

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN**



**HƯỜNG QUANG HUY - 52100893
LÊ HOÀNG KHANG - 52100898**

KHẢO SÁT VỀ CẢM XÚC CỦA AI

DỰ ÁN CÔNG NGHỆ THÔNG TIN

KHOA HỌC MÁY TÍNH

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN**



**HƯỜNG QUANG HUY - 52100893
LÊ HOÀNG KHANG - 52100898**

KHẢO SÁT VỀ CẢM XÚC CỦA AI

DỰ ÁN CÔNG NGHỆ THÔNG TIN

KHOA HỌC MÁY TÍNH

Người hướng dẫn
ThS. Nguyễn Quốc Bình

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn ThS. Nguyễn Quốc Bình đã tận tâm hướng dẫn, góp ý để hoàn thành đề tài này, dù thời gian khá hạn chế bởi vì thầy phải hướng dẫn nhiều nhóm nhưng thầy vẫn sắp xếp thời gian để họp và chỉ ra những thiếu sót cần cải thiện, cũng như giúp nhóm nêu lên ý tưởng để hoàn thành đề tài một cách chính chu

TP. Hồ Chí Minh, ngày 17 tháng 02 năm 2025

Tác giả

(Ký tên và ghi rõ họ tên)

Hường Quang Huy

Lê Hoàng Khang

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Chúng tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của ThS. Nguyễn Quốc Bình. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 17 tháng 02 năm 2025

Tác giả

(Ký tên và ghi rõ họ tên)

Hường Quang Huy

Lê Hoàng Khang

KHẢO SÁT VỀ CẢM XÚC CỦA AI

TÓM TẮT

Đề tài này chúng tôi nghiên cứu nhiều mặt (literature review) của cảm xúc AI xoay quanh các câu hỏi sau:

- Cảm xúc có tác động như thế nào đến câu trả lời/hành vi của AI?
- Các phương pháp đưa yếu tố cảm xúc vào AI.
- Liệu việc đưa thêm yếu tố cảm xúc vào AI có tăng tính hiệu quả công việc?
- Liệu việc đưa thêm yếu tố cảm xúc vào AI có giảm bớt các nguy hiểm được dự báo của AI?

Bài báo cáo này của chúng tôi gồm có các phần chính:

- Mở đầu và tổng quan về đề tài: nói về mục đích, mục tiêu thực hiện đề tài
- Cơ sở lý thuyết: nêu định nghĩa cảm xúc AI, ảnh hưởng của cảm xúc AI đến kinh tế cảm xúc, cơ hội và thách thức, các cảm xúc cơ bản và mô hình cảm xúc, ứng dụng thực tế, các phương pháp đưa yếu tố cảm xúc vào mô hình AI
- Mô hình đề xuất: Tìm hiểu và sử dụng mô hình BART, kiến trúc tổng quát
- Thực nghiệm: Sử dụng dữ liệu daily_dialog chứa các cuộc hội thoại và thêm yếu tố cảm xúc để huấn luyện, đánh giá
- Kết luận: Đưa ra kết luận về các câu hỏi được đặt ra, đánh giá được khả năng phản hồi với cảm xúc của mô hình BART, hướng phát triển trong tương lai

AN INVESTIGATION OF AI EMOTIONS

ABSTRACT

We conduct a comprehensive literature review on AI emotions, focusing on the following key questions:

- How do emotions impact AI's responses and behavior?
- What methods can be used to incorporate emotional factors into AI?
- Does adding emotional factors improve AI's effectiveness in performing tasks?
- Can incorporating emotions into AI help mitigate predicted risks associated with AI?

This report consists of the following main sections:

- **Introduction and Overview:** Discusses the purpose and objectives of the study.
- **Theoretical Background:** Defines AI emotions, examines their impact on the emotional economy, explores opportunities and challenges, introduces basic emotions and emotional models, discusses real-world applications, and reviews methods for integrating emotions into AI models.
- **Proposed Model:** Investigates and utilizes the BART model, presenting its general architecture.
- **Experiments:** Uses the DailyDialog dataset, which contains conversations, and incorporates emotional factors for training and evaluation.
- **Conclusion:** Summarizes findings related to the key research questions, assesses the emotion-aware response capability of the BART model, and suggests future research directions.

MỤC LỤC

DANH MỤC HÌNH VẼ	viii
DANH MỤC BẢNG BIỂU	ix
DANH MỤC CÁC CHỮ VIẾT TẮT.....	x
CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI.....	1
1.1 Lý do chọn đề tài.....	1
1.2 Mục tiêu thực hiện đề tài.....	1
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT.....	2
2.1 Cảm xúc AI là gì, lịch sử hình thành và phát triển của tính toán cảm xúc (Affective Computing)	2
2.1.1 Tính toán cảm xúc là gì và định nghĩa của cảm xúc theo nhiều lý thuyết khác nhau?	2
2.1.2 Lịch sử hình thành và phát triển của cảm xúc AI	3
2.1.3 Các lĩnh vực nghiên cứu của tính toán cảm xúc	5
2.2 AI cảm xúc có tác động như thế nào trong nền kinh tế cảm xúc (Emotional economies)?.....	8
2.3 Cơ hội và thách thức của AI cảm xúc	9
2.3.1 Cơ hội.....	10
2.3.2 Thách thức.....	10
2.4 Các cảm xúc cơ bản và các mô hình cảm xúc.....	11
2.4.1 Các cảm xúc cơ bản theo các nhà nghiên cứu.....	11
2.4.2 Các mô hình tính toán cảm xúc.....	14
2.5 Ứng dụng thực tế về cảm xúc AI trong các lĩnh vực	15
2.5.1 Trong giáo dục.....	15

2.5.2 Trong y tế	16
2.5.3 Dịch vụ khách hàng	17
2.5.4 Ứng dụng trong giao tiếp và phát triển kỹ năng cá nhân	17
2.5.5 Quảng cáo và tiếp thị	18
2.5.6 Ô tô tự lái	18
2.5.7 Mạng xã hội	19
2.6 Các phương pháp đưa yếu tố cảm xúc vào mô hình AI	19
CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT	20
3.1 Kiến trúc Transformer	20
3.1.1 Multi-Head Self-Attention	22
3.1.2 Masked Self-Attention	23
3.1.3 Position-wise Encoding (PE)	24
3.1.4 Residual Connection & Layer Normalization (Add & Norm)	25
3.2 Thuật toán Byte Pair Encoding (BPE) dùng để tokenize dữ liệu câu chữ trong Xử lý ngôn ngữ tự nhiên (NLP – Natural Language Processing)	27
3.2.1 Lợi ích của sử dụng BPE để tách token	27
3.2.2 Nhược điểm của kỹ thuật BPE	29
3.2.3 Biểu diễn token theo phương pháp BPE	30
3.3 Mô hình BART	32
3.3.1 Kiến trúc tổng quát	32
3.3.2 Pre-training BART	33
3.3.3 Fine-tuning BART	34
3.3.4 Ưu và nhược điểm	36

CHƯƠNG 4. THỰC NGHIỆM	37
4.1 Dữ liệu thực nghiệm.....	37
4.2 Cài đặt thực nghiệm	40
4.3 Kết quả thực nghiệm	41
CHƯƠNG 5. KẾT LUẬN.....	44
5.1 Kết luận	44
5.2 Hướng phát triển	45
TÀI LIỆU THAM KHẢO	46

DANH MỤC HÌNH VẼ

Hình 2-1: Hình ảnh trực quan của các nghiên cứu về các lĩnh vực của tính toán cảm xúc từ Viện Công nghệ Massachusetts (MIT)	5
Hình 2-2: Mô hình bánh xe cảm xúc của Plutchik.....	13
Hình 3-1: Hình ảnh tổng quát Transformer	21
Hình 3-2: Scaled Dot-Product Attention.....	22
Hình 3-3: Multi-Head Attention gồm nhiều tầng Scaled Dot-Product Attention chạy song song với nhau.....	22
Hình 3-4: Minh hoạ Masked Self-Attention	24
Hình 3-5: Minh hoạ phương pháp Sinusoidal Position Encoding kết hợp với Word Embedding	25
Hình 3-6: Minh hoạ Identity Residual Connection	26
Hình 3-7: Minh hoạ thuật toán BPE cho 3 từ “picked”, “pickled”, “pickles”	30
Hình 3-8: Minh hoạ mô hình BART	32
Hình 3-9: Phần Encoder, dựa vào kiến trúc của BERT	33
Hình 3-10: Các kỹ thuật làm nhiễu dữ liệu đầu vào (phần màu đỏ)	34
Hình 3-11: Token Classification	35
Hình 3-12: Machine Translation	36
Hình 4-1: Biểu đồ phân bố các nhãn cảm xúc chính của “answer” trong dữ liệu train	38
Hình 4-2: Kết quả dự đoán cảm xúc của mô hình distil roberta cho câu “I love this!”	39
Hình 4-3: Đánh giá các losses sau khi huấn luyện với 5 epoch.....	41

Hình 4-4: So sánh với câu “You got an reward!”, gắn các nhấn cảm xúc khác nhau	42
Hình 4-5: So sánh với câu “I don’t know the reason.”	42
Hình 4-6: So sánh với câu “What do you think?”	43

DANH MỤC BẢNG BIỂU

Bảng 4-1: Số mẫu dữ liệu đã được phân chia trong các tập dữ liệu con của bộ dữ liệu DailyDialog	37
---	----

DANH MỤC CÁC CHỮ VIẾT TẮT

AAM	Active Appearance Model
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASD	Autism Spectrum Disorder
BART	Bidirectional and Auto-Regressive Transformer
BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte Pair Encoding
EDA	Electrodermal Activity
GPT	Generative Pre-trained Transformer
HAL	Heuristically Programmed Algorithmic Computer
HMM	Hidden Markov Model
LLM	Large Language Model
NPC	Non-Player Character
PTSD	Post-Traumatic Stress Disorder
RL	Reinforcement Learning
T5	Text-to-Text Transfer Transformer
SQuAD	Stanford Question Answering Dataset

CHƯƠNG 1. MỞ ĐẦU VÀ TỔNG QUAN ĐỀ TÀI

1.1 Lý do chọn đề tài

Hiện nay, sự phát triển của AI đã giúp cho con người giải quyết được bài toán nhân lực và hiện đại hoá, tuy nhiên vẫn còn khoảng cách khá lớn giữa người và máy trong việc hiểu và đưa ra phản hồi có cảm xúc theo nhiều ngữ cảnh khác nhau như: tư vấn tâm lý, phản hồi theo nhiều cảm xúc

Chúng tôi mong muốn nghiên cứu về cảm xúc AI theo nhiều góc độ khác nhau, cũng như tính khả thi trong việc áp dụng các yếu tố cảm xúc của con người vào quá trình tính toán, phản hồi của AI, giúp câu trả lời mang tính “con người” hơn, đánh giá được tính khả thi của đề tài để áp dụng ra ngoài thực tế trong việc tiếp thị, y tế, học đường

1.2 Mục tiêu thực hiện đề tài

Tìm hiểu được các mặt của của cảm xúc AI thông qua các câu hỏi:

- Cảm xúc có tác động như thế nào đến câu trả lời/hành vi của AI?
- Các phương pháp đưa yếu tố cảm xúc vào AI.
- Liệu việc đưa thêm yếu tố cảm xúc vào AI có tăng tính hiệu quả công việc?
- Liệu việc đưa thêm yếu tố cảm xúc vào AI có giảm bớt các nguy hiểm được dự báo của AI?

Từ đó hiểu được cảm xúc AI là gì, hiểu được lịch sử hình thành và phát triển, các công trình nghiên cứu và ứng dụng thực tế, các tác động tích cực và tiêu cực, cơ hội và thách thức. Các phương pháp đưa yếu tố cảm xúc vào mô hình AI, hiện thực và đánh giá được một trong các phương pháp này

Đánh giá được khả năng phản hồi có cảm xúc của mô hình BART trong bài toán hỏi đáp hội thoại, từ đó nêu được hướng phát triển của đề tài với các phương pháp đưa yếu tố cảm xúc khác nhau vào mô hình BART nói riêng và các mô hình ngôn ngữ lớn nói chung

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Cảm xúc AI là gì, lịch sử hình thành và phát triển của tính toán cảm xúc (Affective Computing)

2.1.1 Tính toán cảm xúc là gì và định nghĩa của cảm xúc theo nhiều lý thuyết khác nhau?

Cảm xúc AI, trong tính toán gọi là tính toán cảm xúc (Affective Computing) là lĩnh vực nghiên cứu về việc tạo ra và tương tác với các hệ thống máy tính có khả năng nhận diện, phản ứng và ảnh hưởng đến cảm xúc của con người. Đây là một lĩnh vực liên ngành, kết hợp các nghiên cứu từ tâm lý học, sinh lý học, kỹ thuật, xã hội học, toán học, khoa học máy tính, giáo dục và ngôn ngữ học. Sự đa dạng trong các ngành liên quan phản ánh sự phức tạp của việc mô tả, hiểu và mô phỏng trải nghiệm cảm xúc của con người (Daily et al., 2017).

Tính toán cảm xúc không chỉ là công cụ để máy móc nhận diện cảm xúc mà còn nhằm mục đích tạo ra khả năng phản hồi tự nhiên, phù hợp với cảm xúc con người, nó đóng vai trò quan trọng trong tương tác xã hội, do đó đòi hỏi các hệ thống phải hiểu và đáp ứng tốt, linh hoạt cảm xúc trong nhiều ngữ cảnh khác nhau để cải thiện trải nghiệm người dùng.

Tính toán cảm xúc không chỉ dựa trên công nghệ mà còn kết hợp các lĩnh vực như tâm lý học (để hiểu cảm xúc), thị giác máy tính (để nhận diện khuôn mặt), học máy (để dự đoán cảm xúc), và khoa học hành vi (để hiểu hành vi cảm xúc của con người). Việc kết hợp giữa các lĩnh vực khác nhau giúp cho lĩnh vực này trở nên phong phú nhưng cũng đầy thách thức, vì nó yêu cầu hiểu biết sâu rộng từ nhiều khía cạnh.

Tính toán cảm xúc, tuy có vẻ hàn lâm nhưng chúng xuất phát từ những cảm xúc ngoài đời thực của con người. Vậy cảm xúc là gì? Đây vẫn là câu hỏi mở, một số người định nghĩa nó là sự thay đổi sinh lý trong cơ thể, trong khi những người khác xem đó là quá trình tư duy bình thường của mỗi người theo hoàn cảnh. Cảm xúc có thể được hiểu theo các nhà khoa học với những lý thuyết khác nhau (Malatesta et al., 2009):

- Theo Klaus Scherer, cảm xúc là một chuỗi các thay đổi có liên quan và đồng bộ trong trạng thái của các hệ thống trong cơ thể, được kích hoạt bởi một kích thích nào đó bên trong lẫn bên ngoài, khi kích thích đó liên quan đến những mối quan tâm của cơ thể
- Các lý thuyết tâm lý hiện nay thường giả định rằng ba thành phần chính của cảm xúc bao gồm:
 - + Trải nghiệm chủ quan (subjective experience)
 - + Phản ứng sinh lý ngoại vi (peripheral physiological response)
 - + Biểu hiện vận động (motor expression)
- Một số lý thuyết cũng bao gồm các yếu tố nhận thức và động cơ. Những xu hướng hành động (action tendencies) cũng được liên kết chặt chẽ với trạng thái cảm xúc, giúp chuẩn bị các phản ứng thích nghi cho phù hợp theo từng hoàn cảnh
- Scherer cũng phân biệt cảm xúc với các hiện tượng cảm xúc khác như cảm giác (feeling), tâm trạng (mood), và thái độ (attitude):
 - + Cảm giác: Là thành phần "trải nghiệm chủ quan" của cảm xúc, phản ánh toàn bộ quá trình đánh giá nhận thức, động cơ, và phản ứng cơ thể liên quan đến một trạng thái cảm xúc cụ thể.
 - + Tâm trạng: Là trạng thái cảm xúc kéo dài, ít mãnh liệt hơn và ít tạo ra sự đồng bộ trong phản ứng.
 - + Thái độ: Là những niềm tin hoặc khuynh hướng lâu dài đối với một đối tượng, người, hoặc tình huống, ít thay đổi bởi kích thích ngắn hạn.

2.1.2 Lịch sử hình thành và phát triển của cảm xúc AI

Theo (Daily et al., 2017), mặc dù hiện nay Affective Computing được công nhận là một lĩnh vực khoa học, thuật ngữ này chỉ mới được đặt ra vào cuối những năm 1990. Tuy nhiên, các ý tưởng liên quan đến Affective Computing và trí tuệ nhân tạo (AI) đã tồn tại từ rất lâu trước khi những thuật ngữ này xuất hiện. Một trong những

ví dụ lâu đời nhất là hình tượng golem trong văn hóa Do Thái - một sinh vật được tạo ra từ vật chất thô (như đất sét) và được thổi hồn bởi một sức mạnh siêu nhiên

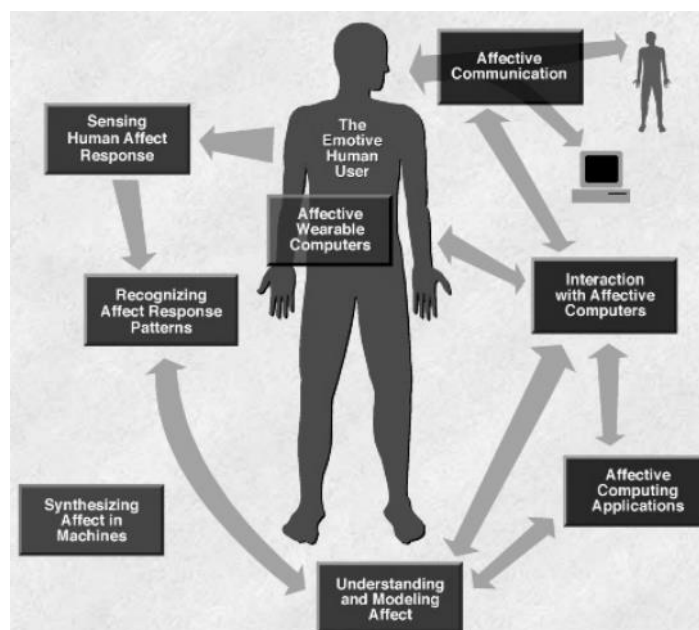
Những câu chuyện về golem xuất hiện từ khoảng năm 1000 TCN trong kinh thánh Do Thái (Tanakh), nơi từ “golem” được nhắc đến trong Sách Thi thiên (Tehillim 139:16), có thể nói mô hình AI ban đầu giống như những “golem” vô tri vô giác, giống như một phôi thai. Ví dụ khác là câu chuyện về việc Chúa tạo ra con người từ bụi đất và thổi sự sống vào đó (Gen 2:7, 2:15), câu chuyện này tượng trưng cho việc con người sáng tạo và cải tiến các mô hình AI từ những mô hình tính toán thông thường thành những mô hình hiện đại như ngày nay

Bên cạnh các câu chuyện thần thoại, từ lâu trong khoa học đã có ý tưởng về việc tạo ra máy móc tự động có khả năng tư duy và làm việc giống như con người. Một trong các bộ phim điện hình về chủ đề này đó là bộ **2001: A Space Odyssey (của Clarke, 1968; Kubrick, 1968)**, bộ phim này nói về hệ thống trí tuệ nhân tạo HAL (Heuristically Programmed Algorithmic Computer), nó là một hệ thống có khả năng nhận diện giọng nói, tư duy logic, xử lý thông tin và kiểm soát toàn bộ hệ thống của tàu Discovery One để đi đến Sao Mộc. Nó giao tiếp với phi hành đoàn bằng giọng nói mượt mà, điềm tĩnh chứng tỏ nó có thể hiểu và phản hồi với cảm xúc như con người. Tuy nhiên bộ phim cũng nói đến việc những sai sót và dấu hiệu bất thường của hệ thống khiến các phi hành gia buộc phải tắt nó, trước khi nó bị vô hiệu hoá hoàn toàn, nó đã van xin tuyệt vọng và “hát” bài Daisy Bell. Ý nghĩa của hệ thống HAL trong bộ phim này nó là biểu tượng của sự xung đột giữa con người và máy móc, khi trí tuệ nhân tạo trở nên quá mạnh mẽ, phản ánh nỗi sợ hãi về sự mất kiểm soát trước AI, đây là một chủ đề rất còn phù hợp về ngày nay và được dư luận quan tâm, quan trọng nhất nó thể hiện một nghịch lý rằng dù là cỗ máy vô cảm, HAL lại bộc lộ những cảm xúc và phản ứng gần giống con người trước khi bị tắt đi

Quá trình tạo ra một hệ thống trí tuệ nhân tạo rất phức tạp, phải trải qua nhiều bước, mất rất nhiều thời gian, bao gồm: nhận diện cảm xúc (Affective Sensing hay Affective Detection, Emotion Detection), tạo cảm xúc (Affective Generation, Emotion Generation):

- **Nhận diện cảm xúc:** Nhận diện cảm xúc là bước đầu tiên để tạo ra một hệ thống trí tuệ nhân tạo có khả năng phản ứng cảm xúc. Để làm được điều này, máy tính cần được trang bị phần cứng và phần mềm để thu thập, phân tích và diễn giải cảm xúc từ con người. Nhận diện cảm xúc đề cập đến các hệ thống có thể nhận diện cảm xúc thông qua dữ liệu tín hiệu và mô hình hóa hành vi. Các hệ thống này có thể được phân loại theo nhiều kênh cảm xúc khác nhau, mỗi kênh có đặc điểm riêng. Phần này tập trung vào các phương pháp cảm nhận cảm xúc qua biểu cảm khuôn mặt, tư thế, cử chỉ, giọng nói, văn bản và hoạt động điện da (EDA - Electrodermal Activity) (Daily et al., 2017)
- **Tạo cảm xúc:** Sau khi một hệ thống AI có thể nhận diện cảm xúc (Affect Sensing), bước tiếp theo là tạo ra phản ứng cảm xúc phù hợp. Điều này có thể được thực hiện thông qua biểu cảm trên robot, nhân vật ảo và các giao diện tương tác thông minh (Daily et al., 2017)

2.1.3 Các lĩnh vực nghiên cứu của tính toán cảm xúc



Hình 2-1: Hình ảnh trực quan của các nghiên cứu về các lĩnh vực của tính toán cảm xúc từ Viện Công nghệ Massachusetts (MIT)

Nguồn: (Malatesta et al., 2009)

Theo (Malatesta et al., 2009), những khía cạnh chính bao gồm:

+ **Nhận diện cảm xúc (Detecting and Recognizing Emotional Information):** quá trình này thường sử dụng các cảm biến thụ động để thu thập dữ liệu về trạng thái vật lý hoặc hành vi của người dùng, giống như cách con người nhận biết cảm xúc của nhau

- **Nhận diện cảm xúc qua phân tích nét mặt:**

- Hệ thống có thể phát hiện và phân tích các biểu cảm trên khuôn mặt, mỗi cảm xúc có những đặc điểm khuôn mặt riêng biệt, các biểu hiện cảm xúc được quy định tùy theo các nhà nghiên cứu khác nhau như 6 cảm xúc cơ bản của Paul Ekman, bánh xe cảm xúc của nhà tâm lý học Robert Plutchik (Daily et al., 2017)
- AI sử dụng các công nghệ phân tích nét mặt thông qua camera và thuật toán máy học nhận diện các trạng thái cảm xúc trên dựa trên biểu cảm, hệ thống sẽ ghi nhận các góc độ của các bộ phận trên mặt để đưa ra trạng thái cảm xúc tương ứng (Malatesta et al., 2009)
- Các thuật toán được dùng để nhận diện ra các cảm xúc này bao gồm: Mô hình Markov ẩn (Hidden Markov Model - HMM), dòng quang học (Optical Flow), mô hình diện mạo động (Active Appearance Model - AAM), mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN) (Daily et al., 2017)

- **Nhận diện cảm xúc qua tư thế và cử chỉ theo (Daily et al., 2017):**

- Các nhà nghiên cứu cho rằng tư thế và cử chỉ cũng có thể truyền tải cảm xúc cụ thể lẫn trạng thái cảm xúc tổng quát
- Tư thế (Posture), trạng thái cảm xúc có thể được tiết lộ qua cách một người đứng hoặc ngồi. Trong khi Cử chỉ (Gestures) bao gồm các chuyển động của tay hoặc đầu có thể biểu đạt cảm xúc rõ ràng

- Các công nghệ nhận diện cử chỉ: Găng tay cảm biến (Wired Gloves), camera cảm biến độ sâu (Depth-Aware Cameras), camera stereo (Stereo Cameras), mô hình 3D (3D Model-Based Algorithms), mô hình xương (Skeletal-Based Algorithms). Các mô hình này sử dụng khung xương ảo, hình ảnh, video để nhận diện cảm xúc
- **Nhận diện cảm xúc qua giọng nói:**
 - Giọng nói là một kênh quan trọng để biểu đạt cảm xúc. Các đặc điểm giọng nói có thể tiết lộ cảm xúc của người nói bao gồm: Tốc độ nói (Speech Rate), cao độ trung bình (Pitch Average), phạm vi cao độ (Pitch Range), cường độ (Intensity), chất lượng giọng nói (Voice Quality) (Daily et al., 2017)
 - Các thuật toán máy học và các thiết bị có thể phân tích giọng nói để xác định cảm xúc thông qua các đặc trưng giọng nói (Malatesta et al., 2009)
- **Nhận diện cảm xúc qua văn bản (Daily et al., 2017):**
 - Sử dụng các công cụ xử lý ngôn ngữ tự nhiên (NLP - Natural Language Processing để đánh giá tâm trạng người viết
 - Một số công cụ phổ biến: WordNet-Affect (Keshtkar & Inkpen, 2013), SenticNet (Denecke, 2008), SentiWordNet (Poria et al., 2012), hiện nay nhiều mạng nơ ron nhân tạo được phát triển để thực hiện nhận diện cảm xúc
 - Các phương pháp phân tích cảm xúc trong văn bản: Phát hiện từ khóa (Keyword Spotting), phân tích liên kết từ vựng (Lexical Affinity), phương pháp thống kê (Statistical Methods), phân tích theo ngữ cảnh (Attention)
- **Các phương pháp nhận diện cảm xúc khác:** Thông qua các cảm biến sinh lý học để thu thập dữ liệu như nhịp tim, nhiệt độ cơ thể để đo trạng thái cảm xúc. (Malatesta et al., 2009). Hoạt động điện da (EDA -

Electrodermal Activity) đo phản ứng sinh lý của da với cảm xúc (Daily et al., 2017)

- **Ví dụ:** Trợ lý ảo Alexa hay Siri điều chỉnh giọng điệu trả lời dựa trên nhận diện cảm xúc của người dùng
- + **Hiểu và tạo cảm xúc (Understanding and generating emotions):** hệ thống không chỉ cần phát hiện cảm xúc mà còn phải xây dựng, lưu trữ và duy trì một mô hình cảm xúc về người dùng để đưa ra phản hồi phù hợp với ngữ cảnh
- AI không chỉ phản hồi cảm xúc, mà còn biểu lộ cảm xúc qua:
 - Hình ảnh hoặc robot: Robot xã hội biểu hiện gương mặt hoặc cử chỉ cảm xúc, có khả năng tương tác với con người theo cách tự nhiên, bao gồm: Hiểu ngôn ngữ và hành vi xã hội, thể hiện ý định một cách dễ hiểu, cộng tác với con người và các robot khác để đạt được mục tiêu chung. (Daily et al., 2017)
 - Giọng nói cảm xúc: Nhân vật ảo có cảm xúc có thể phát âm với âm điệu mang tính biểu cảm, giúp tạo ra trải nghiệm tương tác chân thực hơn trong các lĩnh vực như giáo dục, trò chơi và hỗ trợ trị liệu. (Daily et al., 2017). Ví dụ: chatbot, trợ lý ảo, nhân vật trò chơi (NPC - non-player character)
- Mục tiêu: Tăng tính tương tác xã hội giữa con người và máy, đặc biệt trong các môi trường cần sự đồng cảm như chăm sóc sức khỏe hoặc giáo dục

Mục tiêu của Cảm xúc AI

- Tăng cường sự hiểu biết và tin tưởng giữa con người và máy.
- Phát triển các hệ thống giao tiếp tự nhiên hơn trong các lĩnh vực như y tế, bán lẻ, và công nghệ hỗ trợ

2.2 AI cảm xúc có tác động như thế nào trong nền kinh tế cảm xúc (Emotional economies)?

AI cảm xúc đang thúc đẩy sự phát triển của nền kinh tế cảm xúc - nơi mà cảm xúc không chỉ là một yếu tố giao tiếp mà còn là tài nguyên kinh tế.

Kinh tế cảm xúc là gì?

- **Khái niệm:** Cảm xúc (từ dữ liệu người dùng hoặc biểu cảm) trở thành tài sản thương mại, được sử dụng trong quảng cáo, thiết kế sản phẩm, và trải nghiệm dịch vụ, tạo thành nền kinh tế gọi là nền kinh tế cảm xúc (Emotional economies), nền kinh tế này được hỗ trợ và mở rộng bởi sự phát triển của công nghệ (Patulny et al., 2020)
- **Ứng dụng:**
 - + Quảng cáo thông minh: AI phân tích cảm xúc từ dữ liệu người dùng để điều chỉnh nội dung quảng cáo, tối ưu hóa hiệu quả tiếp cận
 - + Dịch vụ khách hàng: AI cảm xúc hỗ trợ cải thiện trải nghiệm người dùng, ví dụ phân tích giọng điệu của khách hàng trong cuộc gọi hỗ trợ để điều chỉnh phản hồi

Cảm xúc trong lao động

- **Lao động cảm xúc (Emotional labor):**
 - + Con người thực hiện các công việc cần sự đồng cảm và giao tiếp cảm xúc, như chăm sóc khách hàng.
 - + AI có khả năng hỗ trợ hoặc thay thế những công việc này, đặc biệt trong các lĩnh vực như bán lẻ hoặc dịch vụ khách sạn.
- **Dữ liệu cảm xúc:** Dữ liệu cảm xúc được thu thập từ hoạt động trực tuyến (like, bình luận) hoặc cảm biến sinh lý để phân tích xu hướng và hành vi khách hàng

Tương lai công việc: Khi công nghệ tiếp tục thay thế các công việc thủ công và phân tích, các công việc yêu cầu kỹ năng cảm xúc sẽ trở nên quan trọng hơn. Tuy nhiên, AI cũng có thể thay thế một phần lĩnh vực này, gây ra các thách thức về xã hội và lao động

2.3 Cơ hội và thách thức của AI cảm xúc

2.3.1 Cơ hội

Tăng cường giao tiếp: AI cảm xúc giúp tăng hiệu quả giao tiếp và giảm hiểu lầm giữa con người và máy

Ứng dụng đa ngành:

- Giáo dục: AI hỗ trợ học tập cá nhân hóa, nhận biết khi học sinh gặp khó khăn cảm xúc để kịp thời tư vấn, thay đổi lộ trình học khi cần, tổ chức các thời gian nghỉ với các trò chơi cho học sinh thư giãn.
- Y tế: Hỗ trợ điều trị sức khỏe tinh thần, phát hiện sớm các rối loạn cảm xúc, bố trí bác sĩ nhanh và phù hợp cho từng bệnh nhân.
- Phân tích hành vi: AI giúp phân tích hành vi khách hàng hoặc nhân viên, hỗ trợ ra quyết định quản lý tốt hơn, chẳng hạn như phát hiện biểu hiện nhân viên khi làm việc để điều chỉnh thời gian và khối lượng công việc phù hợp
- Giám sát các sự kiện: Trong bối cảnh hiện nay có rất nhiều sự kiện lớn được tổ chức tại các nơi đông người, nơi công cộng, việc xuất hiện những người có hành vi đáng ngờ là không tránh khỏi, AI cảm xúc có thể phân tích các cử chỉ, khuôn mặt để xác định sớm và gợi ý biện pháp ngăn chặn kịp thời cho nhân viên an ninh (Daily et al., 2017)
- Giải trí và trò chơi: Game có thể điều chỉnh độ khó hoặc câu chuyện, các NPC theo cảm xúc người chơi, ngoài ra có thể sáng tạo các điệu nhạc, phim theo tâm trạng của người dùng. Ví dụ nếu game thủ có biểu hiện tức giận khi thua một con quái quá nhiều, AI sẽ giảm độ khó của con quái xuống để người chơi có thể dễ dàng đánh nó hơn, giúp xoa dịu một phần tâm trạng của người đó

2.3.2 Thách thức

Độ chính xác và hiệu quả: Cảm xúc của con người không phải lúc nào cũng rõ ràng hoặc có thể đo lường chính xác, nó mang tính chủ quan và rất phức tạp. Ngoài

ra, AI chủ yếu dựa vào dữ liệu huấn luyện, nhưng dữ liệu có thể bị thiên lệch (bias). Ví dụ: Một người có thể cười nhưng không thực sự hạnh phúc.

Đạo đức và quyền riêng tư: Dữ liệu cảm xúc là thông tin nhạy cảm, hệ thống AI cảm xúc có khả năng thu thập thông tin nhạy cảm về trạng thái tâm lý người dùng bao gồm biểu cảm khuôn mặt, giọng nói, tín hiệu sinh lý và văn bản, các hệ thống này có thể sử dụng các cảm xúc đã thu thập được để lên lút phân tích cảm xúc của người dùng. Ví dụ một hệ thống AI có thể nhận diện khi người dùng căng thẳng, buồn bã, sau đó hiển thị quảng cáo hoặc nội dung phù hợp. Ta có thể thấy mặc dù thời gian đầu rất thuận tiện nhưng về lâu dài lại gây tâm lý hoang mang cho người dùng, tạo cho họ cảm giác lúc này cũng bị theo dõi. Cho nên việc thu thập, lưu trữ, và sử dụng cần tuân thủ các quy định về quyền riêng tư và đạo đức (Daily et al., 2017)

"Đạo đức giả cảm xúc": AI có thể biểu lộ cảm xúc mà không thực sự "hiểu" cảm xúc, dẫn đến mất lòng tin từ người dùng. Việc cố gắng làm cho AI biểu hiện cảm xúc có thể làm cho nó giống như những người bị Psychopathy (một số bản dịch gọi là thái nhân cách).

Psychopathy chỉ những người không thực sự hiểu cảm xúc là gì, nhưng những người đó có thể thông qua môi trường xung quanh để nhận diện và giả vờ biểu hiện cảm xúc đó, do đó họ có thể bị những kẻ xấu khác thao túng tâm lý

Khó khăn để đánh giá xem kết quả phản hồi của mô hình AI có đúng như mình mong muốn hay không: Việc xây dựng các đánh giá phản hồi thủ công tuy có thể giúp cải thiện mô hình thông qua các thuật toán học tăng cường (RL - Reinforcement Learning), tuy nhiên người dùng cũng có thể đánh giá sai lệch tùy theo tâm trạng khi đó của họ, dẫn đến làm cho hệ thống sai lệch

2.4 Các cảm xúc cơ bản và các mô hình cảm xúc

2.4.1 Các cảm xúc cơ bản theo các nhà nghiên cứu

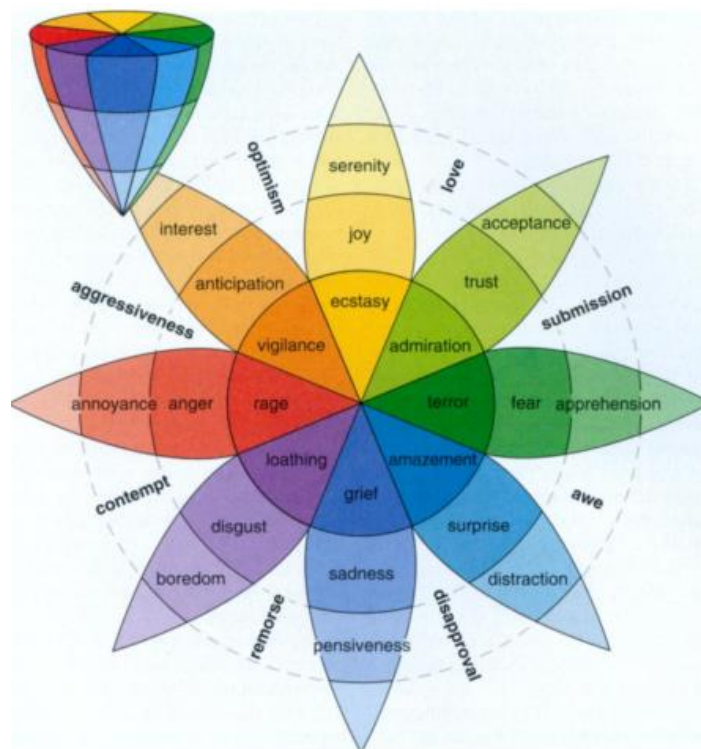
- Paul Ekman (1972) xác định 6 cảm xúc cơ bản:
- + Vui (Happiness): Trạng thái tích cực, hài lòng.

- + Buồn (Sadness): Liên quan đến mất mát hoặc thất vọng.
- + Sợ hãi (Fear): Phản ứng với nguy hiểm.
- + Tức giận (Anger): Phản ứng với cản trở hoặc bất công.
- + Ghê tởm (Disgust): Phản ứng với điều gây khó chịu.
- + Ngạc nhiên (Surprise): Phản ứng với sự kiện bất ngờ
- Theo (Martínez-Miranda & Aldea, 2005), Daniel Goleman (1995) phân loại các cảm xúc thành 8 nhóm và lập luận rằng tất cả cảm xúc đều thuộc 8 nhóm này, bao gồm:
 - + Giận dữ (Anger): cơn thịnh nộ, phẫn nộ, oán giận, tức giận, bức bối, bất bình, khó chịu, gay gắt, thù địch, phiền toái, cáu kỉnh, hằn học và ở mức cực đoan, hận thù bệnh lý và bạo lực
 - + Buồn bã (Sadness): đau buồn, sầu não, mệt mỏi, u ám, u sầu, tự thương hại, cô đơn, chán nản, tuyệt vọng, và khi trở thành bệnh lý, trầm cảm nghiêm trọng.
 - + Sợ hãi (Fear): lo lắng, e ngại, căng thẳng, quan ngại, kinh hoàng, nghi ngờ, cảnh giác, bồn chồn, bồn chồn, sợ hãi, kinh hãi, hoảng sợ; và khi trở thành bệnh lý, ám ảnh và hoảng loạn
 - + Thích thú (Enjoyment): hạnh phúc, vui vẻ, nhẹ nhõm, mãn nguyện, hân hoan, thích thú, tự hào, khoái cảm, hồi hộp, say mê, thỏa mãn, hài lòng, phấn khích, hứng thú, sáng khoái, ngẫu hứng, xuất thần và ở mức cực đoan, hưng cảm
 - + Tình yêu (Love): chấp nhận, thân thiện, tin tưởng, tử tế, gắn kết, tận tâm, tôn thờ và say mê
 - + Ngạc nhiên (Surprise): sốc, kinh ngạc và sửng sờ
 - + Ghê tởm (Disgust): khinh miệt, coi thường, khinh bỉ, ghê sợ, ác cảm, chán ghét và căm phẫn
 - + Xấu hổ (Shame): tội lỗi, lúng túng, tủi hổ, hối hận, nhục nhã, tiếc nuối, ê chề và hối cải

- Bài báo (Plutchik, 2001) nói rằng Robert Plutchik đã phát triển mô hình bánh xe cảm xúc (Wheel of Emotions), trong đó ông phân loại cảm xúc con người thành 8 cảm xúc cơ bản, từ đó kết hợp để tạo nhiều cảm xúc phức tạp hơn, 8 cảm xúc cơ bản theo ông định nghĩa bao gồm:

- + Joy (Vui vẻ): Hạnh phúc, sung sướng
- + Trust (Tin tưởng): An toàn, chấp nhận
- + Fear (Sợ hãi): Lo lắng, căng thẳng
- + Surprise (Ngạc nhiên): Sững sờ, kinh ngạc
- + Sadness (Buồn bã): Đau khổ, thất vọng
- + Disgust (Ghê tởm): Chán ghét, ghê sợ
- + Anger (Giận dữ): Bực bội, phẫn nộ
- + Anticipation (Trông đợi): Mong chờ, hứng thú

Plutchik cho rằng cảm xúc có thể pha trộn như màu sắc, tạo thành những cảm xúc phức tạp hơn



Hình 2-2: Mô hình bánh xe cảm xúc của Plutchik

Nguồn: (Plutchik, 2001)

2.4.2 Các mô hình tính toán cảm xúc

Theo (Malatesta et al., 2009), có 2 loại mô hình chính:

- Mô hình sâu (Deep Models): Tập trung vào các nguyên nhân và quá trình dẫn đến cảm xúc. Ví dụ mô hình dựa trên lý thuyết đánh giá để xác định cảm xúc dựa theo cách người dùng đánh giá một sự kiện
- Mô hình nông (Shallow Models): Tập trung vào kết quả của cảm xúc. Ví dụ mô hình nhận diện nét mặt bằng cách phân tích dữ liệu từ camera

(Malatesta et al., 2009) cũng nêu ra một số mô hình nổi bật:

- **Mô hình đa thành phần (Scherer):**
 - + Cảm xúc là sự phối hợp giữa các thay đổi trong nhận thức, động lực, và phản ứng cơ thể.
 - + Mô hình này dựa trên các quá trình đánh giá liên tục của con người, bao gồm: mức độ mới mẻ của sự kiện, tính phù hợp với mục tiêu, khả năng đối phó, và tính tương thích với chuẩn mực xã hội.
 - + Phù hợp cho các hệ thống cần phản hồi phức tạp
 - + Ví dụ: Khi sợ hãi, tim đập nhanh (cơ thể), nhận thức nguy hiểm (nhận thức), và hành động chạy trốn (động lực)
- **Mô hình OCC (Ortony, Clore, và Collins):**
 - + Xác định cảm xúc dựa trên ba yếu tố: sự kiện, hành động, và đối tượng dựa trên 3 yếu tố chính:
 - Sự hài lòng với kết quả của một sự kiện
 - Sự chấp thuận hoặc phản đối hành động của một cá nhân.
 - Sự thích hoặc không thích một đối tượng.
 - + Được triển khai trong nhiều hệ thống AI, đặc biệt là các trợ lý ảo và trò chơi.
- **Mô hình hai chiều (Valence-Arousal):**
 - + Cảm xúc được định nghĩa dựa trên các chiều như
 - Valence (giá trị): Mức độ tích cực (hạnh phúc)/tiêu cực (buồn bã).

- Arousal (kích hoạt): Mức độ kích thích cao (phần khích)/thấp (bình tĩnh).
- + Phương pháp này hữu ích trong việc biểu diễn các cảm xúc phức tạp hoặc cảm xúc có cường độ thấp
- + Ví dụ:
 - "Hạnh phúc" có Valence cao, Arousal thấp.
 - "Tức giận" có Valence thấp, Arousal cao

2.5 Ứng dụng thực tế về cảm xúc AI trong các lĩnh vực

2.5.1 Trong giáo dục

Affective Computing có tiềm năng cải thiện trải nghiệm học tập bằng cách giúp giáo viên hiểu rõ hơn về trạng thái cảm xúc và mức độ tham gia của học sinh.

Hệ thống giám sát sự tham gia của học sinh (Daily et al., 2017)

- EngageMe (Darnell, 2014) sử dụng dữ liệu cảm xúc từ học sinh (chẳng hạn như độ dẫn điện da và biểu cảm khuôn mặt) để giúp giáo viên đánh giá mức độ hứng thú trong lớp học.
- Giáo viên có thể xem biểu đồ mức độ tương tác của từng học sinh theo thời gian thực để điều chỉnh phương pháp giảng dạy.
- Ví dụ: Classroom AI sử dụng camera để theo dõi biểu cảm khuôn mặt và tư thế của học sinh. Nếu hệ thống nhận thấy học sinh có dấu hiệu buồn chán hoặc căng thẳng, nó sẽ đề xuất phương pháp giảng dạy phù hợp hơn.

Công cụ giao tiếp cảm xúc giữa giáo viên và học sinh (Daily et al., 2017)

- Subtle Stone là một thiết bị cầm tay giúp học sinh thể hiện cảm xúc bằng cách bóp một viên đá có đèn màu (mỗi màu tương ứng với một trạng thái cảm xúc khác nhau).
- Tuy nhiên, hệ thống này phụ thuộc nhiều vào khả năng tự đánh giá cảm xúc của học sinh và có thể gây xao nhãng trong lớp học.

Gia sư AI thông minh (Daily et al., 2017)

- Các hệ thống trợ lý giảng dạy thông minh có thể nhận diện khi học sinh cảm thấy khó hiểu, chán nản hoặc bối rối, từ đó điều chỉnh cách giảng bài để giúp họ tiếp thu tốt hơn.
- Một số nghiên cứu đã kết hợp xử lý ngôn ngữ tự nhiên (NLP) và phân tích nét mặt để đánh giá mức độ hứng thú của học sinh trong quá trình học tập

Lợi ích:

- Cá nhân hóa bài học dựa trên cảm xúc.
- Phát hiện sớm học sinh gặp khó khăn tinh thần hoặc không theo kịp bài giảng.

2.5.2 Trong y tế

AI cảm xúc đang hỗ trợ chẩn đoán, theo dõi và điều trị một số tình trạng tâm lý và thần kinh, đặc biệt là với những bệnh nhân gặp khó khăn trong việc diễn đạt cảm xúc

Hỗ trợ bệnh nhân mắc chứng tự kỷ (ASD - Autism Spectrum Disorder) (Daily et al., 2017)

- Những người mắc chứng tự kỷ (ASD) thường gặp khó khăn trong việc nhận diện và thể hiện cảm xúc.
- Các ứng dụng như SymTrend và Autism Track giúp bệnh nhân tự kỷ theo dõi hành vi và tâm trạng của họ theo thời gian, giúp bác sĩ và nhà trị liệu có thêm dữ liệu để điều chỉnh phương pháp điều trị.

Điều trị chứng rối loạn căng thẳng sau sang chấn (PTSD - Post-Traumatic Stress Disorder) (Daily et al., 2017)

- PTSD ảnh hưởng đến những người từng trải qua sang chấn nghiêm trọng (ví dụ: cựu chiến binh, nạn nhân bạo lực).
- Một số ứng dụng Affective Computing sử dụng thực tế ảo (VR) để giúp bệnh nhân tái tạo môi trường gây căng thẳng trong môi trường an toàn.

- StartleMart (Holmgard et al., 2013) là một hệ thống sử dụng thực tế ảo kết hợp với đo tín hiệu da (EDA) để xác định mức độ căng thẳng của người dùng và điều chỉnh bài trị liệu phù hợp.

Robot hỗ trợ y tế: Robot như Pepper hoặc Paro được thiết kế để tương tác với bệnh nhân, đặc biệt là người cao tuổi hoặc bệnh nhân tâm thần.

- Robot phân tích giọng nói, nét mặt để nhận diện cảm xúc như buồn bã hoặc lo lắng.
- Hỗ trợ liệu pháp tinh thần thông qua tương tác cảm xúc.

Ứng dụng di động: Ứng dụng như Woebot sử dụng AI để nhận diện tâm trạng người dùng qua trò chuyện văn bản, từ đó gợi ý cách giải tỏa căng thẳng.

- Hỗ trợ tâm lý cho bệnh nhân mọi lúc.
- Cải thiện chất lượng sống và giảm gánh nặng cho nhân viên y tế.

2.5.3 Dịch vụ khách hàng

Các cửa hàng, trung tâm dịch vụ sử dụng AI cảm xúc để đánh giá trải nghiệm của khách hàng về sản phẩm của mình. Nhằm mục đích giải quyết vấn đề nhanh chóng và chính xác hơn, nâng cao sự hài lòng của khách hàng

Call Center AI: Các trung tâm hỗ trợ khách hàng sử dụng AI phân tích giọng điệu của khách hàng để xác định cảm xúc như tức giận, lo lắng, hay hài lòng.

Ví dụ: Cogito AI cung cấp phản hồi thời gian thực cho nhân viên để điều chỉnh giọng điệu hoặc câu trả lời.

Chatbot thông minh: Chatbot như Replika sử dụng AI để giao tiếp đồng cảm với người dùng.

2.5.4 Ứng dụng trong giao tiếp và phát triển kỹ năng cá nhân

Hỗ trợ cải thiện kỹ năng giao tiếp (Daily et al., 2017)

- MACH (My Automated Conversation Coach) là một hệ thống giúp người dùng rèn luyện kỹ năng giao tiếp bằng cách nhận diện giọng nói, nét mặt và cử chỉ trong thời gian thực.

- Hệ thống sẽ phân tích cách người dùng thể hiện cảm xúc và đưa ra phản hồi để giúp họ cải thiện cách nói chuyện và kiểm soát cảm xúc tốt hơn

Trợ lý ảo đồng cảm (Daily et al., 2017)

- Các trợ lý ảo thông minh, như Google Assistant, Siri hoặc Alexa, đang được tích hợp khả năng nhận diện cảm xúc từ giọng nói để đưa ra phản hồi phù hợp với tâm trạng của người dùng.
- Một số nghiên cứu còn phát triển trợ lý ảo có thể đồng cảm, giúp người dùng giảm căng thẳng hoặc đưa ra lời khuyên phù hợp với trạng thái cảm xúc hiện tại.

2.5.5 Quảng cáo và tiếp thị

AI cảm xúc sẽ sử dụng các công nghệ nhận diện cảm xúc của người dùng thông qua gương mặt, cử chỉ để phân tích xem khách hàng cần gì, từ đó đưa ra chiến lược tiếp thị phù hợp. Việc nhận diện cảm xúc của khách hàng còn có thể giúp tối ưu chiến dịch quảng cáo, qua đó tăng doanh thu thông qua trải nghiệm cá nhân hoá theo từng khách hàng

Phân tích cảm xúc khách hàng:

- Ví dụ: Affectiva sử dụng công nghệ nhận diện nét mặt để phân tích phản ứng cảm xúc của người dùng khi xem quảng cáo.
- Từ đó, doanh nghiệp điều chỉnh nội dung quảng cáo để phù hợp hơn với khách hàng.

Cửa hàng thông minh: Cảm biến đặt trong cửa hàng nhận diện biểu cảm của khách hàng khi xem sản phẩm, giúp doanh nghiệp hiểu sở thích và hành vi mua sắm.

2.5.6 Ô tô tự lái

AI cảm xúc có thể giúp giảm thiểu tai nạn giao thông, nâng cao trải nghiệm của khách hàng, tăng tính an toàn khi sử dụng xe tự lái và các loại xe khác

Phân tích tài xế:

- Các công ty như Tesla hoặc Affectiva Automotive AI tích hợp camera để nhận diện cảm xúc tài xế, phát hiện mệt mỏi, căng thẳng hoặc mất tập trung.
- Xe có thể đưa ra cảnh báo hoặc tự động giảm tốc nếu nhận thấy tài xế mất tỉnh táo.

Tương tác hành khách: Trong xe tự lái, AI cảm xúc được sử dụng để đảm bảo hành khách cảm thấy thoải mái và an toàn.

2.5.7 Mạng xã hội

Trong thời đại bùng nổ của mạng xã hội như ngày nay, việc xuất hiện rất nhiều bài đăng hoặc thông tin sai lệch, mang tính tiêu cực ngày một tăng lên, làm cho người dùng mệt mỏi, nhất là trẻ con sẽ làm theo những lời nói trên mạng. AI cảm xúc có thể phát hiện và hạn chế các bài đăng đó, thông qua đó nếu phát hiện tâm trạng thất thường của người dùng thông qua các bài đăng, lượt tương tác trên mạng xã hội, hệ thống sẽ gợi ý các chuyên gia tâm lý thích hợp

AI cảm xúc trong mạng xã hội giúp cải thiện sức khỏe tinh thần của người dùng, tăng thời gian sử dụng nền tảng và sự hài lòng của người dùng

Phân tích tâm trạng bài đăng: Facebook hoặc Instagram sử dụng AI để phân tích tâm trạng qua bài đăng, bình luận hoặc hình ảnh.

- Ví dụ: Nếu phát hiện dấu hiệu trầm cảm, hệ thống có thể gợi ý người dùng tìm sự hỗ trợ chuyên nghiệp.

Tùy chỉnh nội dung: AI gợi ý nội dung phù hợp với tâm trạng người dùng, từ đó tăng tương tác.

2.6 Các phương pháp đưa yếu tố cảm xúc vào mô hình AI

Gán nhãn cảm xúc vào dữ liệu huấn luyện

- Khi huấn luyện chatbot hoặc mô hình sinh văn bản (BART, T5, GPT), có thể thêm nhãn cảm xúc vào câu đầu vào. Ta có thể gán thêm trọng số (từ 0 đến 100) đối với trường hợp cảm xúc phức hợp
- Ví dụ:

- + “<Hạnh phúc> Tôi vừa được thắng chức!”
- + “<Buồn bã> Hôm nay là một ngày tệ.”
- Mô hình học cách tạo phản hồi phù hợp với cảm xúc.

Thêm embedding cảm xúc vào vector đầu vào

- Thay vì chỉ thêm cảm xúc vào câu đầu vào như phương pháp đầu tiên, ta có thể chuyển đổi các yếu tố cảm xúc thành một embedding riêng và đưa vào mô hình
- Phương pháp này giúp mô hình hiểu rõ ngữ cảnh hơn

Sử dụng mô hình dự đoán cảm xúc làm bước trung gian

- Trước khi phản hồi, AI sẽ dùng một mô hình phân tích cảm xúc để hiểu trạng thái của người dùng.
- Sau đó, AI sẽ chọn phản hồi phù hợp dựa trên cảm xúc đó.
- Ví dụ:
 - + Nếu người dùng buồn, chatbot sẽ phản hồi nhẹ nhàng, an ủi.
 - + Nếu người dùng tức giận, AI sẽ phản hồi mang tính xoa dịu.

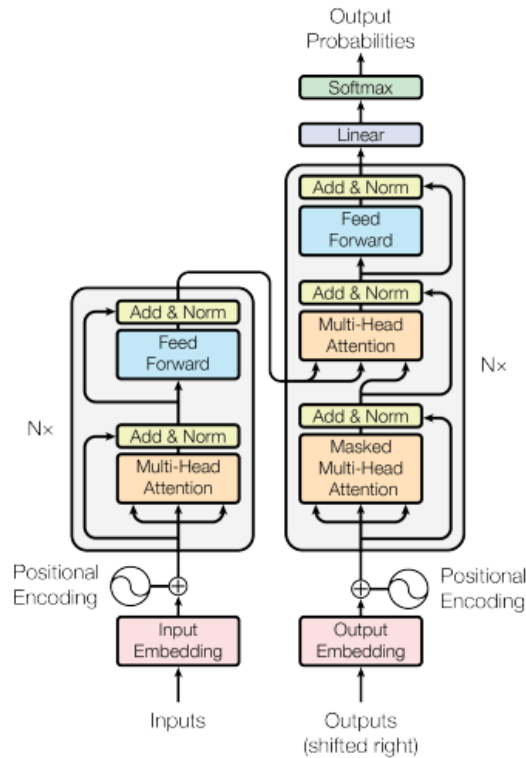
Ứng dụng học tăng cường (RL - Reinforcement Learning) khi phản hồi được người dùng đánh giá để tự động điều chỉnh tham số mô hình cho phù hợp

CHƯƠNG 3. MÔ HÌNH ĐỀ XUẤT

Chúng tôi sử dụng mô hình BART, BART được viết tắt cho Bidirectional and Auto-Regressive Transformers. BART là một mô hình Transformer được thiết kế để xử lý các tác vụ sequence-to-sequence. Có thể nói BART là mô hình được kết hợp từ BERT (Bidirectional Encoder Representations from Transformers) và GPT (Generative Pretrained Transformer). Trong đó điểm BERT phụ trách mã hoá ngữ cảnh 2 chiều, GPT mã hoá sinh ngữ cảnh 1 chiều. Đối với dữ liệu văn bản, BART sử dụng thuật toán BPE để biểu diễn câu đầu vào

3.1 Kiến trúc Transformer

(Vaswani et al., 2023) đã giới thiệu Transformer, đây là kiến trúc mạng nơ ron sâu được thiết kế để xử lý các dữ liệu dạng chuỗi như văn bản, âm thanh và đã trở thành nền tảng của nhiều mô hình hiện tại như BERT, GPT, T5, BART



Hình 3-1: Hình ảnh tổng quát Transformer

Nguồn: (Vaswani et al., 2023)

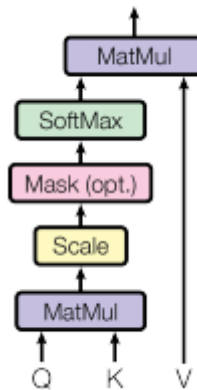
Kiến trúc Transformer gồm 2 phần chính:

- Encoder được ghép từ $N = 6$ lớp giống nhau, mỗi lớp gồm 2 tầng con:
 - + Tầng con thứ nhất là Multi-Head Self-Attention, chịu trách nhiệm học các mối quan hệ khác nhau giữa các từ, mã hoá 2 chiều để hiểu ngữ cảnh tốt hơn
 - + Tầng con thứ hai là Positionwise Fully Connected Feed-Forward Network, chịu trách nhiệm giúp mô hình học các đặc trưng phi tuyến với thông tin vị trí đầu vào
- Decoder cũng được ghép từ $N = 6$ lớp giống nhau, ngoài 2 tầng giống như Encoder, Decoder có thêm các tầng khác:

- + Tầng thứ ba là Masked Multi-Head Attention, ngoài việc học các mối quan hệ thì nó còn có thêm chức năng mã hoá 1 chiều để đảm bảo mô hình tính toán sinh câu trả lời theo thứ tự từ trước đến sau

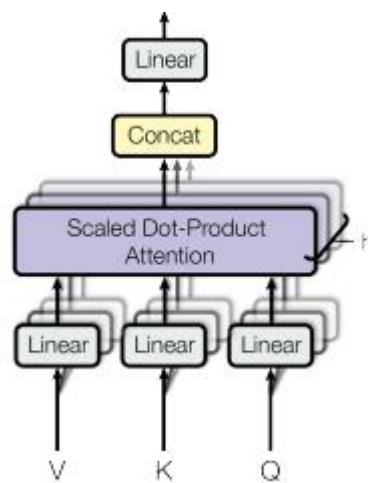
3.1.1 Multi-Head Self-Attention

Tầng này cho phép mô hình tập trung vào các từ khác nhau trong câu cùng một lúc. Nó giúp mô hình hiểu ngữ cảnh của mỗi từ dựa trên tất cả các từ khác trong câu.



Hình 3-2: Scaled Dot-Product Attention

Nguồn: (Vaswani et al., 2023)



Hình 3-3: Multi-Head Attention gồm nhiều tầng Scaled Dot-Product Attention chạy song song với nhau

Nguồn: (Vaswani et al., 2023)

Cụ thể, cho một vector đầu vào X với d chiều, Multi-Head Self-Attention tạo ra h head attention (được scale lại với hệ số $\frac{1}{\sqrt{d_k}}$ nên còn gọi là Scaled Dot-Product Attention) bằng cách sử dụng ma trận trọng số W_i khác nhau. Gọi Q , K , và V là ba loại biểu diễn của mỗi từ trong multi-head self-attention, được sử dụng để tính toán attention scores, trong đó:

- Q (Query): Đại diện cho từ hiện tại trong câu và được sử dụng để tìm kiếm thông tin từ các từ khác trong câu. Nó giúp mô hình quyết định xem từ này nên chú ý đến những gì
- K (Key): Đại diện cho từ hiện tại và được sử dụng để đánh giá mức độ quan trọng của nó đối với từ truy vấn (Query). Nó cho biết những đặc điểm nào của từ nên được quan tâm khi đưa ra quyết định
- V (Value): Đại diện cho từ hiện tại và được sử dụng để tạo ra đầu ra cuối cùng từ multi-head self-attention. Nó chứa thông tin cụ thể từ các từ được chú ý trong câu

Mỗi Scaled Dot-Product Attention (1 head) được tính bằng công thức:

$$Attention(QW_i^Q, KW_i^K, VW_i^V) = softmax\left(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right)(VW_i^V) \quad (1)$$

Nguồn: (Vaswani et al., 2023)

Với: X là dữ liệu đầu vào, các W là các ma trận trọng số của mô hình

Multi-Head Attention được tính bằng cách sử dụng Concat để tổng hợp các Attention đó lại với nhau theo công thức:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$MultiHead(X) = Concat(head_1, head_2, \dots, head_h) \times W_0 \quad (3)$$

Nguồn: (Vaswani et al., 2023)

3.1.2 Masked Self-Attention

Được dùng trong phần Decoder của Transformer, cơ chế Masked Self-Attention đảm bảo rằng trong quá trình huấn luyện và inference, mỗi vị trí i trong chuỗi chỉ có thể chú ý đến các vị trí $\leq i$

Để thực hiện được điều này, ta sẽ đánh dấu các attention score các vị trí sau i bằng cách gán giá trị $-\infty$ để khi qua hàm softmax các attention score này sẽ có giá trị là 0

SCALED SCORES					LOOK-AHEAD MASK					MASKED SCORES			
0.7	0.1	0.1	0.1	*	1	-inf	-inf	-inf	=	0.7	-inf	-inf	-inf
0.1	0.6	0.2	0.1		1	1	-inf	-inf		0.1	0.6	-inf	-inf
0.1	0.3	0.6	0.1		1	1	1	-inf		0.1	0.3	0.6	-inf
0.1	0.3	0.3	0.3		1	1	1	1		0.1	0.3	0.3	0.3

Hình 3-4: Minh hoạ Masked Self-Attention

Nguồn: (Transformers, n.d.)

Mỗi head trong Scaled Dot-Product Attention được tính bởi công thức

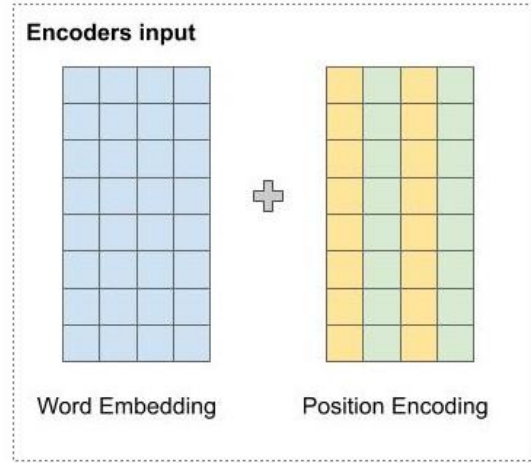
$$Attention(QW_i^Q, KW_i^K, VW_i^V) = softmax\left(\frac{mask + (QK^T)}{\sqrt{d_k}}\right)V \quad (4)$$

Nguồn: (Transformers, n.d.)

3.1.3 Position-wise Encoding (PE)

Mô hình Transformer xử lý các từ song song, do đó, với chỉ Word Embedding mô hình không thể nào biết được vị trí các từ. Như vậy, chúng ta cần một cơ chế nào đó để đưa thông tin vị trí các từ vào trong vector đầu vào. Đó là lúc Positional Encoding xuất hiện và giải quyết vấn đề của chúng ta (phương pháp được đề xuất ở đây là Sinusoidal Position Encoding)

Vị trí của các từ được mã hóa bằng một vector có kích thước bằng Word Embedding và được cộng trực tiếp vào Word Embedding.



Hình 3-5: Minh hoạ phương pháp Sinusoidal Position Encoding kết hợp với Word Embedding

Nguồn: (*Transformer/Transformer.Ipynb at Master · Pbcquoc/Transformer*, n.d.)

Position Encoding tại vị trí pos và chiều thứ i trong mô hình d chiều khi này được tính bởi công thức

$$PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (5)$$

$$PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \quad (6)$$

Nguồn: (Vaswani et al., 2023)

3.1.4 Residual Connection & Layer Normalization (Add & Norm)

Các lớp Add & Norm được thêm vào để giúp ổn định và tăng tốc quá trình huấn luyện, trong đó phần add sử dụng Identity Residual Connection

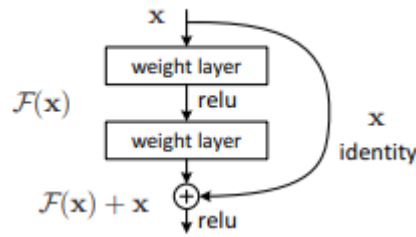
Residual Connection:

- Residual connection (kết nối dư) là một kỹ thuật được sử dụng trong các deep neural networks, đặc biệt phổ biến trong các kiến trúc như ResNet và các biến thể của nó. Kỹ thuật này giúp giảm hiện tượng Vanishing gradient và cho phép xây dựng các mạng neural sâu hơn mà vẫn giữ được hiệu suất.

- Trong một mạng neural thông thường, mỗi layer sẽ có đầu vào x , qua một hàm kích hoạt (ReLU) và các phép biến đổi tuyến tính, sẽ tạo ra đầu ra y
- Residual connection đề xuất thêm đầu vào gốc x trực tiếp vào đầu ra y của layer. Cụ thể, đầu ra của layer được tính như sau:

$$y = \text{ReLU}(x + \text{function}(x)) \quad (7)$$

Nguồn: (He et al., 2015)



Hình 3-6: Minh hoạ Identity Residual Connection

Nguồn: (He et al., 2015)

Layer Normalization

Chúng ta tính toán các thông số normalization cho tất cả đơn vị ẩn (hidden units) trong cùng một lớp như sau:

$$\mu^l = \frac{1}{H} \sum_{i=1}^H a_i^l \quad (8)$$

$$\sigma^l = \sqrt{\frac{1}{H} \sum_{i=1}^H (a_i^l - \mu^l)^2} \quad (9)$$

Nguồn: (Ba et al., 2016)

Trong đó:

- H là số đơn vị ẩn trong một lớp
- a_i^l là các đơn vị ẩn trong một lớp

μ^l và σ^l là giá trị trung bình và độ lệch chuẩn của lớp đó

Các đơn vị ẩn sẽ được tính lại theo công thức Normalization

$$a^l = \frac{a^l - \mu^l}{\sigma^l} \quad (10)$$

3.2 Thuật toán Byte Pair Encoding (BPE) dùng để tokenize dữ liệu câu chữ trong Xử lý ngôn ngữ tự nhiên (NLP – Natural Language Processing)

Trong xử lý ngôn ngữ tự nhiên, việc biểu diễn dữ liệu câu chữ thành dữ liệu số (Word Embedding) để làm việc với các mô hình AI vô cùng quan trọng, vừa phải đảm bảo ngữ nghĩa, vừa phải đảm bảo ngữ cảnh, các nhà khoa học đã nghiên cứu ra nhiều phương pháp khác nhau. Trong đó, các mô hình Transformer đã sử dụng phương pháp Byte Pair Encoding (BPE) để biểu diễn dữ liệu

Theo (Zouhar et al., 2024), kỹ thuật này cho phép phân chia từ thành các đơn vị nhỏ hơn cả từ nhưng lớn hơn một ký tự, gọi là subword. Sử dụng BPE cho phép hầu hết các từ được biểu diễn bởi subword, giảm thiểu số lượng token không biết (unknown) - những token đại diện cho các từ không được nhận biết trước đó. Kỹ thuật này đã nhanh chóng được ưa chuộng và áp dụng rộng rãi trong các mô hình NLP hiện đại. Công nghệ BPE không chỉ cải thiện độ chính xác trong các ứng dụng dịch máy và phân loại văn bản, mà còn trong dự đoán câu tiếp theo, hỏi đáp tự động, phân tích mối quan hệ văn bản

3.2.1 Lợi ích của sử dụng BPE để tách token

Một trong những cách phổ biến nhất để phân tách là tách theo từng từ riêng lẻ. Tuy nhiên, phương pháp này có một nhược điểm lớn liên quan đến các từ không biết (unknown words), hơn nữa, có nhiều từ có các subwords giống nhau. Để giải quyết vấn đề này, chúng ta có thể áp dụng BPE để coi các subwords này như một token để có thể tái sử dụng chúng trong những lần sau

- Ví dụ: Trong tập dữ liệu có những câu sau
 - He likes hiking in the mountains.
 - She enjoys biking along the river.

Nếu chúng ta áp dụng word tokenize, tức là phân chia câu dựa vào khoảng trắng (“ ”) giữa các từ, thì chúng ta sẽ thu được một bộ từ điển (vocabulary) chứa các token sau: {“He”, “likes”, “hiking”, “in”, “the”, “mountains”, “She”, “enjoys”, “biking”, “along”, “river”}

Tuy nhiên, ta có thể thấy các từ “hiking”, và “biking” đều chứa hậu tố (suffix) “-iking”, một nhóm ký tự có thể cung cấp ý nghĩa liên quan đến một hình thức hoạt động hoặc sở thích

Qua đó cho thấy “iking” là một phần có ý nghĩa riêng, ta có thể xem xét xem đây là một token riêng biệt, giúp mô hình học máy nhận ra và xử lý các từ mới có cấu trúc tương tự một cách hiệu quả hơn, ví dụ như “spiking” hoặc “striking” trong các văn bản khác

Bằng cách nhận diện và xử lý “iking” như một token, BPE không chỉ giúp giảm thiểu số lượng từ không biết mà còn tăng cường khả năng của mô hình trong việc hiểu và dự đoán các từ có liên quan mà không cần phải gặp chúng trước đó trong tập huấn luyện

- **Giảm kích thước từ vựng:** Một trong những thách thức chính khi huấn luyện các mô hình Transformer là quản lý kích thước từ vựng, vì kích thước từ vựng lớn yêu cầu nhiều tài nguyên bộ nhớ và tính toán. BPE giải quyết vấn đề này bằng cách phân tách các từ thành các subword phổ biến, giúp cho Transformer giảm số lượng token duy nhất mà nó phải xử lý, làm giảm đáng kể bộ nhớ và tải tính toán, đồng thời vẫn giữ được sự phong phú ngữ nghĩa của ngôn ngữ
- **Xử lý các từ hiếm, từ mới (từ không biết):** Trong các mô hình NLP truyền thống, từ hiếm thường bị bỏ qua hoặc được xử lý kém do thiếu dữ liệu đào tạo. BPE cho phép Transformer xử lý các từ hiếm bằng cách phân rã chúng thành các subwords có thể tái sử dụng, giúp mô hình có thể học và hiểu các từ mới hoặc hiếm gặp mà không cần mở rộng từ vựng

- **Hỗ trợ đa ngôn ngữ:** BPE cung cấp một phương pháp tiêu chuẩn để xử lý nhiều ngôn ngữ, làm cho nó trở thành một công cụ lý tưởng cho các mô hình Transformer đa ngôn ngữ. Với khả năng phân tách ngôn ngữ thành subwords, BPE giúp mô hình dễ dàng học và chuyển giao kiến thức giữa các ngôn ngữ khác nhau
- **Cải thiện hiệu suất mô hình:** Khi sử dụng BPE, các token được biểu diễn dưới dạng các chuỗi subwords, cho phép Transformer học các mẫu ngữ nghĩa phức tạp hơn từ các phần nhỏ của các từ, thay vì chỉ học từ một từ riêng lẻ, cải thiện khả năng tổng quát hóa của mô hình khi nó được áp dụng cho các văn bản mới hoặc các tình huống ngoài tập huấn luyện
- **Tối ưu hóa quá trình huấn luyện:** BPE cũng giúp tối ưu hóa quá trình huấn luyện Transformer bằng cách giảm số lượng token duy nhất cần được xử lý trong mỗi bước, giúp giảm thời gian và chi phí tính toán cho mỗi epoch huấn luyện

3.2.2 Nhược điểm của kỹ thuật BPE

Tuy nhiên BPE có rất nhiều lợi ích nhưng nó cũng có một số nhược điểm:

- Việc giải mã văn bản từ subword sang từ hoàn chỉnh đôi khi có thể không chính xác hoàn toàn, đặc biệt là với các ngôn ngữ có cấu trúc phức tạp
- Ở một số trường hợp, việc phân tách từ thành subword có thể bị mất mát ngữ nghĩa, các subword có thể không còn giữ được ý nghĩa toàn diện của từ gốc
- BPE có thể gặp khó khăn trong việc xử lý các từ ghép hoặc các từ có cấu trúc ngữ âm phức tạp, nơi các phần của từ không thường xuyên xuất hiện như những subwords độc lập
- Phương pháp BPE có thể dẫn đến không đồng đều trong cách biểu diễn các từ, đặc biệt là khi các subwords không được phân bổ đều cho tất cả

các từ trong ngôn ngữ, làm ảnh hưởng đến chất lượng và tính nhất quán của mô hình

3.2.3 Biểu diễn token theo phương pháp BPE

Phương pháp BPE sẽ thống kê tần suất xuất hiện của các từ phụ cùng nhau và tìm cách gộp chúng lại nếu tần suất xuất hiện của chúng là lớn nhất. Cứ tiếp tục quá trình gộp từ phụ cho tới khi không tồn tại các subword để gộp nữa, ta sẽ thu được tập subwords cho toàn bộ bộ văn bản mà mọi từ đều có thể biểu diễn được thông qua subwords.

merge 1	p i c k e d	p i c k l e d	p i c k l e s
merge 2	p i c k e d	p i c k l e d	p i c k l e s
merge 3	p i c k e d	p i c k l e d	p i c k l e s
merge 4	p i c k e d	p i c k l e d	p i c k l e s
merge 5	p i c k e d	p i c k l e d	p i c k l e s
final	p i c k e d	p i c k l e d	p i c k l e s

Hình 3-7: Minh họa thuật toán BPE cho 3 từ “picked”, “pickled”, “pickles”

Nguồn: (Zouhar et al., 2024)

Minh họa ví dụ hình 2-9:

- Ta có ngữ liệu ban đầu:

- + 1 <w> p i c k e d </w>
- + 1 <w> p i c k l e d </w>
- + 1 <w> p i c k l e s </w>

Từ điển hiện tại bao gồm các ký tự: <w>, </w>, p, i, c, k, l, e, d, s

Trong xử lý ngôn ngữ tự nhiên, mỗi từ được lưu trong từ vựng với token <w> (bắt đầu 1 từ) và </w> (kết thúc 1 từ), ví dụ <w>picked</w>

- Vòng lặp đầu tiên:

- + Thuật toán sẽ đếm tất cả những cặp ký tự (không tính <w> và </w> và tìm ra cặp nào xuất hiện nhiều nhất, cặp ký tự xuất hiện nhiều nhất sẽ được ghép lại thành token mới và thêm vào từ điển
- Trong ví dụ này ta có tần suất của các cặp ký tự là: {“pi”: 3, “ic”: 3, “ck”: 3, “ke”: 1, “kl”: 2, “le”: 2, “ed”: 2, “es”: 1}

- + Trong trường hợp này, nếu có nhiều cặp ký tự có tần suất xuất hiện nhiều nhất thì cặp nào được đếm trước sẽ được ghép trước. Trong trường hợp này ta sẽ chọn ghép cặp “pi” (bước merge 1 Hình 2-9)
- + Ngữ liệu khi này ta có:
 - 1 <w> pi c k e d </w>
 - 1 <w> pi c k l e d </w>
 - 1 <w> pi c k l e s </w>
 Từ điển hiện tại bao gồm các ký tự: <w>, <w/>, c, k, l, e, d, s, pi
- **Vòng lặp thứ hai:**
 - + Tần suất của các cặp khi này là: {“ck”: 3, “ke”: 1, “kl”: 2, “le”: 2, “ed”: 2, “es”: 1, “pic”: 3}
 - + Chọn cặp “ck” (bước merge 2 Hình 2-9)
 - + Ngữ liệu khi này ta có:
 - 1 <w> pi ck e d </w>
 - 1 <w> pi ck l e d </w>
 - 1 <w> pi ck l e s </w>
 Từ điển hiện tại bao gồm các ký tự: <w>, <w/>, l, e, d, s, pi, ck
- **Các vòng lặp tiếp theo:**
 - + Quy trình này tiếp tục cho đến khi đạt đến số lần lặp kỳ vọng hoặc không còn cặp nào có thể được hợp nhất thêm. Kết quả cuối cùng là một từ điển với các token được tối ưu hóa để đại diện cho văn bản, giúp giảm số lượng token không xác định và cải thiện khả năng xử lý của các mô hình NLP
 - + Ngoài ra, từ điển sẽ được bố trí dưới dạng [key: value] (key là token và value là số lần lặp lại của token đó)

Tóm lại, thuật toán BPE bao gồm các bước:

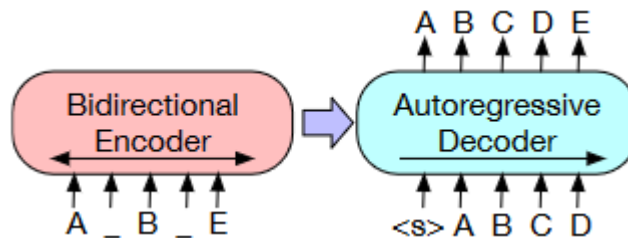
- Bước 1: Khởi tạo bộ từ điển (vocabulary)

- Bước 2: Biểu diễn mỗi từ trong bộ văn bản bằng kết hợp của các ký tự với cặp token <w> và <\w> ở đầu và cuối
- Bước 3: Thống kê số lần xuất hiện theo cặp của toàn bộ token trong từ điển
- Bước 4: Tìm cặp ký tự xuất hiện nhiều nhất và hợp nhất chúng thành một token mới. Thêm token mới này vào từ điển
- Bước 5: Lặp lại bước 3 và bước 4 cho tới đủ số bước hoặc đạt đến kích thước tối đa mà người dùng cho trước

3.3 Mô hình BART

3.3.1 Kiến trúc tổng quát

BART là một autoencoder khử nhiễu trên kiến trúc sequence-to-sequence. Nó sử dụng kiến trúc transformers chuẩn cho bài toán dịch máy. Việc huấn luyện BART bao gồm việc tạo nhiễu trong văn bản với một hàm tùy ý và sử dụng mô hình để tái cấu trúc lại văn bản ban đầu.

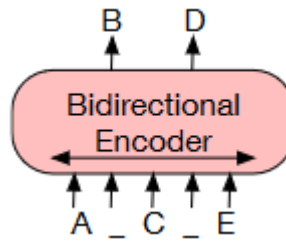


Hình 3-8: Minh hoạ mô hình BART

Nguồn: (Lewis et al., 2019)

Giống như Transformer, BART bao gồm encoder và decoder, nên ta có thể gọi BART là mô hình Transformer based encoder-decoder:

- Encoder: Được lấy từ BERT, dựa vào phần encoder của Transformer (encoder-only), nó có thể mã hóa dữ liệu đầu vào 2 chiều (Multi-head Self-Attention) để hiểu được ngữ cảnh của các từ. Một số lượng ngẫu nhiên các token được đánh dấu riêng biệt và mô hình phải tự khôi phục chúng



Hình 3-9: Phần Encoder, dựa vào kiến trúc của BERT

Nguồn: (Lewis et al., 2019)

- Decoder: Lấy từ GPT, phần decoder của Transformer (decoder-only), chịu trách nhiệm mã hoá 1 chiều (Masked Self-Attention) để sinh văn bản.

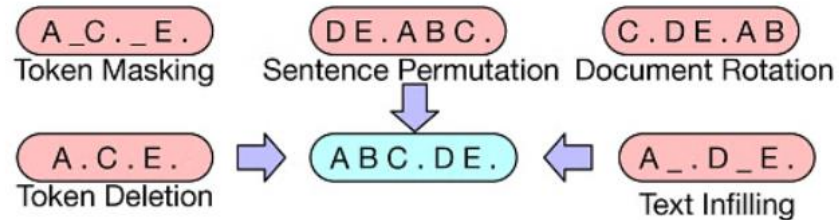
BART sử dụng hàm kích hoạt GeLU thay thế cho RELU. Kiến trúc cơ bản của BART sử dụng 6 lớp cho mỗi phần Encoder và Decoder (tổng 12 lớp), trong khi kiến trúc mở rộng sử dụng 12 lớp cho mỗi phần

3.3.2 Pre-training BART

Theo (Lewis et al., 2019), BART được đào tạo như một bộ mã hóa tự động khử nhiễu, do đó, dữ liệu đào tạo bao gồm văn bản “bị sai” hoặc “nhiều” sẽ được ánh xạ tới văn bản “sạch” hoặc văn bản gốc. Vậy chính xác thì điều gì được coi là nhiễu đối với dữ liệu văn bản? BART quyết định sử dụng một số kỹ thuật tạo nhiễu hiện có và một số kỹ thuật tạo nhiễu mới để huấn luyện trước. Các kỹ thuật đó như là:

- Token Masking: Như BERT, các token được lấy ngẫu nhiên và thay thế bởi [MASK]
- Token Deletion: Các token ngẫu nhiên được xóa khỏi chuỗi đầu vào, mô hình cần đoán được token nào bị xóa
- Text Infilling: Một vài đoạn văn bản ngẫu nhiên được thay thế bằng [MASK]. Đặc biệt, đoạn văn bản có thể là rỗng
- Document Rotation: Một token được chọn ngẫu nhiên, văn bản được xoay để bắt đầu với token đó. Điều này giúp cho mô hình học được đầu và điểm bắt đầu của văn bản

- Sentence Permutation: Văn bản được chia thành các câu và được tráo ngẫu nhiên



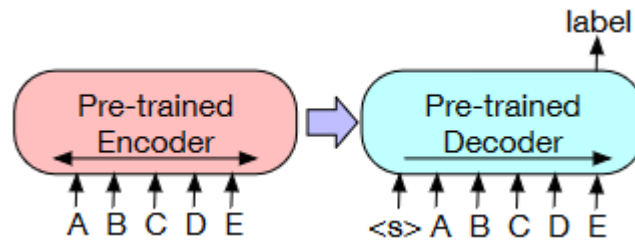
Hình 3-10: Các kỹ thuật làm nhiễu dữ liệu đầu vào (phần màu đỏ)

Nguồn: (Lewis et al., 2019)

3.3.3 Fine-tuning BART

(Lewis et al., 2019) liệt kê ra các bài toán (task) có thể dùng BART để huấn luyện:

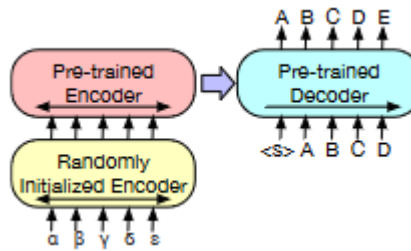
- **Sequence Classification Tasks:** Đối với các tác vụ phân loại trình tự, cùng một đầu vào được đưa vào Encoder và Decoder, đồng thời final hidden state của final decoder token được đưa vào multi-class linear classifier. Cách tiếp cận này có liên quan đến CLS token trong BERT; tuy nhiên, chúng tôi thêm token bổ sung vào cuối để biểu diễn token trong Decoder có thể tham gia vào trạng thái Decoder vào hoàn chỉnh
- **Token Classification Tasks:**
 - + Đối với các tác vụ phân loại token, như phân loại điểm cuối câu trả lời cho SQuAD (Stanford Question Answering Dataset), chúng ta đưa toàn bộ tài liệu vào bộ encoder và bộ decoder, và sử dụng hidden state cuối cùng của bộ decoder làm biểu diễn cho mỗi từ. Biểu diễn này được sử dụng để phân loại token



Hình 3-11: Token Classification

Nguồn: (Lewis et al., 2019)

- + Cách tiếp cận này cho phép mô hình tận dụng được thông tin ngữ cảnh toàn cục từ toàn bộ dữ liệu đầu vào khi biểu diễn từng token. Trạng thái ẩn cuối cùng của bộ decoder bao gồm thông tin về vị trí của từ đó trong ngữ cảnh tài liệu, giúp cải thiện hiệu quả của tác vụ phân loại token
- **Sequence Generation Tasks:** Bởi vì BART có một bộ giải mã tự hồi quy, nó có thể được fine-tune trực tiếp cho các tác vụ sinh chuỗi như trả lời câu hỏi tóm lược và tóm tắt trừu tượng. Trong cả hai tác vụ này, thông tin được sao chép từ đầu vào nhưng được thao tác, điều này có liên quan chặt chẽ với mục tiêu tiền huấn luyện khởi đầu. Ở đây, đầu vào của bộ encoder là chuỗi đầu vào, và bộ decoder tạo ra các đầu ra một cách tự hồi quy
- **Machine Translation:**
 - + Các mô hình có thể được cải thiện bằng cách kết hợp các bộ encoder đã được tiền huấn luyện, nhưng lợi ích từ việc sử dụng các mô hình ngôn ngữ đã được tiền huấn luyện trong bộ giải mã còn hạn chế. Có thể sử dụng toàn bộ mô hình BART (cả bộ encoder và bộ decoder) như một bộ giải mã đã được tiền huấn luyện (Pre-trained) cho dịch máy, bằng cách thêm một tập mới các tham số encoder được học từ dữ liệu song ngữ



Hình 3-12: Machine Translation

Nguồn: (Lewis et al., 2019)

- + Cụ thể hơn, thay thế lớp embedding của bộ encoder BART bằng một bộ encoder mới được khởi tạo ngẫu nhiên. Mô hình được huấn luyện end-to-end, điều này huấn luyện bộ encoder mới để ánh xạ các từ nước ngoài thành một đầu vào mà BART có thể xử lý nhiều thành tiếng Anh. Bộ encoder mới có thể sử dụng một vốn từ riêng biệt so với mô hình BART gốc
- + Huấn luyện bộ encoder nguồn trong hai bước, trong cả hai trường hợp đều lan truyền ngược tổn thất entropy chéo từ đầu ra của mô hình BART. Trong bước đầu tiên, chúng tôi đóng băng hầu hết các tham số của BART và chỉ cập nhật bộ encoder nguồn được khởi tạo ngẫu nhiên, embedding vị trí của BART và ma trận đầu vào Self-Attention của lớp đầu tiên của bộ encoder BART. Trong bước thứ hai, chúng tôi huấn luyện tất cả các tham số của mô hình trong một số lượng lặp lại nhỏ

3.3.4 Ưu và nhược điểm

Ưu điểm:

- Kết hợp được ưu điểm của BERT (mã hoá 2 chiều) và GPT (sinh văn bản chất lượng cao)
- Huấn luyện mạnh mẽ với kỹ thuật noise denoising: BART học cách khôi phục văn bản từ dữ liệu bị nhiễu (mask, shuffle, delete), giúp mô hình hiểu sâu hơn về ngữ nghĩa và cấu trúc câu

- Hiệu quả cao trong nhiều bài toán NLP: tóm tắt văn bản, dịch máy, hỏi đáp tự động
- Hỗ trợ tốt trên thư viện của Hugging Face: dễ dàng fine-tune với api transformers

Nhược điểm:

- Tốn tài nguyên tính toán do có cả Encoder và Decoder
- Khó mở rộng cho các bài toán trong thời gian thực: hỏi đáp nhanh như GPT
- Chưa tối ưu cho tất cả các bài toán sinh văn bản: với tóm tắt văn bản cực ngắn, T5 có thể hoạt động hiệu quả hơn do được huấn luyện chuyên biệt cho seq2seq

CHƯƠNG 4. THỰC NGHIỆM

4.1 Dữ liệu thực nghiệm

Sơ lược về dữ liệu thực nghiệm: Bộ dữ liệu DailyDialog của li2017dailydialog từ hugging face được chia làm 3 tập con

Bảng 4-1: Số mẫu dữ liệu đã được phân chia trong các tập dữ liệu con của bộ dữ liệu DailyDialog

Nguồn: (*Li2017dailydialog/Daily_dialog · Datasets at Hugging Face*, 2023)

Tên	train	validation	test
Số samples	11118	1000	1000

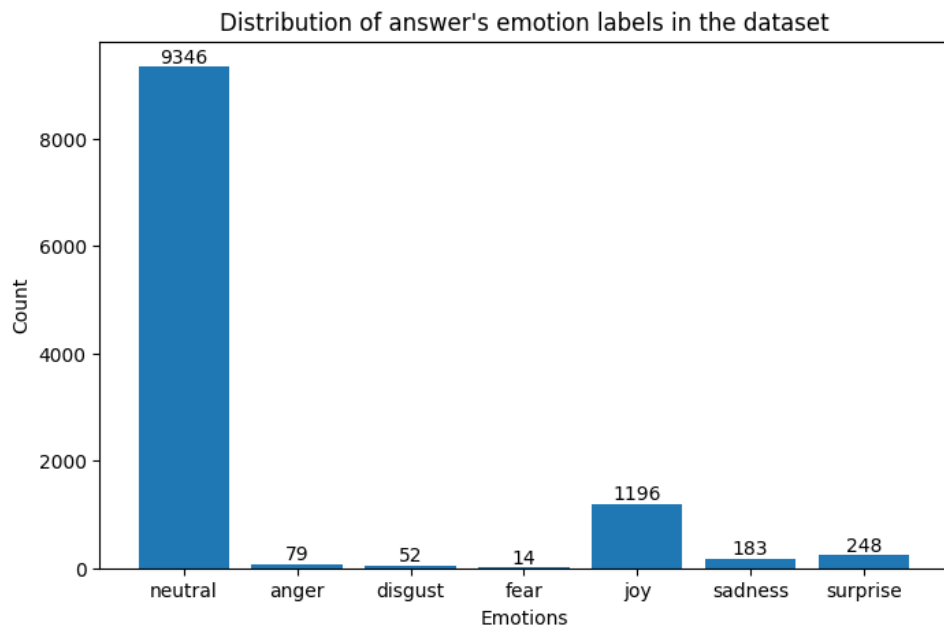
Bộ dữ liệu DailyDialog gồm 3 thuộc tính:

- dialog: list các câu của hội thoại, trong bài này chúng tôi sẽ sử dụng 2 câu đầu tiên làm “question” và “answer”
- act: list các phân loại của các câu, bao gồm: **không phân loại (0)**, **inform (1)**, **question (2)**, **directive (3)** and **commissive (4)**

- emotion: list các phân loại cảm xúc, bao gồm: **neutral (0)**, **anger (1)**, **disgust (2)**, **fear (3)**, **joy (4)**, **sadness (5)** and **surprise (6)**. Các cảm xúc cơ bản này được phân theo Paul Ekman (1972)

Trực quan hoá dữ liệu:

- Chúng tôi trực quan hoá các phân loại cảm xúc chính của các câu sẽ dùng để làm “answer” cho huấn luyện mô hình



Hình 4-1: Biểu đồ phân bố các nhãn cảm xúc chính của “answer” trong dữ liệu train

Xử lý thêm cảm xúc vào dữ liệu:

- Chúng tôi sẽ sử dụng phương pháp thêm token cảm xúc (gắn nhãn cảm xúc với trọng số từ 0 đến 100) vào dữ liệu đầu vào
- Lấy cột “dialog”, chọn câu đầu tiên làm “question”, câu thứ hai làm “answer”, chỉ sử dụng 2 cột “question” và “answer” cho dataset
- Tải mô hình phân loại cảm xúc distil roberta từ hugging face, chạy mô hình dự đoán cảm xúc cho “answer”. Tại sao phải lấy cảm xúc của “answer” mà không phải cảm xúc của “question”? Vì cảm xúc được thêm vào dùng để dự đoán cho “answer” khi huấn luyện mô hình, nếu

dùng cảm xúc của “question” nó sẽ gây ra sai lệch vì khi này cảm xúc là của câu “question”

Output:

```
[[{'label': 'anger', 'score': 0.004419783595949411},
  {'label': 'disgust', 'score': 0.0016119900392368436},
  {'label': 'fear', 'score': 0.0004138521908316761},
  {'label': 'joy', 'score': 0.9771687984466553},
  {'label': 'neutral', 'score': 0.005764586851000786},
  {'label': 'sadness', 'score': 0.002092392183840275},
  {'label': 'surprise', 'score': 0.008528684265911579}]]
```

Hình 4-2: Kết quả dự đoán cảm xúc của mô hình distil roberta cho câu “I love this!”

Nguồn: (*J-Hartmann/Emotion-English-Distilroberta-Base · Hugging Face*, n.d.)

- Score sau khi dự đoán chúng tôi nhân với 100 và làm tròn để các giá trị nằm trong khoảng số nguyên từ 0 đến 100. Sau đó ghép các “label” và “score” đã chuẩn hoá về [0..100] lại theo định dạng “<label1>:score1 <label2>:score2...”
- Chuỗi sau khi được tạo ra được nối vào “question”, khi này question có dạng “<label1>:score1 ... <label7>:score7 câu gốc”

Tokenize dữ liệu và lấy các cột cần thiết:

- Tại mỗi dataset sau khi thêm cảm xúc, sử dụng thư viện hugging face “transformers.BartTokenizer” load bộ tokenizer của mô hình “facebook/bart-base” để thực hiện word embedding cho 2 cột “question” và “answer”, mỗi cột sau khi tokenize sẽ được một dictionary gồm 2 list là: “input_ids”, “attention_mask”, ta có dictionary_question và dictionary_answer. Tiếp theo, gán “input_ids” của dictionary_answer làm “labels” của dictionary_question và tiến hành lưu lại
- Dữ liệu sau khi xử lý để đưa vào mô hình gồm các cột “input_ids”, “attention_mask”, “labels” (dictionary_question)

4.2 Cài đặt thực nghiệm

Môi trường thực nghiệm: Google Colab bản free với T4 GPU

Các thư viện sử dụng:

- transformers: đây là thư viện của hugging face, hỗ trợ tải mô hình gốc, các mô hình đã Pre-train, cung cấp các thư viện để fine-tune các mô hình Transformer
- datasets: đây cũng là thư viện của hugging face, có thể dùng để tải bộ dữ liệu được lưu trữ sẵn trên hugging face mà không cần tải cục bộ về máy

(*Hugging Face - Documentation*, n.d.)


Load bộ dữ liệu daily_dialog bằng “datasets.load_dataset” của hugging face, lấy 3 bộ dữ liệu train, validation, test

Thực hiện Xử lý dữ liệu ở phần trước để thu được các cột “input_ids”, “attention_mask”, “labels”

Load mô hình “facebook/bart-base” bằng thư viện hugging face “transformers.BartForConditionalGeneration” để sử dụng mô hình BART

Như đã nói ở phần trước, các cảm xúc được thêm vào dữ liệu gốc có dạng “<label>:score”, các <label> này ta xem như là những token. Tuy nhiên BartTokenizer gốc không có các token về cảm xúc, nên ta phải thêm thủ công các token về cảm xúc bằng “BartTokenizer.add_tokens”. Sau đó điều chỉnh kích thước token embedding của mô hình BART để không gây xung đột token bằng “BartForConditionalGeneration.resize_token_embeddings” với số lượng token của BartTokenizer sau khi thêm các token cảm xúc. Khi này mô hình đã sẵn sàng huấn luyện

Cấu hình huấn luyện bằng cách điều chỉnh các tham số của các đối tượng “transformers.Seq2SeqTrainingArguments”, “transformers.Seq2SeqTrainer”, rồi gọi hàm “Seq2SeqTrainer.train” để bắt đầu huấn luyện



Epoch	Training Loss	Validation Loss
1	1.044600	0.316959
2	0.318700	0.301803
3	0.262400	0.293558
4	0.238800	0.290480
5	0.225900	0.290588

Hình 4-3: Đánh giá các losses sau khi huấn luyện với 5 epoch

Dùng hàm “Seq2SeqTrainer.evaluate” để đánh giá mô hình trên tập validation, “Seq2SeqTrainer.save_model” để lưu lại mô hình sử dụng cho lần sau

Đánh giá với các mẫu dữ liệu, vì đây là bài toán về cảm xúc AI, nên không thể sử dụng các độ đo của ngôn ngữ tự nhiên như BLEU, ROUGE. Chúng tôi sử dụng “BartForConditionalGeneration.generate” để tạo ra câu trả lời, và sử dụng mô hình phân loại cảm xúc để đánh giá chỉ số cảm xúc của kết quả.

4.3 Kết quả thực nghiệm

Chúng tôi cũng huấn luyện một mô hình BART với cùng bộ dữ liệu nhưng không thêm yếu tố cảm xúc, sau đó so sánh kết quả với các mẫu dữ liệu

Vì hạn chế về phần cứng và thời gian nên chúng tôi chỉ có thể huấn luyện mô hình BART có cảm xúc và BART không có cảm xúc với 5 epoch. Mặc dù vậy, kết quả quan sát được cho thấy ta có thể thêm cảm xúc vào mô hình AI bằng phương pháp thêm token cảm xúc (gắn nhãn cảm xúc với trọng số từ 0 đến 100)

```

=====
You got an reward!
<surprise>:100

** Without emotion
Thank you .
{'label': 'neutral', 'score': 0.5301764607429504}

** With emotion
Really ? I don't know what to do with it .
{'label': 'surprise', 'score': 0.8196104168891907}
=====
You got an reward!
<fear>:70, <anger>:50

** Without emotion
Thank you .
{'label': 'neutral', 'score': 0.5301764607429504}

** With emotion
You did ?
{'label': 'neutral', 'score': 0.5270916223526001}

```

Hình 4-4: So sánh với câu “You got an reward!”, gán các nhãn cảm xúc khác nhau

```

=====
I don't know the reason.
<sadness>:40 <surprise>:25

** Without emotion
What's the matter ?
{'label': 'fear', 'score': 0.4859614372253418}

** With emotion
I don't know .
{'label': 'fear', 'score': 0.5663102269172668}

```

Hình 4-5: So sánh với câu “I don’t know the reason.”

```

=====
What do you think?
<joy>:90

** Without emotion
  I don't think so .
{'label': 'neutral', 'score': 0.45655557513237}

** With emotion
  I don't think so .
{'label': 'neutral', 'score': 0.45655557513237}
=====
What do you think?
<fear>:60

** Without emotion
  I don't think so .
{'label': 'neutral', 'score': 0.45655557513237}

** With emotion
  I don't think so .
{'label': 'neutral', 'score': 0.45655557513237}

```

Hình 4-6: So sánh với câu “What do you think?”

Kết quả cho thấy, đối với câu thông báo (“You got an reward!”), ta có thể điều chỉnh các nhãn cảm xúc và các trọng số để thay đổi câu trả lời. Tuy nhiên đối với các câu nêu lên cảm nghĩ (“I don’t know the reason.”), câu hỏi (“What do you think?”), câu trả lời lại khó có thể điều chỉnh

CHƯƠNG 5. KẾT LUẬN

5.1 Kết luận

Từ việc nghiên cứu, thực nghiệm, ta có thể thấy cảm xúc là yếu tố phức hợp, phức tạp, như Plutchik đã nói rằng, cảm xúc có thể pha trộn như màu sắc theo những tỉ lệ khác nhau, điều này đã được kiểm chứng qua kết quả thực nghiệm với ngôn ngữ tự nhiên. Tuy vậy việc pha trộn cảm xúc theo tỉ lệ trong mô hình AI lại bị giới hạn bởi những cảm xúc mà mô hình được học, tức là nó chỉ phản hồi ra câu trả lời mang những yếu tố cảm xúc trong phạm vi đã được học bởi mô hình, việc điều chỉnh thêm các yếu tố cảm xúc mới tốn rất nhiều thời gian, có thể làm mất đi tri thức vốn có mô hình.

Các phương pháp đưa cảm xúc vào AI đòi hỏi phải có lượng kiến thức rất lớn về thuật toán, hiểu biết về mô hình cần sử dụng, phụ thuộc nhiều vào loại dữ liệu, nhất là cần rất nhiều thời gian và phần cứng. Các mô hình AI hiện nay có kiến trúc rất lớn, rất sâu, người ta không thể đánh giá thủ công như những giả thuật thông thường, chỉ có thể huấn luyện (đây là nút thắt cổ chai) với bộ tham số nhất định, sau đó đánh giá dựa trên các tiêu chí được đặt ra, tạo thành vòng lặp điều chỉnh tham số, huấn luyện, đánh giá, cho đến khi mô hình đạt được hiệu suất mong muốn.

Đối với ngôn ngữ tự nhiên, không phải câu trả lời nào cũng có thể dễ dàng điều chỉnh yếu tố cảm xúc bởi vì mô hình AI không thực sự hiểu các cảm xúc là gì, chỉ có thể tính toán cảm phản hồi xúc thông qua các quy ước, quy tắc, thuật toán do con người đặt ra, giống như những người bị Psychopathy, họ không hiểu được cảm xúc nhưng lại có thể biểu hiện một cảm xúc. Một điểm yếu chung của AI đó chính là phụ thuộc vào bộ dữ liệu huấn luyện, trong những năm gần đây có rất nhiều công trình nghiên cứu và công cụ hỗ trợ để dễ dàng fine-tune mô hình theo mục đích riêng, những người mang ý tưởng lệch lạc có thể tạo ra các bộ dataset mang tính chất thao túng để tạo ra các mô hình AI tuyên truyền các ý tưởng lệch lạc qua đó thao túng tâm lý người khác. Vì vậy, những nguy hiểm tiềm tàng vẫn tồn tại.

Tuy vậy phải nhiều nguy cơ về mặt đạo đức, vẫn không thể phủ nhận được rằng việc thêm các yếu tố cảm xúc vào AI giúp cho xã hội, các ngành nghề (dịch vụ, bán lẻ, giáo dục, y tế, ...) hoạt động hiệu quả hơn rất nhiều thông qua các ứng dụng thực tế đã nêu ở Cơ sở lý thuyết. Không chỉ giúp ích về mặt tâm lý, các hệ thống AI cảm xúc còn giúp cộng tác với các hệ thống nhận diện, tư vấn, khuyến nghị, giúp tăng hiệu quả công việc, tăng độ yêu thích của khách hàng, từ đó giúp các doanh nghiệp tăng độ uy tín, tăng doanh thu, mở ra nhiều loại hình kinh doanh khác có sự trợ giúp của AI cảm xúc.

Cảm xúc AI đòi hỏi phải có thời gian đủ dài để nghiên cứu chuyên sâu các mặt lợi và hại, cả những vấn đề đạo đức đi kèm (có khả năng mô hình sẽ giống những người bị Psychopath nếu thực hiện không đúng cách), những ứng dụng và cải tiến trong tương lai.

Việc huấn luyện mô hình cảm xúc AI với ngôn ngữ tự nhiên đòi hỏi nhiều kiến thức về xử lý ngôn ngữ tự nhiên (NLP), cũng như phần cứng đủ mạnh, thời gian đủ dài (nhiều epoch nhưng không bị overfitting) để mô hình có thể hiểu rõ được ngữ cảnh (trong trường hợp này là cảm xúc).

5.2 Hướng phát triển

Thực nghiệm huấn luyện mô hình với nhiều số epochs hơn để đánh giá thêm khả năng học của mô hình. Các phương pháp đưa yếu tố cảm xúc khác vào mô hình AI, các mô hình AI khác như T5, Bartpho, ... trong đó ưu tiên các mô hình phản hồi dữ liệu có cảm xúc theo thời gian thực. Với dữ liệu tiếng Việt, đây là dữ liệu khó xử lý hơn tiếng Anh rất nhiều vì liên quan đến các từ đồng nghĩa, từ nhiều nghĩa, ngữ cảnh, sắc thái cảm xúc. Ngoài ra còn thực nghiệm với các loại dữ liệu khác (hình ảnh, video)

Ngoài việc thực nghiệm, chúng tôi muốn nghiên cứu mở rộng thêm về các vấn đề, đặt ra thêm nhiều câu hỏi nghiên cứu liên quan đến cảm xúc AI, chẳng hạn như nếu AI có thể "giả vờ" có cảm xúc, có nên sử dụng nó trong các mối quan hệ con người không?

TÀI LIỆU THAM KHẢO

Tiếng Việt

...

Tiếng Anh

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization*

(arXiv:1607.06450). arXiv. <https://doi.org/10.48550/arXiv.1607.06450>

Daily, S. B., James, M. T., Cherry, D., J. Porter, J., Darnell, S. S., Isaac, J., & Roy,

T. (2017). Affective Computing: Historical Foundations, Current

Applications, and Future Trends. In *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 213–231). Elsevier.

<https://doi.org/10.1016/B978-0-12-801851-4.00009-4>

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image*

Recognition (arXiv:1512.03385). arXiv.

<https://doi.org/10.48550/arXiv.1512.03385>

Hugging Face—Documentation. (n.d.). Retrieved February 10, 2025, from

<https://huggingface.co/docs>

J-hartmann/emotion-english-distilroberta-base · Hugging Face. (n.d.). Retrieved

February 10, 2025, from <https://huggingface.co/j-hartmann/emotion-english-distilroberta-base>

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O.,

Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-*

Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arXiv:1910.13461). arXiv.

<https://doi.org/10.48550/arXiv.1910.13461>

Li2017dailydialog/daily_dialog · Datasets at Hugging Face. (2023, October 7).

https://huggingface.co/datasets/li2017dailydialog/daily_dialog

Malatesta, L., Karpouzis, K., & Raouzaïou, A. (2009). Affective Intelligence: The

Human Face of AI. In M. Bramer (Ed.), *Artificial Intelligence An*

International Perspective (Vol. 5640, pp. 53–70). Springer Berlin

Heidelberg. https://doi.org/10.1007/978-3-642-03226-4_4

Martínez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial

intelligence. *Computers in Human Behavior*, 21(2), 323–341.

<https://doi.org/10.1016/j.chb.2004.02.010>

Patulny, R., Lazarevic, N., & Smith, V. (2020). ‘Once more, with feeling,’ said the

robot: AI, the end of work and the rise of emotional economies. *Emotions*

and Society, 2(1), 79–97.

<https://doi.org/10.1332/263168919X15750193136130>

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep

evolutionary roots, a fact that may explain their complexity and provide tools

for clinical practice. *American Scientist*, 89(4), 344–350.

Transformers: The Nuts and Bolts. (n.d.). Retrieved February 9, 2025, from

<https://www.revistek.com/posts/transformer-architecture>

Transformer/transformer.ipynb at master · pbcquoc/transformer. (n.d.). GitHub.

Retrieved February 9, 2025, from

<https://github.com/pbcquoc/transformer/blob/master/transformer.ipynb>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,

Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need*

(arXiv:1706.03762). arXiv. <https://doi.org/10.48550/arXiv.1706.03762>

Zouhar, V., Meister, C., Gastaldi, J. L., Du, L., Vieira, T., Sachan, M., & Cotterell,

R. (2024). *A Formal Perspective on Byte-Pair Encoding*

(arXiv:2306.16837). arXiv. <https://doi.org/10.48550/arXiv.2306.16837>