Enhancing Sentiment Analysis through Multimodal Fusion: A BERT-DINOv2 Approach

Taoxu Zhao[1], Meisi Li[1], Kehao Chen[1], Liye Wang[1], Xucheng Zhou[1], Kunal Chaturvedi[2], Mukesh Prasad[2], Ali Anaissi[1,2], and Ali Braytee[2]

[1] The University of Sydney, Camperdown, Australia
[2] University of Technology Sydney, Ultimo, Australia

Abstract. Multimodal sentiment analysis enhances conventional sentiment analysis, which traditionally relies solely on text, by incorporating information from different modalities such as images, text, and audio. This paper proposes a novel multimodal sentiment analysis architecture that integrates text and image data to provide a more comprehensive understanding of sentiments. For text feature extraction, we utilize BERT, a natural language processing model. For image feature extraction, we employ DINOv2, a vision-transformer-based model. The textual and visual latent features are integrated using proposed fusion techniques, namely the Basic Fusion Model, Self-Attention Fusion Model, and Dual-Attention Fusion Model. Experiments on three datasets—the Memotion 7k dataset, MVSA-single dataset, and MVSA-multi dataset—demonstrate the viability and practicality of the proposed multimodal architecture.

Keywords: Sentiment analysis · Fusion models · Multimodal learning · Self-Attention mechanism.

1 Introduction

Multimodal sentiment analysis works process data and combine the different data modalities first, then use a single model to analyse the combination of feature vectors of the data input, but lack the consideration of the correlation between data modalities. For example, facial and other body features can be implemented with the speaker's words, and different facial expressions with the same words can have different meanings. Recently, many multimodal sentiment analysis approaches have been proposed. The authors [2] integrate SentiBank's capability to extract 1200 adjective-noun pairs for image classification with SentiStrength's text-based sentiment scoring [20]. CNN-Multi [11] leverages pretrained CNNs for text and image feature extraction, concatenating these features for subsequent classification through four fully connected layers. Similarly, DNN-LR [26] utilizes pre-trained convolutional neural networks (CNNs) for extracting text and image representations, employing logistic regression for sentiment classification. MultiSentiNet's approach [24], combines long short term memory (LSTM) model and visual geometry group (VGG) model for feature extraction, integrating results via weighted average or image feature-guided attention.
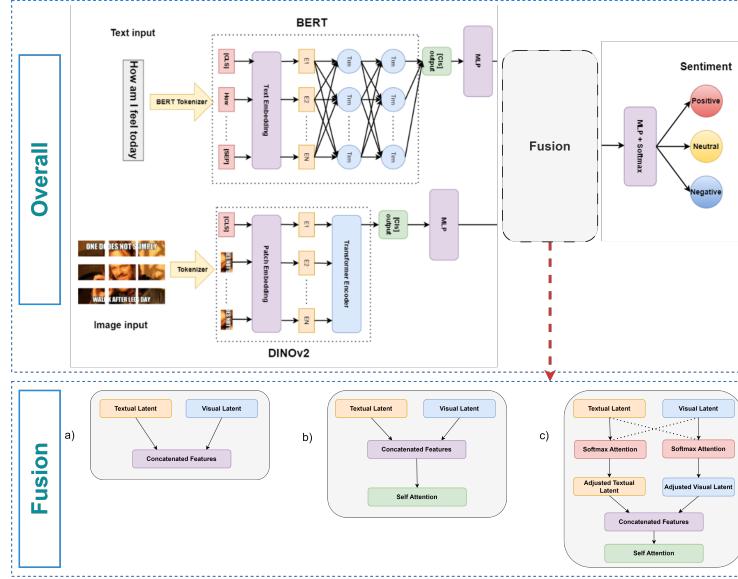
Fig. 1: The overall architecture of the proposed framework (Above). The fusion methodology of the framework (Below): a) Basic Fusion Model; b) Self-Attention Fusion Model; c) Dual-Attention Fusion Model

The enhancement of unimodal statistic models, which can serve as components within multimodal frameworks, is progressing through the adoption of innovative methodologies. Notably, the emergence of Bidirectional Encoder Representations from Transformers (BERT) [6] and its variants have demonstrated superior performance over preceding models, such as recurrent neural networks (RNNs) and LSTMs in text-based sentiment classification due to its refined understanding of textual syntax and semantics. In parallel, the introduction of Vision Transformers [7], represents a significant extension of the transformer model architecture, originally conceived for NLP applications, into computer vision. These models leverage the strengths of large pre-trained datasets, showing enhanced efficacy than CNNs in many image-related tasks. However, we find that vision transformer models may be able to better deal with the vision feature than CNNs.

A good latent representation greatly influences the sentiment analysis results and provides an excellent foundation for the following fusion phase. Therefore, our research aims to harness the capabilities of these newly advanced techniques through their integration via innovative fusion methodologies. Modal fusion is one of the main challenges in multimodal constructions. A single concatenation method [22] can effectively integrate the multimodal information. Huang et al. [9] proposed a model using a "Deep Multimodal Attentive Fusion" method to deal with the fusion issue. They also pointed out the drawbacks of the early fusion methods and emphasized the great performance of late fusion methods as they cannot always unlock the full potential of each modal data. Yu et al. [25] used

the attention technique adding on the BERT model to deal with target-oriented tasks. Tsai et al. [21] proposed cross-attention technique to combine the different latent representations effectively, enhancing the performance of each single modality. Recently, Lee et al. [13] explored multimodal learning by leveraging BERT and DINOv2 for modeling social interactions. However, their focus is on aligning language-visual cues for referent tracking and speaker identification, whereas our work aims at sentiment analysis and proposes novel attention-based fusion techniques tailored for this specific task. The methodologies presented in these papers have guided us in establishing a late fusion strategy and have inspired us to employ attention-based fusion methods.

We select two advanced unimodal models: BERT [6] and DINOv2 [17] (a vision transformer-based model) and then propose three fusion methods to create a multimodal (textual and visual) sentiment analysis model combining the extracted features from separate (text and image) modalities, named Basic Fusion Model, Self-Attention Fusion Model, and Dual-Attention Fusion Model. These methods are designed to synergize the distinct features extracted from separate modalities.

Our contributions:

– Combined extracted features from multimodal data using different fusion methods, such as the Basic Fusion Model, Self-Attention Fusion Model, and Dual-Attention Fusion Model.
– Extensive experiments to compare the effectiveness of the methods using three multimodal sentiment analysis datasets.

2   Method

Here, first, we explain the unimodal information extraction process for both textual and visual context information, respectively. Next, we detail the various fusion methods that combine latent features from both text and image data to be processed through several fully connected layers to predict sentiment analysis. The overall framework is shown in Figure 1.

2.1   Textual Features

BERT [6] is an existing innovative model meticulously designed for a series of natural language understanding tasks. BERT's strength is its pre-training tasks and bidirectional representation, which make it well-known for various downstream NLP tasks. Traditional models like LSTM read the text input in one direction, either forward or backward, which makes them less effective in understanding the contextual relationship between words [15]. BERT's bidirectional attention mechanism allows it to consider both the preceding and following text, offering a more comprehensive understanding of each word within its context [6].

This pre-training involves tasks such as Next Sentence Prediction (NSP) and Masked Language Modeling (MLM), which equip the model with a strong understanding of language semantics and structure [6]. As a result, the pre-trained

BERT model can serve as a robust starting point for a wide range of NLP tasks, from text summarization to question answering. Given BERT's advantages, it's evident that its capability to understand textual context is excellent. Therefore, it is well-suited to serve as our text feature extraction model. In the proposed architecture, we utilize a BERT layer for initial text processing. This involves feeding a sequence of input tokens $X_t = \{x_1, x_2, ..., x_n\}$ into the BERT model, which produces a sequence of output embeddings $H = \{h_1, h_2, ..., h_n\}$. Each embedding h_i is a 768-dimensional vector that captures the contextual information of the corresponding token within the entire sequence.

To adapt these embeddings for our multimodal sentiment analysis, we apply a linear transformation directly to the entire sequence of output embeddings from BERT, reducing the dimensionality of each vector from 768 to 256. The transformation defined in Eq. 1 is designed to condense the rich linguistic information encoded by BERT into a more compact and relevant representation for further analysis.

$$t_i = Wh_i + b \tag{1}$$

where W is a 768×256 weight matrix, b is a bias vector, and t_i represents the transformed embedding of the ith token.

For the final textual latent representation t, defined in Eq. 2, we use the transformed CLS embedding t_{CLS}, which represents a condensed view of the entire input sequence:

$$t = t_{\text{CLS}} \tag{2}$$

This approach leverages the CLS token's embedding directly after merging all the necessary information for sentiment analysis following its transformation. The representation, t, is then used as part of the fusion process.

2.2   Visual Features

DINOv2 [17] is the name of a series of pre-trained visual models, which are trained using Vision Transformers (ViT) [7] on LVD-142M dataset. The key distinction between the ViT and traditional transformer architectures is that ViT divides an image into individual patches and treats these patches as tokens. In this way, the image is viewed as a sequence of tokens. ViT processes the entire image as a unified sequence, allowing it to capture global information instead of CNNs, focusing only on local information through the convolutional kernels. As a result, ViT often outperforms CNNs in various scenarios.

However, when referring to DINOv2, it not only refers to the ViT models but also a training methodology, a self-supervised learning approach [17]. Self-supervised learning achieves excellent segmentation results and robust performance, without fine-tuning, across various computer vision tasks. DINOv2 can learn from any collection of images and acquire features that current standard methodologies can't learn, making it one of the most advanced visual models today.

In the architecture, DINOv2 serves as the foundational layer for extracting complex features directly from image inputs. Each input image I is first divided into a sequence of patches $\{p_1, p_2, ..., p_m\}$, which are then linearly embedded into tokens and processed through the DINOv2 transformer network. This process (see Eq. 3) transforms each image into a sequence of patch embeddings, providing a rich representation of the visual information:

$$G = \text{DINOv2}(I) = \text{DINOv2}(p_1, p_2, ..., p_m) \tag{3}$$

Upon processing images through the transformer, the proposed model employs a custom linear transformation layer, which maps the transformer's output from a 384-dimensional space to a 256-dimensional space for each patch embedding. This transformation is designed to adapt the broad capabilities of the DINOv2 by refining the feature representations to a more manageable size for the analysis of sentiments:

$$v_i = W g_i + b \tag{4}$$

where W is a 384×256 weight matrix, b is a bias vector, and v_i represents the transformed embedding of the ith patch.

For the final visual latent representation v, defined in Eq. 5, we use the transformed global feature embedding v_{CLS}, which represents a condensed view of the entire image:

$$v = v_{\text{CLS}} \tag{5}$$

This approach leverages the global feature embedding directly. It is then used as part of the fusion process with textual features extracted from the BERT model in our sentiment analysis framework.

2.3   Attentional Feature Fusion

The proposed architecture integrates and leverages the strengths of both textual and visual representation models for multimodal sentiment analysis. The integration allows for the extraction and fusion of high-level features from both text and images, facilitating a comprehensive understanding of multimodal data. For this, we propose three fusion methods. A Basic Fusion Model that concatenates the textual and visual latent representation (as shown in Figure 1 a). Let t be the textual latent representation obtained from the BERT model, and v be the visual latent representation obtained from the DINOv2 model. Both t and v are vectors. The simple concatenation c is defined in Eq. 6

$$c = [t, v] \tag{6}$$

where $[t, v]$ represents the concatenation of vectors t and v. The dimension of c will be the sum of the dimensions of t and v. To capture the mutual information of the concatenated latent representation, we introduce a self-attention layer after the concatenation layer as the second method, namely the Self-Attention Fusion

Model (as shown in Figure 1 b). The output s of the self-attention mechanism is defined in Eq. 7 as,

$$s = \text{Softmax}\left(\frac{cW_Q(cW_K)^T}{\sqrt{d_k}}\right) cW_V \tag{7}$$

where c is the concatenated vector given from Eq. 6. W_Q, W_K, and W_V are the weight matrices for the query, key, and value, respectively, which are applied to the vector c. d_k is the dimension of the key.

The third method is the Dual-Attention Fusion Model (as shown in Figure 1 c). Based on the second fusion method, it additionally uses information from another modality to adjust the latent representation vector of each modality, employing a cross-modal attention mechanism. First, queries, keys, and values for both modalities are computed using Eq. 8,

$$\begin{aligned} Q_t = tW_{Q_t}, \quad & K_t = tW_{K_t}, \quad V_t = tW_{V_t}, \\ Q_v = vW_{Q_v}, \quad & K_v = vW_{K_v}, \quad V_v = vW_{V_v}. \end{aligned} \tag{8}$$

Next, we apply softmax attention for cross-modal adjustments as shown in Eq. 9,

$$\begin{aligned} t' &= \text{Softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right) V_v \\ v' &= \text{Softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right) V_t \end{aligned} \tag{9}$$

Finally, we apply concatenation and self-attention as defined in Eq. 10,

$$s' = \text{Softmax}\left(\frac{c'W_Q(c'W_K)^T}{\sqrt{d_k}}\right) c'W_V \tag{10}$$

where c' is the concatenation of t' and v'.

3   Experiments

3.1   Datasets

Memotion 7k Dataset [19] consists of 6992 samples, each paired with an image and the corresponding caption, representing a complete 'meme'. The images are scraped using the Google Images search engine, and the annotation process is carried out with the help of five annotators. The final annotations are adjudicated based on a majority voting scheme. The dataset is part of a challenge that includes three subtasks: analyzing memes for sentiment, which can be positive, negative, or neutral.

MVSA Datasets [16] collected from X, are designed for sentiment analysis on multi-view social data. The MVSA-Single dataset contains 4,869 pairs of images and texts labelled by a single annotator. The MVSA-Multi dataset consists of 19,600 text-image pairs, each labeled by three annotators, ensuring a richer

and more robust sentiment analysis. Similarly to Memotion 7k Dataset, MVSA datasets have three classes: positive, neutral, and negative.

3.2 Experiment settings

Specifically, for the Memotion dataset, we employed macro F1 as our evaluation metric, requiring us to pay attention to the performance across each class. For MVSA-single and MVSA-multi, we used accuracy and F1-score. We utilized focal loss [14] as the loss function and adjusted the γ parameter (γ=2, 3, 4) to ensure the model adequately focuses on the minority classes. However, for the MVSA dataset, our primary evaluation metrics are weighted F1 and accuracy.

To improve the performance of our model, we tuned a set of hyperparameters to facilitate model convergence. Across all datasets, we experimented with different learning rates: 0.01, 0.001, 0.0001, and 0.00001. Adam optimizer is used with dropout rates set to 0, 0.2, and 0.5. Cross-entropy loss function has been used in the experiments.

4 Results and Discussion

In this section, we present experiments conducted on the proposed fusion methods across three different datasets and compare them with benchmark models and state-of-the-art models so that the usability and generalizability of models can be ensured. Additionally, the ablation study is performed to provide in-depth research on the model's fusion methods. At last, we show experiments conducted while tuning hyperparameters for all models across all datasets.

4.1 Comparison with the state of the art

To better demonstrate the usability and SOTA performance of our proposed model in this domain, we compare it with the literature corresponding to the relevant datasets using the same experimental setup.

MVSA datasets The comparative analysis presented in Table 1 showcases the performance of various models on the MVSA-single and MVSA-multi datasets. Notably, the integration of BERT and DINOv2 models, through concatenation, achieves the best performance on the MVSA-single dataset, with an Accuracy of 0.73 and an F1 score of 0.71, surpassing all other models including MultiSentiNet's attention-based approach (MultiSentiNet-Att) [24], which leads on the MVSA-multi dataset with an accuracy of 0.68 and an F1 score of 0.68.

This approach outperforms the previously established benchmarks, including SentiBank [2], CNN-Multi [4], DNN-LR models [26] and HSAN [23], and even the advanced MultiSentiNet [24] and Dual-Pipeline [3] models. The BERT+DINOv2 model with additional self-attention mechanisms also shows strong performance,

Table 1: Results for MVSA-single and MVSA-multi Compared to the Literature

Model	MVSA-single		MVSA-multi	
	Acc	F1	Acc	F1
SentiBank (image only) [2]	0.45	0.43	0.55	0.51
SentiStrength (text only) [2]	0.49	0.48	0.50	0.55
SentiBank+SentiStrength [2]	0.52	0.50	0.65	0.55
HSAN [23]	-	0.66	-	0.67
DNN-LR [26]	0.61	0.61	0.67	0.66
CNN-Multi [4]	0.61	0.58	0.66	0.64
MultiSentiNet-Avg [24]	0.66	0.66	0.67	0.66
MultiSentiNet-Att [24]	0.69	0.69	0.68	0.68
Dual-Pipeline [3]	0.57	0.56	**0.73**	0.69
Ours (Basic Fusion Model)	**0.73**	**0.71**	0.68	0.67
Ours (Self-Attention Fusion Model)	0.72	0.70	0.68	0.67
Ours (Dual-Attention Fusion Model)	0.72	**0.71**	0.67	0.66

underlining the potential of attention mechanisms in multimodal sentiment analysis. This comparison underscores the advances in multimodal sentiment analysis, demonstrating that the fusion of high-performing models like BERT and DINOv2, especially when combined with sophisticated techniques such as self-attention, can lead to substantial improvements. It is worth noting that our model falls short of the state of the art model in the MVSA-multi dataset, which warrants further discussion in the subsequent section. Overall, the performance of our model architectures remains strong, robust, and adaptable.

Memotion 7k dataset As detailed in Table 2, our approach is benchmarked against the competition baseline [19] and the leading models from the leaderboard. The leading models from the competition, as highlighted in the literature, showcase a wide array of computational strategies. Vkeswani IITK from Keswani et al. [10] exemplifies this diversity by employing a spectrum of algorithms, ranging from fundamental linear classifiers like FFNN and Naive Bayes to more sophisticated transformers such as MMBT [18] and BERT [6], applied across both text-only and combined text-image inputs. Similarly, Guo et al. [8] utilizes an ensemble methodology, leveraging textual features extracted via Bi-GRU, BERT, or ELMo in conjunction with visual features sourced from Resnet50. The approach "Aihaihara" as mentioned in [19] opts for a concatenated approach, merging visual cues from a VGG-16 pre-trained model with textual insights from an n-gram language model, illustrating the varied approaches towards integrating visual and linguistic data. Das and Mandal [5] employed a multi-task learning framework that combined ResNet18 for the visual task and bidirectional LSTM and GRU for textual tasks. Bejan [1] combined the BERT model for text understanding and VGG16 pre-trained model for extracting visual information.

The results demonstrate that our model, leveraging a sophisticated fusion of softmax attention and self-attention mechanisms, not only surpasses the competition baseline but also edges out the top-performing models from the leaderboard with a macro F1 score of 0.3552. This achievement underscores the ef-

Table 2: Result for Memotion 7k

Model	Macro F1
Competition Baseline [19]	0.2176
Vkeswani IITK [10]	0.3546
Guoym [8]	0.3519
Aihaihara [19]	0.3501
Sourya Diptadas [5]	0.3488
MemoSYS [1]	0.3475
Ours (Basic Fusion Model)	0.3237
Ours (Self-Attention Fusion Model)	0.3436
Ours (Dual-Attention Fusion Model)	**0.3552**

fectiveness of our model in handling the intricacies of multimodal sentiment analysis, showcasing its potential to advance the state of the art in this field.

4.2   Breakdown Analysis and Discussion

Examining the breakdown performance plot (as shown in Figure 3) in the training process, we can easily discover the issue. Due to Adam's adaptive learning rate [12], the model converges well right from the start, especially for the positive class, reaching near-optimal performance almost in the first epoch. The bias towards predicting the positive class helps maintain high levels of weighted F1 and accuracy. However, for the neutral and negative classes, the model learns fewer relevant features by comparison. From the cases in MVSA-multi, we have come to realize that in addition to understanding the overall performance of the model, the predictive capability for each class is also crucial.

However, the evaluation criteria for the Memotion dataset and the MVSA dataset differ; Memotion focuses on predictions for each class, while MVSA emphasizes overall performance. Therefore, in this section, we aim to explore the predictive performance for different classes, and using the Memotion dataset is a good choice for this purpose. For further investigating how the best model maintains its learning for each class during the model convergence process, we monitored the changes in the F1 score for each class (as shown in Figure 2).
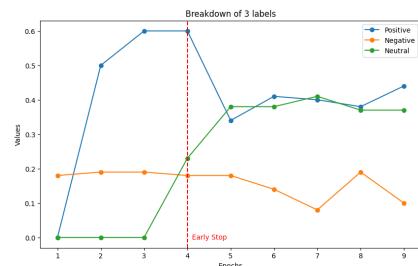

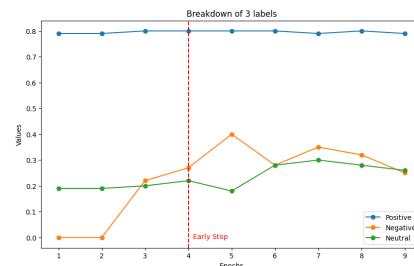Fig. 2: Breakdown performance for the best model for Memotion 7k dataset

Fig. 3: Breakdown performance for the best model for MVSA-multi dataset

From Figure 2, in the epoch where early stopping occurs, we find that the model still inevitably achieves the highest F1 score on the positive class, as it is the dominant class. However, we can see that although there is a large difference in the number of samples between the neutral and negative classes, their predictive performance is similar and not far off from that of the positive class. This suggests that focal loss has effectively captured the minority classes in the dataset without ignoring them.

Example 1

Example 2

Fig. 4: Example of misclassified labels

Further, we selected two fairly typical photos from the set and found that all of them feature individuals who are smiling. This will likely make the model perceive them as expressing positive emotions (as shown in Figure 4). However, another possible reason for the model's incorrect prediction could be issues with the dataset's annotation. As can be seen from Example 1 in Figure 4, although the model has classified a 'neutral' label as 'positive', when we consider what he is saying, from a human perspective, this should be a 'negative' class because his smile is dry. By using image embeddings that include text, a more nuanced representation of the visual and textual data can be captured, which will help improve the accuracy of emotion classification. This approach will enhance the model's ability to discern more subtle emotional expressions and contextual factors that influence perceived emotions, leading to more accurate predictions.

5   Conclusion

Our model shows outstanding performance over multiple datasets, outperforming both baseline and exceeding (or reaching) state-of-the-art performance. This indicates that the model we proposed, along with the three fusion methods, can be well suited to multimodal sentiment analysis. In all three datasets, each fusion strategy yielded a best-performing model, indicating that our fusion methods are highly adept at integrating mutual information across multiple modalities. Although our model's performance is close to or surpasses current SOTA models, there are areas for future improvement. In the future, we will incorporate additional modal inputs like audio and explore advanced fusion techniques and different feature extraction models, including variants of BERT and CNN frameworks.

Acknowledgment

References

1. Bejan, I.: Memosys at semeval-2020 task 8: Multimodal emotion analysis in memes. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1172–1178 (2020)
2. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. pp. 223–232 (10 2013). https://doi.org/10.1145/2502081.2502282
3. Braytee, A., Yang, A.S.C., Anaissi, A., Chaturvedi, K., Prasad, M.: A novel dual-pipeline based attention mechanism for multimodal social sentiment analysis. In: Companion Proceedings of the ACM on Web Conference 2024. p. 1816–1822. WWW '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3589335.3651967, https://doi.org/10.1145/3589335.3651967
4. Cai, G., Xia, B.: Convolutional neural networks for multimedia sentiment analysis. In: Li, J., Ji, H., Zhao, D., Feng, Y. (eds.) Natural Language Processing and Chinese Computing. pp. 159–167. Springer International Publishing, Cham (2015)
5. Das, S.D., Mandal, S.: Team neuro at semeval-2020 task 8: multi-modal fine grain emotion classification of memes using multitask learning. arXiv preprint arXiv:2005.10915 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
8. Guo, Y., Huang, J., Dong, Y., Xu, M.: Guoym at semeval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1120–1125 (2020)
9. Huang, F., Zhang, X., Zhao, Z., Xu, J., Li, Z.: Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems **167** (03 2019). https://doi.org/10.1016/j.knosys.2019.01.019
10. Keswani, V., Singh, S., Agarwal, S., Modi, A.: Iitk at semeval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes. arXiv preprint arXiv:2007.10822 (2020)
11. Kim, Y.: Convolutional neural networks for sentence classification (2014)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
13. Lee, S., Lai, B., Ryan, F., Boote, B., Rehg, J.M.: Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14585–14595 (2024)
14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
15. Manasa, P., Malik, A., Batra, I.: Detection of twitter spam using glove vocabulary features, bidirectional lstm and convolution neural network. SN COMPUT. SCI. **5**, 206 (2024). https://doi.org/10.1007/s42979-023-02518-1, https://doi.org/10.1007/s42979-023-02518-1

16. Niu, T., Zhu, S., Pang, L., El-Saddik, A.: Sentiment analysis on multi-view social data. In: MultiMedia Modeling. p. 15–27 (2016)
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
18. Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.P., Hoque, E.: Integrating multimodal information in large pretrained transformers (2020)
19. Sharma, C., Paka, Scott, W., Bhageria, D., Das, A., Poria, S., Chakraborty, T., Gambäck, B.: Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In: Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020). Association for Computational Linguistics, Barcelona, Spain (Sep 2020)
20. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**, 2544–2558 (12 2010). https://doi.org/10.1002/asi.21416
21. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences (2019)
22. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval-an empirical odyssey. In: CVPR (2019)
23. Xu, N.: Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 152–154 (2017). https://doi.org/10.1109/ISI.2017.8004895
24. Xu, N., Mao, W.: Multisentinet: A deep semantic network for multimodal sentiment analysis. pp. 2399–2402 (11 2017). https://doi.org/10.1145/3132847.3133142
25. Yu, J., Jiang, J.: Adapting bert for target-oriented multimodal sentiment classification. In: International Joint Conference on Artificial Intelligence (2019), https://api.semanticscholar.org/CorpusID:199465957
26. Yu, Y., Lin, H., Meng, J., Zhao, Z.: Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms **9**, 41 (06 2016). https://doi.org/10.3390/a9020041