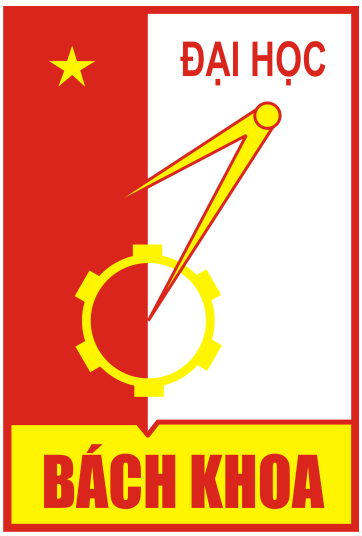
ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN - TIN



PHÂN TÍCH CẢM XÚC NGƯỜI DÙNG

ĐỒ ÁN I

Chuyên ngành: Hệ thống thông tin quản lý

Giảng viên hướng dẫn: **Ths. Lê Quang Hòa**
Sinh viên thực hiện: **Lê Đức Việt**
Mã sinh viên: **20216968**

HÀ NỘI – 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

2. Kết quả đạt được

3. Ý thức làm việc của sinh viên:

Hà Nội, ngày ... tháng ... năm 2021

Giảng viên hướng dẫn

ThS. Lê Quang Hòa

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn tới Ths. Lê Quang Hòa, giảng viên Bộ môn Toán tin, Khoa Toán Tin, Đại học Bách khoa Hà Nội đã tận tình hướng dẫn và chỉ bảo trong quá trình thực hiện đồ án.

Bên cạnh đó, em xin chân thành cảm ơn các thầy cô Khoa Toán Tin đã luôn quan tâm và có những chỉ đạo sát sao trong suốt quá trình thực hiện đồ án.

Do còn nhiều thiếu xót về kiến thức, kỹ năng cũng như kinh nghiệm thực tế, bài báo cáo của em không thể tránh khỏi những sai sót. Em rất mong nhận được những ý kiến đóng góp, phê bình của quý thầy cô để bài báo cáo này ngày càng hoàn thiện hơn. Quá trình thực hiện Đồ án I lần này đã làm cho em nhận ra giá trị của sự tự giác và chủ động trong công việc. Em hiểu rằng chúng ta cần tự tìm hiểu, nghiên cứu và rèn luyện kỹ năng mình còn thiếu sót. Em rất hy vọng rằng trong các dự án nghiên cứu tiếp theo, em sẽ vẫn được hướng dẫn và đồng hành cùng thầy. Một lần nữa, em xin bày tỏ lòng biết ơn sâu sắc và cảm kích đến sự tận tâm và hỗ trợ đắc lực của thầy, đã giúp em vượt qua khó khăn và tiến bộ trong quá trình học tập.

Em xin chân thành cảm ơn!

Mục lục

Bảng ký hiệu và chữ viết tắt

BERT Bidirectional Encoder Representations from Transformers

NLP Natural Language Processing

TL Transfer Learning

CNN Convolutional Neural Network

DNN Deep Neural Network

RNN Convolutional Neural Network

CNN Recurrent Neural Network

SVM Support Vector Machine

VLSP Vietnamese Language and Speech Processing

MLM Masked Language Model

LLM Large Language Model

Danh sách bảng

Danh sách hình vẽ

Tóm tắt

Phân tích cảm xúc của người dùng là một thách thức phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên. Bắt đầu từ dữ liệu văn bản, ta cần xác định và trích xuất thông tin về cảm xúc, có thể là tích cực, tiêu cực hoặc trung tính. Đây là một trong những bài toán quan trọng và phức tạp trong NLP, phân tích văn bản nhằm định lượng các trạng thái và thông tin chủ quan từ văn bản.

Transfer learning (TL) là một phương pháp quan trọng trong học máy, tập trung vào việc áp dụng kiến thức từ việc giải quyết một vấn đề vào việc giải quyết một vấn đề khác có liên quan. Trong lĩnh vực NLP, các mô hình dựa trên TL đã chứng minh hiệu quả bằng cách sử dụng kiến thức được học trước về ngôn ngữ. Nhờ đó, các mô hình này giúp tăng cường sự hiểu biết về cấu trúc của từ và câu, hỗ trợ việc nắm bắt ngữ nghĩa và kết nối giữa các yếu tố dễ dàng hơn

Đồ án này trình bày về lý thuyết và các phương pháp để giải quyết bài toán phân tích cảm xúc. Đồng thời là các kết quả từ quá trình nghiên cứu tập trung vào việc áp dụng mô hình BERT vào bài toán Phân tích cảm xúc của người dùng trên đa dạng bộ dữ liệu với các văn cảnh khác nhau trong việc sử dụng dịch vụ tại khách sạn. Trong tiến trình thực nghiệm đã kết hợp tiến hành đánh giá, so sánh với những kết quả tốt nhất ứng với mỗi dữ liệu.

GIỚI THIỆU

Phân tích cảm xúc đóng vai trò quan trọng trong việc thu thập ý kiến khách hàng từ nhiều nguồn như đánh giá sản phẩm, phản hồi dịch vụ, mạng xã hội, v.v. Nhờ vậy, doanh nghiệp có thể hiểu rõ hơn về thị hiếu khách hàng, cải thiện dịch vụ và chiến lược marketing.. Một nhiệm vụ cơ bản trong phân tích cảm xúc là đánh giá tính phân cực của một văn bản nhất định ở cấp độ tài liệu, câu hoặc tính năng/khía cạnh — xác định xem ý kiến được thể hiện trong một văn bản, một câu hoặc một tính năng/khía cạnh của một thực thể là tích cực, tiêu cực hay trung lập. Phân tích cảm xúc dựa trên khía cạnh nhằm xác định ý kiến hoặc tình cảm được thể hiện trên các tính năng hoặc khía cạnh khác nhau của các thực thể, như điện thoại di động, dịch vụ nhà hàng, hoặc chất lượng hình ảnh của máy ảnh. Mỗi tính năng hoặc khía cạnh là một thuộc tính hoặc thành phần của một thực thể, ví dụ: màn hình của điện thoại di động, dịch vụ nhà hàng, hoặc chất lượng hình ảnh của máy ảnh.

Kết quả của đồ án này sẽ cung cấp thông tin hữu ích cho các nhà nghiên cứu và thực hành về hiệu quả của BERT trong phân tích cảm xúc, góp phần thúc đẩy ứng dụng NLP trong các lĩnh vực khác nhau như marketing, chăm sóc khách hàng, v.v.

Đặt vấn đề

Sự bùng nổ của Internet và xu hướng mua sắm trực tuyến đã tạo nên nhu cầu to lớn cho việc thu thập và phân tích đánh giá của người dùng. Các trang web thương mại điện tử, mạng xã hội và diễn đàn trực tuyến tràn ngập ý kiến phản hồi về sản phẩm, dịch vụ từ người tiêu dùng.

Vai trò then chốt của đánh giá trực tuyến:

- Đối với khách hàng: Đánh giá giúp khách hàng đưa ra quyết định mua

sắm sáng suốt hơn. Họ có thể tham khảo ý kiến từ những người dùng khác, so sánh sản phẩm và dịch vụ, từ đó lựa chọn sản phẩm phù hợp nhất với nhu cầu của mình.

- Đối với doanh nghiệp: Phản hồi của khách hàng là nguồn thông tin quý giá để cải thiện sản phẩm, dịch vụ và chiến lược kinh doanh. Doanh nghiệp có thể lắng nghe ý kiến khách hàng, khắc phục những điểm yếu và phát huy những điểm mạnh để nâng cao chất lượng sản phẩm và dịch vụ, từ đó thu hút thêm khách hàng và tăng doanh thu.

Tuy nhiên, việc quản lý và phân tích đánh giá trực tuyến gặp nhiều thách thức:

- Lượng lớn dữ liệu: Doanh nghiệp phải đối mặt với vô số đánh giá từ nhiều nguồn khác nhau, khiến việc phân tích thủ công trở nên tốn kém và mất thời gian.

- Tính phức tạp của ngôn ngữ: Phản hồi của người dùng thường mang tính chủ quan, đa dạng về cách diễn đạt và sử dụng ngôn ngữ, gây khó khăn cho việc phân tích tự động.

Vì vậy, một số hệ thống đã được phát triển để phân tích các bình luận của người dùng. Trước khi khách hàng đặt hàng hoặc chọn nhà hàng cho các sự kiện, họ thường chú ý đến phản hồi của những khách hàng trước đó để đưa ra quyết định chính xác. Ngoài ra, với lĩnh vực nhà hàng, người tiêu dùng cũng quan tâm đến từng khía cạnh cụ thể của vấn đề để đưa ra quyết định, chẳng hạn như chất lượng thức ăn, dịch vụ, không gian, giá cả,... Bằng cách phân tích chi tiết các khía cạnh, chúng ta có thể tận dụng được nhiều thông tin từ đánh giá của người dùng. Nhận thấy tầm quan trọng của việc này, tôi cần một hệ thống có thể phân tích ý kiến theo khía cạnh của bình luận người dùng trong lĩnh vực nhà hàng.

Hiện nay, bài toán phân tích cảm xúc theo người dùng được quan tâm trong nhiều lĩnh vực khác nhau, từ giáo dục đến khảo sát ý kiến xã hội và đặc biệt là trong lĩnh vực dịch vụ và kinh doanh. Tuy nhiên, đối với tiếng Việt, chưa có nhiều nguồn ngữ liệu được xây dựng để phục vụ cho cộng đồng nghiên cứu khoa học.

Lý do chọn đề tài

Do nhu cầu phát triển của xã hội ngày càng tăng, nhất là về lĩnh vực kinh tế cũng như công nghệ. Việc phân tích cảm xúc trong văn bản được ứng dụng trong hàng loạt các vấn đề như: quản trị thương hiệu doanh nghiệp, thương hiệu sản phẩm, quản trị quan hệ khách hàng, khảo sát ý kiến xã hội học hay dễ hiểu hơn là phân tích đánh giá của khách hàng về một sản phẩm nào đó, ... Việc dự đoán là cực kì quan trọng vì ý kiến của người dùng ngày càng trở nên có giá trị hơn. Thị hiếu, sự quan tâm của cộng đồng là yếu tố ảnh hưởng chính đến các sản phẩm như phim, sách, thiết bị điện tử, ... Do đó, vấn đề này được sự quan tâm không chỉ từ các nhà nghiên cứu mà còn từ phía các công ty. Họ cần một hệ thống phân tích ý kiến khách hàng về sản phẩm một cách tự động để nhanh chóng nắm bắt được cảm nhận và thị hiếu của người tiêu dùng để nâng cao khả năng cạnh tranh với đối thủ cạnh và thích nghi với môi trường kinh doanh thường xuyên có biển động. Những thông tin này không chỉ hữu dụng trong tiếp thị, xếp hạng đánh giá sản phẩm mà còn hỗ trợ trong việc nhận biết vấn đề để xây dựng và phát triển sản phẩm.

Còn trong nghiên cứu, việc xây dựng hệ thống phân tích cảm xúc người dùng là một bước tiến lớn xong công động xử lý ngôn ngữ tự nhiên, giúp giải quyết được nhiều vấn đề đang mắc phải. Xây dựng mô hình giải quyết bài toán phân tích cảm xúc người dùng.

Phát biểu bài toán

Trong nghiên cứu này, mục tiêu chính là tìm hiểu và nghiên cứu về việc phân tích bình luận/đánh giá từ người dùng về trải nghiệm tại khách sạn, được thu thập từ Tripadvisor. Bài toán chính là xác định trạng thái cảm xúc từ các câu bình luận, với các trạng thái cảm xúc được quan tâm là Positive, Neutral, Negative

Đầu vào của bài toán bao gồm 20 nghìn đánh giá của khách hàng về trải nghiệm tại khách sạn được thu thập từ Tripadvisor.

Trong đồ án này, tôi đã sử dụng model BERT trong bài toán Phân tích cảm xúc.

Cấu trúc đồ án

PHẦN 1: CƠ SỞ LÝ THUYẾT

Giới thiệu về các khái niệm liên quan, bài toán xử lí ngôn ngữ tự nhiên, bài toán phân tích cảm xúc người dùng, tình hình nghiên cứu cả trong và ngoài nước.

PHẦN 2: PHƯƠNG PHÁP TIẾP CẬN

Trình bày về cơ sở lý thuyết mô hình và các phương pháp sẽ được sử dụng để thực nghiệm, các bước tiếp cận đến bài toán phân tích cảm xúc người dùng.

PHẦN 3: KẾT QUẢ THỰC NGHIỆM

Thử nghiệm và trình bày kết quả, thống kê kết quả đạt được. Nhận xét và đánh giá mô hình.

Chương 1 CƠ SỞ LÝ THUYẾT

1.1 Xử lí ngôn ngữ tự nhiên

1.1.1 Khái niệm

Xử lý ngôn ngữ [11] là một lĩnh vực quan trọng trong xử lý thông tin, nơi dữ liệu đầu vào thường là dữ liệu ngôn ngữ, bao gồm văn bản và âm thanh. Loại dữ liệu này đang trở thành loại dữ liệu chính của con người và được lưu trữ dưới dạng điện tử. Đặc điểm chung của chúng là không có cấu trúc hoặc chỉ có cấu trúc bán chặt chẽ, không thể được lưu trữ dưới dạng bảng dữ liệu. Do đó, chúng ta cần phải xử lý chúng để chuyển từ dạng không có cấu trúc thành dạng có thể hiểu được.

Xử lý ngôn ngữ tự nhiên (Natural Language Processing = NLP) [11] là một lĩnh vực của trí tuệ nhân tạo và ngôn ngữ học tính toán, mục tiêu là tập trung vào việc xử lý tương tác giữa con người và máy tính để máy tính có thể hiểu hoặc mô phỏng được ngôn ngữ của con người.

Xử lý ngôn ngữ tự nhiên hướng dẫn máy tính thực hiện và hỗ trợ con người trong các công việc liên quan đến ngôn ngữ, như dịch thuật, phân tích dữ liệu văn bản, nhận diện tiếng nói, tìm kiếm thông tin và tóm tắt văn bản.

Xử lý ngôn ngữ tự nhiên có thể được chia ra thành hai nhánh lớn, không hoàn toàn độc lập, bao gồm xử lý tiếng nói (speech processing) và xử lý văn bản (text processing). Xử lý tiếng nói tập trung nghiên cứu, phát triển các thuật toán, chương trình máy tính xử lý ngôn ngữ của con người ở dạng tiếng nói (dữ liệu âm thanh). Các ứng dụng quan trọng của xử lý tiếng nói bao gồm nhận dạng tiếng nói và tổng hợp tiếng nói. Nếu như nhận dạng tiếng nói là chuyển ngôn ngữ từ dạng tiếng nói sang dạng văn bản thì ngược lại, tổng hợp tiếng nói chuyển ngôn ngữ từ dạng văn bản

thành tiếng nói. Xử lý văn bản tập trung vào phân tích dữ liệu văn bản. Các ứng dụng quan trọng của xử lý văn bản bao gồm tìm kiếm và truy xuất thông tin, dịch máy, tóm tắt văn bản tự động, hay kiểm lỗi chính tả tự động. Xử lý văn bản đôi khi được chia tiếp thành hai nhánh nhỏ hơn bao gồm hiểu văn bản và sinh văn bản. Nếu như hiểu liên quan tới các bài toán phân tích văn bản thì sinh liên quan tới nhiệm vụ tạo ra văn bản mới như trong các ứng dụng về dịch máy hoặc tóm tắt văn bản tự động.

1.1.2 Các bước xử lý

- Phân tích hình thái [4] : Trong bước này, mỗi từ sẽ được phân tích và các ký tự không phải là chữ (như dấu câu) sẽ được tách ra khỏi các từ. Trong tiếng Anh và nhiều ngôn ngữ khác, các từ được phân tách bằng dấu cách. Tuy nhiên, trong tiếng Việt, dấu cách được sử dụng để phân tách các tiếng (âm tiết) chứ không phải là từ. Tương tự với các ngôn ngữ như tiếng Trung, tiếng Hàn, tiếng Nhật, phân tách từ trong tiếng Việt là một công việc không hề đơn giản.

- Phân tích cú pháp [4] : Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ với nhau. Có thể có những dãy từ bị loại bỏ vì không tuân theo các quy tắc ngữ pháp.

- Phân tích ngữ nghĩa [4] : Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bước phân tích cú pháp.

- Tích hợp văn bản [4] : Ngữ nghĩa của một câu có thể phụ thuộc vào các câu trước đó, đồng thời cũng có thể ảnh hưởng đến các câu sau đó.

- Phân tích thực nghĩa [4] : Cấu trúc thể hiện ý nghĩa của ngôn từ sẽ được diễn dịch lại để xác định ý nghĩa thực sự của nó.

Tuy nhiên, ranh giới giữa 5 bước xử lý này cũng rất mong manh. Chúng có thể được thực hiện từng bước một hoặc cùng một lúc - tùy thuộc vào thuật toán và ngữ cảnh cụ thể.

1.1.3 Thuật ngữ

Tokenization [4]: Là quá trình tách một cụm từ, câu, đoạn văn, một hoặc nhiều tài liệu văn bản thành các đơn vị nhỏ hơn. Mỗi đơn vị nhỏ hơn này được gọi là Tokens.

Token [4]: Là các khối xây dựng của NLP và tất cả các mô hình NLP đều xử lý văn bản thô ở cấp độ các Tokens. Chúng được sử dụng để tạo từ vựng trong một kho ngữ liệu (một tập dữ liệu trong NLP). Từ vựng này sau đó được chuyển thành số (ID) và giúp chúng ta lập mô hình. Tokens có thể là bất cứ thứ gì – một từ (word), một từ phụ (sub-word) hoặc thậm chí là một ký tự (character).

Nhập nhằng - Ambiguity [12] (Ở nhiều cấp độ: lexical - từ vựng, morphological - hình vị, syntactic - cú pháp, semantic - ngữ nghĩa, domain - lĩnh vực).

Khử nhập nhằng thế đại từ [12] - Anaphora (Sự sử dụng đại từ để thay thế cho một danh từ hoặc một nhóm từ khác). Ví dụ: "Con khỉ ăn chuối vì nó đói." Đại từ "nó" thay thế cho "khỉ" hoặc "chuối".

Túi từ - Bag of Words [12]: Một mô hình thường được sử dụng trong phân loại văn bản (Text Classification), biểu diễn thông tin dưới dạng tập hợp các từ và tần suất xuất hiện của mỗi từ trong văn bản. Bag of Words thường được sử dụng như một đặc trưng để huấn luyện các mô hình phân loại.

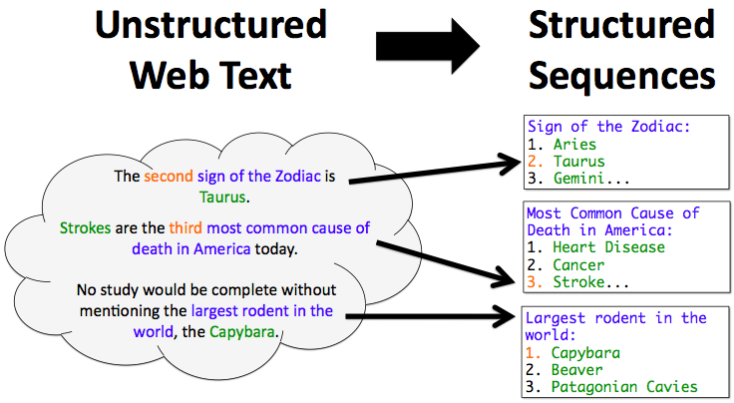
Từ loại [4] - Part-of-speech (POS) Tag: Một từ có thể được phân loại thành một hoặc nhiều lớp từ vựng hoặc một phần của lời nói như Danh từ, Động từ, Tính từ và Bài viết, để đặt tên cho một số. Thẻ POS là một biểu tượng đại diện cho một phạm trù từ vựng như vậy - NN (Danh từ), VB (Động từ), JJ (Tính từ), AT (Bài viết). Một trong những bộ thẻ lâu đời nhất và được sử dụng phổ biến nhất là bộ thẻ Brown Corpus.

Ngữ liệu [4] - Corpus/Corpora (tập dữ liệu ngôn ngữ): Là tập hợp các dữ liệu ngôn ngữ thực tế, được sử dụng để xây dựng và kiểm tra các mô hình và giải thuật trong xử lý ngôn ngữ tự nhiên. Có nhiều loại Corpora như ngữ liệu song ngữ, ngữ liệu song song...

Phân tích ngữ nghĩa - Explicit Semantic Analysis (ESA) [12] là phương pháp giúp máy tính hiểu được ý nghĩa của văn bản. ESA sử dụng các khái niệm từ WordNet để biểu diễn văn bản dưới dạng một tập hợp các vectơ ngữ nghĩa.

Trích xuất Thông tin - Information Extraction [12] - là tiến trình rút trích ra các thông tin có cấu trúc một cách tự động từ các nguồn dữ liệu

không cấu trúc hay bán cấu trúc như các tài liệu văn bản, các trang web...

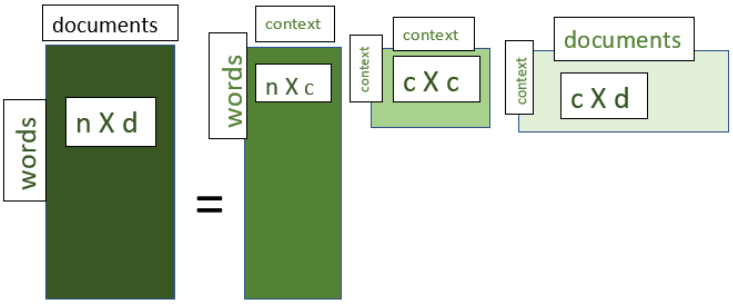


Hình 1.1: Trích xuất thông tin [4]

Phân tích cấu trúc - Lexical Analysis [4]: Các tập hợp gồm các từ và cụm từ trong một ngôn ngữ được gọi là một bộ từ vựng của một ngôn ngữ.

Phân bổ Dirichlet ẩn - Latent Dirichlet Allocation (LDA) [12]: Phương pháp thuộc lớp các mô hình Topic Modeling, LDA được xây dựng dựa trên giả định rằng mỗi chủ đề (topic) là một phân phối xác suất của các từ, mỗi văn bản là sự kết hợp của nhiều chủ đề và mỗi từ được phân bố vào một trong các chủ đề này.

Phân tích ngữ nghĩa ẩn - Latent Semantic Analysis (LSA) [12]: Một phương pháp phân tích ngữ nghĩa được áp dụng để khám phá mối quan hệ giữa các từ và văn bản trong các tập dữ liệu lớn . LSA giả định rằng các từ có ý nghĩa tương đồng sẽ thường xuất hiện trong các văn bản có nội dung tương tự.



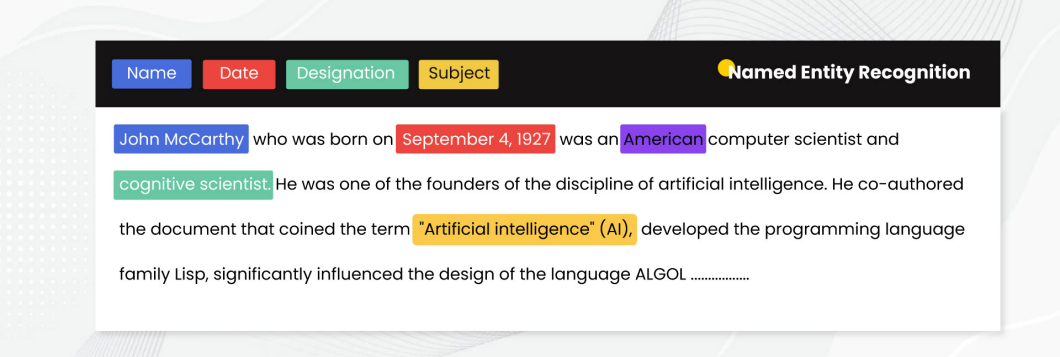
Hình 1.2: Phân tích ngữ nghĩa ẩn [14]

Phân tích hình thái - Morphological analysis [4]: Ngôn ngữ tự nhiên bao gồm một số lượng rất lớn các từ được xây dựng dựa trên các khối xây

dựng cơ bản được gọi là hình thái (hoặc thân), các đơn vị ngôn ngữ nhỏ nhất có ý nghĩa. Phân tích hình thái quan tâm đến việc khám phá và phân tích cấu trúc bên trong của các từ bằng máy tính.

Phân tích cú pháp - Parsing [4]: Trong nhiệm vụ phân tích cú pháp, một trình phân tích cú pháp xây dựng cây phân tích cú pháp đã cho một câu. Có một số trình phân tích cú pháp giả định sự tồn tại của một tập hợp các quy tắc ngữ pháp để phân tích cú pháp nhưng 28 trình phân tích cú pháp gần đây đủ thông minh để suy ra các cây phân tích cú pháp trực tiếp từ dữ liệu đã cho bằng cách sử dụng các mô hình thống kê phức tạp. Hầu hết các trình phân tích cú pháp cũng hoạt động trong một cài đặt được giám sát và yêu cầu câu phải được gắn thẻ POS trước khi nó có thể được phân tích cú pháp. Phân tích cú pháp thống kê là một lĩnh vực nghiên cứu tích cực trong NLP.

Nhận diện thực thể có tên - Named Entity Recognition (NER) [12] - là tiến trình xác định và phân loại các phần tử trong văn bản vào các danh mục được định nghĩa trước như tên người, tổ chức, địa điểm ...



Hình 1.3: Nhận diện thực thể có tên [12]

Phân tích ngữ dụng - Pragmaties [12]: từ "sentence" trong phân tích văn phạm có nghĩa là câu, trong luật pháp có nghĩa là án tù. Do vậy, ta cần xem xét toản bộ văn bản để đưa ra ý nghĩa chính xác.

Tác vụ phía sau - Downstream task : Là những tác vụ supervised-learning được cải thiện dựa trên những pretrained model. VD: Chúng ta sử dụng lại các biểu diễn từ học được từ những pretrained model trên bộ văn bản lớn vào một tác vụ phân tích cảm xúc huấn luyện trên bộ văn bản có kích thước nhỏ hơn. Áp dụng pretrain-embedding đã giúp cải thiện mô hình. Như vậy tác vụ sử dụng pretrain-embedding được gọi là downstream task.

Điểm khái quát đánh giá mức độ hiểu ngôn ngữ - GLUE score benchmark : GLUE score benchmark là một tập hợp các chỉ số được xây dựng để đánh giá khái quát mức độ hiểu ngôn ngữ của các model NLP. Các đánh giá được thực hiện trên các bộ dữ liệu tiêu chuẩn được qui định tại các convention về phát triển và thúc đẩy NLP.

Quan hệ văn bản - Textual Entailment : Là tác vụ đánh giá mối quan hệ định hướng giữa hai văn bản. Nhãn output của các cặp câu được chia thành đối lập (contradiction), trung lập (neutral) hay có quan hệ đi kèm (textual entailment).

Ngữ cảnh - Contextual: Là ngữ cảnh của từ. Một từ được định nghĩa bởi một cách phát âm nhưng khi được đặt trong những câu khác nhau thì có thể mang ngữ nghĩa khác nhau. Ngữ cảnh có thể coi là môi trường xung quanh từ để góp phần định nghĩa từ

Tiền xứ lý dữ liệu - Pre-processing [12], xử lý sơ bộ văn bản: xóa bỏ những kí tự, những mã điều khiển, những vùng không cần thiết cho hệ thẳng gồm: tách đoạn/câu tử (paragraph/sentence/word segmentation), làm sạch (cleaning). tích hợp (integreation), chuyên đổi (transformation).

1.2 Ngữ cảnh và vai trò trong NLP

Ngôn ngữ được hình thành từ âm thanh để diễn đạt suy nghĩ của con người. Trong giao tiếp, các từ thường không đứng riêng lẻ mà kết hợp với các từ khác để tạo thành câu hoàn chỉnh. Sự kết hợp này giúp truyền đạt nội dung và ý nghĩa hiệu quả hơn so với từng từ riêng lẻ.

Ngữ cảnh trong câu đóng vai trò quan trọng trong việc xác định ý nghĩa của từ. Hiểu được điều này, các thuật toán xử lý ngôn ngữ tự nhiên (NLP) tiên tiến (SOTA) đã nỗ lực đưa ngữ cảnh vào mô hình, và BERT là một ví dụ điển hình.

Các phương pháp embedding từ trong NLP có thể được phân cấp như sau:

Non-context (không bối cảnh): Đây là các thuật toán không xem xét ngữ cảnh khi biểu diễn từ, như word2vec, GloVe, và fastText. Mỗi từ chỉ có một biểu diễn vector duy nhất, không thay đổi theo ngữ cảnh.

Uni-directional (Một chiều): Các thuật toán này bắt đầu xem xét ngữ cảnh của từ, nhưng chỉ theo một chiều, từ trái qua phải hoặc từ phải qua trái. Các phương pháp nhúng từ dựa trên RNN là ví dụ điển hình.

Bi-directional (Hai chiều): Ngữ nghĩa của một từ được xác định bởi toàn bộ ngữ cảnh xung quanh, cả trước và sau từ đó. Các mô hình sử dụng kỹ thuật transformer như BERT, ULMFit, OpenAI GPT là đại diện cho phương pháp này. Các mô hình bidirectional này đã đạt được những kết quả tiên tiến (SOTA) trên hầu hết các tác vụ trong GLUE benchmark.

Việc sử dụng mô hình bidirectional như BERT giúp biểu diễn ngữ nghĩa của từ chính xác hơn nhờ xem xét toàn bộ ngữ cảnh, từ đó cải thiện hiệu quả của các tác vụ NLP, đặc biệt trong bài toán phân tích cảm xúc.

1.3 Tổng quan về phân tích cảm xúc

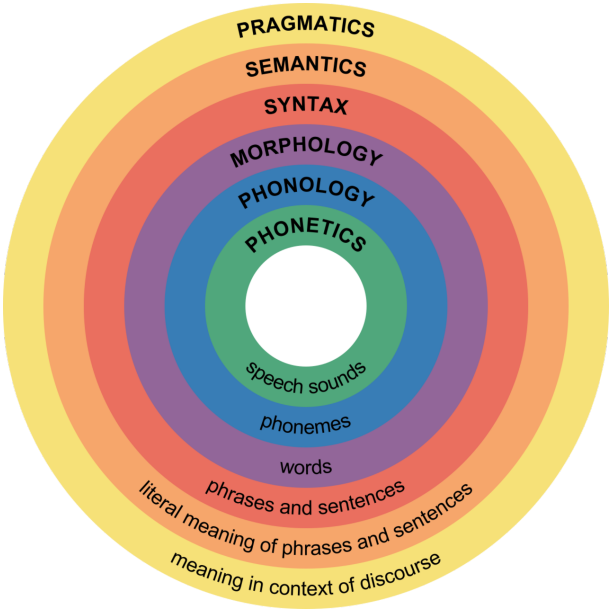
1.3.1 Phân tích cảm xúc là gì?

Trong những năm gần đây, Phân tích cảm xúc (SA) còn được gọi là khai thác ý kiến đã thu hút sự quan tâm đặc biệt từ cộng đồng nghiên cứu trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), cả trong và ngoài cộng đồng nghiên cứu ở nhiều quốc gia. Đây là quá trình xác định và phân loại các cảm xúc khác nhau - ví dụ như tích cực, tiêu cực hoặc trung tính - hoặc các trạng thái cảm xúc như vui, buồn, tức giận - để hiểu thái độ của con người đối với một chủ thể cụ thể hoặc một vấn đề nhất định.

Phân tích cảm xúc cũng là một phần quan trọng trong lĩnh vực NLP. Nó không chỉ có ý nghĩa lớn trong học thuật và nghiên cứu, mà còn đóng

vai trò cực kỳ quan trọng trong các ngành công nghiệp - đặc biệt là trong lĩnh vực dịch vụ, nơi việc nhận biết hành vi và thái độ của khách hàng đối với sản phẩm và dịch vụ mà họ sử dụng là vô cùng quan trọng. Tuy nhiên, mọi người thể hiện các cảm nhận của mình thông qua ngôn ngữ tự nhiên vốn thường có sự nhập nhằng về ngữ nghĩa đã gây không ít khó khăn trong việc xử lý thông tin. Bên cạnh đó, người dùng còn sử dụng các từ viết tắt, từ lóng hay các kí hiệu biểu cảm như '=))', ':(', '>,<', ... để thể hiện trạng thái cảm xúc của họ.

Hiện nay, bài toán phân tích cảm xúc thường được thực hiện ở ba cấp độ chính: cấp độ câu văn (sentence-level), cấp độ văn bản (document-level), và cấp độ khía cạnh (aspect-level). Ở cấp độ câu văn, mục tiêu của bài toán là phân loại một câu văn thành các lớp tiêu cực (negative), tích cực (positive), hoặc trung tính (neutral). Cấp độ văn bản được sử dụng để xác định mức độ cảm xúc của một đoạn văn (bao gồm hai hay nhiều câu văn) là tiêu cực, tích cực, hoặc trung tính. Và cấp độ khía cạnh được sử dụng để xác định mức độ cảm xúc cho mỗi khía cạnh của thực thể được đề cập trong một văn bản. Trong phạm vi của đồ án này, nghiên cứu sẽ tập trung chỉ vào cấp độ câu văn.



Hình 1.4: Biểu đồ cấp độ ngôn ngữ học [15]

Trong lĩnh vực NLP, bài toán phân tích cảm xúc thuộc về cả ngữ dụng học (Pragmatics) và ngữ nghĩa học (Semantics). Vị trí của bài toán này

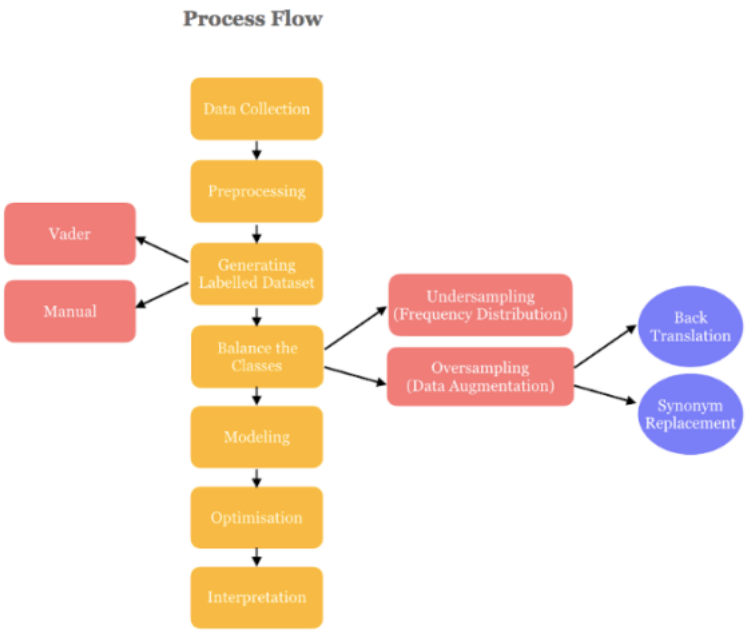
trong lĩnh vực NLP có thể được xem như một ứng dụng cụ thể trong lĩnh vực này, đóng vai trò quan trọng trong việc hiểu và xử lý ngôn ngữ tự nhiên.

1.3.2 Phân tích cảm xúc hoạt động thế nào?

Phân tích cảm xúc là một ứng dụng thuộc công nghệ xử lý ngôn ngữ tự nhiên (NLP) có khả năng đào tạo để giúp phần mềm máy tính hiểu văn bản theo các cách tương tự như con người. Quá trình phân tích thường trải qua môt số giai đoạn trước khi đưa ra kết quả cuối cùng.

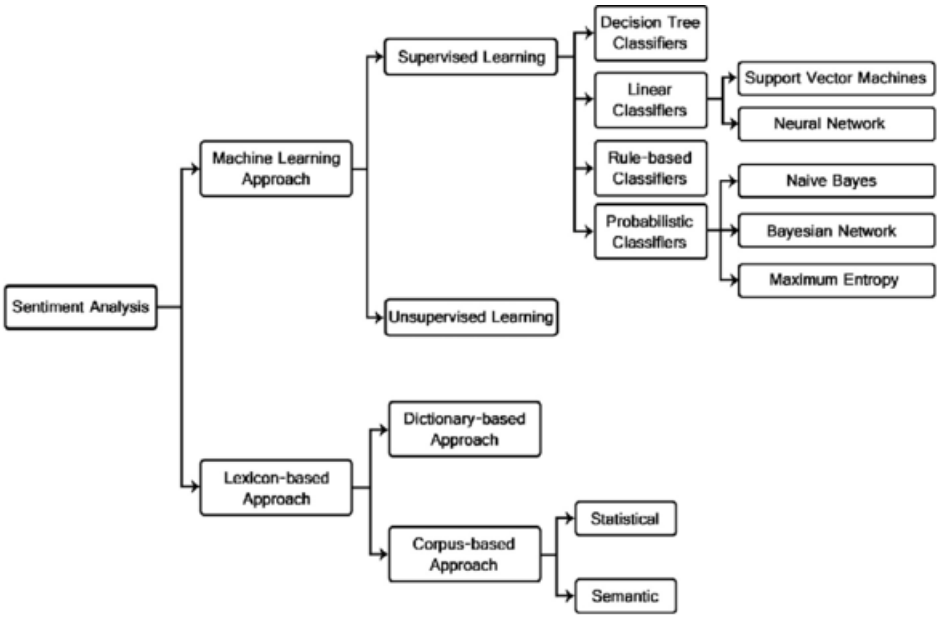
Trong giai đoạn tiền xử lý, phân tích cảm xúc xác định các từ khóa để nêu bật thông điệp chủ đạo trong văn bản. Thông thường, việc gán nhãn từ điển hoặc corpora được thực hiện thủ công và các mô hình phân loại sau đó được huấn luyện với các tập dữ liệu lớn để phân loại các từ hoặc cụm từ mới. Có những cách tiếp cận khác để phân tích tình cảm tập trung vào việc khai thác các câu hoặc toàn bộ tài liệu, thay vì phụ thuộc vào tính chẵn lẻ của các từ. Cách tiếp cận này thường hoạt động với corpora của các tài liệu văn bản. Vấn đề thiết yếu với phân loại tài liệu (phân loại phân cực) là nó phải xác định các đặc điểm tình cảm tổng thể của toàn bộ tài liệu, trong khi tình cảm được thể hiện có thể được bao gồm chỉ trong một câu hoặc từ.

Phân tích từ khóa: Các công nghệ NLP phân tích sâu hơn các từ khóa được trích xuất và cho chúng một số điểm cảm xúc. Điểm cảm xúc là một thang đo cho biết yếu tố cảm xúc trong hệ thống phân tích cảm xúc. Thang đo này cung cấp nhận thức tương đối về cảm xúc được thể hiện trong văn bản nhằm phục vụ mục đích phân tích. Ví dụ: khi phân tích đánh giá của khách hàng, các nhà nghiên cứu sử dụng số 10 để đại diện cho sự hài lòng và số 0 để đại diện cho sự thất vọng.



Hình 1.5: Bài toán phân tích cảm xúc

Các kỹ thuật Machine Learning được áp dụng cho các vấn đề phân tích cảm xúc chia thành hai loại: (1) các mô hình tiêu chuẩn và (2) các mô hình học sâu . Các mô hình truyền thống liên quan đến các thuật toán học máy truyền thống , như bộ phân loại Naive Bayes, bộ phân loại entropy tối đa, và máy vector hỗ trợ (SVM). Những thuật toán này nhận các đặc điểm từ ngữ, các đặc điểm dựa trên từ điển cảm xúc, các phần từ loại, cũng như các tính từ và trạng từ làm đầu vào. Độ chính xác của các hệ thống này phụ thuộc vào các đặc điểm được chọn. Các mô hình học sâu có thể cung cấp hiệu suất cao hơn so với các phương pháp truyền thống. CNN, DNN và RNN là những mô hình học sâu có thể được sử dụng cho phân tích cảm xúc. Những phương pháp này xử lý các vấn đề phân loại ở mức tài liệu, cụm từ và mức khía cạnh. Phần tiếp theo sẽ đề cập đến các phương pháp này của học sâu.



Hình 1.6: Phân loại kỹ thuật phân tích cảm xúc

1.4 Phân loại bài toán phân tích cảm xúc

1.4.1 Phân tích cảm xúc

Phân tích cảm xúc - Emotion Analysis nhằm mục đích xác định trạng thái cảm xúc chung của một văn bản. Điều này bao gồm việc đánh giá tính phân cực của văn bản (tích cực, tiêu cực, trung lập), nhưng cũng mở rộng để phát hiện các cảm xúc cụ thể (như giận dữ, vui vẻ, buồn bã, v.v.), mức độ khẩn cấp (khẩn cấp, không khẩn cấp), và thậm chí cả ý định (quan tâm hoặc không quan tâm). Loại phân tích cảm xúc mà trong đó các loại cảm xúc (vui vẻ, thất vọng, giận dữ, và buồn bã) được phát hiện và phân loại gọi là phát hiện cảm xúc.

Có một số khó khăn trong việc phân loại này. Người dùng có thể diễn đạt cảm xúc của họ bằng nhiều từ khác nhau. Họ có thể sử dụng một từ có nghĩa xấu để biểu lộ sự vui vẻ. Những ví dụ khó nhất cho các mô hình phân loại ở đây là; chẳng hạn, câu "Tôi kết nối với dịch vụ khách hàng quá muộn, điều này làm tôi bực bội" là một câu tiêu cực, trong khi câu "Bạn đang làm tôi phấn khích" lại là một câu tích cực.

1.4.2 Phân tích cảm xúc đa ngôn ngữ

Đây là một phiên bản khác của phân tích cảm xúc cung cấp hỗ trợ đa ngôn ngữ. Ý nghĩa ở đây là thực hiện phân tích cảm xúc trong nhiều ngôn ngữ khác nhau.

Đầu tiên, sử dụng bộ phân loại ngôn ngữ để phát hiện ngôn ngữ của văn bản và sau đó chạy mô hình phân tích cảm xúc phù hợp cho ngôn ngữ đó.

Thứ hai, phát triển một mô hình ngôn ngữ đa ngôn ngữ và tinh chỉnh mô hình này để nó có thể hoạt động trong nhiều ngôn ngữ.

1.4.3 Phân tích cảm xúc theo cấp độ

Nếu độ chính xác của tâm trạng là quan trọng, các danh mục có thể được phân chia chi tiết hơn. Có thể thực hiện một phân loại rộng hơn, không chỉ dừng lại ở tích cực và tiêu cực:

- Rất tích cực
- Tích cực
- Trung lập
- Tiêu cực
- Rất tiêu cực

Phân loại này thường được sử dụng trong các đánh giá và nhận xét nơi mà có thang điểm 5 sao.

Rất tích cực = 5 sao

Rất tiêu cực = 1 sao

1.4.4 Phân tích cảm xúc theo khía cạnh

Thường thì khi phân tích cảm xúc của các văn bản, ta tập trung vào việc xác định xem nhận xét/ý kiến đó là tích cực hay tiêu cực. Nhưng chúng ta không tập trung vào điều gì là tích cực hay tiêu cực trong văn bản đó.

Để nói rõ hơn, trong biểu đạt "Tôi không thích sản phẩm này chút nào, kích thước quá nhỏ", người dùng không hài lòng với sản phẩm và phàn nàn về kích thước của nó. Trong một phân tích cảm xúc thông thường, câu này được phân loại là tiêu cực, nhưng trong phân tích cảm xúc theo khía cạnh, phần "kích thước quá nhỏ" cũng được chú trọng.

1.4.5 Phân tích theo ý định

Phân tích ý định tập trung vào những gì người dùng muốn làm. Hiểu được người dùng muốn làm gì sẽ giúp chúng ta hướng dẫn họ tốt hơn.

Chẳng hạn, việc có thể hiểu rằng một khách hàng đang duyệt một trang web thương mại điện tử có ý định mua sắm cũng cho phép chúng ta đề xuất các sản phẩm phù hợp cho họ.

Một trong những lĩnh vực được sử dụng nhiều nhất là các hệ thống trợ lý thông minh trong các ứng dụng. Nó cho phép chúng ta hướng dẫn người dùng đến những nơi phù hợp trong ứng dụng theo yêu cầu của họ và chúng ta có thể cung cấp một trải nghiệm ứng dụng tốt hơn cho người dùng.

1.5 Tình hình nghiên cứu trong và ngoài nước

Từ những năm 2000 trở đi, phân tích ý kiến và phân tích cảm xúc đã trở thành một lĩnh vực thu hút sự quan tâm và phát triển của cả nhà nghiên cứu và các ứng dụng thực tiễn. Ý tưởng về phân tích cảm xúc (sentiment analysis) được giới thiệu lần đầu trong một nghiên cứu của Nasukawa và Yi, trong khi khái niệm phân tích ý kiến (opinion mining) được đề xuất đầu tiên trong một nghiên cứu của Dave, Lawrence và Pennock. Tuy nhiên, nghiên cứu của Pang và Lillian Lee [6] được coi là mốc quan trọng cho sự phát triển của phân tích cảm xúc, đã đạt được kết quả đáng ghi nhận và đóng góp cho sự phát triển của lĩnh vực này. Các chủ đề nghiên cứu trong phân tích cảm xúc rất đa dạng, bao gồm phân tích đánh giá phim, sản phẩm, nhà hàng, món ăn và nhiều hơn nữa. Để giải quyết bài toán phân tích cảm xúc, các nghiên cứu đã sử dụng các phương pháp như máy học, thống kê và phương pháp dựa trên luật kết hợp với dữ liệu ngôn ngữ. Nhờ vào những nỗ lực này, các nghiên cứu đã đạt được những tiến bộ đáng kể trong việc hiểu và phân tích cảm xúc trong ngôn ngữ.Kể từ đó, bài toán này đã trở nên ngày càng được quan tâm và tiếp tục phát triển.

Với ngôn ngữ tiếng Việt, các nghiên cứu về phân tích cảm xúc câu vẫn đang tiếp tục phát triển. Ví dụ, Kieu và Pham [7] đã giới thiệu một phương pháp phân loại cảm xúc dành cho tiếng Việt, dựa trên hệ thống luật và mô tả các thực nghiệm trên bộ dữ liệu đánh giá sản phẩm máy tính. Duyen

và đồng nghiệp [8] đã sử dụng các thuật toán máy học như SVMI, MEM để phân loại đánh giá khách sạn từ Agoda. Van và đồng nghiệp [9] đã sử dụng SVM để phân loại các bình luận trên Facebook. Tran và Phan [10] đã đưa ngôn ngữ bối cảnh vào câu để cung cấp thêm thông tin cho phân tích cảm xúc. Các nghiên cứu này đóng góp vào việc nghiên cứu và phát triển phân tích cảm xúc trong tiếng Việt. Ngoài ra, nhóm Nghiên Cứu VLSP cũng đã có những nghiên cứu quan trọng, bên cạnh đó còn có các nghiên cứu sinh Việt Nam tại JAIST. Tuy nhiên, nghiên cứu vẫn chưa đi đến mức độ hệ thống và chưa có định hướng rõ ràng, hạn chế chỉ dừng lại ở cấp độ tiến sĩ, thạc sĩ và các bài báo nghiên cứu khoa học có tính chất tìm hiểu. Và, hiện nay, các bộ dữ liệu về phân tích cảm xúc người dùng đã dần được tập trung và xây dựng nhằm phục vụ cộng đồng nghiên cứu. Điển hình, năm 2016, cộng đồng xử lý ngôn ngữ tự nhiên (Vietnamese Language and Speech Processing- VLSP) đã tiến hành tổ chức cuộc thi phân tích cảm xúc người dùng trên các phản hồi mua hàng bởi Huyen và các cộng sự . Bên cạnh đó, gần đây nhất, AIVIVN, một nền tảng giúp tổ chức các cuộc thi machine learning cho cộng đồng, đã giới thiệu bộ dữ liệu phân loại sắc thái bình luận trên các trang thương mại điện tử

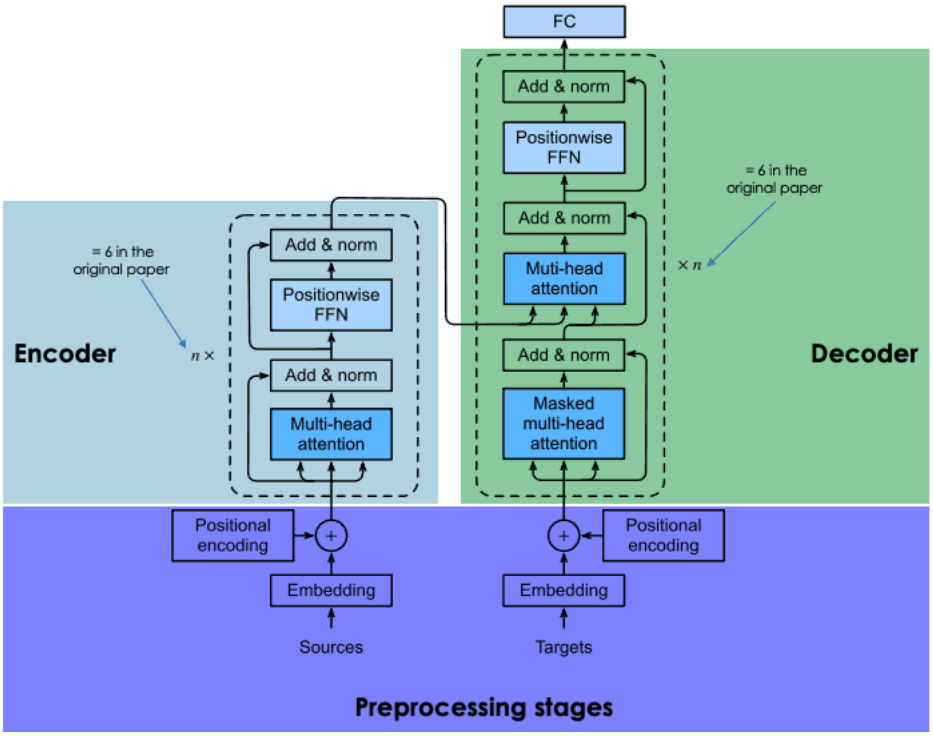
Chương 2 PHƯƠNG PHÁP TIẾP CẬN

2.1 Mô hình BERT

BERT [17] (Bidirectional Encoder Representations from Transformers) là mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer. BERT được thiết kế để huấn luyện trước các biểu diễn từ . Điểm đặc biệt ở BERT đó là nó có thể điều hòa cân bằng ngữ cảnh của câu văn theo cả 2 chiều trái và phải.

Chìa khóa thành công của BERT là cấu trúc Transformer của nó. Trước khi Transformer ra đời, việc lập mô hình ngôn ngữ tự nhiên là một nhiệm vụ rất khó khăn. Bất chấp sự phát triển của các mạng lưới thần kinh phức tạp – cụ thể là mạng lưới thần kinh tái phát hoặc tích chập – kết quả chỉ thành công một phần. Thách thức chính nằm ở cơ chế mạng lưới thần kinh được sử dụng để dự đoán từ còn thiếu trong câu. Vào thời điểm đó, các mạng thần kinh tiên tiến dựa trên kiến trúc bộ mã hóa-giải mã , một cơ chế mạnh mẽ nhưng tiêu tốn nhiều thời gian và tài nguyên, không phù hợp cho tính toán song song[19]. Lưu ý đến những thách thức này, các nhà nghiên cứu của Google đã phát triển máy biến áp Transformers, một kiến trúc thần kinh cải tiến dựa trên cơ chế chú ý.

BERT là model biểu diễn ngôn ngữ được google giới thiệu vào năm 2018. BERT là một pretrain-model, là mô hình xử lý ngôn ngữ tự nhiên (NLP) có thể sử dụng cơ chế tiền xử lý dữ liệu bằng cách chuyển từ một mô hình chung đã được huấn luyện trên một lượng lớn dữ liệu không có nhãn. Chúng ta sử dụng lại các biểu diễn từ học được từ những pretrained model trên bộ văn bản lớn vào một tác vụ phân tích cảm xúc huấn luyện trên bộ văn bản có kích thước nhỏ hơn. BERT là một mô hình pretrained xử lí ngôn ngữ tự nhiên dựa trên ngữ cảnh 2 chiều. Trong khi các kỹ thuật



Hình 2.1: Transformers [19]

như Word2vec, FastText hay Glove chỉ tìm biểu diễn của từ thông qua ngữ cảnh chung của chúng, tức là chỉ một vector cho mỗi từ, thì việc tạo một biểu diễn cho mỗi từ dựa trên các từ khác trong câu (nhiều vector biểu diễn một từ tùy thuộc vào câu) sẽ mang lại kết quả chính xác hơn. Khi học chuyển giao BERT ta sẽ tận dụng lại kiến trúc từ mô hình pretrained và bổ sung một số layers phía sau để phù hợp với nhiệm vụ huấn luyện, đồng thời các tham số của các layers gốc sẽ được fine-tunning lại. Cơ chế attention của Transformer sẽ truyền toàn bộ các từ trong câu văn đồng thời vào mô hình một lúc mà không cần quan tâm đến chiều của câu. Do đó Transformer được xem như là huấn luyện hai chiều (bidirectional) mặc dù trên thực tế chính xác hơn chúng ta có thể nói rằng đó là huấn luyện không chiều (non-directional).

Kiến trúc của mô hình BERT là một kiến trúc đa tầng gồm nhiều lớp Bidirectional Transformer encoder dựa trên bản mô tả đầu tiên của Vaswani et al. (2017) và sự phát hành trong thư viện tensor2tensor.

Hiện tại có nhiều phiên bản khác nhau của model BERT. Các phiên bản đều dựa trên việc thay đổi kiến trúc của Transformer tập trung ở 3 tham số: L: số lượng các block sub-layers trong transformer, H: kích thước

của embedding véc tơ (hay còn gọi là hidden size), A: Số lượng head trong multi-head layer, mỗi một head sẽ thực hiện một self-attention. Tên gọi của 2 kiến trúc bao gồm:

- BERT BASE : L=12, H=768, A=12, Tổng tham số = 110 triệu
- BERT LARGE : L=24, H=1024, A=16, Tổng tham số = 340 triệu

Có một chú thích nhỏ rằng, một Transformer 2 chiều thường được gọi là Transformer encoder trong khi các phiên bản Transformer chỉ sử dụng ngữ cảnh bên trái thường được gọi là Transformer decoder vì nó có thể được sử dụng để tạo ra văn bản.

2.1.1 Ý tưởng cỗi lõi của BERT:

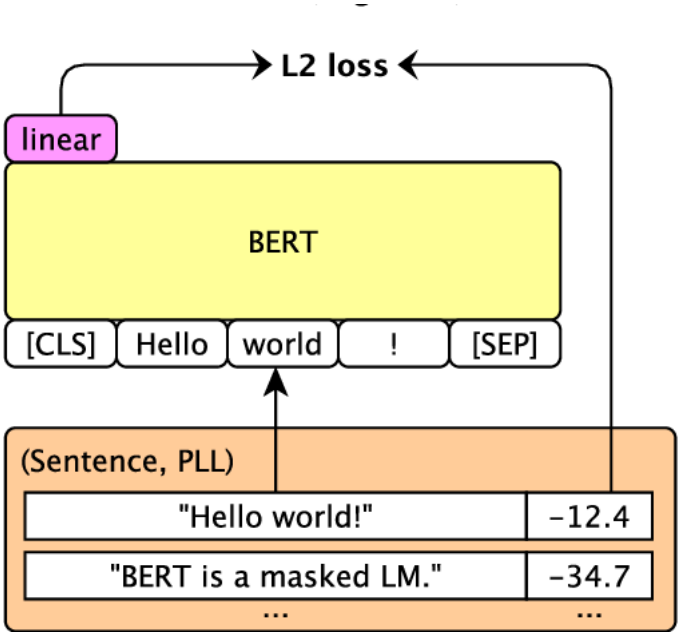
Mạng thần kinh tái phát và tích chập sử dụng tính toán tuần tự để tạo ra dự đoán. Nghĩa là, họ có thể dự đoán từ nào sẽ theo sau một chuỗi các từ nhất định sau khi được đào tạo trên bộ dữ liệu khổng lồ. Theo nghĩa đó, chúng được coi là các thuật toán đơn hướng hoặc không có ngữ cảnh.

Ngược lại, các mô hình sử dụng nguồn biến áp như BERT, cũng dựa trên kiến trúc bộ mã hóa-giải mã, là hai chiều vì chúng dự đoán các từ dựa trên các từ trước và các từ sau. Điều này đạt được thông qua cơ chế tự chú ý, một lớp được tích hợp trong cả bộ mã hóa và bộ giải mã. Mục tiêu của lớp chú ý là nắm bắt các mối quan hệ ngữ cảnh tồn tại giữa các từ khác nhau trong câu đầu vào.

- VD: Ta có 1 câu văn bản sau: "The man headed to the store and purchased a ... of shoes."

Một mô hình ngôn ngữ thông thường sẽ dự đoán từ ... là "box" hoặc "pair". Nhưng với BERT thì không như vậy. Thay vì dự đoán từ tiếp theo trong một chuỗi, BERT sử dụng một kỹ thuật mới có tên Masked LM (MLM): nó che giấu ngẫu nhiên các từ trong câu và sau đó cố gắng dự đoán chúng. Che giấu có nghĩa là mô hình nhìn theo cả hai hướng và sử dụng toàn bộ ngữ cảnh của câu, cả môi trường xung quanh bên trái và bên phải, để dự đoán từ bị che giấu. Không giống như các mô hình ngôn ngữ trước đó, BERT tính đến cả mã thông báo trước và mã thông báo tiếp theo cùng một lúc. Các mô hình dựa trên LSTM kết hợp từ trái sang phải và từ phải sang trái hiện có đã thiếu "đồng thời" này. (Mặc dù có thể chính

xác hơn khi nói rằng BERT là không định hướng)



Hình 2.2: Masked LM (MLM) [20]

Một điểm đặc biệt ở BERT mà các model embedding trước đây chưa từng có đó là kết quả huấn luyện có thể fine-tuning được. Chúng ta sẽ thêm vào kiến trúc model một output layer để tùy biến theo tác vụ huấn luyện.

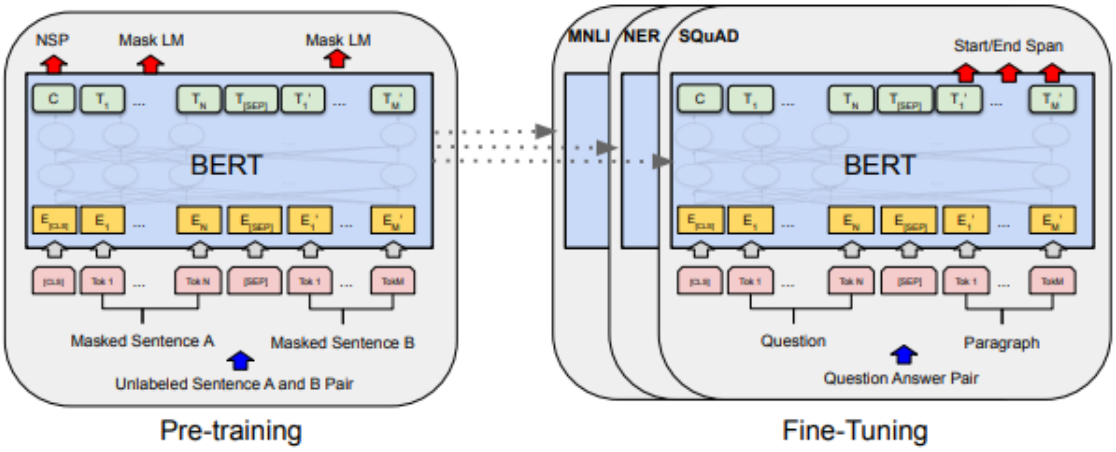
Tiến trình áp dụng fine-tuning sẽ như sau:

Bước 1: Embedding toàn bộ các token của cặp câu bằng các véc tơ nhúng từ pretrain model. Các token embedding bao gồm cả 2 token là [CLS] và [SEP] để đánh dấu vị trí bắt đầu của câu hỏi và vị trí ngăn cách giữa 2 câu. 2 token này sẽ được dự báo ở output để xác định các phần Start/End Spand của câu output.

Bước 2: Các embedding véc tơ sau đó sẽ được truyền vào kiến trúc multi-head attention với nhiều block code (thường là 6, 12 hoặc 24 blocks tùy theo kiến trúc BERT). Ta thu được một véc tơ output ở encoder.

Bước 3: Để dự báo phân phối xác suất cho từng vị trí từ ở decoder, ở mỗi time step chúng ta sẽ truyền vào decoder véc tơ output của encoder và véc tơ embedding input của decoder để tính encoder-decoder attention. Sau đó projection qua liner layer và softmax để thu được phân phối xác suất cho output tương ứng ở time step .

Bước 4: Trong kết quả trả ra ở output của transformer ta sẽ cố định kết



Hình 2.3: Pre-training and Fine-Tuning [17]

quả của câu Question sao cho trùng với câu Question ở input. Các vị trí còn lại sẽ là thành phần mở rộng Start/End Span tương ứng với câu trả lời tìm được từ câu input.

2.1.2 Các biến thể của BERT

- **RoBERTa:** Viết tắt của "Robustly Optimized BERT Approach". RoBERTa là một biến thể của BERT do Meta hợp tác với Đại học Washington phát triển. Được xem là phiên bản nâng cao của BERT ban đầu, RoBERTa được huấn luyện trên một bộ dữ liệu lớn gấp 10 lần so với bộ dữ liệu của BERT. Về kiến trúc, điểm khác biệt chính là việc sử dụng phương pháp học mặt nạ động thay vì mặt nạ tĩnh. Kỹ thuật này bao gồm việc sao chép dữ liệu huấn luyện và che giấu dữ liệu đó 10 lần, mỗi lần với một chiến lược mặt nạ khác nhau, giúp RoBERTa học được các biểu diễn từ vựng mạnh mẽ và tổng quát hơn.
- **DistilBERT:** Kể từ khi các mô hình ngôn ngữ lớn (LLM) đầu tiên ra đời vào cuối những năm 2010, xu hướng phát triển các LLM ngày càng lớn và phức tạp hơn. Điều này có ý nghĩa vì có một mối quan hệ trực tiếp giữa kích thước của mô hình và độ chính xác của nó. Tuy nhiên, mô hình càng lớn thì càng cần nhiều tài nguyên để vận hành, khiến cho việc tiếp cận trở nên khó khăn hơn. DistilBERT được thiết kế để làm cho BERT trở nên dễ tiếp cận hơn bằng cách cung cấp một phiên

bản nhỏ hơn, nhanh hơn, rẻ hơn và nhẹ hơn.

- **ALBERT:** Viết tắt của "A Lite BERT", ALBERT được thiết kế đặc biệt để tăng hiệu quả của BERT trong quá trình huấn luyện. Do việc huấn luyện các mô hình lớn hơn thường gặp phải các hạn chế về bộ nhớ, thời gian huấn luyện kéo dài và sự suy giảm hiệu suất không mong muốn, các nhà phát triển ALBERT đã phát triển hai kỹ thuật giảm tham số để giảm yêu cầu về bộ nhớ và tăng tốc độ trong quá trình huấn luyện.
- **PhoBert:** Đây là một pre-trained được huấn luyện monolingual language, tức là chỉ huấn luyện dành riêng cho tiếng Việt. Việc huấn luyện dựa trên kiến trúc và cách tiếp cận giống RoBERTa của Facebook được Facebook giới thiệu giữa năm 2019.PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT.PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder. Do cách tiếp cận theo tư tưởng của RoBERTa, PhoBERT chỉ sử dụng task Masked Language Model để train, bỏ đi task Next Sentence Prediction.

2.1.3 Ứng dụng thực tế và tác động đến NLP

Được cung cấp năng lượng bởi máy biến áp Transformer, BERT có thể đạt được kết quả tiên tiến trong nhiều nhiệm vụ NLP. Dưới đây là một số bài kiểm tra mà BERT vượt trội:

- Trả lời câu hỏi: BERT là một trong những chatbot sử dụng máy biến áp đầu tiên, mang lại kết quả ấn tượng.
- Phân tích tình cảm: Ví dụ: BERT đã thành công trong việc dự đoán dấu câu tích cực hoặc tiêu cực cho các bài đánh giá phim.
- Tạo văn bản: Là tiền thân của chatbot thế hệ tiếp theo, BERT đã có thể tạo các văn bản dài với những lời nhắc đơn giản.
- Tóm tắt văn bản: Tương tự, BERT có thể đọc và tóm tắt văn bản từ các lĩnh vực phức tạp, bao gồm luật pháp và chăm sóc sức khỏe.

- Dịch ngôn ngữ: BERT đã được đào tạo về dữ liệu được viết bằng nhiều ngôn ngữ. Điều đó làm cho nó trở thành một mô hình đa ngôn ngữ, có nghĩa là rất phù hợp cho việc dịch ngôn ngữ.
- Tự động hoàn thành nhiệm vụ: BERT có thể được sử dụng cho các tác vụ tự động hoàn thành, chẳng hạn như trong email hoặc dịch vụ nhắn tin.

Nhiều LLM đã được thử nghiệm trong các bộ thử nghiệm, nhưng không có nhiều LLM được đưa vào các ứng dụng đã được thiết lập tốt. BERT là một trong những LLM hiện đại đầu tiên nhưng khác xa với lỗi thời, BERT vẫn là một trong những LLM thành công nhất và được sử dụng rộng rãi. Nhờ tính chất nguồn mở của nó, ngày nay, có nhiều biến thể và hàng trăm phiên bản BERT được đào tạo trước được thiết kế cho các nhiệm vụ NLP cụ thể.

2.2 Mô hình BERT trong bài toán phân tích cảm xúc

Đầu vào

- **Thu thập dữ liệu**: Các bộ dữ liệu về đánh giá/bình luận của khách hàng về nhà hàng hoặc khách sạn, bình luận trên các trang mạng xã hội, các bài báo,... nhằm thực hiện mục đích phân tích cảm xúc.
- **Tiền xử lý**: Làm sạch dữ liệu bằng cách loại bỏ URL, email, chữ số, ký tự đặc biệt vì chúng không đóng góp vào việc phát hiện cảm xúc mà chỉ tạo ra nhiễu. Để mô hình mô phỏng cách con người hiểu cảm xúc, không nên theo quy tắc thông thường của việc loại bỏ từ dừng và lemmatization trong NLP.
- **Gán nhãn**: Gán nhãn dữ liệu là một trong những bước quan trọng nhất vì nó hướng dẫn quá trình học có giám sát. Có bộ dữ liệu có thể đã được gán nhãn hoặc tự gán nhãn thủ công, tự gán nhãn dựa trên thang đo đánh giá năm sao,... Các hệ thống xử lý ngôn ngữ tự nhiên dựa trên quy tắc như VADER có thể được sử dụng để theo dõi cảm xúc ban đầu, giúp việc xem xét của con người ở giai đoạn tiếp theo dễ dàng hơn.

- **Cân bằng dữ liệu**: Học trên tập dữ liệu mất cân bằng có xu hướng ưu tiên cho lớp chiếm đa số, dẫn đến độ chính xác sai lệch. Điều này đặc biệt có vấn đề khi ta quan tâm đến việc phân loại chính xác lớp thiểu số. Hầu hết các chatbot ngày nay phục vụ hỗ trợ khách hàng, do đó các tin nhắn là các truy vấn rất đa dạng và thường không mang cảm xúc. Cần giảm mẫu các mẫu Trung tính (dựa trên phân phối tần suất) và tăng mẫu các lớp khác để cân bằng tập dữ liệu. Tăng mẫu các lớp thiểu số bằng các kỹ thuật Tăng cường Dữ liệu. Để cải thiện tổng quát hóa, Tăng cường Dữ liệu là một chiến lược nổi tiếng. Tăng cường các mẫu Tích cực và Tiêu cực có ý nghĩa bằng cách Dịch ngược với nhiều ngôn ngữ và thay thế từ đồng nghĩa cho các lớp thiểu số giúp giảm nhãn thủ công nhiều văn bản bằng cách tạo ra các văn bản tương tự mới, và tăng độ chính xác đáng kể.

2.2.1 Phương pháp BERT Tokenizer

- **WordPiece Tokenization**
  - WordPiece là một thuật toán tokenization đặc biệt, được phát triển để giải quyết vấn đề về từ vựng lớn trong ngôn ngữ tự nhiên.
  - Subword Tokenization: Thay vì chia nhỏ văn bản thành các từ đầy đủ, WordPiece chia từ thành các đơn vị con, thường là các nhóm ký tự có ý nghĩa. Ví dụ, từ "playing" có thể được chia thành "play" và "##ing". Ký hiệu "##" chỉ ra rằng phần tử sau nó là một phần của từ gốc.
- **Handling Unknown Words**
  - Unknown Tokens (UNK): Khi gặp từ không có trong từ điển, tokenizer BERT có thể phân tách từ đó thành các phần nhỏ hơn để tìm các subword tokens có trong từ điển. Điều này giúp giảm thiểu số lượng token không xác định (UNK) và giúp mô hình xử lý văn bản một cách hiệu quả hơn.
- **Special Tokens**
  - [CLS] Token: Được thêm vào đầu mỗi chuỗi văn bản để biểu diễn

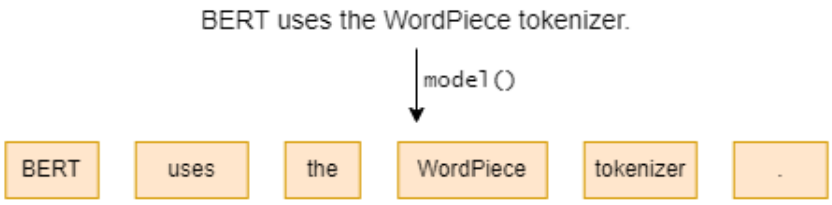
thông tin tổng quan của toàn bộ câu, đặc biệt hữu ích cho các tác vụ phân loại câu.

  - [SEP] Token: Được sử dụng để phân tách các câu hoặc các đoạn văn trong cùng một chuỗi văn bản đầu vào, rất quan trọng cho các tác vụ như trả lời câu hỏi hoặc tương tác giữa các câu.

- **Token ID Mapping**
  - Vocabulary: Tokenizer có một từ điển (vocabulary) ánh xạ mỗi token đến một số nguyên duy nhất (token ID). Đây là bước chuyển đổi từ token thành các số nguyên mà mô hình BERT có thể xử lý.
  - Input IDs: Sau khi tokenization, mỗi token được chuyển thành ID tương ứng để tạo thành chuỗi đầu vào cho mô hình.

- **Truncation and Padding**
  - Truncation: Đối với các chuỗi văn bản dài hơn mức tối đa cho phép, tokenizer sẽ cắt ngắn (truncate) chuỗi để vừa với kích thước đầu vào của mô hình.
  - Padding: Để đảm bảo rằng tất cả các chuỗi đầu vào có cùng độ dài, tokenizer thêm các token đệm (padding tokens) vào cuối chuỗi ngắn hơn.

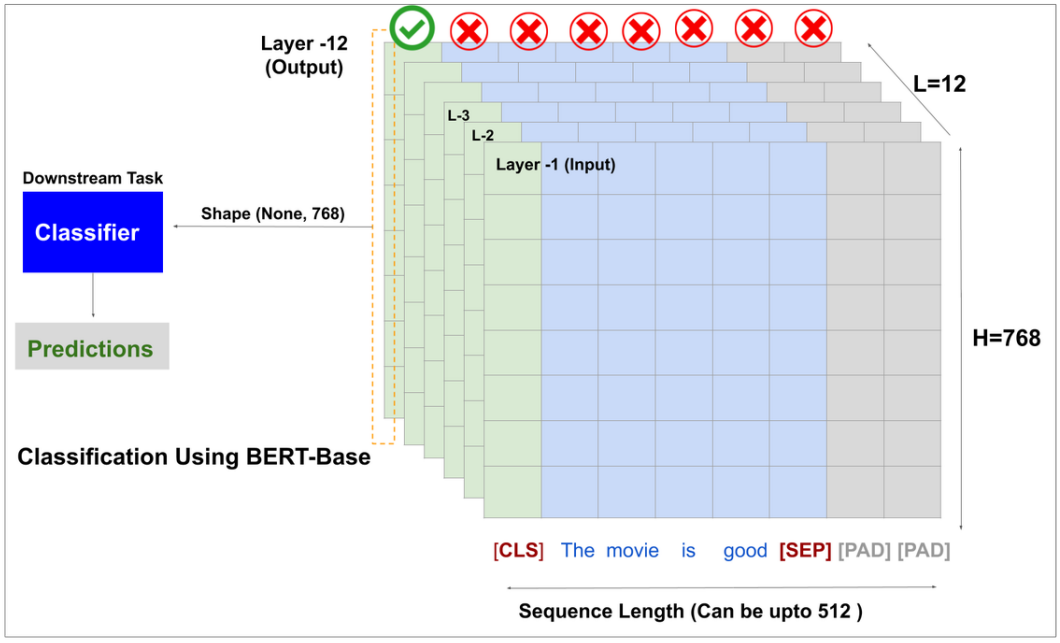
- **Example of Tokenization Process**
  - Ví dụ với câu "I love playing football":
  - Tokenization: ["I", "love", "playing", "football"]
  - WordPiece: ["I", "love", "play", "##ing", "foot", "##ball"]
  - Add Special Tokens: ["[CLS]", "I", "love", "play", "##ing", "foot", "##ball", "[SEP]"]
  - Token IDs: [101, 1045, 2293, 2377, 2075, 2376, 3531, 102]



Hình 2.4: BERT Tokenizer

2.2.2 Mô hình hóa BERT

- **Input IDs (Các mã định danh đầu vào)** – Các mã định danh đầu vào thường là các tham số duy nhất cần được truyền vào mô hình làm đầu vào. Các chỉ số token là các biểu diễn số của các token tạo nên các chuỗi sẽ được sử dụng làm đầu vào cho mô hình.
- **Attention Mask (Mặt nạ chú ý)** – Mặt nạ chú ý được sử dụng để tránh thực hiện chú ý trên các chỉ số token đệm. Giá trị của mặt nạ có thể là 0 hoặc 1, 1 cho các token UNMASKED, 0 cho các token MASKED.
- **Token Type IDs (Các mã định danh loại token)** – Chúng được sử dụng trong các trường hợp như phân loại chuỗi hoặc trả lời câu hỏi. Vì những trường hợp này yêu cầu hai chuỗi khác nhau được mã hóa trong cùng một mã định danh đầu vào. Các token đặc biệt, chẳng hạn như token phân loại `[CLS]` và token phân tách `[SEP]` được sử dụng để tách các chuỗi.

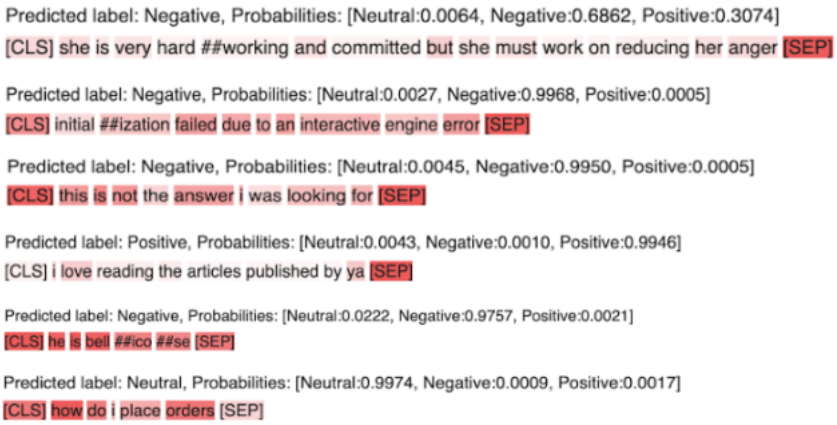


Hình 2.5: Mô hình hóa BERT [18]

Đầu ra

Lớp đầu ra của BERT thường là một lớp softmax để phân loại các vector ngữ nghĩa thành các nhãn cảm xúc (ví dụ: tích cực, tiêu cực, trung tính).

Đối với việc học có giám sát, các giá trị Accuracy, Precision, Recall, F1 dùng để đánh giá mức độ hiệu quả và chính xác của mô hình.



Hình 2.6: Ví dụ đầu ra BERT trong phân tích cảm xúc

Chương 3 KẾT QUẢ THỰC NGHIỆM

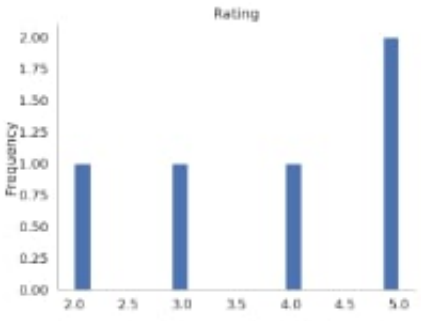
3.1 Bộ dữ liệu

Bộ dữ liệu được sử dụng trong đồ án này bao gồm năm nghìn đánh giá/bình luận của khách hàng về trải nghiệm tại khách sạn đươc thu thập từ Tripadvisor.

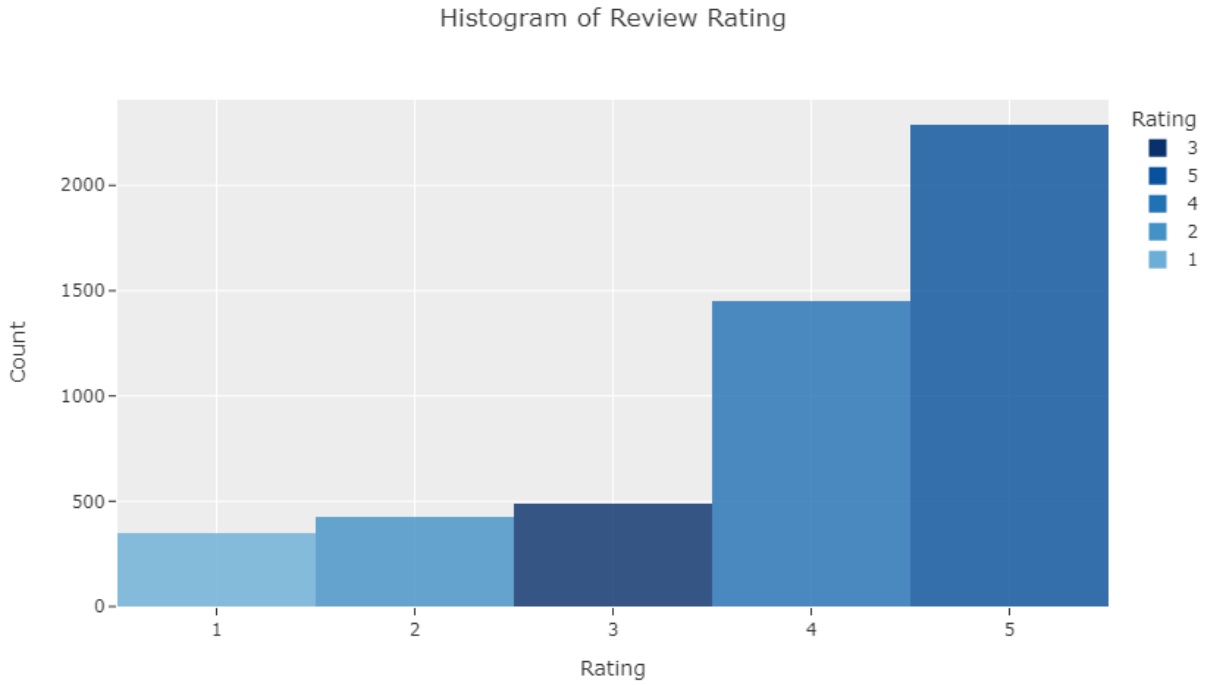
Bộ dữ liệu bao gồm hai cột: Review và Rating.

	Review	Rating
0	nice hotel expensive parking got good deal sta...	4
1	ok nothing special charge diamond member hilto...	2
2	nice rooms not 4* experience hotel monaco seat...	3
3	unique, great stay, wonderful time hotel monac...	5
4	great stay great stay, went seahawk game aweso...	5

Distributions



Hình 3.1: Bộ dữ liệu



Hình 3.2: Bộ dữ liệu

3.2 Tiền xử lí dữ liệu

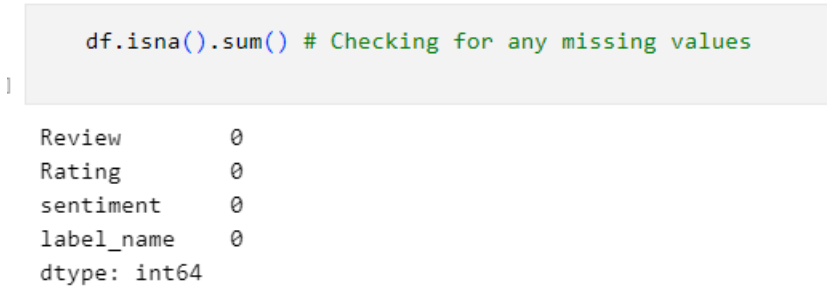
Tiền xử lý là một bước thiết yếu trước khi đi vào quá trình huấn luyện mô hình. Đặc biệt hơn, trong môi trường trực tuyến trên mạng xã hội, các tài liệu thường không được viết bằng văn bản chính thức. Điều này đặc biệt đúng với thanh thiếu niên, những người thường sử dụng nhiều biểu tượng cảm xúc, dạng rút gọn của từ, ký hiệu và ký tự đặc biệt, từ viết sai chính tả, lỗi ngữ pháp, hoặc từ ghép. Trước khi được đưa vào mô hình, dữ liệu phải trải qua các bước tiền xử lý cần thiết:

- Xóa các biểu tượng và ký tự đặc biệt: Các ký tự đặc biệt không mang ý nghĩa phân loại và có thể gây nhiễu trong quá trình phân tích. Chuyển tất cả chữ về dạng chữ thường: Mỗi số và ký tự đặc biệt đều được biểu diễn bằng một dãy nhị phân trong bộ nhớ máy tính. Chữ in hoa và chữ thường có mã Unicode khác nhau, dù về mặt ngữ nghĩa là giống nhau, nhưng máy tính có thể không phân biệt được trong dữ liệu đầu vào, dẫn đến kết quả dự đoán bị ảnh hưởng. Do đó, việc chuyển tất cả chữ về dạng chữ thường là hợp lý cho hệ thống phân tích và dự đoán.

- Loại bỏ 'stopword': trong tiếng Anh các từ này có thể kể đến như the, is, at, on, which, in, some, many hay trong tiếng Việt là các từ cái, các, cả,.... Các từ này thường sẽ được loại bỏ để giảm kích thước của bộ từ vựng.

- Mỗi câu bao gồm một số từ có cường độ và hành vi khác nhau. Trong bước trước, chúng ta đã tính toán loại cực tính tức là tích cực, tiêu cực và trung tính. Mỗi loại sẽ có một đám mây từ hiển thị các từ khác nhau trong danh mục đó. Wordcloud hiển thị tất cả các từ và tần số cho biết kích thước của các từ tương ứng. Từ lớn hơn đại diện cho tần suất xuất hiện cao trong văn bản.

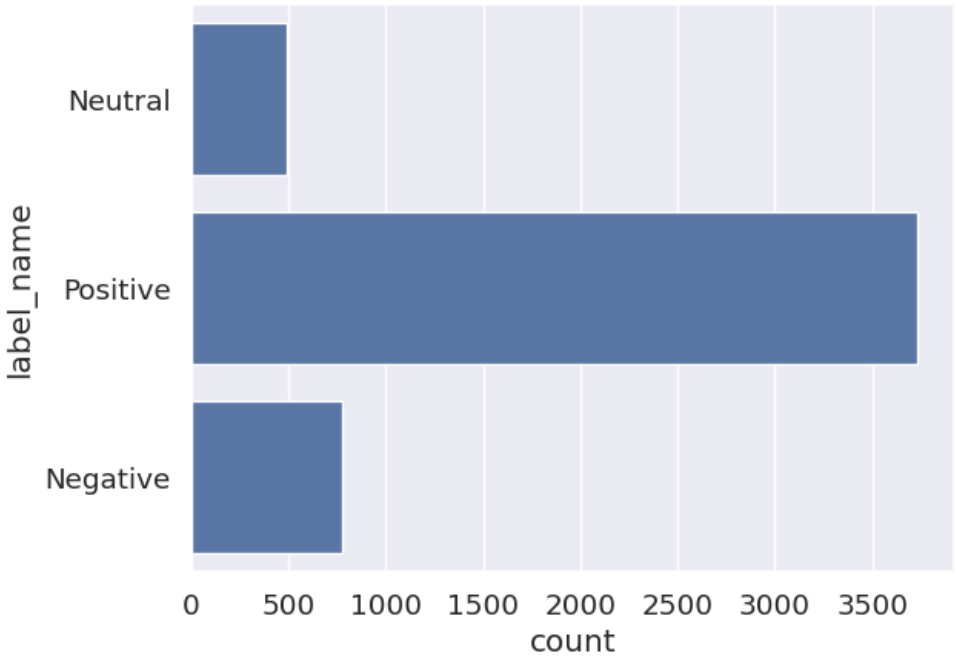
- Xóa dòng dữ liệu trống: Tập dữ liệu thu thập có thể chứa nhiều dòng dữ liệu trống, và dữ liệu trống không có ý nghĩa trong quá trình phân tích, gây lãng phí bộ nhớ lưu trữ.



Hình 3.3: Loại bỏ NULL

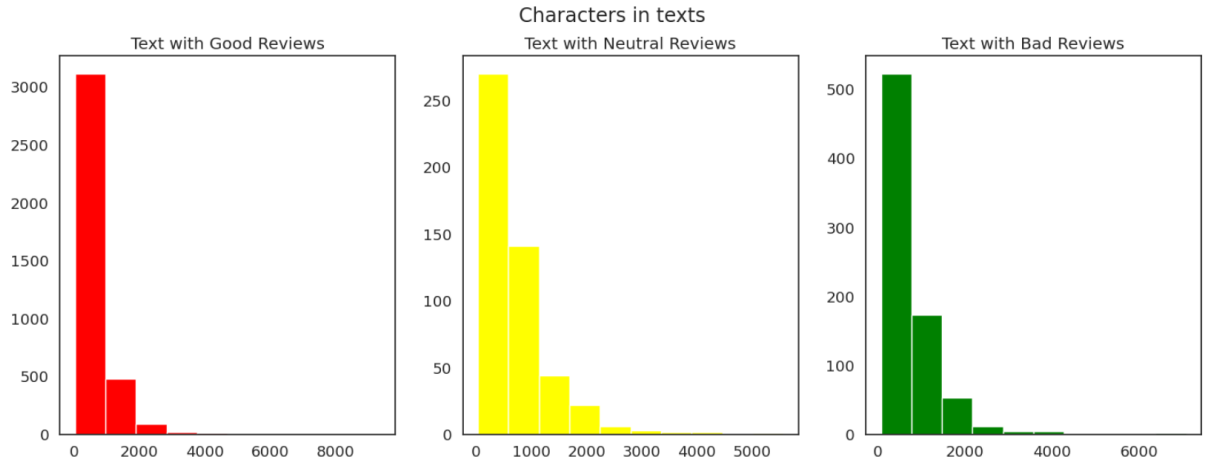
- Gán nhãn dữ liệu: Để thực hiện quá trình gán nhãn dữ liệu trước khi đưa vào huấn luyện, nghiên cứu áp dụng phương pháp phân loại cảm xúc theo điểm số đánh giá (Rating) của khách hàng để phân chia tập dữ liệu đã thu thập được thành hai bộ dữ liệu được gán nhãn theo quy tắc sau: Rate $<=$ 2: bình luận nào đánh giá dưới 2 sao sẽ được dán nhãn là tiêu cực (negative). Rate $=$ 3: bình luận nào đánh giá 3 sao sẽ được dán nhãn là trung tính (neutral) Rate $>$ 3: bình luận nào đánh giá trên 3 sao sẽ được dán nhãn là tích cực (positive).

- Chuyển đổi nhãn dữ liệu về dạng số: Negative $=$ 0, Neutral $=$ 1, Positive $=$ 2.

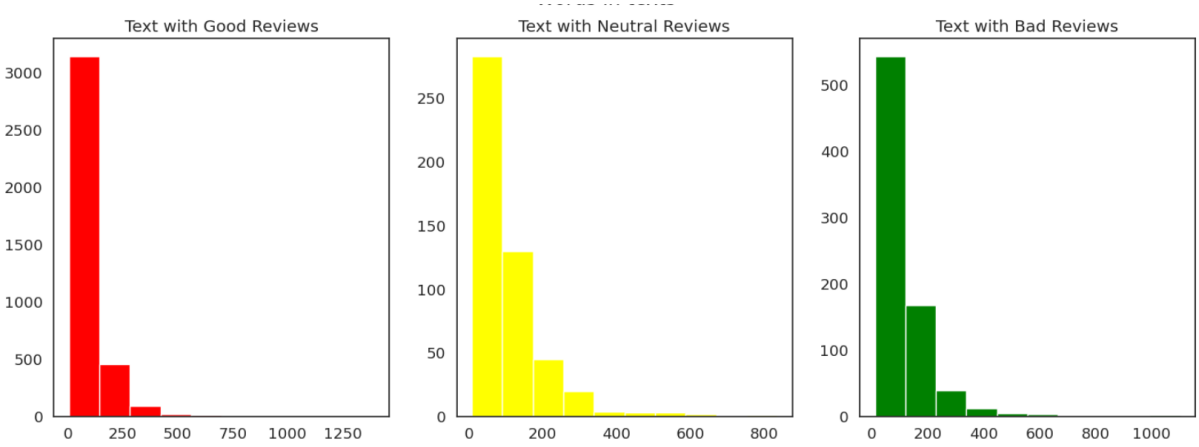


Hình 3.4: Gán nhãn dữ liệu

3.3 Biểu diễn biểu đồ trực quan

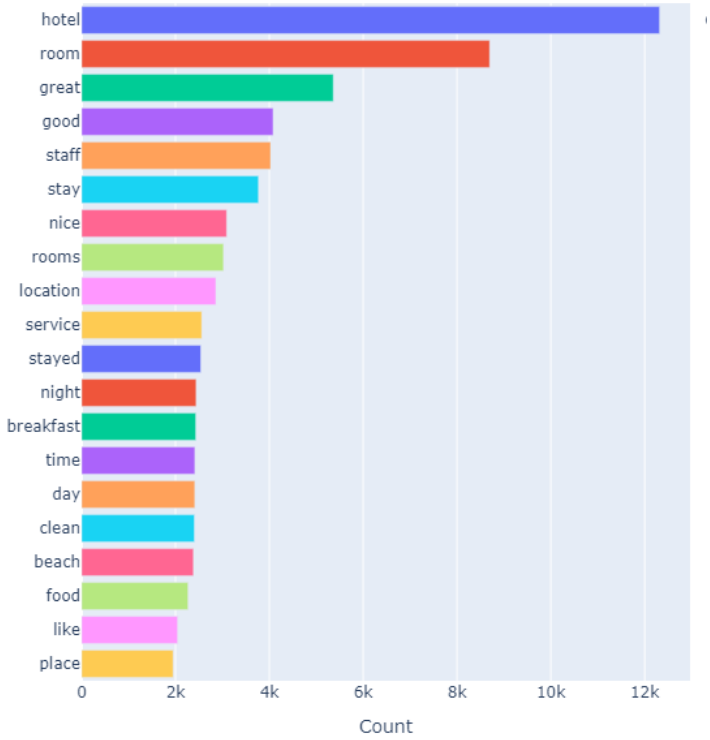


Hình 3.5: Số lượng kí tự trong bộ dữ liệu

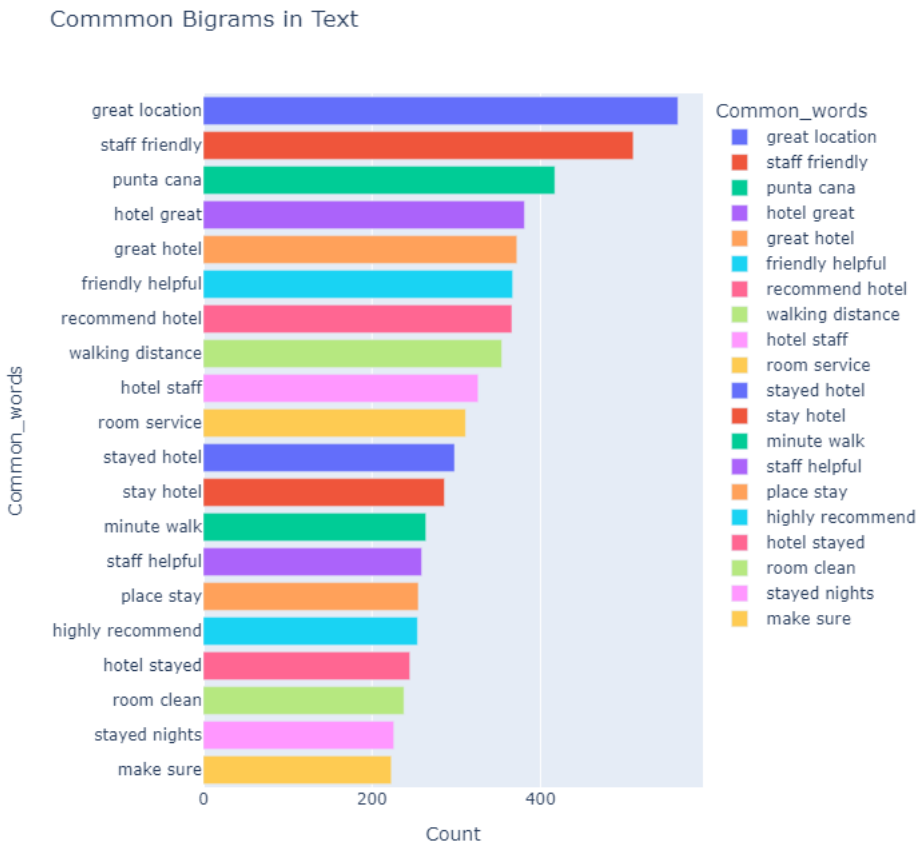


Hình 3.6: Số lượng từ trong bộ dữ liệu

Biểu diễn các biểu đồ N-gram: Unigram và Bigram



Hình 3.7: Các từ ngữ xuất hiện nhiều nhất



Hình 3.8: Các từ ghép xuất hiện nhiều nhất

3.4 Áp dụng mô hình

3.4.1 Chia bộ dữ liệu train và test

- Tỉ lệ tập train chiếm 0.75, tập test 0.25.

category	label	data_type	text
Negative	0	train	658
		val	116
Neutral	1	train	417
		val	74
Positve	2	train	3175
		val	580

Bảng 3.1: Chia tập dữ liệu

3.4.2 Thực hiện Tokenizer và Encoding dữ liệu:

- Trước khi tiến hành chạy mô hình, ta sẽ bắt đầu tách từ Tokenizer sử dụng luôn thư viện BertTokenizer sẵn có, tách các input text thành một tập hợp các token riêng biệt.

- Sau đó ta thực hiện encoding dữ liệu theo phương pháp BERT, tokenizer các token đặc biệt như token bắt đầu ([CLS]), token kết thúc ([SEP]), và các token padding ([PAD]) vào các câu.

- Padding và Truncating: Các câu sẽ được đệm (padding) hoặc cắt ngắn (truncating) để đảm bảo tất cả các câu đầu vào có cùng độ dài, phù hợp với yêu cầu của mô hình BERT.

3.4.3 Xây dựng mô hình

- Ta sẽ sử dụng mô hình BertForSequenceClassification từ thư viện Transformers và xây dựng hàm Traning Loop để tiến hành huấn luyện mô hình.

Mô hình BERT sẽ có một lớp phân loại thêm vào cuối cùng để dự đoán nhãn cảm xúc.

```
model = BertForSequenceClassification.from_pretrained(
    'bert-base-uncased',
    num_labels = len(label_dict),
    output_attentions = False,
    output_hidden_states = False
)
```

Hình 3.9: Xây dựng mô hình

Hàm Traning Loop để bắt đầu huấn luyện mô hình BERT:

```
for epoch in tqdm(range(1, epochs+1)):
    model.train()

    loss_train_total = 0
    progress_bar = tqdm(dataloader_train,
                        desc="Epoch {:1d}".format(epoch),
                        leave=False,
                        disable=False)

    for batch in progress_bar:
        model.zero_grad()
        batch = tuple(b.to(device) for b in batch)
        inputs = {
            'input_ids'       : batch[0],
            'attention_mask'  : batch[1],
            'labels'          : batch[2]
        }

        outputs = model(**inputs)
        loss = outputs[0]
        loss_train_total += loss.item()
        loss.backward()
```

Hình 3.10: Xây dựng mô hình

- Training Loop: Mô hình được huấn luyện qua nhiều epoch.

- Trong mỗi epoch: Dữ liệu huấn luyện được đưa qua mô hình theo từng batch. Loss được tính toán và gradient được lan truyền ngược (backpropagation) để cập nhật các trọng số của mô hình.

- Evaluation: Sau mỗi epoch, mô hình được đánh giá trên tập dữ liệu kiểm tra. Các chỉ số đánh giá như accuracy, precision, recall, và F1-score được tính toán để theo dõi hiệu quả của mô hình.

- Loss Function: Sử dụng hàm mất mát CrossEntropyLoss, phù hợp cho bài toán phân loại đa lớp.

- Optimizer: Sử dụng AdamW Optimizer, một biến thể của thuật toán Adam, được điều chỉnh để làm việc tốt với các mô hình Transformer.

- Learning Rate Scheduler: Sử dụng get_linear_schedule_with_warmup để điều chỉnh tốc độ học trong quá trình huấn luyện. Ban đầu, tốc độ học sẽ tăng dần (warmup) và sau đó giảm dần theo từng epoch.

- Sau giai đoạn huấn luyện ban đầu, mô hình có thể được tinh chỉnh thêm (fine-tuning) bằng cách điều chỉnh các hyperparameter hoặc sử dụng các kỹ thuật như dropout để giảm overfitting.

- Sau khi hoàn thành quá trình huấn luyện và tinh chỉnh, mô hình được lưu lại để sử dụng trong các bước dự đoán sau này.

3.5 Đánh giá mô hình

Accuracy: Độ chính xác trung bình của mô hình BERT là tỷ lệ giữa kết quả dự đoán với dữ liệu thực tế. Trong kết quả này, độ chính xác đạt 87%.

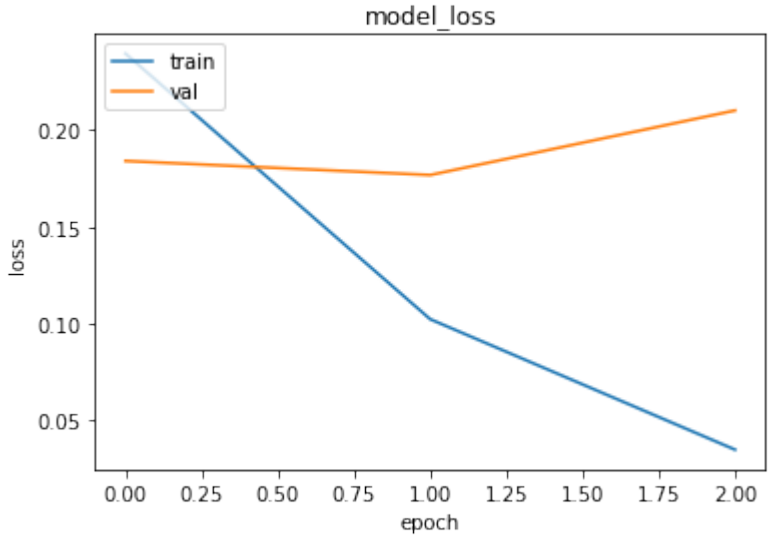
Precision: Được định nghĩa là số lượng dự đoán chính xác hoặc có liên quan trong số tất cả các dự đoán dựa trên lớp tích cực. Ví dụ, lớp cảm xúc "positive" có độ chính xác là 95%.

Recall: Chỉ số này thể hiện Recall càng cao, tức là số điểm là positive bị bỏ sót càng ít. Recall của lớp cảm xúc "positive" là 93%.

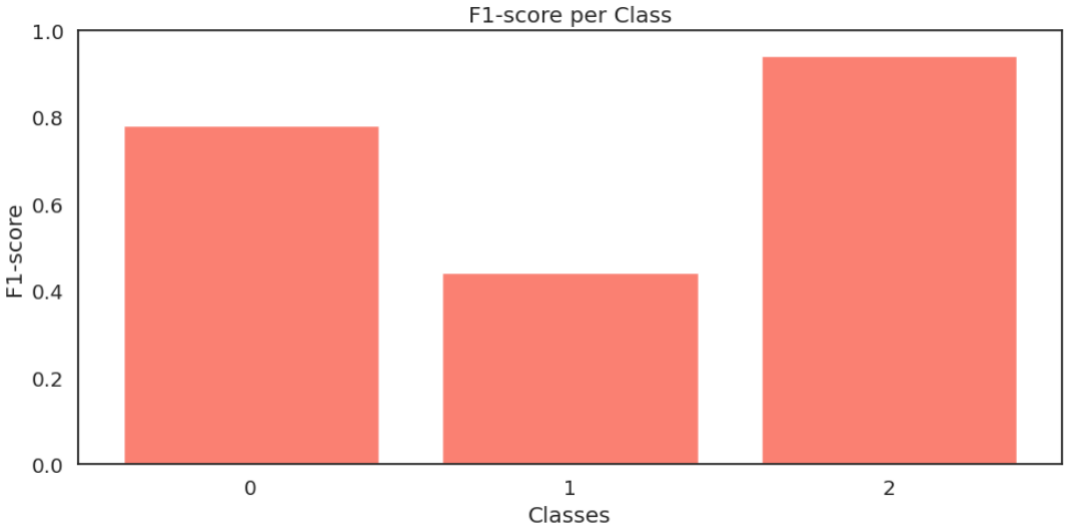
F1-score: Có một số trường hợp cần tối ưu hóa cân bằng giữa độ chính xác (precision) và khả năng thu hồi (recall). Điểm F1 là giá trị trung bình hài hòa của precision và recall và giúp tối ưu hóa mô hình cho cả hai yếu tố này. Ví dụ, F1-score của lớp "positive" là 0.94, cho thấy sự cân bằng tốt giữa precision và recall.

	Precision	**Recall**	**F1-Score**	**Support**
0	0.73	0.83	0.78	102
1	0.45	0.44	0.44	75
2	0.95	0.93	0.94	573
Accuracy			0.87	750
Macro Avg	0.71	0.73	0.72	750
Weighted Avg	0.87	0.87	0.87	750

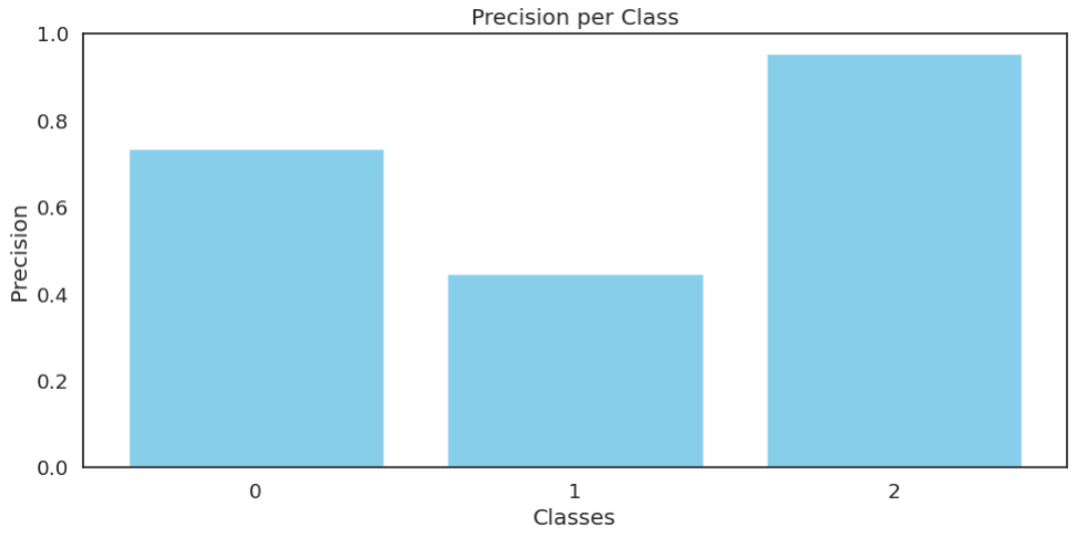
Bảng 3.2: Classification Report



Hình 3.11: Biểu đồ model loss



Hình 3.12: Biểu đồ chỉ số F1-Score



Hình 3.13: Biểu đồ chỉ số Precision

Hình 3.14: Biểu đồ chỉ số Recall

3.6 Mô hình dự đoán

```
model = BertForSequenceClassification.from_pretrained("bert-base-uncased",
                                                      num_labels=len(label_dict),
                                                      output_attentions=False,
                                                      output_hidden_states=False)
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased', do_lower_case=True)

# Định dạng câu văn bạn muốn dự đoán
sentence = "This is amazing."

# Tiền xử lý câu văn
inputs = tokenizer(sentence, return_tensors="pt", padding=True, truncation=True, max_length=128)
input_ids = inputs["input_ids"]
attention_mask = inputs["attention_mask"]

# Dự đoán
with torch.no_grad():
    outputs = model(input_ids, attention_mask=attention_mask)

# Lấy dự đoán
predictions = torch.argmax(outputs.logits, dim=1).item()
print("Predicted label:", predictions)
```

```
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Predicted label: 2
```

Hình 3.15: Dự đoán

Ta sẽ thử nhập vào 1 câu văn bất kì (sentence), và sử dụng model đã train ở phía trên dự đoán nhãn cảm xúc cho câu văn đó.

Với Predicted label = 0 ứng với Negative, = 1 ứng với Neutral, = 2 ứng với Positive.

Sau khi hoàn thành quá trình huấn luyện, mô hình đạt được kết quả như mong đợi với độ chính xác cao trên tập dữ liệu kiểm tra. Kết quả cho thấy mô hình BERT có khả năng phân tích cảm xúc từ các đánh giá của khách hàng với độ chính xác cao, đặc biệt trong các lớp cảm xúc tích cực và tiêu cực.

KẾT LUẬN

Tổng Quan

Đồ án này đã thực hiện việc phân tích cảm xúc của người dùng dựa trên các đánh giá và phản hồi dịch vụ tại khách sạn, thông qua việc áp dụng mô hình BERT (Bidirectional Encoder Representations from Transformers). Đây là một mô hình mạnh mẽ trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), cho phép hiểu rõ ngữ cảnh của từ và câu trong văn bản.

Kết Quả Đạt Được

Xử Lý Dữ Liệu

- Dữ liệu đã được thu thập từ các nguồn đánh giá khách sạn và được gán nhãn dựa trên điểm số đánh giá của khách hàng:
  - Đánh giá ≤ 2 sao: tiêu cực.
  - Đánh giá $= 3$ sao: trung tính.
  - Đánh giá > 3 sao: tích cực.
- Sau khi gán nhãn, dữ liệu được chuyển đổi về dạng số và chia thành các tập train và test với tỉ lệ 75% và 25%.

Áp Dụng Mô Hình

- Mô hình BERT được áp dụng để phân tích cảm xúc với các bước như tokenization, encoding dữ liệu, huấn luyện mô hình từ tập train.
- Mô hình BERT được huấn luyện và đánh giá, cho thấy độ chính xác trung bình đạt được là 87%.

Đóng Góp Của Đồ Án

Đồ án đã chứng minh hiệu quả của việc sử dụng mô hình BERT trong phân tích cảm xúc, đặc biệt là trong việc phân tích các đánh giá của người dùng về dịch vụ khách sạn. Kết quả này không chỉ giúp cải thiện chất lượng dịch vụ mà còn cung cấp thông tin quan trọng cho các chiến lược marketing của doanh nghiệp.

Hướng Phát Triển Tương Lai

Trong tương lai, việc mở rộng và cải thiện mô hình phân tích cảm xúc có thể bao gồm:

- Sử dụng các mô hình tiên tiến hơn hoặc kết hợp nhiều mô hình để tăng độ chính xác.
- Thu thập và phân tích dữ liệu từ nhiều nguồn hơn để có cái nhìn toàn diện hơn về cảm xúc của người dùng.
- Áp dụng phân tích cảm xúc vào các lĩnh vực khác ngoài dịch vụ khách sạn, như thương mại điện tử, mạng xã hội, v.v.

Đồ án này đã đặt nền móng vững chắc cho việc sử dụng công nghệ NLP trong phân tích cảm xúc, mở ra nhiều cơ hội phát triển trong nghiên cứu và ứng dụng thực tế.

Chỉ mục

Tài liệu tham khảo

[1] Lê Si Lắc, *"A Research on Sentiment Analysis"*, 2021.

[2] Agustini, *"Sentiment Analysis on Social Media using Machine Learning-Based Approach"*, June 2021.

[3] Shashank Kalluri, *"Deep Learning Based Sentiment Analysis"*, January 2023.

[4] Doaa Mohey El-Din Mohamed Hussein, *"Analyzing Scientific Papers Based on Sentiment Analysis"*, June 2016.

[5] Ngoc C. Lê, Nguyen The Lam, Son Hong Nguyen, Duc Thanh Nguyen, *"On Vietnamese Sentiment Analysis: A Transfer Learning Method"*, July 2020.

[6] Bo Pang and Lillian Lee, *"Opinion mining and Sentiment analysis"*, 2007.

[7] Binh Thanh Kieu and Son Bao Pham, *"Sentiment analysis for Vietnamese"*, 2010.

[8] Nguyen Thi Duyen, Ngo Xuan Bach and Tu Minh Phuong, *"An Empirical Study on Sentiment analysis for Vietnamese"*, 2014.

[9] Vi Ngo Van, Minh Hoang Van, Tam Nguyen Thanh, *"Sentiment analysis for Vietnamese using Support Vector Machines with application on Facebook comments"*, Proc. VLSP 2016.

[10] Thien Khai Tran and Tuoi Thi Phan, *"Computing Sentiment Scores of Verb Phrases for Vietnamese"*, 2016.

[11] Lê Thanh Hương, *Bài giảng xử lý ngôn ngữ tự nhiên*, Đại học Bách Khoa Hà Nội.

[12] Sinh viên nghiên cứu khoa học Eureka, *Phân tích cảm xúc trong Tiếng Việt*, 2018

[13] FPT Digital, *Xử lý ngôn ngữ tự nhiên: Công nghệ giúp máy tính hiểu và giao tiếp với con người*, 2020

[14] Jashijim, *Latent Semantic Analysis* , 2021

[15] Upen, *Difference Between Semantics and Pragmatics* , 2018

[16] EMIR KOÇAK, *In depth series: Sentiment Analysis w Transformers* , Kaggle, 2022

[17] Pham Dinh Khanh, *Bài 36: BERT Model*, Khoa học dữ liệu - Khanh's blog , May 2020

[18] Engati Team, *How to use BERT for sentiment analysis?*, Engati Simply Intelligent, 2023

[19] Javier Canales Luna, *What is BERT? An Intro to BERT Models*, Learn Data and AI Skill, 2023

[20] Samia Khalid, *BERT Explained: A Complete Guide with Theory and Tutorial*,Towards Machine Learning, 2019