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[1,3], and Ali Braytee[2]

[1] The University of Sydney, School of Computer Science, Camperdown, Australia
[2] University of Technology Sydney, School of Computer Science, Ultimo, Australia
[3] University of Technology Sydney, TD School, Ultimo, Australia

Abstract. This paper proposes a 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

The enhancement of unimodal statistical 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 has 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 visual features than CNNs.

A good latent representation greatly influences the sentiment analysis results and provides an excellent foundation for the following fusion phase. Modal fusion is one of the main challenges in multimodal constructions. A single concatenation method [16] can effectively integrate the multimodal information. Huang et al. [9] proposed Deep Multimodal Attentive Fusion, which uses the internal correlation between visual and textual features for sentiment analysis. They also

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

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. [19] used the attention technique on the BERT model to deal with target-oriented tasks. Tsai et al. [15] proposed a cross-attention technique to combine the different latent representations effectively, enhancing the performance of each single modality. Recently, Lee et al. [11] 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. In this paper, we use two advanced unimodal models: BERT [6] and DINOv2 [13] and then propose three fusion methods to create a multimodal sentiment analysis model combining the extracted features from text and image modalities, named Basic Fusion, Self-Attention Fusion, and Dual-Attention Fusion. The main contributions are as follows:

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

2 Method

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 predict sentiment analysis. The overall framework is shown in Fig. 1.

2.1 Textual Features

We utilize a BERT layer [6] 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. For the final textual latent representation t, we use the transformed CLS embedding t_{CLS}, which represents a condensed view of the entire input sequence. This approach leverages the CLS token's embedding directly after merging all the necessary information for sentiment analysis following its transformation. The representation is then used as part of the fusion process.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

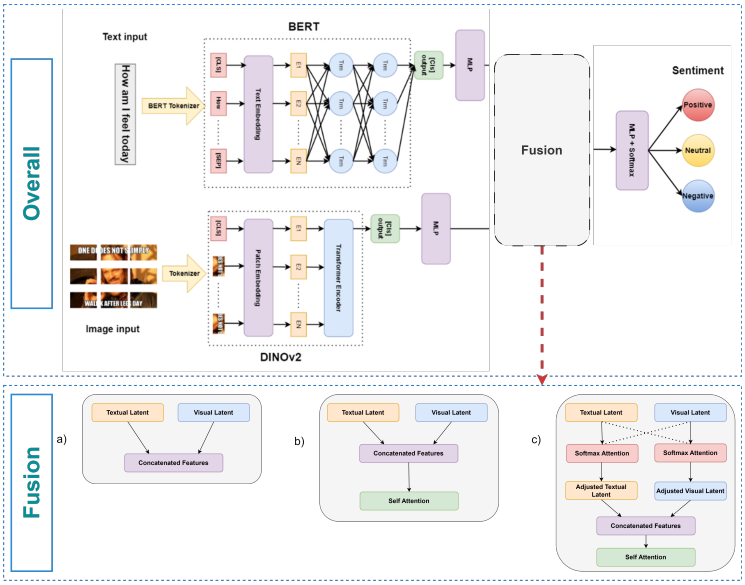

Fig. 1: The overall architecture of the proposed framework (Above). The fusion methodology of the framework (Below): a) Basic Fusion Model; b) Self-Attention Fusion Model; c) Dual-Attention Fusion Model

2.2 Visual Features

In the architecture, DINOv2 [13] serves as the foundational layer for extracting complex features directly from image inputs. Each input image is first divided into a sequence of patches, which are then linearly embedded into tokens and processed through the DINOv2 transformer network. 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. For the final visual latent representation v, we use the transformed global feature embedding v_{CLS}, which represents a condensed view of the entire image. 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

We propose three fusion methods. A Basic Fusion Model that concatenates the textual and visual latent representation (Fig. 1a). 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. 1

$$c = [t, v] \tag{1}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

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 (Fig. 1b). The output s of the self-attention mechanism is defined in Eq. 2 as,

$$s = \text{Softmax}\left(\frac{cW_Q(cW_K)^T}{\sqrt{d_k}}\right)cW_V \tag{2}$$

where c is the concatenated vector given from Eq. 1. 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 (Fig. 1c). 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. 3,

$$\begin{aligned} Q_t &= tW_{Q_t}, \quad K_t = tW_{K_t}, \quad V_t = tW_{V_t}, \\ Q_v &= vW_{Q_v}, \quad K_v = vW_{K_v}, \quad V_v = vW_{V_v}. \end{aligned} \tag{3}$$

Next, we apply softmax attention for cross-modal adjustments as shown in Eq. 4,

$$t' = \text{Softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right)V_v \quad v' = \text{Softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right)V_t \tag{4}$$

Finally, we apply concatenation and self-attention as defined in Eq. 5,

$$s' = \text{Softmax}\left(\frac{c'W_Q(c'W_K)^T}{\sqrt{d_k}}\right)c'W_V \tag{5}$$

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

3 Experiments

3.1 Datasets

Memotion 7k Dataset [14] consists of 6992 samples, each paired with an image and the corresponding caption, representing a complete 'meme'. The dataset is part of a challenge that includes three subtasks: analyzing memes for sentiment, which can be positive, negative, or neutral. MVSA Datasets [12] 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 the Memotion 7k Dataset, MVSA datasets have three classes: positive, neutral, and negative.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

3.2 Experiment settings

For the Memotion dataset, we employed macro F1 as our evaluation metric. For MVSA-single and MVSA-multi, we used accuracy and F1-score. We utilized focal loss as the loss function and adjusted the γ parameter (γ=2, 3, 4) to ensure the model adequately focuses on the minority classes. 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

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) [18], which leads on the MVSA-multi dataset with an accuracy of 0.68 and an F1 score of 0.68.

Table 1: Results on MVSA-Single and MVSA-Multi datasets

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

This approach outperforms the previously established benchmarks, including SentiBank [2], CNN-Multi [4], DNN-LR models [20], and HSAN [17], and even the advanced MultiSentiNet [18] and Dual-Pipeline [3] models. The BERT and DINOv2 model with additional self-attention mechanisms also shows strong performance, 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

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

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 shown in Table 2, the proposed models are compared to the state-of-the-art methods [1, 5, 8, 10, 14]. Dual-Attention model achieves the highest Macro F1 score of 0.3552, surpassing both the competition baseline of 0.2176 and several state-of-the-art methods. Notably, it outperforms strong multimodal approaches such as Vkeswani IITK [10], which has a Macro F1 of 0.3546, Guoym [8] with 0.3519, and Aihaihara [14], which has a Macro F1 of 0.3501, all of which incorporate sophisticated combinations of textual and visual features using advanced transformers or ensemble strategies. Compared to our Basic Fusion and Self-Attention variants, the Dual-Attention framework delivers a marked improvement, demonstrating the benefit of simultaneously modeling both cross-modal interactions and intra-modal salience. Furthermore, although several top-performing systems employ powerful feature extractors such as BERT, ResNet, or VGG-16, their relatively close performance suggests diminishing returns from architecture complexity alone. Our results highlight that refined fusion strategies, rather than just deeper encoders, can drive meaningful performance gains in multimodal sentiment classification.

Table 2: Results for the compared methods on Memotion 7k

Model	Macro F1
Competition Baseline [14]	0.2176
Vkeswani IITK [10]	0.3546
Guoym [8]	0.3519
Aihaihara [14]	0.3501
Sourya Diptadas [5]	0.3488
MemoSYS [1]	0.3475
Ours (Basic Fusion)	0.3237
Ours (Self-Attention)	0.3436
Ours (Dual-Attention)	**0.3552**

Example 1

Example 2

Fig. 2: Misclassified memes

To highlight the limitations of our proposed method, we selected two misclassified images from the testing set. In both of the memes, the individuals are smiling. This likely made the model perceive them as expressing positive emotions (Fig. 2). However, another possible reason for the model's incorrect prediction could be issues with the dataset's annotation. As shown in Example 1 in Fig. 2, 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

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

'negative' class because his smile is dry. Similarly, as shown in Example 2, the individual is giving a thumbs up with a wide smile, which can easily be interpreted as a display of positive emotion. However, the humor and sarcasm embedded in the meme's text "Liam Approves!" may suggest a satirical or ironic context rather than a genuinely positive emotional state. This highlights the limitations of models that do not integrate textual cues effectively. 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 proposed multimodal sentiment analysis framework is built on robust unimodal encoders, BERT for text and DINOv2 for images, followed by fusion through three hierarchical strategies: Basic Fusion, Self-Attention Fusion, and Dual-Attention Fusion. Each fusion mechanism is designed to progressively enhance the model's capacity to capture intra- and inter-modal relationships. The results indicated that our fusion methods are highly adept at integrating mutual information across multiple modalities. In the future, we will explore dynamic fusion strategies such as gating mechanisms, transformer-based cross-modal attention, and graph-based modality alignment.

Acknowledgment

We acknowledge the contributions of Yufan Lin to this project.

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)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25

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. 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)
12. Niu, T., Zhu, S., Pang, L., El-Saddik, A.: Sentiment analysis on multi-view social data. In: MultiMedia Modeling. p. 15–27 (2016)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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
18. 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
19. 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
20. 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

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_25