ĐỀ TÀI: Ứng dụng mô hình Pho-BERT trong Phân Tích Cảm Xúc khách hàng thông qua bình Luận Sản Phẩm Sách trên Sàn Thương Mại Điện Tử Tiki

**GANTT CHART**

**COMMITMENT**

**ACKNOWLEDGEMENTS**

**ABSTRACT**

Trong bối cảnh thương mại điện tử phát triển nhanh chóng, khả năng hiểu và cải thiện trải nghiệm khách hàng đã trở thành yếu tố quyết định sự thành công của doanh nghiệp. Đồ án này tập trung vào việc phân tích đánh giá của khách hàng trên Tiki, một trong những nền tảng thương mại điện tử hàng đầu tại Việt Nam. Mục tiêu chính của nghiên cứu là khai thác dữ liệu bình luận để thực hiện phân tích tâm trạng, từ đó cung cấp những thông tin hữu ích giúp Tiki tối ưu hóa sản phẩm và dịch vụ, nâng cao sự hài lòng và lòng trung thành của khách hàng.

Để đạt được điều này, chúng tôi áp dụng mô hình học sâu PhoBERT, phiên bản tối ưu hóa của BERT cho tiếng Việt. Mô hình này cho phép chúng tôi xác định và phân loại chính xác cảm xúc trong các bình luận của khách hàng, cũng như phát hiện các xu hướng phản hồi. Bằng cách phân tích những tâm trạng này, chúng tôi mong muốn khám phá các mẫu quan trọng phản ánh trải nghiệm của người tiêu dùng, giúp Tiki điều chỉnh các chiến lược kinh doanh một cách hiệu quả. Việc ứng dụng PhoBERT không chỉ nâng cao độ chính xác của phân tích tâm trạng mà còn mang lại hiệu quả trong việc xử lý khối lượng dữ liệu lớn so với các phương pháp truyền thống.

Giao diện người dùng được phát triển cho dự án này được thiết kế đơn giản và trực quan, cho phép người dùng dễ dàng nhập bình luận và nhận kết quả phân tích tâm trạng ngay lập tức. Thiết kế thân thiện với người dùng này nhằm cung cấp trải nghiệm liền mạch, khuyến khích người tiêu dùng tham gia tích cực vào quá trình phản hồi. Bằng cách tạo điều kiện thuận lợi cho việc tiếp cận những thông tin về tâm trạng, Tiki có thể hiểu rõ hơn về nhận thức và sở thích của khách hàng, từ đó xây dựng mối quan hệ tương tác hơn với người dùng.

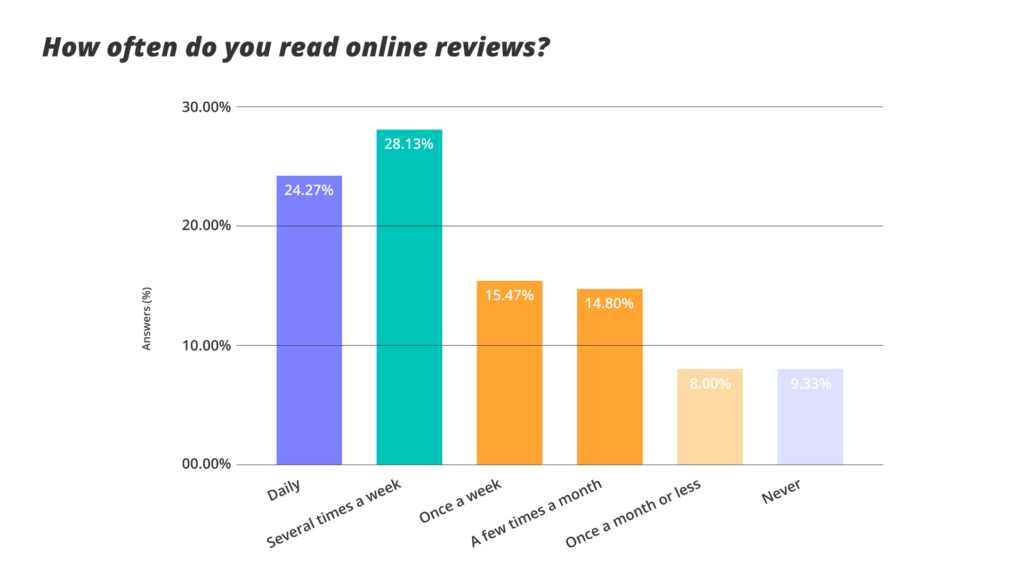
Kết quả nghiên cứu không chỉ cung cấp cho Tiki những thông tin có thể hành động để cải thiện dịch vụ mà còn mở ra hướng đi mới cho các nghiên cứu tiếp theo trong lĩnh vực thương mại điện tử. Những hiểu biết thu được từ dự án này có thể góp phần vào việc hiểu rõ hơn về các xu hướng tâm trạng của khách hàng, thúc đẩy sự đổi mới và phát triển bền vững trong ngành. Tổng thể, nghiên cứu này nhấn mạnh tầm quan trọng của việc ứng dụng các kỹ thuật phân tích tiên tiến nhằm nâng cao trải nghiệm khách hàng trong một thị trường ngày càng cạnh tranh.

# CHAPTER 1. OVERVIEW OF THE PROJECT

Chương 1 cung cấp cái nhìn tổng quan về thương mại điện tử và vai trò quan trọng của trải nghiệm khách hàng trong môi trường cạnh tranh hiện nay. Chương này giới thiệu nền tảng Tiki, phân tích các yếu tố ảnh hưởng đến sự hài lòng của khách hàng và nêu rõ mục tiêu nghiên cứu. Bên cạnh đó, chương còn đề cập đến phương pháp nghiên cứu, quy trình và các công cụ sẽ được sử dụng, tạo cơ sở cho các chương tiếp theo.

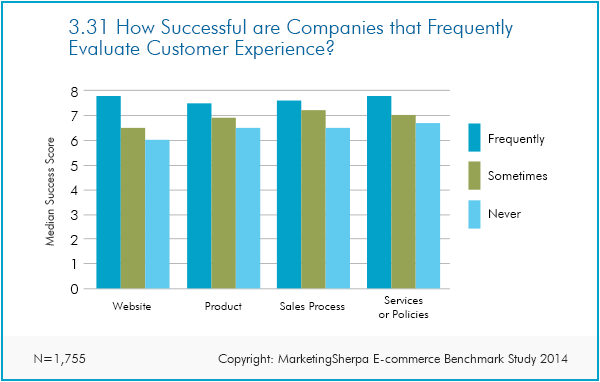
## 1.1. Reason for choosing the project

Đối với nền kinh tế cạnh tranh sôi nổi hiện nay, doanh nghiệp, tổ chức nào chiếm được đa phần thị hiếu người dùng, doanh nghiệp, tổ chức đó chắc chắn sẽ phát triển. Trải nghiệm của khách hàng đối với một tổ chức kinh doanh đóng vai trò quan trọng trong việc ảnh hưởng đến quyết định tiêu thụ, giới thiệu hay quay trở lại với tổ chức đó. Những đánh giá tích cực sẽ giúp doanh nghiệp nhận biết được những yếu tố tạo nên trải nghiệm tích cực của khách hàng khi trải nghiệm sản phẩm, dịch vụ , thúc đẩy lòng trung thành và giữ chân khách hàng lâu hơn. Còn đối với những đánh giá tiêu cực, doanh nghiệp nắm bắt kịp thời những yếu tố xấu ảnh hưởng tiêu cực đến doanh số, dẫn đến truyền miệng những thông tin tiêu cực về doanh nghiệp. Đặc biệt, trong thời đại công nghệ số hiện nay, khách hàng luôn cần có sự tìm hiểu, nắm bắt được chất lượng sản phẩm, dịch vụ họ muốn tiêu thụ của bất kỳ doanh nghiệp nào. Người tiêu dùng có quyền truy cập vào vô số thông tin và có khả năng nghiên cứu sản phẩm hoặc dịch vụ trực tuyến và đọc các đánh giá trước khi mua hàng. Theo một cuộc khảo sát của dự án Internet & American Life của Pew Research Center’s Internet & American Life Project, khoảng 58% người lớn tại Mỹ sẽ thực hiện công việc nghiên cứu sản phẩm và dịch vụ trên Internet trước khi quyết định mua hàng và khoảng 24% khách hàng đã đăng bình luận và đánh giá trực tuyến sau khi trải nghiệm sản phẩm, dịch vụ.



Hình. Khảo sát các khách hàng về tần suất đọc bình luận đánh giá

Biểu đồ trên thể hiện tần suất đọc đánh giá trực tuyến của 750 người tại các quốc gia Đức, Úc, Thụy Sĩ, Pháp và Ý cho thấy đa phần những người thực hiện khảo sát đều đọc đánh giá trực tuyến mỗi ngày hoặc nhiều lần mỗi tuần. Khi đã xác định được đánh giá trực tuyến có ảnh hưởng đáng kể, điều quan trọng là phải thu thập chúng thường xuyên.



Hình. Điểm thành công trung bình của các công ty khi thực hiện đánh giá trải nghiệm khách hàng thường xuyên

Biểu cột đánh giá mức độ thành công của các công ty thuộc phạm vi của nghiên cứu Marketing Sherpa Ecommerce Benchmark Study cho thấy rằng, các công ty thực hiện đánh giá trải nghiệm của khách hàng một cách thường xuyên có mức điểm thành công cao hơn việc công ty chỉ thỉnh thoảng hay không bao giờ đánh giá trải nghiệm của khách hàng.

Trong bối cảnh thương mại điện tử ngày càng phát triển và cạnh tranh khốc liệt, việc phân tích bình luận đánh giá của khách hàng trên nền tảng Tiki trở thành một đề tài nghiên cứu quan trọng. Những bình luận này cung cấp thông tin quý báu không chỉ về chất lượng sản phẩm mà còn về trải nghiệm mua sắm tổng thể trên Tiki. Qua việc phân tích dữ liệu đánh giá, Tiki có thể cải thiện sản phẩm và dịch vụ, tối ưu hóa chiến lược tiếp thị, và quản lý quan hệ khách hàng một cách hiệu quả hơn. Chính vì vậy, nghiên cứu này được lựa chọn nhằm khai thác tiềm năng to lớn của dữ liệu bình luận trong việc nâng cao hiệu suất và tăng cường khả năng cạnh tranh của Tiki trên thị trường thương mại điện tử.

## 1.2. Objectives of the study

Nghiên cứu được thực hiện nhằm đạt được những mục tiêu chính như sau:

Mục tiêu 1: Huấn luyện mô hình với phương pháp học sâu nhằm phân tích cảm xúc khách hàng với dữ liệu đầu vào là các đánh giá của khách hàng về sản phẩm trên nền tảng thương mại điện tử Tiki

Mục tiêu 2: Thiết kế và xây dựng một bộ giao diện tương tác với người dùng, giúp người dùng nắm bắt thông tin về những đánh giá về sản phẩm từ dữ liệu lịch sử.

Dựa trên những mục tiêu cụ thể nhóm nghiên cứu đã đặt ra ngay từ ban đầu, quy trình phân tích, huấn luyện mô hình và xây dựng giao diện tương tác sẽ trở nên có định hướng hơn.

## 1.3. Scope of work (Đối tượng, phạm vi nghiên cứu)

1.3.1 Đối tượng nghiên cứu

Đối tượng được nhóm nghiên cứu sử dụng để phục vụ cho việc phân tích cảm xúc của khách hàng chính là tập dữ liệu đánh giá của khách hàng về các cuốn sách trên nền tảng thương mại điện tử nổi tiếng tại Việt Nam, đó là Tiki.

1.3.2. Phạm vi nghiên cứu

Nghiên cứu phân tích trên những đánh giá được thu thập và tổng hợp trong bộ dữ liệu đánh giá của khách hàng.

1.4. Research Methods

Dữ liệu được nhóm nghiên cứu sử dụng để phân tích là tập dữ liệu đánh giá của khách hàng về các sản phẩm sách trên nền tảng thương mại điện tử Tiki. Nghiên cứu đặt ra mục tiêu phân tích cảm xúc của khách hàng và xây dựng mô hình deep learning cho môn học này.

* Phương pháp 1: Thu thập dữ liệu thứ cấp về đánh giá của khách hàng từ nền tảng Tiki, xác định và lựa chọn các thuộc tính cần thiết cho việc phân tích để đạt được mục tiêu nghiên cứu về cảm xúc của khách hàng.
* Phương pháp 2: Xây dựng quy trình xử lý dữ liệu, xây dựng mô hình học sâu để phân tích cảm xúc của khách hàng bằng tiếng Việt ứng dụng PhoBERT, từ đó khám phá các chủ đề, quan điểm và mức độ hài lòng từ các phản hồi.

1.5. Implementation process.

Quá trình triển khai nghiên cứu sẽ được thực hiện theo các bước sau:

Bước 1: Thu thập dữ liệu:

Tiến hành thu thập dữ liệu đánh giá của khách hàng về sách trên nền tảng thương mại điện tử Tiki. Đảm bảo tính phù hợp của dữ liệu cho đề tài.

Bước 2: Xử lý và làm sạch dữ liệu:

Nghiên cứu quy trình xử lý dữ liệu ngôn ngữ, thực hiện các bước tiền xử lý dữ liệu như loại bỏ các đánh giá trùng lặp, xử lý các giá trị thiếu, và chuẩn hóa văn bản đánh giá. Sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để làm sạch và chuẩn hóa dữ liệu văn bản.

Bước 3: Phân tích dữ liệu:

Xây dựng và huấn luyện mô hình học sâu, ứng dụng PhoBERT để phân tích cảm xúc của khách hàng từ dữ liệu đánh giá. Áp dụng các thuật toán NLP để khám phá các chủ đề, quan điểm và mức độ hài lòng của khách hàng từ các phản hồi bằng tiếng Việt.

Bước 4: Kiểm tra và tối ưu hóa mô hình:

Thực hiện kiểm tra và hiệu chỉnh mô hình để cải thiện độ chính xác và hiệu quả của mô hình phân tích cảm xúc. Đánh giá tính chính xác và độ tin cậy của kết quả mô hình, testing và đảm bảo mô hình phân tích cảm xúc khách hàng đạt hiệu quả cao.

Bước 5: Trình bày kết quả:

Thiết kế và xây dựng bộ giao diện tương tác với người dùng, giúp người dùng có thể nhập đánh giá để mô hình phân loại, dự đoán.

1.6. Report structure

**Chapter 1: Tổng quan về dự án**

Chương đầu tiên cung cấp cái nhìn tổng quan về dự án nghiên cứu, bao gồm lý do chọn đề tài, mục tiêu nghiên cứu, phạm vi nghiên cứu và đối tượng nghiên cứu. Chương này cũng trình bày các phương pháp nghiên cứu, bao gồm cách thu thập dữ liệu và quy trình thực hiện nghiên cứu.

**Chapter 2: Tổng quan tài liệu**

Chương này tổng hợp và phân tích các bài nghiên cứu liên quan đã được công bố trước đó. Nhóm nghiên cứu đã tìm kiếm và phân tích các bài báo và sách có nội dung hỗ trợ và đề xuất cho dự án, dựa trên các từ khóa liên quan.

**Chapter 3: Cơ sở lý thuyết**

Chương này trình bày các khái niệm cơ bản và lý thuyết liên quan đến nghiên cứu, bao gồm Deep Learning, BERT, Pho-BERT, NLP và phân tích cảm xúc (Sentiment Analysis). Các phương pháp và công nghệ được sử dụng trong nghiên cứu cũng được giải thích trong chương này.

**Chapter 4: Mô hình đề xuất và thực nghiệm**

Chương này giới thiệu mô hình đề xuất để giải quyết vấn đề cụ thể, bao gồm các yêu cầu và chỉ số đo lường hiệu quả (KPIs). Chương cũng mô tả các nguồn dữ liệu, quy trình Extract, Transform, Load (ETL), và cấu trúc kho dữ liệu được xây dựng. Các kết quả từ từng bước của mô hình và thực nghiệm sẽ được trình bày để làm rõ quy trình và kết quả đạt được.

**Chapter 5: Triển khai mô hình**

Chương này tập trung vào việc triển khai mô hình thành ứng dụng, bao gồm thiết kế giao diện người dùng (User Interface) và quy trình triển khai giao diện. Các công cụ và kỹ thuật sử dụng để xây dựng giao diện cũng được mô tả trong chương này.

**Chapter 6: Kết luận**

Chương cuối cùng đánh giá hiệu suất của mô hình và đưa ra những kết luận cuối cùng từ nghiên cứu. Chương này cũng phân tích những kiến thức mới mà nhóm nghiên cứu đã thu được, đánh giá đóng góp của đề tài và nhận định những hạn chế cũng như hướng phát triển trong tương lai.

**Chapter 2. LITERATURE REVIEW(mỗi phần 3 bài báo)**

Chương này tổng quan các nghiên cứu trước đây liên quan đến deep learning, phân tích dữ liệu văn bản, PhoBERT, phân tích tâm trạng trong thương mại điện tử. Chương trình bày các mô hình và phương pháp đã được áp dụng, đồng thời phân tích kết quả và hạn chế của những nghiên cứu này. Qua đó, chương làm nổi bật tầm quan trọng của việc áp dụng phân tích tâm trạng để hiểu rõ hơn về hành vi khách hàng, tạo cơ sở cho nghiên cứu của chúng tôi trên nền tảng Tiki.

2.1. Deep Learning

Đề cập về Deep Learning, quyển sách “Dive into Deep Learning” phát biểu như sau, Machine learning là nghiên cứu về các thuật toán có khả năng học hỏi từ kinh nghiệm. Khi một thuật toán machine learning tích lũy nhiều kinh nghiệm hơn, thường là dưới dạng dữ liệu quan sát hoặc tương tác với môi trường, hiệu suất của nó sẽ cải thiện. Điều này khác biệt so với nền tảng thương mại điện tử xác định sẵn của chúng ta, nơi mà logic kinh doanh được áp dụng một cách đồng nhất, bất kể lượng kinh nghiệm tích lũy là bao nhiêu, cho đến khi các nhà phát triển tự nhận thấy và quyết định cần cập nhật phần mềm. Các nguyên tắc cơ bản của machine learning, đặc biệt là deep learning, một tập hợp các kỹ thuật mạnh mẽ đang thúc đẩy đổi mới trong các lĩnh vực đa dạng như thị giác máy tính, xử lý ngôn ngữ tự nhiên, chăm sóc sức khỏe và di truyền học. Cho thấy ứng dụng của Deep Learning và Machine Learning là vô cùng rộng.

Trong tài liệu “Deep Learning” của Ian Goodfellow (Đại học MIT) nói rằng deep learning là một phương pháp trong machine learning đã dựa nhiều vào kiến thức về não người, thống kê và toán học ứng dụng trong suốt vài thập kỷ qua. Trong những năm gần đây, deep learning đã chứng kiến sự gia tăng đáng kể về độ phổ biến và tính hữu ích, chủ yếu nhờ vào sự phát triển của máy tính mạnh mẽ hơn, dữ liệu lớn hơn và các kỹ thuật để đào tạo các mạng nơ-ron sâu hơn. Những năm tới đầy thử thách và cơ hội để cải thiện deep learning hơn nữa và đưa nó đến những lĩnh vực mới.

Ví dụ về một nghiên cứu ứng dụng của deep learning và sự phát triển của nó, có thể tham khảo nghiên cứu “Towards Deep Learning Models Resistant to Adversarial Attacks”, nghiên cứu này đã đề xuất một phương pháp mới để tăng cường khả năng chống lại các tấn công đối kháng của mô hình học sâu (Adversarial Attacks), sử dụng cách tiếp cận tối ưu hóa mạnh (robust optimization) thông qua công thức điểm yên ngựa (saddle point formulation). Các tác giả chứng minh rằng bài toán này có thể giải quyết hiệu quả bằng phương pháp Projected Gradient Descent (PGD), và nhấn mạnh tầm quan trọng của dung lượng mô hình trong việc tạo ra các mạng neural mạnh mẽ. Thực nghiệm trên các bộ dữ liệu MNIST và CIFAR10 cho thấy phương pháp này có thể tạo ra các mô hình có khả năng chống lại nhiều loại tấn công đối kháng, mở ra triển vọng cho việc xây dựng các mạng neural an toàn trong tương lai.

Có thể thấy deep learning đang trở thành một trong những lĩnh vực đổi mới quan trọng nhất trong khoa học máy tính và trí tuệ nhân tạo. Với khả năng học hỏi từ dữ liệu và cải thiện hiệu suất qua từng trải nghiệm, deep learning đã vượt ra ngoài các phương pháp truyền thống, mở ra nhiều cơ hội ứng dụng trong các lĩnh vực như thị giác máy tính, xử lý ngôn ngữ tự nhiên và nhiều lĩnh vực khác. Sự phát triển của công nghệ này không chỉ dựa vào sự gia tăng sức mạnh tính toán và dữ liệu lớn mà còn nhờ vào những kỹ thuật tiên tiến như tối ưu hóa mạnh và cải thiện mô hình. Các nghiên cứu hiện tại chứng minh sự linh hoạt và tiềm năng tiếp tục của deep learning trong việc giải quyết các thách thức phức tạp và nâng cao hiệu suất mô hình. Do đó, deep learning không chỉ là một công nghệ quan trọng trong hiện tại mà còn là nền tảng cho sự đổi mới và phát triển trong tương lai, mở ra nhiều triển vọng cho việc áp dụng trong các lĩnh vực khác nhau và cải thiện cuộc sống con người.

2.2. Phân tích cảm xúc khách hàng trong lĩnh vực thương mại điện tử

Trong nghiên cứu “Customer experience and satisfaction: Importance of customer reviews and customer value on buying preference” kết luận rằng các nhà hàng nên khuyến khích khách hàng để lại đánh giá trên Zomato và phản hồi cả phản hồi tích cực và tiêu cực. Zomato nên cân nhắc triển khai tính năng cho phép khách hàng lọc đánh giá dựa trên các tiêu chí cụ thể, chẳng hạn như chất lượng thực phẩm hoặc dịch vụ. Các nhà hàng nên tập trung vào việc cải thiện xếp hạng chung của mình bằng cách giải quyết các khiếu nại phổ biến được đề cập trong các đánh giá tiêu cực. Từ những đề xuất cho thấy việc thu thập và phân tích đánh giá của các khách hàng thực sự cần thiết cho sự phát triển lâu dài của công ty.

Hơn nữa, nghiên cứu “Determinants of Ecommerce customer satisfaction, trust and loyalty in Saudi Arabia” đã điều tra xem sự hài lòng và lòng tin của khách hàng có đóng vai trò trung gian đáng kể đối với lòng trung thành của người tiêu dùng Ả Rập Xê Út đối với các dịch vụ thương mại điện tử B2C hay không. Sự hài lòng của khách hàng thương mại điện tử được nhận thấy có ảnh hưởng đáng kể đến lòng trung thành của khách hàng và đóng vai trò là trung gian một phần giữa từng cấu trúc chất lượng giao diện người dùng và chất lượng thông tin và lòng trung thành của khách hàng.

Trong một nghiên cứu khác “The effect of customer feedback ratings on purchase decision in ecommerce” đã xem xét vô số yếu tố có khả năng ảnh hưởng đến số tiền khách hàng sẵn sàng mua hàng, nghiên cứu này đã chỉ ra rằng xếp hạng và quyết định mua hàng không có mối quan hệ hình chữ U ngược mà có mối quan hệ đa thức với quyết định mua hàng. Họ đã chỉ ra rằng xếp hạng tăng 1% sẽ khiến khách hàng sẵn sàng trả thêm 859,5 yên. Họ cũng thấy rằng xếp hạng tăng 1% dẫn đến tiện ích tăng 0,344.

Trong bối cảnh cạnh tranh ngày càng gay gắt của thị trường thương mại điện tử, việc phân tích đánh giá của khách hàng trở thành một yếu tố cực kỳ quan trọng. Đánh giá của khách hàng không chỉ là nguồn thông tin quý giá giúp doanh nghiệp hiểu rõ hơn về trải nghiệm và nhu cầu của khách hàng mà còn là công cụ mạnh mẽ để xây dựng và duy trì uy tín thương hiệu. Thông qua việc phân tích các phản hồi từ khách hàng, doanh nghiệp có thể phát hiện ra những vấn đề tồn đọng trong sản phẩm hoặc dịch vụ, từ đó cải thiện chất lượng và nâng cao sự hài lòng của khách hàng. Hơn nữa, đánh giá tích cực có thể tạo dựng lòng tin và thu hút thêm nhiều khách hàng tiềm năng, trong khi việc giải quyết kịp thời các đánh giá tiêu cực sẽ giúp giảm thiểu tác động xấu đến hình ảnh của doanh nghiệp. Do đó, phân tích đánh giá của khách hàng không chỉ là một nhiệm vụ cần thiết mà còn là một chiến lược quan trọng để thúc đẩy sự phát triển bền vững của doanh nghiệp trên nền tảng thương mại điện tử.

2.3. Phân tích dữ liệu văn bản

Nghiên cứu “Sentiment Analysis of the Top 5 E-commerce Platforms in Indonesia using Text Mining and Natural Language Processing (NLP)” đã ứng dụng phân tích dữ liệu văn bản và xử lý ngôn ngữ tự nhiên để phân tích sentiment của người dùng đối với 5 nền tảng thương mại điện tử lớn nhất ở Indonesia dựa trên việc phân tích các phản hồi với 3 hình thái là ‘tích cực’, ‘tiêu cực’ và ‘trung lập’. Trong bài báo này, các kỹ thuật liên quan đến text mining quan trọng được sử dụng bao gồm loại bỏ ngưng từ (stopwords), tạo ma trận term-document, phân tích tần suất từ và cuối cùng là trực quan hóa bằng word cloud. Họ đã đưa ra được kết quả là có 59.46% người dùng đánh giá tích cực về các nền tảng thương mại điện tử ở Indonesia. Nghiên cứu cũng đúc kết được phương pháp và cách thức ứng dụng phân tích dữ liệu văn bản vào phân tích cảm xúc của khách hàng.

Bên cạnh đó, nghiên cứu “Research on Sentiment Analysis of E-commerce Live Comments based on Text Mining” cũng sử dụng phân tích dữ liệu văn bản để phân tích cảm xúc của khách hàng từ bình luận trực tiếp trong các buổi phát trực tiếp trên sàn thương mại điện tử với mục tiêu là khai thác thông tin về xu hướng cảm xúc của người dùng đối với sản phẩm, dịch vụ và trải nghiệm mua sắm. Quy trình khai thác dữ liệu văn bản này khác với nghiên cứu trước đó ở việc xây dựng từ điển cảm xúc và gán giá trị phân cực cảm xúc cho từng từ, kết hợp với thuật toán máy học như SVM, Naive Bayes để huấn luyện mô hình, còn lại trong quy trình tương tự với nghiên cứu trước đó. Kết quả của nghiên cứu này đã đóng góp góc nhìn sâu hơn về cảm xúc và thái độ của người dùng, đồng thời cung cấp thông tin cho doanh nghiệp thương mại điện tử tối ưu hóa hoạt động kinh doanh và cải thiện từ điển cảm xúc.

Ngoài ra, bài nghiên cứu “Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm” cũng đề xuất phương pháp khai thác ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm bằng cách thu thập bình luận từ website Foody.vn. Sử dụng các thuật toán học máy, nghiên cứu đã tiến hành phân tích các bình luận dạng văn bản để khai phá ý kiến và trực quan hóa kết quả, hỗ trợ ra quyết định cho các doanh nghiệp. Kết quả cho thấy độ chính xác của phương pháp đạt 90%, cung cấp thông tin hữu ích giúp các cửa hàng và nhà quản trị hiểu rõ hơn về ưu nhược điểm của sản phẩm và dịch vụ. Nghiên cứu này cũng so sánh các thuật toán để tìm ra mô hình tốt nhất dựa trên F-Score và tạo các báo cáo trực quan. Trong tương lai, nghiên cứu sẽ mở rộng với hệ thống tự động cập nhật và xử lý dữ liệu lớn, cùng với triển khai ứng dụng trên thiết bị di động để hỗ trợ doanh nghiệp ra quyết định hiệu quả hơn.

2.4. Pho-Bert

Đề tài nghiên cứu “Sentiment Analysis Implementing BERT-based Pre-trained Language Model for Vietnamese” tập trung vào việc phân tích cảm xúc từ phản hồi của sinh viên để cải thiện chất lượng giáo dục. Nghiên cứu sử dụng mô hình PhoBERT, một phiên bản tối ưu hóa của mô hình BERT cho tiếng Việt, nhằm khắc phục hạn chế của các mô hình phân loại cảm xúc hiện đại chỉ tập trung vào tiếng Anh. Bằng cách áp dụng các kỹ thuật fine-tuning để mở rộng mô hình cho phân loại đa lớp, phương pháp của nghiên cứu đạt kết quả xuất sắc trên tập dữ liệu UIT-VSFC với F1-score là 93,92% và độ chính xác là 94,28%. Kết quả này không chỉ hỗ trợ cải thiện giáo dục tại Việt Nam mà còn mở ra cơ hội nghiên cứu cho ngôn ngữ thiếu tài nguyên như tiếng Việt.

Tiếp theo, bài báo "Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data" nhằm phát triển một hệ thống thông minh để phát hiện các nội dung thù ghét và xúc phạm, từ đó xây dựng một môi trường lành mạnh và an toàn. Bài báo nhận thấy các nghiên cứu hiện tại trong lĩnh vực này còn tồn tại nhiều thiếu sót như kỹ thuật tiền xử lý kém hiệu quả, không quan tâm đến vấn đề mất cân bằng dữ liệu, mô hình có hiệu suất khiêm tốn, và thiếu ứng dụng thực tiễn. Để khắc phục những vấn đề này, các tác giả đã đề xuất một số giải pháp. Trước hết, họ giới thiệu một kỹ thuật tiền xử lý hiệu quả để làm sạch các bình luận từ mạng xã hội Việt Nam. Sau đó, họ phát triển một mô hình phát hiện ngôn từ thù ghét mới, kết hợp giữa mô hình PhoBERT tiền huấn luyện và Text-CNN. Các kỹ thuật EDA cũng được áp dụng để giải quyết vấn đề mất cân bằng dữ liệu và cải thiện hiệu suất của các mô hình phân loại. Các thí nghiệm so sánh cho thấy mô hình PhoBERT-CNN vượt trội hơn các phương pháp hiện đại nhất, đạt điểm F1 lần lượt là 67.46% và 98.45% trên hai tập dữ liệu chuẩn ViHSD và HSD-VLSP. Cuối cùng, họ xây dựng một ứng dụng phát trực tiếp HSD để minh họa tính thực tiễn của hệ thống. Nhìn chung, bài báo tập trung vào việc phát triển mô hình PhoBERT-CNN để phát hiện ngôn từ thù ghét trong tiếng Việt và chứng minh tính hiệu quả cũng như ứng dụng thực tiễn của hệ thống thông qua các thí nghiệm và ứng dụng thực tế.

Cuối cùng, bài báo "Combining PhoBERT and SentiWordNet for Vietnamese Sentiment Analysis" tập trung vào nhiệm vụ phân tích cảm xúc, một trong những nhiệm vụ quan trọng nhất của Xử lý Ngôn ngữ Tự nhiên (NLP), nơi các mô hình học máy được huấn luyện để phân loại văn bản theo tính cực của ý kiến. Nhiều mô hình đã được đề xuất để giải quyết nhiệm vụ này, trong đó các mô hình PhoBERT tiền huấn luyện được coi là tiên tiến nhất cho tiếng Việt. Phương pháp tiền huấn luyện của PhoBERT dựa trên RoBERTa, tối ưu hóa phương pháp tiền huấn luyện của BERT để đạt hiệu suất mạnh mẽ hơn. Trong bài báo này, các tác giả giới thiệu một phương pháp mới kết hợp PhoBERT và SentiWordNet để phân tích cảm xúc từ các đánh giá tiếng Việt. Mô hình phân tích cảm xúc được đề xuất sử dụng PhoBERT, một sự tối ưu hóa mạnh mẽ cho tiếng Việt từ mô hình BERT nổi bật, và SentiWordNet, một tài nguyên từ vựng được thiết kế đặc biệt để hỗ trợ các ứng dụng phân loại cảm xúc. Kết quả thí nghiệm trên các tập dữ liệu VLSP 2016 và AIVIVN 2019 cho thấy hệ thống phân tích cảm xúc của họ đạt hiệu suất tốt so với các mô hình khác.

**Chapter 3. THEORETICAL BACKGROUND**

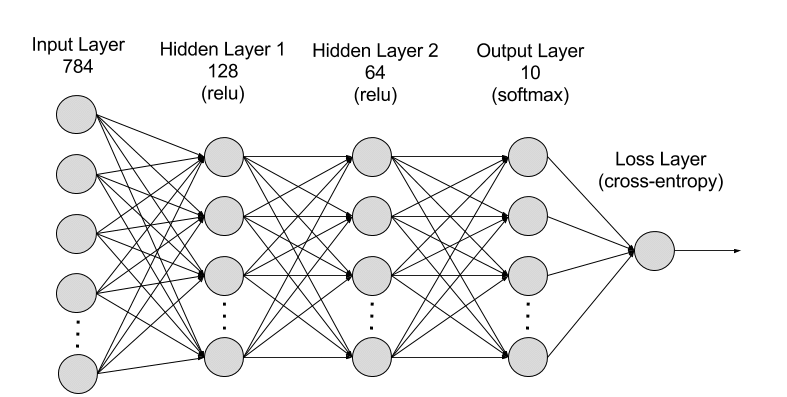
Chương 3 cung cấp cơ sở lý thuyết cho nghiên cứu, bao gồm các khái niệm về học sâu, đặc biệt là các mô hình như BERT, PhoBERT, xử lý ngôn ngữ tự nhiên và thảo luận về hiệu quả của các mô hình, tạo nền tảng cho nghiên cứu phía sau.

3.1.Deep Learning

Học sâu là một nhánh của trí tuệ nhân tạo (AI), mô phỏng cách thức hoạt động của bộ não con người để xử lý và phân tích dữ liệu. Các mô hình học sâu có khả năng nhận diện và phân loại các mẫu phức tạp trong hình ảnh, văn bản, âm thanh và các dạng dữ liệu khác, tạo ra thông tin chuyên sâu và dự đoán chính xác. Công nghệ này có thể tự động hóa các tác vụ phức tạp như mô tả hình ảnh hoặc chuyển đổi âm thanh thành văn bản.

Mô hình học sâu đóng vai trò quan trọng trong sự phát triển của trí tuệ nhân tạo (AI) bằng cách cung cấp cho máy tính khả năng học hỏi và tư duy như con người. Công nghệ này được ứng dụng rộng rãi trong các sản phẩm và dịch vụ hàng ngày như trợ lý kỹ thuật số, điều khiển bằng giọng nói, phát hiện gian lận và nhận dạng khuôn mặt. Học sâu cũng là nền tảng cho các công nghệ tiên tiến như xe tự lái và thực tế ảo.

Học sâu có nhiều ứng dụng trong các lĩnh vực như ô tô, hàng không vũ trụ, y tế, và sản xuất. Ví dụ, xe tự lái sử dụng học sâu để nhận diện biển báo giao thông và người đi bộ; hệ thống quốc phòng sử dụng để phân tích ảnh vệ tinh; y học sử dụng để phát hiện tế bào ung thư trong chẩn đoán hình ảnh. Các ứng dụng khác bao gồm nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên và các công cụ đề xuất cá nhân hóa.

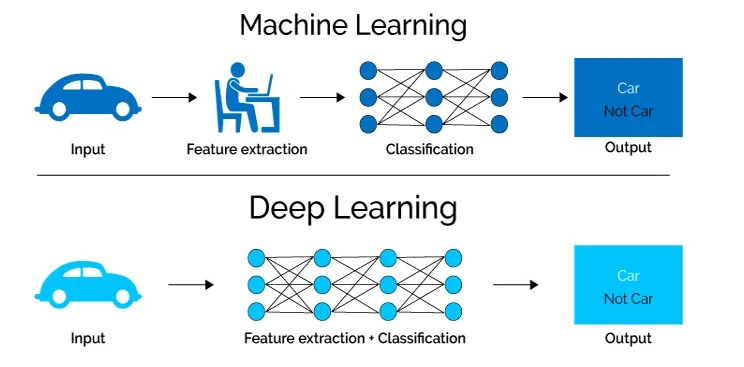


Hình 3. Mô hình mạng nơ-ron sâu

Mô hình học sâu dựa trên các mạng nơ-ron nhân tạo, mô phỏng cấu trúc của bộ não con người. Mạng nơ-ron này bao gồm các lớp đầu vào, lớp ẩn và lớp đầu ra. Các lớp ẩn xử lý thông tin ở nhiều cấp độ khác nhau, cho phép mạng học sâu phân tích và đưa ra dự đoán từ dữ liệu phức tạp.

Một mạng nơ-ron học sâu gồm có lớp đầu vào để nhận dữ liệu, lớp ẩn để xử lý dữ liệu và lớp đầu ra để xuất kết quả. Các lớp ẩn có thể có hàng trăm lớp, mỗi lớp xử lý một khía cạnh khác nhau của dữ liệu, giúp mạng học sâu phân tích vấn đề từ nhiều góc độ và đưa ra kết quả chính xác hơn.

Học sâu là một phần của máy học, ra đời nhằm cải thiện hiệu quả của các kỹ thuật máy học truyền thống. Các phương pháp máy học truyền thống yêu cầu rất nhiều công sức trong việc dán nhãn và xử lý dữ liệu. Ngược lại, mô hình học sâu có khả năng tự học và cải thiện dần theo thời gian mà không cần quá nhiều dữ liệu đã được dán nhãn.



Hình 3. Cách thức hoạt động của Machine Learning và Deep Learning

So với máy học, mô hình học sâu có thể xử lý hiệu quả dữ liệu phi cấu trúc, khám phá các mối quan hệ và mẫu ẩn trong dữ liệu, và học không giám sát. Điều này giúp các mô hình học sâu phân tích dữ liệu một cách sâu rộng hơn, đưa ra dự đoán chính xác hơn và tự động cải thiện theo thời gian mà không cần can thiệp thủ công nhiều.

Một số thách thức của học sâu bao gồm yêu cầu lượng lớn dữ liệu chất lượng cao và năng lực xử lý lớn. Dữ liệu đầu vào cần phải được làm sạch và xử lý trước khi có thể sử dụng để đào tạo mô hình học sâu. Ngoài ra, các thuật toán học sâu yêu cầu cơ sở hạ tầng mạnh mẽ để hoạt động hiệu quả, nếu không, quá trình xử lý có thể tốn rất nhiều thời gian.

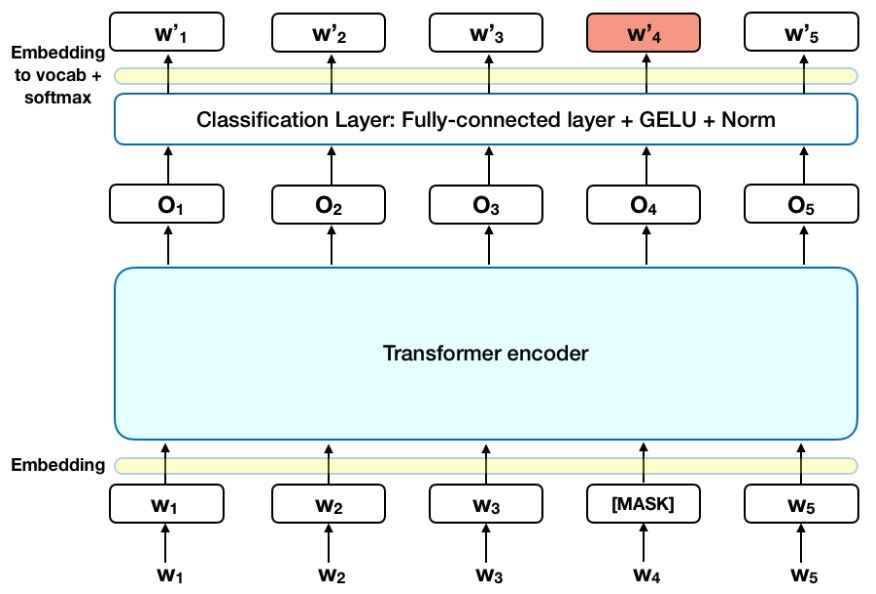
3.2. Bert

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ được phát triển bởi Google AI. BERT đánh dấu một bước đột phá lớn trong lĩnh vực Machine Learning nhờ khả năng ứng dụng vào nhiều bài toán xử lý ngôn ngữ tự nhiên (NLP) như trả lời câu hỏi và suy luận ngôn ngữ tự nhiên, với hiệu quả rất cao.

Một trong những thách thức lớn nhất của NLP là xử lý dữ liệu không đồng nhất. Trên Internet có vô vàn dữ liệu, nhưng chúng thường được sử dụng cho các mục đích riêng biệt. Điều này dẫn đến việc chúng ta chỉ có một lượng nhỏ dữ liệu thích hợp cho từng bài toán cụ thể, trong khi các mô hình Deep Learning lại cần lượng dữ liệu lớn để đạt kết quả tốt. Transfer Learning, một kỹ thuật mới, giúp khắc phục vấn đề này bằng cách xây dựng các mô hình chung từ dữ liệu khổng lồ trên Internet và tinh chỉnh cho từng bài toán cụ thể. BERT là một đại diện xuất sắc của Transfer Learning trong NLP, mang lại hiệu quả cao và hoàn toàn miễn phí.

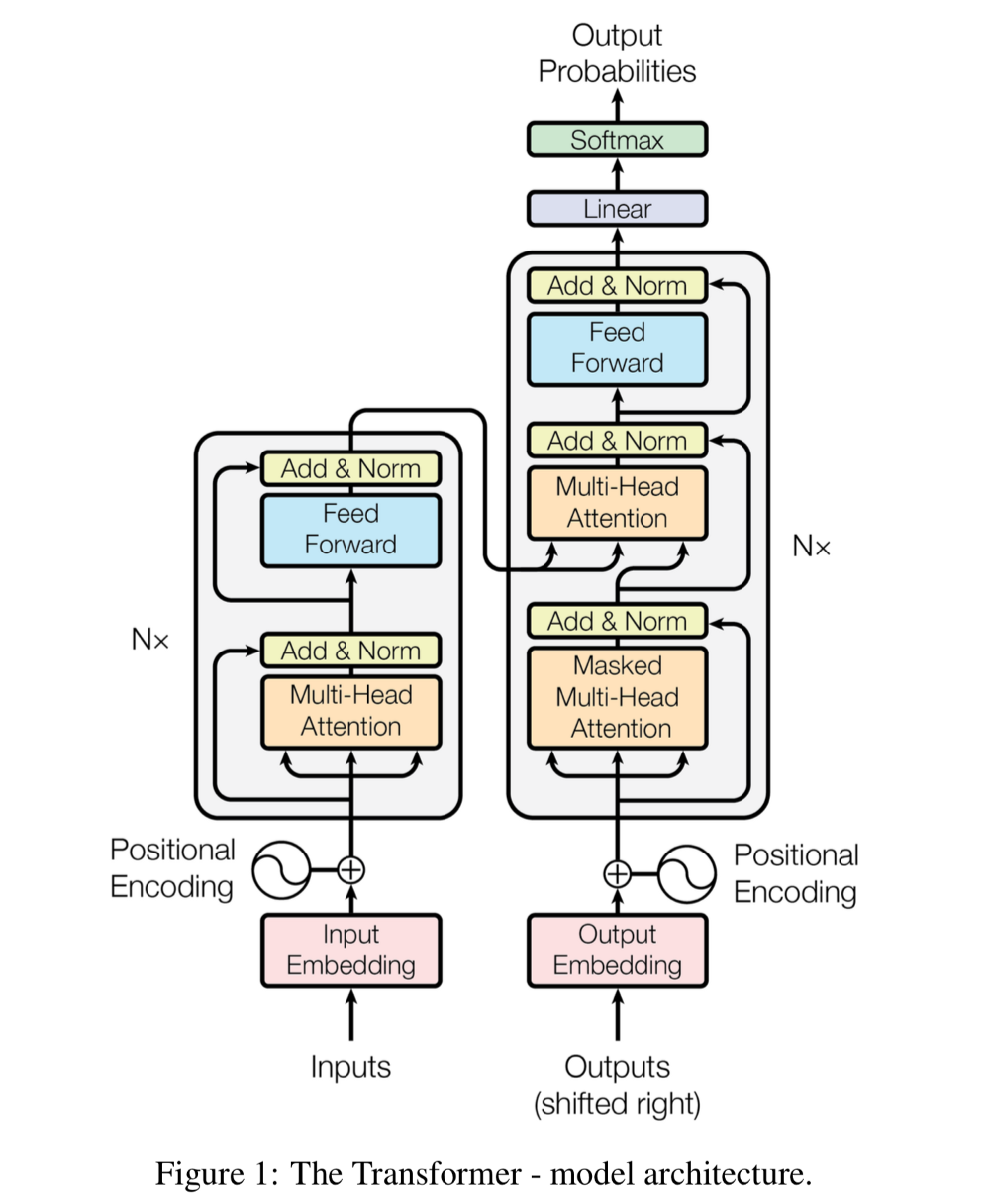
BERT hoạt động dựa trên Transformer, một mô hình sử dụng cơ chế attention để học mối tương quan giữa các từ trong văn bản. Transformer gồm hai phần chính: Encoder và Decoder, trong đó BERT chỉ sử dụng phần Encoder. Khác với các mô hình directional chỉ đọc dữ liệu theo một chiều (trái sang phải hoặc ngược lại), BERT đọc toàn bộ dữ liệu trong một lần, giúp mô hình học ngữ cảnh của từ tốt hơn bằng cách sử dụng thông tin từ cả hai phía.

Cách hoạt động của Encoder trong BERT bắt đầu từ Input của Encoder, là một chuỗi các token được biểu diễn thành chuỗi các vector. Output của mô hình là chuỗi các vector có kích thước bằng với kích thước đầu vào. Trong quá trình huấn luyện, BERT sử dụng hai chiến lược chính để vượt qua giới hạn của các mô hình directional truyền thống: Masked LM (MLM) và Next Sentence Prediction (NSP).



Hình . Mô hình BERT với Masked LM

Với Masked LM (MLM), trước khi đưa vào BERT, 15% số từ trong chuỗi được thay thế bằng token [MASK]. Mô hình sẽ dự đoán từ bị thay thế dựa trên ngữ cảnh của các từ xung quanh. Quá trình này bao gồm việc thêm một lớp phân loại, nhân các vector đầu ra với ma trận embedding và tính toán xác suất của mỗi từ trong tập từ vựng bằng softmax. Hàm lỗi của BERT chỉ tập trung đánh giá các từ được đánh dấu [MASK], giúp mô hình hiểu ngữ cảnh tốt hơn dù hội tụ chậm hơn.



Hình 3. Kiến trúc mô hình Transformer

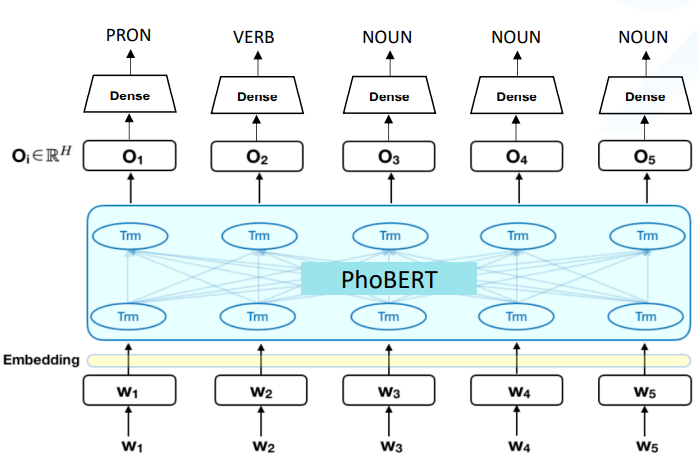
Còn đối với Next Sentence Prediction (NSP), trong chiến lược này, mô hình sử dụng một cặp câu làm dữ liệu đầu vào và dự đoán xem câu thứ hai có phải là câu tiếp theo của câu thứ nhất hay không. Dữ liệu huấn luyện gồm 50% cặp câu liên tiếp và 50% cặp câu ngẫu nhiên. Các bước xử lý bao gồm chèn token [CLS] và [SEP], đánh dấu các token trong từng câu, và thêm vector embedding biểu diễn vị trí của token trong câu. Sau đó, toàn bộ câu đầu vào được đưa vào Transformer để tính toán xác suất IsNextSequence bằng softmax.

Một điểm đặc biệt là BERT có thể được fine-tune cho nhiều bài toán khác nhau. Với bài toán Classification, ta thêm một lớp phân loại với input là output của Transformer cho token [CLS]. Với bài toán Question Answering, mô hình nhận đầu vào là đoạn văn bản và câu hỏi, và được huấn luyện để đánh nhãn cho câu trả lời trong đoạn văn bản. Với bài toán Named Entity Recognition (NER), mô hình được huấn luyện để dự đoán nhãn cho mỗi token, như tên người, tổ chức, địa danh,...

3.3. Pho-Bert

PhoBERT là một mô hình ngôn ngữ tiên tiến được thiết kế đặc biệt cho tiếng Việt. Được phát triển bởi VinAI Research, PhoBERT là phiên bản tối ưu nhất của mô hình BERT dành riêng cho tiếng Việt, giúp giải quyết nhiều bài toán xử lý ngôn ngữ tự nhiên (NLP) như phân loại văn bản, phân tích cảm xúc, và gán nhãn thực thể. Cái tên "PhoBERT" lấy cảm hứng từ món ăn phở nổi tiếng của Việt Nam, thể hiện tính độc đáo và bản địa hóa của mô hình.

Tiếng Việt có cấu trúc ngữ pháp và từ vựng phức tạp, dẫn đến việc áp dụng các mô hình ngôn ngữ chung không đạt hiệu quả cao. Trước khi có PhoBERT, các nhà nghiên cứu phải sử dụng các mô hình không chuyên biệt, dẫn đến kết quả không chính xác. PhoBERT khắc phục điều này bằng cách được huấn luyện trên một lượng lớn dữ liệu tiếng Việt, giúp mô hình hiểu sâu hơn về ngữ cảnh và cấu trúc ngữ pháp của tiếng Việt. Điều này cải thiện đáng kể hiệu quả trong các bài toán NLP liên quan đến tiếng Việt.



Hình 3. Mô hình PhoBERT

PhoBERT dựa trên kiến trúc Transformer giống như BERT, sử dụng cơ chế attention để học mối quan hệ giữa các từ trong câu. Có hai phiên bản của PhoBERT: PhoBERT-base và PhoBERT-large, tương tự như BERT, với số lượng tham số khác nhau để phù hợp với các yêu cầu và tài nguyên tính toán khác nhau. PhoBERT sử dụng phương pháp pre-training dựa trên RoBERTa, tối ưu hóa quy trình huấn luyện của BERT để đạt hiệu suất cao hơn.

PhoBERT sử dụng chiến lược Masked Language Model (MLM) trong quá trình huấn luyện, tương tự như BERT. Trong MLM, một phần từ trong câu được thay thế bằng token [MASK], và mô hình phải dự đoán từ bị che dấu dựa trên ngữ cảnh xung quanh. Điều này giúp mô hình học được ngữ cảnh và mối quan hệ giữa các từ tốt hơn.

PhoBERT đã được áp dụng rộng rãi trong nhiều bài toán NLP tiếng Việt và đã đạt được những kết quả đáng kể. PhoBERT đã thiết lập tiêu chuẩn mới cho các nhiệm vụ NLP tiếng Việt như gán nhãn từ loại (Part-of-speech tagging), phân tích phụ thuộc (Dependency parsing), nhận diện thực thể có tên (Named-entity recognition), và suy luận ngôn ngữ tự nhiên (Natural language inference).

PhoBERT vượt trội so với các phương pháp đơn ngữ và đa ngữ trước đây, đạt được hiệu suất tốt nhất trong các bài toán NLP tiếng Việt. Điều này chứng minh rằng một mô hình ngôn ngữ được huấn luyện riêng cho một ngôn ngữ cụ thể có thể mang lại kết quả tốt hơn so với các mô hình không chuyên biệt. PhoBERT không chỉ là công cụ mạnh mẽ cho các nhà nghiên cứu và phát triển trong lĩnh vực NLP, mà còn là minh chứng cho sự tiến bộ trong việc phát triển các mô hình ngôn ngữ cho các ngôn ngữ ít được chú ý hơn.

3.4. Natural Language Processing (NLP)

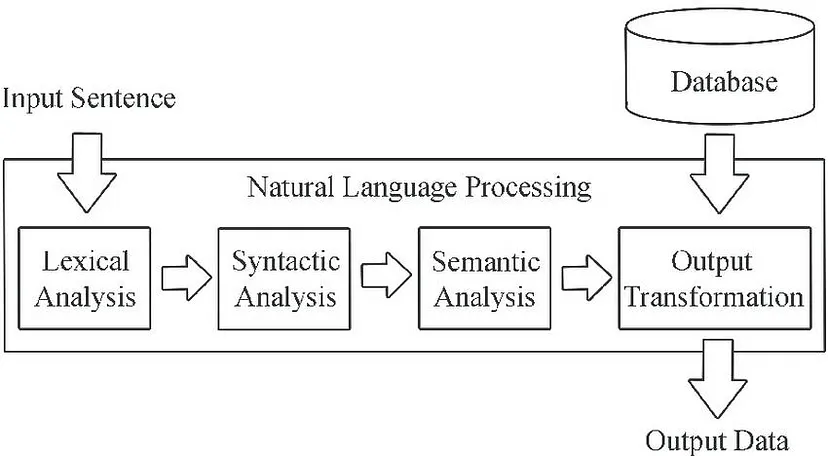
Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực thuộc khoa học máy tính và trí tuệ nhân tạo, tập trung vào việc cung cấp cho máy tính khả năng diễn giải, hiểu và tương tác với ngôn ngữ tự nhiên của con người. Công nghệ NLP cho phép xử lý và phân tích khối lượng lớn dữ liệu thoại và văn bản từ nhiều kênh liên lạc khác nhau như email, tin nhắn văn bản, bảng tin trên mạng xã hội, tệp video, tệp âm thanh và nhiều hơn nữa. Bằng cách sử dụng phần mềm NLP, các tổ chức có thể tự động xử lý và phân tích dữ liệu này, xác định ý định hoặc cảm xúc trong các tin nhắn, và phản hồi một cách hiệu quả.

NLP đóng vai trò rất quan trọng trong việc phân tích đầy đủ dữ liệu văn bản và giọng nói một cách hiệu quả. Công nghệ này có khả năng xử lý các nét khác biệt trong phương ngữ, tiếng lóng và các điểm bất thường về ngữ pháp thường thấy trong các cuộc hội thoại hàng ngày. Các công ty sử dụng công nghệ NLP cho nhiều tác vụ tự động như xử lý, phân tích và lưu trữ các tài liệu lớn, phân tích phản hồi của khách hàng hoặc bản ghi âm của tổng đài, chạy chatbot cho dịch vụ khách hàng tự động, và phân loại, trích xuất văn bản.

NLP có thể được tích hợp vào các ứng dụng tương tác trực tiếp với khách hàng để giao tiếp hiệu quả hơn. Ví dụ, chatbot sử dụng NLP để phân tích và phân loại các truy vấn của khách hàng, trả lời tự động các câu hỏi thường gặp và chuyển các truy vấn phức tạp đến bộ phận hỗ trợ khách hàng. Điều này giúp giảm chi phí, tiết kiệm thời gian cho nhân viên và cải thiện mức độ hài lòng của khách hàng.

Các doanh nghiệp sử dụng NLP để đơn giản hóa, tự động hóa và hợp lý hóa các hoạt động. Ví dụ, trong lĩnh vực bảo hiểm, pháp lý và chăm sóc sức khỏe, công nghệ NLP giúp chỉnh sửa thông tin nhận dạng cá nhân và bảo vệ dữ liệu nhạy cảm. Trong tương tác với khách hàng, chatbot và bot thoại sử dụng NLP để giao tiếp giống con người hơn, nâng cao chất lượng dịch vụ khách hàng. Các nhà tiếp thị sử dụng NLP để phân tích cảm nhận của khách hàng đối với sản phẩm hoặc dịch vụ của công ty, từ đó đưa ra những chiến lược kinh doanh phù hợp.

NLP kết hợp ngôn ngữ học điện toán, máy học và các mô hình học sâu để xử lý ngôn ngữ của con người. Ngôn ngữ học điện toán nghiên cứu và xây dựng các mô hình ngôn ngữ con người bằng công cụ máy tính và phần mềm. Máy học đào tạo máy tính bằng dữ liệu mẫu để nâng cao hiệu quả. Học sâu, một nhánh của máy học, tập trung vào việc dạy máy tính học và suy nghĩ như con người thông qua mạng nơ-ron.



Hình 3. Các bước trong quy trình NLP

Quá trình triển khai NLP bắt đầu bằng việc thu thập và chuẩn bị dữ liệu văn bản hoặc giọng nói từ các nguồn khác nhau. Phần mềm NLP sử dụng các kỹ thuật xử lý trước như token hóa, rút gọn từ, phục hồi nguyên thể từ và loại bỏ từ dừng để chuẩn bị dữ liệu cho các ứng dụng khác nhau. Sau đó, các nhà nghiên cứu sử dụng dữ liệu được xử lý trước và máy học để đào tạo các mô hình NLP. Cuối cùng, các mô hình này được triển khai hoặc tích hợp vào môi trường sản xuất hiện có để nhận dữ liệu đầu vào và đưa ra dự đoán cho các tác vụ cụ thể.

Các tác vụ NLP bao gồm gán nhãn từ loại, xử lý nhập nhằng nghĩa của từ, nhận dạng giọng nói, máy dịch, nhận dạng thực thể và phân tích cảm xúc. Ví dụ, gán nhãn từ loại giúp máy tính hiểu được cách các từ ngữ hình thành mối quan hệ về nghĩa với nhau trong câu. Nhận dạng thực thể xác định tên duy nhất cho con người, địa điểm, sự kiện, công ty và nhiều hơn nữa. Phân tích cảm xúc giúp diễn giải cảm xúc được truyền tải qua dữ liệu văn bản.

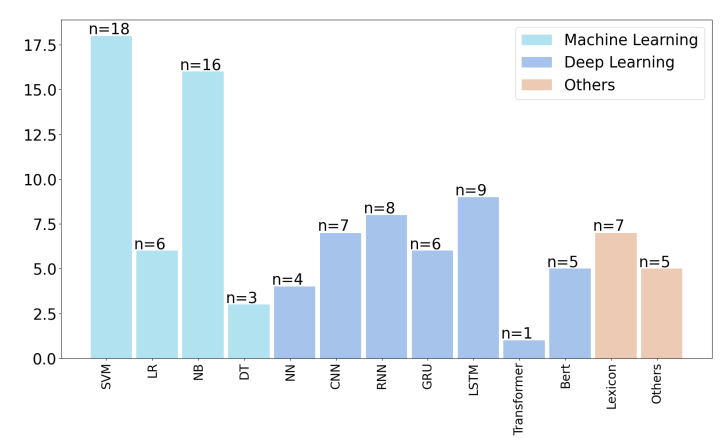
NLP có thể được triển khai theo nhiều phương thức khác nhau, bao gồm phương pháp có giám sát và không giám sát. Phương pháp có giám sát đào tạo phần mềm với tập dữ liệu đầu vào và đầu ra được gắn nhãn, trong khi phương pháp không giám sát sử dụng mô hình ngôn ngữ thống kê để dự đoán mẫu xảy ra khi được cung cấp dữ liệu đầu vào không gắn nhãn. Hiểu ngôn ngữ tự nhiên (NLU) và tạo ngôn ngữ tự nhiên (NLG) là hai nhánh con của NLP, tập trung vào phân tích ý nghĩa đằng sau các câu và tạo ra văn bản hội thoại giống như con người.

3.5. Phân tích cảm xúc (Sentiment Analysis)

Phân tích cảm xúc (Sentiment Analysis) là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (NLP) nhằm xác định và phân loại cảm xúc được biểu đạt trong văn bản. Với sự phát triển mạnh mẽ của các nền tảng trực tuyến, việc hiểu rõ cảm xúc của người dùng trở nên cực kỳ quan trọng đối với các tổ chức để đưa ra các quyết định thông minh.

Phân tích cảm xúc có thể được áp dụng trong nhiều lĩnh vực như kinh doanh, tài chính và áp dụng trong lĩnh vực bán lẻ ở nghiên cứu này. Việc hiểu rõ cảm xúc của khách hàng về sản phẩm và dịch vụ giúp cải thiện sự hài lòng của khách hàng và tăng cường uy tín thương hiệu. Điển hình như trong lĩnh vực bán lẻ, phân tích cảm xúc từ các bài báo, bài đăng hoặc các bình luận của khách hàng trên mạng xã hội hoặc các trang thương mại điện tử có thể giúp doanh nghiệp đánh giá cách nhìn nhận của người tiêu dùng về sản phẩm đồng thời có hướng phát triển tốt hơn.

Phân tích cảm xúc dựa trên nhiều kỹ thuật khác nhau như học máy, học sâu, và học tập hợp. Các kỹ thuật này bao gồm tiền xử lý dữ liệu, trích xuất đặc trưng, và phân loại. Các mô hình học sâu như LSTM và CNN đã được áp dụng rộng rãi và đạt được hiệu suất cao trong phân tích cảm xúc.



Hình 4. Các mô hình học máy và học sâu thường dùng.

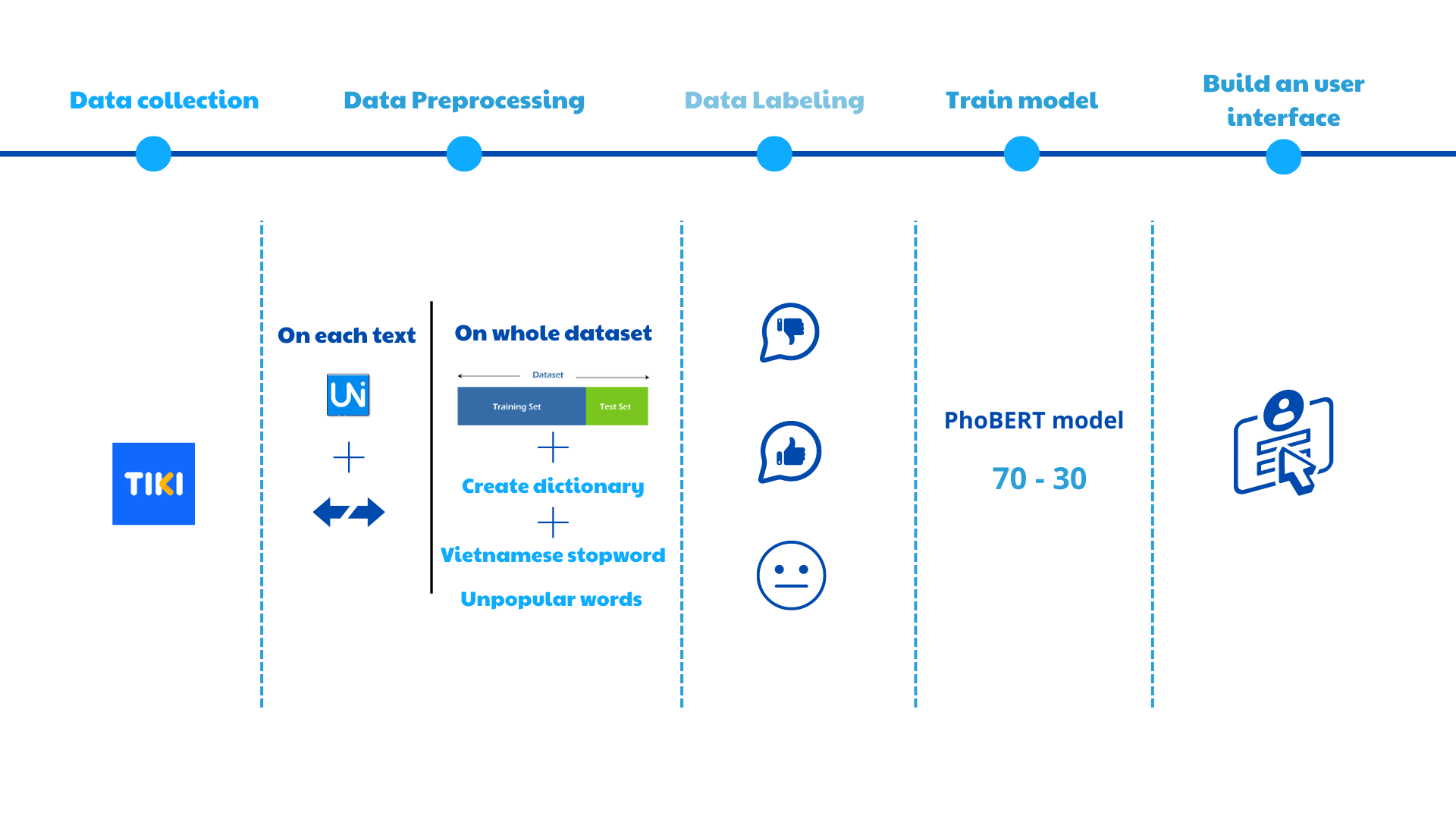
Một trong những thách thức lớn của phân tích cảm xúc là sự phức tạp của ngôn ngữ tự nhiên và sự đa dạng của các biểu đạt cảm xúc. Các nghiên cứu đã chỉ ra rằng việc sử dụng các mô hình ngôn ngữ chuyên biệt cho từng ngôn ngữ có thể cải thiện đáng kể hiệu quả của phân tích cảm xúc. Do đó để có cái nhìn sâu sắc nhất về phân loại cảm xúc thông qua bình luận tiếng việt, nghiên cứu đã tận dụng phoBERT - một mô hình Pre-trained về ngôn ngữ tiếng Việt của VinAi nhằm tối ưu hóa hiệu quả phân tích cảm xúc khách hàng.

Phân tích cảm xúc không chỉ là một công cụ mạnh mẽ cho các nhà nghiên cứu và phát triển trong lĩnh vực NLP, mà còn là minh chứng cho sự tiến bộ trong việc phát triển các mô hình ngôn ngữ cho các ngôn ngữ ít được chú ý hơn trên thế giới.

# Chapter 4. PROPOSED MODEL AND EXPERIMENT

## Chương này mô tả quy trình thực hiện mô hình PhoBERT để phân tích tâm trạng trên dữ liệu bình luận của Tiki. Chương bắt đầu với việc trình bày các bước chuẩn bị dữ liệu, bao gồm tiền xử lý và phân chia tập dữ liệu. Sau đó, chương giải thích cách cấu hình mô hình và các tham số đã sử dụng trong quá trình huấn luyện. Cuối cùng là kết quả đánh giá hiệu suất của mô hình, bao gồm độ chính xác, độ nhạy và các chỉ số khác, nhằm xác định hiệu quả của phân tích tâm trạng.

## 4.1. Bài toán và mô hình đề xuất



Hình. Mô hình đề xuất (do nhóm nghiên cứu đề xuất)

Trong bối cảnh ngày càng phát triển của thương mại điện tử và các nền tảng mạng xã hội, việc phân loại cảm xúc từ bình luận của khách hàng trở nên cực kỳ quan trọng. Để thực hiện điều này một cách hiệu quả, chúng tôi sử dụng mô hình PhoBERT, một mô hình ngôn ngữ dựa trên BERT được tối ưu hóa cho tiếng Việt. Dưới đây là các bước tiền xử lý dữ liệu để huấn luyện mô hình này nhằm phân loại cảm xúc từ bình luận của khách hàng.

Đầu tiên Nhóm sẽ thực hiện thu thập dữ liệu từ nguồn dữ liệu được thu thập từ nhiều nguồn khác nhau như mạng xã hội, diễn đàn, và các trang web đánh giá sản phẩm. Mỗi bình luận phải kèm theo nhãn cảm xúc tương ứng, ví dụ như tích cực, tiêu cực, hoặc trung lập. Điều này giúp đảm bảo rằng mô hình được huấn luyện trên dữ liệu đa dạng và phong phú.

Sau khi thu thập thành công, nhóm tiến hành tiền xử lý dữ liệu bằng một loạt các thao tác tiền xử lý dữ liệu văn bản đi từ cấp độ chi tiết từng từ đến toàn bộ tập dữ liệu. Đối với cấp độ từng văn bản, nhóm nghiên cứu thực hiện các thao tác xử lý như: chuẩn hóa Unicode, chuẩn hóa nguyên âm, dấu câu,... Đối với tiền xử lý toàn bộ tập dữ liệu, nhóm nghiên cứu thực hiện phân chia lại tập huấn luyện và tập kiểm tra và thực hiện các thao tác xử lý dữ liệu văn bản cơ bản như loại bỏ stop word, các ký tự đặc biệt, các từ ít phổ biến,...

Tiếp theo, Mỗi bình luận sẽ được gắn nhãn và định dạng dữ liệu tương ứng với các lượt rating và đánh giá của người dùng. Điều này giúp việc tải dữ liệu vào mô hình PhoBERT trở nên dễ dàng và thuận tiện.

Sau đó sẽ tiến hành bước quan trọng nhất của đề tài là huấn luyện mô hình phân loại cảm xúc ứng dụng phương pháp học sâu PhoBERT, tiến hành đi vào quy trình huấn luyện cho mô hình. Tại bước này thực hiện phân tách dữ liệu train, test và tokenizer để chuyển đổi văn bản thành các token mà mô hình có thể hiểu được. Tiếp theo là huấn luyện mô hình phân loại cảm xúc dựa trên bộ dữ liệu train, đánh giá hiệu suất mô hình bằng bộ dữ liệu test.

Bước cuối cùng để hoàn thành đồ án, nhóm sẽ xây dựng giao diện tương tác với người dùng để áp dụng ứng dụng mô hình phân loại cảm xúc vào thiết kế và xây dựng một giao diện tương tác với người dùng.

## 4.2. Phương pháp thu thập dữ liệu

Đối với bộ dữ liệu được sử dụng trong nghiên cứu bao gồm bộ dữ liệu về các cuốn sách được đăng bán trên nền tảng thương mại điện tử Tiki và bộ dữ liệu về các bình luận đánh giá về chất lượng của các sản phẩm sách mà họ đã từng tiêu thụ. Nhóm nghiên cứu kế thừa dữ liệu từ một nghiên cứu được đăng tải công khai nên nền tảng Kaggle.com với tiêu đề *Tiki Books Dataset.*  Bộ dữ liệu bao gồm các tệp dữ liệu nhỏ chứa thông tin về 2024 cuốn sách bestselling và đánh giá của khách hàng trên nền tảng thương mại điện tử Tiki.

Dữ liệu ban đầu bao gồm 3 tệp như sau:

| File name | Column name | Description | File type |
| --- | --- | --- | --- |
| book\_data | product\_id | ID of the product in the Tiki database (unique) | CSV |
|  | title | Name of the book, maybe contain republish time |
|  | authors | Same with it's name |
|  | original\_price | Price at the first time |
|  | current\_price | Price at present if having a discount |
|  | quantity | Total number of books sold of all time |
|  | category | Kind of book |
|  | n\_review | Number of reviews |
|  | avg\_rating | Average rating (max 5.0) |
|  | pages | Total pages of each book |
|  | manufacturer | Name of publisher |
|  | cover\_link | Link to the website to buy the book on Tiki |
| book\_id | id | ID of all products when I crawl. You can ignore | CSV |
| comments | product\_id | Same with book\_data file | CSV |
|  | comment\_id | Each comment has individual id |
|  | title | Keyword of comment |
|  | thank\_count | Number of like of other people |
|  | customer\_id | Each customer has individual id |
|  | rating | Average rating of the comment |
|  | content | Same with it's name |

4.3. Tiền xử lý dữ liệu

4.3.1. **Tiền xử lý dữ liệu đánh giá**

4.3.1.1. Chuyển đổi từ viết tắt về nguyên bản

Đầu tiên, nhóm nghiên cứu tiến hành định nghĩa từ viết tắt hoặc tiếng lòng và chuyển đổi chúng về trạng thái đầy đủ ban đầu như ko - không, ok - được,...

Một số từ viết tắt, từ lóng được nhóm nghiên cứu phân tích và định nghĩa:

| **Từ viết tắt** | **Từ gốc** |
| --- | --- |
| ko | không |
| k | không |
| không | không |
| hok có | không có |
| ho có | không có |
| cx | cũng |
| ok | được |
| ntn | Như thế nào |
| r | rồi |
| vs | với |
| bt | bình thường |
| dc | được |
| đc | được |
| tl | trả lời |
| thk | thích |

Sau khi định nghĩa và xây dựng hàm thì dữ liệu bình luận ban đầu đã được xử lý sơ bộ về mặt từ viết tắt, từ lóng xuất hiện trong câu. Bằng cách xây dựng một hàm xử lý các từ viết tắt, từ lóng:

* Sử dụng vòng lặp for duyệt qua các cặp từ viết tắt và dạng đầy đủ của từ đó trong từ điển
* Sử dụng hàm re.sub từ thư viện re (Regular Expression) của python để tìm kiếm và truy xuất các “abbr” thành “full\_form” của nó trong phạm vi từ điển
* Ngoài ra, xây dựng một biểu thức chính quy để đảm bảo chit thay thế nhgữn từ viết tắt chính xác (được phân chia bằng ranh giới từ “\b” mà không phải là một phần của từ khác. Ví dụ: hàm sẽ chỉ được kích hoạt khi “ASAP” đứng một mình, còn nếu “ASAP” xuất hiện trong các từ khác như “ASAPly” thì hàm chuyển đổi sẽ không hoạt động.

4.3.1.2. Mã hóa dấu câu và nguyên âm

Để thực hiện được thao tác mã hóa nguyên âm trong các văn bản Tiếng Việt, nhóm nghiên cứu thực hiện xây dựng thủ công một thư viện bảng nguyên âm và bảng ký tự dấu câu trong Tiếng Việt.

| **Nguyên âm có dấu** | **Định nghĩa nguyên âm** |
| --- | --- |
| 'a', 'à', 'á', 'ả', 'ã', 'ạ' | 'a' |
| 'ă', 'ằ', 'ắ', 'ẳ', 'ẵ', 'ặ' | 'aw' |
| 'â', 'ầ', 'ấ', 'ẩ', 'ẫ', 'ậ' | 'aa' |
| 'e', 'è', 'é', 'ẻ', 'ẽ', 'ẹ' | 'e' |
| 'ê', 'ề', 'ế', 'ể', 'ễ', 'ệ' | 'ee' |
| 'i', 'ì', 'í', 'ỉ', 'ĩ', 'ị' | 'i' |
| 'o', 'ò', 'ó', 'ỏ', 'õ', 'ọ' | 'o' |
| 'ô', 'ồ', 'ố', 'ổ', 'ỗ', 'ộ' | 'oo' |
| 'ơ', 'ờ', 'ớ', 'ở', 'ỡ', 'ợ' | 'ow' |
| 'u', 'ù', 'ú', 'ủ', 'ũ', 'ụ' | 'u' |
| 'ư', 'ừ', 'ứ', 'ử', 'ữ', 'ự' | 'uw' |
| 'y', 'ỳ', 'ý', 'ỷ', 'ỹ', 'ỵ' | 'y' |

| **Tên dấu câu** | **Ký tự dấu câu** |
| --- | --- |
| '' | ngang |
| 'f' | huyền |
| 's' | sắc |
| 'r' | hỏi |
| 'x' | ngã |
| 'j' | nặng |

* Đối với mã hóa dấu câu

Tại đây, nhóm nghiên cứu sử dụng hàm normalize trong python với module “unicodedata” để chuẩn hóa Unicode, cụ thể là chuẩn hóa văn bản sang dạng "Normalization Form Composed" - NFC. Tức là mỗi dấu câu sẽ được hiển thị bằng mã Unicode của các dấu khi kết hợp trong quá trình huấn huyện về dạng NFC.

* Đối với mã hóa nguyên âm

Đầu tiên, nhóm nghiên cứu thực hiện kiểm tra tính hợp lệ của các từ Tiếng Việt trong các văn bản bằng cơ chế xem xét vị trí của các nguyên âm trong từ. Từ tiếng Việt hợp lệ là từ có các nguyên âm đứng liền kề nhau theo quy tắc ngữ pháp tiếng Việt. Tại đây có 1 trường hợp đặc biệt nhóm nghiên cứu đề xuất cần được chuẩn hóa liên quan đến nguyên âm đó là: các trường hợp đặc biệt với các từ có “qu” hoặc “gi”.

***Xử lý trường hợp đặc biệt chỉ có 1 nguyên âm hặc “qu”, “gi”***

Đầu tiên, thực hiện kiểm tra và xác định các từ có “qu” hoặc “gi”. Sau khi đã xác định được:

* Các từ có nhiều nguyên âm: Thực hiện đặt dấu câu phù hợp vào các từ đó bằng cách định nghĩa vị trí của nguyên âm đó trong từ và đặt dấu câu vào vị trí phù hợp nếu có.
* Các từ có 2 nguyên âm: Đặt dấu câu vào nguyên âm phù hợp (thường là nguyên âm cuối cùng nếu nó không phải là nguyên âm cuối của từ).
* Các từ có 3 nguyên âm trở lên: Đặt dấu câu vào nguyên âm thứ 2

Cuối cùng là ghép các ký tự đã được chuẩn hóa lại thành từ hoàn chỉnh

***Xử lý dấu câu***

Kết hợp các thao tác chuyển đổi tất cả các từ về dạng thường và tách câu thành danh sách các từ dựa trên khoảng trắng. Mỗi từ sẽ được xử lý riêng biệt.

Nhóm nghiên cứu thực hiện tách từ ra khỏi các dấu xung quanh:

* Phần 1: Bắt đầu bằng các ký tự không phải chữ (dấu câu) ở đầu từ.
* Phần 2: Phần chính của từ (các ký tự chữ cái).
* Phần 3: Các ký tự không phải chữ (dấu câu) ở cuối từ.

Thực hiện kiểm tra trên từ đã tách nếu ký tự trong từ đó đang được xét có dấu, lưu lại dấu này và biến dau\_cau và thay thế nó bằng nguyên âm không dấu được nhóm nghiên cứu xây dựng trước.

4.3.1.3. Chuyển đổi thành từ thường và tách từ

Sau khi đã hoàn thành mã hóa dấu câu cho các nguyên âm trong văn bản, nhóm nghiên cứu thực hiện chuyển đổi tất cả các ký tự về dạng thường không in hoa và tách các từ trong văn bản.

4.3.1.4. Chuẩn hóa câu

Với các từ chứa các ký tự nguyên âm có dấu không mong muốn như các nguyên âm có dấu trong từ điểm nguyên âm nhóm đã xây dựng trước đó, thực hiện loại bỏ các ký tự không mong muốn và thay thế các khoảng trắng thừa chuyển đổi chúng về dạng một khoảng trắng duy nhất.

4.3.2. Tiền xử lý dữ liệu trên toàn bộ tập dữ liệu

4.3.2.1. Chuẩn hóa và xuất dữ liệu vào tập train và test

Tại bước này, thực hiện khởi tạo danh sách các nhãn (cột) trong toàn bộ tập dữ liệu. Mỗi tên cột sẽ đại diện cho một nhãn (category).

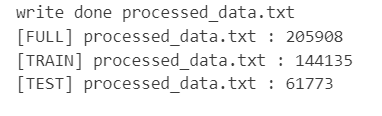


Image . Data after preprocessing

4.3.2.2. Xử lý dữ liệu đánh giá trùng lặp và đánh giá trống

Với một số dòng dữ liệu đánh giá bị đánh trùng lặp về ID của nó, nhóm nghiên cứu thực hiện trích xuất và loại bỏ đi các dòng dữ liệu đánh giá bị trừng và chỉ giữ lại 1 dòng duy nhất. Ngoài ra, sự xuất hiện của các dòng dữ liệu đánh giá trống khiến cho dữ liệu không được cân bằng, dễ ảnh hưởng đến hiệu suất huấn luyện mô hình, nhóm nghiên cứu cũng thực hiện loại bỏ các dòng đánh giá có cột “content” rỗng.

Với danh sách category đó, nhóm nghiên cứu thực hiện chuyển các nhãn về đúng tập dữ liệu phù hợp (tập huấn luyện hay tập kiểm tra) sau đó tạo ra hai file tập huấn luyện và kiểm tra với đường dẫn tương ứng.

4.3.2.3. Xây dựng tập dữ liệu train và test mới

Tại đây, thực hiện kết hợp dữ liệu từ các tệp huấn luyện và kiểm tra trong thư mục thành một tệp dữ liệu tổng hợp lớn. Sau đó, chia tệp dữ liệu tổng hợp này thành hai phần: tập huấn luyện và tập kiểm tra theo tỷ lệ 70% dữ liệu được sử dụng cho huấn luyện và 30% cho kiểm tra. Quá trình này giúp chuẩn bị dữ liệu cho việc huấn luyện và đánh giá mô hình học máy một cách có tổ chức và hiệu quả.

4.3.2.4. Tạo từ điển từ dữ liệu văn bản

Từ tập huấn luyện đã được chia trước đó, nhóm nghiên cứu thực hiện tạo một bộ từ điển gồm các từ trong tệp văn bản, với mỗi từ là một khóa và giá trị của khóa là số lần xuất hiện của từ đó. Nhóm nghiên cứu thu được kết quả về kích thước của từ điển, tức là số lượng từ khác nhau trong từ điển là 28743.

Sau khi tổng hợp và tạo từ điển, nhóm nghiên cứu thực hiện xử lý từ điển bằng cách loại bỏ stop word và các từ không phổ biến hoặc chứa số, sau đó lưu từ điển cuối cùng vào tệp.

* Xóa bỏ stop word

Sử dụng tệp stop word trong tiếng việt được cung cấp một cách công khai trên nhiều nền tảng *vietnamese-stopwords* để lọc ra những từ không nằm trong danh sách này và tạo một từ điển mới. Và kích thước của từ điển sau bước này giảm xuống còn 28404.

* Loại bỏ các từ ít phổ biến và các từ chứa số từ từ điển

Tại đây, nhóm nghiên cứu thực hiện tạo một từ điển mới chỉ chứa các từ có số lần xuất hiện lớn hơn hoặc bằng 30 và không chứa ký tự số. Và kích thước của từ điển sau bước này giảm xuống còn 4754.

## 4.4. Gắn nhãn chủ đề

Trước khi thực hiện gắn nhãn chủ đề, dữ liệu văn bản được chuẩn bị thông qua việc tải và tiền xử lý. Đầu tiên, một cột nhãn ("Label") được tạo ra dựa trên cột đánh giá sao ("Rating"). Cụ thể, các đánh giá từ 1 đến 2 sao được gán nhãn "tiêu cực", đánh giá 3 sao được gán nhãn "trung lập", và đánh giá từ 4 đến 5 sao được gán nhãn "tích cực". Tiếp theo, chỉ những đánh giá có tiêu đề thuộc các nhóm hợp lệ như "Cực kì hài lòng", "Hài lòng", "Bình thường", "Rất không hài lòng", và "Không hài lòng" mới được giữ lại, các tiêu đề khác sẽ bị loại bỏ.

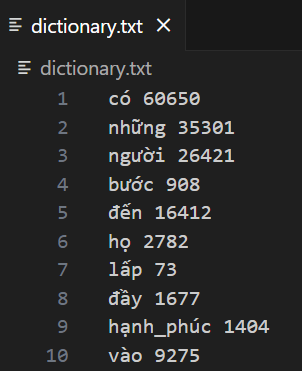
Để đảm bảo tính chính xác của dữ liệu, một bộ lọc cảm xúc được áp dụng nhằm loại bỏ các trường hợp đánh giá có tiêu đề tích cực nhưng lại có rating mặc định là thấp do người dùng quên thay đổi. Cụ thể, bộ lọc này chỉ giữ lại các đánh giá tích cực khi tiêu đề là "Cực kì hài lòng" hoặc "Hài lòng" và nhãn là "tích cực"; đánh giá trung lập khi tiêu đề là "Bình thường" và nhãn là "trung lập"; và đánh giá tiêu cực khi tiêu đề là "Rất không hài lòng" hoặc "Không hài lòng" và nhãn là "tiêu cực". Sau đó, các bản ghi có "comment id" trùng lặp được loại bỏ để đảm bảo mỗi đánh giá là duy nhất. Cuối cùng, dữ liệu đã qua xử lý được lưu lại dưới dạng file 'last\_dataset.csv' để sử dụng trong các bước tiếp theo của quá trình phân tích và xây dựng mô hình. Tập dữ liệu ‘last\_dataset.csv’ chứa các đánh giá về sản phẩm với các thông tin liên quan như tiêu đề, số lượng cảm ơn, đánh giá sao, và nội dung đánh giá.

Sau khi tải dữ liệu, cột ‘processed\_data’ được đổi tên thành ‘train’ để làm dữ liệu đầu vào cho quá trình xử lý. Các nhãn cảm xúc (‘label’) như “tích cực”, “trung lập”, và “tiêu cực” cũng được chuyển đổi thành các giá trị số 2, 1, và 0 tương ứng để chuẩn bị cho mô hình học máy.

Bảng 4. Định nghĩa nhãn chủ đề

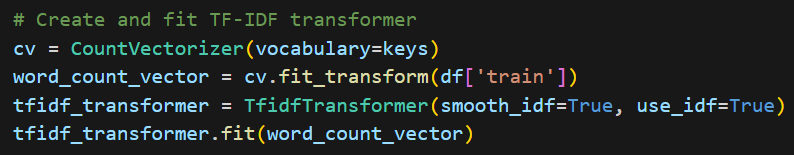
| **Nhãn chủ đề** | **Tên chủ đề** | **Mô tả ngắn gọn** |
| --- | --- | --- |
| 0 | Tiêu cực | Cảm xúc tiêu cực |
| 1 | Trung lập | Cảm xúc trung lập |
| 2 | Tích cực | Cảm xúc tích cực |

Quá trình gắn nhãn chủ đề sử dụng phương pháp TF-IDF (Term Frequency-Inverse Document Frequency) để biểu diễn các văn bản dưới dạng vector số. Trước hết, một tập từ điển (vocabulary) được tạo từ danh sách các từ trong file ‘dictionary.txt’.



Hình 4. Từ điển của bộ dữ liệu

Tập từ điển này sau đó được sử dụng để đếm tần suất xuất hiện của các từ trong từng văn bản thông qua ‘CountVectorizer’. Kết quả là một ma trận đếm từ được chuyển đổi thành ma trận TF-IDF, đại diện cho mức độ quan trọng của mỗi từ trong văn bản so với toàn bộ tập dữ liệu.



Hình .

Mỗi dòng văn bản sau khi xử lý sẽ được tính toán TF-IDF bằng cách gọi hàm ‘computing\_tfidf’, hàm này sử dụng mô hình ‘CountVectorizer’ và ‘TfidfTransformer’ đã được huấn luyện trước đó để chuyển đổi văn bản thành vector TF-IDF.



Mỗi ô trong cột tfidf là một từ điển Python (dictionary) với các từ khóa là từ trong văn bản và giá trị là điểm TF-IDF của từ đó. Các thành phần như “Từ khóa” là các từ trong văn bản được đưa vào từ điển như là các khóa (key), “Điểm TF-IDF” là các giá trị tương ứng với từng từ cho biết mức độ quan trọng của từ đó trong văn bản. Điểm này được tính dựa trên sự kết hợp của tần suất xuất hiện của từ (Term Frequency - TF) và độ hiếm của từ trong toàn bộ tập dữ liệu (Inverse Document Frequency - IDF).

Cách Tính Điểm TF-IDF cụ thể:

1. Term Frequency (TF): Đo lường tần suất của từ trong văn bản. TF = (Số lần từ xuất hiện trong văn bản) / (Tổng số từ trong văn bản).
2. Inverse Document Frequency (IDF): Đo lường độ hiếm của từ trong toàn bộ tập dữ liệu. IDF = log((Tổng số tài liệu) / (Số tài liệu chứa từ)).
3. TF-IDF Score: Là tích của TF và IDF. Công thức là:

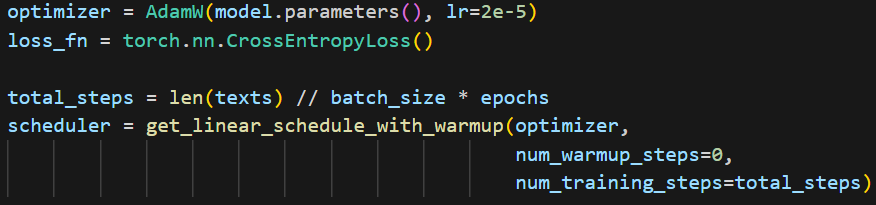
TF-IDF = TF × IDF

## 4.5. Xây dựng mô hình

Để xây dựng mô hình phân loại cảm xúc, nhóm đã sử dụng kết hợp giữa mô hình ngôn ngữ PhoBERT và các đặc trưng TF-IDF. PhoBERT là một mô hình ngôn ngữ được huấn luyện trước trên dữ liệu tiếng Việt, giúp mô hình có khả năng hiểu ngữ cảnh của các từ trong câu một cách chính xác.

Mô hình phân loại cảm xúc được xây dựng dựa trên lớp ‘SentimentClassifier’. Trong lớp này, đầu tiên là mô hình PhoBERT, sau đó là một lớp dropout để giảm thiểu hiện tượng overfitting, và cuối cùng là hai lớp fully connected để kết hợp đặc trưng của PhoBERT với các đặc trưng TF-IDF, trước khi đưa ra dự đoán cuối cùng.

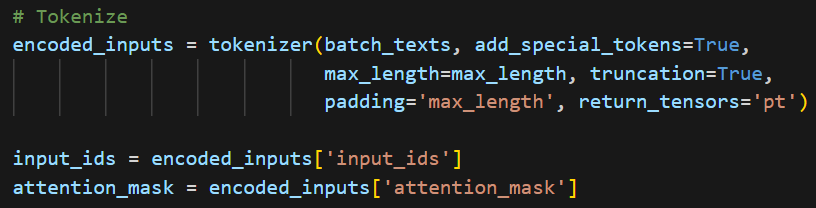
Mô hình được huấn luyện với tập dữ liệu đã được phân tách thành tập huấn luyện và tập kiểm thử. Các đặc trưng văn bản được token hóa và chuyển đổi thành định dạng phù hợp với mô hình PhoBERT. Đồng thời, các đặc trưng TF-IDF cũng được chuẩn bị sẵn để kết hợp với các đặc trưng ngôn ngữ của PhoBERT. Quá trình huấn luyện sử dụng ‘AdamW’ làm optimizer, với hàm mất mát là ‘CrossEntropyLoss’. Để điều chỉnh tốc độ học, ‘get\_linear\_schedule\_with\_warmup’ được sử dụng.



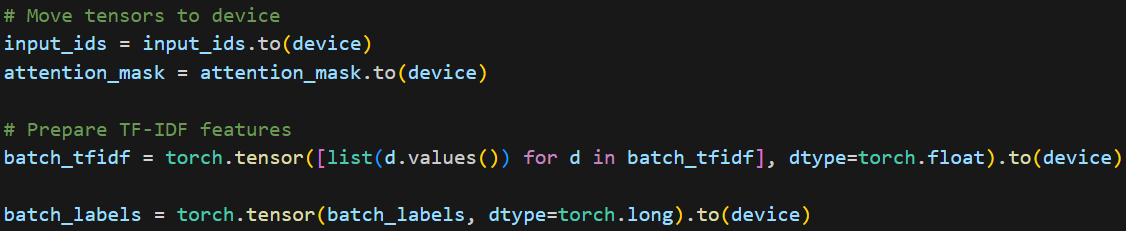
* **`optimizer`**: Tạo một đối tượng AdamW optimizer với tốc độ học là 2e-5. Đây là thuật toán tối ưu hóa được sử dụng để cập nhật trọng số của mô hình.
* **`loss\_fn`**: Tạo hàm mất mát CrossEntropyLoss, thường dùng cho bài toán phân loại với nhiều lớp.
* **`total\_steps`**: Tính số bước huấn luyện tổng cộng dựa trên số lượng văn bản, kích thước lô (batch size), và số epoch.
* **`scheduler`**: Tạo một đối tượng get\_linear\_schedule\_with\_warmup để điều chỉnh tốc độ học theo từng bước huấn luyện.

Mô hình được huấn luyện qua nhiều epoch, và hiệu suất của mô hình được đánh giá thông qua độ chính xác (accuracy) và độ mất mát (loss) sau mỗi epoch. Trong mỗi epoch, mô hình thực hiện các bước huấn luyện bao gồm:

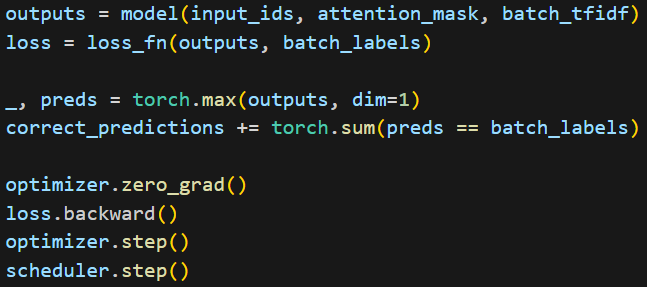
* **Token hóa văn bản**



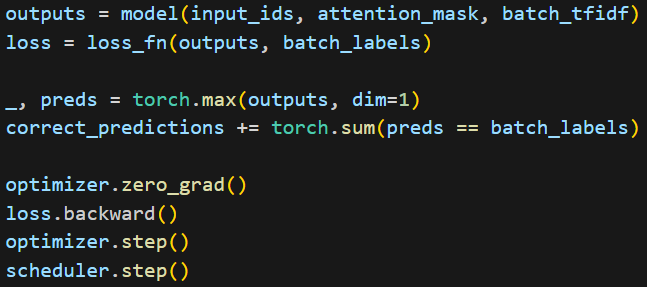
* **`encoded\_inputs`**: Token hóa văn bản trong lô, trả về các chỉ số của từ và mặt nạ attention.
* **`input\_ids`**, **`attention\_mask`**: Các tensor đầu vào và mặt nạ attention được chuyển đến thiết bị (CPU hoặc GPU).
* **Chuyển đổi TF-IDF thành tensor**



* **`batch\_tfidf`**: Chuyển đổi đặc trưng TF-IDF thành tensor và đưa vào thiết bị.
* **`batch\_labels`**: Chuyển đổi nhãn thành tensor và đưa vào thiết bị.
* **Tính toán dự đoán và mất mát**

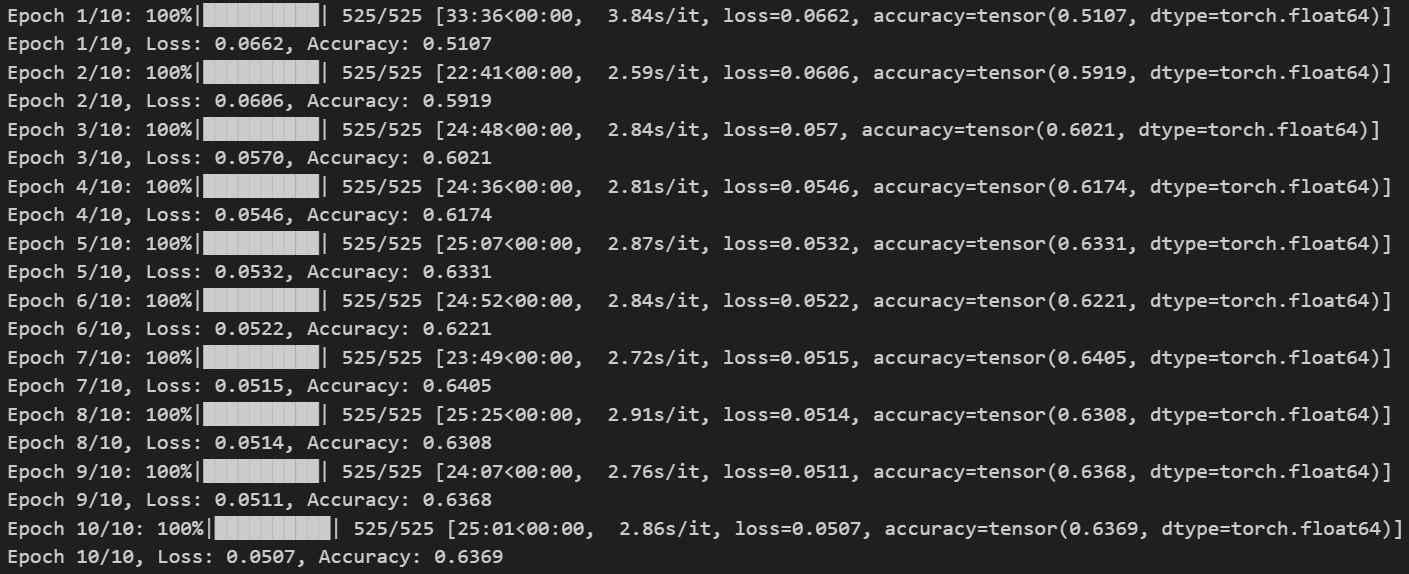


* **`outputs`**: Tính toán dự đoán từ mô hình với đầu vào `input\_ids`, `attention\_mask`, và batch\_tfidf.
* **`loss`**: Tính toán mất mát giữa dự đoán và nhãn thực tế.
* **Cập nhật trọng số mô hình**



* **`preds`**: Lấy dự đoán của mô hình (lớp có xác suất cao nhất).
* **`correct\_predictions`**: Cộng dồn số lượng dự đoán đúng.
* **`optimizer.zero\_grad()`**: Đặt gradient của các trọng số về 0.
* **`loss.backward()`**: Tính gradient của mất mát.
* **`optimizer.step()`**: Cập nhật trọng số của mô hình.
* **`scheduler.step()`**: Cập nhật lịch trình học.

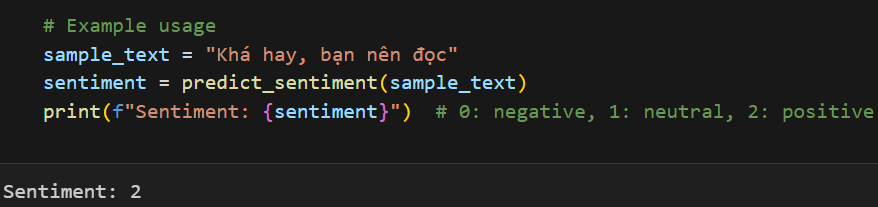
Sau khi huấn luyện, mô hình được đánh giá dựa trên tập dữ liệu kiểm thử để xác định hiệu suất của nó. Độ chính xác của mô hình được cải thiện qua từng epoch, cho thấy mô hình đang dần học được các đặc trưng quan trọng của dữ liệu. Mô hình sau khi đạt được kết quả tốt nhất sẽ được lưu lại để sử dụng cho việc dự đoán sau này.



Hình . Kết quả của mô hình huấn luyện

Kết quả huấn luyện cho thấy mô hình đạt được độ chính xác từ 51% đến 64% qua các epoch, cho thấy mô hình cải thiện dần theo thời gian.

Cuối cùng, nhóm sử dụng một hàm predict\_sentiment, được xây dựng để kiểm thử mô hình với các đoạn văn bản mới. Hàm này sẽ thực hiện token hóa văn bản, tính toán đặc trưng TF-IDF, và sau đó sử dụng mô hình đã huấn luyện để đưa ra dự đoán về cảm xúc của văn bản đó.



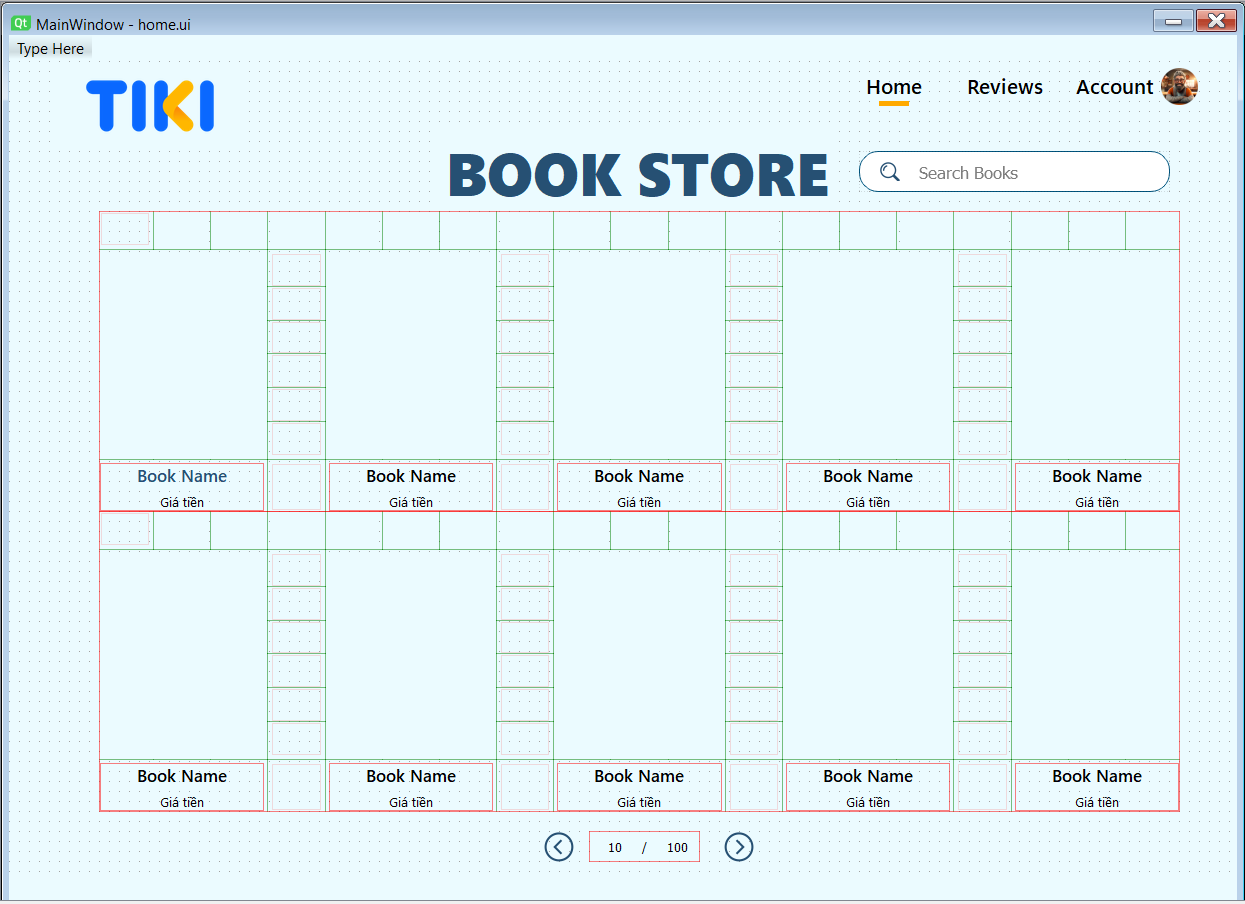
**Chapter 5. MODEL APPLICATION (Triển khai mô hình thành ứng dụng)**

Chương 5 mô tả triển khai ứng dụng phân tích tâm trạng sau khi huấn luyện mô hình. Ứng dụng có giao diện, trình bày kết quả thử nghiệm và phản hồi từ người dùng, đánh giá hiệu quả của mô hình và giao diện.

**5.1. User Interface**

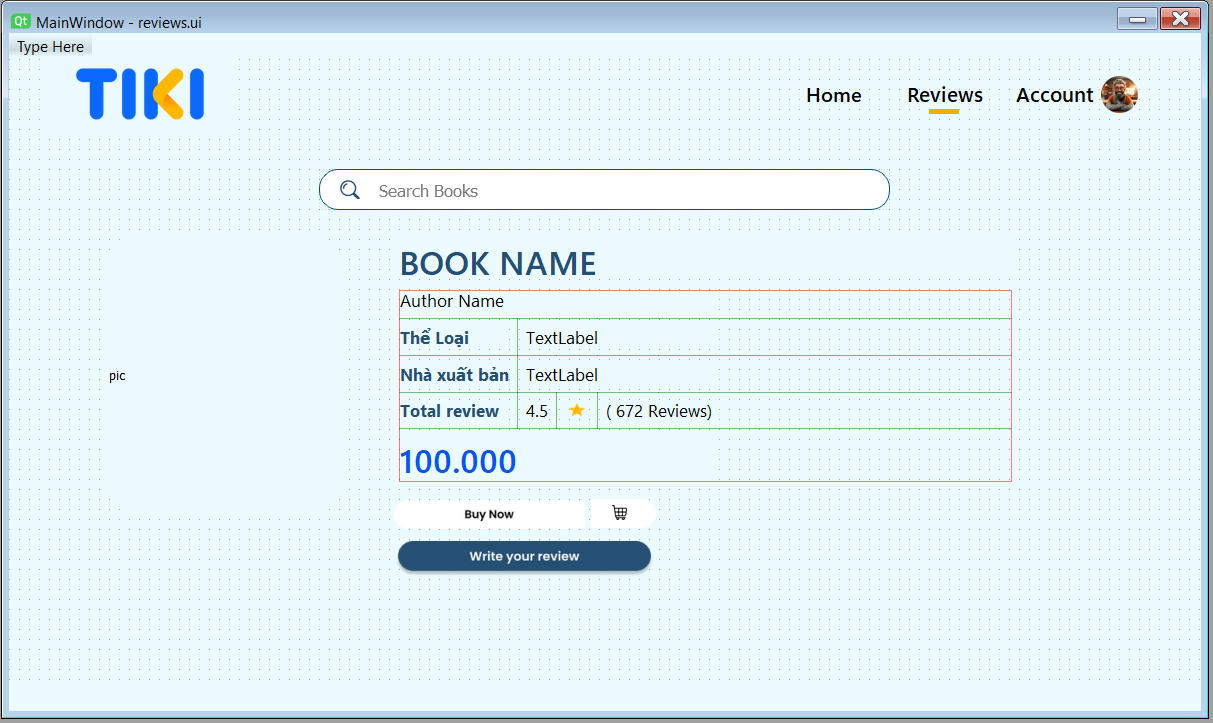
Sau khi thực hiện huấn luyện mô hình, nhóm nghiên cứu tiến hành thực hiện triển khai thiết kế một ứng dụng trực quan để có thể trình bày ứng dụng một cách dễ hiểu. Đối với phần thiết kế User Interface(UI), nhóm đã dùng phần mềm Qt Designer để thiết kế ứng dụng.

Đối với trang chủ của ứng dụng nhóm thực hiện thiết kế 1 cửa hàng sách hiển thị thông tin tên sách, giá và hình ảnh trang bìa của sách. Ngoài ra ở đây còn hiển thị mục tìm kiếm sách để khách hàng có thể tìm kiếm những cuốn sách phù hợp với sở thích của mình.

****

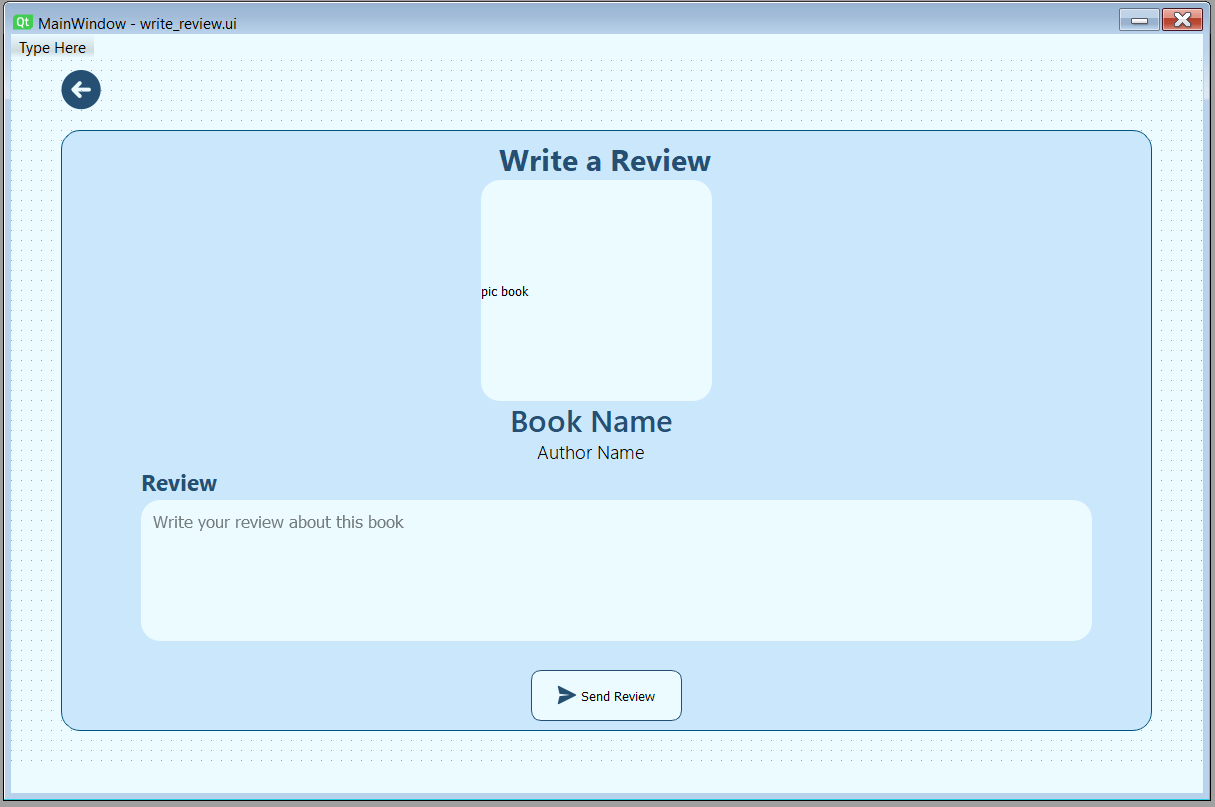
Hình . Trang chủ của cửa hàng sách

Sau đó, người dùng có thể thực hiện trỏ chuột vào ảnh hoặc tên sách để xem thông tin chi tiết về tác phẩm. Trong phần thông tin chi tiết của sách sẽ hiển thị các thông tin cơ bản như: tên sách, tên tác giả, thể loại, nhà xuất bản và tổng số bình luận và số lượng đánh giá trung bình. Tại đây, người dùng có thể thực hiện viết bình luận của mình bằng cách ấn vào nút “write your review”.

****

Hình . Trang thông tin chi tiết của sách

Sau khi thực hiện yêu cầu viết đánh giá, ứng dụng sẽ hiển thị trang mới để khách hàng thực hiện đánh giá tại textbox và thực hiện gửi bằng cách nhấn nút “Send Review”.

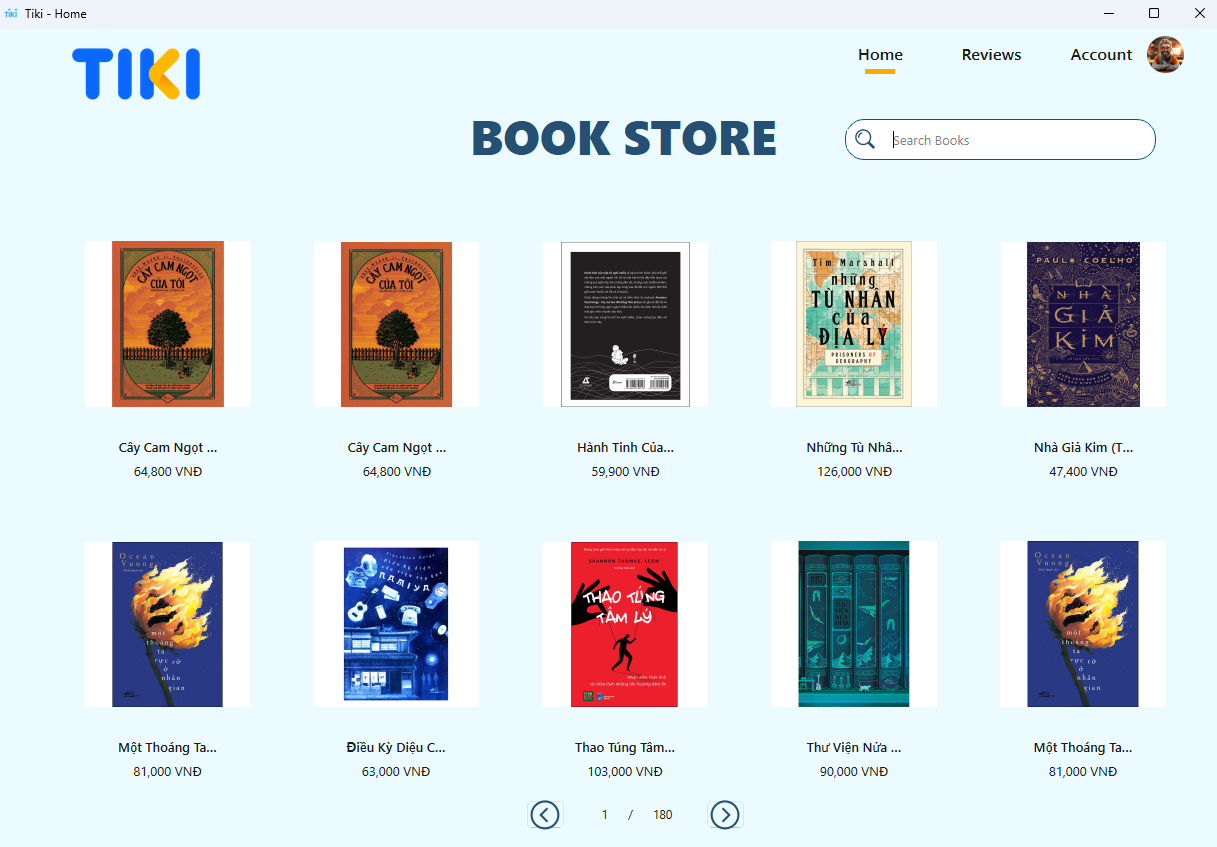
****

Hình . Trang viết đánh giá sách

**5.2.Interface implementation**

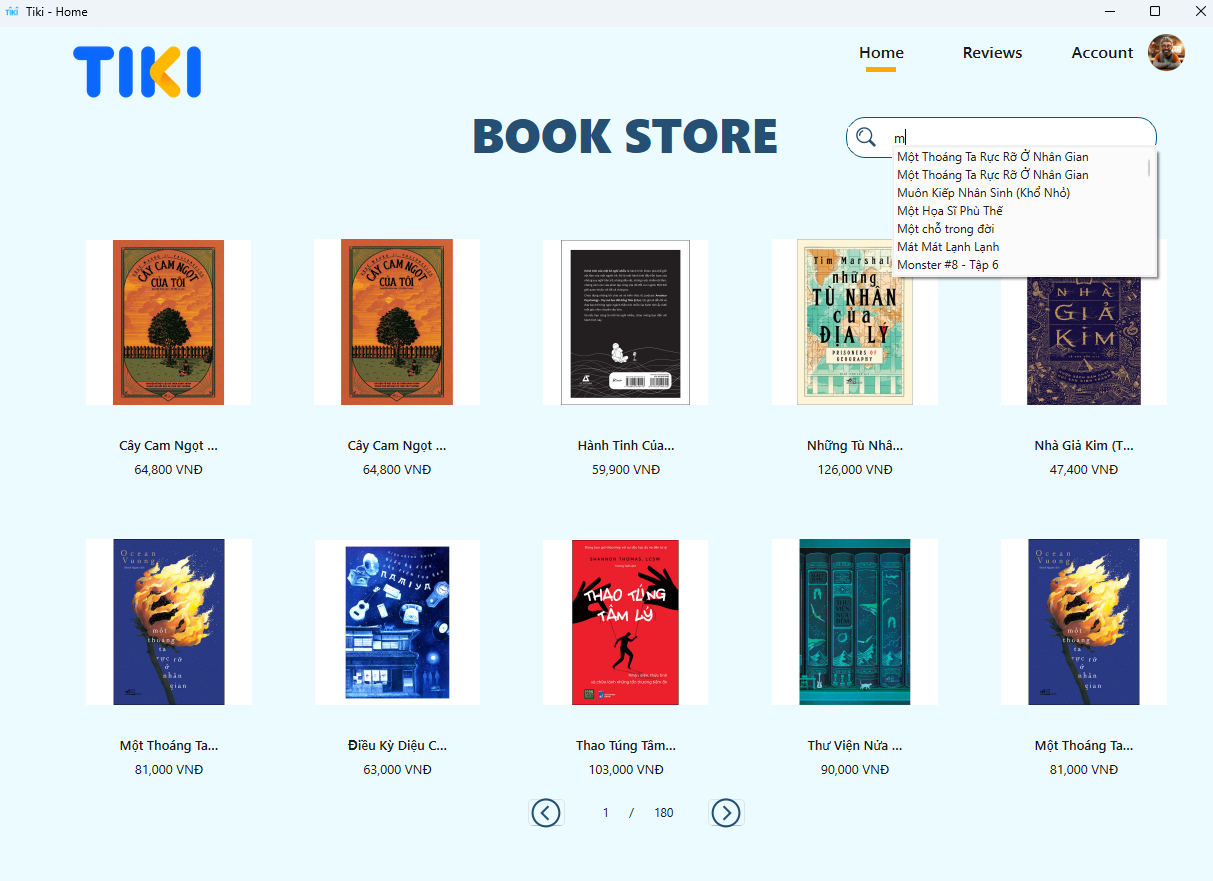
Sau khi thực hiện thiết kế ứng dụng bằng Qt Designer, nhóm đã thực hiển triển khai phần mềm bằng thư viện PyQt6.

5.2.1. Home Page



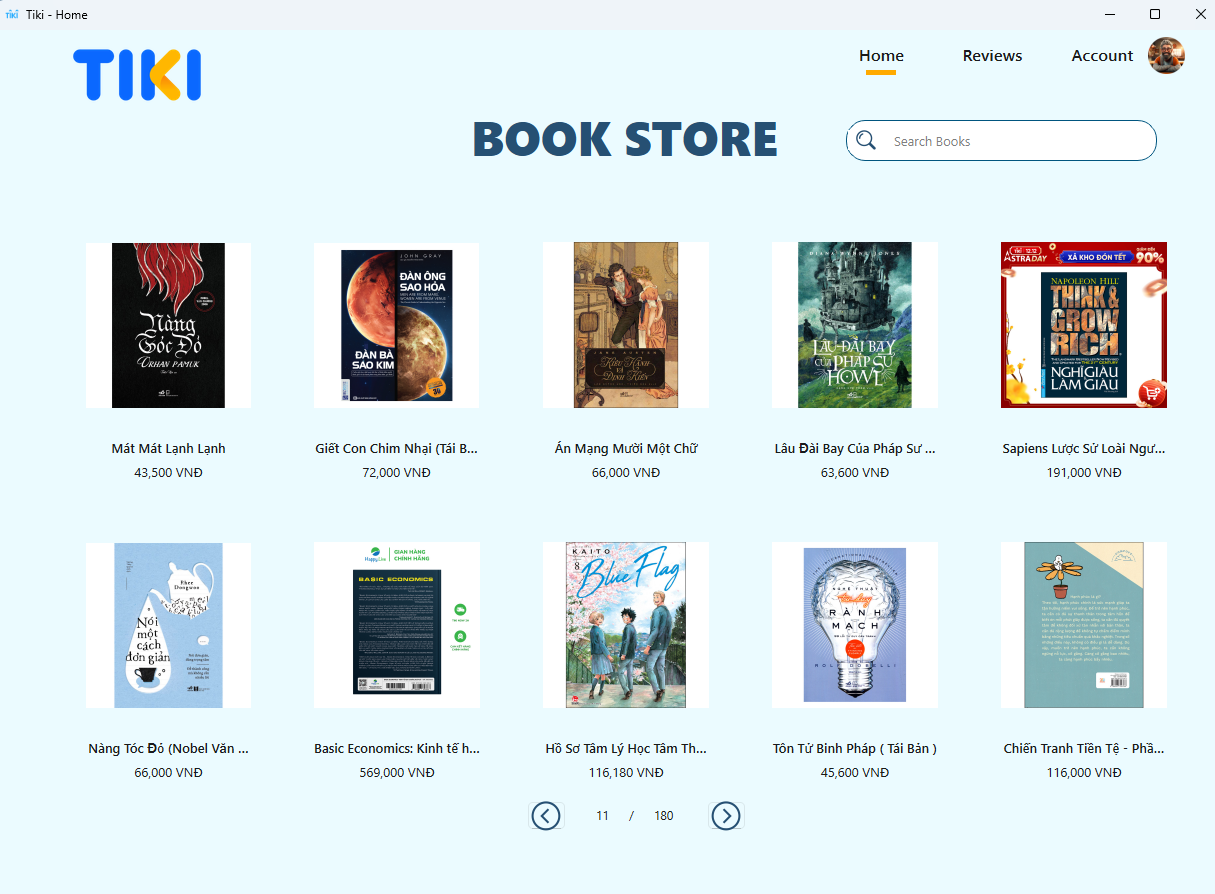
hình. Home Page

Ở giao diện Home, có thể quan sát được các quyển sách có trong dữ liệu, thể hiện tên, hình ảnh, mức giá. Tại đây, người dùng có thể click vào sản phẩm mà họ muốn review để chuyển đến trang sản phẩm. Đối với nút Reviews ở góc trên bên phải, nếu click vào người dùng mặc định sẽ được chuyển tới trang của quyển sách đầu tiên trong dãy sản phẩm là “Cây cam ngọt của tôi”.



Hình. Search bar ở Home Page

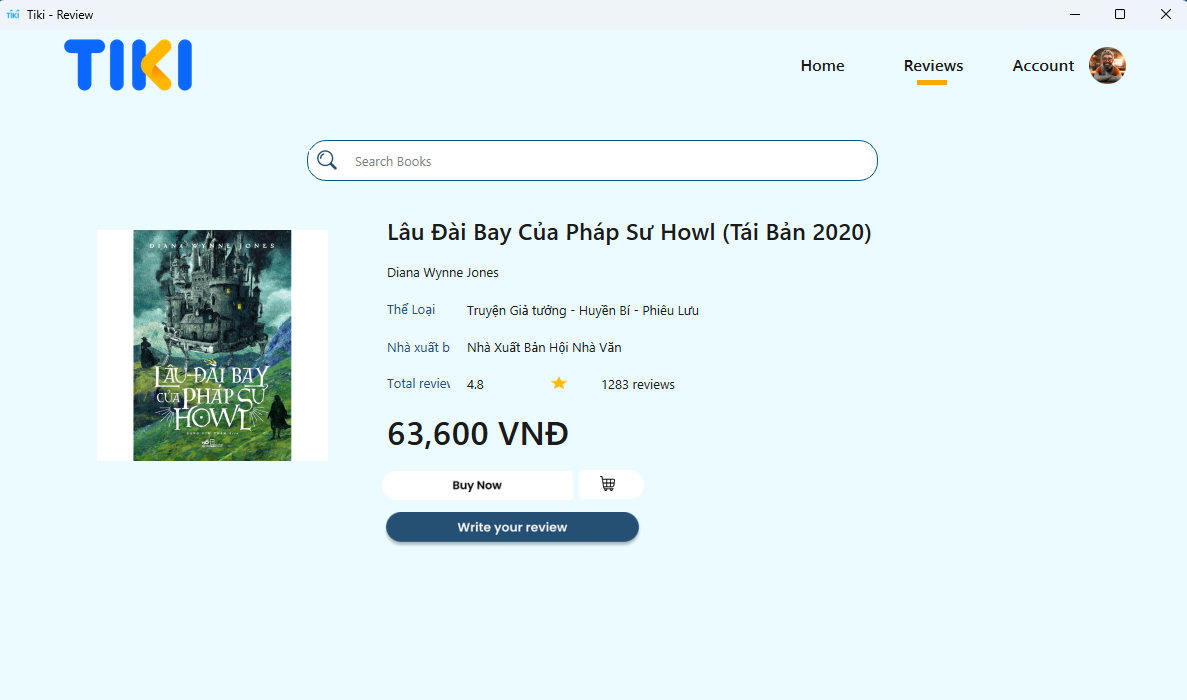
Người dùng có thể tìm kiếm sản phẩm muốn xem ở trang Home và truy cập đến sản phẩm đó khi nhấn vào nút search.



Hình. Thay đổi sản phẩm khi nhấn nút Next hay Previous

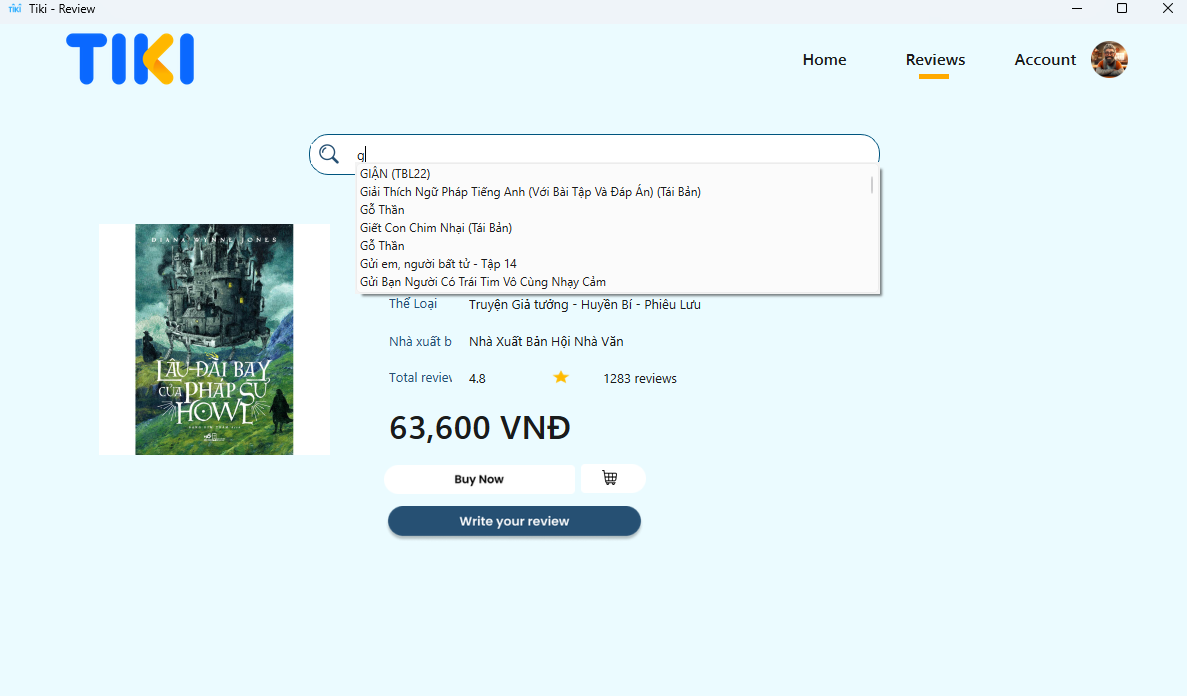
Có rất nhiều sản phẩm, người dùng có thể lướt với 180 trang sản phẩm.

5.2.2. Review page



hình. Review Page

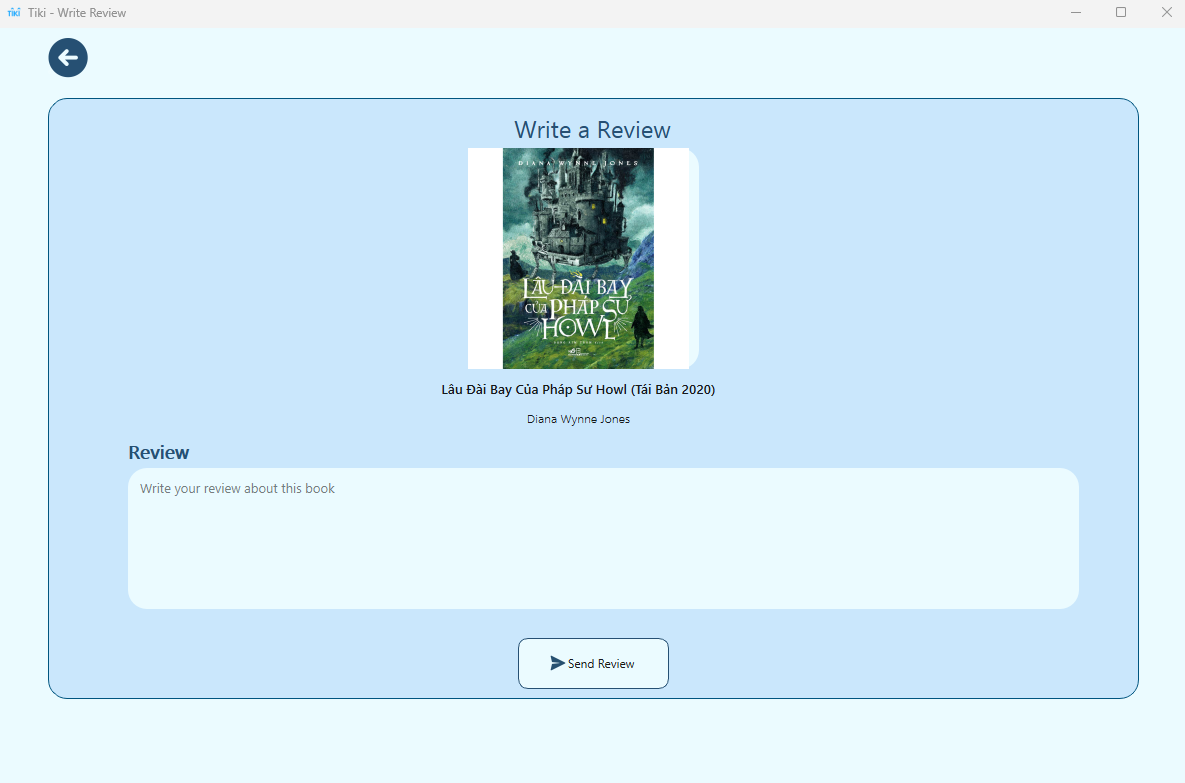
Được thiết kế để người dùng có thể mua sản phẩm, nhưng các chức năng của đồ án là giúp người dùng thực hiện review sản phẩm. Tại đây có thể xem thông tin về tác giả, nhà xuất bản, sao đánh giá và lượng người đánh giá. Review Page cũng có search bar, chức năng tương tự search bar tại Home Page.



Hình. chức năng search ở Review page

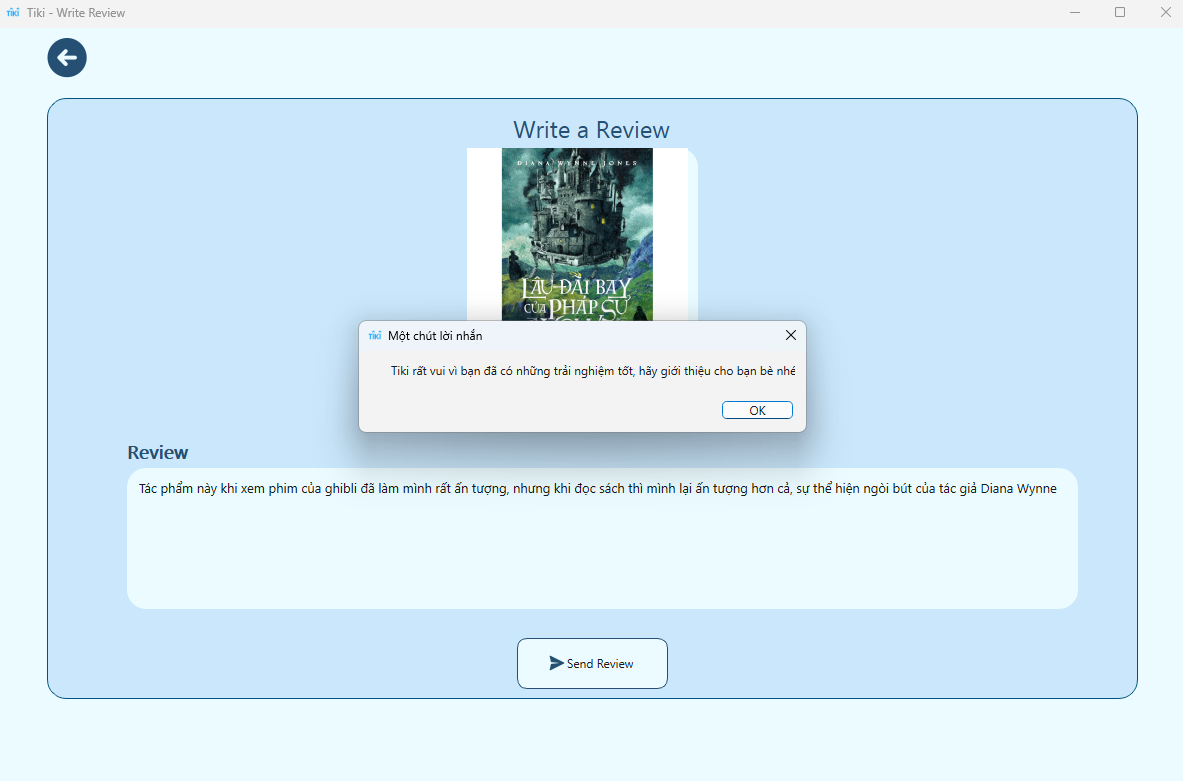
Khi người dùng muốn review sản phẩm, họ có thể nhấn vào nút “Write your review” phía dưới để chuyển đến trang Write Review Page.

5.2.3. Write Review Page



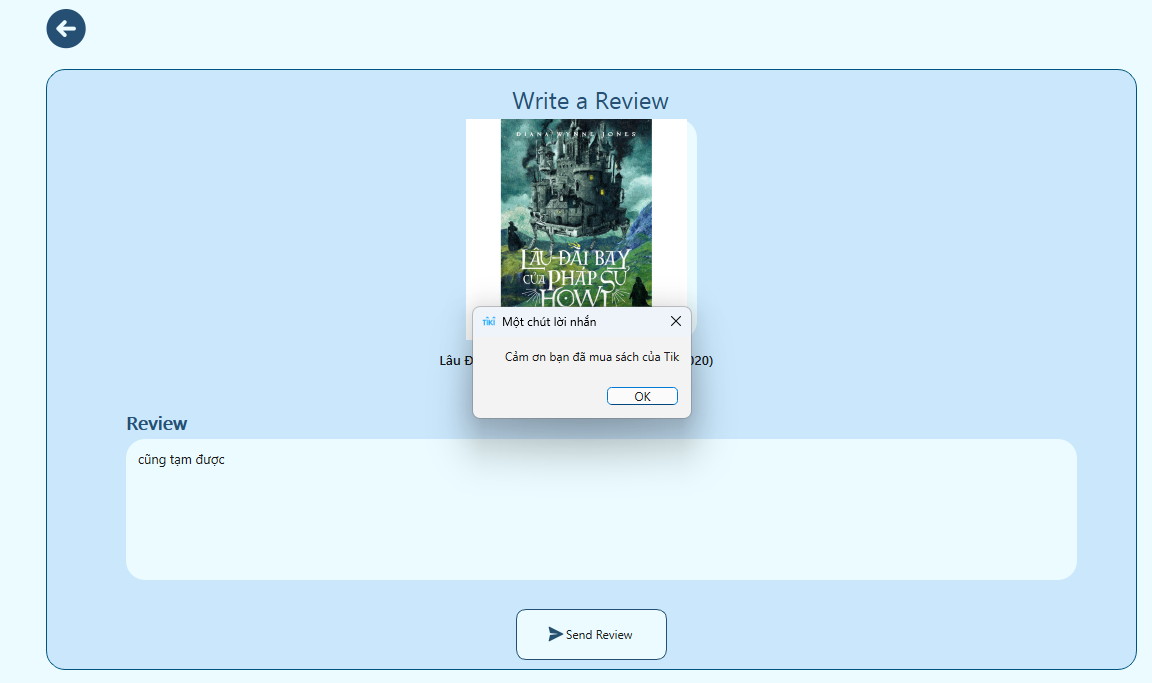
Hình. Write Review Page

Tại trang này, người dùng có thể viết review về sản phẩm, ví dụ với tác phẩm “lâu đài bay của pháp sư howl”, có nhận xét như sau “Tác phẩm này khi xem phim của ghibli đã làm mình rất ấn tượng, nhưng khi đọc sách thì mình lại ấn tượng hơn cả, sự thể hiện ngòi bút của tác giả Diana Wynne Jones làm mình đi từ bất ngờ này đến bất ngờ khác, mình thực sự đắm chìm vào thế giới xinh đẹp ấy.” Khi người dùng nhấn Send Review, câu review của họ sẽ được gửi đi và lưu vào data, ứng dụng sẽ đánh giá comment đó là tích cực, tiêu cực hay trung tính và hiển thị lên message box.

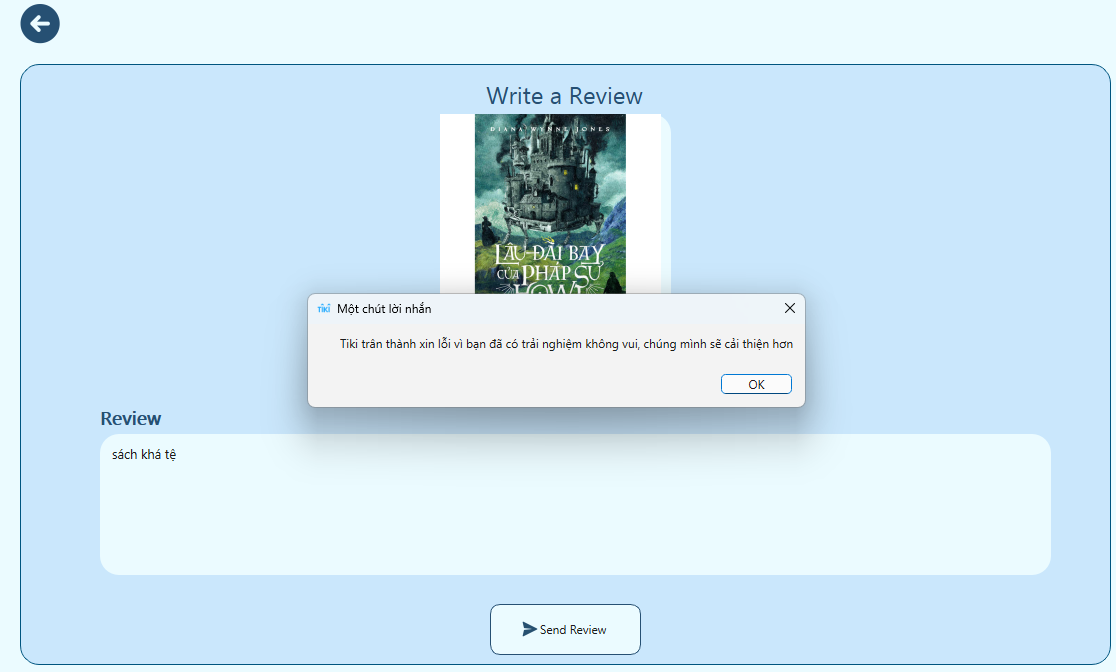


hình. MessageBox sau khi người dùng nhấn gửi comment. (MessageBox tích cực)

Tùy vào câu comment ở mức độ nào thì Messgae box ẽ đưa ra câu trả lời khác nhau.



Hình. MessageBox trung tính



Hình. Messagebox tiêu cực

**Chapter 6. CONCLUSION**

Chương 6 tóm tắt kết quả nghiên cứu về phân tích tâm trạng, nhấn mạnh hiệu quả của mô hình PhoBERT cũng chỉ ra hạn chế như độ chính xác và mở rộng cải tiến ứng dụng mô hình sang các lĩnh vực khác để nâng cao trải nghiệm người dùng.

**6.1. Conclusion**

Nghiên cứu này đã đạt được những kết quả quan trọng trong việc ứng dụng mô hình Pho-BERT để phân tích cảm xúc khách hàng dựa trên bình luận sản phẩm sách trên sàn thương mại điện tử Tiki. Từ những đánh giá được đăng tải bởi người dùng trên nền tảng Tiki, nhóm nghiên cứu đã thành công phân loại cảm xúc các đánh giá đó. Mô hình Pho-BERT, vốn được tối ưu hóa cho tiếng Việt, đã cho phép nhóm nghiên cứu thực hiện phân tích cảm xúc với độ chính xác cao, nhận diện các cảm xúc tích cực, tiêu cực và trung tính trong các bình luận của khách hàng. Điều này giúp tạo ra một cái nhìn rõ ràng hơn về sự hài lòng của khách hàng đối với các sản phẩm sách, từ đó hỗ trợ các nhà quản lý trong việc cải thiện chất lượng dịch vụ và sản phẩm.

Bên cạnh đó, nhóm nghiên cứu cũng thành công xây dựng giao diện tương tác người dùng cho phép tìm kiếm tên sách và xem các đánh giá liên quan đã giúp tăng cường khả năng tiếp cận và sử dụng thông tin cho người dùng. Giao diện này không chỉ cung cấp một cách dễ dàng để tìm kiếm thông tin về sách mà còn hiển thị các đánh giá một cách rõ ràng và trực quan, từ đó giúp người dùng đưa ra quyết định mua sắm thông minh hơn.

Những cải tiến này không chỉ nâng cao trải nghiệm người dùng trên nền tảng Tiki mà còn mở ra cơ hội để áp dụng các kỹ thuật phân tích cảm xúc trong các lĩnh vực khác. Kết quả từ nghiên cứu này cung cấp nền tảng vững chắc cho các nghiên cứu tiếp theo và ứng dụng thực tiễn, góp phần vào sự phát triển của các công cụ phân tích cảm xúc và cải thiện giao diện người dùng trong môi trường thương mại điện tử.

**6.2. Hạn chế**

Mặc dù đề tài ứng dụng mô hình Pho-BERT trong phân tích cảm xúc khách hàng đã đạt được những kết quả đáng khích lệ, vẫn tồn tại một số hạn chế quan trọng cần lưu ý.

* Kích thước của bộ dữ liệu huấn luyện còn khá nhỏ, điều này có thể hạn chế khả năng của mô hình trong việc nhận diện và phân tích đầy đủ các sắc thái cảm xúc trong bình luận. Một bộ dữ liệu phong phú và đa dạng hơn sẽ giúp mô hình học được nhiều khía cạnh của cảm xúc và từ vựng tiếng Việt, từ đó cải thiện độ chính xác và khả năng tổng quát của phân tích.
* Dữ liệu mẫu được sử dụng trong nghiên cứu có thể mang tính chủ quan, vì chúng chủ yếu dựa trên các bình luận từ người dùng cụ thể trên nền tảng Tiki. Điều này có thể dẫn đến sự thiên lệch trong phân tích, vì các bình luận này không nhất thiết phải phản ánh đầy đủ sự đa dạng của cảm xúc trong cộng đồng người dùng rộng lớn. Do đó, kết quả phân tích cảm xúc có thể không hoàn toàn chính xác và không đại diện cho toàn bộ cảm xúc của khách hàng trên sàn thương mại điện tử.
* Mô hình Pho-BERT có thể gặp khó khăn trong việc xử lý các trường hợp ngữ nghĩa phức tạp hoặc các tình huống cảm xúc cực đoan. Các tình huống như vậy thường yêu cầu sự hiểu biết sâu sắc về ngữ cảnh và các yếu tố văn hóa hoặc tâm lý, điều mà mô hình hiện tại có thể không hoàn toàn nắm bắt được. Ví dụ, cảm xúc trong các tình huống cực đoan như sự tức giận mạnh mẽ hoặc sự hạnh phúc tột độ có thể đòi hỏi khả năng nhận diện các tín hiệu ngữ nghĩa và ngữ cảnh đặc thù mà mô hình chưa được huấn luyện để nhận diện chính xác.

Khi mô hình không được tối ưu hóa cho những yếu tố này, kết quả phân tích cảm xúc có thể không phản ánh đúng mức độ và sắc thái cảm xúc thực sự của người dùng. Điều này làm giảm khả năng của hệ thống trong việc cung cấp các phân tích chi tiết và chính xác về phản hồi của khách hàng, đặc biệt trong các tình huống phức tạp hoặc mới lạ.

**6.3. Hướng phát triển**

Trong tương lai, đề tài ứng dụng mô hình Pho-BERT trong phân tích cảm xúc khách hàng có thể mở rộng theo nhiều hướng phát triển.

* Cải thiện hiệu suất mô hình: tinh chỉnh thuật toán hoặc kết hợp với các mô hình học sâu khác để nhận diện các cảm xúc phức tạp hơn
* Tích hợp phân tích trên nhiều danh mục sản phẩm: từ mô hình đã được huấn luyện, có thể sử dụng nó để áp dụng phần tích cho nhiều mặt hàng khác trên các nền tảng thương mại điện tử.
* Kết hợp phân tích dữ liệu văn bản đa ngôn ngữ: ngoài ngôn ngữ Tiếng Việt, nhóm nghiên cứu có thể tìm hiểu và ứng dụng các công nghệ phân tích cảm xúc dưới ngôn ngữ khác bằng những kiến thức đã thu nạp được từ nghiên cứu này.

Những hướng phát triển này không chỉ làm gia tăng giá trị của nghiên cứu hiện tại mà còn mở ra cơ hội để áp dụng công nghệ phân tích cảm xúc vào nhiều lĩnh vực và ứng dụng khác nhau.