

ĐẠI HỌC CÔNG NGHỆ
ĐẠI HỌC QUỐC GIA HÀ NỘI



Cao Đinh Hoàng Minh

**EXPLAINABLE AI: INTEGRATED
GRADIENTS CHO BÀI TOÁN
PHÂN TÍCH QUAN ĐIỂM**

KHÓA LUẬN TỐT NGHIỆP

Chuyên ngành: Khoa học máy tính chương trình chất lượng cao

HÀ NỘI - 2023

ĐẠI HỌC CÔNG NGHỆ
ĐẠI HỌC QUỐC GIA HÀ NỘI

Cao Đinh Hoàng Minh

EXPLAINABLE AI: INTEGRATED
GRADIENTS CHO BÀI TOÁN
PHÂN TÍCH QUAN ĐIỂM

KHÓA LUẬN TỐT NGHIỆP

Chuyên ngành: Khoa học máy tính chương trình chất lượng cao

Cán bộ hướng dẫn: TS. Nguyễn Văn Vinh

HÀ NỘI - 2023

LỜI CẢM ƠN

Trước hết, tôi xin gửi lời cảm ơn chân thành và sâu sắc nhất đến TS. Nguyễn Văn Vinh, người đã hết lòng hướng dẫn, định hướng và động viên tôi trong suốt quá trình thực hiện khóa luận tốt nghiệp này. Những kiến thức, kinh nghiệm và tinh thần trách nhiệm của thầy đã giúp tôi hoàn thành khóa luận một cách tốt nhất.

Tôi cũng muốn gửi lời cảm ơn đến Ban Giám hiệu và các Thầy/Cô giáo ở chuyên ngành Khoa học Máy tính, đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập và nghiên cứu tại trường.

Tôi cũng xin gửi lời cảm ơn tới anh Nguyễn Hoàng Minh Công trong nhóm nghiên cứu, cũng như bạn bè và đồng nghiệp đã giúp đỡ, chia sẻ kiến thức và kinh nghiệm trong quá trình thực hiện khóa luận.

Đặc biệt, tôi xin gửi lời cảm ơn tới gia đình, bố mẹ và người thân đã luôn tin tưởng, ủng hộ và động viên tôi trong suốt quá trình học tập và thực hiện khóa luận. Tình yêu, sự quan tâm và động viên của họ đã giúp tôi vượt qua những khó khăn, thử thách trong quá trình này.

Cuối cùng, tôi xin chân thành cảm ơn tất cả những người đã giúp đỡ, hỗ trợ tôi trong quá trình thực hiện khóa luận tốt nghiệp. Tôi rất trân trọng những đóng góp, ý kiến và sự giúp đỡ của mọi người, và xin được giữ lại trong lòng những ơn nghĩa này.

Xin chân thành cảm ơn!

LỜI CAM ĐOAN

Với mục đích học tập, nghiên cứu để nâng cao kiến thức và trình độ chuyên môn nên tôi đã làm khóa luận này một cách nghiêm túc và hoàn toàn trung thực dưới sự hướng dẫn của giảng viên TS. Nguyễn Văn Vinh. Tôi xin cam đoan nội dung khóa luận không sao chép từ các tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ trong tài liệu tham khảo ở cuối khóa luận. Nếu bị phát hiện có bất kỳ sự gian lận, sao chép nào, tôi xin hoàn toàn chịu trách nhiệm trước quy định của Trường Đại học Công Nghệ - Đại học Quốc gia Hà Nội.

Hà Nội, ngày 26 tháng 5 năm 2023
Sinh viên

TÓM TẮT

TÓM TẮT: Phân tích quan điểm (Sentiment Analysis) là một lĩnh vực của xử lý ngôn ngữ tự nhiên, nhằm phân loại tự động những quan điểm thể hiện qua dữ liệu văn bản, như đánh giá, bài đăng trên các mạng xã hội và bài báo. Mặc dù các mô hình học sâu đã đạt được thành công nhất định trong phân tích quan điểm, việc giải thích những mô hình đó một cách đáng tin cậy vẫn còn là một điều bí ẩn. Integrated Gradients (tạm dịch: Tích hợp độ dốc) là một kỹ thuật tiềm năng để giải thích các dự đoán của mạng nơ-ron, giúp giải quyết những vấn đề này. Khóa luận này tập trung nghiên cứu và áp dụng phương pháp Integrated Gradients vào bài toán phân tích quan điểm, nhằm đưa ra lời giải thích cho quá trình học máy dựa trên trí tuệ nhân tạo giải thích được (Explainable AI). Mục tiêu chính của nghiên cứu là cung cấp một công cụ giúp người dùng hiểu rõ hơn về cách thức hoạt động của các mô hình học máy trong bài toán này, đồng thời đưa ra những đề xuất để cải thiện hiệu quả và độ chính xác của mô hình.

Từ khóa: *Explainable AI, Integrated Gradients, Sentiment Analysis.*

Mục lục

LỜI CẢM ƠN	i
LỜI CAM ĐOAN	ii
TÓM TẮT	iii
CHƯƠNG 1. MỞ ĐẦU	1
1.1 Đặt vấn đề	1
1.2 Mục đích nghiên cứu	2
1.3 Đối tượng nghiên cứu	2
1.4 Đóng góp của khóa luận	3
1.5 Cấu trúc khóa luận	4
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	5
2.1 Phân tích cảm xúc	5
2.1.1 Khái niệm	5
2.1.2 Các cấp độ của phân tích cảm xúc	6
2.1.3 Nhiệm vụ của phân tích cảm xúc	7
2.1.4 Phương pháp phân loại cảm xúc	11
2.2 Một vài mô hình học máy sử dụng trong bài toán phân tích cảm xúc	14
2.2.1 BERT	14
2.2.2 DistilBERT	20
2.3 Trí tuệ nhân tạo giải thích được	22
2.3.1 Giới thiệu về trí tuệ nhân tạo giải thích được	22
2.3.2 Tầm quan trọng của trí tuệ nhân tạo giải thích được	22
2.3.3 Phân loại	23
2.3.4 Một số kỹ thuật để giải thích mô hình trí tuệ nhân tạo	25
2.3.5 Thách thức và hạn chế của trí tuệ nhân tạo giải thích được	30
2.3.6 Tương lai của trí tuệ nhân tạo giải thích được	30
2.4 Integrated Gradient	31
2.4.1 Định nghĩa	31
2.4.2 Biểu thức đang tính gì	32

2.4.3 Nguồn gốc và vai trò của điểm dữ liệu cơ sở	33
2.4.4 Một số cách chọn điểm dữ liệu cơ sở	34
2.4.5 Ưu điểm và nhược điểm của Integrated Gradient	37
CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT	38
3.1 Xây dựng mô hình BERT	38
3.1.1 Tập dữ liệu	39
3.1.2 Bộ mã hóa từ	39
3.1.3 Huấn luyện mô hình	40
3.2 Xây dựng mô hình DistilBERT	41
3.2.1 Tập dữ liệu	42
3.2.2 Huấn luyện mô hình	43
3.3 Kiểm thử và giải thích mô hình	43
3.3.1 Kiểm thử	43
3.3.2 Giải thích mô hình	44
CHƯƠNG 4. ỨNG DỤNG VÀO BÀI TOÁN PHÂN TÍCH QUAN ĐIỂM	47
4.1 Giới thiệu bài toán	47
4.2 Kết quả thực nghiệm	48
4.3 Phân tích và đánh giá khả năng giải thích của Integrated gradients	50
4.3.1 BERT	50
4.3.2 DistilBERT	52
4.4 Ứng dụng Integrated Gradients vào thực tế	54
CHƯƠNG 5. KẾT LUẬN	57

Danh sách hình vẽ

2.1 Các mức độ phân tích cảm xúc	6
2.2 Các nhiệm vụ của phân tích cảm xúc	8
2.3 Cách BERT nhúng đầu vào	15
2.4 Giải thích công thức của IG	32
2.5 Trực quan hóa Maximum baseline	35
2.6 Trực quan hóa Blur baseline	36
2.7 Trực quan hóa Uniform baseline	36
2.8 Trực quan hóa Gaussian baseline	37
3.9 Chỉ số của BERT sau khi huấn luyện	41
3.10 Ví dụ về cách khởi tạo của Integrated Gradients	45
3.11 Gọi hàm explain của Integrated Gradients để giải thích mô hình	46
4.12 Minh họa đầu ra của Integrated Gradients	47
4.13 IG giải thích BERT-base	48
4.14 IG giải thích DistilBERT	49
4.15 Lấy top 5 từ cho mức độ liên quan và không liên quan đến đầu ra	55
4.16 Integrated Gradients giải thích một câu dài hơn	55
4.17 Tập hợp các từ có ảnh hưởng đến đầu ra (điểm số không âm)	56
4.18 Tập hợp các từ không liên quan đến đầu ra (điểm số âm)	56

Bảng các từ viết tắt

Từ viết tắt	Giải nghĩa tiếng Anh	Giải nghĩa tiếng Việt (tạm dịch)
AI	Artificial Intelligence	Trí tuệ nhân tạo
xAI, XAI	eXplainable Artificial Intelligence	Trí tuệ nhân tạo giải thích được
SA	Sentiment Analysis	Phân tích cảm xúc
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
IG	Integrated Gradients	Tích hợp độ dốc
BERT	Bi-directional Encoder Representations from Transformers	Biểu diễn Thẻ hiện Mã hóa Hai chiều từ Transformer
DistilBERT	Distilled version of BERT	Phiên bản rút gọn của BERT
LIME	Local Interpretable Model-agnostic Explanation	Giải thích cục bộ về mô hình bất khả tri giải thích được
SHAP	SHapley Additive exPlanations	Giải thích giá trị thêm vào của Shapley

CHƯƠNG 1. MỞ ĐẦU

1.1 Đặt vấn đề

Trong thời đại công nghệ thông tin phát triển mạnh mẽ như hiện nay, trí tuệ nhân tạo (Artificial Intelligence - AI) và học máy (Machine Learning - ML) đã trở thành những công cụ quan trọng trong việc phân tích và xử lý dữ liệu lớn. Tuy nhiên, một nhược điểm lớn của các mô hình học máy đó là khả năng giải thích vì sao có thể đưa ra được kết quả đó [47]. AI giải thích được (XAI) ra đời nhằm giải quyết những khó khăn này trong việc phân tích cảm xúc. XAI là một lĩnh vực cung cấp các phương pháp để giải thích quá trình ra quyết định của các mô hình học máy phức tạp. Việc giải thích và minh bạch hóa các mô hình dự đoán đóng vai trò quan trọng trong việc tạo niềm tin cho người dùng và những người quản lý. Một phương pháp đáng chú ý trong việc giải thích các mô hình học máy chính là Integrated Gradient.

Integrated Gradient (IG) được sử dụng để giải thích quá trình ra quyết định của các mô hình dự đoán phức tạp, đặc biệt là các mạng nơ-ron, bằng cách tính toán đạo hàm của đầu vào và sự thay đổi của chúng để làm nổi bật tầm quan trọng hoặc mức độ gây nhiễu của các đặc trưng đầu vào. Bằng cách giúp hiểu rõ hơn về cách thức hoạt động của mô hình dự đoán, IG không chỉ tăng cường tính minh bạch mà còn đáp ứng nhu cầu của xã hội hiện đại trong việc đưa ra những quyết định dựa trên dữ liệu và phân tích chuyên sâu.

Trong phạm vi khóa luận này, tôi chỉ áp dụng IG trong các bài toán về phân tích cảm xúc, là bài toán về việc tự động xác định cảm xúc tích cực, tiêu cực, hay trung lập của người dùng dựa trên những dữ liệu văn bản này [42]. phân tích cảm xúc có ứng dụng rộng rãi trong các lĩnh vực như tiếp thị, quảng cáo, dự báo thị trường, hỗ trợ quyết định của doanh nghiệp, phân tích xã hội và nhiều hơn nữa. Với sự phát triển của các mô hình học sâu (Deep Learning), khả năng phân tích cảm xúc đã được cải thiện đáng kể, mang lại kết quả chính xác hơn và ứng dụng

rộng rãi hơn [41].

Để làm rõ cách thức hoạt động của các mô hình phân tích cảm xúc, IG đánh giá đóng góp của từng từ trong câu đối với xu hướng cảm xúc được dự đoán, xem từ nào là nguyên nhân chính dẫn đến một nhận định là tích cực/tiêu cực (gán bởi một số dương), hoặc xem từ nào trong văn bản có nội dung không liên quan đến đối tượng cần nhận định đánh giá (gán bởi một số âm). Nếu dùng một ánh xạ để trực quan hóa các con số này thành 2 loại màu đại diện cho số dương và số âm, chúng ta sẽ thu được cách giải thích mô hình trực quan hơn và dễ tiếp cận đối với mọi người.

1.2 Mục đích nghiên cứu

Integrated Gradient (IG) giúp người dùng có cái nhìn sâu sắc hơn về cách mà mô hình phân tích cảm xúc hoạt động, từ đó nâng cao độ chính xác và tin cậy của kết quả phân tích. Các nhà khoa học dữ liệu, kỹ sư học máy, chuyên viên phân tích, những người làm trong lĩnh vực công nghệ thông tin, y tế, tài chính và các lĩnh vực khác có nhu cầu giải thích mô hình học máy đều có thể áp dụng IG vào ngành nghề của họ để làm tăng tính minh bạch khi sử dụng các mô hình học máy.

Ngoài ra, việc áp dụng IG cho phép người dùng không chuyên về công nghệ tiếp cận và sử dụng các công cụ phân tích cảm xúc một cách hiệu quả hơn, giảm bớt rào cản và khó khăn trong việc tận dụng công nghệ để phục vụ cho nhu cầu phân tích cảm xúc của mình. Hơn nữa, kỹ thuật này cũng mở ra khả năng áp dụng cho nhiều lĩnh vực khác, không chỉ giới hạn ở phân tích cảm xúc, giúp nâng cao chất lượng và hiệu quả của các dự đoán mô hình.

1.3 Đối tượng nghiên cứu

Trong phạm vi của khóa luận này, tôi xin giới thiệu:

- **Integrated Gradients (IG)**, một phương pháp giải thích kết quả dự đoán của mô hình học sâu. IG giúp ta hiểu rõ hơn về những đặc trưng quan trọng nhất của dữ liệu đầu vào trong việc đưa ra dự đoán của mô hình học máy. Cụ thể, kỹ thuật này cung cấp một phương pháp trực quan để xác định sự đóng góp của từng đặc trưng đầu vào đến kết quả dự đoán. IG được tính bằng cách

tích phân gradient của hàm giá trị dự đoán theo đường cong nội suy giữa một điểm đầu vào và một điểm tham chiếu.

- Ứng dụng của IG trong việc giải thích một số mô hình tiêu biểu được sử dụng trong bài toán phân tích cảm xúc (BERT và DistilBERT). IG sẽ chỉ ra những từ nào có ảnh hưởng tới đầu ra và những từ nào không quan và có tính gây nhiễu trong dữ liệu đầu vào.

Để sử dụng IG để giải thích một mô hình phân tích cảm xúc hoạt động như thế nào, ta cần có các đầu vào sau:

- Dữ liệu đầu vào của mô hình học máy là câu (đoạn) văn thể hiện cảm xúc khi đánh giá một dịch vụ
- Dữ liệu đầu ra của mô hình học máy, xem mô hình này gán nhãn gì cho đoạn văn trên
- Cấu trúc mô hình (thường là các tham số trong mô hình đó)

Khi đó IG sẽ trả về đầu ra là một dãy các số thực, mỗi số (theo thứ tự) đánh giá điểm quan trọng của từng đặc trưng (từng token) trong bài đánh giá trên.

1.4 Đóng góp của khóa luận

Trong phạm vi khóa luận này, chúng ta sẽ:

- Tìm hiểu về AI giải thích được và các phương pháp để giải thích một mô hình học máy được coi là "hộp đen".
- Tìm hiểu về một vài mô hình phổ biến được sử dụng trong bài toán phân tích cảm xúc (BERT và DistilBERT)
- Đề xuất một hệ thống có thể giải thích mô hình phân tích cảm xúc
- Áp dụng Integrated Gradients để giải thích một số mô hình như BERT và DistilBERT được sử dụng trong bài toán về phân tích cảm xúc bằng cách cung cấp một cái nhìn trực quan về sự đóng góp của từng từ trong văn bản đối với kết quả phân tích cảm xúc.
- Đánh giá khả năng giải thích của Integrated Gradients đối với mô hình trong phân tích cảm xúc.

1.5 Cấu trúc khóa luận

Phần còn lại của khóa luận được chia thành các chương như sau:

Chương 2: Cơ sở lý thuyết: Giới thiệu về các khái niệm và lý thuyết cơ bản liên quan đến SA, XAI, IG. Các khái niệm và lý thuyết cơ bản được trình bày trong chương này sẽ là nền tảng cho các chương tiếp theo trong khóa luận.

Chương 3: Phương pháp đề xuất: Chương này đề xuất một hệ thống dùng để giải thích hành vi của một số mô hình được sử dụng trong bài toán phân tích cảm xúc.

Chương 4: Ứng dụng vào bài toán phân tích cảm xúc: Chương này áp dụng IG cho mô hình học sâu trong bài toán phân tích cảm xúc (BERT, DistilBERT), phân tích và đánh giá kết quả.

Chương 5: Kết luận: Đánh giá và kết luận hiệu quả của IG và định hướng phát triển trong tương lai.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Chương này trình bày các khái niệm và lý thuyết cơ bản liên quan đến bài toán phân tích quan điểm (Sentiment Analysis hay SA), trí tuệ nhân tạo có thể giải thích được (Explainable Artificial Intelligence hay xAI), và Integrated Gradients (IG). Các khái niệm và lý thuyết cơ bản được trình bày dưới đây sẽ là nền tảng cho các chương tiếp theo trong khóa luận.

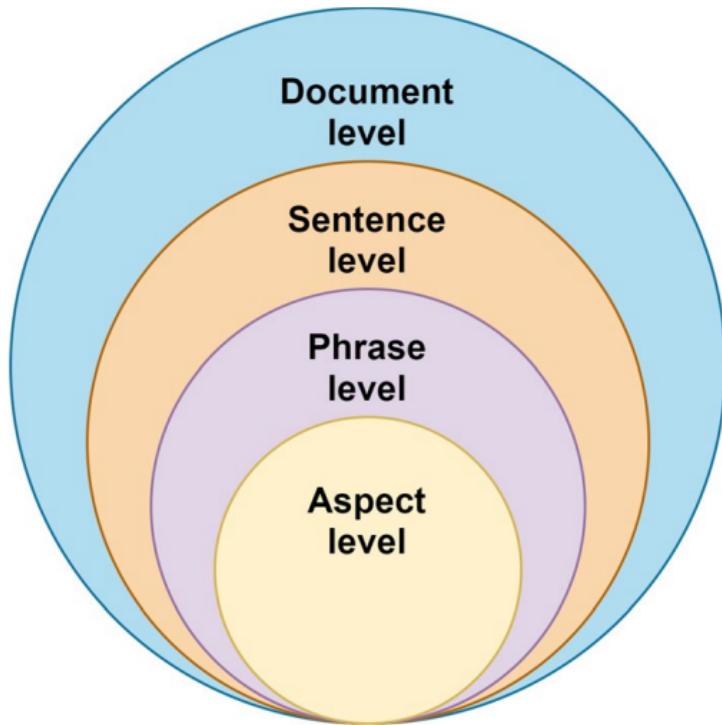
2.1 Phân tích cảm xúc

2.1.1 Khái niệm

phân tích cảm xúc (Sentiment Analysis hay SA) là một lĩnh vực quan trọng trong khoa học dữ liệu và xử lý ngôn ngữ tự nhiên, giúp đánh giá và phân loại cảm xúc, ý kiến hoặc quan điểm của người dùng trong văn bản. Mục đích của SA là xác định xu hướng tích cực, tiêu cực hay trung lập của một đoạn văn bản, từ đó giúp các tổ chức, doanh nghiệp hiểu được cảm nhận của khách hàng, người dùng về sản phẩm, dịch vụ, hoặc chủ đề cụ thể.

Các phương pháp phân tích cảm xúc thường dựa trên các kỹ thuật học máy có giám sát, không giám sát, hay học sâu. Học máy có giám sát đòi hỏi một tập dữ liệu gồm các đánh giá đã được gắn nhãn cảm xúc để đào tạo mô hình, trong khi học máy không giám sát thì không cần tập dữ liệu nhãn. phân tích cảm xúc có thể được thực hiện ở nhiều cấp độ khác nhau, như phân loại cảm xúc toàn bộ văn bản, phân tích cảm xúc ở cấp độ câu, hay phân tích cảm xúc ở cấp độ khía cạnh cụ thể trong văn bản.

SA được ứng dụng rộng rãi trong nhiều lĩnh vực như quản lý quan hệ khách hàng, giám sát uy tín thương hiệu, phân tích phản hồi của người dùng trên mạng xã hội, đánh giá chất lượng sản phẩm và dịch vụ, và hỗ trợ trong việc đưa ra các quyết định kinh doanh dựa trên ý kiến khách hàng.



Hình 2.1: Các mức độ phân tích cảm xúc

2.1.2 Các cấp độ của phân tích cảm xúc

Phân tích cảm xúc được nghiên cứu ở nhiều cấp độ khác nhau: cấp độ văn bản, cấp độ câu, cấp độ cụm từ và cấp độ khía cạnh. phân tích cảm xúc ở những cấp độ từ nhỏ đến lớn sẽ có những đặc điểm riêng [71](Hình 2.1)

Phân tích cảm xúc ở **cấp độ văn bản** được thực hiện trên toàn bộ văn bản, và chỉ đưa ra một cực tính duy nhất cho toàn bộ văn bản. Loại phân tích cảm xúc này không được sử dụng nhiều. Nó có thể được áp dụng để phân loại các chương hay trang của một quyển sách là tích cực, tiêu cực hay trung lập. Tại cấp độ này, cả các phương pháp học có giám sát và không giám sát đều có thể được sử dụng để phân loại văn bản [6]. phân tích cảm xúc đa ngữ cảnh và đa ngôn ngữ là hai vấn đề đáng chú ý nhất trong phân tích cảm xúc ở cấp độ văn bản [58]. phân tích cảm xúc theo ngữ cảnh cụ thể đã được chứng minh có độ chính xác đáng kể nhưng vẫn rất nhạy cảm với ngữ cảnh. Trong những công việc này, vector đặc trưng là một tập hợp các từ cần phải đặc thù cho ngữ cảnh và bị giới hạn.

Ở **cấp độ câu văn**, mỗi câu được phân tích và xác định cực tính tương ứng. Điều này rất hữu ích khi có nhiều loại cảm xúc khác nhau và phức tạp trong cùng một văn bản [76]. Cấp độ phân loại này liên quan đến phân loại chủ quan [53]. Cực tính của mỗi câu sẽ được xác định độc lập bằng cách sử dụng cùng các phương

pháp như cấp độ văn bản, nhưng với dữ liệu đào tạo và nguồn tài nguyên xử lý lớn hơn. Cực tính của mỗi câu có thể được tổng hợp để tìm ra cảm xúc của văn bản hoặc sử dụng riêng lẻ. Đôi khi, phân tích cảm xúc ở cấp độ văn bản không đủ cho những ứng dụng cụ thể [5]. Trong các nghiên cứu trước đây về phân tích ở cấp độ câu, công việc chủ yếu tập trung vào việc tìm kiếm các câu mang tính chủ quan. Tuy nhiên, các nhiệm vụ khó khăn hơn, chẳng hạn như làm việc với các câu có điều kiện hoặc các nhận định mơ hồ [19], phân tích cảm xúc ở cấp độ câu lại trở nên rất quan trọng.

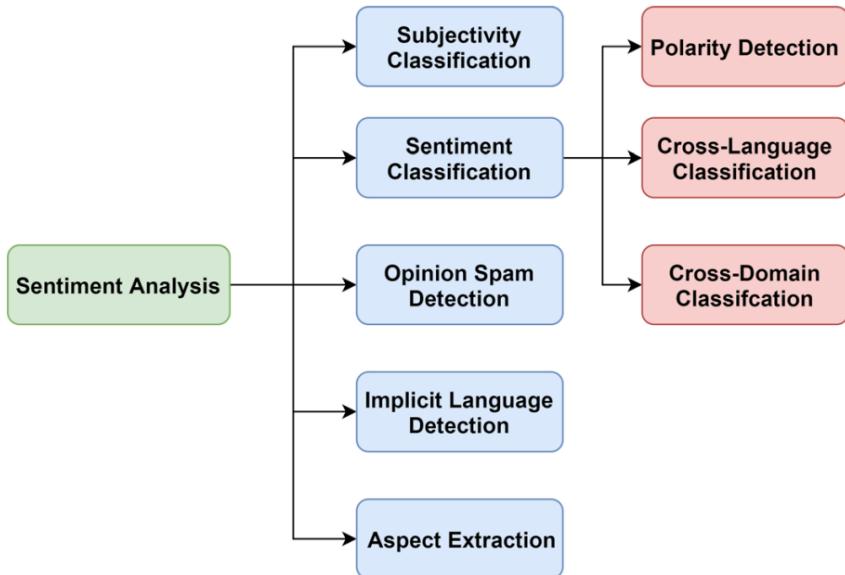
Phân tích cảm xúc cũng có thể được thực hiện ở **cấp độ cụm từ**, nơi các từ chỉ ý kiến được khai thác và phân loại. Mỗi cụm từ có thể chứa nhiều khía cạnh hoặc chỉ một khía cạnh. Điều này có thể hữu ích trong việc đánh giá sản phẩm qua nhiều cụm từ hoặc về câu; ở đây, người ta nhận thấy rằng một khía cạnh duy nhất được thể hiện trong từng cụm từ riêng biệt [66]. Đây là một chủ đề nóng của các nhà nghiên cứu trong thời gian gần đây. Trong khi phân tích ở cấp độ văn bản tập trung vào việc phân loại toàn bộ văn bản theo tính chủ quan, tích cực hoặc tiêu cực, phân tích ở cấp độ câu lại có lợi hơn, vì một văn bản chứa cả các câu tích cực và tiêu cực. Từ là đơn vị cơ bản nhất của ngôn ngữ; cực tính của nó có liên quan chặt chẽ với tính chủ quan của câu hoặc văn bản mà nó xuất hiện. Một câu chứa tính từ có khả năng cao là câu mang tính chủ quan [23]. Ngoài ra, từ được chọn để diễn đạt đại diện cho các đặc điểm dân số của cá nhân, như giới tính và độ tuổi, cũng như mong muốn, địa vị xã hội và tính cách, cũng như các đặc điểm tâm lý và xã hội khác [21]. Do đó, từ đóng vai trò là nền tảng cho phân tích cảm xúc văn bản.

Phân tích cảm xúc thường được thực hiện ở **cấp độ khía cạnh**. Mỗi câu có thể chứa nhiều khía cạnh; do đó, phân tích cảm xúc ở cấp độ khía cạnh dành sự chú ý chính đến tất cả các khía cạnh được sử dụng trong câu và gán cực tính cho tất cả các khía cạnh. Sau đó, cảm xúc tổng hợp được tính toán cho cả câu [59, 44].

2.1.3 Nhiệm vụ của phân tích cảm xúc

Trong phân tích cảm xúc có nhiều loại bài toán (nhiệm vụ) khác nhau cần giải quyết (hình 2.2) như:

Phân loại chủ quan (Subjectivity classification) được coi là bước đầu tiên quan trọng trong lĩnh vực phân tích cảm xúc. Mục đích của phân loại chủ quan



Hình 2.2: Các nhiệm vụ của phân tích cảm xúc

là nhận diện các cụm từ mang tính cảm xúc và ý tưởng chủ quan trong văn bản. Các từ như 'khó khăn', 'tuyệt hảo' và 'đắt đỏ' được xác định là chủ quan [35].

Thông qua việc phân tích những chỉ dẫn này, các đối tượng văn bản chủ quan hay khách quan có thể được phân biệt. Trong bài báo của [35], họ xác định liệu có chủ đề cụ thể nào trong văn bản hay không. Mục tiêu của phân loại chủ quan là loại bỏ những thông tin khách quan không mong muốn ra khỏi các quá trình xử lý tiếp theo [33].

Bằng cách loại bỏ các từ ngữ khách quan, phân loại chủ quan giúp tập trung vào những phần tử mang tính cảm xúc, từ đó nâng cao độ chính xác của phân tích cảm xúc trong các bước tiếp theo. Phân loại chủ quan là một bước tiền đề quan trọng để phát hiện và phân tích cảm xúc người dùng dựa trên ngôn ngữ tự nhiên, điều này đóng vai trò rất quan trọng trong việc nắm bắt ý kiến của khách hàng, phân tích đánh giá sản phẩm và dịch vụ, hoặc theo dõi độ hài lòng của người dùng.

Phân loại cảm xúc (Sentiment classification) là một bài toán nổi tiếng trong SA. Xác định cực tính (polarity) là một trong những nhiệm vụ phụ của phân loại cảm xúc, và thuật ngữ 'phân tích ý kiến' (opinion analysis) thường được sử dụng khi nói đến phân tích cảm xúc. Nhiệm vụ này nhằm xác định cảm xúc của mỗi đoạn văn bản, thường là tích cực hoặc tiêu cực [70].

Trong nghiên cứu [74], các tác giả điều tra bối cảnh ở mức ý kiến, với các khía cạnh ý kiến đơn lẻ và ý kiến chung được mô tả một cách chi tiết, cũng có thể bao gồm ý kiến trùng lập trong một số trường hợp. Với một mô hình phân loại đã

được huấn luyện, phân tích đa ngữ cảnh (cross-domain analysis) sẽ dự đoán cảm xúc của một ngữ cảnh mục tiêu. Việc trích xuất các đặc trưng bất biến theo ngữ cảnh và phân phối chúng là một phương pháp thường được sử dụng [51]. Phân tích đa ngôn ngữ (cross-language analysis) được thực hiện tương tự bằng cách huấn luyện mô hình trên một tập dữ liệu từ một ngôn ngữ nguồn, sau đó đánh giá nó trên một tập dữ liệu từ một ngôn ngữ khác có dữ liệu hạn chế.

Một trong những thách thức mà phân tích cảm xúc phải đối mặt là sự mơ hồ của cực tính từ. Các bài báo [68, 63] đã chỉ ra rằng các mô hình dựa trên truy xuất cung cấp một lựa chọn thay thế cho các chiến lược dựa trên học máy để phát hiện cực tính từ. Tính toán ảnh hưởng (affective computing) và phân tích cảm xúc cũng có tiềm năng lớn như công nghệ phụ trợ cho các hệ thống khác [10]. Chúng có thể mở rộng khả năng của hệ thống quản lý quan hệ khách hàng và hệ thống đề xuất, bằng cách cho phép khám phá ra những tính năng mà khách hàng đặc biệt yêu thích hoặc loại bỏ những mặt hàng đã nhận được phản hồi rất không thuận lợi khỏi danh sách đề xuất.

Thông qua phân loại cảm xúc, các doanh nghiệp và tổ chức có thể hiểu được tâm trạng, ý kiến và cảm nhận của người dùng về sản phẩm, dịch vụ hoặc chính sách một cách chính xác và nhanh chóng. Điều này giúp họ đưa ra những quyết định phù hợp với mong muốn của khách hàng, tối ưu hóa chiến lược kinh doanh và cải thiện chất lượng dịch vụ. Phân loại cảm xúc cũng đóng vai trò quan trọng trong việc theo dõi xu hướng trên mạng xã hội, giám sát đánh giá sản phẩm và phân tích thị trường, từ đó giúp các doanh nghiệp cạnh tranh và thích ứng với thị trường đầy biến động.

Phát hiện bình luận spam (Opinion Spam Detection) đã trở thành một thách thức đáng kể trong phân tích cảm xúc do sự tăng trưởng của thương mại điện tử và các nền tảng đánh giá. Bình luận spam, còn được gọi là đánh giá giả mạo, là những nhận xét được viết tốt ủng hộ hoặc phê bình một sản phẩm vì lợi ích của họ. Phát hiện bình luận spam nhằm xác định ba đặc điểm riêng biệt của một đánh giá giả mạo: nội dung đánh giá, metadata của đánh giá và kinh nghiệm sản phẩm thực tế [15]. Các thuật toán học máy thường được sử dụng để đánh giá văn bản đánh giá nhằm phát hiện sự gian lận. Đánh giá số sao hoặc điểm, địa chỉ IP của người dùng, vị trí địa lý của người dùng và thông tin khác là một số Metadata được sử dụng trong việc phát hiện bình luận spam. Tuy nhiên, trong nhiều trường hợp, thông tin này không thể truy cập để phân tích. Tuy nhiên kinh nghiệm và

kiến thức thực tế sẽ được đưa vào để giải quyết. Ví dụ, nếu một sản phẩm có đánh giá và xếp hạng kém đang được đánh giá cao trong một khoảng thời gian liên tục, điều này có thể bị sự nghi ngờ và phân tích để phát hiện bình luận spam.

Phát hiện hàm ý (Implicit Language Detection): Những phương thức hành văn không rõ ràng và mơ hồ này trong ngôn ngữ thường rất khó để phát hiện, thậm chí đôi khi còn khó với chính con người. Tuy nhiên, ngôn ngữ ẩn này là một khía cạnh thiết yếu của câu và có thể hoàn toàn lật ngược ý nghĩa và độc lập của câu. Ví dụ, xem xét cụm từ 'Hay quá, tôi bị sa thải rồi'. Cụm 'Hay quá' rất tích cực, nhưng lại mô tả sự mỉa mai hoặc châm biếm khi kết hợp với các phần sau, tức là 'tôi bị sa thải rồi', khiến cho cụm từ 'tôi bị sa thải rồi' trở nên tiêu cực hơn. Việc phân tích các nhóm ký tự tạo thành các biểu tượng cảm xúc (như :), :(, :D) là những phương pháp cổ điển hơn để phát hiện ngôn ngữ ẩn [18, 20].

Trích xuất khía cạnh (Aspect Extraction): phân tích cảm xúc theo khía cạnh là một phương pháp quan trọng trong phân tích cảm xúc, bao gồm ba bước chính: trích xuất khía cạnh, phân loại cực tính và tổng hợp. Đây là bước khởi đầu và cũng là điểm khác biệt so với phân tích cảm xúc thông thường. Khía cạnh có thể được trích xuất thông qua tập hợp các khía cạnh được định nghĩa trước, dựa trên lĩnh vực mà nó được áp dụng.

Ngoài ra, còn có các phương pháp tiên tiến hơn như phương pháp dựa trên tần suất, phương pháp dựa trên cú pháp, và các phương pháp học máy có giám sát và không giám sát. Trong các đánh giá [34], một số từ được sử dụng phổ biến hơn những từ khác và những từ này thường mang tính khía cạnh. Phương pháp này có thể trở thành cách tiếp cận mạnh mẽ nhưng cũng có một số hạn chế, chẳng hạn như không phải tất cả các danh từ phổ biến đều liên quan đến khía cạnh, hoặc các khía cạnh ít được nhắc đến có thể bị bỏ qua, ví dụ như các từ 'bát cơm', 'đồng', 'đi lại' đều không có xu hướng cảm xúc của bất kỳ khía cạnh nào.

Để khắc phục những vấn đề này, có thể kết hợp một tập quy tắc với phương pháp dựa trên tần suất. Tuy nhiên, việc điều chỉnh thủ công các thông số này là công việc vất vả và tốn thời gian. Thay vào đó, có thể sử dụng phương pháp dựa trên cú pháp, vì nó khắc phục được nhược điểm của phương pháp dựa trên tần suất khi không phát hiện các khía cạnh ít xuất hiện [4]. Ví dụ, từ 'đẹp' trong cụm 'Điện thoại rất đẹp' được xem là tính từ ám chỉ khía cạnh 'Điện thoại'. Để áp dụng phương pháp này, cần thu thập nhiều dữ liệu được gán nhãn, bao gồm tất cả các mối quan hệ cú pháp để đào tạo thuật toán.

2.1.4 Phương pháp phân loại cảm xúc

Ba cách tiếp cận chủ yếu được sử dụng để phân tích cảm xúc bao gồm Cách tiếp cận dựa trên từ vựng, học máy và phương pháp lai giữa từ vựng và học máy. Ngoài ra còn một số phương pháp khác như mạng nơ-ron, phân tích cảm xúc theo nhiều khía cạnh (aspect-based SA), học chuyển giao (transfer learning), phân tích cảm xúc đa phương tiện (multimodal SA) [71].

Phương pháp dựa trên từ vựng: Từ vựng là tập hợp các mã thông báo, điểm số của từng từ được sẽ là một số thực nằm trong đoạn [-1, +1], trong đó +1 đại diện cho rất tích cực, 0 đại diện có sự trung tính và -1 đại diện cho rất tiêu cực. Trong Phương pháp dựa trên từ vựng, đối với một đánh giá hoặc văn bản đã cho, sẽ tính điểm số của mỗi từ, và điểm số tích cực, tiêu cực, trung tính được cộng lại riêng biệt. Cuối cùng, tính cực tổng thể được gán cho văn bản dựa trên giá trị cao nhất của các điểm số trên từng từ. Do đó, tài liệu trước tiên được chia thành các mã từ đơn lẻ, sau đó tính cực của mỗi mã được tính toán và tổng hợp cuối cùng.

Kỹ thuật dựa trên từ vựng rất phù hợp cho phân tích cảm xúc ở cấp độ câu văn. Vì không yêu cầu dữ liệu huấn luyện, nó có thể được coi là một kỹ thuật không giám sát. Mặt khác, nhược điểm chính của kỹ thuật này là phụ thuộc vào lĩnh vực, vì các từ có thể có nhiều nghĩa và ý nghĩa, do đó một từ tích cực trong một lĩnh vực có thể là tiêu cực trong lĩnh vực khác. Ví dụ, xét từ 'nhỏ' và các câu 'Màn hình TV quá nhỏ' và 'Máy ảnh này cực kỳ nhỏ', từ 'nhỏ' trong câu đầu tiên là tiêu cực, vì mọi người thường ưa thích màn hình lớn, trong khi ở câu thứ hai, nó lại tích cực, vì nếu máy ảnh nhỏ, nó sẽ dễ mang theo. Vấn đề này có thể được giải quyết bằng cách phát triển một từ vựng cảm xúc đặc thù cho lĩnh vực hoặc bằng cách thích ứng một từ vựng hiện có.

Ưu điểm của phương pháp dựa trên từ vựng là không yêu cầu dữ liệu huấn luyện và được một số chuyên gia xem là phương pháp không giám sát [75]. Nhược điểm chính của phương pháp dựa trên từ vựng là nó rất phụ thuộc vào lĩnh vực và các từ liên quan đến một lĩnh vực không thể được sử dụng trong một lĩnh vực khác [48]. Ví dụ, xem xét từ 'to lớn', nó có thể là tích cực hoặc tiêu cực dựa trên lĩnh vực mà nó được sử dụng. Trong 'hàng đợi xem phim rất lớn', từ này có thể được coi là tích cực, trong khi trong 'có sự chậm trễ lớn trong mạng' từ này có thể được coi là tiêu cực. Do đó, tính cực nên được gán cho các từ một cách cẩn thận, xem xét lĩnh vực.

Có chủ yếu hai phương pháp được sử dụng trong Phương pháp dựa trên từ vựng là phương pháp dựa trên kho dữ liệu văn bản và phương pháp thống kê.

- **Phương pháp dựa trên kho dữ liệu** (Corpus-based approach) sử dụng các mẫu ngữ nghĩa và cú pháp để xác định cảm xúc của câu. Phương pháp bắt đầu với một tập hợp các thuật ngữ cảm xúc đã định nghĩa và định hướng, sau đó tìm kiếm các mẫu cú pháp hoặc tương tự để tìm ra các đơn vị cảm xúc và định hướng trong một kho ngữ liệu lớn. Đây là phương pháp phụ thuộc vào bối cảnh cụ thể và cần nhiều dữ liệu được gắn nhãn để huấn luyện. Tuy nhiên, nó giúp giải quyết vấn đề các từ ý kiến có định hướng phụ thuộc vào ngữ cảnh. [50] áp dụng kho ngữ liệu với hai loại phương pháp: Thống kê và Ngữ nghĩa. Phương pháp Thống kê dựa vào tần suất xuất hiện của từ trong văn bản tích cực hoặc tiêu cực. Phương pháp Ngữ nghĩa tính điểm tương tự giữa các từ, sử dụng Wordnet để tìm đồng nghĩa và trái nghĩa.
- **Phương pháp dựa trên từ điển** (Dictionary-based method) bao gồm danh sách các từ ý kiến được thu thập một cách thủ công [14, 32]. Ý tưởng chính đằng sau phương pháp này là những từ đồng nghĩa chắc chắn có cùng cực tính, và từ trái nghĩa sẽ trái ngược nhau về cảm xúc. Các kho ngữ liệu lớn như từ điển đồng nghĩa, trái nghĩa hoặc Wordnet được tìm kiếm để tìm các từ đồng nghĩa và trái nghĩa, sau đó nối vào nhóm hoặc danh sách hạt giống đã chuẩn bị trước đó. Trong giai đoạn đầu tiên, tập từ ban đầu được thu thập thủ công với cảm xúc của chúng. Sau đó, danh sách được mở rộng bằng cách xem xét các từ đồng nghĩa và trái nghĩa trong các nguồn từ vựng có sẵn [30, 62]. Tiếp theo, các từ được thêm vào danh sách lặp đi lặp lại và danh sách được mở rộng. Đánh giá thủ công hoặc sửa chữa có thể được thực hiện ở giai đoạn cuối để đảm bảo chất lượng của nó. Các công cụ dựa trên phương pháp từ điển bao gồm SentiWordNet, Bing Liu's Sentiment Lexicon, SentiStrength, Opinion Finder và National Taiwan University Sentiment Dictionary. Tuy nhiên, phương pháp từ điển chỉ hiệu quả với kích thước từ điển nhỏ và có khó khăn trong việc tìm từ ý kiến đặc thù cho từng lĩnh vực.

Phương pháp dựa theo học máy: Các thuật toán Học Máy có thể được sử dụng để phân loại cảm xúc. Phân tích cảm xúc là quá trình xác định và định lượng cảm xúc của văn bản hoặc âm thanh bằng cách sử dụng xử lý ngôn ngữ tự nhiên, phân tích văn bản, ngôn ngữ học tính toán và các kỹ thuật khác. Có hai phương pháp chính trong Học Máy để phân tích cảm xúc:

- Học Máy có giám sát
- Học Máy không giám sát dựa trên từ điển

Nhiệm vụ này có thể được thực hiện bằng cả hai phương pháp học có giám sát và không giám sát. Các chiến lược không giám sát cho phân tích cảm xúc bằng cách sử dụng cơ sở tri thức, ôn tập, cơ sở dữ liệu và từ điển bao gồm kiến thức chi tiết đã được chọn và chuẩn bị đặc biệt cho phân tích cảm xúc. Các phương pháp học có giám sát được sử dụng phổ biến hơn do kết quả chính xác của chúng. Các thuật toán này cần được đào tạo trên một tập huấn luyện trước khi áp dụng cho dữ liệu thực tế. Các đặc trưng có thể được trích xuất từ dữ liệu văn bản.

Kỹ thuật Học Máy sử dụng các yếu tố cú pháp và/hoặc ngôn ngữ để giải quyết phân loại cảm xúc như một vấn đề phân loại văn bản tiêu chuẩn sử dụng các yếu tố cú pháp và/hoặc ngôn ngữ. Mô hình phân loại liên kết các đặc trưng của hồ sơ gốc với một trong nhãn lớp. Mô hình sau đó được sử dụng để dự đoán nhãn lớp cho một trường hợp không xác định lớp. Khi một trường hợp chỉ được gán một nhãn, chúng ta có một thách thức phân loại khó. Khi một giá trị xác suất của nhãn được gán cho một trường hợp, điều này được gọi là vấn đề phân loại mềm. Học Máy cho phép các hệ thống có được khả năng mới mà không cần được lập trình cụ thể để làm điều đó. Các thuật toán phân tích cảm xúc có thể được đào tạo để đọc ngoài các định nghĩa đơn giản để hiểu thông tin ngữ cảnh, châm biếm và từ ngữ sai lầm.

Phương pháp lai giữa từ vựng và học máy: Phương pháp lai kết hợp giữa học máy và phương pháp dựa trên từ điển. Lai là thuật ngữ chỉ sự kết hợp giữa các kỹ thuật học máy và dựa trên từ điển để phân tích cảm xúc. Phương pháp lai giữa hai phương pháp kể trên rất phổ biến, với từ điển cảm xúc đóng một vai trò đáng kể trong đa số hệ thống. Phân tích cảm xúc là một phương pháp lai, bao gồm cả các phương pháp thống kê và dựa trên kiến thức để nhận dạng cực. Trong bài báo của [28], họ đề xuất một phương pháp lai giữa học máy sử dụng SVM và hai kỹ thuật lựa chọn đặc trưng sử dụng bộ tối ưu hóa da vũ trụ và các thuật toán Relief [13]. Nhiệm vụ phân tích cảm xúc [2] đề xuất sử dụng phương pháp lai dựa trên học máy bao gồm RF và SVM. Họ đã chỉ ra rằng các mô hình cá nhân của SVM và RF có độ chính xác lần lượt là 81,01% và 82,03%, trong khi phương pháp lai kết hợp cả hai thuật toán có độ chính xác gần 84% trong bộ dữ liệu đánh giá sản phẩm cung cấp bởi amazon.com. Một số nhà nghiên cứu đã đề xuất một kiến trúc lai kết hợp cả hai kỹ thuật dựa trên từ điển và học tự động để cải thiện kết

quả. Đây vẫn là một chủ đề nóng cho các nhà nghiên cứu, và cần nhiều nghiên cứu hơn.

Trong bài báo [28], họ sử dụng dữ liệu Twitter để đào tạo. Có đến 6900 tweet được trích xuất để đào tạo bằng API Twitter. Kết quả cho thấy mô hình của họ vượt trội hơn hầu hết các mô hình khác trong khi giảm tổng số đặc trưng lên đến 96%. Họ cũng chỉ ra khả năng của các phương pháp lai và kết luận rằng các mô hình lai có thể vượt trội hơn tất cả các mô hình với kiến trúc phù hợp và lựa chọn tham số siêu phù hợp [13]. Mô hình lai vượt trội hơn cả hai mô hình trong tất cả các chỉ số và so sánh khác [28].

2.2 Một vài mô hình học máy sử dụng trong bài toán phân tích cảm xúc

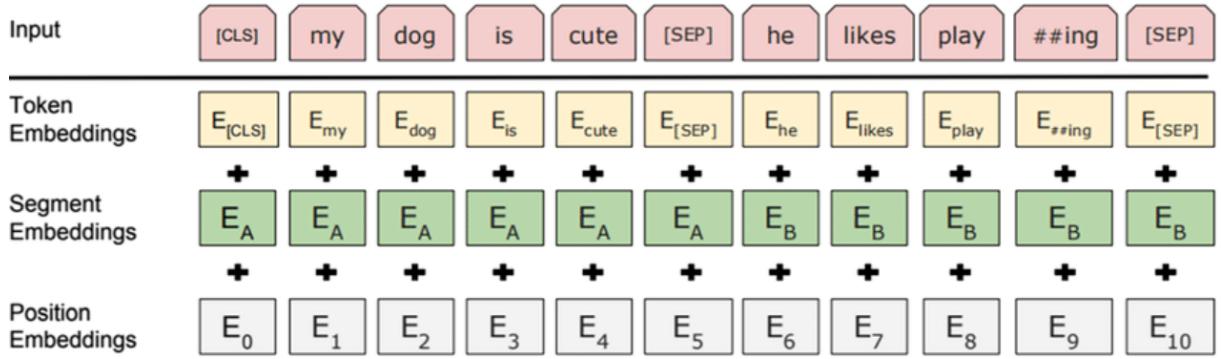
Các mô hình trong SA có vai trò quan trọng trong việc phát hiện và phân tích ý kiến của người dùng về các sản phẩm, dịch vụ, sự kiện, ... Dưới đây là một vài mô hình tiêu biểu được sử dụng nhiều trong các bài toán phân tích cảm xúc.

2.2.1 BERT

BERT được đề xuất bởi [17]. Đây là một mô hình mạng thần kinh được thiết kế cho các tác vụ hiểu ngôn ngữ tự nhiên (Natural Language Understanding) và được đào tạo trước trên một nguồn dữ liệu lớn hơn 3 tỷ từ bao gồm BooksCorpus và Wikipedia tiếng Anh.

Một trong những lý do chính cho sự thành công của BERT là việc sử dụng đào tạo trước (pretrain) tự giám sát chuyên sâu và cơ chế self-attention từ kiến trúc transformer. Tự chú ý cho phép mỗi từ trong câu có thể chú ý đến tất cả các từ khác, giúp BERT đạt được hiểu biết hai chiều sâu sắc về ngữ cảnh của từ. Điều này khác với các phương pháp biểu diễn từ trước đây, là độc lập với ngữ cảnh (ví dụ như WordNet, GloVe) hoặc chỉ có thể bao phủ ngữ cảnh một chiều (ví dụ như ELMo [17]).

BERT là mô hình ngôn ngữ đầu tiên có thực sự hiểu biết hai chiều về ngữ cảnh của ngôn ngữ, điều này rất hữu ích cho các tác vụ khó như trả lời câu hỏi. Kiến trúc của BERT bao gồm nửa đầu tiên của encoder trong kiến trúc transformer, bao gồm một lớp nhúng và một chồng N lớp bộ mã hóa transformer.



Hình 2.3: Cách BERT nhúng đầu vào

Về thông số đầu vào và đầu ra, BERT yêu cầu đầu vào là văn bản đã được mã hóa token bằng mã hóa riêng của nó. Đầu ra của BERT là biểu diễn vector của văn bản đầu vào, có thể được sử dụng cho các tác vụ NLU như phân loại, trả lời câu hỏi và tạo ngôn ngữ [17].

- **Đầu vào và đầu ra:** Đầu vào của mô hình là một dãy các từ được tokenized $\{w_1, w_2, \dots, w_n\}$, và được số hóa thành một vector $\{x_1, x_2, \dots, x_n\}$.

Đầu ra là một danh sách các vectơ $\{z_1, z_2, \dots, z_n\} \in \mathbb{R}^H$, mỗi vectơ tương ứng với một mã thông báo đầu vào của cùng một chỉ mục. Với H là kích thước trạng thái ẩn, một tham số được xác định trước khi mô hình được khởi tạo, ví dụ 512 hoặc 1024.

Đầu tiên, các chuỗi đầu vào được xử lý thành các subword token thông qua tokenizer.

- **Token embeddings:** tương tự như ý tưởng của word embeddings (WordNet, GloVe), biểu diễn vector one-hot của một token đầu vào được nhúng vào không gian có kích thước H . Lớp nhúng này sẽ được học trong quá trình tiền huấn luyện.
- Khi đầu vào là một cặp câu, các segment embeddings sẽ được sử dụng để phân biệt token đến từ câu đầu tiên hay thứ hai.
- Các positional embeddings giúp mô hình nhận thức được thứ tự của các token đầu vào. Hình 2.3 minh họa quá trình đó.

Các token đặc biệt cũng được thêm vào chuỗi đầu vào bởi tokenizer. Mỗi token này có mục đích riêng của nó, đặc biệt là cho hai tác vụ tiền huấn luyện được mô tả sau đây.

- **Token CLS**, viết tắt của classification, luôn được thêm vào vị trí đầu tiên và đầu ra của nó được sử dụng cho các tác vụ phân loại.
- **Token SEP**, viết tắt của separation, được sử dụng để mô hình phân biệt một cặp chuỗi (với segment embeddings).
- **Token PAD**, viết tắt của padding, được sử dụng để ghép các chuỗi có độ dài khác nhau. Các attention masks sẽ được tạo ra để xóa bỏ attention không cần thiết đến các token PAD.
- **Mã hóa từ vựng** (Tokenizer) là công cụ để chuyển đổi văn bản thành một số biểu diễn số học hợp lý, ví dụ như một danh sách các mã thông báo, để cho mô hình xử lý. Hợp lý ở đây có nghĩa là các mã thông báo phải được thu gọn, nén hiệu quả và dễ hiểu. Ví dụ, ta muốn chia nhỏ cụm từ 'A good restaurant' thành 'A, good, restaurant', mà không phải là 'A, g, o, o, d, r, e, s, t, a, u, r, a, n, t'.

Văn bản có thể được chia thành danh sách các mã thông báo có thể là từ, các phần tử của từ hoặc các ký tự. Chia nhỏ từ rất đơn giản và có thể được thực hiện bằng cách tách các khoảng trống và dấu chấm câu đơn giản và một số quy tắc, nhưng có thể dẫn đến một bộ từ vựng rất lớn (vì các từ tương tự nhau như 'play', 'playing' được xử lý khác biệt). Chia nhỏ theo ký tự đơn giản, linh hoạt với bộ từ vựng khiêm tốn nhưng làm cho việc tạo ra biểu diễn có ý nghĩa cho mô hình trở nên khó khăn và do đó thường liên quan đến mất hiệu suất. Hơn nữa, chuỗi đã được chia nhỏ sẽ có độ dài dài hơn, dẫn đến chi phí tính toán tăng lên. Sử dụng phương pháp chia nhỏ từ và ký tự kết hợp, được gọi là chia nhỏ phần tử, là tối ưu hơn (ví dụ, 'playing' có thể được chia thành 'play' và 'ing'). Chia nhỏ phần tử có những lợi thế về kích thước từ vựng hợp lý, chi phí tính toán hiệu quả, cho phép biểu diễn có ý nghĩa và một mức độ linh hoạt.

Có ba loại chia nhỏ các cụm từ độc lập ngôn ngữ chính: BytePair Encoding (BPE) [60], WordPiece [73] và SentencePiece unigram [39], được cho là mới nhất và phổ biến nhất cho các bộ biến đổi hiện nay. Các thuật toán chia nhỏ phần tử chính bao gồm từ vựng, chiến lược đào tạo và chiến lược mã hóa. Các tokenizer cần được đào tạo trên một tập văn bản lớn để học một tập từ vựng tối ưu, có thể bao gồm các phần tử đã học, trước khi có thể chia nhỏ (hoặc mã hóa) văn bản bằng cách sử dụng một chiến lược mã hóa cụ thể chẳng hạn như các bước chuẩn hóa văn bản, chia nhỏ trước và xử lý sau (thêm các mã

thông báo đặc biệt).

BERT sử dụng tokenizer WordPiece [73]. Tương tự như BPE, thuật toán đào tạo khởi tạo từ vựng dưới dạng một tập hợp nhỏ các ký tự có thể có trong tập văn bản và sau đó dựa trên thống kê tần số để hợp nhất các ký tự thành các từ, với một khái niệm bổ sung về ký tự đặc biệt `#`. Tại mỗi bước đào tạo, tất cả các cặp mã thông báo trong từ vựng được ghi điểm dựa trên tần số f và cặp có điểm số lớn nhất sẽ được hợp nhất. Sau đó, quá trình đào tạo tiếp tục cho đến khi đạt được một số lượng phần tử đã được xác định trước. Để mã hóa văn bản, cho mỗi từ, cần tìm phần tử dài nhất có thể từ đầu rồi làm tương tự cho phần còn lại của từ.

- **Tiền huấn luyện:** Có hai phương pháp tiền huấn luyện (pretraining) sử dụng phương pháp đào tạo tự giám sát (self-supervised), đó là Masked Language Model (MLM) và Next Sentence Prediction (NSP). MLM tập trung vào việc đào tạo mô hình để dự đoán các từ bị che giấu trong câu, trong khi NSP tập trung vào việc đào tạo mô hình để dự đoán xem hai câu có liên quan đến nhau hay không. Trong quá trình đó, một layer gọi là language model head chỉ dùng để tiền huấn luyện và sẽ bị bỏ đi khi bắt đầu nhiệm vụ đằng sau (down-stream tasks).
 - Trong MLM, khoảng 15% của các từ trong câu được ngẫu nhiên chọn để bị che giấu bằng ký hiệu [MASK]. Mô hình được đào tạo để dự đoán các từ này bằng cách sử dụng thông tin từ các từ còn lại trong câu văn. Điều này giúp mô hình học được mối quan hệ giữa các từ trong câu và tạo ra một biểu diễn có ý nghĩa cho mỗi từ.
 - Trong NSP, hai câu ngẫu nhiên được ngăn cách bởi token [SEP]. Mô hình được đào tạo để đưa ra dự đoán xem hai câu có liên quan đến nhau hay không. Để làm điều này, hai câu được chọn ngẫu nhiên từ một tài liệu văn bản, và mô hình được đào tạo để dự đoán xem câu thứ hai có phải là câu tiếp theo của câu đầu tiên hay không.

MLM và NSP đóng vai trò quan trọng trong quá trình pretraining cho BERT, giúp mô hình học được các đặc trưng ngôn ngữ tổng quát và đa dạng. Theo nghiên cứu của các tác giả của BERT, cả hai nhiệm vụ này đóng vai trò quan trọng trong việc cải thiện hiệu suất của mô hình trên nhiều tác vụ ngôn ngữ khác nhau.

Ví dụ ta có câu: 'The cat sat on the mat. It was a sunny day outside.'

Đối với MLM, khoảng 15% của các từ trong câu được ngẫu nhiên chọn để bị che giấu bằng ký hiệu [MASK]. Đối với NSP, hai câu được chọn ngẫu nhiên từ một tài liệu văn bản và token [SEP] sẽ được đặt giữa hai câu đó, mô hình được đào tạo để dự đoán xem câu thứ hai có phải là câu tiếp theo của câu đầu tiên hay không. Vì vậy, câu trên có thể được biến đổi thành:

'The [MASK] sat on the [MASK] [SEP] It was a [MASK] day outside.'

Mô hình BERT được đào tạo để dự đoán các từ bị che giấu bằng cách sử dụng thông tin từ các từ còn lại trong câu. Ví dụ, khi mô hình nhìn vào từ đầu tiên '[MASK]', nó có thể dự đoán rằng từ đó có thể là 'cat', 'dog' hoặc 'mouse' dựa trên ngữ cảnh của câu. Mô hình BERT cũng được đào tạo để dự đoán rằng câu thứ hai 'It was a sunny day outside.' là một câu tiếp theo hợp lý của câu đầu tiên 'The cat sat on the mat.'

- **Tập dữ liệu để huấn luyện BERT**

Để huấn luyện mô hình BERT, các tập dữ liệu lớn được sử dụng để cung cấp cho mô hình các ví dụ về ngôn ngữ tổng quát và đa dạng. Dưới đây là một số tập dữ liệu phổ biến được sử dụng để huấn luyện mô hình BERT:

- **Wikipedia** là một nguồn dữ liệu lớn về văn bản với hàng triệu bài viết trên nhiều chủ đề khác nhau. Các nghiên cứu cho thấy rằng đào tạo BERT trên tập dữ liệu Wikipedia đã cải thiện đáng kể hiệu suất của mô hình trên nhiều tác vụ ngôn ngữ.
- **BookCorpus** là một tập dữ liệu được tạo ra bằng cách tự động tải xuống và xử lý các cuốn sách từ trang web thư viện điện tử. Đây là một nguồn dữ liệu lớn về văn bản với khoảng 11.000 cuốn sách và hơn 800 triệu từ.
- **Common Crawl** là một dự án lưu trữ web lớn với hàng tỷ trang web được lưu trữ. Tập dữ liệu Common Crawl có thể được sử dụng để tạo ra các tập dữ liệu lớn cho việc đào tạo mô hình BERT.
- **SNLI** (The Stanford Natural Language Inference) là một tập dữ liệu được sử dụng để đào tạo và đánh giá các mô hình ngôn ngữ tổng quát. Tập dữ liệu này bao gồm các cặp câu được gắn nhãn về mối quan hệ giữa chúng, bao gồm 'entailment' (kết quả), 'contradiction' (mâu thuẫn) và 'neutral' (trung tính).
- **GLUE** (General Language Understanding Evaluation) là một tập dữ liệu được sử dụng để đánh giá hiệu suất của các mô hình ngôn ngữ tổng quát

trên nhiều tác vụ khác nhau, bao gồm phân loại văn bản, dự đoán liên kết giữa các câu, hỏi đáp và phân tích ngôn ngữ tự nhiên.

Các tập dữ liệu này đều cung cấp các ví dụ về ngôn ngữ tổng quát và đa dạng để huấn luyện mô hình BERT. Tuy nhiên, để đạt được hiệu suất tốt nhất trên các tác vụ ngôn ngữ cụ thể, một số tập dữ liệu có thể cần được tạo ra hoặc điều chỉnh để phù hợp với mục đích sử dụng cụ thể của mô hình.

• BERT trong bài toán phân tích cảm xúc

BERT đã mang lại những cải tiến đột phá trong nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP) khác nhau, trong đó có phân tích cảm xúc (SA). Phân tích cảm xúc là tác vụ xác định cảm xúc, suy nghĩ hoặc ý kiến được bày tỏ trong một đoạn văn bản. BERT giúp cải thiện độ chính xác của phân tích cảm xúc bằng cách sử dụng cấu trúc Transformer để hiểu ngữ cảnh từ hai hướng (cả văn bản trước và sau từ đang xét). Kết quả, BERT đã giúp cải thiện đáng kể kết quả của các tác vụ phân tích cảm xúc trong nhiều bài toán NLP, từ đánh giá sản phẩm đến phân tích cảm xúc trên mạng xã hội. BERT đã trở thành một trong những công cụ tiên tiến nhất trong lĩnh vực NLP, đặc biệt là phân tích cảm xúc, bởi khả năng hiểu ngữ nghĩa phức tạp và ngữ cảnh của văn bản. Để áp dụng BERT cho phân tích cảm xúc, ta cần tiến hành hai bước chính: tiền huấn luyện (pre-training) và tinh chỉnh (fine-tuning).

- **Tiền huấn luyện:** BERT được tiền huấn luyện trên một lượng lớn văn bản không gắn nhãn (unlabeled) bằng cách sử dụng hai kỹ thuật: dự đoán từ bị che khuất (masked language modeling) và dự đoán câu tiếp theo (next sentence prediction). Mục đích của tiền huấn luyện là học được các biểu diễn ngôn ngữ phong phú và có tính đại diện.
- **Tinh chỉnh:** Sau khi đã tiền huấn luyện, BERT có thể được tinh chỉnh cho tác vụ phân tích cảm xúc bằng cách sử dụng một vài tập dữ liệu gắn nhãn (labeled) về quan điểm. Quá trình này giúp mô hình học cách liên kết biểu diễn ngôn ngữ đã học được với các nhãn quan điểm (ví dụ: tích cực, tiêu cực, hoặc trung lập).

Các biến thể của BERT, như RoBERTa, ALBERT và DistilBERT, cũng được đề xuất sau đó để giải quyết một số hạn chế về tài nguyên tính toán và kích thước mô hình của BERT gốc [43, 40, 57]. Chúng cũng đã được áp dụng thành công trong phân tích cảm xúc, mang lại hiệu suất tương đương hoặc cao hơn so với BERT gốc [17, 64, 77].

2.2.2 DistilBERT

Phiên bản DistilBERT [57] là một biến thể tinh gọn của mô hình BERT gốc. Dù vẫn duy trì kiến trúc Transformer đặc trưng của BERT, DistilBERT đã giảm thiểu đáng kể kích thước và độ phức tạp. Thông qua việc áp dụng kỹ thuật 'distillation' (nén mô hình), DistilBERT đã giảm được lượng tham số và thời gian huấn luyện mô hình, đồng thời vẫn đạt được độ chính xác tương đối cao so với phiên bản BERT gốc.

- **Đặc điểm của DistilBERT** Một số đặc điểm nổi bật của DistilBERT có thể kể đến như:

- **Kích thước nhỏ gọn:** DistilBERT giảm kích thước mô hình gốc BERT xuống khoảng 40%, từ 340 triệu tham số xuống còn khoảng 66 triệu tham số.
- **Tốc độ huấn luyện nhanh hơn:** Nhờ kích thước nhỏ gọn hơn, thời gian huấn luyện của DistilBERT cũng nhanh hơn so với BERT, giúp tiết kiệm tài nguyên và thời gian tính toán.
- **Hiệu năng tốt:** Mặc dù nhỏ gọn hơn, DistilBERT vẫn giữ được 97% hiệu năng của BERT trên tập dữ liệu GLUE benchmark, cho thấy nó vẫn rất hiệu quả trong các tác vụ xử lý ngôn ngữ tự nhiên.

Kỹ thuật 'distillation' trong DistilBERT [57] được thực hiện bằng cách huấn luyện một mô hình nhỏ hơn (DistilBERT) với dữ liệu đầu ra từ một mô hình lớn hơn (BERT). Quá trình này giúp mô hình nhỏ hơn học được mối quan hệ giữa các đặc trưng đầu vào và đầu ra mà không cần tiếp xúc trực tiếp với dữ liệu huấn luyện gốc [29]. Kết quả là mô hình nhỏ hơn nhanh chóng hội tụ và đạt được hiệu suất tương đương với mô hình lớn hơn.

DistilBERT đã trở thành một công cụ hữu ích trong việc triển khai các ứng dụng xử lý ngôn ngữ tự nhiên đòi hỏi hiệu năng cao mà vẫn đảm bảo tiết kiệm tài nguyên và thời gian tính toán. DistilBERT được sử dụng trong nhiều tác vụ NLP như phân loại văn bản, phân tích cảm xúc, trích xuất thông tin, dịch máy, đặc biệt là trong các môi trường có giới hạn về tài nguyên như thiết bị di động hoặc edge computing [57].

DistilBERT cũng đã truyền cảm hứng cho việc phát triển các mô hình rút gọn khác như DistilGPT [52, 43] và nhiều mô hình khác trong gia đình Hugging Face Transformers [72]. Các mô hình này đều áp dụng cách tiếp cận tương tự

để giảm kích thước và tăng tốc độ huấn luyện, đồng thời giữ được hiệu năng cao trong các tác vụ xử lý ngôn ngữ tự nhiên.

Tóm lại, DistilBERT là một bước tiến đáng kể trong việc tối ưu hóa và giảm kích thước của các mô hình NLP dựa trên kiến trúc Transformer. Việc giữ được hiệu năng cao trong khi tiết kiệm tài nguyên và thời gian huấn luyện đã tạo nên sự phổ biến của DistilBERT và các mô hình rút gọn khác trong ứng dụng thực tế.

- **DistilBERT trong bài toán phân tích cảm xúc** DistilBERT đã được chứng minh là hiệu quả trong nhiều tác vụ xử lý ngôn ngữ tự nhiên (NLP), bao gồm cả phân tích cảm xúc [57]. Phân tích cảm xúc là tác vụ phân loại văn bản dựa trên cảm xúc hoặc quan điểm được bày tỏ trong nội dung văn bản đó [49]. Thông thường, phân tích cảm xúc được sử dụng để xác định liệu một đoạn văn bản có tính cảm tích cực, tiêu cực, hay trung lập. Cách dùng mô hình DistilBERT trong phân tích cảm xúc về cơ bản tương tự như BERT:
 - **Tiền xử lý dữ liệu:** Các văn bản cần được tiền xử lý, bao gồm loại bỏ ký tự đặc biệt, chuyển đổi chữ hoa thành chữ thường, tách từ và mã hóa văn bản dưới dạng token.
 - **Tinh chỉnh mô hình (Fine-tuning):** DistilBERT đã được tiền huấn luyện trên một lượng lớn dữ liệu văn bản [57]. Để áp dụng cho tác vụ phân tích cảm xúc, mô hình cần được tinh chỉnh trên một tập dữ liệu chứa các câu được gán nhãn quan điểm (tích cực, tiêu cực, trung lập) [64]. Quá trình này giúp mô hình học được cách liên kết các đặc trưng của văn bản với nhãn quan điểm tương ứng.

Nhờ sự linh hoạt, hiệu quả và nhẹ nhàng của DistilBERT, nhiều ứng dụng thực tế đã được triển khai thành công trong lĩnh vực phân tích cảm xúc [57]. Diễn hình là các ứng dụng liên quan đến:

- Phân tích cảm xúc của người dùng đối với sản phẩm, dịch vụ hoặc nội dung thông qua bình luận trên các trang web, diễn đàn và mạng xã hội.
- Phân tích ý kiến khách hàng trong khảo sát, phản hồi và đánh giá.
- Giám sát và phân tích cảm xúc của công chúng đối với chính sách, sự kiện hay vấn đề xã hội.
- Phân loại bài viết, tin tức hay bình luận dựa trên quan điểm để đưa ra đề xuất phù hợp cho người dùng.

DistilBERT trong phân tích cảm xúc cung cấp một giải pháp hiệu quả cho các doanh nghiệp và tổ chức muốn đánh giá và theo dõi cảm xúc của khách hàng và công chúng. Nhờ kích thước nhỏ gọn và tốc độ xử lý nhanh hơn, DistilBERT trở thành lựa chọn phù hợp cho các ứng dụng cần triển khai trên thiết bị di động hay các môi trường có giới hạn tài nguyên [57].

Với những ưu điểm của DistilBERT, nó đáp ứng được nhu cầu của các tổ chức và doanh nghiệp trong việc thu thập thông tin về thái độ và quan điểm của khách hàng cũng như đánh giá phản ứng của công chúng đối với các chính sách và sự kiện xã hội. Đặc biệt, DistilBERT còn giúp cải thiện hiệu suất của các hệ thống giới thiệu tin tức cá nhân, nhờ vào khả năng phân loại quan điểm của nó.

2.3 Trí tuệ nhân tạo giải thích được

2.3.1 Giới thiệu về trí tuệ nhân tạo giải thích được

Trí tuệ nhân tạo giải thích được (XAI) là một ngành của trí tuệ nhân tạo nhằm đạt được tính minh bạch và giải thích được các quyết định và hành động của mô hình AI. XAI giúp giải thích các quá trình hoạt động của AI, giúp con người hiểu rõ cách thức hoạt động, lý do của kết quả đưa ra, và đồng thời tạo sự tin tưởng vào AI [3]. Mục tiêu của XAI là đem lại sự minh bạch, tin cậy và giải thích cho các quyết định của AI, điều này rất quan trọng trong các lĩnh vực nhạy cảm như chẩn đoán y khoa, pháp lý và quyết định tài chính [24].

2.3.2 Tầm quan trọng của trí tuệ nhân tạo giải thích được

Việc đưa ra giải thích cho các mô hình AI ngày càng trở nên cần thiết, bởi vì:

- **Đáp ứng yêu cầu pháp lý:** Theo Điều 22 của GDPR (Luật bảo vệ dữ liệu chung của EU), các cá nhân có quyền không bị phân biệt đối xử dựa trên các quyết định hoàn toàn tự động.
- **Tăng tính tin cậy:** Giải thích các quyết định của AI giúp người dùng tin tưởng hơn vào kết quả và tiếp nhận dễ dàng hơn công nghệ AI [54].

- **Nâng cao hiệu quả:** Việc hiểu rõ hơn về hoạt động của mô hình AI giúp các nhà phát triển và chuyên gia trong lĩnh vực cải tiến và tối ưu hóa các mô hình hơn [24].

Tóm lại, XAI là một lĩnh vực quan trọng giúp tạo ra sự minh bạch và tin cậy trong quá trình ra quyết định của các mô hình AI. XAI bao gồm cả việc phát triển các mô hình học máy khả giải thích và kỹ thuật giải thích cho các mô hình hiện tại. Sự tiến bộ trong XAI đang được hỗ trợ bởi các công cụ và thư viện mã nguồn mở, giúp ngày càng nhiều nhà phát triển và người dùng tiếp cận và áp dụng XAI trong thực tế.

2.3.3 Phân loại

Các phương pháp giải thích máy học có thể được phân loại dựa trên nhiều tiêu chí khác nhau. Trong bài báo của Molnar (2019), tác giả phân loại các phương pháp giải thích máy học thành nhiều nhóm.

- **Intrinsic hay Post hoc:** Tiêu chí này phân biệt liệu khả năng giải thích có đạt được bằng cách hạn chế độ phức tạp của mô hình học máy (Intrinsic) hay bằng cách áp dụng các phương pháp phân tích mô hình sau khi huấn luyện (Post hoc). Khả năng giải thích Intrinsic đề cập đến các mô hình học máy có cấu trúc đơn giản, ví dụ như cây quyết định ít lớp hoặc mô hình tuyến tính thừa [54]. Khả năng giải thích Post hoc đề cập đến việc áp dụng phương pháp giải thích sau khi huấn luyện mô hình, ví dụ như tính điểm quan trọng của các đặc trưng (Feature Importance) [47].
- **Kết quả của phương pháp giải thích:** Các phương pháp giải thích khác nhau có thể phân biệt dựa trên kết quả của chúng [47], bao gồm:
 - **Thống kê tóm tắt feature:** Nhiều phương pháp giải thích cung cấp số liệu thống kê tóm tắt cho từng đặc trưng, ví dụ như điểm quan trọng (Importance Score) hoặc kết quả phức tạp hơn như mức độ tương tác giữa các cặp đặc trưng.
 - **Trực quan hóa tóm tắt features:** Hầu hết các số liệu thống kê tóm tắt đặc trưng cũng có thể được trực quan hóa.
 - **Model internals:** Phương pháp giải thích các mô hình dễ hiểu từ bên trong, ví dụ như trọng số trong mô hình tuyến tính hoặc cấu trúc cây quyết định (Features và ngưỡng được sử dụng để phân chia dữ liệu).

- **Điểm dữ liệu:** Các phương pháp trả về điểm dữ liệu (tồn tại hoặc mới tạo) để giúp mô hình dễ hiểu hơn, ví dụ như giải thích phản thực tế (Counterfactual Explanations) [69]. Để giải thích dự đoán của một điểm dữ liệu, phương pháp này tìm ra một điểm dữ liệu tương tự bằng cách thay đổi một số đặc trưng mà kết quả dự đoán thay đổi theo cách liên quan (ví dụ: lật lớp dự đoán) [69]. Một ví dụ khác là xác định các nguyên mẫu của các lớp dự đoán [47]. Để hữu ích, các phương pháp giải thích đầu ra các điểm dữ liệu mới yêu cầu bắn thân các điểm dữ liệu đó có thể được diễn giải. Điều này hoạt động tốt đối với hình ảnh và văn bản, nhưng ít hữu ích hơn đối với dữ liệu dạng bảng có hàng trăm đặc trưng.
- **Mô hình cụ thể hay mô hình bất khả tri:** Các công cụ giải thích dành riêng cho mô hình được giới hạn ở lớp mô hình cụ thể (Model specific) [47]. Ví dụ, việc giải thích các trọng số hồi quy trong một mô hình tuyến tính là một cách giải thích cụ thể theo mô hình, vì - theo định nghĩa - việc giải thích các giá trị nội tại của mô hình dễ hiểu luôn dành riêng cho mô hình [47]. Các công cụ chỉ hoạt động để giải thích cho mạng thần kinh cũng là mô hình cụ thể. Các công cụ không liên quan đến mô hình (Model Agnostic) có thể được sử dụng trên mọi mô hình học máy và được áp dụng sau khi mô hình đã được huấn luyện (Post hoc) [54]. Những phương pháp bất khả tri này thường hoạt động bằng cách phân tích các cặp đặc trưng đầu vào và đầu ra [47]. Theo định nghĩa, các phương pháp này không thể có truy cập vào nội bộ mô hình như trọng số (weights) hoặc thông tin cấu trúc [47].
- **Cục bộ hay toàn cục:** Liệu phương pháp giải thích có giải thích một dự đoán cá thể (Local) hoặc toàn bộ hành vi của mô hình (Global)? Hoặc là phạm vi ở đâu đó ở giữa? Các phương pháp giải thích cục bộ (Local Explanations) tập trung vào việc giải thích một dự đoán cá thể trong khi các phương pháp giải thích toàn cục (Global Explanations) cung cấp thông tin về cách thức hoạt động của mô hình trên toàn bộ dữ liệu [54, 47]. Ví dụ, LIME [54] và IG [65] cung cấp giải thích cục bộ cho từng dự đoán cá thể, trong khi phương pháp Feature Importance cung cấp thông tin toàn cục về tầm quan trọng của từng đặc trưng đối với mô hình [47]. Ngoài ra, cũng có các phương pháp giải thích ở mức độ trung gian, chẳng hạn như giải thích cụm (cluster-based explanations) hoặc giải thích phân lớp (class-based explanations [22]), trong đó giải thích được cung cấp cho một nhóm các điểm dữ liệu có đặc điểm tương tự hoặc thuộc cùng một lớp dự đoán.

Bằng cách phân loại các phương pháp giải thích dựa trên các tiêu chí như intrinsic hay post hoc, kết quả của phương pháp giải thích, mô hình cụ thể hay mô hình bất khả tri, và cục bộ hay toàn cục; ta có thể nắm bắt một cách hệ thống các công cụ và phương pháp hiện có để diễn giải các mô hình học máy, từ đó giúp đưa ra lựa chọn phù hợp cho từng bài toán cụ thể và đáp ứng nhu cầu giải thích của người dùng.

2.3.4 Một số kỹ thuật để giải thích mô hình trí tuệ nhân tạo

Có thể phân loại các nhóm kỹ thuật trong XAI thành 5 nhóm chính: **Độ quan trọng của đặc trưng**, **Mô hình đại diện**, **Giải thích bằng ví dụ**, **Giải thích bằng nguồn gốc** và **Suy diễn luật** [16].

- **Độ quan trọng của đặc trưng** (Feature Importance) là một nhóm kỹ thuật XAI giúp giải thích các quyết định của mô hình AI dựa trên độ quan trọng của các đặc trưng (features) trong dữ liệu đầu vào. Những kỹ thuật này nhằm xác định đặc trưng nào góp phần nhiều nhất vào quyết định cuối cùng của mô hình. Các phương pháp phổ biến trong nhóm này bao gồm:
 - **IG** (Tích hợp độ dốc) là một phương pháp tính Feature Importance bằng cách tính khoảng biến thiên của đầu ra mô hình khi thay đổi giá trị của từng feature[65]. Các feature gây ảnh hưởng lớn nhất đến đầu ra sẽ được coi là quan trọng nhất. Phương pháp này có ưu điểm là có thể áp dụng cho mọi loại mô hình ML, kể cả mô hình học sâu, và cho ra kết quả chính xác hơn so với các phương pháp đơn giản như điều chỉnh trọng số.
 - **Độ quan trọng của hoán vị** (Permutation Importance): Kỹ thuật này đánh giá độ quan trọng của mỗi đặc trưng bằng cách hoán vị giá trị của đặc trưng đó rồi đo lường sự thay đổi trong hiệu suất của mô hình sau hoán vị [8].
 - **Giảm độ tinh khiết trung bình** (Mean Decrease Impurity) / **Độ quan trọng dựa trên Gini** (Gini Importance): Đây là một phương pháp đánh giá độ quan trọng của đặc trưng trong các mô hình cây quyết định, bằng cách tính tổng giảm độ tinh khiết (impurity) trung bình khi đặc trưng được sử dụng để tách dữ liệu [7].
 - **Chuẩn hóa L1** (Least Absolute Shrinkage and Selection Operator) là một kỹ thuật điều chỉnh hồi quy tuyến tính (linear regression) giúp chọn

lọc và giảm độ quan trọng của các đặc trưng không liên quan bằng cách áp dụng ràng buộc L1 [67].

- **Loại bỏ đặc trưng tuần tự** (Recursive Feature Elimination hay RFE) là một kỹ thuật tiến hành loại bỏ tuần tự các đặc trưng ít quan trọng từ mô hình và đánh giá hiệu suất của mô hình sau mỗi lần loại bỏ, nhằm tìm ra tập con tối ưu của đặc trưng [27].
- **Giải thích giá trị thêm của Shapley:** Kỹ thuật này dựa trên khái niệm giá trị Shapley trong lý thuyết trò chơi hợp tác để đo lường đóng góp của mỗi đặc trưng trong dự đoán của mô hình. SHAP giúp giải thích các quyết định của mô hình thông qua việc phân tích độ quan trọng của từng đặc trưng trong mỗi trường hợp cụ thể [45].

Độ quan trọng của đặc trưng giúp người dùng hiểu rõ hơn về vai trò của từng đặc trưng trong việc đưa ra quyết định của mô hình AI. Điều này không chỉ hỗ trợ trong việc giải thích mô hình, mà còn giúp cải thiện chất lượng dữ liệu, tối ưu hóa mô hình và phát hiện các vấn đề tiềm ẩn như độ chênh (bias) hoặc các đặc trưng không liên quan. Tuy nhiên, cần lưu ý rằng các kỹ thuật Feature Importance chỉ giải thích vai trò của từng đặc trưng trong mô hình một cách độc lập, chưa đánh giá được sự tương tác giữa các đặc trưng. Để đánh giá các tương tác giữa các đặc trưng và giải thích mô hình một cách toàn diện hơn, có thể kết hợp với các kỹ thuật XAI khác.

- **Mô hình đại diện** (Surrogate Model) là một nhóm các kỹ thuật XAI dựa trên việc xây dựng một mô hình đơn giản hơn (ví dụ: cây quyết định, hồi quy tuyến tính, ...) để đại diện cho mô hình AI phức tạp (như mạng nơ-ron sâu). Mục đích của mô hình đại diện là giúp giải thích các quyết định của mô hình AI thông qua việc đưa ra các quy tắc hoặc đặc trưng dễ hiểu hơn. Các kỹ thuật phổ biến trong nhóm này bao gồm:
 - **Giải thích cục bộ về mô hình bất khả tri giải thích được** (LIME) là một kỹ thuật giải thích mô hình AI bằng cách xây dựng một mô hình tuyến tính đơn giản hơn để đại diện cho mô hình AI trong một khu vực cục bộ xung quanh điểm dữ liệu cần giải thích [54].
 - **Giải thích dựa trên quy tắc cục bộ** (Local Rule-based Explanations) là một phương pháp tạo ra các quy tắc địa phương dựa trên các cây quyết định để giải thích các quyết định của mô hình AI trong từng trường hợp cụ thể [25].

- **Mô neo** là một kỹ thuật giải thích mô hình AI dựa trên việc xác định các điều kiện (còn gọi là Anchors) trên các đặc trưng đầu vào, giúp dự đoán của mô hình trở nên ổn định và đáng tin cậy hơn trong một khu vực cục bộ [55].
- **Cây quyết định đại diện** (Decision Tree Surrogate) sử dụng cây quyết định (Decision Tree) như một mô hình đại diện để giải thích mô hình AI phức tạp. Mô hình cây quyết định sẽ được huấn luyện để đạt hiệu suất cao nhất khi dự đoán kết quả của mô hình AI gốc.

Mô hình đại diện giúp làm rõ quy trình ra quyết định của mô hình AI phức tạp thông qua việc đưa ra các quy tắc hoặc đặc trưng dễ hiểu hơn. Tuy nhiên, chúng có thể chỉ mang tính xấp xỉ và không hoàn toàn chính xác trong việc giải thích mô hình AI gốc. Do đó, khi sử dụng mô hình đại diện, cần cân nhắc đến độ chính xác và tính nhất quán của mô hình đại diện.

- **Giải thích bằng ví dụ** (Example-driven) là một nhóm kỹ thuật XAI dựa trên việc tìm kiếm và trình bày các ví dụ cụ thể từ dữ liệu huấn luyện để giải thích các quyết định của mô hình AI. Các kỹ thuật phổ biến trong nhóm này bao gồm:
 - **Giải thích bằng phản ví dụ** (Counterfactual Explanations): Counterfactual Explanations đưa ra các ví dụ dựa trên việc thay đổi nhỏ nhất các giá trị đầu vào, dẫn đến kết quả dự đoán khác biệt so với kết quả gốc [69].
 - **Mẫu và chỉ trích** (Prototype and Criticism): Kỹ thuật này đề xuất việc sử dụng các 'mẫu' (prototype) và 'chỉ trích' (criticism) để giải thích mô hình AI. Các mẫu là những ví dụ tiêu biểu đại diện cho một nhóm hoặc lớp, trong khi chỉ trích là những ví dụ đặc biệt khác biệt so với mẫu [36].
 - **Hàm ảnh hưởng** (Influence Functions) giải thích các dự đoán của mô hình AI bằng cách xác định các điểm dữ liệu quan trọng nhất đối với kết quả dự đoán. Các điểm dữ liệu này được sử dụng để xác định mức độ ảnh hưởng của từng điểm dữ liệu đối với kết quả dự đoán cuối cùng của mô hình. Các giải thích này có thể giúp người dùng hiểu tốt hơn về tầm quan trọng của từng điểm dữ liệu đối với kết quả dự đoán [38].

Example-driven giúp người dùng hiểu hơn về cách thức hoạt động của mô hình AI thông qua việc trình bày các ví dụ cụ thể. Tuy nhiên, cần lưu ý rằng việc trình bày quá nhiều ví dụ có thể khiến người dùng cảm thấy rối và khó tiếp cận.

- **Giải thích bằng nguồn gốc** (Provenance-based XAI) là một hướng tiếp cận giải thích quyết định của các mô hình học máy thông qua việc truy xuất nguồn gốc của dữ liệu và quá trình biến đổi dữ liệu. Mục đích chính của provenance-based XAI là đem lại sự minh bạch và tin cậy cho các mô hình AI, giúp người dùng hiểu rõ hơn về cách mà các quyết định được đưa ra và hệ thống học máy sử dụng dữ liệu như thế nào để đưa ra dự đoán. Các kỹ thuật phổ biến trong Provenance-based XAI bao gồm:
 - **Nguồn gốc dữ liệu** (Data Provenance) theo dõi nguồn gốc của dữ liệu đầu vào và quá trình tiền xử lý (pre-processing) để đảm bảo chất lượng và tính phù hợp của dữ liệu sử dụng trong mô hình học máy. Data provenance giúp cho việc xác định các yếu tố gây nhiễu, thiếu sót trong dữ liệu, đồng thời giúp đánh giá và kiểm soát tác động của dữ liệu đầu vào lên kết quả đầu ra.
 - **Nguồn gốc mô hình** (Model Provenance) theo dõi và lưu trữ thông tin về quá trình huấn luyện, kiến trúc mô hình, thuật toán học, và các siêu tham số (hyperparameters) của mô hình học máy [11]. Model provenance hỗ trợ việc tái tạo kết quả, so sánh hiệu suất giữa các mô hình và giúp hiểu các yếu tố ảnh hưởng đến hiệu suất của mô hình.
 - **Nguồn gốc đặc trưng** (Feature Provenance) phân tích sự phụ thuộc giữa các đặc trưng (features) đầu vào và kết quả đầu ra của mô hình, giúp người dùng hiểu rõ hơn về vai trò của các đặc trưng trong việc đưa ra quyết định [37]. Một ví dụ điển hình về kỹ thuật này là tính toán độ quan trọng của các đặc trưng (feature importance) trong mô hình cây quyết định hoặc mô hình học sâu.
- **Suy diễn luật** (Declarative Induction) trong XAI là một phương pháp giúp tăng cường khả năng giải thích cho các mô hình học máy thông qua việc suy diễn luật từ dữ liệu huấn luyện. Điều này giúp người dùng hiểu được quy trình ra quyết định của mô hình và đánh giá tính đúng đắn của nó. Declarative Induction đã trải qua nhiều sự phát triển và cải tiến trong hơn ba thập kỷ qua. Trong kỹ thuật này, hệ thống học máy sử dụng một loạt các quy tắc được xây dựng từ dữ liệu huấn luyện. Những quy tắc này được đưa ra dưới dạng các luận đề (premises) và kết luận (conclusions). Một quy tắc cụ thể được áp dụng khi tất cả các luận đề của nó được thỏa mãn. Điều này giúp người dùng dễ dàng theo dõi quá trình ra quyết định của mô hình. Các kỹ thuật phổ biến

trong Declarative Induction bao gồm một số phương pháp sau:

- **Cây quyết định** (Decision Trees): Đây là một trong những phương pháp phổ biến nhất trong Declarative Induction. Cây quyết định biểu diễn các quyết định dưới dạng cây, với các nút đại diện cho thuộc tính và các nhánh đại diện cho giá trị của thuộc tính đó. Các nút lá chứa kết luận hoặc dự đoán của mô hình. Các thuật toán phổ biến để xây dựng cây quyết định bao gồm ID3, C4.5 và CART.
- **Học dựa trên quy tắc** (Rule-based Learning): Trong kỹ thuật này, mô hình học một tập các luật có dạng IF-THEN từ dữ liệu. Các luật này được áp dụng tuần tự hoặc song song để đưa ra quyết định hoặc dự đoán. Một số thuật toán học dựa trên luật phổ biến bao gồm RIPPER, PRISM, FOIL và CN2.
- **Lập trình Logic Ước lượng** (Inductive Logic Programming) là một phương pháp kết hợp giữa logic mệnh đề (propositional logic) và logic bậc 1 (first-order logic). Mô hình học từ dữ liệu dưới dạng các mệnh đề logic và tìm ra các luật suy diễn (rules of inference) trong một ngôn ngữ logic cho trước. Một số công cụ và thuật toán phổ biến trong ILP bao gồm Progol, FOIL và Aleph.
- **Học luật kết hợp** (Association Rule Learning): Phương pháp này tìm kiếm các mối liên quan mạnh giữa các biến trong dữ liệu và xây dựng các luật kết hợp dưới dạng 'IF X THEN Y'. Thuật toán Apriori và FP-Growth là hai thuật toán phổ biến trong học luật kết hợp.
- **Thuật toán di truyền** (Genetic Algorithms): Trong Genetic Algorithms, mô hình học từ dữ liệu bằng cách sử dụng các kỹ thuật di truyền như lai ghép, đột biến và lựa chọn tự nhiên để tìm kiếm các giải pháp tối ưu trong không gian giải pháp. Genetic Algorithms thường được áp dụng để học các luật hoặc tham số của mô hình.

Một số ứng dụng của kỹ thuật Declarative Induction trong XAI bao gồm:

- **Phân loại:** Hệ thống có thể học từ dữ liệu với nhãn và đưa ra quyết định dựa trên các quy tắc suy diễn.
- **Khai phá luật kết hợp:** Hệ thống có thể tìm ra mối liên hệ giữa các thuộc tính trong dữ liệu và xác định các quy tắc phổ biến trong tập dữ liệu.

- **Dự báo:** Hệ thống có thể sử dụng các quy tắc được học từ dữ liệu để dự đoán giá trị của một biến mục tiêu.

2.3.5 Thách thức và hạn chế của trí tuệ nhân tạo giải thích được

Mặc dù XAI đã đạt được nhiều thành tựu, vẫn còn tồn tại các thách thức và hạn chế như:

- **Đánh đổi giữa độ chính xác và khả năng giải thích:** Các mô hình AI phức tạp như mạng nơ-ron sâu thường khó giải thích hơn các mô hình đơn giản hơn như cây quyết định. Tuy nhiên, các mô hình phức tạp thường cho độ chính xác cao hơn [56].
- **Ngôn ngữ giải thích phù hợp:** Việc chọn ngôn ngữ và cách thức trình bày giải thích phù hợp với người dùng là một thách thức lớn, đòi hỏi sự hiểu biết sâu sắc về nhu cầu và kiến thức của người dùng [46].
- **Tính nhất quán giữa các giải thích:** Tính nhất quán giữa các giải thích: Các phương pháp giải thích khác nhau có thể đưa ra các giải thích đối lập cho cùng một kết quả, gây khó khăn trong việc đánh giá và lựa chọn giải thích chính xác [1].

2.3.6 Tương lai của trí tuệ nhân tạo giải thích được

Tương lai của XAI đầy hứa hẹn, với nhiều hướng nghiên cứu và ứng dụng tiềm năng:

- **Nâng cao khả năng giải thích của mô hình AI:** Các nhà nghiên cứu tiếp tục tìm kiếm phương pháp giải thích mới và cải tiến các phương pháp hiện có để đạt được độ minh bạch cao hơn cho mô hình AI [26].
- **Kết hợp giữa XAI và các lĩnh vực khác:** Việc kết hợp XAI với các lĩnh vực khác như bảo mật, đạo đức AI, và quản lý rủi ro sẽ giúp tạo ra các ứng dụng AI toàn diện và bền vững hơn [9].
- **Phát triển XAI dựa trên ngữ cảnh và người dùng:** Tùy chỉnh giải thích cho từng người dùng và ngữ cảnh sẽ giúp XAI trở nên linh hoạt và hiệu quả hơn trong việc đáp ứng nhu cầu và kiến thức của người dùng [31].

- **Ứng dụng XAI trong các ngành công nghiệp:** Việc áp dụng XAI trong các ngành công nghiệp khác nhau như y tế, tài chính, và tự động hóa sẽ giúp đánh giá hiệu quả của XAI trong thực tiễn và mở rộng lĩnh vực ứng dụng [12].

Tóm lại, XAI là một lĩnh vực đầy tiềm năng và thách thức, giúp con người hiểu rõ hơn về hoạt động của các mô hình AI. Việc nghiên cứu và phát triển các phương pháp giải thích trong XAI sẽ không chỉ đáp ứng yêu cầu pháp lý và tăng sự tin tưởng của người dùng, mà còn giúp cải tiến và tối ưu hóa hiệu quả của các mô hình AI trong tương lai.

2.4 Integrated Gradient

2.4.1 Định nghĩa

IG là một phương pháp giải thích cho các mô hình học sâu, đặc biệt là mạng nơ-ron. Phương pháp này đo lường tác động của từng đặc trưng trong một mô hình học sâu bằng cách tính đạo hàm riêng của kết quả đối với giá trị của đặc trưng đó dọc theo một đường kết nối điểm dữ liệu cần giải thích và một điểm dữ liệu cơ sở [65].

Để tính IG, ta cần thực hiện các bước sau:

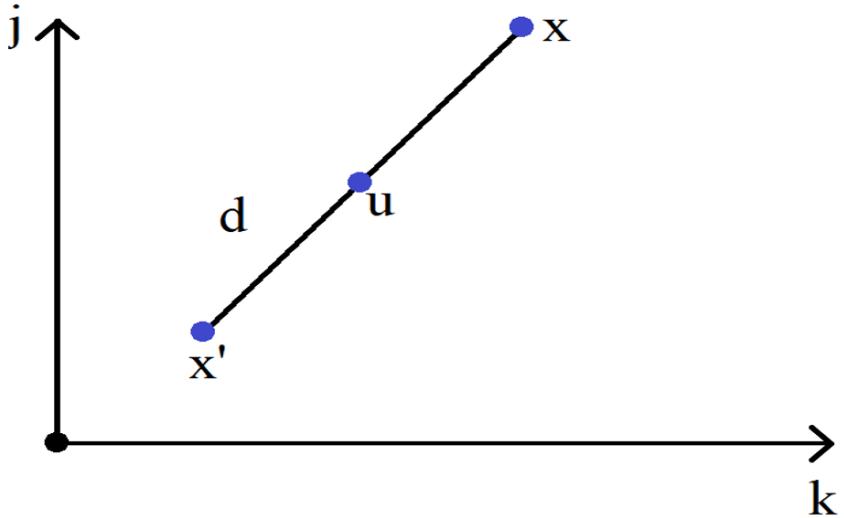
- Chọn một điểm dữ liệu cơ sở (baseline). Điểm dữ liệu cơ sở thường là một điểm không có thông tin hoặc kết quả dự đoán bằng giá trị trung bình.
- Tính đạo hàm riêng của kết quả đối với mỗi đặc trưng dọc theo đường nối giữa điểm dữ liệu và điểm dữ liệu cơ sở.
- Tích hợp (cộng dồn) đạo hàm riêng trên đường nối để tính IG cho mỗi đặc trưng.

Công thức của IG tính cho sự phân bổ của đầu ra mô hình vào feature thứ i của điểm dữ liệu đầu vào (tạm gọi là a_i) được tính như sau:

$$a_i = \phi_i(f, x, x') = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \quad (2.1)$$

Trong đó:

- $f(x)$ là đầu ra của mô hình tại nhãn một nhãn bất kì cần xét với đầu vào của mô hình tại x .



Hình 2.4: Giải thích công thức của IG

- x' là điểm dữ liệu cơ sở
- $\int_{\alpha=0}^1$ là phép tích phân từ 0 đến 1 theo biến α .

Ở đây ta sẽ cần làm rõ công thức:

- Biểu thức đang tính gì?
- Tại sao lại cần điểm dữ liệu cơ sở?

2.4.2 Biểu thức đang tính gì

Để dễ hình dung, ta coi dữ liệu đầu vào chỉ có 2 feature, khi đó có thể biểu diễn x và x' trên mặt phẳng tọa độ Okj , với d là đoạn thẳng nối x và x' .

Xét điểm u nằm trên đoạn d (hình 2.4) có tỉ lệ khoảng cách là $\alpha = \frac{u-x'}{x-x'}$, do đó:

$$u - x' = \alpha(x - x') \Leftrightarrow u = x + \alpha(x - x') \quad (2.2)$$

Vì phân 2 vế, ta được:

$$du = (x - x')d\alpha \Rightarrow du_i = (x_i - x'_i)d\alpha \quad (2.3)$$

Khi đó α sẽ trải dài từ 0 (khi $u \equiv x'$) đến 1 (khi $u \equiv x$), quay lại với biểu thức tính IG, ta có:

$$\phi_i(f, x, x') = (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (2.4)$$

$$= (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial f(u)}{\partial x_i} d\alpha \quad (2.5)$$

$$= \int_{\alpha=0}^1 \frac{\partial f(u)}{\partial x_i} (x_i - x'_i) d\alpha \quad (2.6)$$

$$= \int_{u_i=x'_i}^{x_i} \frac{\partial f(u)}{\partial x_i} du_i \quad (2.7)$$

Ta có thể hiểu $\frac{\partial f(u)}{\partial x_i}$ là biểu thức tính độ dốc (gradient) của đầu ra mô hình ưng với điểm dữ liệu đầu vào u nằm trên đoạn thẳng nối x' và x theo feature thứ i của dữ liệu đầu vào x . Do đó tổng tất cả các gradient của đầu ra theo tất cả các dữ liệu đầu vào nằm trên d theo feature i của dữ liệu đầu vào x sẽ là:

$$\int_{u_i=x'_i}^{x_i} \frac{\partial f(u)}{\partial x_i} du_i \quad (2.8)$$

Tóm lại ta có thể hiểu công thức tính IG như sau: Để tính điểm số quan trọng của đặc trưng i trong đầu vào, ta sẽ theo dõi sự thay đổi đầu ra (Gradient) khi giá trị của đặc trưng i đó thay đổi từ dữ liệu cơ sở đến dữ liệu đầu vào. Sau đó tích hợp tất cả các độ dốc đó bằng phép toán tích phân. Cuối cùng ta nhân kết quả đó với khoảng cách giữa dữ liệu cơ sở và dữ liệu đầu vào (tính theo đặc trưng i) sẽ ra được điểm số quan trọng của đặc trưng i .

2.4.3 Nguồn gốc và vai trò của điểm dữ liệu cơ sở

Nguồn gốc của điểm dữ liệu cơ sở bắt đầu đến từ học thuyết trò chơi của Shapley [61]. Học thuyết đặt vấn đề rằng: Cho một nhóm người tham gia một trò chơi. Sau khi kết thúc trò chơi, cả nhóm chiến thắng và giành được phần thưởng. Câu hỏi được đặt ra là: Phần thưởng ấy nên chia cho mỗi người như thế nào?

Tác giả học thuyết-Shapley cho rằng: Sẽ tiến hành chia nhóm đó thành các nhóm nhỏ hơn (có thể giao nhau) tham gia lại trò chơi. Sau khi kết thúc trò chơi, lượng phần thưởng giảm sút đi sẽ tỉ lệ thuận với đóng góp của một phần nhóm không tham gia vào trò chơi. Để cho dễ hiểu, giả sử nhóm gồm 10 người tham gia, phần thưởng nhóm sau khi trò chơi kết thúc đạt được là X. Bây giờ loại bỏ người thứ i, chỉ cho 9 người còn lại tham gia, phần thưởng họ nhận được bị sụt đi so với X một khoảng X_i . Mặt khác, khi loại bỏ người thứ j khác i, cho 9 người chơi còn

lại tham gia, phần thưởng họ nhận được bị sụt đi so với X một khoảng $X_j < X_i$. Ta có thể nhận định rằng: Đóng góp của người i vào trò chơi lớn hơn so với đóng góp của người j . Do đó, ta có thể hiểu là: Muốn xem một người đóng góp vào trò chơi nhiều hay ít so với những người còn lại, ta chỉ đơn giản bỏ người đó ra khỏi nhóm, cho tất cả những người còn lại tham gia trò chơi, xem lượng phần thưởng bị sụt giảm đi nhiều hay ít so với trường hợp loại bỏ người khác khỏi trò chơi.

Khi áp dụng vào lĩnh vực học máy, để kiểm tra xem sự “quan trọng” của đặc trưng đầu vào trong việc giúp mô hình đưa ra dự đoán, ta đơn giản là “loại bỏ” đặc trưng ấy đi, xem đầu ra của dự đoán giảm đi bao nhiêu, sự “quan trọng” của đặc trưng ấy sẽ tỉ lệ thuận qua sự sụt giảm đầu ra của dự đoán so với khi chưa loại bỏ là nhiều hay ít.

Shapley cũng đã đưa ra thuật toán để giải quyết vấn đề “chia thưởng” này, tuy nhiên thuật toán ấy khi đưa vào học máy lại trở nên tốn rất nhiều chi phí về thời gian và bộ nhớ. Do vậy, IG đã được ra đời để khắc phục nhược điểm của thuật toán này. Trong khi SHAP cần “loại bỏ” đặc trưng ấy ra khỏi dữ liệu đầu vào để xem đầu ra mô hình bị sụt giảm đi bao nhiêu mới có thể đánh giá độ quan trọng của một đặc trưng trong việc giúp mô hình đưa ra dự đoán, thì IG không làm vậy. Chúng ta đều biết, bản chất các dữ liệu đầu vào khi đưa vào mô hình đều là các ma trận, việc loại bỏ một đặc trưng (1 phần tử ma trận) đi gần như là việc không thể, nó sẽ ảnh hưởng tới cấu trúc ma trận, và người lập trình sẽ phải định hình lại các phân bố các phần tử để ma trận đầu vào không mất đi cấu trúc vốn có của nó. Thoáng qua khi đến đây, mọi người thường đưa ra ý kiến rằng, nên thay giá trị tại vị trí phần tử đó thành 0. Nhưng với sự đa dạng của dữ liệu hiện nay, 0 có phải là một giá trị thay thế đặc trưng tối ưu hay không. Từ đó, điểm dữ liệu cơ sở ra đời như một câu trả lời cho vấn đề này.

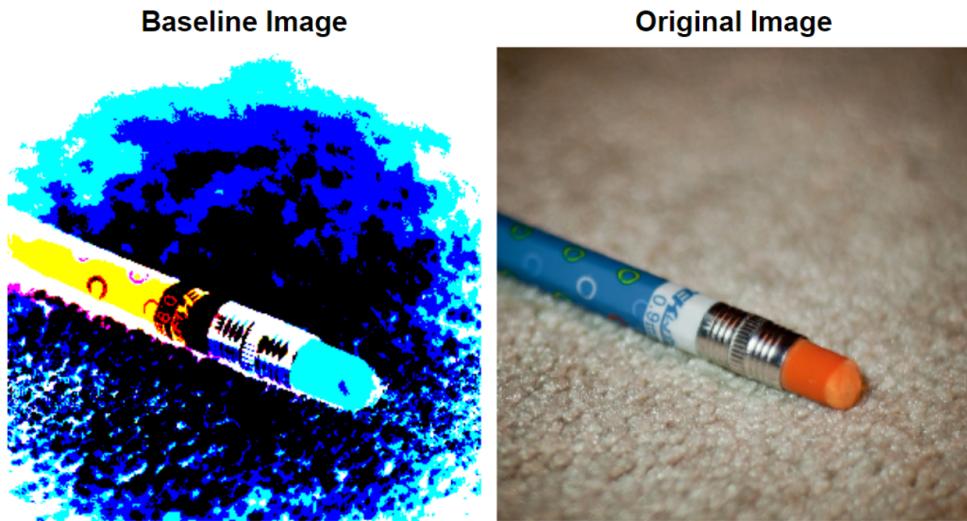
2.4.4 Một số cách chọn điểm dữ liệu cơ sở

Về bản chất của IG, điểm dữ liệu cơ sở có thể được chọn là một giá trị tùy ý (cùng kích cỡ với đầu vào). Phương pháp sơ cấp nhất thường được mọi người chọn cho điểm dữ liệu cơ sở dữ liệu đầu vào đó là véc-tơ 0 (điểm dữ liệu cơ sở tĩnh). Ngoài ra, có một số phương pháp chọn điểm dữ liệu cơ sở động khác như: **điểm dữ liệu cơ sở xa nhất**, **điểm dữ liệu cơ sở mập mờ**, **điểm dữ liệu cơ sở phân phôi đều**, **điểm dữ liệu cơ sở gaussian**.

Trước khi đề cập từng phương pháp, ta coi các dữ liệu đầu vào khi cho vào

mô hình là các véc-tơ được mã hóa từ các từ trong câu theo dictionary được tạo từ data, mỗi từ được mã hóa là một số tự nhiên không lớn hơn kích thước từ điển.

- **Điểm dữ liệu cơ sở xa nhất** (Maximum baseline): Phương pháp này được chọn sao cho khoảng cách L1 từ điểm dữ liệu cơ sở đến dữ liệu đầu vào gốc là lớn nhất, nhưng các giá trị phần tử trên điểm dữ liệu cơ sở vẫn nằm trong đoạn [0, kích thước từ điển].



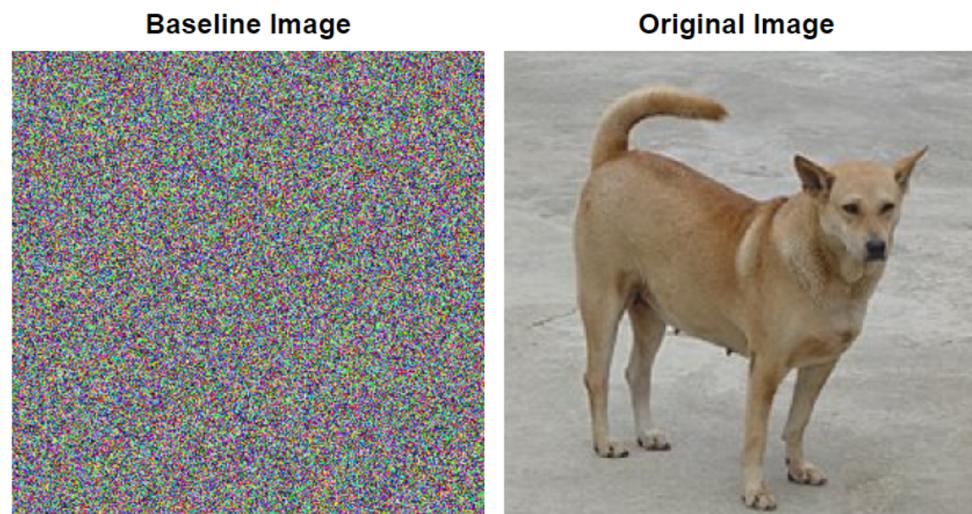
Hình 2.5: Trực quan hóa Maximum baseline

- **Điểm dữ liệu cơ sở mập mờ** (Blur baseline): Phương pháp này được chọn bằng cách tại một phần tử trên dữ liệu đầu vào, ta thay giá trị của nó bằng trung bình cộng giá trị các phần tử cạnh nó. Phương pháp này có một nhược điểm lớn đó là nó bị bias với các cụm các phần tử kề nhau với giá trị chênh lệch nhau lớn. Để làm rõ hơn ta giả sử dữ liệu đầu vào đầu vào là $[0, 1, 2, 8, 2, 10]$. Áp dụng phương pháp với điểm dữ liệu cơ sở mập mờ với khoảng cách blurr là 1. Ta thu được điểm dữ liệu cơ sở dữ liệu đầu vào là $[0.50, 1.00, 3.67, 4.00, 7.33, 6.00]$, ta có thể thấy, giá trị phần tử thứ 2 của dữ liệu đầu vào cơ sở là 1.00 đúng bằng giá trị tương ứng tại dữ liệu đầu vào, dẫn đến phân bổ IG cho phần tử thứ nhất sẽ bằng 0, trong khi giá trị điểm dữ liệu cơ sở tại phần tử thứ 5 thu được là 7.33, cách giá trị tương ứng tại dữ liệu đầu vào gốc là 5.33, một khoảng khá lớn, việc tính IG cho phần tử này sẽ tăng bias lên so với các phần tử khác, do công thức tính IG cho phần tử thứ i chứa nhân tử $x_i - x'_i$. Tóm lại, với các phần tử có giá trị gần với giá trị các phần tử xung quanh thì IG cho phần tử này sẽ bị bias thấp và ngược lại.
- **Điểm dữ liệu cơ sở phân phối đều** (Uniform baseline): Phương pháp này



Hình 2.6: Trực quan hóa Blur baseline

được chọn bằng cách lấy giá trị trên các phần tử của dữ liệu đầu vào theo một phân bố đều với phạm vi như phạm vi của dữ liệu đầu vào ($[0, 256]$ với ảnh, $[0, \text{kích thước từ điển}]$ đối với văn bản).



Hình 2.7: Trực quan hóa Uniform baseline

- **Điểm dữ liệu cơ sở Gaussian** (Gaussian baseline): Phương pháp này là phương pháp tổng quát hơn so với Điểm dữ liệu cơ sở phân phối đều, khi các phần tử trên điểm dữ liệu cơ sở được lấy tuân theo phân phối chuẩn với kì vọng đúng bằng giá trị phần tử tương ứng bên dữ liệu đầu vào, độ lệch chuẩn tại các phần tử trên điểm dữ liệu cơ sở là như nhau.



Hình 2.8: Trực quan hóa Gaussian baseline

2.4.5 Ưu điểm và nhược điểm của Integrated Gradient

Ưu điểm:

- IG áp dụng được cho nhiều loại mô hình học sâu, như mạng nơ-ron tích chập (CNN) và mạng nơ-ron truy hồi (RNN) [65].
- Phương pháp này cung cấp mức độ quan trọng của từng đặc trưng trong kết quả dự đoán, giúp hiểu được tác động của các đặc trưng lên kết quả của mô hình [65].

Nhược điểm:

- Tính toán đòi hỏi nhiều tài nguyên và thời gian, đặc biệt khi xử lý dữ liệu lớn hoặc mô hình phức tạp [65].
- Kết quả của IG có thể bị ảnh hưởng bởi sự lựa chọn của điểm dữ liệu cơ sở. Tuy nhiên, điểm này có thể được giải quyết bằng cách chọn điểm dữ liệu cơ sở phù hợp hoặc sử dụng nhiều điểm dữ liệu cơ sở và lấy trung bình kết quả [65].

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

Chương này đề xuất một hệ thống có thể giải thích mô hình phân tích cảm xúc. Hệ thống này bao gồm mô hình được sử dụng trong bài toán phân tích cảm xúc và phương pháp để giải thích cách hoạt động của mô hình đó. Hệ thống gồm 3 phần: Xây dựng mô hình, kiểm thử mô hình và giải thích mô hình.

Hệ thống này sẽ chọn ra 2 mô hình tiêu biểu trong các bài toán về phân tích cảm xúc: BERT và DistilBERT, sau đó sẽ dùng Integrated Gradients để giải thích hành vi của 2 mô hình này

3.1 Xây dựng mô hình BERT

Mô hình BERT-base là một phiên bản BERT đã được huấn luyện trước¹ và được cung cấp thông qua thư viện Transformers của Hugging Face. Mô hình này có 12 tầng Transformer, 768 đơn vị ẩn (hidden layer) và 12 head attention, tổng cộng khoảng 110 triệu tham số.

Mô hình này được huấn luyện trên một tập dữ liệu lớn bao gồm BooksCorpus (800 triệu từ) và English Wikipedia (2,500 triệu từ). Quá trình huấn luyện được thực hiện bằng cách 'che' một số từ ngẫu nhiên trong câu và yêu cầu mô hình dự đoán chúng dựa trên ngữ cảnh, cùng với một số tác vụ khác.

Do tập dữ liệu hiện tại của BERT có nội dung rất đa dạng, do đó ta sẽ tinh chỉnh lại (fine-tune) mô hình này bằng cách huấn luyện lại với một tập dữ liệu khác.

¹https://huggingface.co/google/bert_uncased_L-12_H-768_A-12

3.1.1 Tập dữ liệu

Để phù hợp với bài toán phân tích cảm xúc, ta dùng bộ dữ liệu đánh giá phim từ IMDB (Internet Movie Database)¹. Đây là một tập hợp lớn chứa 50,000 đánh giá phim trực tuyến. Bộ dữ liệu này rất đa dạng và phong phú, đại diện cho một loạt các thể loại phim và phản hồi của người xem từ khắp nơi trên thế giới.

Mỗi đánh giá trong bộ dữ liệu là một đoạn văn tự nhiên, thường có độ dài từ một đến một vài đoạn, đánh giá một bộ phim cụ thể. Những đánh giá này đều đã được người dùng gửi lên IMDB và bao gồm cả đánh giá chuyên nghiệp cũng như đánh giá của người dùng thông thường.

Các đánh giá được gán nhãn là 'tích cực' hoặc 'tiêu cực' dựa trên số lượng sao mà người dùng đánh giá cho phim. Một đánh giá được xem là 'tích cực' nếu phim nhận được 7 sao trở lên và được xem là 'tiêu cực' nếu phim nhận được ít hơn 4 sao. Đánh giá từ 4 đến 6 sao không được sử dụng trong bộ dữ liệu này để đảm bảo rằng sự phân biệt giữa 'tích cực' và 'tiêu cực' rõ ràng.

Bộ dữ liệu được chia đều, với 25,000 đánh giá cho mục đích huấn luyện và 25,000 đánh giá còn lại dành cho việc thử nghiệm. Trong mỗi tập, số lượng các đánh giá tích cực và tiêu cực là ngang nhau, giúp đảm bảo tính công bằng khi huấn luyện và kiểm tra mô hình.

3.1.2 Bộ mã hóa từ

Bộ mã hóa từ WordPiece là một phương pháp phổ biến để mã hóa văn bản trong các mô hình học máy, đặc biệt là trong các mô hình ngôn ngữ dựa trên transformer như BERT.

WordPiece tiếp cận việc mã hóa từ bằng cách chia từ ra thành các 'mảnh' nhỏ hơn. Ban đầu, mỗi từ được coi là một mảnh riêng biệt. Sau đó, thuật toán tìm kiếm cặp ký tự liên tiếp xuất hiện nhiều nhất và gộp chúng lại thành một mảnh. Quá trình này lặp lại cho đến khi đạt đến số lượng mảnh tối đa được chỉ định trước hoặc không còn cặp ký tự nào để gộp.

Giả sử chúng ta có từ 'unhappiness'. Ban đầu, từ này có thể được chia thành các mảnh 'un', '##happi', '##ness'. Khi một từ được chia, mảnh đầu tiên giữ nguyên và các mảnh sau được thêm '##' ở đầu để biểu thị rằng chúng là phần

¹<https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>

của từ.

Một trong những lợi ích chính của WordPiece là nó cho phép mô hình xử lý một cách linh hoạt với các từ không thường xuất hiện hoặc không có trong từ điển. Thay vì chỉ định một token đặc biệt cho các từ không xác định (như UNK trong các mô hình từ ngữ trước đó), WordPiece cho phép mô hình phân rã từ không xác định thành các mảnh mà nó đã biết.

WordPiece được sử dụng rộng rãi trong các mô hình ngôn ngữ transformer như BERT, và cũng được sử dụng trong các mô hình dịch máy như Google's Neural Machine Translation system.

3.1.3 Huấn luyện mô hình

Khởi tạo tham số:

- *Learning_rate* (tốc độ học): 0.00002. Đây là tham số phù hợp với thuật toán tối ưu hóa Adam.
- *Epoch*: 1. Lý do chọn epoch sẽ được giải thích sau khi huấn luyện mô hình.
- *Batch_size* (kích thước lô): 4.
- *Sequence_length* (độ dài chuỗi): 512 do mô hình BERT thường sử dụng tối đa 512 token.

Xác định thuật toán tối ưu hóa và hàm mất mát phù hợp với mô hình phân loại 2 lớp:

- Thuật toán tối ưu hóa: Adam, với *Learning_rate* đã được xác định ở trên.
- Hàm mất mát: tf.keras.losses.BinaryCrossentropy
- Metric (đo lường hiệu suất): tf.keras.metrics.BinaryAccuracy

Biên dịch mô hình:

```
model.compile(  
    optimizer = tf.keras.optimizers.Adam(learning_rate=2e-5, epsilon=1e-8),  
    loss = tf.keras.losses.BinaryCrossentropy(from_logits=True),  
    metrics = [tf.keras.metrics.BinaryAccuracy('accuracy')])
```

Huấn luyện mô hình:

```
bert_history = model.fit(  
    ds_train_encoded,  
    epochs = 1,  
    validation_data = ds_test_encoded  
)
```

Sau khi huấn luyện, ta có thể thấy trong hình 3.9, đã có dấu hiệu của overfitting, dù chỉ sau 1 epoch. Điều này có thể khiến mô hình không đạt được trạng thái hoạt động tốt nhất

```
loss, accuracy = model.evaluate(ds_test_encoded)  
  
print(f'Test Loss: {loss}')  
print(f'Test Accuracy: {accuracy}')  
  
6250/6250 [=====] - 166s 26ms/step - loss: 0.1647 - accuracy: 0.9385  
Test Loss: 0.16468968987464905  
Test Accuracy: 0.938480019569397
```

```
loss, accuracy = model.evaluate(ds_train_encoded)  
  
print(f'Train Loss: {loss}')  
print(f'Train Accuracy: {accuracy}')  
  
6250/6250 [=====] - 166s 26ms/step - loss: 0.0969 - accuracy: 0.9698  
Train Loss: 0.09694787859916687  
Train Accuracy: 0.969760000705719
```

Hình 3.9: Chỉ số của BERT sau khi huấn luyện

3.2 Xây dựng mô hình DistilBERT

DistilBERT là một mô hình học máy dựa trên kiến trúc của BERT, một mô hình học máy dựa trên Transformer được phát triển bởi Google. DistilBERT, như tên gọi của nó (kết hợp từ 'distill' và 'BERT'), là phiên bản 'tinh giản' của BERT, được tạo ra nhằm giảm độ phức tạp của mô hình gốc mà vẫn giữ được một mức độ hiệu suất cao.

DistilBERT giữ lại khoảng 95% hiệu suất của BERT gốc, nhưng chỉ có khoảng 60% số tham số và kích thước của BERT. Điều này giúp tăng tốc độ suy luận và giảm bớt yêu cầu về bộ nhớ và khả năng tính toán khi sử dụng mô hình.

Để tạo ra DistilBERT, các nhà nghiên cứu đã áp dụng kỹ thuật gọi là 'knowledge distillation'. Trong quá trình này, một mô hình lớn (ở đây là BERT) được dùng như một 'giáo viên' để hướng dẫn một mô hình nhỏ hơn (ở đây là DistilBERT) cách tạo ra các dự đoán tốt. Cụ thể, mô hình nhỏ hơn được huấn luyện để mô phỏng cách 'giáo viên' đưa ra dự đoán, thay vì cố gắng dự đoán trực tiếp từ dữ liệu. Về mặt kiến trúc, DistilBERT có 6 tầng Transformer (so với 12 tầng của BERT-base), 768 đơn vị ẩn và 12 head attention, tổng cộng khoảng 66 triệu tham số.

Trong các mô hình được huấn luyện trước, mô hình distilBERT-base-uncased-finetuned-sst-2-english¹ (để cho tiện, ta gọi tắt là DistilBERT vì phạm vi khóa luận chỉ bao đến phiên bản này của DistilBERT) là một lựa chọn phù hợp cho bài toán phân tích quan điểm.

3.2.1 Tập dữ liệu

Mô hình DistilBERT này sử dụng bộ dữ liệu SST-2 (Stanford Sentiment Treebank-2), là một tập dữ liệu được phát triển bởi nhóm nghiên cứu tại Đại học Stanford. Tập dữ liệu này được xây dựng nhằm mục đích phục vụ cho bài toán phân loại cảm xúc trong xử lý ngôn ngữ tự nhiên. Đó là lý do ta chọn mô hình này làm thí nghiệm để giải thích bằng Integrated Gradients

Nguồn gốc của SST-2 xuất phát từ tập dữ liệu ban đầu là SST-1, (Stanford Sentiment Treebank-1), được lấy từ dự án Treebank của Đại học Pennsylvania. Tuy nhiên, SST-2 tập trung vào việc phân loại câu thành hai nhãn: 'positive' (tích cực) và 'negative' (tiêu cực), trong khi SST-1 có nhiều nhãn chi tiết hơn.

Dữ liệu trong SST-2 chủ yếu bao gồm các câu mẫu được thu thập từ các nguồn tin tức và đánh giá phim. Các câu mẫu này phản ánh ý kiến hoặc cảm nhận của người viết đối với một chủ đề hoặc sự kiện cụ thể.

Việc gán nhãn trong SST-2 được thực hiện bằng cách thuê người đánh giá chuyên nghiệp để đánh giá cảm xúc của mỗi câu. Những người đánh giá này được yêu cầu gán nhãn câu theo hai nhãn 'positive' hoặc 'negative' dựa trên cảm nhận của họ về câu đó. Quá trình gán nhãn được thực hiện với sự hướng dẫn và kiểm soát chặt chẽ để đảm bảo tính đáng tin cậy của tập dữ liệu.

¹<https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english>

Mô hình DistilBERT ở đây sử dụng bộ dữ liệu SST-2 từ Hugging Face¹, tập dữ liệu này gồm ba phần chính: Train, Validation và Test.

- Tập dữ liệu Train: Đây là tập dữ liệu được sử dụng để huấn luyện mô hình. Tập dữ liệu Train của SST-2 chứa 67,349 câu, bao gồm 37569 câu tích cực và 29780 câu tiêu cực
- Tập dữ liệu Validation: Được sử dụng để kiểm tra hiệu suất của mô hình trong quá trình huấn luyện và tinh chỉnh siêu tham số. Tập dữ liệu Validation của SST-2 chứa 872 câu, bao gồm 444 câu tích cực và 428 câu tiêu cực
- Tập dữ liệu Test: Được sử dụng để đánh giá độ chính xác của mô hình đã được huấn luyện. Tập dữ liệu Test của SST-2 chứa 1,821 câu.

3.2.2 Huấn luyện mô hình

Mô hình DistilBERT được tinh chỉnh bởi các tham số như sau:

- *Learning_rate* (tốc độ học): 0.00001. Tốc độ học này bằng một nửa so với mô hình BERT kể trên.
- *Epoch*: 3.
- *Batch_size* (kích thước lô): 32.
- *Sequence_length* (độ dài chuỗi): 128.

3.3 Kiểm thử và giải thích mô hình

Sau khi mô hình học máy đã được huấn luyện xong, chúng ta cần thử mô hình này với tập các dữ liệu mới, sau đó sẽ dùng Integrated Gradients để giải thích hành vi của chúng, tại sao mô hình lại dự đoán nhãn như thế

3.3.1 Kiểm thử

Trong hệ thống này ta sẽ cho 2 mô hình: BERT-base và DistilBERT dự đoán nhãn đầu ra của 30 câu nói về cảm xúc, đánh giá về một dịch vụ nào đó. Hình 4.13 và hình 4.14 minh họa việc dự đoán đầu ra của 2 mô hình này với nhãn 0 (tiêu cực) hoặc 1 (tích cực)

¹<https://huggingface.co/datasets/sst2>

3.3.2 Giải thích mô hình

Sau khi mô hình đã được dự đoán xong, ta sẽ sử dụng thư viện Integrated Gradients của Abili-Explain để giải thích hành vi của các mô hình học máy.

Abili-explain là một công cụ Python được sử dụng để giải thích các quyết định của các mô hình học máy dựa trên phương pháp tích phân liên tục. Được xây dựng trên nền tảng thư viện TensorFlow, thư viện này cung cấp một cách tiếp cận mạnh mẽ để xác định đóng góp của mỗi đặc trưng đầu vào đến quyết định cuối cùng của mô hình. Thư viện Integrated Gradients của Abili-Explain cung cấp các công cụ và chức năng để áp dụng phương pháp tích phân liên tục vào các mô hình học máy đã được huấn luyện. Có 2 bước để sử dụng kỹ thuật Integrated Gradients trong thư viện Abili.

Để khởi tạo Integrated Gradients, ta cần các tham số sau:

- **Cấu trúc mô hình:** Ở đây khi giải thích 2 mô hình (BERT và DistilBERT) ta lấy chính các mô hình đó làm tham số thứ nhất này
- **Một lớp (layer) trong mô hình đó:** Trong cả 2 mô hình, lớp nhúng (embeddings) chiếm phần lớn số tham số của mô hình, nên ta chọn lớp nhúng là lớp được giải thích hành vi
- **Phương pháp xấp xỉ tích phân:** Integrated Gradients sử dụng tích phân nên ta cũng cần một phương pháp để xấp xỉ tích phân. Trong đó phương pháp cầu phương Gauss–Legendre (Gauss–Legendre quadrature) là phương pháp cho độ chính xác cao.
- **Số bước để tính tích phân:** Tham số này biểu thị số lượng số hạng trong tổng dùng để xấp xỉ tích phân. Giá trị tham số này càng cao, độ chính xác càng lớn.
- **Batch_size:** Nếu các mô hình học máy có batch size thì Integrated Gradient cũng có batch size với vai trò tương tự. Trong mỗi lần tính toán, IG sẽ giải thích *batch_size* câu cùng một lúc.

Hình 3.10 là ví dụ cho việc khởi tạo hàm tính toán IG với phương pháp xấp xỉ Gauss-Legendre, tính tích phân bằng tổng của 32 thành phần nhỏ hơn, giải thích được 1 câu trên 1 lần tính toán.

```

✓ [19] # 1st parameter for IG function
# Making sentiment predictions with pretrained BERT Model
auto_model = AutoModelWrapper(auto_model_bert)

# 2nd parameter for IG function
# Layer with respect to which the gradients are calculated
block = auto_model.layers[0].layers[0].transformer.layer[0]

# 3rd parameter for IG function
# Method for the integral approximation
method = "gausslegendre"

# 4th parameter for IG function
# Number of step in the path integral approximation from the baseline
n_steps = 32

# 5th parameter for IG function
# Batch size for the internal batching
internal_batch_size = 1

```



```

✓ [20] # IG function
ig = IntegratedGradients(
    model           = auto_model,
    layer          = block,
    method         = method,
    n_steps        = n_steps,
    internal_batch_size = internal_batch_size
)

```

Hình 3.10: Ví dụ về cách khởi tạo của Integrated Gradients

Khi chọn tham số thứ 2 cho hàm khởi tạo, ta cần đi sâu vào các thuộc tính của biến chứa cấu trúc mô hình (như hình 3.10 là *auto_model*). Sau đó cần đảm bảo class của tham số này là TFEmbeddings, TFBertEmbeddingss hoặc các class khác tương tự về Embeddings của mô hình.

Sau khi mô hình được huấn luyện và kiểm thử xong, ta sẽ gọi một hàm trong thư viện của IG để giải thích cách mô hình dự đoán đầu ra. Hàm này cần 3 tham số chính:

- **Dữ liệu văn bản đầu vào (đầu vào của mô hình học máy):** Đây cũng chính là phần dữ liệu chúng ta đánh giá mức độ quan trọng của từng đặc trưng.

- **Đầu ra được mô hình dự đoán:** Ở đây IG chỉ giải thích làm sao mô hình có thể dự đoán được đầu ra của mô hình mà không thể biết được đầu ra đó có thực sự đúng hay không.
- **Dữ liệu cơ sở:** Tham số này có cùng kích thước với dữ liệu đầu vào, tương trưng cho dữ liệu trung tính. IG sẽ dựa vào dữ liệu cơ sở này để làm hệ quy chiếu và đánh giá mức độ quan trọng của từng đặc trưng trong dữ liệu đầu vào.

```
[ ] # Test sentence for the sentiment analysis
z_test_sample = ['This is the best movie i have ever watch, there is nothing bad to say about that film']

# Process and tokenize the sentences
z_test_sample = process_sentences(z_test_sample, tokenizer, max_len)
print(z_test_sample)

{'input_ids': array([[2023, 2003, 1996, 2190, 3185, 1045, 2031, 2412, 3422, 1010, 2045,
                     2003, 2498, 2919, 2000, 2360, 2055, 2008, 2143]]),
 'attention_mask': array([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]])}

[ ] # 1st parameter for explain function
# Instance for which integrated gradients attribution are computed
x_test_sample = z_test_sample['input_ids']

# 2nd parameter for explain function
# Get tensors for model prediction
kwargs = {k: tf.constant(v) for k, v in z_test_sample.items() if k == 'attention_mask'}

# 4th parameter for explain function
# Get the prediction outputs
predictions = auto_model(x_test_sample, **kwargs).numpy().argmax(axis=1)

# Explain function
explanation = ig.explain(
    X                  = x_test_sample,
    forward_kwargs     = kwargs,
    baselines         = None,
    target            = predictions
)

[ ] predictions
array([1])
```

Hình 3.11: Gọi hàm explain của Integrated Gradients để giải thích mô hình

Hình 3.11 là ví dụ cho việc gọi hàm giải thích của IG với câu văn 'This is the best movie i have ever watch, there is nothing bad to say about that film', dữ liệu cơ sở là 0 và đầu ra được dự đoán là 1. Khi tiến hành kiểm thử với 30 câu văn, ta sẽ gọi hàm giải thích cho từng câu để IG đánh giá mức độ quan trọng của các từ trong câu đó.

CHƯƠNG 4. ỨNG DỤNG VÀO BÀI TOÁN PHÂN TÍCH QUAN ĐIỂM

Trong chương này, tôi sẽ áp dụng mô hình IG vào bài toán phân tích quan điểm, một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên. Mục đích của chương này là đánh giá khả năng giải thích của IG trong việc làm rõ vai trò của từng đặc trưng đầu vào (từng từ trong câu), đối với kết quả dự đoán quan điểm của mô hình học máy.

4.1 Giới thiệu bài toán

Trong phạm vi khóa luận này, tôi sẽ giải thích mô hình về NLP có tốc độ huấn luyện và độ chính xác cao, đó là BERT và DistilBERT. Đầu vào của bài toán bao gồm:

- **Dữ liệu văn bản:** Một câu (hoặc đoạn văn) thể hiện quan điểm hoặc cảm xúc của người nói.
- **Đầu ra dự đoán:** Mô hình sau khi được huấn luyện trước sẽ nhận dữ liệu văn bản ở trên và cho đầu ra tương ứng.

`Predicted label = [1]`

`this is the best movie i have ever watch, there is nothing bad to say about that film`

Hình 4.12: Minh họa đầu ra của Integrated Gradients

Với đầu vào đó, đầu ra của IG sẽ giải thích làm sao DistilBERT có thể đưa ra được dự đoán như vậy, bằng cách gán từng từ trong văn bản bằng con số thực, điểm số của một từ càng nhỏ, từ đó càng không liên quan tới đầu ra, ngược lại từ đó có ảnh hưởng càng lớn đến đầu ra của mô hình. Trong hình 4.12, từ 'best' và 'ever' được nhận định là có ảnh hưởng đến đầu ra nhất, trong khi đó từ 'bad' và 'about' thì ngược lại, không liên quan đến nhãn tích cực của câu.

```

Positive = 42.897042632102966% -> Label [0]
the delivery was late and the product was damaged.
-----
Positive = 95.43190002441406% -> Label [1]
the customer service at the restaurant was exceptional.
-----
Positive = 2.0474281162023544% -> Label [0]
the airline staff was not helpful during our travel delay.
-----
Positive = 95.67959308624268% -> Label [1]
I found the tech support to be very responsive and helpful.
-----
Positive = 80.5235207080841% -> Label [1]
the gym staff is always friendly and supportive.
-----
Positive = 21.48238569498062% -> Label [0]
the service at the car repair shop was not satisfactory.
-----
Positive = 82.19560384750366% -> Label [1]
I had a great experience with the customer service team.
-----
Positive = 77.56950259208679% -> Label [1]
the cleaning service did an excellent job with my apartment.
-----
Positive = 6.245623156428337% -> Label [0]
the customer support was unable to resolve my issue.
-----
Positive = 83.55286121368408% -> Label [1]
the hotel provided excellent room service during our stay.
-----
Positive = 71.42757773399353% -> Label [1]
the medical staff was caring and made the visit less stressful.
-----
Positive = 36.349135637283325% -> Label [0]
I had a poor experience with the airport security staff.
-----
Positive = 88.58258724212646% -> Label [1]
the service at the spa was exceptional and relaxing.
-----
Positive = 42.44862198829651% -> Label [0]
the mechanic was honest and fixed my car efficiently.
-----
Positive = 66.96258783340454% -> Label [1]
the food delivery was quick and the food was still hot.

```



```

Positive = 52.127885818481445% -> Label [1]
the taxi service was unreliable and the driver was impolite.
-----
Positive = 93.55314373970032% -> Label [1]
the concert staff ensured a safe and enjoyable experience for everyone.
-----
Positive = 95.0993299484253% -> Label [1]
the property management service is always responsive to our needs.
-----
Positive = 73.66368174552917% -> Label [1]
the amusement park staff was very friendly and helpful.
-----
Positive = 2.220984175801277% -> Label [0]
the delivery service did not handle the package with care.
-----
Positive = 5.422874167561531% -> Label [0]
the security service at the event was not up to the mark.
-----
Positive = 97.55687117576599% -> Label [1]
the staff at the post office was very helpful in resolving my issue.
-----
Positive = 3.855549916625023% -> Label [0]
the customer service at the internet company was very poor.
-----
Positive = 77.90365815162659% -> Label [1]
the hospital staff provided excellent care during my stay.
-----
Positive = 8.620268106460571% -> Label [0]
the customer service at the phone company was not able to resolve my issue.
-----
Positive = 31.078442931175232% -> Label [0]
the technical support was unable to fix the issue with my computer.
-----
Positive = 53.20456027984619% -> Label [1]
the management at the hotel was not responsive to our complaints.
-----
Positive = 98.70761036872864% -> Label [1]
the service at the pet store was excellent and the staff was very helpful.
-----
Positive = 9.487418830394745% -> Label [0]
the staff at the grocery store was not friendly.
-----
Positive = 98.51463437080383% -> Label [1]
the customer service at the online store was quick to respond and very helpful.

```

Hình 4.13: IG giải thích BERT-base

4.2 Kết quả thực nghiệm

Để đánh giá hiệu quả của IG trong bài toán phân tích cảm xúc, tôi đã tiến hành thử nghiệm trên một tập dữ liệu gồm 30 câu bình luận về 1 sản phẩm, dịch vụ nào đó. Với mỗi câu, cả 2 mô hình BERT và DistilBERT sẽ đưa ra dự đoán về sắc thái cảm xúc (nhãn 1 là tích cực, nhãn 0 là tiêu cực), và IG sẽ dựa vào đó để tìm từ có ảnh hưởng đến đầu ra đó. IG sẽ tô những từ quan trọng bằng màu xanh và những từ không liên quan (hoặc gây nhiễu) màu hồng với độ đậm nhạt tùy theo giá trị số thực mà IG tính cho mỗi từ. Hình 4.13 và hình 4.14 là kết quả của IG khi giải thích hai mô hình này. (Lưu ý những từ in xanh không nhất thiết phải là từ ảnh hưởng tới cảm xúc tích cực của câu văn, mà là ảnh hưởng tới cảm xúc mà mô hình dự đoán)

```

Positive = 0.07577495998702943% -> Label [0]
the delivery was late and the product was damaged
-----
Positive = 99.95248317718506% -> Label [1]
the customer service at the restaurant was exceptional
-----
Positive = 0.05651916726492345% -> Label [0]
the airline staff was not helpful during our travel delay
-----
Positive = 99.04409646987915% -> Label [1]
i found the tech support to be very responsive and helpful
-----
Positive = 99.96076226234436% -> Label [1]
the gym staff is always friendly and supportive
-----
Positive = 0.029699105652980506% -> Label [0]
the service at the car repair shop was not satisfactory
-----
Positive = 99.93947744369507% -> Label [1]
i had a great experience with the customer service team
-----
Positive = 99.20182824134827% -> Label [1]
the cleaning service did an excellent job with my apartment
-----
Positive = 0.028197732171975076% -> Label [0]
the customer support was unable to resolve my issue
-----
Positive = 99.90059733390808% -> Label [1]
the hotel provided excellent room service during our stay
-----
Positive = 96.56701683998108% -> Label [1]
the medical staff was caring and made the visit less stressful
-----
Positive = 0.10126741835847497% -> Label [0]
i had a poor experience with the airport security staff
-----
Positive = 99.98053908348083% -> Label [1]
the service at the spa was exceptional and relaxing
-----
Positive = 99.89727735519409% -> Label [1]
the mechanic was honest and fixed my car efficiently
-----
Positive = 5.236156657338142% -> Label [0]
the food delivery was quick and the food was still hot
-----
```



```

Positive = 0.08977926336228848% -> Label [0]
the taxi service was unreliable and the driver was impolite
-----
Positive = 99.98421669006348% -> Label [1]
the concert staff ensured a safe and enjoyable experience for everyone
-----
Positive = 99.55741167068481% -> Label [1]
the property management service is always responsive to our needs
-----
Positive = 99.95805621147156% -> Label [1]
the amusement park staff was very friendly and helpful
-----
Positive = 0.041952249011956155% -> Label [0]
the delivery service did not handle the package with care
-----
Positive = 0.1276203547604382% -> Label [0]
the security service at the event was not up to the mark
-----
Positive = 99.88718628883362% -> Label [1]
the staff at the post office was very helpful in resolving my issue
-----
Positive = 0.028168471180833876% -> Label [0]
the customer service at the internet company was very poor
-----
Positive = 99.13213849067688% -> Label [1]
the hospital staff provided excellent care during my stay
-----
Positive = 0.029290889506228268% -> Label [0]
the customer service at the phone company was not able to resolve my issue
-----
Positive = 0.026040829834528267% -> Label [0]
the technical support was unable to fix the issue with my computer
-----
Positive = 0.032020913204178214% -> Label [0]
the management at the hotel was not responsive to our complaints
-----
Positive = 99.91828799247742% -> Label [1]
the service at the pet store was excellent and the staff was very helpful
-----
Positive = 0.0324438966345042% -> Label [0]
the staff at the grocery store was not friendly
-----
Positive = 98.13352823257446% -> Label [1]
the customer service at the online store was quick to respond and very helpful
-----
```

Hình 4.14: IG giải thích DistilBERT

4.3 Phân tích và đánh giá khả năng giải thích của Integrated gradients

Trong phần này tôi sẽ phân tích kết quả mà hệ thống đã đạt được ở trên và đề xuất một cách đánh giá mức độ hiệu quả của Integrated Gradients khi giải thích mô hình học máy.

4.3.1 BERT

Với bộ dữ liệu 30 câu mỗi ngày, BERT đạt độ chính xác 90.00% (27/30 câu đúng) và điểm số F1 vào khoảng 0.9143. Nếu ở mỗi câu văn, ta chọn ra một vài từ quan trọng nhất trong câu và xem IG liệu có thể thấy các từ đó nằm ở trong top 5 từ có điểm đánh giá cao nhất trong từng câu không, thì ta thấy tỉ lệ bắt chính xác từ khóa của IG vào khoảng 78.85% (41/52 từ bắt đúng)

Dưới đây là nhận xét về cách IG giải thích BERT dự đoán nhãn cho 30 câu (hình 4.13):

- 'Late' đúng, 'Product' có liên quan nhưng không phải quan trọng nhất, quan trọng nhất ngoài 'late' phải là 'damaged'. Từ 'the' không liên quan là hợp lý. Giải thích tốt ngoại trừ 'damaged'.
- Chỉ có từ 'exceptional' quan trọng nhất. Giải thích rất tốt
- Cụm từ 'not helpful' quan trọng nhất. Từ 'during' nhấn mạnh thời điểm nhân viên ko hữu ích, không thật sự quan trọng bằng not 'helpful'. Từ delay nên có màu đậm hơn, vì chuyến bay trễ thường là không tốt.
- Nhấn mạnh được 4 từ cuối rất quan trọng và các từ đầu tiên không thực sự quan trọng. Từ 'the' không nên là màu xanh, dù là xanh nhạt. Giải thích tốt.
- Từ 'staff' nên là màu xanh nhạt, còn lại giải thích rất tốt.
- 'Not satisfactory' là quan trọng nhất. Còn 'service' và 'car repair shop' nên là màu xanh nhạt. Giải thích không chuẩn lắm.
- Giải thích rất tốt, 'great' là từ khoá chính và cũng được đánh giá là quan trọng, tuy nhiên nên có điểm số quan trọng cao hơn nữa cho từ này.
- Chỉ có 'excellent' là quan trọng nhất, các từ khác không quan trọng. Giải thích rất tốt

- 'Was unable to resolve' được nhấn mạnh rất chuẩn xác. Giải thích rất tốt
- Mặc dù từ 'excellent' được dự đoán là quan trọng nhất, nhưng cách đánh giá từ 'during' đối nghịch với nhau. Chứng tỏ IG không ổn định
- 'Caring' được đánh giá chuẩn nhưng 'less' cũng quan trọng không kém. Các từ màu hồng chấp nhận được. Giải thích tốt ngoại trừ từ 'less'
- Nhìn chung tô màu ổn, tuy nhiên 'experience' nên có màu đậm hơn, các từ đằng sau nó nên có màu nhạt hơn
- Relaxing nên có màu đậm cùng với 'exceptional'. Từ 'the' thứ hai không nên là màu xanh, dù là xanh nhạt. Ngoài ra giải thích tốt.
- Câu này sai nhẫn, tuy nhiên từ quan trọng nhất ('honest' và 'efficiently') lại được tô màu hồng, chứng tỏ nó không liên quan đến đầu ra sai đó. IG phụ thuộc vào đầu ra nên cũng bị ảnh hưởng từ sai sót của mô hình.
- Cụm từ 'still hot' quan trọng, từ quick cũng nên tô xanh. Từ 'delivery' và 'food' nên để xanh nhạt. Còn lại giải thích tốt.
- Câu này sai nhẫn, từ 'unreliable' và 'impolite' là nguyên nhân chính dẫn đến nhẫn 0 của câu này. Cách đánh giá điểm số cũng không chuẩn.
- Giải thích tốt, từ safe nên có điểm số lớn hơn (màu đậm hơn).
- Giới từ 'to' không nên là từ quan trọng nhất. Nếu đổi điểm đánh giá của hai từ 'responsive' và 'to' cho nhau thì điểm số đánh giá sẽ chuẩn hơn.
- Giải thích các từ tốt. Từ 'staff' không nên là màu hồng đậm nhất vì nó là chủ thể của những hành vi 'very friendly' và 'helpful'. Còn lại giải thích rất tốt.
- Giải thích ổn. Nếu đổi chỗ 2 điểm đánh giá của từ 'did' và 'handle' thì tốt hơn.
- Giải thích tốt. 'Was not up' là quan trọng nhất.
- Đánh giá đúng cụm từ 'was helpful' là quan trọng nhất. Cụm từ 'resolving my issue' nên có màu xanh nhạt.
- Đánh giá ổn, chữ 'poor' và 'was' nên đổi đánh giá điểm cho nhau. mặc dù 2 từ này được đánh giá là 2 từ quan trọng nhất, từ 'poor' vẫn quan trọng hơn
- Đánh giá tốt, nhất là từ 'excellent'.

- Đánh giá tốt, nhất là từ 'not'.
- Đánh giá tốt, nhất là từ 'unable' và 'fix'.
- Mặc dù mô hình gán nhãn sai nhưng IG vẫn cho rằng cụm từ 'was not responsive' không liên quan đến đầu ra. Có thể thấy trong một số trường hợp, IG có thể giúp ta nhận biết mô hình gán nhãn sai hay không.
- Đánh giá tốt, nhất là từ 'excellent' và 'helpful'.
- Đánh giá tốt, cụm từ 'not friendly' nên là quan trọng nhất thay vì chữ 'was'.
- Đánh giá đúng cụm từ 'very helpful', tuy nhiên từ quick cũng nên có màu xanh đậm.

Tóm lại, ta thấy IG rất mạnh dạn trong việc chỉ ra những từ ngữ không ảnh hưởng tới đầu ra của mô hình BERT, tuy nhiên chính điều đó lại làm giảm mức độ giải thích chuẩn xác của IG.

4.3.2 DistilBERT

Với bộ dữ liệu 30 câu mới này, DistilBERT đạt độ chính xác 96.67% (29/30 câu đúng) và điểm số F1 vào khoảng 0.9697. Nếu ở mỗi câu văn, ta chọn ra một vài từ quan trọng nhất trong câu và xem IG liệu có thể thấy các từ đó nằm ở trong top 5 từ có điểm đánh giá cao nhất trong từng câu không, thì ta thấy tỉ lệ bắt chính xác từ khóa của IG vào khoảng 92.31% (48/52 từ bắt đúng)

Dưới đây là nhận xét về cách IG giải thích DistilBERT dự đoán nhãn cho 30 câu (hình 4.14):

- 'Late' và 'damaged' là quan trọng nhất. Giải thích rất tốt
- Chỉ có từ 'exceptional' quan trọng nhất. Giải thích rất tốt
- Cụm từ 'not helpful' quan trọng nhất. Từ delay cũng quan trọng nhưng không bằng 'not helpful'. Giải thích tốt.
- Từ 'responsive' nên là màu xanh đậm, còn lại giải thích rất tốt.
- Từ 'friendly' nên là màu xanh đậm, còn lại giải thích rất tốt.
- 'Not satisfactory' là quan trọng nhất. Giải thích tốt.
- Giải thích rất tốt. 'Great experience' là quan trọng nhất.

- Chỉ có 'excellent' là quan trọng nhất. Giải thích rất tốt
- 'Was unable to resolve' được nhấn mạnh rất chuẩn xác. Giải thích rất tốt.
- 'Excellent' là quan trọng nhất. Giải thích tốt.
- Giải thích rất tốt, không chỉ 'caring' mà còn làm chuyển đi đỡ bị áp lực hơn cũng rất quan trọng.
- Nhìn chung tô màu tốt, tuy nhiên 'experience' nên có màu đậm hơn, các từ đằng sau nó nên có màu nhạt hơn
- Từ 'the' không nên là màu xanh, dù là xanh nhạt. Còn lại giải thích rất tốt.
- 'Honest' và 'efficiently' là quan trọng nhất. Giải thích rất tốt.
- Mặc dù mô hình dự đoán sai nhưng IG vẫn đưa ra được 2 từ 'quick' và 'still' là quan trọng nhất.
- 'unreliable' từ dịch vụ là nguyên nhân chính, tài xế 'impolite' là nguyên nhân thứ hai. Hai từ này nên có điểm số quan trọng cao nhất. Giải thích này chấp nhận được.
- Giải thích tốt, từ safe nên có điểm số lớn hơn (màu đậm hơn).
- Nếu đổi điểm đánh giá của hai từ 'responsive' và 'our' cho nhau thì điểm số đánh giá sẽ hoàn hảo hơn.
- Cụm từ 'very friendly' nên có điểm số quan trọng tương đương 'helpful'. Còn lại giải thích rất tốt.
- Giải thích tốt. 'Did not handle' là quan trọng nhất.
- Giải thích tốt. 'Was not up' là quan trọng nhất.
- Cụm từ 'was helpful' nên là yếu tố quan trọng nhất thay cho giới từ 'at'. Còn lại giải thích tốt.
- Đánh giá ổn, chữ 'service' và 'internet' nên có điểm số thấp hơn chút. Giải thích tốt
- Đánh giá tốt, nhất là từ 'excellent'.
- Từ 'not' nên là yếu tố quan trọng nhất thay cho từ 'phone'. Còn lại giải thích tốt.

- Từ 'unable' và 'fix' là quan trọng nhất, từ 'computer' nên có màu xanh nhạt.
- Cụm từ 'was not responsive' là quan trọng nhất. Giải thích rất tốt.
- Từ 'helpful' nên có điểm số quan trọng tương đương 'excellent'. Còn lại giải thích rất tốt.
- Đánh giá tốt, cụm từ 'not friendly' là quan trọng nhất.
- Từ 'quick' và 'helpful' phải là từ quan trọng nhất, không phải từ 'service'. Nếu đổi điểm đánh giá của hai từ 'service' và 'quick' cho nhau thì điểm số đánh giá sẽ tốt hơn.

Nhìn chung IG có biểu hiện trên DistilBERT tốt hơn so với BERT, tuy nhiên nó vẫn gặp khó khăn trong việc nhận biết nhiều hơn một tính từ về cảm xúc (những từ khóa chính quyết định đến đánh giá chung hay nhãn của câu). Mặt khác, IG khi giải thích mô hình DistilBERT ít khi chỉ ra những từ không có tác động hoặc không liên quan đến đầu ra, so với khi giải thích BERT.

4.4 Ứng dụng Integrated Gradients vào thực tế

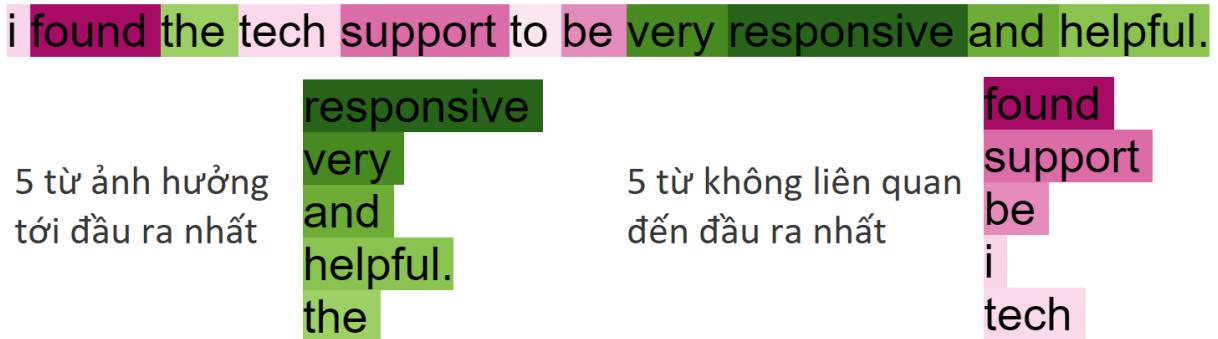
Việc diễn giải hành vi của các mô hình học máy đóng vai trò quan trọng trong việc hiểu và đo lường cảm xúc, ý kiến và phản hồi của mọi người, đặc biệt trong lĩnh vực nghiên cứu thị trường và đánh giá sản phẩm. Áp dụng phương pháp Integrated Gradients để giải thích các mô hình học máy và trực quan hóa kết quả đóng góp rất nhiều cho người dùng. Qua việc trực quan hóa giá trị đóng góp của từng từ, người dùng có thể dễ dàng hiểu cách mà mô hình đưa ra dự đoán cảm xúc và tăng cường tính minh bạch và tin cậy của mô hình. Điều này cung cấp cho họ cái nhìn sâu hơn về cơ chế hoạt động của mô hình và giúp họ có được thông tin giá trị để đưa ra quyết định và cải thiện chất lượng sản phẩm và dịch vụ.

Để thuận tiện trong việc hiển thị kết quả trên giao diện người dùng cuối ta có một vài cách như sau:

- **Cách 1:** Đối với mỗi bài đánh giá, chọn tối đa 5 từ có ảnh hưởng lớn nhất đến kết quả và tối đa 5 từ không liên quan đến kết quả. Hiển thị các từ này dưới dạng danh sách như ví dụ ở hình 4.15

- **Cách 2:** Đối với mỗi bài đánh giá, tạo hai đám mây từ. Đám mây từ thứ nhất chứa các từ có điểm số quan trọng dương (hình 4.17), và đám mây từ thứ hai chứa các từ có điểm số âm (hình 4.18). Trong cả hai đám mây, các từ có giá trị tuyệt đối lớn hơn sẽ được hiển thị với kích cỡ chữ lớn hơn, thu hút sự chú ý của người dùng rằng từ này có ảnh hưởng lớn (đối với đám mây thứ nhất) hoặc không liên quan (đối với đám mây thứ hai) đến đầu ra của mô hình học máy.

Predicted label = [1]



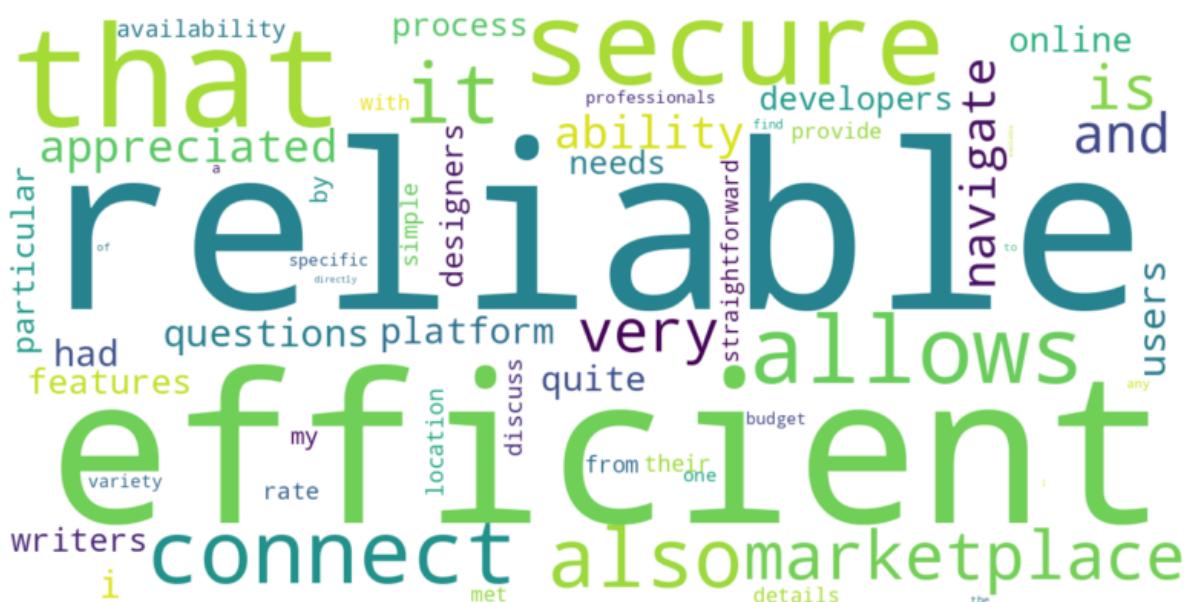
Hình 4.15: Lấy top 5 từ cho mức độ liên quan và không liên quan đến đầu ra

Predicted label = [1]

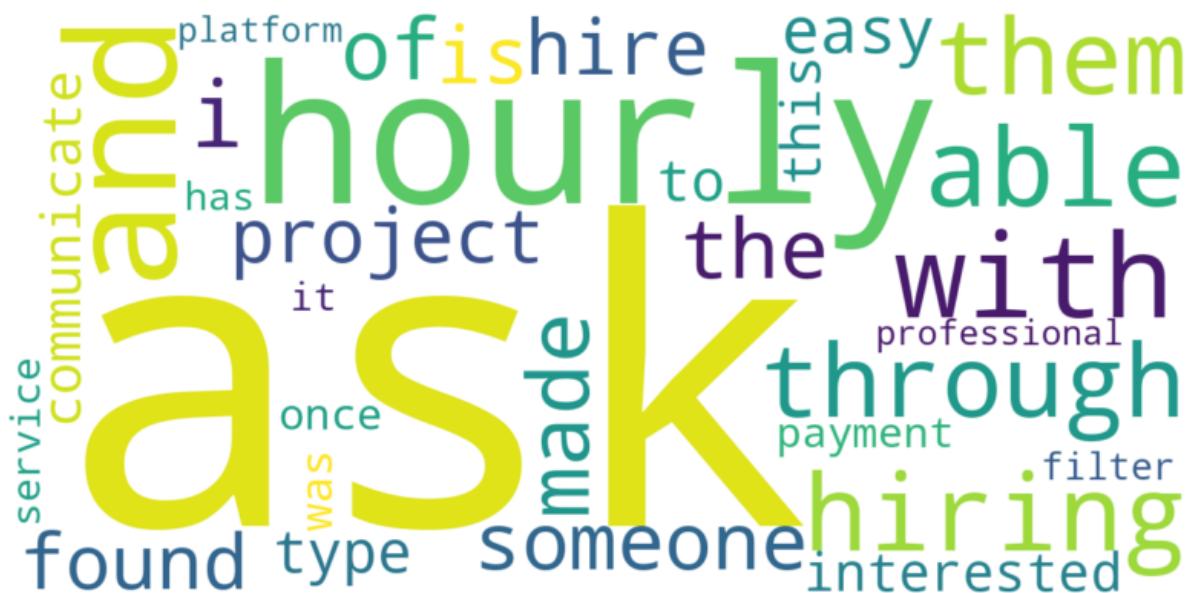
the service is quite efficient and reliable it is a type of online marketplace that allows users to connect with professionals who provide a particular service the platform is easy to navigate and has a variety of professionals available for hire from writers to designers to developers one of the features i appreciated was the ability to filter professionals by their location hourly rate and availability this made it easy to find someone who met my specific needs and budget the process of hiring a professional was straightforward once i found someone i was interested i was able to communicate with them directly through the platform to discuss the project details and ask any questions i had the payment process was also very simple and secure

Hình 4.16: Integrated Gradients giải thích một câu dài hơn

Bằng cách sử dụng cách hiển thị trực quan như vậy, người dùng sẽ có cái nhìn rõ ràng hơn về các từ quan trọng và không quan trọng đối với kết quả đánh giá. Điều này giúp họ nhanh chóng nhận biết các yếu tố ảnh hưởng và cải thiện sự hiểu biết về kết quả của mô hình phân tích cảm xúc.



Hình 4.17: Tập hợp các từ có ảnh hưởng đến đầu ra (điểm số không âm)



Hình 4.18: Tập hợp các từ không liên quan đến đầu ra (điểm số âm)

CHƯƠNG 5. KẾT LUẬN

Trong khóa luận này, chúng tôi đã tập trung nghiên cứu và ứng dụng Integrated Gradients, để giải quyết bài toán phân tích cảm xúc. Qua việc tìm hiểu về một vài mô hình phổ biến được sử dụng trong bài toán phân tích cảm xúc (BERT và DistilBERT) và huấn luyện chúng, ta có thể đưa ra những nhận định ban đầu về khả năng giải thích của Integrated Gradients. Kết quả thử nghiệm cho thấy:

- Khả năng giải thích của Integrated Gradients nói chung: Có thể nhận dạng được những từ mang tính phân cực rõ ràng. Theo IG, những từ này cũng là nguyên nhân khiến mô hình dự đoán đầu vào ở mức trên 90% (là tích cực) cho câu tích cực và dưới 10% (là tích cực) cho câu tiêu cực.
- Khả năng giải thích DistilBERT của Integrated Gradients tương đối đơn giản, tập trung vào những từ quan trọng hơn là những từ gây nhiễu. Trong khi đó khả năng giải thích DistilBERT của Integrated Gradients tập trung vào cả những từ quan trọng và những từ gây nhiễu, tuy nhiên còn thiếu ổn định do có một vài sai sót.
- Việc áp dụng Integrated Gradients không chỉ có lợi ích về việc minh bạch hóa mô hình học máy, mà còn cung cấp một vài phương thức để truyền tải cách giải thích mô hình học máy đến giao diện người dùng cuối.

Tuy nhiên, khóa luận còn một số hạn chế như sau:

- Mô hình chỉ được thử nghiệm trên một tập dữ liệu hạn chế. Việc áp dụng mô hình này cho các bài toán phân tích cảm xúc khác vẫn cần được kiểm chứng thêm.
- Mức độ đa dạng của dữ liệu chưa cao và chưa đủ độ khó (có nhiều thông tin nhiễu không liên quan tới bày tỏ cảm xúc, độ dài câu có thể lớn hơn nhiều lần)

- Tập dữ liệu kiểm thử (30 câu) chỉ bao gồm các từ không bị phân đoạn trong quá trình mã hóa từ (tokenizer). Khi một từ bị tách thành nhiều token, việc đánh giá điểm số chung cho cả từ là không đơn giản.
- Bài toán phân tích cảm xúc có thể phân tích dựa trên nhiều yếu tố chứ không phải là một yếu tố đánh giá cảm xúc chung nhất (ví dụ như "giá thành phải chăng" và "không gian chật hẹp" là 2 yếu tố khác nhau nói về 2 cảm xúc đối lập nhau)

Nhằm khắc phục nhược điểm và nâng cao hiệu quả của mô hình, chúng tôi đề xuất một số định hướng phát triển trong tương lai như sau:

- Mở rộng nghiên cứu và thử nghiệm mô hình trên các tập dữ liệu lớn và đa dạng hơn, việc đánh giá nên dựa theo nhiều yếu tố khác nhau.
- Sử dụng các mô hình để phục vụ cho tập dữ liệu tiếng Việt như phoBERT hoặc BERT-multilingual.
- Tìm hiểu và kết hợp các phương pháp Explainable AI khác với IG để tạo ra một hệ thống giải thích đa chiều, giúp người dùng dễ dàng hiểu hơn về quá trình ra quyết định của mô hình.
- Thử nghiệm và so sánh hiệu quả giữa DistilBERT và các kiến trúc mô hình học sâu khác (như BERT, RoBERTa, GPT) để tìm ra mô hình tối ưu nhất cho bài toán phân tích cảm xúc.
- Xây dựng các ứng dụng thực tế, kết nối với các hệ thống đánh giá, giám sát trực tuyến, để đưa mô hình vào thực tiễn, giúp doanh nghiệp và người dùng có cái nhìn tổng quan hơn về quan điểm của khách hàng và cải thiện chất lượng sản phẩm, dịch vụ.

Như vậy, thông qua khóa luận này, chúng tôi đã khảo sát và ứng dụng thành công phương pháp giải thích mô hình AI bằng IG trong bài toán phân tích cảm xúc, mở ra nhiều triển vọng trong việc tạo ra các mô hình AI dễ hiểu và có tính minh bạch cao hơn. Đồng thời, việc nghiên cứu và áp dụng các định hướng phát triển kể trên sẽ góp phần đưa mô hình phân tích cảm xúc của chúng tôi lên một tầm cao mới, giúp nâng cao chất lượng và hiệu quả của mô hình trong nhiều lĩnh vực ứng dụng khác nhau.

Tài liệu tham khảo

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt **and** Been Kim. “Sanity Checks for Saliency Maps”. *inAdvances in Neural Information Processing Systems: volume 31.* NeurIPS. 2018, pp. 9505–9515. URL: <https://arxiv.org/pdf/1810.03292>.
- [2] Yassine Al Amrani, Mariam Lazaar **and** Khalid El El Kadiri. “Random Forest and Support Vector Machine Based Hybrid Approach to Sentiment Analysis”. *inProcedia Computer Science:* Vol. 127 (2018), pp. 511–520. URL: <https://doi.org/10.1016/j.procs.2018.01.098>.
- [3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila **and** Francisco Herrera. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. *inInformation Fusion:* Vol. 58 (2020). URL: <https://arxiv.org/pdf/1910.10045>.
- [4] Xuefeng Bai, Pengfei Liu **and** Yue Zhang. “Investigating Typed Syntactic Dependencies for Targeted Sentiment Classification Using Graph Attention Neural Network”. *inIEEE/ACM Transactions on Audio, Speech, and Language Processing:* Vol. 29 (2020), pp. 503–514. URL: <https://doi.org/10.1109/TASLP.2020.2963501>.
- [5] Salim Behdenna, Faouzi Barigou **and** Ghalem Belalem. “Document level sentiment analysis: a survey”. *inEAI Endorsed Transactions on Context-aware Systems and Applications:* Vol. 4.13 (2018), pp. e2. URL: <https://eudl.eu/doi/10.4108/eai.13-7-2018.164058>.
- [6] Parminder Bhatia, Yangfeng Ji **and** Jacob Eisenstein. “Better document-level sentiment analysis from RST discourse parsing”. *in(2015):* URL: <https://arxiv.org/abs/1509.01599>.
- [7] Leo Breiman. *Classification and Regression Trees.* 1 edition. Routledge, 1984. URL: <https://doi.org/10.1201/9781315139470>.
- [8] Leo Breiman. “Random forests”. *inMachine Learning:* Vol. 45.1 (2001), pp. 5–32. URL: <https://doi.org/10.1023/A:1010933404324>.
- [9] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O’Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigearaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl

- Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio **and** Markus Anderljung. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims”. *in arXiv preprint arXiv:2004.07213*: (2020). URL: <https://arxiv.org/pdf/2004.07213.pdf>.
- [10] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay **and** Alessandro Feraco. “Affective Computing and Sentiment Analysis”. *in A Practical Guide to Sentiment Analysis*: 2017, pp. 1–10. URL: https://doi.org/10.1007/978-3-319-57358-4_1.
 - [11] Lucian Carata, Sherif Akoush, Nikilesh Balakrishnan, Thomas Bytheway, Ripduman Sohan, Margo Seltzer **and** Andy Hopper. “A primer on provenance”. *in Communications of the ACM*: Vol. 57.5 (2014), pp. 52–60. URL: <https://doi.org/10.1145/2602649.2602651>.
 - [12] Diogo V Carvalho, Eduardo M Pereira **and** Jaime S Cardoso. “Machine Learning Interpretability: A Survey on Methods and Metrics”. *in Electronics*: Vol. 8.8 (july 2019), pp. 832. URL: <https://doi.org/10.3390/electronics8080832>.
 - [13] Jui-Rui Chang, Hsin-Yuan Liang, Li-Sung Chen **and** Chih-Wei Chang. “Novel Feature Selection Approaches for Improving the Performance of Sentiment Classification”. *in Journal of Ambient Intelligence and Humanized Computing*: (2020), pp. 1–14. URL: <https://doi.org/10.1007/s12652-020-02654-4>.
 - [14] Ilia Chetviorkin **and** Natalia Loukachevitch. “Extraction of Russian sentiment lexicon for product metadomain”. *in Proceedings of COLING 2012*: 2012, pp. 593–610.
 - [15] Matthew Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Alexander N Richter **and** Husam Al Najada. “Survey of review spam detection using machine learning techniques”. *in Journal of Big Data*: Vol. 2.1 (2015), pp. 1–24. URL: <https://doi.org/10.1186/s40537-015-0023-0>.
 - [16] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas **and** Prithviraj Sen. “A Survey of the State of Explainable AI for Natural Language Processing”. *in in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*: december 2020, pp. 447–459. URL: <https://arxiv.org/pdf/2010.00711.pdf>.
 - [17] Jacob Devlin, Ming-Wei Chang, Kenton Lee **and** Kristina Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: volume 1. Association for Computational Linguistics, june 2019, pp. 4171–4186. URL: <https://doi.org/10.18653/v1/N19--1423>.
 - [18] Zhiwen Fang, Qingping Zhang, Xue Tang, Aina Wang **and** Christian Baron. “An Implicit Opinion Analysis Model Based on Feature-Based Implicit Opinion Patterns”. *in Artificial Intelligence Review*: Vol. 53.6 (2020), pp. 4547–4574. URL: <https://doi.org/10.1007/s10462-020-09897-0>.
 - [19] Alessio Ferrari **and** Andrea Esuli. “An NLP approach for cross-domain ambiguity detection in requirements engineering”. *in Automated Software Engineering*: Vol. 26.3 (2019), pp. 559–598. URL: <https://doi.org/10.1007/s10515-019-00287-5>.
 - [20] Elena Filatova. “Irony and sarcasm: corpus generation and analysis using crowdsourcing”. *in LREC*: 2012, pp. 392–398. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.

- [21] Lukas Flek. “Returning the N to NLP: towards contextually personalized classification models”. *inProceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: 2020, pp. 7828–7838. URL: <https://www.aclweb.org/anthology/2020.acl-main.684>.
- [22] Ruth C. Fong **and** Andrea Vedaldi. “Interpretable Explanations of Black Boxes by Meaningful Perturbation”. *inin Proceedings of the IEEE International Conference on Computer Vision*: 2017, pp. 3449–3457. URL: <https://arxiv.org/pdf/1704.03296>.
- [23] Karen I Fredriksen-Goldsen **and** Hyun-Jun Kim. “The science of conducting research with LGBT older adults—an introduction to aging with pride: National health, aging, and sexuality/gender study (NHAS)”. *inThe Gerontologist*: Vol. 57 (2017), pp. S1–S14. URL: <https://doi.org/10.1093/geront/gnw180>.
- [24] Lydia H Gilpin, David Bau, Ben Yuan, Ayesha Bajwa, Matt Specter **and** Lalana Kagal. “Explaining explanations: An overview of interpretability of machine learning”. *inIEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*: (2018), pp. 80–89. URL: <https://arxiv.org/pdf/1806.00069>.
- [25] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti **and** Dino Pedreschi. “A Survey of Methods for Explaining Black Box Models”. *inACM Computing Surveys*: Vol. 51(5).93 (2018), pp, 1–42. URL: <https://arxiv.org/pdf/1802.01933>.
- [26] David Gunning **and** David W Aha. “DARPA’s Explainable Artificial Intelligence (XAI) Program”. *inAI Magazine*: Vol. 40.2 (2019), pp. 44–58. URL: <https://doi.org/10.1609/aimag.v40i2.2850>.
- [27] Isabelle Guyon, Jason Weston, Stephen Barnhill **and** Vladimir Vapnik. “Gene selection for cancer classification using support vector machines”. *inMachine learning*: Vol. 46.1–3 (2002), pp. 389–422. URL: <https://doi.org/10.1023/A:1012487302797>.
- [28] Mohammad Al Hassonah, Ramadan Al-Sayyed, Ali Rodan, Ahmad Z Ala’M, Ibrahim Aljarrah **and** Hossam Faris. “An Efficient Hybrid Filter and Evolutionary Wrapper Approach for Sentiment Analysis of Various Topics on Twitter”. *inKnowledge-Based Systems*: Vol. 192 (2020), pp. 105353. URL: <https://doi.org/10.1016/j.knosys.2019.105353>.
- [29] Geoffrey Hinton, Oriol Vinyals **and** Jeff Dean. “Distilling the Knowledge in a Neural Network”. *in*(2015): URL: <https://arxiv.org/pdf/1503.02531>.
- [30] Chiew Ho, Masraha Azrifah Azmi Murad, Shyamala Doraisamy **and** Rabiah Abdul Kadir. “Extracting lexical and phrasal paraphrases: a review of the literature”. *inArtificial Intelligence Review*: Vol. 42.4 (2014), pp. 851–894. URL: <https://doi.org/10.1007/s10462-012-9331-0>.
- [31] Andreas Holzinger, Chris Biemann, Constantin Schoenmüller, Andreas Stigler, Daniela Tran **and** Viktoriia Voronkov. “Interactive machine learning: experimental evidence for the human in the algorithmic loop”. *inApplied Intelligence*: Vol. 49 (2019), pp. 2401–2414. URL: <https://doi.org/10.1007/s10489-018-1361-5>.
- [32] Moumita Kaity **and** V Balakrishnan. “Sentiment Lexicons and Non-English Languages: A Survey”. *inKnowledge and Information Systems*: (2020), pp. 1–36. URL: <https://doi.org/10.1007/s10115-020-01448-4>.
- [33] Ahmed Kamal. “Subjectivity classification using machine learning techniques for mining feature-opinion pairs from web opinion sources”. *inarXiv preprint arXiv:1312.6962*: (2013). URL: <https://arxiv.org/abs/1312.6962>.

- [34] Anusha Kanapala, Sonal Pal **and** Raghavendra Pamula. “Text summarization from legal documents: a survey”. *inArtificial Intelligence Review*: Vol. 51.3 (2019), pp. 371–402. URL: <https://doi.org/10.1007/s10462-018-09679-y>.
- [35] Emaliana Kasmuri **and** Halizah Basiron. “Subjectivity analysis in opinion mining—a systematic literature review”. *inInternational Journal of Advanced Soft Computing Applications*: Vol. 9.3 (2017), pp. 133–159. URL: <https://doi.org/10.14569/IJACSA.2017.080317>.
- [36] Been Kim, Rajiv Khanna **and** Oluwasanmi O Koyejo. “Examples are not enough, learn to criticize! Criticism for Interpretability”. *in29th Advances in Neural Information Processing Systems*: byeditorD. Lee, M. Sugiyama, U. Luxburg, I. Guyon **and** R. Garnett. **volume** 1. Curran Associates, Inc., 2016, pp. 2288–2296. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf.
- [37] Jaesik Kim **and** John Canny. “Interpretable learning for self-driving cars by visualizing causal attention”. *in(2017)*: URL: <https://arxiv.org/pdf/1703.10631.pdf>.
- [38] Pang Wei Koh **and** Percy Liang. “Understanding black-box predictions via influence functions”. *inInternational Conference on Machine Learning*: (march 2017), pp. 1885–1894. URL: <https://arxiv.org/pdf/1703.04730>.
- [39] Taku Kudo. “Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates”. *inProceedings of the 56th Annual Meeting of the Association for Computational Linguistics*: **volume** 1. Association for Computational Linguistics, july 2018, pp. 66–75. URL: <https://doi.org/10.18653/v1/P18-1007>.
- [40] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma **and** Radu Soricut. “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations”. *inOpenReview.net*: april 2020. URL: <https://arxiv.org/pdf/1909.11942>.
- [41] Yann LeCun, Yoshua Bengio **and** Geoffrey Hinton. “Deep learning”. *inNature*: Vol. 521.7553 (2015), pp. 436–444. URL: <https://doi.org/10.1038/nature14539>.
- [42] Bing Liu. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012. ISBN: 978-3-031-02145-9. URL: <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>.
- [43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer **and** Veselin Stoyanov. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. *in(2019)*: URL: <https://arxiv.org/pdf/1907.11692>.
- [44] Bin Lu, Michael Ott, Claire Cardie **and** Benjamin K Tsou. “Multi-aspect sentiment analysis with topic models”. *in2011 IEEE 11th International Conference on Data Mining Workshops*: IEEE. 2011, pp. 81–88. URL: <https://doi.org/10.1109/ICDMW.2011.153>.
- [45] Scott Lundberg **and** Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. *in(2017)*: URL: <https://arxiv.org/pdf/1705.07874>.
- [46] Tim Miller. “Explanation in artificial intelligence: Insights from the social sciences”. *inArtificial Intelligence*: Vol. 267 (2019), pp. 1–38. URL: <https://arxiv.org/pdf/1706.07269>.
- [47] Christoph Molnar. *Interpretable Machine Learning*. Leanpub, 2019. URL: <https://christophm.github.io/interpretable-ml-book/>.

- [48] Alejandro Moreo, Manuel Romero, Jose Castro **and** Jose M Zurita. “Lexicon-Based Comments-Oriented News Sentiment Analyzer System”. *inExpert Systems with Applications*: Vol. 39.10 (2012), pp. 9166–9180. URL: <https://doi.org/10.1016/j.eswa.2012.02.076>.
- [49] Bo Pang, Lillian Lee **and** Shivakumar Vaithyanathan. “Thumbs up? Sentiment Classification using Machine Learning Techniques”. *inin Proceedings of the 2002 conference on Empirical methods in natural language processing*: Association for Computational Linguistics, **july** 2002, pp. 79–86. URL: <https://arxiv.org/pdf/cs/0205070>.
- [50] Seung-Ki Park **and** Yong-Moo Kim. “Building Thesaurus Lexicon Using Dictionary-Based Approach for Sentiment Classification”. *in2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*: IEEE. 2016, pp. 39–44. URL: <https://doi.org/10.1109/SERA.2016.7516126>.
- [51] Yifan Peng, Shikang Yan **and** Zhiyong Lu. “Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets”. *inarXiv preprint arXiv:1906.05474*: (2019). URL: <https://arxiv.org/abs/1906.05474>.
- [52] Alec Radford, Karthik Narasimhan, Tim Salimans **and** Ilya Sutskever. “Improving Language Understanding by Generative Pre-Training”. *inOpenAI Blog*: (2018). URL: <https://openai.com/blog/language-unsupervised/>.
- [53] Guoqing Rao, Wenbin Huang, Zheng Feng **and** Qikun Cong. “LSTM with sentence representations for document-level sentiment classification”. *inNeurocomputing*: Vol. 308 (2018), pp. 49–57. URL: <https://doi.org/10.1016/j.neucom.2018.05.033>.
- [54] Marco Tulio Ribeiro, Sameer Singh **and** Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. *inin 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 2016. ISBN: 978-1-4503-4232-2. URL: <https://arxiv.org/pdf/1602.04938>.
- [55] Marco Tulio Ribeiro, Sameer Singh **and** Carlos Guestrin. “Anchors: High-precision model-agnostic explanations”. *inACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: (2018), pp. 152–534. URL: <https://doi.org/10.1609/aaai.v32i1.11491>.
- [56] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. *inNature Machine Intelligence*: Vol. 1.5 (2019), pp. 206–215. URL: <https://arxiv.org/pdf/1811.10154.pdf>.
- [57] Victor Sanh, Lysandre Debut, Julien Chaumond **and** Thomas Wolf. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. *in(2019)*: URL: <https://arxiv.org/pdf/1910.01108>.
- [58] David Saunders. “Domain adaptation for neural machine translation”. phdthesis. University of Cambridge, 2021. URL: <https://www.repository.cam.ac.uk/handle/1810/320066>.
- [59] Kim Schouten **and** Flavius Frasincar. “Survey on aspect-level sentiment analysis”. *inIEEE Transactions on Knowledge and Data Engineering*: Vol. 28.3 (2015), pp. 813–830. URL: <https://doi.org/10.1109/TKDE.2015.2389623>.
- [60] Rico Sennrich, Barry Haddow **and** Alexandra Birch. “Neural machine translation of rare words with subword units”. *inProceedings of the 54th Annual Meeting of the Association for Computational Linguistics*: **volume** 1. Association for Computational Linguistics, **august** 2016, pp. 1715–1725. URL: <https://doi.org/10.18653/v1/P16-1162>.

- [61] Lloyd S Shapley. "A value for n-person games". in *Contributions to the Theory of Games*: Vol. 2.28 (1953), pp. 307–317. URL: <https://doi.org/10.7249/P0295>.
- [62] Jaypee Singh, Shafiq Irani, Nripendra P Rana, Yogesh K Dwivedi, Saumya **and** Pratyush Kumar Roy. "Predicting the "helpfulness"of online consumer reviews". in *Journal of Business Research*: Vol. 70 (2017), pp. 346–355. URL: <https://doi.org/10.1016/j.jbusres.2016.08.008>.
- [63] Rajeev Kumar Singh, Manish Kumar Sachan **and** Rajendra Patel. "360 degree view of cross-domain opinion classification: A survey". in *Artificial Intelligence Review*: Vol. 54.2 (2021), pp. 1385–1506. URL: <https://doi.org/10.1007/s10462-020-09831-4>.
- [64] Chi Sun, Xipeng Qiu **and** Xuanjing Huang. "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence". in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: volume 1*. Association for Computational Linguistics, june 2019, pp. 380–385. URL: <https://doi.org/10.18653/v1/N19-1035>.
- [65] Mukund Sundararajan. "Axiomatic attribution for deep networks". in *Proceedings of the 34th International Conference on Machine Learning*: Vol. 70 (2017), pp. 3319–3328. URL: <https://arxiv.org/pdf/1703.01365.pdf>.
- [66] Thet Thet Thet, Jin-Cheon Na **and** Christopher S. G. Khoo. "Aspect-based sentiment analysis of movie reviews on discussion boards". in *Journal of Information Science*: Vol. 36.6 (2010), pp. 823–848. URL: <https://doi.org/10.1177/0165551510381263>.
- [67] Robert Tibshirani, Michael Saunders, Saharon Rosset **and** Ji Zhu. "Regression shrinkage and selection via the lasso". in *Journal of the Royal Statistical Society: Series B (Methodological)*: Vol. 58.1 (1996), pp. 267–288. URL: <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- [68] Olga Vechtomova. "Disambiguating context-dependent polarity of words: An information retrieval approach". in *Information Processing & Management*: Vol. 53.5 (2017), pp. 1062–1079. URL: <https://doi.org/10.1016/j.ipm.2017.06.001>.
- [69] Sandra Wachter, Brent Mittelstadt **and** Chris Russell. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. URL: <https://arxiv.org/pdf/1711.00399.pdf>.
- [70] Guojun Wang, Jianyong Sun, Jian Ma, Kun Xu **and** Jia Gu. "Sentiment classification: The contribution of ensemble learning". in *Decision Support Systems*: Vol. 57 (2014), pp. 77–93. URL: <https://doi.org/10.1016/j.dss.2013.09.009>.
- [71] Komal K. Wankhade, Nandkishor G. Bendale, Swapnil V. Kamble, Kajal Deshmukh, Nilesh Sawant **and** Yashwant N. Patil. "A survey on sentiment analysis methods, applications, and challenges". in *Computational Linguistics: Theories, Methods and Applications*: 2022. URL: <https://doi.org/10.1007/s10462-022-10144-1>.
- [72] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest **and** Alexander M. Rush. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". in (2020): URL: <https://arxiv.org/pdf/1910.03771.pdf>.

- [73] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Navdeep Patil, Wei Wang, Cliff Young, Jason Smith, Javier Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Matt Hughes **and** Jeff Dean. “Google’s neural machine translation system: Bridging the gap between human and machine translation”. **in**(2016): URL: <https://arxiv.org/pdf/1609.08144>.
- [74] Yuan Xia, Erik Cambria, Amir Hussain **and** Hui Zhao. “Word polarity disambiguation using Bayesian model and opinion-level features”. **in***Cognitive Computation*: Vol. 7.3 (2015), pp. 369–380. URL: <https://doi.org/10.1007/s12559-015-9315-5>.
- [75] Zhang Yan-Yan, Qin Bing **and** Liu Ting. “Integrating Intra-and Inter-document Evidences for Improving Sentence Sentiment Classification”. **in***Acta Autom Sinica*: Vol. 36.10 (2010), pp. 1417–1425. URL: <https://doi.org/10.3724/SP.J.1004.2010.01417>.
- [76] Bishan Yang **and** Claire Cardie. “Context-aware learning for sentence-level sentiment analysis with posterior regularization”. **in***Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: **volume** 1. 2014, pp. 325–335. URL: <https://www.aclweb.org/anthology/P14-1032/>.
- [77] Jingqing Zhang, Yao Zhao, Mohamed Saleh **and** Peter Liu. “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”. **in***june* 2021: URL: <https://arxiv.org/pdf/1912.08777>.