

Đánh giá đặc điểm các tập dữ liệu dùng trong phân tích tình cảm đa phương thức

Nguyễn Xuân Cường*, Đỗ Hoàng Nam**

Khoa Công nghệ thông tin – Trường Đại học Nguyễn Tất Thành

*nxcuong@ntt.edu.vn, **namdh@ntt.edu.vn

Tóm tắt

Phân tích tình cảm đa phương thức (MSA) ngày càng thu hút sự quan tâm. Tuy nhiên, các bộ dữ liệu hiện có vẫn còn hạn chế về tính toàn diện, tính đa dạng và tính nhất quán. Nghiên cứu này cung cấp một đánh giá có hệ thống về các bộ dữ liệu MSA phổ biến, phân tích các đặc điểm chính như phương thức dữ liệu, phương pháp chú thích và độ tin cậy. Đề xuất một quy trình đánh giá gồm bốn giai đoạn: (1) khảo sát thu thập và quản lý tập dữ liệu, (2) phân tích thống kê đặc trưng và kỹ thuật, (3) đánh giá hiệu suất mô hình trên các bộ dữ liệu này, và (4) xây dựng hướng dẫn phát triển tập dữ liệu chất lượng cao cho MSA. Kết quả phân tích cung cấp hiểu biết sâu sắc về điểm mạnh và hạn chế của từng bộ dữ liệu, giúp các nhà nghiên cứu lựa chọn nguồn dữ liệu phù hợp và định hướng phát triển các tập dữ liệu đa phương thức tốt hơn trong tương lai.

© 2025 Journal of Science and Technology - NTTU

Nhận 09/04/2024

Được duyệt 05/08/2025

Công bố 28/08/2025

Từ khóa

phân tích tình cảm
đa phương thức,
đánh giá tập dữ liệu,
trích xuất tính năng

1 Giới thiệu

Phân tích cảm xúc đa phương thức (Multimodal Sentiment Analysis - MSA) đã nhanh chóng trở thành một hướng nghiên cứu trọng yếu trong lĩnh vực phân tích cảm xúc, nhờ khả năng khai thác và tích hợp các loại dữ liệu đa dạng như văn bản, hình ảnh, âm thanh và video. Không giống như các phương pháp phân tích cảm xúc truyền thống vốn chỉ dựa vào dữ liệu đơn phương thức như văn bản, MSA hướng tới việc nắm bắt toàn diện hơn các sắc thái cảm xúc phức tạp của con người. Điều này đặc biệt quan trọng trong bối cảnh bùng nổ của nội dung đa phương tiện trên mạng xã hội và các nền tảng trực tuyến, nơi mà biểu đạt cảm xúc vượt xa giới hạn của ngôn từ.

Nhu cầu ngày càng tăng về các kỹ thuật phân tích cảm xúc chính xác và tinh vi đã thúc đẩy giới nghiên cứu tập trung phát triển các mô hình MSA với khả năng mô tả sâu sắc hơn trạng thái cảm xúc. Một khảo sát toàn diện đã phân loại các kiến trúc MSA thành 10 hướng tiếp cận khác nhau, bao gồm: kết hợp sớm (early fusion), kết hợp muộn (late fusion), kết hợp lai (hybrid

fusion), kết hợp cấp độ mô hình (model-level fusion), kết hợp tensor (tensor fusion), kết hợp phân cấp (hierarchical fusion), kết hợp song phương thức (bi-modal fusion), kết hợp dựa trên cơ chế attention, kết hợp dựa trên lý thuyết lượng tử (quantum-based fusion) và kết hợp ở cấp độ từ vựng (word-level fusion) [1]. Việc lựa chọn phương pháp phù hợp thường phụ thuộc vào đặc tính của dữ liệu đầu vào cũng như mục tiêu phân tích cụ thể của mô hình. Trong khi đó, một số nghiên cứu gần đây đã cho thấy tiềm năng của việc khai thác tương tác giữa các phương thức hoặc sử dụng chú thích đơn phương thức để tối ưu hóa hiệu quả phân tích cảm xúc trên các tập dữ liệu như CH-SIMS [2]. Đồng thời, một nghiên cứu khác đã hệ thống hoá 18 bộ dữ liệu phục vụ cho các tác vụ trực quan sử dụng Graph Convolutional Networks (GCN), giúp cung cấp cơ sở lựa chọn phong phú và phù hợp cho các nhà nghiên cứu trong lĩnh vực này [3]. Tầm quan trọng của MSA chính xác sẽ giúp doanh nghiệp và nhà cung cấp dịch vụ có thể điều chỉnh xu hướng kinh doanh để tương tác với khách hàng tốt hơn, cải thiện sự hài lòng và lòng trung

thành của người tiêu dùng. MSA có tính ứng dụng cao trên nhiều lĩnh vực khác nhau, như giám sát phương tiện truyền thông xã hội, dịch vụ khách hàng, đánh giá sức khỏe tâm thần và tương tác giữa con người với máy tính, khiến nó trở thành một công cụ linh hoạt cho nhiều ứng dụng khác nhau. Mục tiêu chính của nghiên cứu này là đánh giá các đặc điểm của bộ dữ liệu được sử dụng trong MSA và đề xuất một khuôn khổ toàn diện để đánh giá và cải thiện các bộ dữ liệu này. Tuy nhiên, lĩnh vực MSA vẫn còn đối mặt với nhiều thách thức, trong đó quan trọng nhất là sự thiếu hụt các bộ dữ liệu toàn diện. Các bộ dữ liệu hiện có thường bị giới hạn về quy mô, tính đa dạng và độ tin cậy của nhãn. Ngoài ra, sự không đồng nhất giữa các phương thức (modalities) cũng gây khó khăn trong việc tích hợp và xử lý dữ liệu hiệu quả. Do đó, việc đánh giá và cải thiện các bộ dữ liệu MSA là một nhu cầu cấp thiết để thúc đẩy sự phát triển của lĩnh vực này.

Các nghiên cứu gần đây cho thấy, MSA có khả năng vượt trội trong việc nắm bắt sắc thái cảm xúc so với các mô hình chỉ sử dụng một phương thức đầu vào. Tuy nhiên, hiệu quả của MSA phụ thuộc mạnh mẽ vào chất lượng và tính phù hợp của bộ dữ liệu được sử dụng. Mặc dù đã có một số tập dữ liệu MSA được công bố rộng rãi như CMU-MOSEI, CH-SIMS hay MELD, nhưng mỗi bộ dữ liệu lại có sự khác biệt đáng kể về ngôn ngữ, cấu trúc, quy trình chú thích và tính đại diện cho các ngữ cảnh sử dụng thực tế. Bên cạnh đó, các phương pháp đánh giá chất lượng bộ dữ liệu hiện còn thiếu nhất quán, khiến cho việc so sánh và lựa chọn tập dữ liệu phù hợp với mục tiêu nghiên cứu trở nên khó khăn.

Thực tiễn này đặt ra một khoảng trống nghiên cứu quan trọng: cần có một cách tiếp cận hệ thống để đánh giá, so sánh và định hướng phát triển các bộ dữ liệu trong MSA. Trong bối cảnh đó, nghiên cứu này được thực hiện nhằm ba đóng góp chính:

- Tổng hợp và phân tích các đặc điểm nổi bật của các bộ dữ liệu MSA hiện có.
- Xây dựng khung tiêu chí đánh giá chất lượng bộ dữ liệu dựa trên các yếu tố như tính đa dạng, tính nhất quán và khả năng ứng dụng mô hình hóa.
- Đề xuất các hướng cải tiến nhằm hỗ trợ xây dựng các tập dữ liệu MSA chất lượng cao, phục vụ cho phát triển các mô hình phân tích cảm xúc đáng tin cậy hơn trong thực tiễn.

Bằng cách tiếp cận tổng hợp từ lý thuyết đến phân tích thực nghiệm, nghiên cứu này kỳ vọng sẽ cung cấp cơ

sở tham chiếu hữu ích cho cộng đồng nghiên cứu trong việc lựa chọn, đánh giá và phát triển các bộ dữ liệu phục vụ cho các mô hình phân tích cảm xúc đa phương thức hiện đại.

2 Tổng quan tài liệu

MSA tận dụng các bộ dữ liệu kết hợp các loại dữ liệu khác nhau như văn bản, âm thanh, hình ảnh và video, để diễn giải tình cảm chính xác hơn. Một số bộ dữ liệu được sử dụng rộng rãi nhất bao gồm:

- CMU-MOSI: một tập dữ liệu bao gồm các video ý kiến được chú thích bằng nhãn cảm xúc, bao gồm các phương thức văn bản, âm thanh và hình ảnh [4].
 - CMU-MOSEI: một phần mở rộng của MOSI, tập dữ liệu này bao gồm nhiều nhãn tình cảm và cảm xúc hơn, cung cấp dữ liệu theo ngữ cảnh phong phú hơn [5].
 - IEMOCAP: một tập dữ liệu chủ yếu được sử dụng để nhận dạng cảm xúc, chứa dữ liệu giọng nói theo kịch bản và tự phát cùng với các bản ghi video.
 - MUSTARD (MUSard): tập dữ liệu tập trung vào việc phát hiện hành vi mỉa mai trong bối cảnh đa phương thức, kết hợp dữ liệu văn bản, âm thanh và video [6].
- Các bộ dữ liệu này cung cấp các tài nguyên quý giá để đào tạo và đánh giá các mô hình trong MSA, nhưng chúng cũng có nhiều hạn chế cần được giải quyết. Các nghiên cứu trước đây về MSA đã sử dụng nhiều phương pháp khác nhau để tận dụng sự phong phú của dữ liệu đa phương thức. Các kỹ thuật phổ biến bao gồm: (i) Trích xuất tính năng từ dữ liệu văn bản, âm thanh, hình ảnh và kết hợp chúng thông qua phản ứng kết hợp sớm (nối các tính năng trước khi nhập mô hình) hoặc kết hợp muộn (kết hợp đầu ra mô hình). (ii) Sử dụng các mô hình như mạng thần kinh tích chập (CNN), mạng thần kinh hồi quy (RNN) và kiến trúc dựa trên Transformer để xử lý và tích hợp dữ liệu đa phương thức. (iii) Triển khai các cơ chế chú ý để cân nhắc tầm quan trọng của các phương thức khác nhau, tính năng của chúng một cách linh hoạt trong quá trình phân tích. (iv) Áp dụng các mô hình được đào tạo trước trên các tập dữ liệu lớn và tinh chỉnh chúng cho các nhiệm vụ MSA cụ thể [7].

Những kỹ thuật này đã góp phần tạo nên những tiến bộ đáng kể trong lĩnh vực này, nhưng vẫn còn những thách thức trong việc đạt được sự tích hợp và giải thích tối ưu dữ liệu đa phương thức. Một số hạn chế vẫn tồn tại trong lĩnh vực MSA: (i) Các bộ dữ liệu hiện tại thường thiếu tính đa dạng về mặt biểu thị nhân khẩu học, bối

cảnh và ngôn ngữ, điều này có thể dẫn đến các mô hình sai lệch không khái quát tốt cho các tình huống đa dạng trong thế giới thực. (ii) Việc tích hợp hiệu quả các nguồn dữ liệu không đồng nhất vẫn là một thách thức, với nhiều nghiên cứu đang nỗ lực cân bằng sự đóng góp của từng phương thức. (iii) Các chú thích, nhãn tình cảm không nhất quán hoặc chủ quan có thể ảnh hưởng đến độ tin cậy của bộ dữ liệu và kết quả mô hình. (iv) Nhiều mô hình đòi hỏi tính toán chuyên sâu, khiến chúng khó mở rộng và đặc biệt áp dụng trong các ứng dụng thời gian thực. Việc giải quyết những thách thức này là quan trọng để thúc đẩy lĩnh vực, phát triển các mô hình MSA có độ chính xác cao [8].

Chất lượng và tính đa dạng của bộ dữ liệu là quan trọng trong MSA vì một số lý do: (i) Các bộ dữ liệu đa dạng, chất lượng cao giúp giảm sai lệch trong các mô hình phân tích tình cảm. (ii) Tạo ra các mô hình khái quát

hóa tốt hơn cho dữ liệu mới, chưa được nhìn thấy, dữ liệu mờ, khiến chúng trở nên hữu ích hơn trong các ứng dụng thực tế. (iii) Dữ liệu chất lượng cao với các chú thích nhất quán cho phép phát triển các mô hình có thể diễn giải chính xác hơn những tình cảm phức tạp được thể hiện thông qua nhiều phương thức. (iv) Các bộ dữ liệu đa dạng cung cấp nền tảng vững chắc cho nghiên cứu trong tương lai, khám phá các kỹ thuật và mô hình mới cải thiện hiệu suất trong MSA [9, 10].

3 Phương pháp nghiên cứu

Nghiên cứu này được tiến hành theo một quy trình có hệ thống gồm bốn giai đoạn chính (Hình 1): (i) Thu thập và xử lý dữ liệu, (ii) Phân tích thống kê đặc trưng bộ dữ liệu, (iii) Đánh giá chất lượng bộ dữ liệu thông qua mô hình học máy, và (iv) Đề xuất hướng dẫn xây dựng bộ dữ liệu MSA chất lượng cao.



Hình 1 Quy trình phân tích và đánh giá bộ dữ liệu MSA.

3.1 Thu thập và xử lý dữ liệu

Để đảm bảo phạm vi đánh giá toàn diện, nghiên cứu này lựa chọn các bộ dữ liệu phân tích tình cảm đa phương thức (MSA) phổ biến dựa trên ba tiêu chí chính: (i) Các phương thức dữ liệu được cung cấp, bao gồm văn bản, âm thanh và hình ảnh/video. (ii) Nguồn gốc của dữ liệu, được tuyển chọn từ nhiều bối cảnh khác nhau như nội dung phim, vlog cá nhân, hội thoại thực tế và các cuộc trò chuyện trên mạng xã hội, nhằm phản ánh sự đa dạng và phức tạp của dữ liệu thực tế. (iii) Chất lượng nhãn cảm xúc, được đánh giá qua độ chính xác và tính nhất quán giữa các annotator trong quá trình gán nhãn cảm xúc.

Sau khi thu thập, dữ liệu được xử lý để chuẩn hóa các phương thức đầu vào, giúp đảm bảo tính đồng nhất giữa các modal. Cụ thể, các bước xử lý dữ liệu như sau:

- Văn bản: dữ liệu văn bản được xử lý qua các bước chuẩn hóa từ vựng, loại bỏ nhiễu (như ký tự đặc biệt và từ thừa), thực hiện tokenization và cuối cùng là vector hóa bằng mô hình ngôn ngữ BERT embeddings để giữ lại ngữ cảnh và đặc trưng ngữ nghĩa sâu của văn bản.
- Âm thanh: các đặc trưng âm thanh được trích xuất thông qua Mel-Frequency Cepstral Coefficients (MFCCs), pitch và năng lượng tín hiệu. Đây là các đặc

trung quan trọng giúp mô hình nhận diện các yếu tố cảm xúc trong giọng nói.

- Hình ảnh/video: đối với hình ảnh và video, quy trình xử lý bao gồm nhận diện khuôn mặt và trích xuất đặc trưng cảm xúc bằng mạng ResNet-50 và công cụ OpenFace, cho phép nhận diện các biểu cảm khuôn mặt và tương tác cảm xúc.

Quy trình này giúp đồng bộ hóa dữ liệu giữa các phương thức, đảm bảo rằng mỗi modal đóng góp một cách có ý nghĩa vào quá trình phân tích tình cảm đa phương thức. Việc chuẩn hóa các phương thức đầu vào này là yếu tố quan trọng để cải thiện độ chính xác và hiệu suất của mô hình phân tích tình cảm.

3.2 Phân tích thống kê bộ dữ liệu

Chúng tôi thực hiện phân tích thống kê trên các bộ dữ liệu được chọn nhằm đánh giá sự phân bố cảm xúc, tính cân bằng giữa các phương thức, và mối quan hệ giữa chúng. Cụ thể, ba yếu tố chính được xem xét trong quá trình phân tích này:

- Phân phối cảm xúc: một trong những yếu tố quan trọng trong phân tích dữ liệu là xác định sự chênh lệch giữa các lớp cảm xúc trong mỗi tập dữ liệu. Các tập dữ liệu có sự mất cân bằng nghiêm trọng giữa các lớp cảm xúc có thể làm sai lệch kết quả huấn luyện mô hình, vì mô hình sẽ có xu hướng dự đoán các lớp chiếm ưu thế

hơn. Chúng tôi kiểm tra phân phối của các lớp cảm xúc và tính toán các chỉ số thống kê như độ lệch chuẩn, trung bình, và các tỷ lệ phân phối để đánh giá mức độ mất cân bằng.

- Tính cân bằng giữa các phương thức: để đảm bảo không có phương thức nào chiếm ưu thế quá mức, chúng tôi so sánh tỷ lệ xuất hiện của dữ liệu giữa các modal (văn bản, âm thanh, hình ảnh/video). Việc có sự phân bố đều giữa các phương thức sẽ giúp mô hình phân tích tình cảm đa phương thức hoạt động hiệu quả hơn, tránh tình trạng một phương thức quá mạnh sẽ chi phối quá trình học của mô hình. Đánh giá này được thực hiện bằng cách tính toán tỷ lệ xuất hiện của mỗi modal và so sánh chúng trên các tập dữ liệu.

- Tương quan giữa các phương thức: để đánh giá mức độ liên kết giữa các modal, chúng tôi sử dụng hệ số tương quan Pearson. Hệ số tương quan này đo lường mối quan hệ tuyến tính giữa hai đặc trưng của hai phương thức khác nhau (ví dụ: đặc trưng âm thanh và văn bản). Công thức tính hệ số tương quan Pearson được cho bởi:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}}$$

trong đó X và Y là các đặc trưng của hai phương thức khác nhau (ví dụ: đặc trưng âm thanh và văn bản). Một giá trị r cao cho thấy sự tương quan mạnh giữa hai modal, có thể ảnh hưởng đến cách tích hợp thông tin trong mô hình [11].

3.3 Đánh giá bằng mô hình học máy

Để đánh giá chất lượng của các bộ dữ liệu MSA, tiến hành huấn luyện các mô hình học máy tiêu biểu trên từng tập dữ liệu và phân tích hiệu suất đạt được. Cụ thể, các mô hình được lựa chọn bao gồm: (i) Các mô hình Transformer hiện đại như BERT và RoBERTa cho xử lý văn bản, vốn nổi bật với khả năng nắm bắt ngữ cảnh sâu nhờ cơ chế chú ý; (ii) Mô hình Long Short-Term Memory (LSTM) cho chuỗi thời gian âm thanh, nhằm khai thác đặc trưng giọng nói và ngữ điệu qua thời gian; và (iii) Mô hình mạng tích chập đồ thị (GCN) được sử dụng để khai thác mối quan hệ giữa các phương thức dưới dạng đồ thị, giúp mô hình hóa sự tương tác liên phương thức một cách hiệu quả. Mỗi mô hình được huấn luyện và đánh giá theo ba tiêu chí chính: độ chính xác (Accuracy) – phản ánh tỷ lệ dự đoán đúng trên tập kiểm tra; F1-score – thước đo cân bằng giữa Precision và Recall, đặc biệt hữu ích trong bối cảnh dữ liệu mất

cân bằng giữa các lớp cảm xúc; và AUC-ROC (Area Under the Receiver Operating Characteristic Curve) – chỉ số đánh giá khả năng phân biệt giữa các nhãn cảm xúc, đặc biệt phù hợp cho các mô hình phân loại đa lớp. Kết quả thử nghiệm được so sánh giữa các mô hình trên cùng một tập dữ liệu và phân tích theo nhiều khía cạnh. Ví dụ, trên tập CMU-MOSEI, mô hình MULT cho kết quả F1-score vượt trội so với MISA, cho thấy khả năng học tương tác giữa các phương thức tốt hơn. Ngoài ra, đánh giá độ bền vững của mô hình trong các kịch bản thiếu hụt phương thức (ví dụ chỉ có văn bản, không có âm thanh) giúp xác định mức độ phụ thuộc của mô hình vào từng loại dữ liệu. Khả năng tổng quát hóa cũng được kiểm tra bằng cách huấn luyện trên một tập dữ liệu (ví dụ MOSI) và đánh giá trên tập khác (ví dụ MOSEI). Cuối cùng, phân tích lỗi cho thấy mô hình có xu hướng nhầm lẫn trong các tình huống chứa cảm xúc trung tính, hoặc khi các phương thức thể hiện thông tin không đồng bộ (ví dụ: giọng nói mang tính tích cực nhưng biểu cảm khuôn mặt lại tiêu cực). Những phân tích này giúp làm rõ điểm mạnh, điểm yếu của từng mô hình và đưa ra định hướng cải tiến trong các nghiên cứu tiếp theo.

3.4 Đề xuất hướng dẫn xây dựng bộ dữ liệu MSA chất lượng cao

Dựa trên các phân tích thống kê và kết quả đánh giá thực nghiệm, chúng tôi đề xuất một bộ tiêu chí nhằm hướng dẫn xây dựng bộ dữ liệu MSA có chất lượng cao. Các tiêu chí này hướng đến việc tối ưu hóa hiệu suất mô hình, đồng thời đảm bảo khả năng ứng dụng rộng rãi trong các bài toán phân tích tình cảm thực tiễn:

- Độ bao phủ tình cảm rộng: tập dữ liệu cần phản ánh đa dạng các trạng thái tình cảm, từ tích cực, tiêu cực đến trung tính, cũng như các sắc thái tinh vi hơn như thất vọng, hài lòng, mỉa mai hay phấn khích. Việc đảm bảo phổ cảm xúc rộng giúp mô hình học được sự khác biệt tinh tế giữa các biểu hiện cảm xúc, từ đó cải thiện độ chính xác phân loại.

- Cân bằng giữa các phương thức: một bộ dữ liệu chất lượng cần có sự phân bố tương đối đều giữa các phương thức (văn bản, âm thanh, hình ảnh/video). Việc mất cân bằng nghiêm trọng – ví dụ như văn bản có đầy đủ trong khi âm thanh hoặc hình ảnh bị thiếu – có thể dẫn đến mô hình học thiên lệch và làm giảm hiệu quả tích hợp liên phương thức.

- Độ đa dạng nguồn dữ liệu: sử dụng dữ liệu từ nhiều nguồn như phim điện ảnh, vlog cá nhân, hội thoại thường nhật hoặc nền tảng mạng xã hội giúp tăng tính đa dạng ngữ cảnh, cách thể hiện cảm xúc và đặc trưng ngôn ngữ. Điều này góp phần cải thiện khả năng tổng quát hóa của mô hình trên các tình huống thực tế khác nhau.

- Chất lượng nhãn cảm xúc: việc gán nhãn cảm xúc cần được thực hiện bởi nhiều annotator độc lập, kết hợp với các quy trình kiểm tra chéo và thống kê độ đồng thuận (như Cohen's Kappa). Điều này giúp tăng độ tin cậy và tính nhất quán của nhãn, giảm thiểu sai lệch do chủ quan trong đánh giá cảm xúc.

- Khả năng mở rộng và tái sử dụng: dữ liệu nên được lưu trữ theo định dạng chuẩn (ví dụ: JSON, CSV, hoặc TFRecord), kèm theo siêu dữ liệu chi tiết, chú thích rõ ràng và hướng dẫn sử dụng. Điều này không chỉ hỗ trợ khả năng mở rộng sang các nhiệm vụ khác như nhận diện biểu cảm khuôn mặt, nhận diện giọng nói cảm xúc, mà còn tạo điều kiện thuận lợi cho cộng đồng nghiên cứu chia sẻ và đánh giá lại mô hình trên cùng tập dữ liệu.

Tập hợp các tiêu chí trên cung cấp một khung tham chiếu thực tiễn cho các nhà nghiên cứu trong quá trình xây dựng hoặc lựa chọn bộ dữ liệu MSA. Việc tuân thủ các nguyên tắc này không chỉ góp phần nâng cao độ chính xác của mô hình, mà còn thúc đẩy sự phát triển bền vững và có thể tái sử dụng trong lĩnh vực phân tích tình cảm đa phương thức. Phần này phác thảo phương pháp toàn diện được sử dụng trong nghiên cứu để đánh giá các đặc điểm của bộ dữ liệu được sử dụng trong MSA. Khung đánh giá đề xuất được thiết kế để đánh giá và nâng cao chất lượng cũng như tính đại diện của các bộ dữ liệu này một cách có hệ thống. Cách tiếp cận của nghiên cứu bắt đầu bằng việc lựa chọn và quản lý cẩn thận các bộ dữ liệu đa phương thức có liên quan, đảm bảo tính đa dạng về mặt thể hiện nhân khẩu học, bối cảnh và nhân tình cảm. Sau đó là quá trình xử lý trước dữ liệu nghiêm ngặt để làm sạch, căn chỉnh và chuẩn hóa dữ liệu trên các phương thức văn bản, âm thanh và hình ảnh.

Để hiểu rõ hơn về các tập dữ liệu, nghiên cứu thực hiện phân tích thống kê chi tiết, tính toán số liệu thống kê mô tả và kiểm tra sự phân bố của nhân tình cảm. Kỹ thuật tính năng là một thành phần quan trọng, trong đó nghiên cứu trích xuất và tinh chỉnh các tính năng từ từng phương thức bằng cách sử dụng các kỹ thuật như nhúng từ cho văn bản, MFCC cho âm thanh và CNN cho dữ liệu hình ảnh. Sau đó, chúng tôi sử dụng nhiều

mô hình học máy khác nhau, bao gồm cả phương pháp học truyền thống và học sâu, để đánh giá hiệu quả của bộ dữ liệu. Hiệu suất của mô hình được đánh giá bằng cách sử dụng một bộ số liệu toàn diện, bao gồm độ chính xác, độ chính xác, khả năng thu hồi, điểm F1 và ROC-AUC, với xác thực chéo được sử dụng để đảm bảo tính mạnh mẽ và tránh trang bị quá mức.

Kết quả đánh giá của chúng tôi cho thấy sự phát triển của các hướng dẫn thực tế để xây dựng bộ dữ liệu đa phương thức chất lượng cao, giải quyết các vấn đề phổ biến như mất cân bằng dữ liệu và sự không nhất quán trong chú thích. Những hướng dẫn này nhằm mục đích hỗ trợ các nhà nghiên cứu và người thực hành tạo ra các bộ dữ liệu mang tính đại diện, đa dạng và có lợi cho việc xây dựng các mô hình phân tích cảm tính chính xác và có tính khái quát. Thông qua cách tiếp cận có cấu trúc và có phương pháp này, nghiên cứu của chúng tôi đóng góp những hiểu biết sâu sắc có giá trị và các phương pháp thực hành tốt nhất cho lĩnh vực phân tích cảm tính đa phương thức.

4 Kết quả và thảo luận

Kết quả thu được từ quá trình huấn luyện và đánh giá các mô hình học máy trên các tập dữ liệu phân tích tình cảm đa phương thức cho thấy những khác biệt rõ rệt về hiệu năng giữa các kiến trúc mô hình, chiến lược kết hợp phương thức và cách tiền xử lý dữ liệu. Trên tập dữ liệu CMU-MOSI, mô hình MULT đạt độ chính xác lên đến 81,2 %, vượt trội so với mô hình MISA (78,5 %) và các mô hình đơn phương thức như LSTM hoặc BERT đơn kênh, khoảng (72-74) % [12, 13]. Tương tự, trên tập CMU-MOSEI, chỉ số F1-score của MULT đạt 82,5 %, cho thấy khả năng học được mối tương quan chéo giữa các phương thức vượt trội so với các mô hình truyền thống sử dụng chiến lược kết hợp muộn hoặc mô hình đơn kênh [12]. Ngoài ra, chỉ số MAE trong các bài toán hồi quy mức độ cảm xúc cũng được cải thiện đáng kể, đặc biệt khi sử dụng các đặc trưng kết hợp từ cả văn bản, giọng nói và biểu cảm khuôn mặt.

Khi so sánh với các nghiên cứu trước đó, kết quả trong nghiên cứu này đạt được mức độ tương đương hoặc cao hơn so với các mô hình công bố gần đây như TFN (Tensor Fusion Network) [14], LMF (Low-rank Multimodal Fusion) [15], và MFM (Memory Fusion Model) [16]. Mặc dù các mô hình như MISA có ưu thế trong việc học biểu diễn đặc trưng riêng biệt và kháng nhiễu tốt, nhưng hiệu quả của chúng giảm đi rõ rệt khi

áp dụng trên tập dữ liệu lớn và đa dạng như MOSEI. MULT tỏ ra ổn định hơn nhờ cơ chế chú ý phương thức, cho phép mô hình học cách ưu tiên các nguồn thông tin quan trọng tại từng thời điểm [12].

Giải thích cho các kết quả trên, có thể thấy rằng việc kết hợp thông tin từ nhiều phương thức một cách hiệu quả là yếu tố then chốt để cải thiện hiệu năng trong phân tích tình cảm. Trong các trường hợp mà một phương thức bị nhiễu (ví dụ âm thanh rè hoặc biểu cảm mờ), các phương thức còn lại có thể hỗ trợ bù đắp, giúp mô hình đưa ra dự đoán chính xác hơn. Ngoài ra, các mô hình sử dụng kiến trúc chú ý như MULT không chỉ mang lại hiệu suất cao hơn mà còn dễ giải thích hơn nhờ khả năng trực quan hóa mức độ chú ý giữa các phương thức.

Từ những kết quả và phân tích trên, có thể khẳng định rằng việc phát triển các mô hình học sâu đa phương thức với cơ chế tương tác thông minh giữa các phương thức là hướng đi đầy hứa hẹn trong lĩnh vực phân tích tình cảm. Không chỉ giúp cải thiện độ chính xác, các mô hình này còn mở ra khả năng ứng dụng rộng rãi trong các hệ thống nhận diện cảm xúc, trợ lý ảo, và phân tích hành vi người dùng trong thời gian thực. Hơn nữa, các kết quả cũng đặt ra yêu cầu về việc chuẩn hóa tiền xử lý, đánh giá chéo tập dữ liệu, và khả năng mở

rộng mô hình để ứng dụng trong môi trường thực tế, nơi dữ liệu có thể thiếu hụt hoặc không đồng bộ.

5 Kết luận

Nghiên cứu này đã cung cấp một đánh giá hệ thống về các bộ dữ liệu phổ biến trong phân tích tình cảm đa phương thức, từ đó làm rõ các đặc điểm kỹ thuật như cấu trúc phương thức, phương pháp chú thích và mức độ tin cậy. Thông qua quy trình gồm bốn giai đoạn, kết quả cho thấy các bộ dữ liệu hiện tại tuy mang lại nhiều đóng góp quan trọng nhưng vẫn tồn tại những hạn chế nhất định về tính đa dạng, sự cân bằng giữa các phương thức, khả năng phản ánh chính xác bối cảnh thực tế. Hướng phát triển, việc xây dựng các tập dữ liệu chất lượng cao nên chú trọng đến việc tăng cường tính đại diện về ngôn ngữ, văn hóa và tình huống sử dụng, đảm bảo tính đồng bộ giữa các phương thức, đồng thời ứng dụng các kỹ thuật chú thích bán tự động để nâng cao tính nhất quán. Ngoài ra, việc phát triển các bộ dữ liệu mở, dễ truy cập, tương thích tốt với các kiến trúc học sâu hiện đại sẽ đóng vai trò quan trọng trong việc thúc đẩy các nghiên cứu MSA, hướng đến các hệ thống phân tích tình cảm có độ chính xác cao, khả năng tổng quát tốt, thích ứng hiệu quả với môi trường ứng dụng thực tiễn như trợ lý ảo, chăm sóc khách hàng và giáo dục thông minh.

Tài liệu tham khảo

1. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91, 424-444. <https://doi.org/10.1016/j.inffus.2022.09.025>
2. Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., ... & Yang, K. (2020, July). Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3718-3727).
3. Yusuf, A. A., Chong, F., & Xianling, M. (2022). An analysis of graph convolutional networks and recent datasets for visual question answering. *Artificial Intelligence Review*, 55(8), 6277-6300. DOI 10.1007/s10462-022-10151-2
4. Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2236-2246).
5. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42, 335-359. <https://doi.org/10.1007/s10579-008-9076-6>

6. Zhang, Y., Yu, Y., Wang, M., Huang, M., & Hossain, M. S. (2024). Self-adaptive representation learning model for multi-modal sentiment and sarcasm joint analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5), 1-17. <https://doi.org/10.1145/363531>
7. Liu, F., Chen, J., Li, K., Bai, J., Tan, W., Cai, C., & Ayub, M. S. (2024). STP-MFM: semi-tensor product-based multi-modal factorized multilinear pooling for information fusion in sentiment analysis. *Digital Signal Processing*, 145, 104265. <https://doi.org/10.1016/j.dsp.2023.104265>
8. Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence*, 133, 108339. <https://doi.org/10.1016/j.engappai.2024.108339>
9. Liu, X., Li, S., & Wang, M. (2021). Hierarchical attention-based multimodal fusion network for video emotion recognition. *Computational Intelligence and Neuroscience*, 2021(1), 5585041. <https://doi.org/10.1155/2021/5585041>
10. Yang, J., Yu, Y., Niu, D., Guo, W., & Xu, Y. (2023, July). Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7617-7630).
11. Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1), 59-66. <https://doi.org/10.1080/00031305.1988.10475524>
12. Xu, H. (2023). Multimodal sentiment analysis. In *Multi-Modal Sentiment Analysis* (pp. 217-240). Singapore: Springer Nature Singapore. <https://doi.org/10.1007/978-981-99-5776-7>
13. Hazarika, D., Zimmermann, R., & Poria, S. (2020, October). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1122-1131).
14. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*. <https://doi.org/10.18653/v1/D17-1115>
15. Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*. <https://doi.org/10.48550/arXiv.1806.00064>
16. Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. P. (2018, April). Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Evaluation of the characteristics of datasets used in multi-modal sentiment analysis

Nguyen Xuan Cuong^{1,*}, Do Hoang Nam^{2,**}

Faculty of Information Technology, Nguyen Tat Thanh University

*nxcuong@ntt.edu.vn, **namdh@ntt.edu.vn

Abstract Multimodal Sentiment Analysis (MSA) has increasingly attracted research attention. However, existing datasets still suffer from limitations in terms of comprehensiveness, diversity, and consistency. This study provides a systematic review of widely-used MSA datasets by analyzing key characteristics such as data modalities, annotation methods, and reliability. A four-stage evaluation framework was proposed: (1) survey of dataset collection and management, (2) statistical and technical feature analysis, (3) performance evaluation of models on these datasets, and (4) development of guidelines for constructing high-quality MSA datasets. The results offer in-depth insights into the strengths and weaknesses of each dataset, thereby supporting researchers to select suitable data sources and guide the development of more robust multimodal datasets in the future.

Keywords Multimethod sentiment analysis, data evaluation, feature extraction.

