

MULTIMODAL SARCASM DETECTION ON VIETNAMESE SOCIAL MEDIA DATASETS

Pham Trung Tin, Tran Thanh Son
Nguyen Duc Vu, Huynh Van Tin

Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam

Vietnam National University, Ho Chi Minh City, Vietnam

{21522678, 21522557}@gm.uit.edu.vn

vund@uit.edu.vn, tinhv@uit.edu.vn

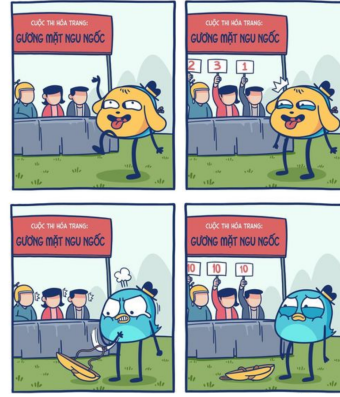
Abstract

Châm biếm là một hiện tượng ngôn ngữ phức tạp, thường được sử dụng để thể hiện ý kiến trái ngược với những gì được nói ra, thường với mục đích hài hước hoặc chỉ trích. Trong bối cảnh sự phát triển mạnh mẽ của nội dung đa phương tiện trên các nền tảng mạng xã hội như Facebook và Instagram, việc phát hiện châm biếm trở nên ngày càng phức tạp do sự kết hợp giữa văn bản, hình ảnh và các biểu tượng cảm xúc. Bài báo cáo này trình bày một nghiên cứu về việc phát hiện châm biếm đa phương thức, sử dụng dữ liệu từ cuộc thi DSC UIT Challenge - Bảng B, với mô hình học sâu kết hợp giữa các đặc trưng ngôn ngữ và hình ảnh. Chúng tôi áp dụng Vision Transformer (ViT) và ResNet152 để trích xuất đặc trưng hình ảnh, trong khi ViBERT và XLM-RoBERTa được sử dụng để nắm bắt đặc trưng ngôn ngữ. Mô hình đa phương thức cho thấy hiệu quả vượt trội so với các mô hình đơn phương thức, với F1-score cao nhất đạt được là 0.424. Kết quả thực nghiệm cho thấy việc kết hợp thông tin từ nhiều modal giúp nắm bắt tốt hơn các ngữ cảnh châm biếm phức tạp, đồng thời khẳng định vai trò quan trọng của đặc trưng ngôn ngữ trong phát hiện châm biếm. Bài báo cáo cũng thảo luận về các thách thức và hướng phát triển trong tương lai cho nghiên cứu này, nhấn mạnh tầm quan trọng của việc phát triển các công cụ tự động hóa để nhận diện châm biếm trong nội dung đa phương tiện.

1 Introduction

Châm biếm là một hình thức giao tiếp đặc biệt, thường được sử dụng để thể hiện sự mỉa mai hoặc chỉ trích một cách tinh tế. Nó không chỉ đơn thuần là việc nói ra một điều gì đó mà còn bao hàm một ý nghĩa ngược lại, thường nhằm mục đích tạo ra tiếng cười hoặc thể hiện sự không đồng tình. Trong thời đại số, với sự

gia tăng của nội dung đa phương tiện trên các nền tảng mạng xã hội, châm biếm đã trở thành một phần không thể thiếu trong giao tiếp trực tuyến. Người dùng thường kết hợp văn bản, hình ảnh, video để truyền tải thông điệp châm biếm, tạo ra những ngữ cảnh phức tạp mà việc phát hiện trở nên khó khăn hơn.



Caption: Kệ thắng là được

Label: Multi-Sarcasm

Hình 1: Ảnh minh họa trường hợp một bài viết châm biếm đa phương tiện. Caption "Kệ thắng là được" mang sắc thái khá bình thản, đơn giản nhưng tự tin. Trong khi hình ảnh thể hiện sắc thái rất thất vọng vì đã thắng theo một cách không hề mong muốn. Điều này thể hiện sự không nhất quán giữa hình ảnh và văn bản

Ngày nay, các ứng dụng mạng xã hội như Facebook cho phép người dùng đăng tải các bài viết đa phương tiện, gồm cả hình ảnh và văn bản. Từ đó, châm biếm có thể xuất hiện không chỉ trong văn bản hay hình ảnh mà có thể là sự kết hợp giữa cả hai, làm cho việc nhận biết sự châm biếm trở nên phức tạp hơn, như ví dụ ở Figure 1. Vì vậy, việc nắm bắt sự không nhất quán giữa các phương thức là rất quan trọng cho việc phát hiện sự châm biếm đa phương thức. Phương pháp phát hiện châm biếm đa phương tiện hiện tại chủ yếu kết hợp các đặc trưng từ nhiều phương thức mà chưa chú trọng đến sự tương tác phức tạp giữa chúng.

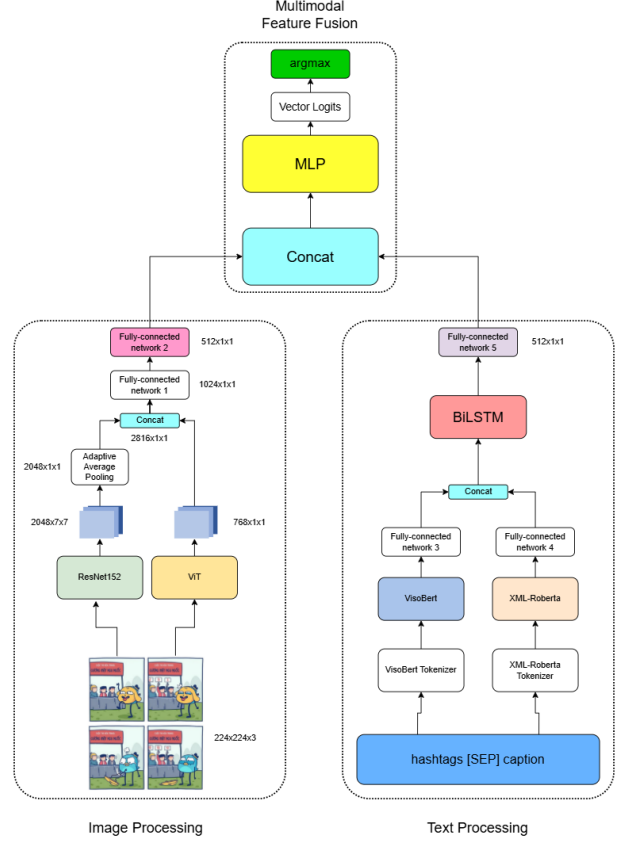
Nghiên cứu trước đây đã chỉ ra rằng châm biếm thường liên quan đến một khái niệm gọi là sự không nhất quán, được sử dụng để gợi ý sự phân biệt giữa thực tế và kỳ vọng. Các nhà nghiên cứu đã phát triển nhiều phương pháp để phát hiện châm biếm, chủ yếu tập trung vào việc phân tích văn bản. Tuy nhiên, với sự phát triển của nội dung đa phương tiện, việc chỉ dựa vào văn bản để phát hiện châm biếm là không đủ. Các phương pháp hiện tại chủ yếu kết hợp các đặc trưng từ nhiều phương thức mà chưa chú trọng đến sự tương tác phức tạp giữa chúng.

Bài báo cáo này nhằm mục đích phát triển một mô hình đa phương thức kết hợp giữa văn bản và hình ảnh để phát hiện châm biếm. Chúng tôi sử dụng các mô hình học sâu tiên tiến như Vision Transformer (ViT) và ResNet-152 để trích xuất đặc trưng từ hình ảnh, trong khi VisoBERT và XLM-RoBERTa được áp dụng để mã hóa văn bản. Mô hình của chúng tôi không chỉ tập trung vào việc trích xuất các đặc trưng riêng lẻ mà còn nắm bắt các mối quan hệ phức tạp giữa văn bản và hình ảnh, từ đó cải thiện độ chính xác trong việc phân loại các loại châm biếm.

Chúng tôi sẽ trình bày chi tiết về phương pháp nghiên cứu, tập dữ liệu sử dụng, các tham số huấn luyện, và kết quả thực nghiệm. Kết quả cho thấy mô hình đa phương thức không chỉ đạt hiệu quả cao hơn so với các mô hình đơn phương thức mà còn mở ra hướng đi cho nghiên cứu phát hiện châm biếm trong bối cảnh đa phương tiện. Bài báo cáo cũng thảo luận về các thách thức trong việc phát triển các công cụ tự động hóa để nhận diện châm biếm, cũng như các hướng nghiên cứu tiềm năng trong tương lai, nhằm nâng cao khả năng hiểu biết và phân tích ngữ nghĩa trong nội dung đa phương tiện.

2 Method

Trong phần này, trước tiên chúng tôi định nghĩa nhiệm vụ multi-modal sarcasm detection. Sau đó, chúng tôi trình bày ngắn gọn về background của các mô hình ViSoBERT (VBERT), XLM-Roberta (XMLR), ViT (Vision Transformer), ResNet-152 (R152) và mô tả chi tiết kiến trúc của mô hình được đề xuất. Hình 2 cung cấp cái nhìn tổng quan về mô hình của chúng tôi.



Hình 2: Overall Model Architecture

2.1 Task Definition

Multi-modal sarcasm classification là một nhiệm vụ phân loại nhằm xác định loại hình châm biếm trong một cặp dữ liệu gồm caption và hình ảnh. Cụ thể, với một tập dữ liệu đa phương thức D , mỗi mẫu $d \in D$ bao gồm một đoạn caption T chứa n từ $\{t_1, t_2, t_3, \dots, t_n\}$ và một hình ảnh liên kết I . Mục tiêu là xây dựng một bộ phân loại có khả năng dự đoán một trong bốn nhãn sau cho các mẫu chưa từng thấy:

- **Not-sarcasm:** caption-hình ảnh không mang ý nghĩa châm biếm.
- **Image-sarcasm:** Ý nghĩa châm biếm được truyền tải chủ yếu qua hình ảnh.
- **Text-sarcasm:** Ý nghĩa châm biếm được truyền tải chủ yếu qua đoạn caption.
- **Multi-sarcasm:** Ý nghĩa châm biếm được tạo nên từ sự kết hợp giữa đoạn caption và hình ảnh.

Nhiệm vụ này yêu cầu phát triển một mô hình *multimodal classification* có khả năng khai thác

hiệu quả các đặc trưng từ cả văn bản T và hình ảnh I , đồng thời tích hợp chúng để nhận diện chính xác các loại châm biếm. Quá trình này không chỉ bao gồm việc trích xuất các đặc trưng từ từng phương thức riêng lẻ, mà còn đòi hỏi mô hình nắm bắt được các mối quan hệ phức tạp giữa hai loại dữ liệu để thực hiện phân loại chính xác.

2.2 Background

Việc tiền huấn luyện (pretraining) các mô hình ngôn ngữ đã được chứng minh là hữu ích cho nhiều bài toán xử lý ngôn ngữ tự nhiên. Trong nghiên cứu này, chúng tôi kết hợp các mô hình hiện đại, bao gồm cả ngôn ngữ và hình ảnh, để giải quyết bài toán phát hiện và phân loại châm biếm đa phương tiện. Cụ thể, các mô hình sử dụng bao gồm:

- **ViSoBERT (VBERT)**: Mô hình ngôn ngữ tiền huấn luyện đơn ngữ đầu tiên được thiết kế cho văn bản tiếng Việt trên mạng xã hội. ViSoBERT được huấn luyện trên một tập dữ liệu lớn và đa dạng từ văn bản mạng xã hội tiếng Việt, sử dụng kiến trúc XLM-R. Mô hình này được tối ưu hóa cho các tác vụ như nhận diện cảm xúc, phát hiện ngôn từ thù địch, phân tích cảm xúc, và phát hiện bình luận spam.
- **XLM-Roberta (XMLR)**: Mô hình đa ngôn ngữ mạnh mẽ, tiền huấn luyện trên dữ liệu ngôn ngữ không đồng nhất. XLM-Roberta được sử dụng để cải thiện chất lượng biểu diễn ngữ nghĩa văn bản tiếng Việt.
- **ViT (Vision Transformer)**: Mô hình học sâu xử lý ảnh tiên tiến, được tiền huấn luyện trên tập dữ liệu lớn, phù hợp để trích xuất đặc trưng từ hình ảnh.
- **ResNet-152 (R152)**: Một biến thể sâu của mạng nơ-ron tích chập (CNN), ResNet-152 được sử dụng để giảm thiểu vấn đề biến mất đạo hàm trong quá trình huấn luyện và trích xuất đặc trưng từ hình ảnh.

Chúng tôi sử dụng kiến trúc kết hợp các mô hình trên:

- **Xử lý hình ảnh**: ResNet-152 và ViT trích xuất đặc trưng từ hình ảnh, sau đó hợp nhất qua một mạng nơ-ron để tạo biểu diễn cuối cùng.

- **Xử lý văn bản**: ViSoBERT và XLM-Roberta mã hóa văn bản, đầu ra từ hai mô hình được kết hợp và đưa qua LSTM để cải thiện biểu diễn.
- **Kết hợp đa phương thức**: Biểu diễn từ phần hình ảnh và văn bản được kết hợp thông qua các tầng mạng, và dự đoán nhãn đầu ra thông qua một bộ phân loại.

2.3 Model Architecture

Trong báo cáo này, chúng tôi đề xuất một mô hình phân loại đa phương thức (MMP), được thiết kế để xử lý đồng thời thông tin từ hình ảnh và văn bản, nhằm phát hiện sự mỉa mai trong nội dung đa phương thức. Kiến trúc mô hình bao gồm ba thành phần chính:

- **Xử lý hình ảnh (Image Processing)**
- **Xử lý văn bản (Text Processing)**
- **Hợp nhất đặc trưng đa phương thức (Multimodal Feature Fusion)**

Mô hình sử dụng các phương pháp hiện đại để trích xuất và hợp nhất đặc trưng từ hai phương thức nhằm đạt được khả năng phân loại tối ưu.

2.3.1 Xử Lý Hình Ảnh (Image Processing)

Mô hình xử lý hình ảnh sử dụng hai kiến trúc mạnh mẽ là ResNet-152 (R152) và Vision Transformer (ViT) để trích xuất đặc trưng từ các khu vực khác nhau và biểu diễn ngữ nghĩa toàn cục của hình ảnh.

Trích xuất đặc trưng từ ResNet-152

Hình ảnh đầu vào được thay đổi kích thước về 224×224 pixel và được đưa qua R152, bỏ qua tầng fully-connected cuối cùng. Các đặc trưng từ tầng cuối được tổng hợp bằng hàm Adaptive Average Pooling thành một vector:

$$\mathbf{F}_{R152} = \text{AdaptivePool}(R152(\mathbf{I})) \in R^{2048}, \quad (1)$$

trong đó \mathbf{I} là hình ảnh đầu vào và \mathbf{F}_{R152} là vector đặc trưng kích thước 2048.

Trích xuất đặc trưng từ Vision Transformer (ViT)

Hình ảnh được chia thành các 16×16 pixel

patch và mỗi patch được chuyển đổi thành token. Token $[CLS]$ từ ViT biểu diễn ngữ nghĩa toàn cục của hình ảnh:

$$\mathbf{F}_{ViT} = ViT(\mathbf{I})[CLS] \in R^{768}. \quad (2)$$

Hợp nhất đặc trưng hình ảnh

Để tận dụng cả hai đặc trưng, chúng tôi ghép nối vector từ R152 và ViT, sau đó đưa qua một mạng fully-connected (FC) để giảm chiều:

$$\mathbf{F}_{img} = FC(concat(\mathbf{F}_{R152}, \mathbf{F}_{ViT})) \in R^{512}. \quad (3)$$

2.3.2 Xử Lý Văn Bản (Text Processing)

Phần xử lý văn bản sử dụng hai mô hình ngôn ngữ tiên tiến là VisoBERT (VBERT) và XLM-Roberta (XLMR), được huấn luyện trước để trích xuất đặc trưng từ văn bản đầu vào (caption và hashtags).

Trích xuất đặc trưng từ VisoBERT và XLM-Roberta

Caption và hashtags sau khi tokenizer được đưa vào hai mô hình. Token $[CLS]$ từ mỗi mô hình được chọn làm đại diện ngữ nghĩa toàn cục:

$$\mathbf{F}_{VBERT} = VBERT(X)[CLS] \in R^{768}, \\ \mathbf{F}_{XLMR} = XLMR(X)[CLS] \in R^{1024}.$$

Sau đó, hai vector này được chiếu về không gian 256 chiều thông qua các tầng fully-connected (FC):

$$\mathbf{F}_{VBERT} = FC_{VBERT}(\mathbf{F}_{VBERT}) \in R^{256}, \\ \mathbf{F}_{XLMR} = FC_{XLMR}(\mathbf{F}_{XLMR}) \in R^{256}.$$

Mã hoá tuần tự bằng LSTM

Hai vector đặc trưng sau khi được ghép nối được đưa qua một mạng LSTM hai chiều (BiLSTM) để khai thác quan hệ ngữ cảnh tuần tự:

$$\mathbf{F}_{LSTM} = BiLSTM(concat(\mathbf{F}_{VBERT}, \mathbf{F}_{XLMR})) \in R^{512}.$$

Kết quả cuối cùng được giảm chiều qua một tầng fully-connected:

$$\mathbf{F}_{txt} = FC(\mathbf{F}_{LSTM}) \in R^{512}. \quad (4)$$

2.3.3 Hợp Nhất Đặc Trưng Đa Phương Thức (Multimodal Feature Fusion)

Sau khi trích xuất đặc trưng từ hình ảnh và văn bản, chúng tôi hợp nhất chúng thành một vector tổng hợp để đưa vào các tầng phân loại.

Ghép nối đặc trưng

Các đặc trưng từ hình ảnh và văn bản được ghép nối:

$$\mathbf{F}_{fusion} = concat(\mathbf{F}_{img}, \mathbf{F}_{txt}) \in R^{1024}. \quad (5)$$

Vector tổng hợp \mathbf{F}_{fusion} sau đó được đưa qua một chuỗi các tầng fully-connected (FC) để giảm chiều và cuối cùng tính toán logits:

$$\mathbf{F}_{hidden} = FC_1(\mathbf{F}_{fusion}) \in R^{512}$$

$$\mathbf{F}_{output} = FC_2(\mathbf{F}_{hidden}) \in R^C$$

Trong đó C là số nhãn (num_labels) của bài toán phân loại. Tầng FC thứ hai sẽ tạo ra vector logits với kích thước C .

2.4 Prediction

Sau khi tính toán được logits, chúng ta sử dụng hàm $\arg \max$ để chọn lớp có xác suất cao nhất và dự đoán nhãn cho mỗi mẫu. Công thức tính toán nhãn dự đoán là:

$$\mathbf{y}_{pred} = \arg \max(\mathbf{F}_{output}, dim = 1)$$

Trong đó:

- \mathbf{F}_{output} là vector logits đầu ra từ tầng phân loại cuối cùng.
- $\arg \max(\mathbf{F}_{output}, dim = 1)$ sẽ trả về chỉ số của phần tử có giá trị lớn nhất trong vector logits, tương ứng với lớp có xác suất cao nhất. Đây là nhãn mà mô hình dự đoán cho mỗi mẫu trong batch.

2.5 Training objectives

Hàm mất mát được tính bằng Cross-Entropy Loss để tối ưu hoá quá trình học:

$$\mathcal{L} = CrossEntropy(\mathbf{F}_{output}, \mathbf{y}), \quad (6)$$

trong đó \mathbf{y} là nhãn thực của mẫu dữ liệu.

2.6 Model Evaluation Metrics

Do tính chất mất cân bằng của dữ liệu, đặc biệt ở các nhãn text-sarcasm và image-sarcasm, chúng tôi sử dụng phương pháp macro-averaging để đánh giá hiệu suất của mô hình. Phương pháp này cho phép đánh giá công bằng trên tất cả các lớp, không phụ thuộc vào số lượng mẫu trong từng lớp.

2.6.1 Macro-Averaged Metrics

Các metric được tính toán độc lập cho từng lớp trước khi lấy trung bình:

$$Precision(Macro) = \frac{1}{C} \sum_{i=1}^C Precision_i \quad (7)$$

$$Recall(Macro) = \frac{1}{C} \sum_{i=1}^C Recall_i \quad (8)$$

$$F1(Macro) = \frac{1}{C} \sum_{i=1}^C F1 - Score_i \quad (9)$$

Trong đó:

- C là tổng số lớp (trong trường hợp này $C = 4$)
- $Precision_i$, $Recall_i$, và $F1 - Score_i$ là các metric được tính cho lớp thứ i

Cho mỗi lớp i , các metric được tính như sau:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (10)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (11)$$

$$F1 - Score_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (12)$$

Trong đó:

- TP_i (True Positive): Số mẫu được phân loại đúng vào lớp i
- FP_i (False Positive): Số mẫu được phân loại nhầm vào lớp i
- FN_i (False Negative): Số mẫu thuộc lớp i nhưng bị phân loại nhầm vào lớp khác

Việc sử dụng macro-averaging có những ưu điểm sau:

- Đối xử công bằng với tất cả các lớp, không phụ thuộc vào số lượng mẫu

- Phản ánh chính xác hiệu suất của mô hình trên các lớp thiểu số (text-sarcasm và image-sarcasm)
- Cung cấp đánh giá toàn diện về khả năng phân loại của mô hình trên tất cả các lớp

3 Experiment

3.1 Dataset

Nghiên cứu này sử dụng tập dữ liệu châm biếm đa phương tiện **ViMMSD** (Vietnamese Multimodal Sarcasm Detection Dataset), được phát hành trong khuôn khổ cuộc thi **DSC UIT Challenge - Bảng B**. Tập dữ liệu được thiết kế để phát hiện sắc thái mỉa mai trong nội dung đa phương tiện, bao gồm các thành phần hình ảnh và văn bản.

Định nghĩa bài toán

Bài toán đặt ra là phát hiện và phân loại nội dung mỉa mai dựa trên đầu vào là một cặp đa phương tiện:

- **Hình ảnh (I)**: một bức ảnh thuộc tập hợp **images**.
- **Văn bản (C)**: một đoạn văn bản mô tả (caption) liên quan đến hình ảnh.

Đầu ra là một nhãn thuộc tập hợp gồm bốn loại: $\{multi-sarcasm, image-sarcasm, text-sarcasm, non-sarcasm\}$.

Phân bố dữ liệu

Tập dữ liệu được chia thành ba tập con để phục vụ các giai đoạn huấn luyện và đánh giá như sau:

- **Training set**: gồm 10,805 mẫu, được sử dụng để huấn luyện mô hình.
- **Public test set**: gồm 1,413 mẫu, được sử dụng để đánh giá hiệu suất trong quá trình phát triển.
- **Private test set**: gồm 1,504 mẫu, được sử dụng để đánh giá cuối cùng trong cuộc thi.

Phân bố nhãn trong tập huấn luyện

Tập huấn luyện bao gồm các mẫu được phân bố theo các nhãn như sau:

- *non-sarcasm*: 6,062 mẫu (56.1%).

- *multi-sarcasm*: 4,224 mẫu (39.1%).
- *image-sarcasm*: 442 mẫu (4.1%).
- *text-sarcasm*: 77 mẫu (0.7%).

Sự mất cân bằng giữa các nhãn cho thấy tính thách thức của bài toán, đặc biệt đối với các nhãn có tần suất thấp như *text-sarcasm* và *image-sarcasm*.

Ví dụ mẫu

Bảng 1 minh họa một số mẫu từ tập dữ liệu ViMMSD. Mỗi mẫu bao gồm một cặp hình ảnh và chú thích, kèm theo nhãn phân loại.

3.2 Baseline Models

Trong nghiên cứu này, chúng tôi xây dựng và đánh giá các mô hình cơ sở để phát hiện châm biếm. Mô hình đa phương thức: Kết hợp thông tin từ văn bản và hình ảnh thông qua cơ chế fusion. Trong đó:

- **Văn bản:** Tận dụng PhoBERT, ViSoBERT (đơn ngôn ngữ) và mBERT, XLM-RoBERTa (đa ngôn ngữ)
- **Hình ảnh:** Sử dụng Vision Transformer (ViT) và ResNet152 để trích xuất đặc trưng tổng quát và chi tiết từ hình ảnh, tối ưu hóa cho nội dung mạng xã hội tiếng Việt.

Mô hình chỉ dùng văn bản sử dụng các đặc trưng từ các mô hình pre-trained qua fully-connected layer để tích hợp các thông tin ngôn ngữ. Mô hình chỉ dùng hình ảnh sử dụng Vision Transformer (ViT) và ResNet152, trích xuất đặc trưng khác biệt, sau đó kết hợp qua fully-connected layer. Mô hình đa phương thức kết hợp đặc trưng văn bản và hình ảnh thông qua lớp fusion fully-connected layer, trực tiếp biểu diễn trong không gian chung.

4 Thử nghiệm Xử lý Văn bản

4.1 Mục tiêu

Mục tiêu của thử nghiệm này là so sánh hiệu quả của các mô hình pretrain khác nhau trong việc xử lý văn bản. Chúng tôi tập trung đánh giá khả năng biểu diễn và phân loại ngữ nghĩa của các mô hình.

4.2 Mô hình Sử dụng

Các mô hình được thử nghiệm bao gồm:

- **Mô hình đơn ngôn ngữ:** PhoBERT, ViSoBERT.
- **Mô hình đa ngôn ngữ:** mBERT, XLM-RoBERTa.

4.3 Thiết lập Kích thước Vector Đặc trưng Đầu ra

Để đảm bảo tính nhất quán và hiệu quả, chúng tôi thiết lập đầu ra của các mô hình như sau:

- Với một mô hình pretrain, đặc trưng đầu ra được chuyển qua tầng fully-connected (FC) để giảm về không gian vector 512 chiều.
- Với hai mô hình kết hợp, đặc trưng đầu ra của từng mô hình được chuyển qua các tầng FC riêng biệt để giảm về không gian vector 256 chiều. Sau đó, các vector này được ghép nối để tạo thành vector đầu ra có kích thước 512 chiều.

4.4 Experimental Settings

Dữ liệu

Tập dữ liệu ViMMSD (Vietnamese Multimodal Sarcasm Detection) được chia theo tỷ lệ 80:20 cho tập huấn luyện và phát triển, với các nhãn: *không châm biếm*, *châm biếm đa phương thức*, *châm biếm hình ảnh*, và *châm biếm văn bản*.

- **Văn bản:** Caption và hashtag.
- **Hình ảnh:** Đi kèm với văn bản.
- **Nhãn phân loại:** Một trong bốn nhãn trên.


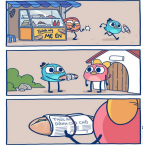
Tiền xử lý

- **Văn bản:** Tokenize bằng ViSoBERT và XLM-RoBERTa, giới hạn 128 token.
- **Hình ảnh:** Chuẩn hóa kích thước 224x224 và giá trị pixel theo ImageNet.

Tham số huấn luyện

Bảng 2 trình bày các tham số huấn luyện cho quá trình training models. Optimizer AdamW và Focal Loss được sử dụng để giảm mất cân bằng dữ liệu. Mô hình được huấn luyện trên GPU NVIDIA Tesla P100.

Bảng 1: Ví dụ minh họa từ tập dữ liệu ViMMSD.

Hình ảnh	Chú thích	Nhãn
	Tóm tắt 3 năm trong 1 bức hình. Hy vọng vào 2023 Cre on pic.	<i>non-sarcasm</i>
	CHÚNG NÓ BIẾT HẾT ĐẤY! Loài quạ có thể nhận diện con người và chúng có hành vi “tặng quà”. Quạ hoang thường sẽ đem những món đồ lấp lánh... - Kiến Know -	<i>multi-sarcasm</i>
	Bảo sao mấy chó ở quê hay chạy đuổi mình chả bao giờ vẫy đuôi. - Kiến Know -	<i>text-sarcasm</i>
	con trai bà bán bánh bánh bánh mì	<i>image-sarcasm</i>

Tham số	Giá trị
Batch Size	32
Learning rate	5×10^{-5}
Weight decay	0.01
Epoch	8
Early stopping	Patience = 3

Bảng 2: Tham số huấn luyện

Mô hình	Precision	Recall	F1-score
ViSoBERT	41.7	44.3	41.6
mBERT	35.1	38.1	35.2
PhoBERT	37.2	39.3	37.3
XLM-R	37.5	40.2	37.6
ViSoBERT + XLM-R	42.0	44.8	42.4

Bảng 3: Kết quả trên tập Private Test

4.5 Kết quả thực nghiệm

Bảng 3 cho thấy kết quả thực nghiệm cho thấy mô hình đa phương thức đạt hiệu quả cao hơn so với các mô hình đơn phương thức.

Nhận xét

- ViSoBERT đạt hiệu quả cao nhất trong các mô hình đơn ngôn ngữ, với F1-score đạt 41.6. Điều này cho thấy ViSoBERT rất phù hợp với dữ liệu mạng xã hội tiếng Việt, vốn thường chứa nhiều yếu tố đặc

trưng như teencode, emoji và phong cách văn phong phi chính thức.

- Trong nhóm mô hình đa ngôn ngữ, XLM-RoBERTa vượt trội hơn mBERT với F1-score đạt 37.6, nhờ khả năng biểu diễn ngữ nghĩa mạnh mẽ và hỗ trợ nhiều ngôn ngữ khác nhau.
- Khi kết hợp đặc trưng từ ViSoBERT và XLM-RoBERTa, F1-score tăng lên 42.4,

vượt trội so với việc chỉ sử dụng riêng lẻ từng mô hình. Điều này cho thấy sự bổ trợ giữa biểu diễn ngữ nghĩa đơn ngôn ngữ của ViSoBERT và khả năng xử lý ngữ cảnh đa ngôn ngữ của XLM-RoBERTa. Việc kết hợp này đặc biệt hiệu quả khi dữ liệu chứa các ngôn ngữ khác ngoài tiếng Việt.

- Tầng fully-connected giúp giảm kích thước vector đặc trưng mà vẫn giữ được các thông tin quan trọng, tạo điều kiện cho việc kết hợp các mô hình với nhau một cách hiệu quả và ổn định.

5 Related Work

5.1 Text-based Sarcasm detection

Việc nhận diện mỉa mai dựa trên văn bản đã thu hút sự quan tâm đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên. Mỉa mai thường thể hiện qua sự khác biệt giữa ý nghĩa bề mặt và ý nghĩa ngụ ý, đòi hỏi các mô hình phải phân tích ngữ cảnh và cảm xúc, dấu hiệu ngôn ngữ, và các yếu tố ngữ nghĩa để phát hiện chính xác.

Tuy nhiên, Bamman và Smith (2015) chỉ ra rằng việc kết hợp thông tin ngôn ngữ với các yếu tố ngoài ngôn ngữ, chẳng hạn như đặc điểm của người nói, người nghe, và môi trường giao tiếp, có thể cải thiện đáng kể độ chính xác của mô hình. Phương pháp này không chỉ tăng hiệu quả nhận diện mà còn làm sáng tỏ các yếu tố tương tác cá nhân góp phần tạo nên hiện tượng mỉa mai trong hội thoại.

Ngoài việc kết hợp thông tin ngoài ngôn ngữ, các nghiên cứu gần đây đã tập trung vào việc sử dụng các phương pháp học sâu để tự động trích xuất đặc trưng từ dữ liệu văn bản. Zhang và đồng đội (2016) đã áp dụng mạng nơ-ron hồi tiếp hai chiều để nắm bắt thông tin ngữ pháp và ngữ nghĩa từ các tweet, đồng thời sử dụng một mạng nơ-ron pooling để tự động trích xuất ngữ cảnh từ các tweet trước đó. Kết quả nghiên cứu cho thấy các đặc trưng tự động từ mạng nơ-ron mang lại hiệu suất cao hơn so với các đặc trưng thủ công truyền thống, đồng thời có sự phân bố lỗi khác biệt, góp phần cung cấp góc nhìn mới về bài toán nhận diện mỉa mai.

Ngoài ra, Mishra và đồng đội (2017) đã mở rộng hướng tiếp cận bằng cách nghiên cứu dữ liệu từ các hành vi. Cụ thể, họ đề xuất một hệ

thống NLP tích hợp các đặc trưng nhận thức từ dữ liệu di chuyển ánh mắt và dữ liệu ánh mắt (gaze data) của người đọc khi tiếp nhận văn bản. Hệ thống này sử dụng mạng nơ-ron tích chập (CNN) để tự động trích xuất các đặc trưng từ cả văn bản và dữ liệu ánh mắt, thay vì dựa vào các đặc trưng được thiết kế thủ công. Kết quả thực nghiệm trên các tập dữ liệu nhân cảm xúc và mỉa mai cho thấy việc kết hợp tự động các đặc trưng từ văn bản và ánh mắt mang lại hiệu suất phân loại cao hơn so với các hệ thống chỉ sử dụng văn bản hoặc các đặc trưng thủ công trước đó.

5.2 Multi-Modal Sarcasm Detection

Với sự phát triển mạnh mẽ của nội dung đa phương tiện trên các nền tảng mạng xã hội, sự mỉa mai ngày càng trở nên tinh vi hơn khi không chỉ được thể hiện qua văn bản mà còn qua hình ảnh, video, và các biểu tượng cảm xúc (emojis). Sự kết hợp giữa các phương tiện này tạo ra những mâu thuẫn giữa các đặc trưng văn bản và đặc trưng đa phương tiện khác, điều này làm cho việc nhận diện sự mỉa mai trở nên phức tạp hơn. Trong khi văn bản có thể mang ý nghĩa mỉa mai rõ rệt, hình ảnh hoặc biểu tượng kèm theo lại có thể cung cấp ngữ cảnh trái ngược, từ đó tạo ra một thông điệp mỉa mai mang tính ẩn dụ. Để giải quyết vấn đề này, các phương pháp phát hiện sự mỉa mai đa phương tiện đã được đề xuất, nhằm kết hợp các đặc trưng của văn bản, hình ảnh và biểu tượng để phân tích và nhận diện sự mỉa mai một cách toàn diện hơn.

Rossano và các đồng đội (2016) đã đề xuất hai phương pháp nhằm phát hiện sự mỉa mai trên các nền tảng mạng xã hội đa phương tiện. Phương pháp đầu tiên khai thác các đặc trưng ngữ nghĩa của hình ảnh, được huấn luyện trên một tập dữ liệu bên ngoài, và sau đó sử dụng hai khung tính toán khác nhau để phát hiện châm biếm đa phương thức giữa ảnh và văn bản. Phương pháp thứ hai điều chỉnh một mạng nơ-ron thị giác, với các tham số ban đầu được huấn luyện trên ImageNet, để xử lý các bài đăng đa phương tiện chứa mỉa mai, tận dụng sự tích hợp giữa văn bản và hình ảnh nhằm nâng cao hiệu quả phân tích. Kết quả của nghiên cứu này cho thấy sự tích hợp giữa các phương tiện có tác động tích cực đến hiệu suất nhận diện sự mỉa mai.

Sau đó, mô hình Hierarchical Fusion Model do Cai (2019) đề xuất đã giải quyết các hạn chế của hai phương pháp trên. Thay vì xử lý từng đặc trưng dữ liệu riêng biệt, Hierarchical Fusion Model không chỉ kết hợp các đặc trưng văn bản và hình ảnh mà còn phân tách ba loại đặc trưng riêng biệt: văn bản, hình ảnh, và thuộc tính hình ảnh. Sau đó các đặc trưng này sẽ được phân tách tái cấu trúc và hợp nhất thông qua một cơ chế hợp nhất phân cấp, mang lại sự biểu diễn toàn diện hơn cho dữ liệu đa phương thức.

Sau đó, Hongliang và đồng đội (2020) đã cải tiến bằng một mô hình dựa trên kiến trúc BERT đã được đề xuất, tập trung vào việc khai thác mâu thuẫn (incongruity) giữa các phương thức (inter-modality) và trong nội tại từng phương thức (intra-modality) – đặc tính cốt lõi của mĩa mai mà các phương pháp trước đây thường bỏ qua. Cụ thể, mô hình sử dụng cơ chế inter-modality incongruity để phát hiện mâu thuẫn giữa các phương thức khác nhau như văn bản và hình ảnh, cùng với intra-modality incongruity để xử lý mâu thuẫn trong cùng một phương thức. Kết quả thực nghiệm cho thấy mô hình đạt hiệu suất vượt trội, khẳng định hiệu quả của việc tập trung vào đặc tính mâu thuẫn trong nhận diện mĩa mai đa phương thức.

6 Conclusion

6.1 Kết quả đạt được

Các thí nghiệm mang lại những phát hiện quan trọng:

- **Hiệu suất của ViSoBERT:** ViSoBERT đạt kết quả tốt nhất trong các mô hình đơn ngôn ngữ, chứng minh sự phù hợp vượt trội với dữ liệu truyền thông xã hội tiếng Việt, đặc biệt khi dữ liệu chứa nhiều yếu tố phi chính thức như teencode, emoji và các phong cách viết đặc trưng.
- **Hiệu suất của XLM-RoBERTa:** Trong nhóm mô hình đa ngôn ngữ, XLM-RoBERTa thể hiện hiệu suất tốt nhờ khả năng biểu diễn ngữ nghĩa mạnh mẽ trong các ngôn ngữ khác nhau.
- **Sự kết hợp giữa ViSoBERT và XLM-RoBERTa:** Kết quả thử nghiệm cho thấy rằng việc kết hợp đặc trưng từ ViSoBERT

(mô hình ngôn ngữ tiếng Việt) và XLM-RoBERTa (mô hình ngôn ngữ đa ngôn ngữ) mang lại hiệu quả tốt hơn so với việc chỉ sử dụng riêng lẻ từng mô hình, đặc biệt trên tập dữ liệu chứa các ngôn ngữ khác ngoài tiếng Việt.

6.2 Hạn chế

- Dữ liệu mất cân bằng giữa các nhãn tiếp tục là thách thức lớn, làm giảm hiệu suất trên một số nhãn ít xuất hiện.
- Một số ngữ cảnh đa ngôn ngữ hoặc chứa yếu tố văn hóa đặc thù chưa được mô hình xử lý tốt.
- Yêu cầu tài nguyên lớn và thời gian huấn luyện dài, đặc biệt khi sử dụng các mô hình kết hợp.

6.3 Đóng góp chính

- Cung cấp đánh giá toàn diện về hiệu quả của các mô hình đơn ngôn ngữ và đa ngôn ngữ trên dữ liệu tiếng Việt, cùng với các chiến lược kết hợp đặc trưng.
- Khẳng định vai trò của ViSoBERT trong việc xử lý dữ liệu phi chính thức và ngôn ngữ mạng xã hội tiếng Việt.
- Đề xuất cách kết hợp mô hình ngôn ngữ đơn ngữ và đa ngữ để tận dụng điểm mạnh của từng mô hình trong các bài toán phức tạp.

6.4 Hướng phát triển trong tương lai

- Thu thập thêm dữ liệu với ngữ cảnh đa dạng hơn, bao gồm cả dữ liệu ngôn ngữ đa ngữ và yếu tố văn hóa.
- Cải tiến phương pháp xử lý các yếu tố phi chính thức như emoji, teencode, và các biểu tượng đặc trưng khác.
- Nâng cấp phương pháp kết hợp mô hình để tăng tính hiệu quả và khả năng tổng quát hóa.
- Tối ưu hóa quy trình huấn luyện để giảm thời gian và chi phí tài nguyên.

References

- [1] Donald E. Knuth (1986) , Addison-Wesley Professional.
- [2] Leslie Lamport (1994) , Addison Wesley, Massachusetts, 2nd ed.
- [3] David Bamman and Noah A. Smith. (2015). "Contextualized Sarcasm Detection on Twitter." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Carnegie Mellon University.
- [4] Meishan Zhang, Yue Zhang, and Guohong Fu. (2016). "Tweet Sarcasm Detection Using Deep Neural Network." *School of Computer Science and Technology, Heilongjiang University, China, Singapore University of Technology and Design*.
- [5] Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. (2017). "Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification Using Convolutional Neural Network." *IBM Research, India, Indian Institute of Technology Bombay, India*.
- [6] Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. (2016). "Detecting Sarcasm in Multimodal Social Platforms." *Proceedings of the 24th ACM International Conference on Multimedia (MM '16)*, Pages 1136-1145. ACM. DOI: <https://doi.org/10.1145/2964284.2964321>.
- [7] Yitao Cai, Huiyu Cai, and Xiaojun Wan. (2019). "Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model." *Institute of Computer Science and Technology, Peking University, The MOE Key Laboratory of Computational Linguistics, Peking University, Center for Data Science, Peking University*.
- [8] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. (2020). "Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection." *Proceedings of the [Tân hội nghị hoặc tạp chí]*, Chinese Academy of Sciences Beijing, China, Institute of Information Engineering Beijing, China.
- [9] Link source code: https://drive.google.com/file/d/1JRL1EqfUKkF0HpAG6EJ_hxHcx1Jw0b2Y/view?usp=drive_link