ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Emotion Recognition in Multimodal Social Data

Lucia Passaro[1], Davide Amadei[1] and Davide Bacciu[1] *

1- Department of Computer Science, University of Pisa

Abstract. Emotion recognition on social media is often approached in unimodal or single-label settings, despite the multimodal nature of online communication. This paper presents a study of multilabel emotion recognition from paired text–image data. We evaluate vision–language encoders and compare them with strong unimodal baselines and a zero-shot multimodal LLM. A simple multimodal classifier built on CLIP achieves the most reliable performance. Data-centric additions such as emoji transcription, caption augmentation, and pseudo-labelling offer limited gains, whereas calibrated decision thresholds have a consistent effect. The results highlight the value of visual cues and show limitations of recent VLMs.

1 Introduction

Social media is a central medium for everyday communication, where users convey affect through short text, images, emojis and mixed image–text posts. Automated analysis of such content supports applications in opinion monitoring, recommendation, and social listening. While sentiment analysis focuses on polarity [13], Emotion Recognition (ER) targets finer affective categories defined in established taxonomies. Most research in ER and Multimodal ER (MER) has concentrated on controlled audiovisual data [6, 12], whereas Multimodal Sentiment Analysis has progressed rapidly with transformer-based architectures [11, 16]. In contrast, ER on image–text posts remains largely unexamined; only limited work has considered this setting [14], and none has evaluated modern Vision–Language Models (VLMs) on real social-media data.

A second limitation is the reliance on single-label annotation, despite well-established evidence that emotions are subjective, ambiguous, and co-occuring. Multilabel ER has been explored primarily in textual settings [1, 2], with isolated studies in other modalities [3] and minimal attention to multimodal content.

Recent advances in vision–language modelling provide new opportunities. Contrastive VLMs such as CLIP, SigLIP and Jina CLIP [10, 15, 7] learn aligned image–text representations and show strong zero-shot transfer. Generative approaches, including BLIP-2 [8] and Vision–Language Large Language Models (VLLMs) such as LLaVA [9] further integrate visual features through lightweight adaptors or direct LLM conditioning. Their effectiveness for multilabel MER, however, has not been examined.

This paper presents, to the best of our knowledge, the first evaluation of multilabel MER on real-world social-media image–text posts using contemporary VLMs and VLLMs. Using the MM-Emo dataset,[1] we benchmark CLIP, SigLIP,

*Work partially supported by the EIC project EMERGE (Grant No. 101070918).

[1] `https://huggingface.co/datasets/luciacpassaro/MM-Emo`

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

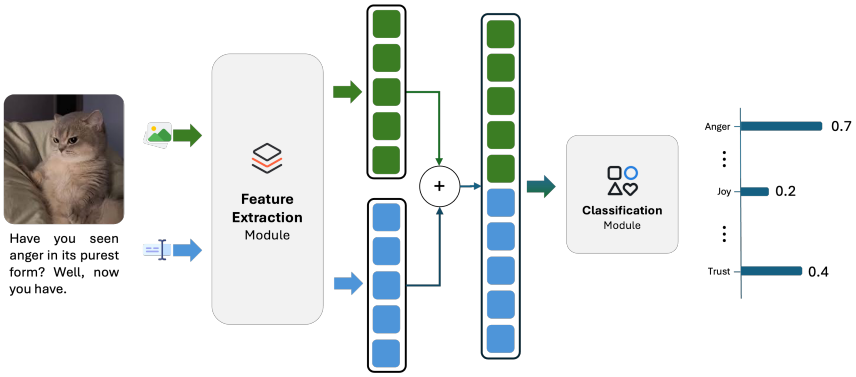

Fig. 1: Proposed model framework.

BLIP-2 and related encoders in a unified framework (Fig. 1); analyse unimodal text and image variants; and examine zero-shot MER with LLaVA.

The study is guided by three research questions: **RQ1 -** *how effectively do modern VLMs handle multilabel MER in realistic multimodal settings, and what do the individual modalities contribute?* **RQ2 -** *to what extent do data-centric strategies and fine-tuning configurations influence performance?* **RQ3 -** *how do supervised multimodal models compare with modern zero-shot VLLMs?*

To address these questions, we evaluate controlled multimodal and unimodal baselines using consistent training objectives and thresholding strategies. Our analysis covers both model-level factors—encoder selection, partial fine-tuning—and data-level factors. We additionally benchmark LLaVA in a zero-shot setting to provide a contemporary reference point.

2 Method

Task. We address multilabel ER from paired image–text posts. Each instance consists of an image I, a text segment T, and one or more labels. The model outputs a vector $\hat{y} \in [0,1]^K$ for $K = 9$ (i.e., 8 emotions + 1 neutral class); labels are predicted via class-specific thresholds.

Multimodal architecture. As illustrated in Figure 1, the framework adopts an early-fusion dual-encoder pipeline. The *feature extraction module* consists of a visual encoder f_θ and a text encoder g_ϕ, which map the input image and text to embeddings z_I and z_T:

$$z_I = f_\theta(I), \qquad z_T = g_\phi(T).$$

The central block of Figure 1 shows how these embeddings are *concatenated* into a joint representation

$$z = [z_I \| z_T],$$

which the *classification module* transforms into multilabel sigmoid outputs, as illustrated on the right of the figure. Both encoders may be kept frozen or fine-

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

tuned jointly with the classifier, following the configurations evaluated in our experiments.

Label space. The classifier predicts the Plutchik emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*) plus a *neutral* class. The multilabel setup captures the richness and co-occurrence typical of social-media content.

Auxiliary modifications. We evaluate several input- or data-level variants without altering the core architecture: (i) *Emoji text processing*: replaces emojis with short textual descriptors to retain affective cues; (ii) *Caption augmentation*: a pretrained captioning model generates a short description of I, appended to T as additional context; (iii) *Pseudo-labelling*: high-confidence predictions supplement gold labels for low-support emotions; (iv) *Partial fine-tuning*: only the classifier and projection layers are trained, reducing overfitting risk.

Unimodal variants. To isolate modality contributions, we ablate the architecture by removing one encoder. The text-only model uses g_ϕ, and the image-only model uses f_θ; both preserve the same classifier.

Zero-shot evaluation. A VLLM receives (I, T) and a prompt listing the emotions and is asked to select applicable labels. This provides a reference point for the performance attainable without supervision or task-specific parameters.

3 Experimental Setting

Dataset. We use the gold-labelled MM-Emo dataset with multilabel annotations. It consists of 900 tweets crowdsourced for multimodal gold labels (five annotators per tweet, multiple emotions allowed) following the Plutchik's taxonomy plus *Neutral*, and are binarised via score thresholding. Text is normalised by removing URLs, mentions, and markup; emojis are either removed or replaced with textual descriptors. Images are centre-cropped and resized according to each encoder's requirements. Caption-augmented variants append captions generated by a pretrained captioning model [8].

Model variants. We evaluate CLIP, SigLIP, Jina CLIP, and BLIP-2 [10, 15, 7, 8] via HuggingFace in an early-fusion dual-encoder setup. Unimodal baselines use text encoders (BERT [4], CLIP-text) or vision encoders (CLIP-ViT, ViT-B/16 [5]). Zero-shot performance is assessed using LLaVA [9] with either per-emotion binary prompts or a single list-style prompt; the best-performing prompt on the validation set is used for testing. We also test encoder freezing, emoji preprocessing, caption augmentation (Caption-CLIP), their combination, and pseudo-labelling (Silver-CLIP) for low-support emotions, all using the same search space.

Training and optimisation. Hyperparameters are selected with Optuna. The search spans batch sizes $8, 16, 32$, learning rates 10^{-5}–10^{-2}, dropout $[0, 0.5]$, 1–3 MLP layers, hidden sizes $128, 256, 512$, and warmup steps $0 \ldots 200$. Models use AdamW with a linear warmup–decay schedule. Encoder variants receive 200 trials, and the top models are refined with 300 trials. Class-specific sigmoid thresholds are further tuned with an additional Optuna sweep (500 trials) on training predictions and selected based on development performance.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

4 Results

Emotion	Support	CLIP Base	Jina CLIP	CLIP Large	SigLIP	BLIP-2
Anger	9.84%	0.62	**0.64**	0.59	0.44	0.12
Anticipation	13.47%	0.37	**0.44**	0.36	0.32	0.12
Disgust	8.29%	0.61	**0.60**	0.57	0.53	0.10
Fear	3.89%	0.14	**0.17**	0.16	0.09	0.00
Joy	28.50%	***0.73***	*0.70*	*0.72*	*0.74*	*0.68*
Neutral	15.80%	**0.51**	0.49	0.48	0.24	0.09
Sadness	6.74%	**0.57**	0.54	0.46	0.37	0.05
Surprise	7.51%	0.02	**0.04**	0.03	0.00	0.02
Trust	5.96%	0.08	**0.12**	0.08	0.00	0.00
Macro avg	386	**0.41**	**0.41**	0.38	0.30	0.13
Micro avg	386	**0.55**	0.53	0.52	0.48	0.39

Table 1: Encoder performance across emotion classes. Reported values correspond to the average of each model's top 10 trials, ranked by F1-score.

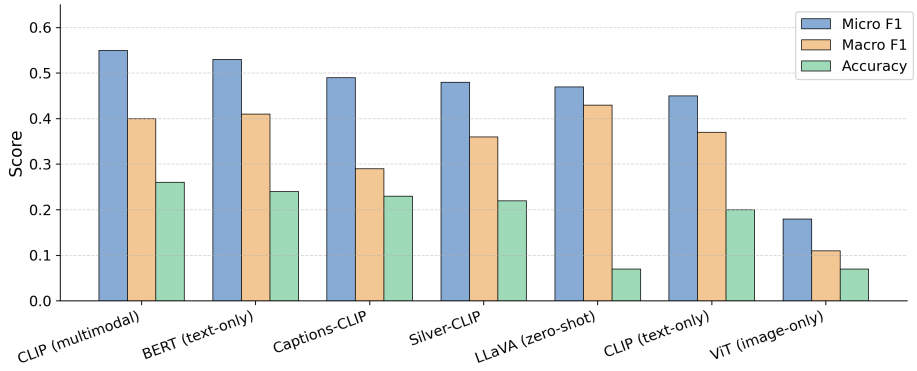

Fig. 2: Comparison of model performance across Micro F1, Macro F1, and Accuracy (9 classes). Models sorted by Micro F1 in descending order.

We report test results for multimodal models, unimodal baselines and zero-shot prompting. All trained systems use class-specific threshold optimisation. The analysis below follows the research questions in the Introduction. Beyond aggregate F1 scores, per-class trends show that encoder choice interacts with label frequency. High-support emotions (e.g., *joy*, *neutral*) remain stable across models, while low-support categories vary widely and are more sensitive to representation quality. CLIP-based models consistently outperform others on these minority classes, whereas BLIP-2 and SigLIP often default to negative predictions despite threshold calibration. Errors frequently occur when text and image

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

cues conflict; in such cases, multimodal CLIP is more robust than text-only baselines, particularly for emotions with visually salient signals.

VLM performance and modality contributions (RQ1). Table 1 compares the main multimodal encoders. Across all metrics, CLIP variants significantly outperform SigLIP and BLIP-2, with CLIP base achieving the highest micro and macroF1. This indicates that contrastively trained dual encoders remain better suited to MER than generative or instruction-tuned architectures in this low-resource setting. Figure 2 further disentangles modality contributions. A strong text-only BERT baseline performs competitively, but multimodal CLIP achieves the best micro and weighted F1 after threshold optimisation. CLIP's text-only branch underperforms its multimodal counterpart, and image-only models perform poorly. This confirms that textual cues dominate performance in MER, but visual features provide complementary information.

Effect of data-centric and training choices (RQ2). Data enrichment methods provide limited benefits. Caption-based augmentation and pseudo-labelling offer minor validation gains but do not transfer to the test set, likely due to noise in auxiliary labels. Emoji preprocessing and encoder freezing have negligible impact. Overall, performance is more sensitive to encoder selection and thresholding than to modest data-centric interventions.

Comparison with zero-shot VLLMs (RQ3). Zero-shot LLaVA achieves competitive macro recall but substantially lower precision and accuracy compared to supervised CLIP. Even with prompt optimisation, LLaVA fails to match the multimodal baseline, suggesting that current VLLMs remain unreliable for fine-grained multilabel MER, where calibrated decision boundaries and modality-specific alignment are crucial.

5 Conclusion and Future Directions

This paper presented the first study of multilabel Emotion Recognition on social-media image–text posts using Vision–Language Models. Across multimodal, unimodal, and zero-shot settings, contrastively trained CLIP encoders provided the most reliable performance, outperforming larger generative and instruction-tuned architectures. Our findings highlight three main points. First, text remains the dominant signal, yet visual cues provide complementary information that improves predictions for several emotions. Second, model scale is not the main bottleneck in this low-resource scenario: larger encoders, caption-based augmentation and pseudo-labelling provides limited or inconsistent gains. Third, zero-shot VLLMs achieve high recall but lack precision, showing that supervised multimodal training with calibrated thresholds remains more dependable.

Future work should focus on expanding and rebalancing multimodal emotion datasets, exploring richer fusion mechanisms, and adapting emerging VLLMs to few-shot or instruction-driven multilabel settings. We also aim to better address underrepresented and visually ambiguous emotions by using training objectives that handle multilabel imbalance more effectively and by developing models that capture finer cross-modal interactions. Finally, evaluating systems on data that

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

evolves over time or varies across platforms will help assess their robustness under real-world shifts in language, visual content and user behaviour.

References

[1] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. F. Gelbukh. Multi-label emotion classification in texts using transfer learning. *Expert Syst. Appl.*, 213(Part):118534, 2023.

[2] I. Ameer, N. Bölücü, G. Sidorov, and B. Can. Emotion classification in texts over graph neural networks: Semantic representation is better than syntactic. *IEEE Access*, 11:56921–56934, 2023.

[3] I. Bendjoudi, F. Vanderhaegen, D. Hamad, and F. Dornaika. Multi-label, multi-task CNN approach for context-based emotion recognition. *Inf. Fusion*, 76:422–428, 2021.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019*, pages 4171–4186, 2019.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[6] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya. Emotion recognition and artificial intelligence: A systematic review (2014-2023) and research recommendations. *Inf. Fusion*, 102:102019, 2024.

[7] A. Koukounas, G. Mastrapas, M. Günther, B. Wang, S. Martens, I. Mohr, S. Sturua, M. K. Akram, J. F. Martínez, S. Ognawala, S. Guzman, M. Werk, N. Wang, and H. Xiao. Jina clip: Your clip model is also your text retriever, 2024.

[8] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[11] U. Singh, K. Abhishek, and H. K. Azad. A survey of cutting-edge multimodal sentiment analysis. *ACM Comput. Surv.*, 56(9):227:1–227:38, 2024.

[12] G. Vijayaraghavan, M. T., P. D., and U. E. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Inf. Fusion*, 105:102218, 2024.

[13] M. Wankhade, A. C. S. Rao, and C. Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.*, 55(7):5731–5780, 2022.

[14] X. Yang, S. Feng, D. Wang, and Y. Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Trans. Multim.*, 23:4014–4026, 2021.

[15] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE, 2023.

[16] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian. Multimodal sentiment analysis with image-text interaction network. *IEEE Trans. Multim.*, 25:3375–3385, 2023.