Phân loại cảm xúc văn bản Tiếng Việt trên mạng xã hội

Nguyen Duy Thinh - Le Cuong Thinh - Doan Cong Tai
22521414 - 22521409 - 22521271

Abstract

Trong thử nghiệm này, chúng tôi tập trung vào việc nghiên cứu mức độ ảnh hưởng của việc tiền xử lí dữ liệu trong bài toán phân loại cảm xúc văn bản mạng xã hội Tiếng Việt trên tập dữ liệu UIT-VSMEC. Sau khi đánh giá các kỹ thuật tiền xử lí qua mô hình PhoBERT, độ chính xác của bài toán được cải thiện rõ rệt. Đặc biệt, chúng tôi đã đạt được độ chính xác cao nhất là 66,02% (F1-Score) sau khi kết hợp các kỹ thuật đạt hiểu quả cao với nhau.

1 Giới thiệu

Trong thời đại bùng nổ thông tin và mạng xã hội, các nền tảng trực tuyến ngày càng trở thành nơi mà người dùng chia sẻ cảm xúc, ý kiến và quan điểm cá nhân. Việc hiểu và phân loại cảm xúc từ các bình luận trên mạng xã hội đóng vai trò quan trọng trong việc hỗ trợ các doanh nghiệp, tổ chức và cá nhân phân tích ý kiến khách hàng, dự đoán xu hướng, hoặc giải quyết các vấn đề liên quan đến truyền thông.

Bài toán phân loại cảm xúc văn bản tập trung vào việc xác định trạng thái cảm xúc với các cảm xúc cơ bản dựa trên nội dung văn bản bình luận. Đây là một bài toán thuộc lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), được sự quan tâm của nhiều nhà nghiên cứu hiện nay và đòi hỏi sự kết hợp giữa xử lý ngôn ngữ và các mô hình học sâu tiên tiến.

Các bình luận trên mạng xã hội thường chứa đựng những thách thức lớn đối với việc xử lý ngôn ngữ tự nhiên, đặc biệt là khi làm việc với các ngôn ngữ có cấu trúc phức tạp như tiếng Việt.

Những văn bản ngắn trên mạng xã hội (Microtext) thường không tuân theo các quy tắc ngữ pháp chặt chẽ. Chúng có thể bao gồm từ viết tắt, từ lóng, lỗi chính tả, biểu tượng cảm xúc (emoji), và các yếu tố đa phương thức khác.

Người dùng mạng xã hội có thể thể hiện cảm xúc một cách gián tiếp, ẩn ý hoặc thông qua sự kết hợp của nhiều phương thức biểu đạt, làm tăng thêm độ phức tạp của bài toán phân loại.

Trong nghiên cứu này, chúng tôi tập trung nghiên cứu tác động của các kỹ thuật tiền xử lý lên hiệu suất của hệ thống phân loại cảm xúc văn bản, được xây dựng dựa trên dữ liệu mạng xã hội Việt Nam. Cụ thể, chúng tôi nghiên cứu và lựa chọn một tập hợp các kỹ thuật tiền xử lý có khả năng làm sạch và chuẩn hóa microtext tiếng Việt. Các kỹ thuật này bao gồm loại bỏ ký tự đặc biệt, chuẩn hóa từ viết tắt, sai chính tả, xử lý biểu tượng cảm xúc, và chuẩn hóa dấu câu.

Chúng tôi xây dựng nhiều phiên bản dữ liệu đầu vào bằng cách áp dụng các tổ hợp khác nhau của các kỹ thuật tiền xử lý bằng cách sử dụng mô hình PhoBERT trên tập dữ liệu chuẩn UIT-VSMEC.

Chúng tôi sử dụng điểm F1 làm thước đo chính để đánh giá hiệu suất của các mô hình. Thông qua việc so sánh điểm F1 giữa các bộ phân loại, chúng tôi xác định được sự kết hợp tối ưu của các kỹ thuật tiền xử lý, từ đó rút ra các khuyến nghị thực tiễn cho việc xử lý dữ liệu mạng xã hội tiếng Việt.

2 Bộ dữ liệu

Để thực hiện bài toán này, chúng tôi lựa chọn bộ dữ liệu UIT-VSMEC (Vietnamese Students' Mental Health Corpus)[2] được xây dựng bởi Trường Đại học Công nghệ Thông tin (UIT). Trong bộ dữ liệu này, kết quả không được tạo ra dưới dạng phân cực: tích cực hay tiêu cực hoặc ở dạng xếp hạng (từ 1 đến 5) mà ở mức độ phân tích tình cảm chi tiết hơn, trong đó kết quả được mô tả bằng nhiều biểu thức như buồn bã, thích thú, giận dữ, ghê tởm, sợ hãi và ngạc nhiên.

Bộ dữ liệu này tập trung vào phân loại cảm xúc trong các bài viết ngắn của học sinh, sinh viên, với mục đích hỗ trợ nghiên cứu và phát triển các ứng dụng chăm sóc sức khỏe tinh thần.

2.1 Tổng quan

Bộ dữ liệu bao gồm 6.927 câu, mỗi câu là một đơn vị độc lập, mang một cảm xúc rõ ràng, được lựa chọn và kiểm định kỹ lưỡng để đảm bảo chất lượng.

Các nhãn cảm xúc (Emotion Labels): Bộ dữ liệu được gắn nhãn với 6 loại cảm xúc chính đại diện cho các trạng thái cảm xúc thường gặp, và nhãn 'Other' nếu bình luận không bao gồm bất kỳ cảm xúc nào hoặc nếu bình luận chứa cảm xúc khác với sáu nhãn cảm xúc trên.

No.	Vietnamese sentences	Emotion
1	Ảnh đẹp quá!	Enjoyment
2	Tao khóc..huhu..Tao rớt rồi	Sadness
3	Khuôn mặt của tên đó vẫn còn ám ảnh tao.	Fear
4	Cái gì cơ? Bắt bỏ tù lũ khốn đó hết!	Anger
5	Thật không thể tin nổi, tại sao lại nhanh tới thế??	Surprise
6	Những điều nó nói làm tao buồn nôn	Disgust
7	Hàng cổ rồi anh ơi	Other

Bảng 1: Ví dụ một vài câu gắn được nhãn cảm xúc.

2.2 Cấu trúc bộ dữ liệu

Bộ dữ liệu UIT-VSMEC được chia thành 3 phần:train , valid và test.Việc phân chia này nhằm đáp ứng cho việc huấn luyện ,điều chỉnh các thông số sao cho phù hợp và kiểm thử trên bộ test.

Bộ dữ liệu được phát hành dưới dạng các file xlsx, bao gồm các cột thông tin:

- Câu văn: Nội dung của câu hoặc đoạn văn ngắn.
- Nhãn cảm xúc: Một trong sáu loại cảm xúc (vui vẻ, buồn bã, tức giận, chán ghét, sợ hãi, ngạc nhiên).

Train set

	Unnamed: 0	Emotion	Sentence
0	188	Other	cho mình xin bài nhạc tên là gì với ạ
1	166	Disgust	cho đáng đời con quỳ . về nhà lôi con nhà mày ...
2	1345	Disgust	lo học đi . yêu đương lol gì hay lại thích học...
3	316	Enjoyment	ước gì sau này về già vẫn có thể như cụ này :))
4	1225	Enjoyment	mỗi lần có video của con là cứ coi đi coi lại ...

Hình 1: Tập train

Valid set

	Unnamed: 0	Emotion	Sentence
0	941	Other	tính tao tao biết , chẳng có chuyện gì có thể ...
1	142	Enjoyment	lại là lào cai , tự hào quê mình quá :))
2	1164	Sadness	bị từ chối rồi
3	182	Enjoyment	tam đảo trời đẹp các mem à
4	868	Other	đọc bình luận của thằng đó không thiếu chữ nào 🤬🤬

Hình 2: Tập valid

Test set

	Unnamed: 0	Emotion	Sentence
0	713	Sadness	người ta có bạn bè nhìn vui thật
1	1827	Surprise	cho nghỉ việc mới đúng sao gọi là kỷ luật
2	1166	Disgust	kinh vãi 🤬
3	228	Fear	nhà thì không xa lắm nhưng chưa bao giờ đi vì ...
4	1942	Anger	bố không thích nộp đấy mày thích ý kiến không

Hình 3: Tập test

Sau khi xem xét qua các tập trong bộ dữ liệu, chúng tôi nhận thấy rằng các bình luận thu thập từ mạng xã hội có sự phân bổ không đồng đều giữa các nhãn cảm xúc. Trong đó, nhãn ”vui vẻ” (Enjoyment) chiếm tỷ lệ cao nhất với 1.965 câu tương ứng 28.36% tổng số dữ liệu. Ngược lại, nhãn ”ngạc nhiên” (Surprise) có số lượng thấp nhất, chỉ 309 câu, chiếm 4.46%.

Tổng cộng bộ dữ liệu có 6.927 câu, phản ánh sự đa dạng trong biểu hiện cảm xúc nhưng vẫn còn mất cân bằng giữa các nhãn, đặc biệt với sự chênh lệch lớn giữa nhãn ”vui vẻ” (Enjoyment) và ”ngạc nhiên” (Surprise).

Emotion	Sentences	Percentage (%)
Enjoyment	1,965	28.36
Disgust	1,338	19.31
Sadness	1,149	16.59
Anger	480	6.92
Fear	395	5.70
Surprise	309	4.46
Other	1,291	18.66
Total	6,927	100.00

Bảng 2: Thống kê số lượng nhãn cảm xúc của bộ dữ liệu

Sự phân bổ này là một thách thức cần cân nhắc khi huấn luyện các mô hình nhận diện cảm xúc, đặc biệt trong việc đảm bảo khả năng học tốt đối với các nhãn có tần suất thấp như ”ngạc nhiên” (Surprise).

3 Tiền xử lí dữ liệu

Sau khi khảo sát qua tập dữ liệu UIT-VSMEC, chúng tôi nhận thấy bộ dữ liệu này chưa được tác giả thực hiện đầy đủ các công đoạn làm sạch, dữ liệu vẫn tồn tại một số yếu tố gây nhiễu đặc trưng của văn bản trên mạng xã hội, như lỗi chính tả, từ viết tắt, từ lóng, biểu tượng cảm xúc, và dấu câu không cần thiết. Điều này có thể ảnh hưởng đến hiệu suất của các hệ thống phân loại cảm xúc khi xử lý văn bản tiếng Việt.

Kỹ thuật xử lí
{1} Xử lí những kí tự bị lặp
{2} Chuẩn hóa thành chữ thường và xóa khoảng trắng dư thừa
{3} Tìm kiếm, sửa lỗi chính tả và chữ viết tắt
{4} Xử lí các emoji và emoticon
{5} Loại bỏ những stop words và những từ lặp

Bảng 3: Các kĩ thuật tiền xử lí sử dụng

Để khắc phục những hạn chế này và cải thiện chất lượng dữ liệu đầu vào, chúng tôi đề xuất một loạt các kỹ thuật tiền xử lý chuyên biệt nhằm làm sạch và chuẩn hóa dữ liệu mạng xã hội Việt Nam (microtext tiếng Việt)[4]. Những kỹ thuật này được thiết kế dựa trên việc phân tích đặc điểm ngôn ngữ và cách biểu đạt phổ biến trên các nền tảng mạng xã hội, từ đó tối ưu hóa hiệu quả của các mô hình

phân loại cảm xúc. Chi tiết của năm kỹ thuật tiền xử lý này được trình bày trong Bảng 3.

Kỹ thuật đầu tiên mà chúng tôi sử dụng là xử lí các kí tự bị lặp trong câu bình luận. Dữ liệu trong UIT-VSMEC chứa nhiều kí tự bị lặp đi lặp lại. Đây là một trường hợp không thể thiếu trong các văn bản bình luận trên mạng xã hội. Với kĩ thuật 1, chúng tôi sẽ thay thế những từ chứa những kí tự lặp thành từ đúng trong tiếng Việt, ví dụ: các từ 'hahaaaa', 'luônnn' sẽ được chuyển đổi thành 'haha', 'luôn' hay ':)))))', ':(((' được chuyển thành ':)' và ':('. Bảng 4 sẽ cho thấy một vài ví dụ về việc áp dụng kỹ thuật này.

Từ gốc	Từ được chuẩn hóa
luônnnn	luôn
hahaaaa	haha
hihiiii	hihi
:)))))	:)
:>>>	:>

Bảng 4: Một vài trường hợp áp dụng kỹ thuật 1

Trên các nền tảng mạng xã hội, người dùng thường xuyên sử dụng chữ hoa để nhấn mạnh cảm xúc (ví dụ: "QUÁ TUYỆT VỜI"). Bên cạnh đó, có nhiều khoảng trắng dư thừa xuất hiện trong cùng một câu, điều này có thể làm ảnh hưởng trong quá trình tách từ hoặc mã hóa từ khi sử dụng các mô hình NLP. Vì thế chúng tôi kết hợp hai điều này tương đương với kỹ thuật 2 góp phần giảm sự thiên lệch mà các kiểu nhấn mạnh này có thể tạo ra, đảm bảo tính khách quan cho mô hình[5]. Các câu sau khi được áp dụng kỹ thuật năm sẽ có cấu trúc như trong bảng 5.

Câu gốc	Câu được chuẩn hóa
ĐẸP KHÔNG ??????	đẹp không ??????
per MISS YOU ..	per miss you..
chí nh x ác lu ôn a k h ậu	chính xác luôn ak hậu
n h ỏ nà y... bị kh ùng t a	nhỏ này... bị khùng ta

Bảng 5: Một vài trường hợp áp dụng kỹ thuật 2

UIT-VSMEC chứa đựng nhiều từ viết sai chính tả, từ lóng và từ viết tắt – những đặc trưng thường thấy trong văn bản siêu ngắn, đặc biệt là trên mạng xã hội. Đây là kết quả của xu hướng tối giản và sáng tạo trong giao tiếp, nhưng lại gây ra không ít khó khăn cho quá trình xử lý ngôn ngữ tự nhiên. Chẳng hạn, "bjo" là cách viết rút gọn của "bây giờ" (now), trong khi "ju jin" được dùng thay cho "giữ gìn" (conserve).

Để khắc phục những vấn đề này, chúng tôi đã xây dựng một từ điển chuyên dụng, bao gồm 127 mục từ, nhằm chuẩn hóa các từ viết tắt và từ viết sai chính tả. Từ điển này cung cấp bản dịch chi tiết cho từng mục, giúp chuyển đổi các dạng không chuẩn hóa thành dạng chính xác và phù hợp. Việc áp dụng từ điển cho phép chúng tôi tự động thay thế các từ viết sai chính tả và từ viết tắt trong dữ liệu, từ đó nâng cao chất lượng và tính đồng nhất của văn bản đầu vào.

Emotion	Từ tiếng Việt
ko, kô, khôg, khg	không
biet, bit, bij, pít	biết
ngta, nta	người ta
kq, kqua	kết
cũnh, cũg, cungc, cungz	cũng

Bảng 6: Một vài trường hợp áp dụng kỹ thuật 3

Một số ví dụ minh họa về cách áp dụng kỹ thuật này được trình bày trong Bảng 6. Sau khi hoàn tất quá trình chuẩn hóa, chúng tôi tiến hành phân tích thống kê trên tập dữ liệu UIT-VSMEC để đánh giá hiệu quả của kỹ thuật này.

Bộ dữ liệu UIT-VSMEC chứa đựng nhiều biểu tượng cảm xúc và emoji, hai yếu tố phổ biến trên các nền tảng mạng xã hội. Biểu tượng cảm xúc là cách biểu đạt cảm xúc hoặc trạng thái thông qua sự sắp xếp các ký tự như dấu chấm câu, chữ cái và số, nhằm tạo hình ảnh đơn giản mô phỏng biểu cảm khuôn mặt. Trong khi đó, emoji là các biểu tượng đồ họa được thiết kế để trực quan hơn, thay vì sử dụng ký tự để xấp xỉ. Mặc dù biểu tượng cảm xúc thường được xem như chuỗi ký tự đơn lẻ và emoji được coi là các thành phần hình ảnh độc lập, cả hai đều đóng vai trò quan trọng trong việc biểu đạt cảm xúc trên mạng xã hội. Tuy nhiên, trong quá trình tiền xử lý dữ liệ7 cho thấy một số ví dụ về dạng từ của một số biểu tượng cảm xúc, emoticon.

Các dạng từ của biểu tượng cảm xúc và emoticon là tiếng Anh, trong khi ngôn ngữ chính của tập dữ liệu là tiếng Việt. Sự khác biệt về ngôn ngữ này dẫn đến việc các biểu tượng cảm xúc và emoticon có thể mang ý nghĩa không đồng nhất hoặc bị hiểu sai so với các từ cảm xúc trong tiếng Việt. Để giải quyết vấn đề này, chúng tôi áp dụng kỹ thuật 5 nhằm dịch các dạng biểu tượng cảm xúc và emoticon từ tiếng Anh sang tiếng Việt, đảm bảo tính nhất quán về ngữ nghĩa trong dữ liệu.

Quy trình dịch được thực hiện thông qua việc xây dựng một từ điển quy tắc dịch dành riêng cho các biểu tượng cảm xúc và emoticon. Từ điển này bao gồm tổng cộng 82 quy tắc dịch, mỗi quy tắc tương ứng với một dạng biểu tượng cụ thể. Các quy tắc này được thiết kế để chuyển đổi chính xác các biểu tượng cảm xúc và emoticon thành các từ

Emojis	Emotion	Dạng từ tiếng anh
😠		:angry_face:
😑	-_-	:expressionless_face:
❤️	<3	:red_heart:
😭	:'(	:loudly_crying_face:

Bảng 7: Một vài trường hợp áp dụng kỹ thuật 4

hoặc cụm từ tiếng Việt mang nghĩa tương đương. Ví dụ, biểu tượng "disappointed_face" được dịch thành từ tiếng Việt "thất vọng," giúp duy trì cảm xúc nguyên bản mà biểu tượng này muốn truyền tải.

Emojis	Dạng từ
:angry_face:	tức giận
:expressionless_face:t	không cảm xúc
:red_heart:	yêu thương
:loudly_crying_face:	khóc lớn

Bảng 8: Một vài trường hợp áp dụng kỹ thuật 4

Để minh họa rõ hơn về cách áp dụng kỹ thuật này, chúng tôi trình bày một số ví dụ cụ thể trong Bảng 8. Việc chuyển đổi này không chỉ cải thiện khả năng hiểu ngữ nghĩa của dữ liệu mà còn tăng độ chính xác của mô hình trong việc nhận diện và phân loại cảm xúc, đặc biệt khi làm việc với dữ liệu mạng xã hội có nhiều yếu tố giao thoa giữa ngôn ngữ.

Đối với kỹ thuật 5, chúng tôi tìm kiếm một danh sách các stopwords và những từ lặp để kiểm tra. Stopwords là những từ xuất hiện thường xuyên trong câu nhưng không mang ý nghĩa cụ thể như: 'là', 'nhé', 'vậy'. Giữ lại những từ này có thể làm giảm hiệu quả của mô hình vì chúng làm tăng khối lượng dữ liệu không liên quan, khiến mô hình khó tập trung vào các từ thực sự mang ý nghĩa cảm xúc. Chúng tôi tiến hành thu thập những từ này bằng cách sau: sắp xết toàn bộ từ vựng theo tần suất xuất hiện; lọc ra các từ có tần suất xuất hiện nhiều hơn 15 lần; không phải là danh từ/cụm danh từ, động từ, tính từ hoặc trạng từ[1]. Danh sách từ sau khi thu thập được bao gồm hơn 45 từ.

Tuy nhiên, việc lọc một số từ dừng và các từ cụ thể trong tập xác thực và kiểm tra sẽ bỏ qua tác động của những từ này. Nếu chúng tôi cũng xóa những từ này khỏi tập huấn luyện, thì bộ phân loại của chúng tôi sẽ không hiểu ngữ cảnh của bình luận. Chúng tôi đánh giá tác động của việc xóa những từ này đối với tập xác thực bằng cách sử dụng bộ phân loại cảm xúc được xây dựng trên tập huấn luyện đã được làm sạch.

4 Phương pháp đánh giá

Để đánh giá mức độ hiệu quả của các kỹ thuật tiền xử lý dữ liệu trên bộ dữ liệu UIT-VSMEC, chúng tôi lựa chọn sử dụng PhoBERT[3], một mô hình ngôn ngữ tiền huấn luyện được thiết kế riêng cho tiếng Việt và được công bố vào năm 2020 bởi VinAI Research. PhoBERT được xây dựng dựa trên kiến trúc Transformer và được huấn luyện trên một Corpus tiếng Việt lớn (20GB văn bản), giúp mô hình nắm bắt tốt các đặc điểm ngữ pháp, từ vựng, và ngữ nghĩa trong tiếng Việt – một ngôn ngữ có cấu trúc đa dạng và phức tạp.

PhoBERT áp dụng cơ chế Masked Language Modeling (MLM), tương tự như BERT, giúp mô hình hiểu rõ ngữ cảnh của từ hoặc cụm từ trong câu. Khả năng này đặc biệt quan trọng đối với bài toán phân loại cảm xúc, nơi ngữ cảnh đóng vai trò quan trọng trong việc xác định chính xác cảm xúc của câu văn. Ví dụ, cùng một từ ngữ có thể mang ý nghĩa tích cực hoặc tiêu cực, tùy thuộc vào ngữ cảnh cụ thể.

- PhoBERT-base: Đây là phiên bản nhỏ với kích thước bao gồm 135 triệu tham số. Mô hình có 12 lớp Transformer, 12 Attention heads và hidden size là 768.
- PhoBERT-large: Đây là phiên bản lớn hơn, có cấu trúc phức tạp hơn. Mô hình bao gồm 370 triệu tham số, 24 lớp Transformer, 16 Attention heads và hidden size là 1024.

Trong nghiên cứu này, chúng tôi lựa chọn sử dụng PhoBERT-base[6] vì kích thước phù hợp với các bài toán nghiên cứu không quá lớn, yêu cầu tốc độ xử lý nhanh và ít phụ thuộc vào tài nguyên tính toán cao cấp. Việc sử dụng PhoBERT-base đảm bảo tính hiệu quả cả về mặt thời gian và tài nguyên, đồng thời vẫn đạt được kết quả đáng tin cậy trên tập dữ liệu UIT-VSMEC.

5 Kết quả nghiên cứu

5.1 Phân phối từ

Sau khi áp dụng dụng các kỹ thuật xử lí, chúng tôi tiến hành phân tích sự thay đổi của số lượng từ câu mỗi nhãn dữ liệu để nghiên cứu sự tác động của cá kỹ thuật này lên bộ dữ liệu. Kết quả phân tích được trình bày qua bảng 9.

Cảm xúc	Số từ gốc	KT 3	KT 4	KT 5
Anger	7169	7224	7359	6274
Disgust	18326	18428	18827	16013
Enjoyment	19550	19699	21123	17051
Fear	4733	4762	4936	4087
Surprise	2247	2259	2373	1829
Sadness	13238	13308	13628	11536
Other	12455	12479	12897	10641

Bảng 9: Phân phối từ sau khi áp dụng các kỹ thuật

Với kỹ thuật 1 (xử lý các ký tự lặp) và kỹ thuật 2 (chuẩn hóa thành chữ thường và xóa khoảng trắng dư thừa), số lượng từ không thay đổi đáng kể. Điều này cho thấy, các kỹ thuật này chủ yếu tập trung vào việc chuẩn hóa và làm sạch văn bản mà không làm ảnh hưởng đến cấu trúc câu hay nội dung chính. Chúng giúp đảm bảo dữ liệu có định dạng thống nhất, từ đó giảm bớt lỗi phát sinh trong quá trình xử lý tiếp theo.

Ngược lại, với kỹ thuật 3, 4 và 5, có sự thay đổi rõ rệt về số lượng từ trong câu, thể hiện rõ tác động của chúng đến cấu trúc dữ liệu.

Với Kỹ thuật 3 (sửa lỗi chính tả và viết tắt) làm tăng nhẹ số lượng từ trong tập dữ liệu. Nguyên

nhân là vì khi sửa lỗi chính tả hoặc mở rộng các từ viết tắt, một số từ mới được thêm vào để thay thế các lỗi ban đầu. Mặc dù số lượng từ tăng lên, nhưng sự cải thiện về chất lượng dữ liệu lại đáng kể. Những từ được sửa chính tả và mở rộng viết tắt giúp tăng tính chính xác trong nhận diện ý nghĩa của câu, làm giảm sự mơ hồ mà các lỗi ban đầu có thể gây ra. Tuy nhiên, việc thêm các từ mới cũng có thể làm giảm mức độ tập trung của mô hình với các từ quan trọng khác, do sự gia tăng số lượng từ trong câu.

Kỹ thuật 4 (chuyển đổi emoji và emoticon) cũng góp phần làm tăng số lượng từ. Khi các biểu tượng cảm xúc được chuyển thành từ hoặc cụm từ mô tả, tập dữ liệu trở nên phong phú hơn về mặt biểu đạt. Các emoji và emoticon thường mang ý nghĩa cảm xúc rõ rệt, việc chuyển đổi chúng thành dạng văn bản giúp mô hình có thêm dữ liệu để phân tích cảm xúc chính xác hơn. Tuy nhiên, điều này cũng đồng nghĩa với việc làm phức tạp thêm tập dữ liệu, đòi hỏi mô hình phải xử lý lượng thông tin lớn hơn và duy trì khả năng nhận diện các yếu tố quan trọng.

Kỹ thuật 5 (loại bỏ stopwords) có tác động mạnh mẽ nhất đến số lượng từ. Số từ trong câu giảm đáng kể sau khi loại bỏ các từ dừng (stopwords). Các từ dừng này tuy có ý nghĩa trong ngữ pháp nhưng thường không mang nhiều thông tin trong phân tích ý nghĩa câu. Việc loại bỏ chúng giúp mô hình tập trung nhiều hơn vào các từ quan trọng còn lại, từ đó tăng khả năng nhận diện và phân tích chính xác các yếu tố chính trong câu. Tuy nhiên, việc này cũng đặt ra một thách thức khi các từ bị loại bỏ có thể ảnh hưởng đến ngữ cảnh tổng thể của câu, đặc biệt là đối với tiếng Việt – một ngôn ngữ phụ thuộc nhiều vào ngữ cảnh.

5.2 Kết quả thực nghiệm

Chúng tôi chọn sử dụng F1-score và Accuracy làm thước đo chính để đánh giá hiệu quả của các kỹ thuật tiền xử lý được áp dụng trên bộ dữ liệu UIT-VSMEC khi sử dụng mô hình PhoBERT. Đây là hai chỉ số quan trọng, phản ánh khả năng cân bằng giữa độ chính xác và độ bao phủ của mô hình, đặc biệt hữu ích trong bài toán phân loại cảm xúc. Kết quả chi tiết của từng trường hợp được trình bày trong Bảng 10.

Khi áp dụng PhoBERT với dữ liệu thô chưa được xử lý, chỉ số Accuracy và F1-score đạt khoảng 60%. Đây là mức nền (baseline) để so sánh hiệu quả của các kỹ thuật xử lý dữ liệu. Với hai kỹ thuật đầu tiên là 1, 2, việc áp dụng hai kỹ thuật này đã giúp tăng đáng kể độ chính xác (+2.12%) và F1-score (+0.89%). Điều này cho thấy rằng hai kỹ thuật này mang lại lợi ích rõ rệt trong việc cải thiện hiệu suất. Nhưng khi thêm kỹ thuật xử lý 5 có vẻ như làm giảm nhẹ hiệu suất so với chỉ dùng kỹ thuật 1 và 2. Điều này gợi ý rằng kỹ thuật 5 có thể không phù hợp, và việc loại bỏ các stopwords làm mất đi một số thông tin hữu ích. Trái lại khi thêm kỹ thuật 3 vào kết hợp với kỹ thuật 1, 2, hiệu suất được cải thiện rõ rệt (+3.61% Accuracy và +3.72% F1-score so với 1, 2). Điều này cho thấy rằng kỹ thuật 3 xử lý các từ viết tắt và sai chính tả đóng vai trò quan trọng trong việc nâng cao hiệu suất. Thay vì sử dụng kỹ thuật 3, kỹ thuật 4 được thêm vào và cải thiện hiệu suất so với baseline, nhưng hiệu quả không bằng kỹ thuật 3. Điều này cho thấy kỹ thuật 3 mạnh hơn kỹ thuật 4, do việc dịch các emoji và emoticon không phù hợp ở một số ngữ cảnh nhất định. Sự kết hợp của các kỹ thuật 1, 2, 3 và 4 mang lại hiệu suất cao nhất, tăng 6.02% Accuracy và 5.13% F1-score so với PhoBERT gốc. Đây là sự kết hợp tối ưu trong thử nghiệm. Khi kết hợp cả 5 kỹ thuật lại với nhau, một lần nữa cho thấy việc thêm kỹ thuật 5 vào đã làm giảm hiệu suất đáng kể, quay trở lại mức gần giống với khi chỉ sử dụng 1 và 2. Điều này xác nhận rằng kỹ thuật 5 không phù hợp trong bài toán này trên bộ dữ liệu áp dụng.

Mô hình	Accuracy	F1-score
PhoBERT	59.93%	60.89%
PhoBERT + kĩ thuật xử lí 1, 2	59.93%	61.78%
PhoBERT + kĩ thuật xử lí 1, 2, 5	62.05%	61.25%
PhoBERT + kĩ thuật xử lí 1, 2, 3	61.62%	65.50%
PhoBERT + kĩ thuật xử lí 1, 2, 4	65.66%	64.58%
PhoBERT + kĩ thuật xử lí 1, 2, 3, 4	65.95%	66.02%
PhoBERT + kĩ thuật xử lí 1, 2, 3, 4, 5	61.90%	61.85%

Bảng 10: Kết quả đánh giá các kỹ thuật xử lí

Hầu hết các kỹ thuật mà nhóm đề xuất đều mang lại hiệu quả tích cực khi áp dụng trên bộ dữ liệu UIT-VSMEC, ngoại trừ việc loại bỏ stopwords. Việc thu thập và xác định danh sách stopwords chưa thực sự chính xác, dẫn đến việc một số từ có thể ảnh hưởng đến nghĩa của câu vẫn còn trong danh sách.

5.3 Các trường hợp dự đoán sai

Bảng 11 cho thấy một vài trường hợp mô hình dữ đoán sai nhãn. Đây là kết quả dự đoán dựa trên mô hình có độ chính xác cao nhất là PhoBERT với các kỹ thuật xử lí 1, 2, 3, 4.

Câu dự đoán	Cảm xúc	Dự đoán
người ta có bạn bè nhìn vui thật	sadness	enjoyment
cho nghỉ viêc mới đúng sao goi là kỷ luật	surprise	anger
cứu em nó với	sadness	fear
nấu mì 2 tôm là vô địch	enjoyment	other

Bảng 11: Các trường hợp dự đoán sai

Với câu đầu tiên, câu này chứa yếu tố mỉa mai hoặc buồn, kết hợp với từ ngữ tích cực (”vui thật”), khiến mô hình dễ nhầm lẫn với cảm xúc Enjoyment. Cảm xúc mỉa mai thường khó phân biệt nếu không có ngữ cảnh rõ ràng. Điều này tạo nên thử thách

cho việc phân loại chính xác cảm xúc trong các tình huống như thế này. Mô hình nhầm lẫn giữa sự ngạc nhiên và phẫn nộ do ngữ điệu phê phán trong câu thứ 2. Từ "kỷ luật" kết hợp với ngữ điệu phê phán có thể khiến mô hình nhận diện là cảm xúc tiêu cực mạnh hơn, chẳng hạn như anger. Đặc biệt, do số lượng câu mang nhãn Anger và Surprise còn hạn chế, mô hình chưa hoàn toàn nhận diện chính xác. Cụm từ "cứu em" trong câu 3 thường gợi lên cảm giác khẩn cấp và sợ hãi (fear), khiến mô hình dễ nhầm lẫn giữa nỗi buồn và cảm giác sợ hãi. Do thiếu thông tin đầy đủ trong câu, việc phân biệt rõ ràng giữa hai cảm xúc này trở nên khó khăn. Câu "nấu mì 2 tôm là vô địch" là một nhận xét vui nhộn nhưng không rõ ràng về cảm xúc (cảm giác Enjoyment không mạnh mẽ). Mô hình có thể thiếu dữ liệu về các câu mang tính giải trí nhẹ nhàng. Câu "Other" thường xuất hiện khi mô hình không tự tin vào dự đoán chính xác.

6 Kết luận

Dữ liệu là một yếu tố then chốt ảnh hưởng trực tiếp đến hiệu suất và độ chính xác của mô hình xử lý ngôn ngữ tự nhiên (NLP), đặc biệt là đối với tiếng Việt. Bộ dữ liệu UIT-VSMEC hiện vẫn còn hạn chế về độ bao quát, điều này ảnh hưởng đến khả năng mô hình học tập và nhận diện chính xác các đặc điểm ngữ nghĩa.

Tuy nhiên, với các kỹ thuật xử lý dữ liệu mà nhóm nghiên cứu đã áp dụng, hiệu suất đã được cải thiện đáng kể. Kết quả đạt được cao nhất là 66.02% (F1-score), thể hiện sự tiến bộ rõ rệt trong việc phân loại và nhận diện thông tin từ dữ liệu. Những cải thiện này cho thấy tiềm năng lớn trong việc tối ưu hóa mô hình và nâng cao hiệu suất xử lý dữ liệu cho bài toán NLP tiếng Việt.

Bên cạnh đó, chúng tôi sẽ tiếp tục phát triển và áp dụng các phương pháp xử lý dữ liệu mới nhằm cải thiện hiệu quả của mô hình. Các phương pháp này sẽ được kiểm tra và áp dụng trên nhiều bộ dữ liệu khác nhau để đánh giá tính hiệu quả trong các ngữ cảnh và bài toán NLP đa dạng. Điều này sẽ giúp mô hình trở nên mạnh mẽ hơn và có khả năng thích nghi tốt hơn với các tình huống phức tạp và đa dạng trong thực tế.

Tài liệu

[1] Nguyen Thanh Hau. ?Phân loại văn bản tự động bằng Machine Learning như thế nào?? inPhân loại cảm xúc: 2018.

[2] Vong Anh Ho andothers. ?Emotion Recognition for Vietnamese Social Media Text? inArXiv: abs/1911.09339 (2019). URL: https://api.semanticscholar.org/CorpusID:208202333.

[3] Dat Quoc Nguyen and Anh Tuan Nguyen. ?PhoBERT: Pre-trained language models for Vietnamese? inFindings of the Association for Computational Linguistics: EMNLP 2020: byeditorTrevor Cohn, Yulan He and Yang Liu. Online: Association for Computational Linguistics, november 2020, pages 1037–1042. DOI: 10.18653/v1/2020.findings-emnlp.92. URL: https://aclanthology.org/2020.findings-emnlp.92/.

[4] Khang Phuoc-Quy Nguyen and Kiet Van Nguyen. ?Exploiting Vietnamese Social Media Characteristics for Textual Emotion Recognition in Vietnamese? in2020 International Conference on Asian Language Processing (IALP): 2020, pages 276–281. DOI: 10.1109/IALP51396.2020.9310495.

[5] Nam Nguyen andothers. ?ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing? inProceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: byeditorHouda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, december 2023, pages 5191–5207. DOI: 10.18653/v1/2023.emnlp-main.315. URL: https://aclanthology.org/2023.emnlp-main.315/.

[6] Nguyen Chien Thang. ?Thử nhận diện cảm xúc văn bản Tiếng Việt với PhoBert? inBERT Series: 2020.