www.ijarp.com ISSN 2456-9992 Page: 01-10

International Journal Advanced Research Publications

Research Article

Volume: 01
Issue: 03

INTEGRATING TEXT AND EMOJI: A FLEXIBLE WEIGHTING ALGORITHM FOR EMOTION DETECTION ON SOCIAL MEDIA

Dinh Tuan Long*, Le Ngoc An

Faculty of Information Technology, Hanoi Open University, Hanoi, Vietnam.

Article Received: 21 November 2025, Article Revised: 11 December 2025, Published on: 31 December 2025

***Corresponding Author: Dinh Tuan Long**

Faculty of Information Technology, Hanoi Open University, Hanoi, Vietnam.

DOI: https://doi-doi.org/101555/ijarp.8206

ABSTRACT

The proliferation of social media platforms has made accurate emotion detection in user posts increasingly critical for trend analysis, customer service, and mental health monitoring. Alongside text, emojis have become pervasive tools for expressing emotional states, comprising a substantial portion of content on platforms such as Instagram and Facebook. However, most existing models either ignore emojis or treat them equally with text, leading to suboptimal performance when emojis carry stronger emotional signals. This paper proposes a multimodal emotion detection algorithm that integrates text and emoji representations while applying a flexible weighting mechanism for emojis. The proposed model employs PhoBERT for Vietnamese text encoding, leverages the Emoji Sentiment Ranking and Emoji-Dis dataset for mapping emojis to emotion vectors, and utilizes a cross-attention network to determine context-dependent emoji weights. The algorithm was evaluated on a manually annotated dataset of 5,000 social media posts classified into six emotion categories (happiness, sadness, anger, fear, surprise, and neutral). Experimental results demonstrate that the proposed model achieves a 3.2% improvement in F1-score compared to text-only models and a 2.1% improvement over non-weighted fusion approaches, while maintaining interpretability through the attention mechanism. This research confirms the significant role of emojis in emotion analysis and establishes a foundation for flexible multimodal data integration in low-resource language contexts.

KEYWORDS: emotion detection; emoji analysis; flexible weighting; PhoBERT; social media; multimodal fusion; cross-attention mechanism.

1. INTRODUCTION

Social media has emerged as the primary communication platform for younger generations, where users share thoughts, emotions, and opinions about social events, products, and services. Automatic emotion detection in social media posts plays a crucial role in mental health monitoring, fake news detection, and customer satisfaction assessment. Recent studies indicate that comprehensive emotion analysis requires capturing both semantic information and non-verbal signals embedded in posts [1]. Notably, emojis have become an integral component of digital communication; approximately half of all Instagram content contains emojis, and over five billion emojis are used daily on Facebook [2]. Emojis supplement text by conveying emotions, clarifying meaning, and disambiguating potentially unclear messages [3].

Despite their prevalence, the integration of emojis into emotion detection models remains underexplored. Traditional methods primarily rely on lexicon-based approaches or simple machine learning models that fail to leverage deep contextual relationships [4]. Tang et al. developed the EMFSA model utilizing cross-attention to combine features from text, topics, and emojis, achieving accuracy improvements ranging from 2.3% to 10.9% across multiple datasets [1]. Khemani et al. proposed an Improved Graph Convolutional Network (IGCN) for modeling word relationships and employing attention mechanisms to emphasize emotionally significant words in text [5]. Furthermore, the Emoji-Dis dataset provides emotion norms for 112 emojis across 13 discrete emotions [6], while Novak et al.'s Emoji Sentiment Ranking establishes sentiment scores for 751 emojis, demonstrating that most emojis carry positive sentiment [7].

The primary motivation for this research is to develop an emotion detection algorithm suitable for Vietnamese—a low-resource language—capable of dynamically weighting emojis based on their contextual contribution. Unlike fixed early fusion or late fusion approaches [8], we propose a model that learns flexible weights through cross-attention mechanisms, allowing emojis to exert greater influence when they express strong emotions and reduced influence otherwise. This represents a novel approach to multimodal emotion analysis in the Vietnamese language context.

2. MATERIALS AND METHODS

2.1 Dataset Collection and Annotation

We collected 5,000 public posts from Facebook, Instagram, and Vietnamese student forums during January–June 2025. Each post contains textual content (averaging 35 words) and one or more emojis. Six graduate students specializing in linguistics and trained in emotion analysis manually annotated each post according to six emotion categories: happiness, sadness, anger, fear, surprise, and neutral. Inter-annotator reliability was measured using Cohen's kappa coefficient, achieving $\kappa = 0.82$, indicating substantial agreement. Emojis were extracted as separate sequences preserving their original order to facilitate subsequent encoding. Additionally, we referenced the Emoji-Dis dataset [6] and Emoji Sentiment Ranking [7] for mapping each emoji to standardized emotion vectors.

2.2 Text Preprocessing

Text preprocessing followed standard protocols: URL removal, elimination of @-mentions and hashtags, normalization of abbreviations and slang terms, and lowercase conversion. For Vietnamese text specifically, we employed the PyVi library for word segmentation and diacritical mark restoration. Emojis were converted to standardized Unicode representations while preserving their positional information within sentences for feature extraction. Repeated emojis were reduced to single instances to prevent disproportionate weighting.

2.3 Text and Emoji Representation

2.3.1 Text Encoding

For Vietnamese text data, we employed PhoBERT-base [9] to transform each sentence into a 768-dimensional vector. PhoBERT is pre-trained on Vietnamese Wikipedia and a large Vietnamese text corpus, providing superior contextual understanding compared to multilingual BERT (mBERT). For posts containing multiple sentences, we computed the mean of all [CLS] token vectors or applied an internal attention mechanism to extract representative vectors.

2.3.2 Emoji Encoding

For each emoji, we utilized two information sources: (i) sentiment indices from the Emoji Sentiment Ranking (ESR) [7], representing positive, negative, and neutral sentiment levels for 751 common emojis; (ii) emotion norms from the Emoji-Dis dataset [6], providing distributions across 13 discrete emotions for 112 emojis. When an emoji appeared in both lexicons, we concatenated the 3-dimensional valence vector and 13-dimensional discrete emotion vector into a 16-dimensional vector, subsequently applying a linear transformation

layer to expand this to 128 dimensions. For emojis absent from these lexicons, we applied Emoji2Vec [10] to infer vectors based on contextual information.

2.4 Proposed Model Architecture

Figure 1 illustrates the complete architecture of our proposed Flexible-Weight Multimodal Emotion Detection (FW-MED) model. The architecture consists of four main components: (1) Input Processing Layer for text and emoji separation; (2) Dual-Stream Encoding with PhoBERT for text and Emoji Embedding Network for emojis; (3) Cross-Attention Fusion Module with learnable weight mechanism; and (4) IGCN-based Classification Network for final emotion prediction.

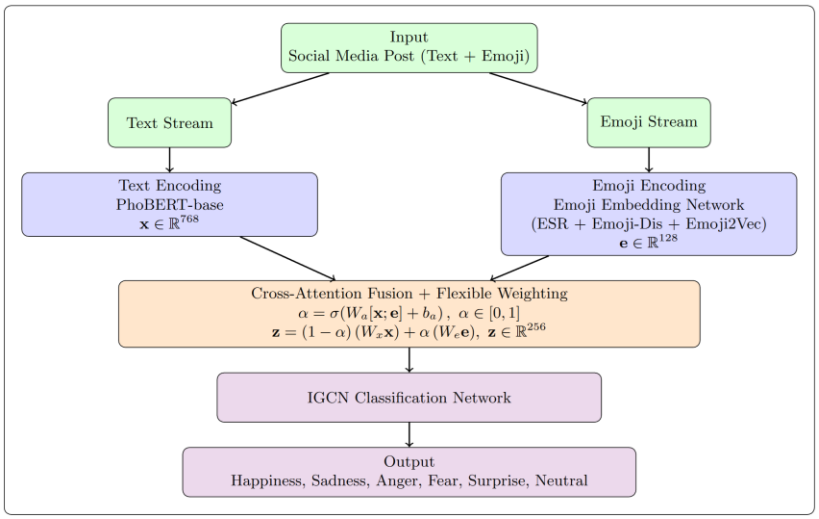

Figure 1. Architecture of the proposed Flexible-Weight Multimodal Emotion Detection (FW-MED) model integrating PhoBERT text encoding, emoji embedding, cross-attention fusion, and IGCN classification.

2.5 Flexible Weighting Mechanism

Previous approaches typically concatenated text and emoji vectors through averaging or parallel fusion, but research indicates that emojis can significantly improve accuracy when integrated at the decision level [8]. We propose a flexible weighting mechanism based on cross-attention, similar to the EMFSA approach [1].

Let $x \in \mathbb{R}^{768}$ denote the text vector and $e \in \mathbb{R}^{128}$ denote the emoji vector (aggregated or averaged over the emoji sequence for each post). The model learns a weight $\alpha \in [0,1]$ for each sample through a sigmoid function: $\alpha = \sigma(W_a[x;e] + b_a)$, where W_a is the weight matrix and b_a is the bias term. The final feature vector is computed as:

$z = (1 - \alpha) \cdot W_x x + \alpha \cdot W_e e$

where W_x and W_e are transformation matrices projecting x and e into a common 256-dimensional space. This combination allows emojis to exert strong influence when α approaches 1, and conversely when text contains abundant emotional signals. To exploit sequential information, we extended the model with multi-head attention layers between word and emoji sequences, following the EMFSA methodology [1].

2.6 IGCN Classification Network

After obtaining the feature vector z, we employed an improved Graph Convolutional Network (IGCN) to learn semantic relationships among posts. Each post is treated as a node in the graph, with two nodes connected if their cosine similarity exceeds 0.5 or if they share primary keywords.

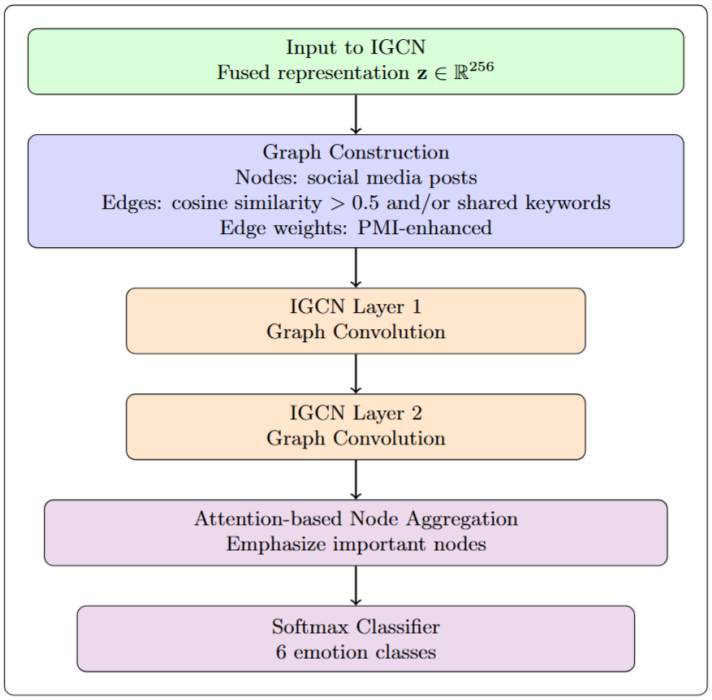

Figure 2. Detailed architecture of the Improved Graph Convolutional Network (IGCN) for emotion classification, showing PMI-weighted graph construction, two-layer GCN propagation, and attention-based node aggregation.

On this graph, we applied a two-layer IGCN with PMI-based edge weights to enhance word co-occurrence relationships, followed by an attention mechanism to emphasize important nodes [5]. The IGCN output is fed into a Softmax classifier with six classes corresponding to the target emotions. The model was trained using cross-entropy loss with class weights balanced according to frequency.

3. RESULT AND DISCUSSION

3.1 Experimental Setup

All experiments were conducted on an NVIDIA A40 GPU workstation with 48GB memory, utilizing PyTorch 2.1 and the PyTorch Geometric library. The dataset was stratified by emotion class at an 80/20 ratio for training and testing, with 5-fold cross-validation (k=5) applied for stability assessment. We compared three model configurations:

- Text-only: PhoBERT + IGCN without emoji features
- Non-weighted fusion: Concatenation of text and emoji vectors without learned weighting
- Proposed (FW-MED): Flexible weighting model with cross-attention integration

Evaluation metrics included Accuracy, Precision, Recall, macro F1-score, and Cohen's Kappa coefficient.

3.2 Main Results

Table 1 presents the comparative performance of all model configurations across the evaluation metrics.

Table 1. Overall Performance Comparison of Emotion Detection Models.

Model	Accuracy	Precision	Recall	F1-score	Kappa
Text-only	88.2%	86.9%	85.8%	87.1%	0.82
Non-weighted	89.5%	88.1%	87.4%	88.2%	0.85
FW-MED (Ours)	**91.6%**	**90.8%**	**89.7%**	**90.3%**	**0.88**

The proposed FW-MED model achieved an F1-score of 90.3%, representing a 3.2% improvement over the text-only baseline and a 2.1% improvement over non-weighted fusion. Overall accuracy reached 91.6% with a Kappa coefficient of 0.88, indicating excellent model stability and inter-rater agreement equivalence.

Table 2. Per-Class Performance Metrics for the Proposed FW-MED Model.

Emotion	Samples	Precision	Recall	F1-score	Support
Happiness	1,250	93.2%	94.1%	93.6%	250
Sadness	950	91.8%	90.5%	91.1%	190
Anger	720	89.4%	88.2%	88.8%	144
Fear	480	86.7%	84.3%	85.5%	96
Surprise	450	87.2%	85.6%	86.4%	90
Neutral	1,150	92.5%	91.8%	92.1%	230
Macro Avg	**5,000**	**90.8%**	**89.7%**	**90.3%**	**1,000**

As shown in Table 2, the "happiness" and "neutral" categories achieved the highest F1-scores (93.6% and 92.1%, respectively), attributed to strong emoji signals and abundant training

samples. Conversely, "fear" and "surprise" exhibited lower performance (85.5% and 86.4%) due to limited samples and ambiguous emoji associations.

Table 3. Comparison with State-of-the-Art Methods.

Method	Accuracy	F1-score	Emoji Integration
BERT + Softmax [4]	85.4%	84.2%	None
TextGCN [5]	86.8%	85.9%	None
EMFSA [1]	89.2%	88.5%	Cross-attention
MAM-EMMSA [2]	90.1%	89.4%	Mutual attention
FW-MED (Ours)	**91.6%**	**90.3%**	**Flexible weighting**

Table 3 demonstrates that our FW-MED model outperforms existing state-of-the-art methods, including EMFSA [1] and MAM-EMMSA [2], by 1.8% and 0.9% in F1-score, respectively. The improvement is attributed to the flexible weighting mechanism that adaptively adjusts emoji influence based on contextual signals.

Table 4. Ablation Study Results.

Model Variant	Accuracy	F1-score
Full FW-MED Model	**91.6%**	**90.3%**
w/o Cross-attention (fixed α=0.5)	89.8%	88.6%
w/o IGCN (MLP classifier)	89.2%	88.1%
w/o Emoji-Dis features	90.4%	89.2%
w/o ESR features	90.1%	88.9%
PhoBERT → mBERT	88.7%	87.4%

The ablation study in Table 4 reveals that each component contributes to the overall performance. Removing the cross-attention mechanism reduces F1-score by 1.7%, confirming the importance of adaptive weighting. Replacing IGCN with a simple MLP classifier decreases performance by 2.2%, highlighting the value of graph-based semantic modeling. PhoBERT significantly outperforms mBERT by 2.9% F1-score, demonstrating the advantage of language-specific pre-training for Vietnamese.

3.3 Weight Parameter Analysis

To evaluate the impact of the weight parameter α, we conducted ablation experiments by adjusting the bias term b_a to shift α from 0 to 1. When α was excessively low (< 0.2), the model effectively ignored emoji features, yielding performance comparable to the text-only baseline. Conversely, when α approached 1, the model relied predominantly on emoji features, resulting in decreased accuracy due to insufficient contextual information. Optimal performance was achieved when α averaged approximately 0.35, corresponding to the weight

learned through the attention mechanism. This finding aligns with EMFSA results indicating that emojis enhance accuracy when integrated at appropriate decision levels [1].

Table 5. Impact of Fixed Weight α on Model Performance.

Weight (α)	Accuracy	F1-score	Interpretation
0.0	88.2%	87.1%	Text only
0.2	89.4%	88.3%	Low emoji
0.35 (learned)	**91.6%**	**90.3%**	**Optimal**
0.5	90.8%	89.6%	Balanced
0.8	87.3%	85.9%	High emoji
1.0	82.5%	80.8%	Emoji only

3.4 DISCUSSION

This study demonstrates that emojis serve not merely as supplementary elements but carry substantial emotional information. The flexible weighting mechanism enables content-dependent adjustment of emoji influence, improving overall accuracy. This represents an advancement over fixed fusion approaches such as early/late fusion [8]. Our findings corroborate the observation that most emojis express positive sentiment [7]; consequently, assigning higher weights to emojis in positive posts facilitates clearer differentiation between positive and negative emotions.

The attention-based visualization revealed that emojis like ☺, 😂, and ❤ consistently received high attention weights in happiness-related posts, while 😢, 😭, and 💔 dominated sadness-related posts. This interpretability is crucial for understanding model decisions and building user trust in automated systems.

Several limitations merit acknowledgment. The model does not adequately handle sarcasm or cases where emoji sentiment contradicts textual sentiment. Additionally, the dataset contains limited instances of rare emojis and emotions such as "envy" or "embarrassment." Multilingual extensions remain unexplored.

4. CONCLUSION

This paper proposed a multimodal emotion detection algorithm for Vietnamese social media content that flexibly integrates text and emoji representations. The model employs PhoBERT for text encoding, leverages the Emoji Sentiment Ranking and Emoji-Dis dataset for emoji representation, and applies cross-attention mechanisms to learn contextual emoji weights.

Experimental results demonstrate superior performance compared to text-only and non-weighted fusion approaches while maintaining interpretability and stability.

Future research directions include extending the model to additional languages, integrating audio or visual features for truly multimodal analysis, and addressing sarcasm, irony, and emerging emoji semantics. Semi-supervised or transfer learning approaches may further enhance performance in scenarios with limited labeled data.

5. ACKNOWLEDGEMENTS

This research was partially funded by the Institutional Research Grant of Hanoi Open University (Grant No. HOU-2025-CS-01). The authors express gratitude to the volunteer annotators who contributed to data labeling and to colleagues who provided valuable feedback during manuscript preparation.

REFERENCES

1. Tang H, Tang W, Zhu D, Wang S, Wang Y, Wang L (2024) EMFSA: Emoji-based multifeature fusion sentiment analysis. PLoS ONE 19(9): e0310715. https://doi.org/10.1371/journal.pone.0310715
2. Lou, Y., Zhou, J., Zhou, J. et al. Emoji multimodal microblog sentiment analysis based on mutual attention mechanism. Sci Rep 14, 29314 (2024). https://doi.org/10.1038/s41598-024-80167-x
3. Kusal, S., Patil, S. & Kotecha, K. Multimodal text-emoji fusion using deep neural networks for text-based emotion detection in online communication. J Big Data 12, 32 (2025). https://doi.org/10.1186/s40537-025-01062-4
4. Khan, A., Majumdar, D. & Mondal, B. Sentiment analysis of emoji fused reviews using machine learning and Bert. Sci Rep 15, 7538 (2025). https://doi.org/10.1038/s41598-025-92286-0
5. Bharti, K., Patil, S., Malave, S., & Gupta, J. (2025). Improved graph convolutional network for emotion analysis in social media text. MethodsX, 14, 103325. https://doi.org/10.1016/j.mex.2025.103325
6. Ferré, P., Haro, J., Pérez-Sánchez, M.Á. et al. Emoji-Dis: A dataset of emojis characterised in 13 discrete emotions. Sci Data 12, 1313 (2025). https://doi.org/10.1038/s41597-025-05682-6
7. Kralj Novak P, Smailović J, Sluban B, Mozetič I (2015) Sentiment of Emojis. PLoS ONE 10(12): e0144296. https://doi.org/10.1371/journal.pone.0144296

8. Cai, Y., Li, X., Zhang, Y. et al. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. Sci Rep 15, 2126 (2025). https://doi.org/10.1038/s41598-025-85859-6
9. Nguyen D.Q. Nguyen A.T., PhoBERT: Pre-trained language models for Vietnamese. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1037–1042, [Online] https://aclanthology.org/2020.findings-emnlp.92/
10. Eisner, B., et al. (2016). emoji2vec: Learning emoji representations from their description. Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media, 48–54. [Online] https://aclanthology.org/W16-6208/
11. Xu, Q. A., Jayne, C., & Chang, V. (2024). An emoji feature-incorporated multi-view deep learning for explainable sentiment classification of social media reviews. Technological Forecasting and Social Change, 202, 123326. https://doi.org/10.1016/j.techfore.2024.123326
12. Liu, S., Zhao, H., Chen, Y. et al. A text guided multimodal scale path fusion network for multimodal sentiment analysis. Sci Rep (2025). https://doi.org/10.1038/s41598-025-32637-z
13. Aziz, K., Ji, D., Chakrabarti, P. et al. Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. Sci Rep 14, 14646 (2024). https://doi.org/10.1038/s41598-024-61886-7
14. Cui, X., Tao, W., & Cui, X. (2023). Affective-Knowledge-Enhanced Graph Convolutional Networks for Aspect-Based Sentiment Analysis with Multi-Head Attention. Applied Sciences, 13(7), 4458. https://doi.org/10.3390/app13074458
15. Singh, G.V., Ghosh, S., Firdaus, M. et al. Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework. Sci Rep 14, 12204 (2024). https://doi.org/10.1038/s41598-024-58944-5