, while in late fusions, different analysis techniques and models are required [9]. A combination of both approaches must be tested and evaluated to determine the most effective method for detecting offensive content, hate speech, and false information in memes.

As illustrated at the top of Figure 2, early fusion combines features from various modalities, while in contrast, late fusion combines decisions from different modalities post-classification, as depicted in the bottom portion of Figure 2. However, early fusion faces limitations like time synchronization, integrating dissimilar input features, and retraining classification systems. Late fusion, on the other hand, combines decisions from different modality classifiers, making it easier to combine decisions and learn characteristics. However, [17] removed text from the image before analysis, potentially losing contextual information. Using [14] has the drawback of being heavily reliant on textual information and may not fully capture the visual aspects of harmful memes. In addition to integrating vision and text modalities and creating fusion models, it is also essential to recognize the relevance of features from each modality to sentiment prediction. Hence, explainability is of utmost importance. Utilizing SHAP (Shapley Additive Explanations) enables the analysis of the explainability of the proposed models, facilitating a comparative assessment of the significance attributed to textual and visual features within each model.

This paper compares and contrasts early and late fusion models in meme analysis using visual and textual features extracted from various pre-trained models. The Memotion dataset, which was proposed in [20], was utilized to examine a variety of scenarios using different combinations of features derived from distinct pre-trained vision and language encoders. Vision encoders VGG-16 and ViT were employed, while language encoders BERT, ALBERT, and DistilBERT were utilized. XGBoost and RF classifiers were used to assess and compare the effectiveness of these fusion methods, and corresponding performance results were reported. Furthermore, a comparative analysis of the explainability of each fusion model has been carried out using SHAP.

The main contributions of this article are:

- An evaluation of early and late fusion models in meme analysis, emphasizing their comparative effectiveness. Through extensive experiments involving diverse combinations of vision and text features across various sentiment scenarios, the research demonstrates the superiority of late fusion over early fusion in meme analysis.
- the comparative analysis of the explainability of early and late fusion models using SHAP, shedding light on the extent

of influence of visual and textual features on multi-modal meme analysis.

## 2 RELATED WORK

Memes include both textual and visual content. This necessitates the need for a multimodal framework for detecting them. These frameworks combine visual and language understanding to analyze images and text in memes. One such framework was presented in [18], which used an ensemble of UNITER-based models that combine image and text information for improved performance. Techniques like upsampling contrastive examples and ensemble learning based on cross-validation enhance the framework's robustness. Margin ranking loss helped in predicting higher probability scores for hateful memes. The authors also experimented with augmenting the model with fine-grained object-detection classes from the YOLO9000 model to identify target groups in the image and text, leading to more accurate predictions. The multimodal framework considered images and text, enabling a more comprehensive analysis of hateful memes. Ensemble learning and upsampling of contrastive examples improved the framework's performance and robustness. Margin ranking loss helped prioritize hateful memes during prediction. Augmenting the model with fine-grained object-detection classes enhanced the identification of target groups, leading to more accurate predictions.

Meme analysis uses fusion methods as they allow the integration of multiple modalities, such as text and images, to convey the full meaning and context of memes [1]. Fusion methods, by integrating images and text, can effectively tackle the challenge posed by memes that incorporate both visual and textual elements, both of which are essential for comprehending the intended message. As a result of the fusion of visual and textual features, memes can be classified more accurately and help in curbing misinformation and hate on social media platforms through automatic censoring. Two types of fusion methodologies, namely early fusion and late fusion, have been utilized for this task.[8].

### 2.1 Early Fusion

There are several advantages to using early fusion methods in data analysis, as has been highlighted in recent literature. By integrating information from multiple sources at the beginning of the analysis, they facilitate comprehensive data processing [9]. Moreover, early fusion approaches have been found to improve computational efficiency and reduce processing time as compared to late fusion methods by consolidating information early in the analysis process. In addition, early fusion results in more interpretable models, which aid in the understanding of underlying patterns in the data [4]. Although early fusion methods have many advantages, they are not without limitations. It appears that they may be more susceptible to noise and outliers in the data, which may compromise the quality of the results [7]. Furthermore, early fusion requires the selection of features and fusion techniques in advance, making it difficult to adapt to different datasets or analytical objectives. Moreover, early fusion can result in information loss, as merging features from various sources too early in the analysis pipeline may fail to identify important correlations and dependencies among the data, resulting in suboptimal outcomes [2].

The authors of [17] proposed a meme analysis method that combines visual and textual features for correlated information. They used fusion techniques like multi-hop attention and stacked attention networks, Efficientnet-v2 as a vision backbone, and RoBERTa/LSTM for text processing. The research in [12] introduced pre-trained visual-language models for the multi-modal classification of hateful memes. These models employed Transformer layers to synchronize text and visual inputs, underwent fine-tuning for grounding visual-text slurs, and utilized gradients to signify the contribution of each modality. However, the models exhibited biases, resulting in false-positive predictions, and the extent of capturing derogatory or slur references remains unclear. To classify hateful social media memes using textual and image data, BERT and ResNeXT-152 Aggregated Residual Transformations based Masked Regions with Convolutional Neural Networks (R-CNN) were utilized [10]. The method achieved an AUCROC score of 78%, showcasing its effectiveness in classifying hateful memes.

Incorporating knowledge and multi-modal early fusion techniques, MeBERT [24] classified memes. It entailed three steps: concept retrieval, concept-sensitive visual representation extraction, and multi-modal representation generation. MeBERT enhanced the semantic representation of memes by integrating conceptual information from external Knowledge Bases. It surpassed state-of-the-art techniques on Memotion and MultiOFF datasets. A prompt-based method in [14] converted multi-modal data into text using captions and attributes and fine-tuned pre-trained language models for masked language modeling to identify harmful memes. Experimental results and evaluation metrics such as Accuracy and Macro-F1 demonstrated that the approach outperforms state-of-the-art methods. VaxMeme [16] is a multi-modal framework that identified vaccine-critical memes on Twitter. There were four modules in this system: Text representation learning, Image representation learning, Attentive representation learning, and Classification learning. With an F1-Score of 84.2%, the framework outperformed state-of-the-art methods, demonstrating the importance of understanding both modalities when identifying vaccine-critical memes.

## 2.2 Late Fusion

Late fusion methods have gained significant attention in the field of multi-modal data analysis despite their potential advantages and limitations. In the course of exploring late fusion techniques, it has become apparent that several key factors influence their effectiveness and applicability in different areas. Late fusion enables the integration of complementary information from diverse modalities, thereby enhancing the performance of classification systems. This approach captures nuanced data patterns effectively [7]. It is possible to select classifiers and features based on a specific dataset and task using late fusion methods [4]. As a result of this adaptability, researchers are able to experiment with more diverse combinations without affecting the fusion process itself. The authors of [11] emphasized the importance of processing each modality independently before fusion to minimize errors and enhance overall system robustness, as late fusion approaches exhibit increased robustness in the face of noise and outliers compared to early fusion methods.

In [13], a meme classification system utilized a late fusion approach, combining data from text and image modalities. Data pre-processing involved removing redundant tags, punctuation, and stop-words, alongside converting words to lowercase and expanding short forms to their full versions. The research introduced a challenging dataset for identifying hateful speech within multi-modal memes, suggesting additional investigation to refine the classification framework. The authors of [3] applied a voter-based fusion technique for late fusion, integrating information from image and text modalities after the learning phase. This method entailed training three identical models on image, text, and a concatenation of both embeddings, facilitating interpretation and testing across different models and tasks. Zhang et al. [22] introduced a late sequential fusion scheme, which integrated textual information into a multi-modal system. They utilized UNITER and fine-tuned BERT predictions to modify the medium confidence zone of XGBoost predictions, thereby improving its ability to detect misogynous memes.

Despite the advantages associated with late fusion methodologies, inherent challenges and limitations arise when employing them in the classification pipeline. An important drawback of late fusion techniques is their heightened complexity, which can significantly complicate implementation and optimization efforts. The management of multiple classifiers, features, and fusion techniques escalates computational and algorithmic complexity, potentially hindering interpretability and scalability. Moreover, late fusion methods entail higher computational costs, necessitating additional processing time and resources compared to early fusion methods [4]. Particularly with large datasets or real-time applications, independently processing each modality and subsequently combining their outputs imposes a substantial computational burden. Additionally, late fusion entails the risk of information loss [11], as integrating individual classifier and feature outputs at a later stage may fail to fully exploit correlations and dependencies between different modalities, thereby yielding suboptimal results.

## 3 METHOD

The methodological approach adopted for meme analysis in this study involved the fusion of visual and textual features. Both early and late fusion techniques were applied to integrate these features during experimentation. Pre-trained vision and language encoders were utilized to extract relevant features. XGBoost and Random Forest (RF) classifiers were selected as the primary models for analysis. Classification tasks were conducted individually for each sentiment category present in the dataset, encompassing sentiments such as humour, sarcasm, offensive language, motivation, and overall sentiment. Furthermore, the explainability of each model was evaluated and compared using SHAP analysis.

### 3.1 Vision and Text Encoders

Various vision and text encoders were employed to extract features from memes for fusion and subsequent classification. However, this paper reports only a few of these models, based on their best performance. Vision encoders included pre-trained VGG-16 [23] and ViT [6]. Similarly, pre-trained BERT [5], ALBERT [15], and DistilBERT [19] were used as text encoders to extract features.

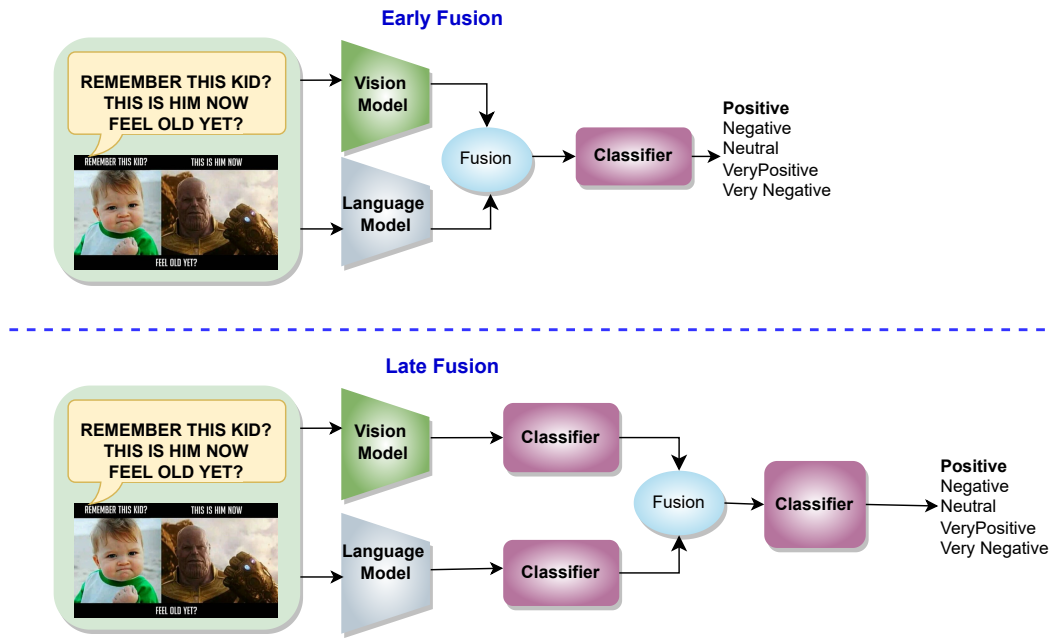


Figure 2: Early and Late Fusion

### 3.2 Early Fusion

One of the fusion strategies utilized in this paper to tackle multi-modal meme analysis is early fusion, depicted in the upper portion of the Figure. 2. Early fusion involved combining vision and text features before classification. To execute early fusion, the experiments outlined in the article involved extracting vision features through pre-trained ViT and VGG-16 models, and text features through pre-trained BERT, ALBERT, and DistilBERT models. The experiments explored various combinations of text and visual features, as depicted in Table.2. Concatenating vision and text feature vectors formed the resulting new feature vectors. These feature vectors were afterward passed to classifiers, including XGBoost and Random Forest, to assess sentiment.

### 3.3 Late Fusion

In meme classification, the experiments employed a late fusion approach. They utilized the same pre-trained models for both feature extraction and classification as those used in the early fusion approach. The late fusion framework independently classified each modality feature. Subsequently, it combined the resulting probability values through concatenation. The resulting feature vectors, which encompassed probability values, underwent additional classification using the same classifier employed in the initial step for sentiment detection. This process is illustrated at the bottom of Figure. 2.

### 3.4 Explainability

SHAP (Shapley Additive Explanations) serves as a powerful instrument, enriching model transparency by offering insights into model behaviour, the significance of features, individual predictions, and

comparisons. It provides a systematic approach to interpreting complex predictions, enabling a better understanding of each feature's impact on the model's output. SHAP values assign importance scores to each feature, indicating its contribution to the model's output. This helps identify influential features in the prediction process, aiding in feature selection, debugging, and understanding data patterns. SHAP values also offer individual explanations, providing insights into why a specific prediction was made and fostering trust in machine learning models, especially in sensitive or high-stakes applications. SHAP also allows for model comparison, enabling the selection of the most suitable model for a task and understanding the trade-offs between complexity and interpretability.

## 4 EXPERIMENTS

A detailed experimental analysis conducted the classification of memes based on extracted visual and textual features from the Memotion dataset.

### 4.1 Dataset

The Memotion dataset [20], was published in 2020, comprises 6992 annotated memes with human-annotated labels for sentiment, emotion type, and intensity. The emotion labels include sarcastic, humour, offensive, motivational, and overall sentiment, categorized into different types. Table. 1 shows the class distribution corresponding to each sentiment.

Except for the *motivational* sentiment, all the other sentiments, including overall sentiment, have multi-class distribution. The *motivational* sentiment has a binary class distribution. The offensive, sarcasm, humour, and overall sentiment have four classes in their distribution. The classes are not balanced, especially regarding sentiments, offensive, and sarcasm. Even in the Overall sentiment scenario, the class distribution is highly imbalanced.

**Table 1: Class distribution corresponding to each sentiment in Memotion dataset [20]**

Sentiment	Classes	Size
Motivational	Non-motivational	4421
	Motivational	2409
Humour	Funny	2394
	Very funny	2176
	Not funny	1618
	Hilarious	642
Offensive	Not offensive	2657
	Slight	2536
	Very offensive	1424
	Hateful offensive	213
Sarcasm	General	3430
	Not sarcastic	1516
	Twisted meaning	1499
	Very twisted	385
Overall sentiment	Positive	3057
	Neutral	2157
	Very positive	1001
	Negative	469
	Very negative	146

## 4.2 Experimental Settings

To extract features from images using the pre-trained vision models, ViT and VGG-16, the images were resized to  $224 \times 224$ . Since the data set itself provided the extracted text from the images, those text data were used to extract text features. The dataset was split into 80 : 20 ratio as train-test sets for model training.

The classified results were evaluated and compared using accuracy, precision, recall, and Macro F1-score metrics. Even though extensive experiments were carried out using various pre-trained vision, language models, and classifiers, the final results were based on the best scenarios.

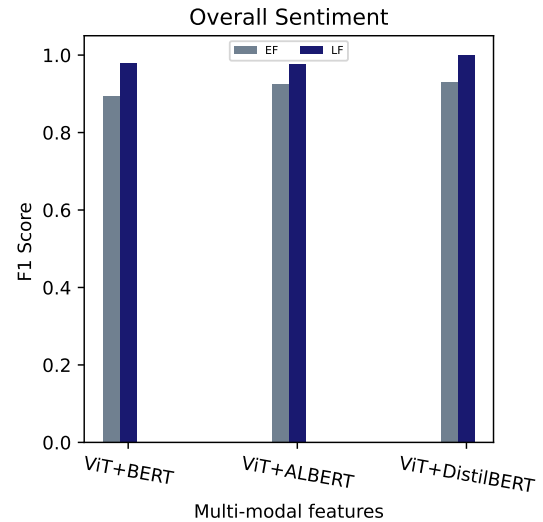
## 5 RESULTS

A comparative analysis result of early and late fusion for multi-modal meme analysis is presented in this section. Table. 2 has the sentiment analysis results using the XGBoost classifier, with a comparison of early and late fusion methods. Figure. 4 illustrates the SHAP summary plot for the early fusion model using the XGBoost classifier. The summary plots for late fusion using XGBoost are presented in Figure. 5.

In Figure 3, the macro F1-score of early and late fusion methods using the RF classifier for the overall sentiment scenario are compared and presented. It can be seen from overall sentiment performance that the late fusion using RF performed better compared to early fusion. Thus, the late fusion improved compared to early fusion, when the memes were classified for various sentiments. It shows the effectiveness of late fusion in multi-modal meme classification.

Performance of late and early fusion methods using XGBoost classifier for different sentiments are presented in Table. 2. It can be seen from the Table that late fusion outperforms early fusion in all scenarios. The highest value for the F1-score was achieved, with motivational sentiment, with ViT and DistilBERT combination.

The early fusion F1-score of 47.38 % improved up to 53.23% in late fusion.



**Figure 3: Comparison of F1-Score for fusion methods using RF classifier, corresponding to overall sentiment. EF= Early fusion and LF = Late fusion**

Figure 4 illustrates the SHAP summary plot for the task overall sentiment analysis with different feature fusion scenarios considered in the experiment while using XGBoost classifier. The plot shows the first 10 features that are more relevant to identifying the five classes in the task. Consider the plot Figure. 4a shows the feature relevance in the early fusion of VGG and BERT features. Out of the 10 features plotted, the positive and negative classes barely used the image features. These both classes relied on BERT features rather than VGG features. When it comes to the VGG+ALBERT scenario, the negative class could be found making use of text features, while the positive class relied upon image features. The positive class used text features in VGG+DistilBERT early fusion. However, the negative class had a slight dependency on image features, too.

When ViT features were combined with text features from BERT, ALBERT, and DistilBERT models, the negative class was mostly influenced by the text features. On the other hand, the positive class was decided mostly by the image features from ViT, even though text features also had a slight impact on the positive class prediction. While considering the other three classes in the overall sentiment task here, when VGG features were combined with text features from different models, the neutral class was mostly influenced by text only. The image feature dependency was very low for this class. Out of the ten features plotted, the very positive class was found to rely on text features rather than image features while influenced by at least 5 features out of 10. The very negative class in the overall sentiment task made use of both features and used 5 or more features out of the 10 features shown in the plot.

Similarly, for other sentiments, when VGG features were combined with any of the text features considered in the experiments, the text features dominated in deciding the classes rather than

**Table 2: Comparison of early and late fusion using XGboost for sentiment analysis on Memotion dataset. A = Accuracy, P = Precision, R = Recall, F = Macro-F1-Score**

Sentiment	Pre-trained Models		Early Fusion				Late Fusion			
	Vision	Language	A	P	R	F	A	P	R	F
Humour	VGG-16	BERT	34.11	24.15	26.06	24.05	32.50	28.32	28.16	28.20
		ALBERT	32.72	25.08	25.36	23.86	30.53	25.83	25.77	25.77
		DistilBERT	33.60	26.93	26.26	24.95	25.72	26.34	26.34	26.31
	ViT	BERT	33.09	23.87	25.26	23.19	25.41	25.35	25.42	25.35
		ALBERT	33.16	28.92	26.02	24.95	29.50	25.53	25.55	25.48
		DistilBERT	32.28	27.55	25.12	23.77	30.67	27.05	26.88	26.90
Offensive	VGG-16	BERT	35.65	21.62	23.63	21.59	34.11	24.43	24.91	24.62
		ALBERT	38.14	24.49	25.29	23.15	34.85	25.78	25.80	25.75
		DistilBERT	37.85	27.56	25.62	24.29	35.80	35.80	35.80	25.27
	ViT	BERT	37.26	23.52	24.86	22.98	37.19	27.34	27.26	27.19
		ALBERT	38.87	25.85	25.91	23.91	35.72	26.43	26.31	26.25
		DistilBERT	38.58	25.73	25.64	23.56	36.09	26.83	26.16	26.07
Sarcasm	VGG-16	BERT	48.17	21.22	24.60	18.49	35.65	24.49	24.41	24.39
		ALBERT	48.24	24.39	25.08	19.76	37.77	26.53	26.36	26.36
		DistilBERT	47.88	21.02	24.57	18.77	35.72	24.78	25.03	24.80
	ViT	BERT	47.88	22.11	24.77	19.34	35.58	24.49	24.49	24.49
		ALBERT	48.46	24.24	25.24	19.99	36.24	25.43	25.38	25.37
		DistilBERT	23.42	22.96	24.72	19.65	36.24	24.41	24.41	24.36
Motivation	VGG-16	BERT	60.76	51.29	50.73	48.23	55.05	49.50	49.53	49.41
		ALBERT	62.01	51.93	50.86	47.11	56.15	50.80	50.80	50.68
		DistilBERT	61.42	51.23	50.59	47.19	57.83	52.26	52.06	51.95
	ViT	BERT	62.01	53.55	51.98	51.98	58.49	52.73	52.43	52.28
		ALBERT	61.86	51.99	50.93	47.47	58.13	52.25	52.00	51.83
		DistilBERT	61.71	51.73	50.82	<b>47.38</b>	59.15	53.68	53.31	<b>53.23</b>
Overall Sentiment	VGG-16	BERT	41.14	17.16	19.72	16.54	33.16	19.50	19.34	19.24
		ALBERT	41.00	20.52	19.78	16.73	33.09	19.18	19.13	18.98
		DistilBERT	43.56	20.68	21.26	18.34	34.99	21.04	20.82	20.72
	ViT	BERT	42.02	20.18	20.18	17.95	33.02	20.54	19.83	19.95
		ALBERT	41.95	20.93	20.34	17.34	33.89	19.55	19.41	19.33
		DistilBERT	42.53	19.78	20.73	17.79	33.89	19.95	19.59	19.48

VGG features. On the other hand, ViT features were combined with BERT, ALBERT, or DistilBERT features; the classes were impacted by both image and text features.

Figure 5 illustrates the plots corresponding to the SHAP analysis on late fusion models using the XGBoost classifier. Out of the five classes in the task, the neutral class depends mostly on all the training features but is prominently dependent on VGG features in the VGG+BERT scenario. When VGG and ALBERT features were used in late fusion, the neutral class exhibited almost similar dependency as in the VGG+BERT. For the other scenarios, VGG+DistilBERT, ViT+BERT, ViT+ALBERT, and ViT+DistilBERT, the neutral class depends upon mostly all features, except a few.

For other classes in the VGG+BERT scenario, the negative class was influenced by all the probabilities from the uni-modal output. On the other hand, the positive class directly depends on the uni-modal probabilities of the same class. With all the other evacuating scenarios, the positive class only depends on the uni-modal probability of the class itself. The negative class in the overall sentiment task showed behaviour almost similar to the positive class in the late fusion SHAP analysis. Most of the uni-modal probabilities from the text and image classification models in the VGG+BERT scenario

influenced the very positive class. When evaluated by combining image features with DistilBERT in late fusion, the very positive class did not reply to the neutral, negative, and negative class probabilities from uni-modal models.

## 5.1 Discussion

The late fusion using both classifiers performed better compared to early fusion. In Humour sentiment classification with RF, there is a slight drop in performance when features used combined ViT and ALBERT. However, the rest of the combinations exhibited improved late fusion performance than early fusion. A similar result was shown when the offensive sentiment scenario was considered. Thus, the late fusion improved compared to early fusion, when the memes were classified for various sentiments. Rather than combining features, it would be better to classify them individually and combine the classifications for better sentiment analysis using memes. Finally, our analysis showed that the highest performance was achieved using late fusion, with an F1-Score of 53.23%. This indicates the potential to design new and innovative methods to improve the performance of meme analysis further.

The positive and negative classes barely used image features in the early fusion of VGG and BERT features. The negative class in



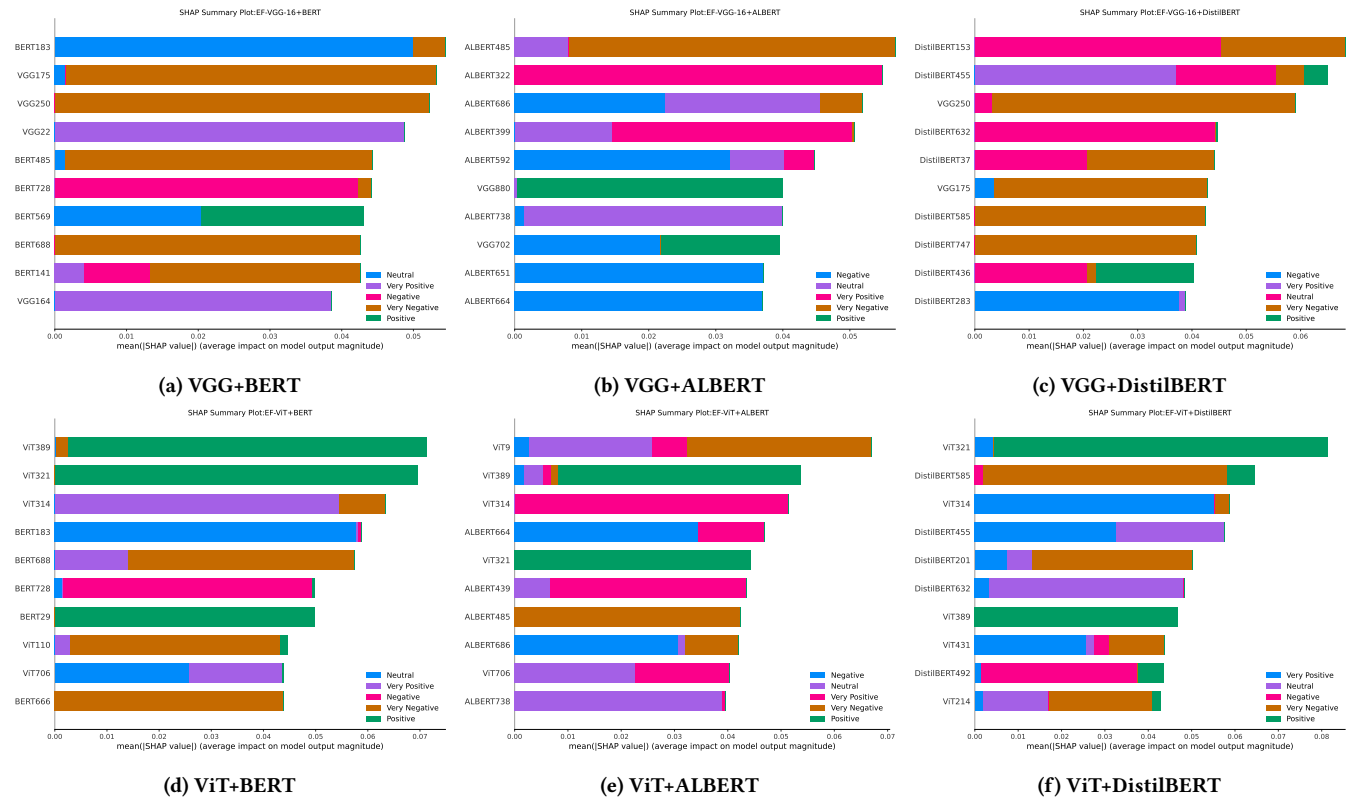


Figure 4: SHAP summary plots for the overall sentiment task with early fusion

the VGG+ALBERT scenario used text features, whereas the positive class relied on image features. VGG+DistilBERT early fusion used text features but slightly depended on image features. In combining ViT features with text from BERT, ALBERT, and DistilBERT models, the negative class was mostly influenced by text features, whereas image features mainly influenced the positive class. In the early fusion, text features were more prominent when combined with VGG. However, the early fusion model using ViT demonstrated improved relevance of image features in the prediction process. Based on this analysis, it appears that the pre-trained VGG model was unable to extract features relevant to the problem, while the ViT was able to obtain features that were more appropriate to the problem at hand.

In late fusion models using the XGBoost classifier, the neutral class depends on all detection probabilities of each class in the uni-modal classification. However, it depends most significantly on VGG features in the VGG+BERT scenario. When VGG and ALBERT features were used in late fusion, the neutral class exhibited a similar dependence. As in other scenarios, the neutral class is determined by all features except a few. It is important to note that the negative class is affected by all probabilities from the uni-modal output, while the positive class depends on the uni-modal probabilities of the same class. In the overall sentiment task, the very negative class showed behaviour similar to that of the positive class in the late fusion SHAP analysis. Almost all uni-modal probabilities from the text and image classification model were associated with very positive classes. It

implies that all the other classes relied on uni-modal classification probabilities from both image and text classification models in late fusion except for the positive and very negative classes. The positive and the very negative class performance depended upon the class probability of these classes in the uni-modal text and classification models using XGBoost. Hence, improving the initial uni-modal classification performance with an image could facilitate the overall performance in late fusion. This also points to the fact that the training features of the uni-modal classification should be the most task-specific to achieve the best classification.

Early and late fusion SHAP analyses demonstrate the necessity of fine-tuning feature extraction models using meme datasets before extracting features for fusion and classification. However, fine-tuning may enhance performance further. Consequently, one of the future directions will be to extract fine-tuned features for meme analysis.

Analyzing memes can raise ethical considerations. Even though memes are often shared publicly on social media platforms, individuals depicted in memes may not have consented to their images being used for analysis. This raises ethical concerns regarding the collection and analysis of data without explicit permission, potentially infringing upon individuals' rights to privacy. Nevertheless, this article used public datasets with proper references, which mitigated ethical and privacy concerns associated with the use of these memes.

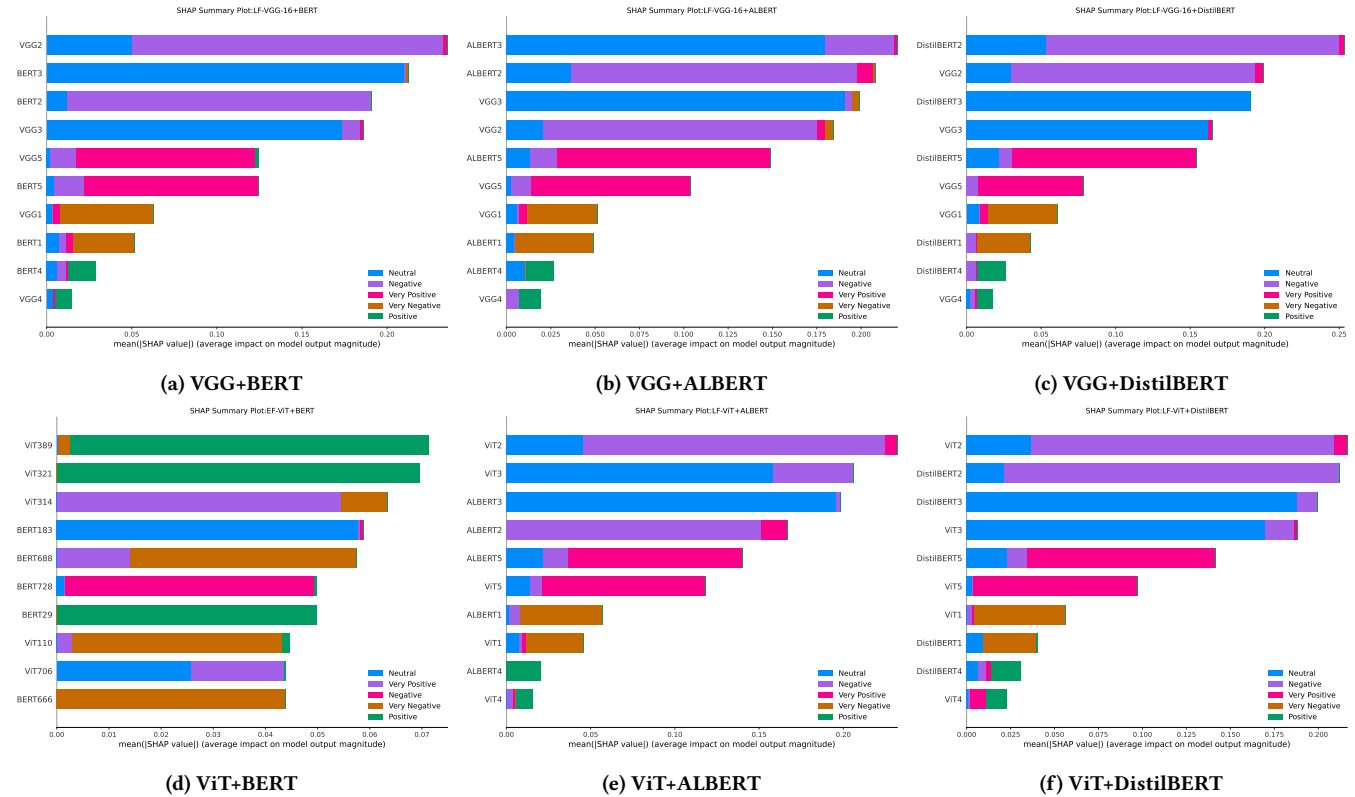


Figure 5: SHAP summary plots for the overall sentiment task with late fusion

## 6 CONCLUSION

Multi-modal fusion techniques are crucial in meme analysis for seamless integration of visual and textual features. Multi-modal data fusion techniques have been demonstrated to be effective in meme classification, with both early and late fusion methods showing distinct advantages and limitations. Hence, a comparative analysis of early and late fusion in multi-modal meme analysis is presented in this paper. The extensive experiments using different classifiers and combinations of different vision and text features on multiple sentiment scenarios showed that late fusion is more effective than early fusion. The analysis of explainability revealed that in late fusion, the classes were predominantly influenced by the respective uni-modal prediction probabilities, indicating the necessity for extracting more appropriate features through additional fine-tuning procedures. Hence, this experiment will be further extended to address meme analysis by using more fusion methods incorporating fine-tuned models for feature extraction. Incorporating multi-modal Large Language Models (LLMs) for meme analysis stands as a prospective future research direction.

## REFERENCES

- [1] Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2021. A multimodal memes classification: A survey and open research issues. In *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*. Springer, 1451–1466.
- [2] George Barnum, Sabera Talukder, and Yisong Yue. 2020. On the benefits of early fusion in multimodal representation learning. *arXiv preprint arXiv:2011.07191* (2020).
- [3] Lisa Bonheme and Marek Grzes. 2020. SESAM at SemEval-2020 task 8: investigating the relationship between image and text in sentiment analysis of memes. (2020).
- [4] Petra Budikova, Michal Batko, and Pavel Zezula. 2017. Fusion strategies for large-scale multi-modal image retrieval. *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXIII* (2017), 146–184.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [7] Konrad Gadjicki, Razieh Khamsehashari, and Christoph Zetzsche. 2020. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 1–6.
- [8] Ankita Gandhi, Kinjal Adharyu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* 91 (2023), 424–444.
- [9] Apeksha Gaonkar, Yogya Chukkappalli, P Jahnnavi Raman, Sahana Srikanth, and Sanjeev Gurugopinath. 2021. A comprehensive survey on multimodal data representation and information fusion algorithms. In *2021 International Conference on Intelligent Technologies (CONIT)*. IEEE, 1–8.
- [10] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Amanullah Yasin, Gautam Srivastava, Dawid Polap, Thippa Reddy Gadekallu, and Zunera Jalil. 2023. Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. *ACM Transactions on Asian and Low-Resource Language Information Processing* (2023).
- [11] Mehrtaash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C Lovell. 2013. Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *Proceedings of the IEEE international conference on computer vision*.



- 3120–3127.
- [12] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*. 3651–3655.
- [13] Rachana Jadhav and Vikas N Honmane. 2021. MEMES CLASSIFICATION SYSTEM USING COMPUTER VISION AND NLP TECHNIQUES. *International Journal of Engineering Applied Sciences and Technology* (2021).
- [14] Junhui Ji, Wei Ren, and Usman Naseem. 2023. Identifying Creative Harmful Memes via Prompt based Approach. In *Proceedings of the ACM Web Conference 2023*. 3868–3872.
- [15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
- [16] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2023. A Multimodal Framework for the Identification of Vaccine Critical Memes on Twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 706–714.
- [17] Thanh Tin Nguyen, Nhat Truong Pham, Ngoc Duy Nguyen, Hai Nguyen, Long H Nguyen, and Yong-Guk Kim. 2022. HCLab at Memotion 2.0 2022: Analysis of sentiment, emotion and intensity of emotion classes from meme images using single and multi modalities. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*.
- [18] Lippe Phillip, Holla Nithin, Chandra Shantanu, Rajamanickam Santhosh, Antoniou Georgios, Shutova Ekaterina, and Yannakoudakis Helen. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871* (2020).
- [19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).
- [20] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. SemEval-2020 Task 8: Memotion Analysis–The Visuo-Lingual Metaphor! *arXiv preprint arXiv:2008.03781* (2020).
- [21] Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274* (2022).
- [22] Jing Zhang and Yujin Wang. 2022. 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)*. 585–596.
- [23] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. 2015. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence* 38, 10 (2015), 1943–1955.
- [24] Qi Zhong, Qian Wang, and Ji Liu. 2022. Combining knowledge and multi-modal fusion for meme classification. In *International Conference on Multimedia Modeling*. Springer, 599–611.