Phân tích tình cảm đa phương thức cho tiếng Việt sử dụng mạng nơ-ron tích chập đồ thị chuẩn hóa

Đỗ Hoàng Nam*, Nguyễn Thị Phong Dung**

Khoa Công nghệ Thông tin, Trường Đại học Nguyễn Tất Thành, Thành phố Hồ Chí Minh
*namdh@ntt.edu.vn, **ntpdung@ntt.edu.vn

Tóm tắt

Phân tích tình cảm đa phương thức (Multimodal Sentiment Analysis - MSA) kết hợp thông tin từ nhiều phương thức, điển hình là văn bản và hình ảnh, nhằm nhận diện chính xác các trạng thái tình cảm của con người. Tuy nhiên, phương pháp MSA hiện nay vẫn gặp hạn chế trong việc mô hình hóa các mối quan hệ cấu trúc ngữ nghĩa phức tạp của tiếng Việt cũng như các tương quan không gian giữa các vùng đặc trưng hình ảnh. Trong nghiên cứu này, đề xuất ViMACSA-GCN, một khung công tác mới cho bài toán MSA tiếng Việt dựa trên mạng nơ-ron tích chập đồ thị. Cụ thể, đặc trưng văn bản và hình ảnh lần lượt được trích xuất bằng PhoBERT và Vision transformer (ViT). Sau đó, các đồ thị đa phương thức được xây dựng để biểu diễn mối quan hệ cấu trúc giữa các thành phần dữ liệu và được tinh chỉnh bằng mạng nơ-ron tích chập đồ thị chuẩn hóa đối xứng. Các biểu diễn đặc trưng được hợp nhất thông qua tầng tuyến tính và sử dụng bộ phân loại để dự đoán ba nhãn tình cảm: Negative, Neutral và Positive. Kết quả thực nghiệm trên tập dữ liệu ViMACSA cho thấy mô hình đạt độ chính xác 87,1 % và F1-score 72,0 %, khẳng định hiệu quả của cách tiếp cận đề xuất đối với MSA tiếng Việt.

Nhận 19/01/2026
Được duyệt 08/02/2026
Công bố 28/02/2026

Từ khóa

Phân tích tình cảm đa phương thức, ViMACSA, mạng nơ-ron tích chập đồ thị, vision transformer, PhoBERT

1 Đặt vấn đề

Sự phát triển nhanh chóng của các nền tảng truyền thông xã hội đã làm gia tăng mạnh mẽ nội dung do người dùng tạo ra, trong đó thông tin tình cảm không còn được biểu đạt đơn thuần qua văn bản mà còn gắn liền với các yếu tố hình ảnh. Điều này thúc đẩy sự ra đời của MSA, với mục tiêu kết hợp các tín hiệu từ văn bản và hình ảnh nhằm diễn giải chính xác trạng thái tình cảm tiềm ẩn của người dùng [1]. Trong những năm gần đây, các phương pháp MSA dựa trên học sâu đã đạt được nhiều tiến bộ đáng kể nhờ sự phát triển của các mô hình tiền huấn luyện và cấu trúc transformer cho cả ngôn ngữ và thị giác. Đối với tiếng Việt, các mô hình ngôn ngữ như PhoBERT đã chứng minh hiệu quả vượt trội trong việc học biểu diễn ngữ nghĩa theo ngữ cảnh [2], trong khi ViT cho thấy khả năng trích xuất đặc trưng hình ảnh cấp cao một cách hiệu quả [3]. Tuy nhiên, phần lớn các phương pháp hiện tại vẫn xử lý dữ liệu đa phương thức dưới dạng chuỗi phẳng, dẫn đến hạn chế trong việc nắm bắt các mối quan hệ cấu trúc phức tạp tồn tại trong cả văn bản và hình ảnh. Đối với tiếng Việt, một ngôn ngữ biệt lập với trật tự từ linh hoạt và giàu ngữ cảnh − việc bỏ qua các phụ thuộc cấu trúc có thể làm suy giảm đáng kể hiệu quả phân tích tình cảm. Các nghiên cứu gần đây cho thấy, mạng nơ-ron tích chập đồ thị (GCN) có khả năng mô hình hóa hiệu

quả các quan hệ phi tuyến trên dữ liệu có cấu trúc, qua đó khắc phục hạn chế của các cấu trúc tuần tự truyền thống [4, 5]. Đồng thời, trong bối cảnh học đa phương thức, biểu diễn dựa trên đồ thị ngày càng được xem là hướng tiếp cận tiềm năng để liên kết và tinh chỉnh các đặc trưng khác loại một cách có hệ thống [6].

Xuất phát từ những quan sát trên, nghiên cứu này đề xuất ViMACSA-GCN, một khung MSA được thiết kế chuyên biệt cho tiếng Việt, trong đó GCN đóng vai trò trung tâm trong việc mô hình hóa và kết hợp các mối quan hệ cấu trúc giữa văn bản và hình ảnh.

2 Các nghiên cứu liên quan

Sự phát triển của MSA đặc trưng bởi sự chuyển đổi từ việc ghép nối các đặc trưng đơn giản sang học biểu diễn phức tạp. Các phương pháp hiện có có thể được phân loại rộng rãi thành hai mô hình chính: các phương pháp dựa trên transformer và các phương pháp dựa trên đồ thị.

2.1 Các phương pháp MSA dựa trên transformer

Transformer đã trở thành cấu trúc chính trong MSA nhờ khả năng mô hình hóa các phụ thuộc tầm xa và căn chỉnh thông tin giữa các phương thức không đồng bộ. Phương pháp MulT [7], giới thiệu cơ chế chú ý chéo được sử dụng để học mối quan hệ giữa các chuỗi đa phương thức không được căn chỉnh, mở ra hướng tiếp cận hiệu quả cho MSA. Tiếp theo đó, phương pháp MISA [8], nhằm tách các biểu diễn đa phương thức thành thành phần bất biến và đặc thù theo từng phương thức, qua đó cải thiện khả năng tổng hợp thông tin tình cảm. Các nghiên cứu gần đây tiếp tục mở rộng cấu trúc transformer bằng cách khai thác cấu trúc phân cấp và cơ chế điều tiết động. Điển hình, phương pháp HGTFM [9], trong đó các tầng transformer được tổ chức theo cấu trúc phân cấp nhằm tăng cường khả năng kết hợp thông tin từ nhiều mức độ trừu tượng khác nhau. Mặc dù đạt được những cải thiện đáng kể về hiệu suất, các mô hình transformer cho MSA vẫn chủ yếu xử lý dữ liệu dưới dạng chuỗi phẳng. Cách tiếp cận này có xu hướng làm mất các mối quan hệ cấu trúc nội tại, chẳng hạn như phụ thuộc cú pháp trong văn bản hoặc bố cục không gian giữa các thành phần hình ảnh, đồng thời dễ bị ảnh hưởng bởi nhiễu trong môi trường truyền thông xã hội.

2.2 Các phương pháp MSA dựa trên đồ thị

Nhằm khắc phục hạn chế về biểu diễn cấu trúc của các mô hình dựa trên transformer, các nghiên cứu gần đây đã chuyển sang khai thác mạng nơ-ron đồ thị (GNN) cho bài toán MSA. Bằng cách mô hình hóa dữ liệu đa phương thức dưới dạng đồ thị, các phương pháp này cho phép biểu diễn tường minh mối quan hệ giữa các thực thể khác nhau và thực hiện suy luận quan hệ thông qua cơ chế truyền thông điệp. Mô hình MSA dựa trên đồ thị kết hợp tái cấu trúc đồ thị và cơ chế chú ý, cho thấy khả năng cải thiện đáng kể hiệu suất so với các cấu trúc tuần tự truyền thống [10]. Các kết quả này khẳng định vai trò quan trọng của biểu diễn đồ thị trong việc nắm bắt các phụ thuộc phi tuyến giữa các phương thức. Tuy nhiên, việc huấn luyện các mạng đồ thị sâu trong bối cảnh dữ liệu đa phương thức vẫn đối mặt với thách thức về sự mất cân bằng lớp, một vấn đề phổ biến trong phân tích tình cảm trên truyền thông xã hội. Tổng quan gần đây chỉ ra rằng, các kỹ thuật chuẩn hóa và cơ chế lan truyền ổn định là yếu tố then chốt để cải thiện khả năng học trên đồ thị mất cân bằng [11].

Dựa trên những tiến bộ này, nghiên cứu đề xuất kế thừa hướng tiếp cận dựa trên đồ thị và tập trung vào việc ổn định quá trình lan truyền đặc trưng thông qua chuẩn hóa Laplacian đối xứng, từ đó nâng cao hiệu quả MSA cho tiếng Việt.

3 Phương pháp nghiên cứu

3.1 Đặc điểm dữ liệu

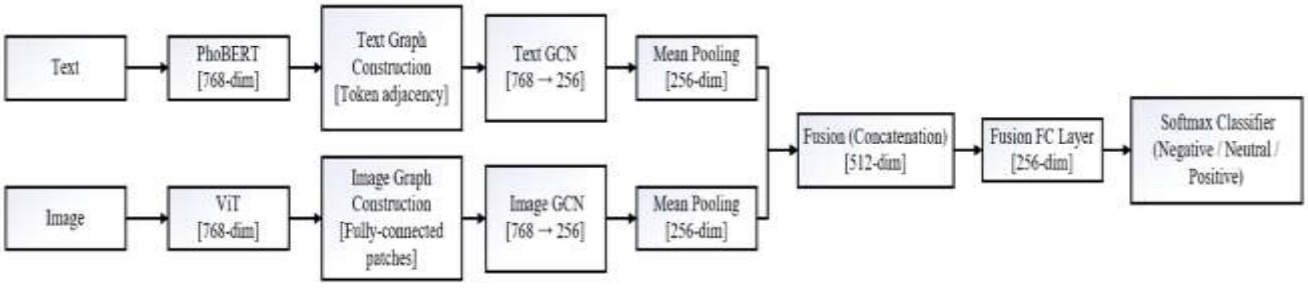
Trong nghiên cứu này, sử dụng bộ dữ liệu có tên Vietnamese Multimodal Aspect-Category Sentiment Analysis (ViMACSA) là một tập dữ liệu đa phương thức mới dành cho bài toán phân tích tình cảm theo khía cạnh trong tiếng Việt, kết hợp văn bản và hình ảnh trong miền khách sạn [12]. Tập dữ liệu gồm 4 876 cặp văn bản − hình ảnh, với 14 618 nhãn tình cảm chi tiết, trong đó mỗi mẫu có thể đi kèm tối đa 7 ảnh. ViMACSA được gán nhãn fine-grained theo từng aspect category (như phòng, dịch vụ, tiện nghi, vị trí,…) cho cả hai phương thức, cho phép mô hình học và khai thác mối quan hệ tình cảm giữa văn bản và hình ảnh. Tập dữ liệu được xây dựng nhằm khắc phục hạn chế của các bộ dữ liệu ACSA truyền thống chỉ dựa trên văn bản, đồng thời hỗ trợ nghiên cứu các phương pháp

kết hợp đa phương thức và đánh giá mô hình phân tích tình cảm tiếng Việt trong bối cảnh thực tế mạng xã hội.

3.2 Tổng quan cấu trúc

Trong phần này, trình bày chi tiết cấu trúc của mô hình ViMACSA-GCN, được thiết kế nhằm khai thác hiệu quả thông tin đa phương thức từ văn bản tiếng Việt và hình ảnh. Mô hình kết hợp các bộ trích xuất đặc trưng mạnh mẽ dựa trên transformer với GCN để mô hình hóa các mối quan hệ ngữ nghĩa và không gian trước khi thực hiện hội tụ đa phương thức cho bài toán phân tích tình cảm.

Cấu trúc tổng thể của ViMACSA-GCN bao gồm bốn thành phần chính: (i) nhánh trích xuất đặc trưng văn bản, sử dụng mô hình ngôn ngữ tiền huấn luyện PhoBERT; (ii) nhánh trích xuất đặc trưng thị giác, dựa trên ViT; (iii) các lớp GCN để tinh lọc và lan truyền thông tin trên cấu trúc đồ thị; (iv) khối kết hợp phương thức và bộ phân loại tình cảm. Cấu trúc này cho phép mô hình vừa nắm bắt đặc trưng riêng biệt của từng modality, vừa khai thác mối quan hệ nội tại thông qua biểu diễn đồ thị. Quy trình đề xuất được minh họa trong Hình 1, bao gồm các nhánh trích xuất đặc trưng văn bản (PhoBERT) và hình ảnh (ViT), các lớp tích chập đồ thị chuẩn hóa và khối hợp nhất đa phương thức cho phân loại tình cảm.



Hình 1 Kiến trúc tổng thể của mô hình ViMACSA-GCN

Với mỗi mẫu dữ liệu đầu vào đa phương thức: $X = \{T, I\}$, trong đó T biểu thị nội dung bình luận văn bản (text) và I là hình ảnh (image) đi kèm, mô hình hướng tới mục tiêu dự đoán nhãn tình cảm: $y \in$ {Negative, Neutral, Positive}.

3.2.1 Nhánh xử lý văn bản dựa trên đồ thị

Để nắm bắt các sắc thái biểu cảm phức tạp trong tiếng Việt, mô hình sử dụng PhoBERT, một mô hình ngôn ngữ tiền huấn luyện dựa trên cấu trúc transformer, được tối ưu hóa cho tiếng Việt.

Với văn bản đầu vào T $(text)$, PhoBERT mã hóa thành một chuỗi các vector đặc trưng theo token:

$$X_{\text{text}} = \{h_1, h_2, \dots, h_n\}$$

trong đó n là số lượng token sau khi tách từ. Mỗi $h_i$là vector đặc trưng của token thứ i, chứa thông tin ngữ nghĩa và ngữ cảnh của token đó.

Thay vì chỉ khai thác thông tin tuần tự, công trình xây dựng đồ thị văn bản (Text Graph) $G_t = (V_t, E_t)$, trong đó, mỗi nút $v \in V_t$ tương ứng với một token; các cạnh $e \in E_t$ được thiết lập dựa trên mối quan hệ lân cận giữa các token trong câu. Cách tiếp cận này cho phép mô hình khai thác hiệu quả cấu trúc cú pháp cục bộ của tiếng Việt, đồng thời tạo điều kiện cho việc lan truyền thông tin ngữ cảnh giữa các token liên quan. Việc sử dụng quan hệ lân cận tuần tự trong đồ thị văn bản là một lựa chọn có chủ đích. Đối với tiếng Việt, các bộ phân tích cú pháp hiện nay vẫn còn hạn chế về độ ổn định, đặc biệt trong dữ liệu mạng xã hội nhiều nhiễu và không chuẩn hóa. Việc tích hợp các quan hệ cú pháp sâu có thể làm gia tăng sai lệch lan truyền trên đồ thị và ảnh hưởng tiêu cực đến quá trình huấn luyện. Do đó, nghiên cứu ưu tiên mô hình hóa các phụ thuộc ngữ cảnh cục bộ, đồng thời đảm bảo tính ổn định và khả năng tổng quát của mô hình. Việc mở rộng đồ thị văn bản với các quan hệ cú pháp sâu hơn được xem là hướng nghiên cứu tiềm năng trong tương lai.

3.2.2 Nhánh xử lý hình ảnh dựa trên đồ thị

Nhánh hình ảnh sử dụng ViT để xử lý hình ảnh đầu vào. Cụ thể, hình ảnh được chia thành các patch không chồng lấp, mỗi patch sau đó được ánh xạ thành một vector đặc trưng:

$$X_{\text{img}} = \{v_1, v_2, \dots, v_m\}$$

trong đó m là số lượng patch ảnh sau khi chia ảnh đầu vào. Mỗi $v_i$là vector đặc trưng của patch thứ i.

Công trình xây dựng đồ thị hình ảnh (Image Graph) $G_i = (V_i, E_i)$ bằng cách xem mỗi patch ảnh là một nút $v \in V_i$. Các cạnh $e \in E_i$được thiết lập theo cơ chế kết

nối đầy đủ (fully-connected graph), tức là mỗi nút đều có liên kết với tất cả các nút còn lại. Thiết kế này giúp mô hình nắm bắt mối quan hệ toàn cục giữa các vùng khác nhau trong ảnh, chẳng hạn như mối liên hệ giữa biểu cảm khuôn mặt và bối cảnh xung quanh, từ đó tăng khả năng suy luận tình cảm thị giác.

3.2.3 Lớp tích chập đồ thị

Để tinh lọc đặc trưng thu được từ các đồ thị văn bản và hình ảnh, mô hình sẽ áp dụng các lớp GCN với cơ chế chuẩn hóa đối xứng. Quá trình lan truyền thông tin tại lớp thứ l được định nghĩa như sau:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right)$$

trong đó: $\tilde{A} = A + I$ là ma trận kề được bổ sung các vòng tự lặp (self-loops); $\tilde{D}$ là ma trận bậc của $\tilde{A}$; $H^{(l)}$*là ma trận đặc trưng của các nút tại lớp* l; $W^{(l)}$*là ma trận trọng số có thể học tại lớp* l; $\sigma(\cdot)$ là hàm kích hoạt ReLU.

Việc sử dụng Laplacian chuẩn hóa đối xứng giúp ổn định quá trình huấn luyện và giảm thiểu hiện tượng bùng nổ gradient, đặc biệt quan trọng khi xử lý dữ liệu đa phương thức có mức độ nhiễu cao.

3.2.4 Kết hợp đa phương thức và phân loại tình cảm

Sau khi thu được các đặc trưng đồ thị đã qua GCN từ hai nhánh, ký hiệu lần lượt là H_{text} và H_{img}, mô hình áp dụng Global Average Pooling để trích xuất các vector biểu diễn toàn cục. Hai vector này được nối lại (concatenation) để tạo thành vector đặc trưng đa phương thức:

$$Z_{\text{fused}} = \left[\text{pool}(H_{\text{text}}) \ \oplus \ \text{pool}(H_{\text{img}})\right]$$

trong đó, $\text{pool}(\cdot)$là phép tổng hợp đặc trưng nhằm chuyển ma trận đặc trưng nút thành một vector biểu diễn toàn cục; $\oplus$ là phép nối vector (concatenation).

Vector hợp nhất Z_{fused} sau đó được đưa qua một tầng fully connected, tiếp theo là hàm Softmax để dự đoán phân phối xác suất trên ba lớp cảm xúc: tích cực (positive), tiêu cực (negative) và trung tính (neutral).

3.2.5 Hàm mất mát và tối ưu hóa

Do tập dữ liệu ViMACSA có sự mất cân bằng nhãn đáng kể, trong đó lớp *Positive* chiếm ưu thế, mô hình sử dụng hàm mất mát entropy chéo có trọng số (Weighted Cross-Entropy Loss (WCEL)) để huấn luyện mô hình. Hàm mất mát được định nghĩa như sau:

$$\mathcal{L} = -\sum_{c=1}^{3} w_c \cdot y_c \log(\hat{y}_c)$$

trong đó: y_c và $\hat{y}_c$ lần lượt là nhãn thực và xác suất dự đoán của lớp c; Trọng số w_c được tính nghịch đảo với tần suất xuất hiện của lớp c.

Cơ chế này giúp mô hình tập trung nhiều hơn vào các lớp thiểu số (*Negative* và *Neutral*), từ đó cải thiện đáng kể chỉ số Macro F1-score và khả năng tổng quát hóa.

3.2.6 Thuật toán ViMACSA-GCN

Thuật toán dưới đây mô tả chi tiết toàn bộ quy trình suy luận của mô hình ViMACSA-GCN, từ dữ liệu đầu vào đa phương thức thô (văn bản tiếng Việt và hình ảnh) đến kết quả phân loại tình cảm cuối cùng. Điểm cốt lõi của phương pháp nằm ở việc nhúng đặc trưng vào không gian đồ thị và tinh lọc biểu diễn thông qua tích chập đồ thị chuẩn hóa, trước khi thực hiện hội tụ đa phương thức.

Thuật toán 1: quy trình phân loại tình cảm đa phương thức ViMACSA-GCN

Đầu vào:

Câu văn bản tiếng Việt T; hình ảnh tương ứng I;

các tham số mô hình đã huấn luyện

$\Theta = \{\Theta_{\text{PhoBERT}}, \Theta_{\text{ViT}}, \Theta_{\text{GCN}}\}$.

Đầu ra:

Nhãn tình cảm dự đoán

$$y \in \{\text{Negative, Neutral, Positive}\}.$$

Bước 1: tiền xử lý và trích xuất đặc trưng cơ sở

- Mã hóa văn bản đầu vào bằng PhoBERT:

$$X_{\text{text}} \leftarrow \text{PhoBERT}(T),$$

trong đó $X_{\text{text}} \in \mathbb{R}^{n \times d_t}$ là chuỗi vector đặc trưng token.

- Trích xuất đặc trưng hình ảnh bằng ViT:

$$X_{\text{img}} \leftarrow \text{ViT}(I),$$

trong đó $X_{\text{img}} \in \mathbb{R}^{m \times d_i}$là tập vector biểu diễn các patch ảnh.

Bước 2: xây dựng cấu trúc đồ thị

- Đồ thị Văn bản G_t: thiết lập ma trận kề A_t dựa trên mối quan hệ tuần tự giữa các token liền kề trong câu.

- Đồ thị Hình ảnh G_i: thiết lập ma trận kề A_i theo cấu trúc kết nối đầy đủ (fully connected) giữa các patch ảnh.

Bước 3: tinh lọc đặc trưng bằng GCN

- Chuẩn hóa ma trận kề cho mỗi đồ thị:

$$\hat{A} = \tilde{D}^{-1/2}(A + I)\tilde{D}^{-1/2},$$

trong đó I là ma trận đơn vị và $\tilde{D}$ là ma trận bậc.

- Lan truyền thông tin qua GCN:

$$H_{\text{text}} \leftarrow \text{ReLU}(\hat{A}_t X_{\text{text}} W_t),$$
$$H_{\text{img}} \leftarrow \text{ReLU}(\hat{A}_i X_{\text{img}} W_i),$$

với $W_t$và $W_i$là các tham số học được của GCN.

Bước 4: kết hợp đa phương thức

- Thực hiện Global Average Pooling:

$$f_{\text{text}} = \text{AvgPool}(H_{\text{text}}), f_{\text{img}} = \text{AvgPool}(H_{\text{img}}).$$

- Kết nối đặc trưng hai miền:

$$Z_{\text{fused}} = [f_{\text{text}} \oplus f_{\text{img}}].$$

Bước 5: dự đoán nhãn tình cảm

- Tính phân phối xác suất:

$$P(y \mid T, I) = \text{Softmax}(\text{Linear}(Z_{\text{fused}})).$$

- Suy ra nhãn tình cảm cuối cùng:

$$y = \arg\max P(y \mid T, I).$$

3.2.7 Thiết lập thực nghiệm

Triển khai mô hình ViMACSA-GCN trên nền tảng PyTorch, sử dụng GPU NVIDIA GeForce RTX 4070 Ti Super 16GB để tăng tốc quá trình huấn luyện và suy luận. Tập dữ liệu ViMACSA được chia theo tỷ lệ chuẩn cho các tập huấn luyện, xác thực và kiểm thử, trong đó tập Test bao gồm 1 000 mẫu, phản ánh đa dạng các sắc thái tình cảm thường gặp trên mạng xã hội Việt Nam. Trong quá trình huấn luyện, mô hình sử dụng bộ tối ưu hóa AdamW, với chiến lược thiết lập tốc độ học khác nhau cho từng nhóm tham số nhằm đảm bảo sự hội tụ ổn định: Learning rate 2×10^{-5} cho các mô hình backbone đã được tiền huấn luyện (PhoBERT và ViT); Learning rate 1×10^{-4}cho các lớp GCN và khối phân loại. Chiến lược này cho phép bảo toàn tri thức ngôn ngữ và thị giác đã học trước đó, đồng thời giúp các lớp đồ thị thích nghi nhanh với đặc thù của bài toán MSA.

4 Kết quả và bàn luận

4.1 Kết quả thực nghiệm

Trình bày kết quả so sánh giữa mô hình ViMACSA-GCN đề xuất và các phương pháp cơ sở (baselines) trên tập kiểm thử ViMACSA ở Bảng 1.

Bảng 1 So sánh hiệu năng trên tập kiểm thử ViMACSA với các chỉ số Accuracy và Macro F1-score

Mô hình	Accuracy (%)	Macro F1-score (%)
Text-only (PhoBERT)	81,2	64,0
Image-only (ViT)	65,8	48,0
MACSA-LSTM	84,5	69,5
ViMACSA-GCN (đề xuất)	**87,1**	**72,0**

Kết quả cho thấy, mô hình ViMACSA-GCN đạt độ chính xác (Acc) 87,1 % và Macro F1-score 72,0 %, vượt trội so với các mô hình đơn phương thức cũng như mô hình đa phương thức khác. Cụ thể, ViMACSA-GCN cải thiện Acc 2,6 % và 2,5 % Macro F1-score so với MACSA-LSTM. Sự cải thiện này cho thấy việc sử dụng tích chập đồ thị với Laplacian chuẩn hóa giúp mô hình ổn định hơn trước dữ liệu nhiễu, vốn rất phổ biến trong môi trường mạng xã hội.

Để đánh giá sâu hơn tác động của hiện tượng mất cân bằng nhãn, mô hình được tiến hành phân tích chi tiết hiệu năng trên từng lớp tình cảm, như thể hiện trong Bảng 2.

Bảng 2 Kết quả đánh giá chi tiết theo từng nhãn tình cảm trên tập kiểm thử ViMACSA.

Nhãn	Precision	Recall	F1-score	Support
Negative	0,93	0,69	0,79	137
Neutral	0,49	0,37	0,42	106
Positive	0,90	0,97	0,94	757

Phân tích chi tiết:

- Lớp Positive đạt F1-score rất cao (0,94), cho thấy mô hình học hiệu quả các đặc trưng tình cảm tích cực – vốn chiếm đa số trong tập dữ liệu.
- Lớp Negative, mặc dù có số lượng mẫu hạn chế (137), vẫn đạt Precision lên tới 0,93. Điều này chứng tỏ ViMACSA-GCN có khả năng nhận diện chính xác các tín hiệu tiêu cực và hạn chế việc dự đoán nhầm sang các lớp khác.
- Lớp Neutral là lớp khó nhất, với F1-score đạt 0,42. Tuy nhiên, so với các phiên bản không sử dụng đồ thị, kết quả này cho thấy GCN đã giúp cải thiện khả năng phân biệt các sắc thái trung tính – vốn thường bị nhầm lẫn với tình cảm tích cực trong tiếng Việt. Nguyên nhân của hiện tượng này đến từ tính mơ hồ ngữ nghĩa của lớp Neutral trong tiếng Việt, khi nhiều biểu đạt trung tính chỉ mang sắc thái nhẹ và dễ bị chi phối bởi các tín hiệu tích cực từ ngữ cảnh hoặc hình ảnh. Bên cạnh đó, việc sử dụng đồ thị đồng nhất và cơ chế hợp nhất tuyến tính có thể chưa đủ linh hoạt để làm nổi bật các tín hiệu tình cảm yếu. Trong tương lai, mô hình có thể được cải

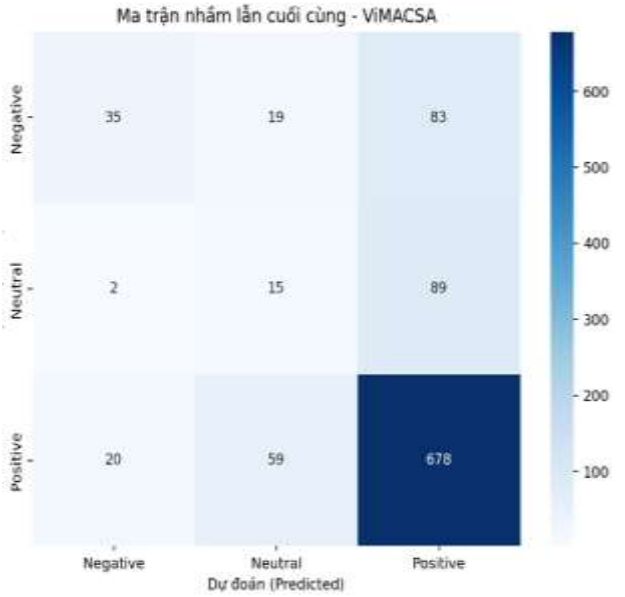
thiện bằng cách tích hợp cơ chế chú ý trong giai đoạn hợp nhất đa phương thức nhằm gán trọng số thích nghi cho từng phương thức, hoặc mở rộng sang đồ thị không đồng nhất để mô hình hóa rõ hơn các loại nút và quan hệ khác nhau giữa văn bản và hình ảnh.

4.2 Bàn luận

Nguyên nhân chính giúp ViMACSA-GCN đạt được hiệu năng cao (Acc 87,1 %) nằm ở khả năng tinh lọc đặc trưng của tích chập đồ thị. Trong khi các cơ chế khác như MACSA-LSTM có thể bị chi phối bởi các từ khóa hoặc vùng ảnh gây nhiễu, GCN với Laplacian chuẩn hóa đối xứng cho phép lan truyền thông tin một cách cân bằng giữa các nút lân cận.

Cụ thể:

- Đồ thị văn bản giúp mô hình nắm bắt các từ khóa tình cảm quan trọng bất kể vị trí của chúng trong câu,
- Đồ thị hình ảnh kết nối các patch đặc trưng, cho phép mô hình hiểu được bối cảnh tổng thể của hình ảnh thay vì chỉ tập trung vào các vùng cục bộ.



Hình 2 Ma trận nhầm lẫn cuối cùng của mô hình ViMACSA-GCN

Sự kết hợp này tạo ra một vector biểu diễn đa phương thức có khả năng phân biệt cao, đặc biệt hiệu quả trong bối cảnh ngôn ngữ mạng xã hội tiếng Việt, nơi dữ liệu thường mang tính không chuẩn hóa và nhiều nhiễu. Mô tả ma trận nhầm lẫn như Hình 2.

Kết quả cho thấy mô hình đạt hiệu suất cao nhất đối với lớp *Positive*, với phần lớn các mẫu được phân loại chính xác, phản ánh khả năng khai thác hiệu quả các tín hiệu tình cảm tích cực từ dữ liệu đa phương thức. Tuy nhiên, vẫn tồn tại một tỷ lệ đáng kể các mẫu thuộc lớp *Neutral* và *Negative* bị dự đoán nhầm sang lớp *Positive*, cho thấy xu hướng thiên lệch về lớp chiếm ưu thế cũng như khó khăn trong việc phân biệt các biểu hiện tình cảm trung tính hoặc tiêu cực ở mức độ tinh tế. Điều này gợi ý rằng, bên cạnh việc mô hình hóa quan hệ nội tại bằng GCN, cần bổ sung các cơ chế hợp nhất và chiến lược huấn luyện nâng cao nhằm cải thiện hiệu suất trên các lớp thiểu số.

4 Kết luận

Trong bài viết này, giới thiệu mô hình ViMACSA-GCN, một khung dựa trên biểu đồ mới để phân tích tâm lý đa phương thức của người Việt. Bằng cách tích hợp sức mạnh ngữ nghĩa của PhoBERT với khả năng biểu diễn trực quan của ViT thông qua GCN chuẩn hóa đối xứng, cách tiếp cận này nắm bắt một cách hiệu quả các mối quan hệ phức tạp giữa cú pháp văn bản và không gian hình ảnh. Kết quả thực nghiệm trên bộ dữ liệu ViMACSA chứng minh rằng mô hình của đề xuất đạt được độ chính xác tiên tiến là 87,10 % và điểm Macro F1 là 72,0 %. Tính ưu việt của kiến trúc GCN so với các mô hình dựa trên cơ chế khác khẳng định cần thiết để giảm thiểu nhiễu và xử lý sự mơ hồ vốn có của nội dung mạng xã hội. Hơn nữa, chiến lược tối ưu hóa có trọng số của mô hình đã giải quyết thành công thách thức về sự mất cân bằng lớp và cải thiện hiệu suất MSA. Trong tương lai, nghiên cứu sẽ mở rộng mô hình theo hướng đồ thị không đồng nhất, tích hợp thêm dữ liệu âm thanh hoặc video, cũng như áp dụng các chiến lược học mất cân bằng nâng cao để cải thiện hiệu năng trên các lớp thiểu số.

Tài liệu tham khảo

1. Sharma, R., Le Tan, N., & Sadat, F. (2018). Multimodal sentiment analysis using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1475-1478). IEEE. https://doi.org/10.1109/ICMLA.2018.00240.

2. Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained Language Models for Vietnamese. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1037-1042). https://doi.org/10.18653/v1/2020.findings-emnlp.92.
3. Dosovitskiy, A. (2020). An image is worth 16 × 6 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929.
4. Zhao, T., Peng, J., Huang, Y., Wang, L., Zhang, H., & Cai, Z. (2023). A graph convolution-based heterogeneous fusion network for multimodal sentiment analysis. *Applied Intelligence*, *53*(24), 30455-30468. DOI:10.1007/s10489-023-05151-w.
5. Phan, C. T., Nguyen, Q. N., Dang, C. T., Do, T. H., & Van Nguyen, K. (2023). ViCGCN: graph convolutional network with contextualized language models for social media mining in Vietnamese. *arXiv preprint arXiv:2309.02902*. https://doi.org/10.48550/arXiv.2309.02902.
6. Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., & Zitnik, M. (2023). Multimodal learning with graphs. *Nature Machine Intelligence*, *5*(4), 340-350. https://doi.org/10.1038/s42256-023-00624-6.
7. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of The Conference. Association For Computational Linguistics. Meeting* (Vol. 2019, p. 6558). https://doi.org/10.18653/v1/p19-1656.
8. Hazarika, D., Zimmermann, R., & Poria, S. (2020). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1122-1131). https://doi.org/10.1145/3394171.3413678.
9. Yang, C., Liang, Z., Yan, D., Hu, Z., & Wu, T. (2025). HGTFM: Hierarchical Gating-Driven Transformer Fusion Model for Robust Multimodal Sentiment Analysis. *IEEE Access*. DOI: 10.1109/ACCESS.2025.3560641.
10. Hu, R., Yi, J., Chen, L., & Jin, Z. (2024). Graph Reconstruction Attention Fusion Network for Multimodal Sentiment Analysis. *IEEE Transactions on Industrial Informatics*. doi: 10.1109/TII.2024.3452204.
11. Ma, Y., Tian, Y., Moniz, N., & Chawla, N. V. (2025). Class-imbalanced learning on graphs: A survey. *ACM Computing Surveys*, *57*(8), 1-16. https://doi.org/10.1145/3718734.
12. Nguyen, Q. H., Nguyen, M. V. T., & Van Nguyen, K. (2025). New benchmark dataset and fine-grained cross-modal fusion framework for Vietnamese multimodal aspect-category sentiment analysis. *Multimedia Systems*, *31*(1), 4. https://doi.org/10.1007/s00530-024-01558-8.

Multimodal Sentiment Analysis for Vietnamese Using Normalized Graph Convolutional Networks

Do Hoang Nam*, Nguyen Thi Phong Dung**
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Viet Nam
*namdh@ntt.edu.vn, **ntpdung@ntt.edu.vn

Abstract Multimodal Sentiment Analysis (MSA) combines information from multiple modes, typically text and images, to accurately identify human emotional states. However, many current MSA methods are limited in modeling the complex semantic structural relationships of Vietnamese language, as well as the spatial correlations between image feature regions. This study proposed ViMACSA-GCN as a new framework for Vietnamese MSA based on graph convolutional neural networks. Specifically, text and image features were extracted using PhoBERT and Vision transformer (ViT), respectively. Then, multimodal graphs were constructed to represent the structural relationships between data components and refined using a symmetrically normalized graph convolutional network. Feature representations were merged through a linear layer and a classifier was used to predict three sentiment labels: Negative, Neutral, and Positive. Experimental results on the ViMACSA dataset showed that the model achieved an accuracy of 87.1% and an F1-score of 72.0%, confirming the effectiveness of the proposed approach for Vietnamese MSA.

Keywords Multimodal sentiment analysis, ViMACSA, Graph convolution network, Vision transformer, PhoBERT