

TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN



PBL6: DỰ ÁN CHUYÊN NGÀNH
KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO

TÊN ĐỀ TÀI:
PHÂN TÍCH CẢM XÚC COMMENT TRÊN
MỘT BÀI VIẾT ĐÁNH GIÁ SẢN PHẨM

Giảng viên hướng dẫn: TS. NINH KHÁNH DUY

Nhóm học phần: 21N15D

HỌ VÀ TÊN SINH VIÊN	MSSV
Trần Công Thiên Hữu	102210296
Lê Huỳnh Đức	102210292
Lê Huỳnh Đức	102210312

Đà Nẵng, 12/2024

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Đánh giá
Trần Công Thiên Hữu (21TCLC_KHDL)	<ul style="list-style-type: none">- Thu thập dữ liệu- Trực quan hóa dữ liệu- Xây dựng server AI, web app- Làm slide, báo cáo	Đã hoàn thành
Lê Huỳnh Đức (21TCLC_KHDL)	<ul style="list-style-type: none">- Nghiên cứu đề tài và hướng triển khai- Thu thập dữ liệu- Xây dựng, test mô hình CNN-BiLSTM- Làm slide, báo cáo	Đã hoàn thành
Lê Huỳnh Đức (21TCLC_KHDL2, Nhóm trưởng)	<ul style="list-style-type: none">- Nghiên cứu đề tài và hướng triển khai- Thu thập dữ liệu- Tiền xử lý dữ liệu- Xây dựng, test mô hình BERT- Làm slide, báo cáo	Đã hoàn thành

MỤC LỤC

MỤC LỤC	2
DANH MỤC HÌNH ẢNH	4
DANH MỤC BẢNG	6
MỞ ĐẦU	7
1 TỔNG QUAN ĐỀ TÀI	7
2 CƠ SỞ LÝ THUYẾT.....	9
2.1 Ý tưởng	9
2.2 Cơ sở lý thuyết.....	9
2.2.1 Giới thiệu về BERT	9
2.2.2 Giới thiệu về CNN-BiLSTM	12
3 THU THẬP VÀ XỬ LÝ DỮ LIỆU.....	16
3.1 Thu thập dữ liệu.....	16
3.1.1 Nguồn dữ liệu	16
3.1.2 Phương pháp thu thập.....	17
3.2 Phân tích dữ liệu	19
3.2.1 Thống kê mô tả.....	19
3.2.2 Khám phá dữ liệu	22
3.3 Tiền xử lý dữ liệu	27
4 CÁC THUẬT TOÁN TRIỂN KHAI.....	29
4.1. Tổng quan về hai thuật toán	29
4.1 Thuật toán	31
4.1.1 Thuật toán sử dụng mô hình BERT.....	31
4.1.2 Thuật toán sử dụng mô hình CNN-BiLSTM.....	38
4.2 So sánh và đánh giá hai thuật toán	44
4.2.1. So sánh kết quả	44
4.2.2. Đánh giá và kết luận	46

5	THIẾT KẾ ỨNG DỤNG	47
5.1	Tổ chức chương trình	47
5.2	Ứng dụng học máy trong chương trình	48
5.2.1	Cách thức hoạt động của hai mô hình	49
5.2.2	Lợi ích khi sử dụng hai mô hình cùng nhau.....	49
5.3	Kết quả.....	49
5.3.1	Giao diện chính của chương trình.....	49
5.3.2	Kết quả thực thi của chương trình.....	55
6	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	57
6.1	Kết luận.....	57
6.2	Hướng phát triển	57
	TÀI LIỆU THAM KHẢO	59

DANH MỤC HÌNH ẢNH

Hình 1. Thông số model BERT.....	10
Hình 2. Cấu trúc model BERT	10
Hình 3. Hình ảnh thực tế dữ liệu crawl	19
Hình 4. Kết quả thu thập dữ liệu.....	19
Hình 5. Phân bố nhãn cảm xúc.....	20
Hình 6. Biểu đồ phân bố nhãn theo phần trăm.....	21
Hình 7. Biểu đồ KDE độ dài bình luận tích cực.....	23
Hình 8. Biểu đồ KDE độ dài bình luận tiêu cực.....	23
Hình 9. Word cloud cho bình luận tiêu cực (label 0)	25
Hình 10. Word cloud cho bình luận tích cực (label 1)	25
Hình 11. Biểu đồ phân phối 30 từ tích cực và tiêu cực phổ biến	27
Hình 12. Quy trình tiền xử lý dữ liệu	27
Hình 13. Sơ đồ tổng quát sử dụng thuật toán mô hình BERT	29
Hình 14. Sơ đồ tổng quát sử dụng thuật toán CNN-BiLSTM	30
Hình 15. Sơ đồ kiến trúc mô hình BERT-BASE.....	32
Hình 16. Confusion matrix trên tập test với mô hình BERT.....	36
Hình 17. Biểu đồ Training của mô hình BERT.....	37
Hình 18. Sơ đồ tiền xử lý	38
Hình 19. Sơ đồ kiến trúc mô hình CNN-BiLSTM	39
Hình 20. Confusion matrix của tập test với model CNN-BiLSTM.....	43
Hình 21. Biểu đồ training của model CNN-BiLSTM	43
Hình 22. Biểu đồ so sánh số lượng dự đoán giống và khác nhau	45
Hình 23. Biểu đồ Heatmap hai model	45
Hình 24. Kiến trúc web	47
Hình 25. Hình ảnh giao diện chính của chương trình	50
Hình 26. Hình ảnh giao diện đang chạy tiến trình.....	51

Hình 27. Hình ảnh phần đầu giao diện kết quả	52
Hình 28. Biểu đồ tỷ lệ tích cực tiêu cực.....	53
Hình 29. Hình ảnh thông tin hiện ra khi đưa trỏ chuột vào biểu đồ.....	54
Hình 30. Hình ảnh phần sau của giao diện kết quả chương trình	54
Hình 31. Kết quả chạy thực tế	56

DANH MỤC BẢNG

Bảng 1. Đánh giá BERT trên tập test	36
Bảng 2. Kết quả phân loại của CNN-BiLSTM trên tập test	42
Bảng 3. Chỉ số đánh giá của hai model trên tập test	44

MỞ ĐẦU

Trong bối cảnh kinh doanh trực tuyến ngày càng phát triển, việc phân tích cảm xúc từ các bình luận đánh giá sản phẩm đóng vai trò quan trọng trong việc hiểu phản hồi của khách hàng. **Mục đích** của đề tài "*Phân tích cảm xúc comment trên một bài viết đánh giá sản phẩm*" là xây dựng một hệ thống tự động phân loại cảm xúc (tích cực, tiêu cực) từ các bình luận trực tuyến, hỗ trợ cá nhân và doanh nghiệp đánh giá chất lượng sản phẩm và cải thiện trải nghiệm người dùng.

Đề tài sử dụng dữ liệu bình luận bằng tiếng Việt thu thập từ các trang thương mại điện tử Lazada và Tiki, áp dụng và so sánh hai mô hình học sâu là BERT và LSTM. **Phạm vi và đối tượng** nghiên cứu tập trung vào bình luận đánh giá sản phẩm, hướng đến nhóm người dùng có nhu cầu thống kê phản hồi để nhận diện cảm xúc như chủ doanh nghiệp mà người mua sắm. **Phương pháp** thực hiện bao gồm thu thập dữ liệu, tiền xử lý, huấn luyện mô hình và đánh giá hiệu quả, kết hợp với việc xây dựng ứng dụng web đơn giản trên nền Flask API để tích hợp các mô hình.

Cấu trúc đề án gồm các phần: tổng quan về đề tài, cơ sở lý thuyết, quy trình nghiên cứu và triển khai, kết quả và thảo luận, cùng phần kết luận và định hướng phát triển. Thông qua đề tài này, hệ thống không chỉ hỗ trợ phân tích cảm xúc mà còn mở ra tiềm năng ứng dụng AI trong xử lý ngôn ngữ tự nhiên cho các lĩnh vực khác.

1 TỔNG QUAN ĐỀ TÀI

Trong bối cảnh kinh doanh trực tuyến phát triển mạnh mẽ, các bình luận đánh giá sản phẩm đóng vai trò quan trọng trong việc xây dựng niềm tin và định hướng quyết định mua sắm của khách hàng. Các doanh nghiệp ngày càng quan tâm đến việc phân tích phản hồi từ khách hàng để cải thiện sản phẩm, nâng cao chất lượng dịch vụ và tối ưu hóa trải nghiệm người dùng. Tuy nhiên, với lượng lớn dữ liệu được tạo ra mỗi ngày, việc phân tích thủ công không còn hiệu quả, dẫn đến nhu cầu sử dụng công nghệ tự động để xử lý và khai thác thông tin từ dữ liệu này.

Phân tích cảm xúc từ bình luận là một bài toán thực tiễn, mang lại nhiều giá trị ứng dụng trong việc đánh giá chất lượng sản phẩm và xác định xu hướng thị trường. Tuy nhiên, trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) bằng tiếng Việt, các công cụ và hệ thống hỗ trợ hiện nay vẫn còn hạn chế. Trước thực tế đó, đề tài “Phân tích cảm xúc comment trên một bài viết đánh giá sản phẩm” được thực hiện với mục tiêu xây dựng một hệ thống tự động phân loại cảm xúc từ các bình luận trực tuyến, giúp doanh nghiệp và cá nhân dễ dàng hiểu rõ cảm nhận của khách hàng.

Đề tài tập trung vào việc thu thập dữ liệu bình luận tiếng Việt từ các trang thương mại điện tử phổ biến như Lazada và Tiki. Hai mô hình học sâu là BERT (Bidirectional Encoder Representations from Transformers) và CNN-BiLSTM (Convolutional Neural Network & Bidirectional Long Short-Term Memory) được áp dụng để phân tích cảm xúc của các bình luận này, đồng thời so sánh hiệu quả của chúng. Kết quả từ các mô hình sẽ được tích hợp vào một ứng dụng web đơn giản dựa trên Flask API, cho phép triển khai và ứng dụng trong thực tế.

Phạm vi nghiên cứu của đề tài tập trung vào các bình luận đánh giá sản phẩm bằng tiếng Việt, với đối tượng hướng đến là các cá nhân hoặc doanh nghiệp cần phân tích phản hồi để nhận diện cảm xúc, chẳng hạn như chủ cửa hàng trực tuyến hoặc người mua sắm quan tâm đến chất lượng sản phẩm. Hệ thống được thiết kế không chỉ để hỗ trợ phân tích cảm xúc mà còn góp phần mở rộng tiềm năng ứng dụng trí tuệ nhân tạo (AI) trong xử lý ngôn ngữ tự nhiên cho các lĩnh vực khác.

Bên cạnh ý nghĩa khoa học, đề tài còn mang lại giá trị thực tiễn, giúp doanh nghiệp nâng cao hiệu quả kinh doanh thông qua việc hiểu rõ cảm xúc và nhu cầu của khách hàng. Hơn nữa, kết quả nghiên cứu có thể được ứng dụng làm nền tảng cho các

nghiên cứu tiếp theo, như phát triển chatbot thông minh hoặc hệ thống hỗ trợ khách hàng tự động.

Phương pháp nghiên cứu bao gồm các bước chính: thu thập dữ liệu từ các nguồn đáng tin cậy, tiền xử lý dữ liệu để chuẩn hóa và loại bỏ nhiễu, huấn luyện mô hình học sâu với bộ dữ liệu đã xử lý, đánh giá hiệu quả của từng mô hình, và cuối cùng là triển khai kết quả thông qua một giao diện web thân thiện với người dùng.

Thông qua việc kết hợp nghiên cứu lý thuyết và thực tiễn, đề tài không chỉ cung cấp một giải pháp toàn diện cho bài toán phân tích cảm xúc từ bình luận sản phẩm mà còn tạo tiền đề cho những ứng dụng công nghệ AI trong tương lai.

2 CƠ SỞ LÝ THUYẾT

2.1 Ý tưởng

Phân tích cảm xúc (Sentiment Analysis) là một lĩnh vực quan trọng trong xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP), nhằm xác định cảm xúc hoặc thái độ (tích cực, tiêu cực, trung tính) được biểu đạt trong văn bản. Đề tài "Phân tích cảm xúc comment trên một bài viết đánh giá sản phẩm" tập trung vào khai thác và áp dụng các mô hình/kỹ thuật học sâu, cụ thể là BERT và CNN-BiLSTM, để xây dựng các mô hình phân loại cảm xúc tự động từ dữ liệu bình luận trực tuyến và tích hợp vào website để mang lại trải nghiệm dễ sử dụng cho người dùng. Ý tưởng này được hình thành từ nhu cầu thực tế trong việc hiểu rõ hơn về phản hồi của khách hàng trên các nền tảng thương mại điện tử.

2.2 Cơ sở lý thuyết

2.2.1 Giới thiệu về BERT

a. Tổng quan về mô hình BERT

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình xử lý ngôn ngữ tự nhiên (NLP) tiên tiến được nhóm nghiên cứu của Google AI giới thiệu vào năm 2018. Đây là bước đột phá lớn trong lĩnh vực NLP nhờ khả năng học biểu diễn ngữ nghĩa hai chiều (bidirectional contextual representations) từ văn bản. BERT đã mở ra một kỷ nguyên mới cho NLP, tạo nền tảng cho nhiều mô hình hiện đại khác.

Mô hình BERT được thiết kế để giải quyết một hạn chế lớn của các mô hình trước đây: chỉ hiểu ngữ cảnh theo một hướng (từ trái sang phải hoặc phải sang trái). Với BERT, mô hình có thể học ngữ cảnh của từ dựa trên cả hai hướng (trước và sau). Điều này giúp BERT hiểu tốt hơn ý nghĩa của từ trong từng bối cảnh cụ thể.

BERT được tiền huấn luyện trên lượng dữ liệu rất lớn, bao gồm:

- BookCorpus: Hơn 800 triệu từ.
- Wikipedia tiếng Anh: Khoảng 2.5 tỷ từ.

b. Kiến trúc mô hình BERT

BERT sử dụng kiến trúc Transformer, được giới thiệu lần đầu trong bài báo "*Attention is All You Need*" vào năm 2017. Cụ thể, BERT sử dụng phần **Encoder** của Transformer để mã hóa văn bản.

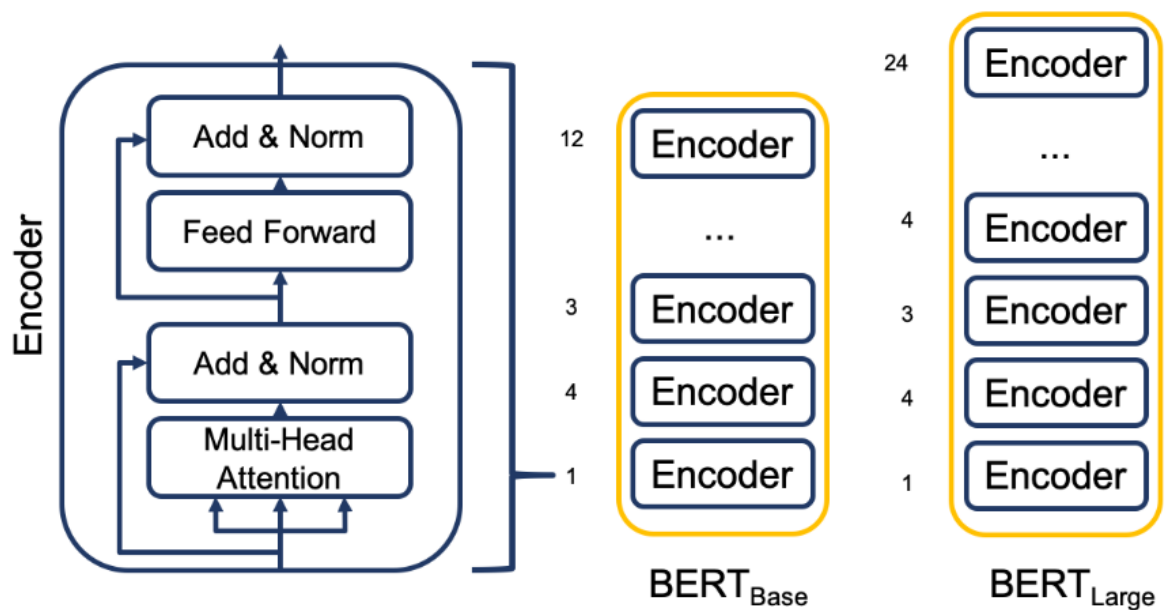
Cấu trúc chính: gồm 2 loại (**BERT-base** và **BERT-large**) với số lượng lớp Transformer Encoder tương ứng là 12 và 24 lớp

Dưới đây là một số thông số liên quan đến 2 loại mô hình:

	BERT Base	BERT Large
Layers	24	112
Hidden Size	768	1024
Heads	12	16
Parameters	110M	340M

Hình 1. Thông số model BERT

Kiến trúc mô hình BERT



Hình 2. Cấu trúc model BERT

Cấu trúc của Encoder Block (Transformer Encoder)

Một lớp **Encoder** bao gồm các thành phần chính sau:

Multi-Head Attention:

- Thành phần này cho phép mô hình tập trung vào các phần khác nhau của chuỗi đầu vào đồng thời. Nó sử dụng nhiều "head" để tạo ra các vector ngữ cảnh khác nhau cho mỗi token.
- **Self-Attention Mechanism:** Trong BERT, Self-Attention được sử dụng, nghĩa là mỗi token tính toán mối quan hệ của nó với tất cả các token khác trong chuỗi đầu vào.

Add & Norm (Residual Connection + Layer Normalization):

- **Residual Connection:** Giúp truyền thông tin từ đầu vào gốc tới đầu ra của mỗi lớp, tránh mất mát thông tin khi qua nhiều lớp.
- **Layer Normalization:** Chuẩn hóa đầu ra để tăng độ ổn định khi huấn luyện.

Feed Forward Neural Network (FFNN): Mỗi Encoder Block chứa một mạng nơ-ron truyền thẳng. Đây là một lớp phi tuyến tính dùng để học các biểu diễn phức tạp hơn từ đầu ra của Attention.

Add & Norm: Sau khi qua Feed Forward, đầu ra được chuẩn hóa một lần nữa với Residual Connection

c. Ưu điểm của mô hình BERT cho phân loại cảm xúc

- **Hiểu ngữ cảnh sâu sắc:** Nhờ vào cơ chế attention và kiến trúc transformer, BERT có khả năng hiểu mối quan hệ giữa các từ trong câu rất tốt, ngay cả khi chúng ở vị trí xa nhau.
- **Khả năng áp dụng đa ngữ cảnh:** BERT có thể sử dụng mô hình đã huấn luyện sẵn trên một ngôn ngữ để áp dụng cho nhiều tác vụ khác nhau mà không cần huấn luyện lại từ đầu.
- **Mô hình tiên tiến:** BERT đã thiết lập các chuẩn mực mới trong việc phân loại cảm xúc, với hiệu suất rất cao so với các phương pháp truyền thống.

d. Hạn chế của mô hình BERT

- **Tài nguyên tốn kém:** Huấn luyện và triển khai BERT đòi hỏi GPU/TPU mạnh mẽ do kích thước mô hình lớn.
- **Độ trễ cao:** BERT không phù hợp cho các ứng dụng thời gian thực (real-time) do mô hình nặng.

- Không hỗ trợ ngữ cảnh dài: BERT bị giới hạn độ dài chuỗi đầu vào (thường là 512 token).

e. Các phiên bản cải tiến của BERT

Nhiều phiên bản cải tiến của BERT đã được phát triển để khắc phục hạn chế hoặc nâng cao hiệu suất:

RoBERTa (Facebook): Tăng cường dữ liệu huấn luyện và loại bỏ NSP.

ALBERT (Google): Giảm kích thước mô hình bằng cách chia sẻ tham số.

DistilBERT : Phiên bản nhỏ gọn, nhanh hơn nhưng hiệu quả gần bằng BERT.

mBERT: Phiên bản đa ngôn ngữ, hỗ trợ nhiều ngôn ngữ trên thế giới.

2.2.2 Giới thiệu về CNN-BiLSTM

Trong lĩnh vực Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing - NLP) và Học Máy (Machine Learning), việc kết hợp các mô hình Convolutional Neural Network (CNN) và Bidirectional Long Short-Term Memory (BiLSTM) đã chứng minh được hiệu quả vượt trội trong nhiều ứng dụng, đặc biệt là trong việc phân tích cảm xúc từ dữ liệu văn bản. **CNN-BiLSTM** là sự kết hợp của hai kiến trúc mạng nơ-ron sâu này, tận dụng những ưu điểm riêng biệt của từng mô hình để cải thiện khả năng hiểu và xử lý ngôn ngữ tự nhiên một cách toàn diện hơn.

a. Tổng Quan về CNN-BiLSTM

- **CNN (Convolutional Neural Network)** là một loại mạng nơ-ron sâu được thiết kế ban đầu để xử lý dữ liệu có cấu trúc lưới, như hình ảnh. Tuy nhiên, CNN cũng được áp dụng rộng rãi trong NLP để trích xuất các đặc trưng cục bộ từ chuỗi từ trong văn bản. Các lớp convolutional trong CNN giúp mô hình nhận diện các mẫu đặc trưng như cụm từ hay biểu đạt cảm xúc trong các vùng khác nhau của câu.
- **BiLSTM (Bidirectional Long Short-Term Memory)** là một biến thể của mô hình LSTM, một loại mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) có khả năng học các phụ thuộc dài hạn trong dữ liệu chuỗi. BiLSTM xử lý dữ liệu theo cả hai hướng: từ trái sang phải (forward) và từ phải sang trái (backward), giúp mô hình nắm bắt được ngữ cảnh từ cả hai phía của mỗi từ trong câu.

Sự kết hợp của CNN và BiLSTM trong CNN-BiLSTM mang lại khả năng trích xuất đặc trưng mạnh mẽ từ văn bản và hiểu được ngữ cảnh sâu hơn, từ đó nâng cao hiệu quả trong các nhiệm vụ phân loại cảm xúc.

b. Cấu Trúc của CNN-BiLSTM

Mô hình CNN-BiLSTM thường được xây dựng với các thành phần chính như sau:

- **Embedding Layer (Lớp Nhúng):** Chuyển đổi các từ trong văn bản thành các vector số học, đại diện cho ý nghĩa ngữ nghĩa của từ. Các kỹ thuật phổ biến bao gồm Word2Vec, GloVe, hoặc các mô hình nhúng từ hiện đại như FastText và BERT.
- **Convolutional Layer (Lớp Chập):** Sử dụng các bộ lọc (filters) để quét qua các chuỗi từ trong văn bản, trích xuất các đặc trưng cục bộ như cụm từ hoặc các biểu hiện cảm xúc. Các bộ lọc này giúp mô hình nhận diện các mẫu đặc trưng quan trọng mà có thể ảnh hưởng đến cảm xúc của câu.
- **Pooling Layer (Lớp Gộp):** Giảm kích thước không gian của các đặc trưng đã trích xuất, giúp giảm số lượng tham số và tránh hiện tượng overfitting. Thường sử dụng kỹ thuật max pooling để giữ lại các đặc trưng quan trọng nhất.
- **Bidirectional LSTM Layer (Lớp BiLSTM):** Xử lý các đặc trưng đã được trích xuất bởi CNN theo cả hai hướng của chuỗi từ, giúp mô hình nắm bắt được ngữ cảnh từ cả phía trước và phía sau của mỗi từ trong câu. Điều này cải thiện khả năng hiểu ngữ cảnh và ý nghĩa tổng thể của câu.
- **Fully Connected Layer (Lớp Kết Nối Đầy Đủ):** Kết nối các đặc trưng từ BiLSTM để tạo ra các đầu ra cuối cùng cho nhiệm vụ phân loại. Lớp này thường sử dụng các kỹ thuật như dropout để giảm thiểu overfitting.
- **Output Layer (Lớp Đầu Ra):** Sử dụng hàm kích hoạt như softmax để phân loại cảm xúc vào các lớp tương ứng (tích cực, tiêu cực, trung lập).

c. Ưu Điểm của CNN-BiLSTM trong Phân Tích Cảm Xúc

- **Khả Năng Trích Xuất Đặc Trưng Mạnh Mẽ:** CNN có khả năng nhận diện các mẫu đặc trưng phức tạp trong văn bản, giúp mô hình hiểu rõ hơn về các biểu hiện cảm xúc trong ngôn ngữ tự nhiên.
- **Nắm Bắt Ngữ Cảnh Hiệu Quả:** BiLSTM giúp mô hình hiểu được ngữ cảnh rộng hơn bằng cách xử lý chuỗi từ theo cả hai hướng, từ đó cải thiện độ chính xác trong việc phân loại cảm xúc.
- **Khả Năng Xử Lý Dữ Liệu Phi Cấu Trúc:** CNN-BiLSTM có thể xử lý tốt các dữ liệu văn bản không cấu trúc, như các đánh giá sản phẩm, nơi mà cảm xúc có thể được biểu đạt qua nhiều cách khác nhau.
- **Linh Hoạt và Mở Rộng:** Mô hình có thể dễ dàng được mở rộng và điều chỉnh để phù hợp với các ngôn ngữ khác nhau, bao gồm cả tiếng Việt, thông qua việc tùy chỉnh các thành phần của mạng.

d. Ứng Dụng của CNN-BiLSTM trong Phân Tích Cảm Xúc

Trong bối cảnh phân tích cảm xúc trên các đánh giá sản phẩm tiếng Việt từ các nền tảng mua sắm trực tuyến như Lazada và Tiki, CNN-BiLSTM được áp dụng để:

- Trích Xuất Đặc Trưng Cảm Xúc: Nhận diện các từ khóa và cụm từ mang tính cảm xúc trong đánh giá, như "tốt", "kém", "yêu thích", "không hài lòng", v.v.
- Hiểu Ngữ Cảnh và Ý Nghĩa: Xác định được cảm xúc chính xác dựa trên ngữ cảnh của câu, giúp phân loại đánh giá một cách chính xác hơn.
- Cải Thiện Độ Chính Xác: Kết hợp CNN và BiLSTM giúp mô hình đạt được độ chính xác cao hơn so với việc sử dụng từng mô hình riêng lẻ, nhờ vào khả năng trích xuất đặc trưng và hiểu ngữ cảnh sâu hơn.
- Xử Lý Các Biểu Đạt Cảm Xúc Phức Tạp: Đối với các đánh giá có cấu trúc phức tạp hoặc chứa nhiều cảm xúc, CNN-BiLSTM có khả năng phân tích và phân loại chính xác hơn nhờ vào sự kết hợp giữa trích xuất đặc trưng và xử lý ngữ cảnh.

e. Thách Thức Khi Áp Dụng CNN-BiLSTM cho Tiếng Việt

- Đặc Thù Ngôn Ngữ: Tiếng Việt có cấu trúc ngữ pháp và từ vựng phức tạp, bao gồm việc sử dụng dấu và từ ghép, đòi hỏi mô hình phải được tối ưu hóa để xử lý đúng các đặc điểm này. Việc này cần có sự điều chỉnh kỹ lưỡng trong quá trình tiền xử lý dữ liệu và thiết kế mô hình.
- Thiếu Dữ Liệu Đánh Nhãn: Việc thu thập và gán nhãn dữ liệu cảm xúc cho tiếng Việt có thể gặp khó khăn, hạn chế hiệu quả của việc huấn luyện mô hình. Dữ liệu đánh giá cần phải được thu thập và xử lý một cách cẩn thận để đảm bảo chất lượng và tính đại diện.
- Tính Hiệu Quả và Tốc Độ Xử Lý: CNN-BiLSTM có thể đòi hỏi tài nguyên tính toán lớn, ảnh hưởng đến tốc độ xử lý khi áp dụng cho khối lượng dữ liệu lớn từ các nền tảng như Lazada và Tiki. Điều này yêu cầu các giải pháp tối ưu hóa mô hình và sử dụng các phần cứng mạnh mẽ để đảm bảo hiệu suất.

f. Kết Luận

CNN-BiLSTM là một kiến trúc mạnh mẽ kết hợp giữa khả năng trích xuất đặc trưng của CNN và khả năng nắm bắt ngữ cảnh của BiLSTM, phù hợp với nhiệm vụ phân tích cảm xúc từ dữ liệu văn bản tiếng Việt. Việc áp dụng mô hình này vào phân tích cảm xúc trên các đánh giá sản phẩm từ Lazada và Tiki không chỉ giúp tự động hóa quá trình phân loại cảm xúc mà còn nâng cao độ chính xác và hiệu quả của hệ thống. Tuy nhiên,

để đạt được kết quả tối ưu, cần chú trọng vào việc tối ưu hóa mô hình cho đặc thù ngôn ngữ tiếng Việt và xử lý các thách thức liên quan đến dữ liệu và tính toán.

Việc triển khai **CNN-BiLSTM** trong dự án phân tích cảm xúc sẽ đóng góp đáng kể vào việc cải thiện chất lượng dịch vụ của các nền tảng thương mại điện tử, đồng thời mở rộng kiến thức và ứng dụng các kỹ thuật học sâu trong lĩnh vực NLP cho ngôn ngữ tiếng Việt.

3 THU THẬP VÀ XỬ LÝ DỮ LIỆU

3.1 Thu thập dữ liệu

3.1.1 Nguồn dữ liệu

- Nguồn dữ liệu cho dự án được thu thập từ hai nền tảng thương mại điện tử phổ biến tại Việt Nam, cụ thể là **Lazada** và **Tiki**. Đây là hai trang web được lựa chọn vì có cơ sở người dùng lớn, hoạt động sôi động, và chứa lượng lớn bình luận khách hàng đa dạng về sản phẩm.

a. Đặc điểm nguồn dữ liệu:

- Lazada:
 - Lazada là một trong những nền tảng thương mại điện tử hàng đầu tại Đông Nam Á, cung cấp một loạt sản phẩm từ các ngành hàng khác nhau như điện tử, thời trang, đồ gia dụng, và mỹ phẩm. Hệ thống đánh giá sản phẩm của Lazada cho phép khách hàng để lại các bình luận, kèm theo xếp hạng sao, hình ảnh, và thậm chí video minh họa. Những bình luận này thường phản ánh chính xác trải nghiệm thực tế của người dùng, bao gồm cả ý kiến tích cực và tiêu cực.
- Tiki:
 - Tiki là một nền tảng thương mại điện tử nổi tiếng tại Việt Nam, được biết đến với hệ thống bán hàng uy tín và chính sách chăm sóc khách hàng tốt. Dữ liệu từ Tiki thường bao gồm các bình luận chi tiết, với nhiều khách hàng chia sẻ cả ưu và nhược điểm của sản phẩm. Nhiều bình luận được bổ sung thêm thông tin cụ thể về sản phẩm như chất lượng, dịch vụ giao hàng, hoặc hỗ trợ sau mua.

b. Lý do chọn nguồn dữ liệu:

- Tính đa dạng: Cả Lazada và Tiki đều có một danh mục sản phẩm đa dạng và phong phú, từ hàng tiêu dùng đến các mặt hàng công nghệ cao cấp. Điều này tạo cơ hội thu thập dữ liệu từ nhiều lĩnh vực khác nhau, giúp mở rộng phạm vi ứng dụng của chương trình.
- Khối lượng bình luận lớn: Hai trang thương mại điện tử này thu hút hàng triệu người dùng mỗi tháng, đồng nghĩa với việc lượng bình luận về sản phẩm rất phong phú. Dữ liệu lớn này cung cấp cái nhìn toàn diện về trải nghiệm của người tiêu dùng.
- Tính thực tiễn: Các bình luận từ Lazada và Tiki được viết bởi những người dùng thực tế, đảm bảo rằng dữ liệu phản ánh đúng cảm xúc và suy nghĩ của người tiêu dùng về sản phẩm. Điều này rất quan trọng trong bài toán phân tích cảm xúc, nơi cần dữ liệu thực để mô hình hóa cảm xúc chính xác.

c. Hạn chế và giải pháp:

- Hạn chế:
 - Các bình luận có thể bị thiên lệch, khi một số khách hàng thường chỉ để lại nhận xét trong các trường hợp cực kỳ tích cực hoặc tiêu cực.
 - Dữ liệu có thể bị hạn chế bởi cơ chế bảo mật hoặc cấu trúc HTML thay đổi của các trang web.
- Giải pháp:
 - Thu thập dữ liệu từ nhiều sản phẩm khác nhau để giảm thiểu tác động của thiên lệch.
 - Sử dụng các kỹ thuật tự động hóa linh hoạt, có khả năng điều chỉnh để phù hợp với các thay đổi cấu trúc trang web...

d. Đóng góp của nguồn dữ liệu:

- Dữ liệu từ Lazada và Tiki không chỉ cung cấp nội dung phong phú để phân tích cảm xúc, mà còn tạo ra cơ sở thực tế để kiểm tra tính chính xác của mô hình. Sự đa dạng trong các ngành hàng và loại sản phẩm giúp chương trình có khả năng áp dụng rộng rãi và phù hợp với nhiều tình huống sử dụng khác nhau.

3.1.2 Phương pháp thu thập

a. Công cụ và công nghệ sử dụng:

- **Ngôn ngữ lập trình:** Python
- **Thư viện hỗ trợ:**
 - selenium: Sử dụng để tự động hóa trình duyệt và tương tác với các phần tử trên trang web, hỗ trợ cuộn trang, nhấp chuột và trích xuất dữ liệu.
 - pandas: Dùng để tổ chức dữ liệu thu thập, xử lý dữ liệu thô và lưu trữ dữ liệu ở định dạng CSV.
 - datetime: Được sử dụng để tạo dấu thời gian (timestamp) nhằm đảm bảo tên tệp CSV là duy nhất.
 - random: Tích hợp độ trễ ngẫu nhiên trong các thao tác như chờ tải trang và nhấp chuột để tránh bị phát hiện bởi hệ thống chống tự động hóa.
- **Trình duyệt:** Google Chrome với giao diện hiện đại và khả năng tương thích cao với Selenium
- **Driver:** ChromeDriver, phiên bản tương ứng với trình duyệt, được sử dụng để kết nối Selenium với trình duyệt.

b. Quy trình thực hiện:

- **Chuẩn bị và thiết lập:**
 - + **Kiểm tra URL:** URL được kiểm tra xem có hợp lệ và có thể truy cập được hay không.

+ **Cấu hình trình duyệt:**

- Vô hiệu hóa các tính năng tự động phát hiện hoạt động tự động hóa của trình duyệt.
- Cải thiện hiệu suất khi thu thập dữ liệu như mở trình duyệt ở chế độ toàn màn hình (maximized) và tùy chỉnh user-agent để mô phỏng trình duyệt thực.

+ **Xử lý lỗi khi tải trang:**

- Script kiểm tra nếu URL không tải được hoặc tải sai (ví dụ: hiển thị data:;) sẽ tự động làm mới trang.

- **Thu thập bình luận:**

+ Truy cập trang web:

- Truy cập URL sản phẩm và đảm bảo rằng trang đã tải hoàn chỉnh trước khi bắt đầu trích xuất dữ liệu.
- Sử dụng phương pháp chờ đợi có điều kiện (WebDriverWait) để đảm bảo các phần tử cần thiết đã xuất hiện.

+ Tự động cuộn trang và nhấp nút "Next":

- Trang được tự động cuộn xuống cuối để tải thêm các bình luận.
- Nếu nút "Next" tồn tại, script sẽ nhấp vào nút để chuyển sang trang tiếp theo. Tương tự, thời gian chờ và thao tác được thực hiện ngẫu nhiên để giảm nguy cơ bị phát hiện là bot.

+ Trích xuất dữ liệu bình luận:

- Các bình luận được nhận diện thông qua lớp CSS (CLASS_NAME) hoặc cấu trúc HTML đặc thù của trang web.

- **Lưu trữ dữ liệu:**

- + Lưu dữ liệu vào một tệp CSV, được đặt tên duy nhất theo thời gian thu thập, đảm bảo khả năng truy xuất.

c. Kết quả thu thập dữ liệu lúc đầu:

- Kết quả thu thập là một tệp CSV chứa các thông tin:
 - + Ngày bình luận
 - + Tên người dùng
 - + Nội dung bình luận
 - + Tên sản phẩm

- + Số lượt thích của bình luận.
- File CSV được lưu với tên định dạng: comments_YYYYMMDD_HHMMSS.csv.

	A	B	C	D	E	F
1	Date	Name	Comment	Product	Liked	Label
2	4 tuần trước	Choi K.Chú	Quá hợp với nhu cầu sử dụng	Variation1:	2	Positive
3	08 thg 7 20	P***.Chú	Dễ sử dụng, chất lượng cao,	Màu sắc:B	0	Positive
4	23 thg 6 20	Thanh N.C	nhỏ gọn rẻ tiền dễ sử dụng Thiết kế s	Màu sắc:B	2	Positive
5	4 tuần trước	Văn H.Chú	Bền và lâu dài, Giá trị tuyệt vời cho s	Màu sắc:B	2	Positive
6	04 thg 6 20	Đinh N.Chú	Sản phẩm chất lượng tốt, chắc chắn	Màu sắc:B	0	Positive
7	1 ngày trước	1***nChú	Bấm móng tay bằng thép không gỉ, B	Màu sắc:B	0	Positive
8	27 thg 11 2	Minh T.Chú	sử dụng tốt, rất tiện lợi.	Màu sắc:B	0	Positive
9	14 thg 8 20	Bang L.Chú	Bộ di động tiện lợi, nhỏ gọn	Màu sắc:B	0	Positive
10	30 thg 3 20	Hoang V.C	giao đúng hàng sử dụng được	Màu sắc:B	1	Positive
11	05 thg 1 20	Ngô Đ.Chú	Sản phẩm hoàn thiện tốt - sử dụng t	Màu sắc:B	1	Positive
12	27 thg 12 2	8***9Chú	Kết cấu: đẹp	Màu sắc:B	1	Positive
13	18 thg 8 20	Trần H.Chú	Dễ sử dụng, Thiết kế mượt mà và ph	Màu sắc:B	2	Positive
14	15 thg 12 2	Phi T.Chú	Hàng đẹp đúng mẫu	Đáng tiền:	2	Positive
15	11 thg 8 20	9***5Chú	Giá trị tuyệt vời cho số tiền bỏ ra, Bộ	Màu sắc:B	0	Positive
16	17 thg 3 20	Nguyễn V.C	hàng giao nhanh, đúng mô tả, giá tốt	Màu sắc:B	0	Positive
17	18 thg 8 20	Trần V.Chú	Dễ sử dụng, chất lượng đúng với số t	Màu sắc:B	0	Positive
18	19 thg 2 20	Y***jChú	Chất lượng: tốt	Màu sắc:B	0	Positive
19	25 thg 3 20	Chuong M.	Dễ sử dụng	Màu sắc:B	0	Positive
20	17 thg 3 20	k***.Chú	cửa hàng tư vấn nhiệt tình	Màu sắc:B	0	Positive
21	31 thg 12 2	Ngô Q.Chú	Đã nhận sản phẩm giao đúng đơn đ	Màu sắc:B	0	Positive

Hình 3. Hình ảnh thực tế dữ liệu crawl

3.2 Phân tích dữ liệu

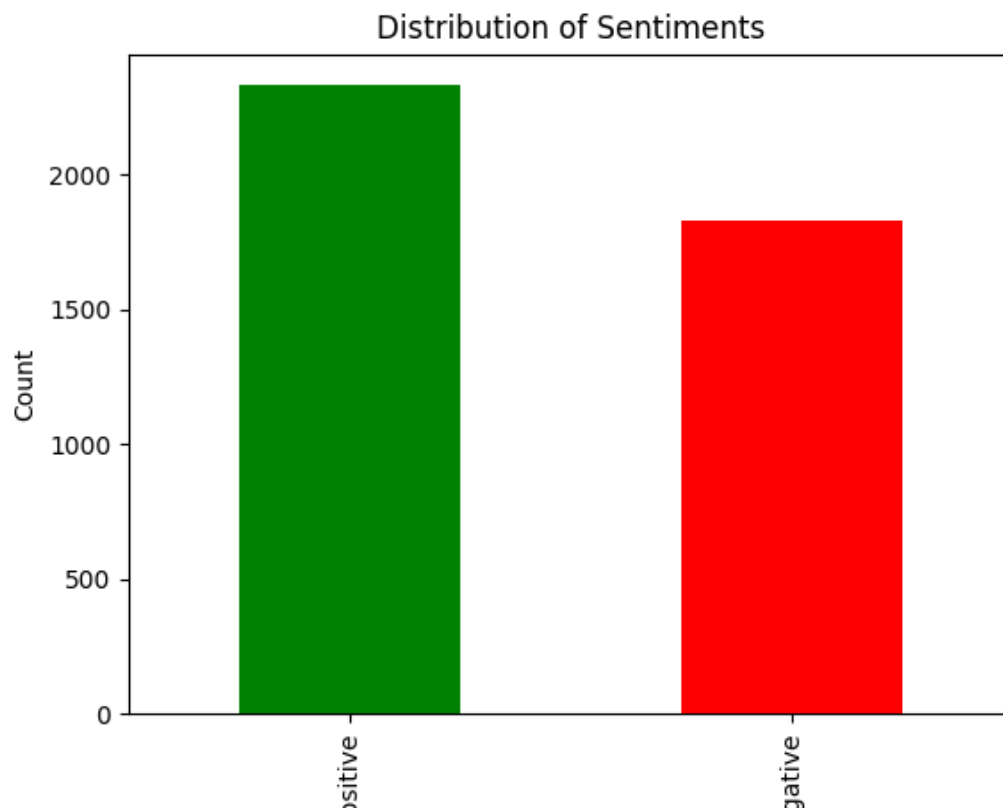
3.2.1 Thống kê mô tả

- Kết quả thu thập dữ liệu (sau khi lọc review):

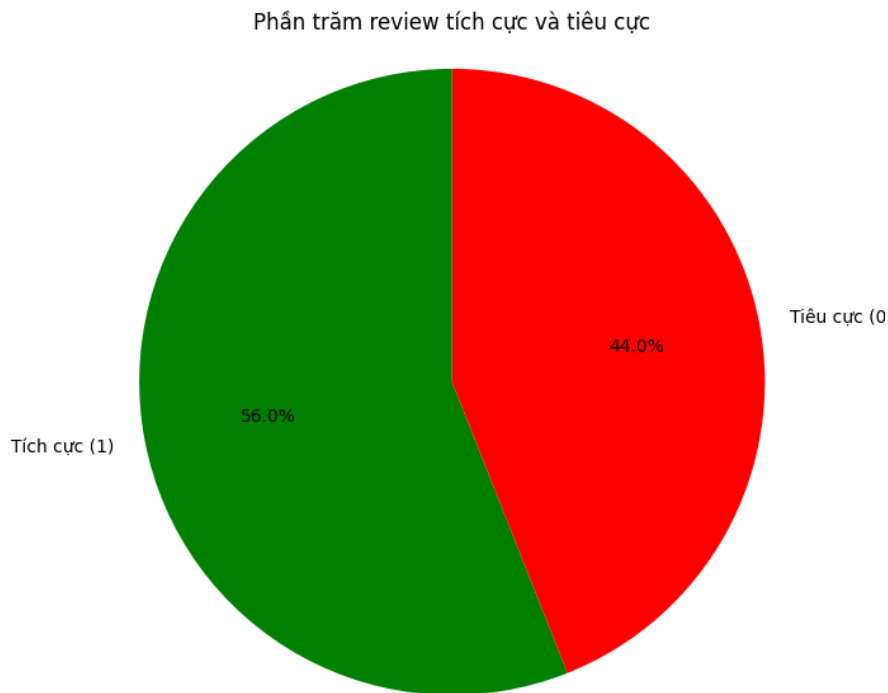
	Số lượng	Phần trăm (%)
1	2333.0	56.014
0	1832.0	43.986
Tổng	4165.0	100.0

Hình 4. Kết quả thu thập dữ liệu

- Trực quan hóa dữ liệu:



Hình 5. Phân bố nhãn cảm xúc



Hình 6. Biểu đồ phân bố nhãn theo phần trăm

- Từ hai biểu đồ trên, có thể rút ra được kết luận sau đây về dữ liệu:

Tập dữ liệu thu thập từ **Lazada** và **Tiki** cho thấy sự phân bố tương đối cân bằng giữa hai nhãn cảm xúc chính: **tích cực (56%)** và **tiêu cực (44%)**. Mặc dù tỷ lệ bình luận tiêu cực nhỉnh hơn một chút so với bình luận tích cực, nhưng sự chênh lệch không đáng kể này lại mang đến nhiều lợi ích quan trọng trong quá trình huấn luyện mô hình học máy.

- Phân bố dữ liệu:

Sự phân bố gần cân bằng giữa hai nhãn cảm xúc giúp dữ liệu trở nên phong phú và đa dạng, đồng thời hạn chế hiện tượng mất cân bằng dữ liệu thường thấy trong các bài toán phân loại. Điều này tạo điều kiện lý tưởng để các mô hình học máy như **BERT** và **CNN-BiLSTM** có thể khai thác đầy đủ đặc trưng của cả hai nhãn, từ đó học được mối quan hệ ngữ nghĩa và cảm xúc một cách sâu sắc, chính xác hơn.

- **Dữ liệu tích cực (56%):** Các bình luận này thường chứa các từ ngữ mang tính khẳng định, hài lòng, hoặc khen ngợi sản phẩm và dịch vụ, cung cấp thông tin về các khía cạnh mà khách hàng đánh giá cao.

- **Dữ liệu tiêu cực (44%):** Phản ánh các vấn đề hoặc điểm yếu mà khách hàng gặp phải, từ đó mang lại cái nhìn toàn diện hơn về trải nghiệm thực tế của người tiêu dùng.

Sự hiện diện của cả hai loại cảm xúc với tỷ lệ gần cân bằng đảm bảo rằng mô hình sẽ không bị "thiên lệch", tức là quá chú trọng vào một nhãn cảm xúc trong khi bỏ qua hoặc đánh giá thấp nhãn còn lại.

- Vai trò của dữ liệu trong mô hình học máy:

- **BERT:**

Với cơ chế **Attention (Chú ý)** đặc trưng, BERT tận dụng tốt nguồn dữ liệu đa dạng để nhận diện các đặc trưng ngữ nghĩa quan trọng trong cả hai nhãn cảm xúc. Điều này giúp mô hình xử lý tốt các bình luận phức tạp, có ngữ cảnh hoặc chứa nhiều tầng nghĩa, chẳng hạn như bình luận có cả ý kiến tích cực và tiêu cực. Nhờ dữ liệu phong phú, BERT có khả năng học sâu hơn về mối quan hệ ngữ nghĩa, đảm bảo độ chính xác cao khi phân tích cảm xúc.

- **CNN-BiLSTM:**

CNN-BiLSTM phát huy ưu thế vượt trội trong việc xử lý các chuỗi dữ liệu liên tiếp. Với tập dữ liệu hiện tại, CNN-BiLSTM tận dụng cấu trúc câu và mối liên hệ ngữ pháp để đưa ra dự đoán chính xác, ngay cả trong các bình luận dài, chứa nhiều ý kiến liên kết. Khả năng học thứ tự và mối quan hệ chuỗi của CNN-BiLSTM càng được tối ưu hóa nhờ sự cân bằng trong phân phối dữ liệu.

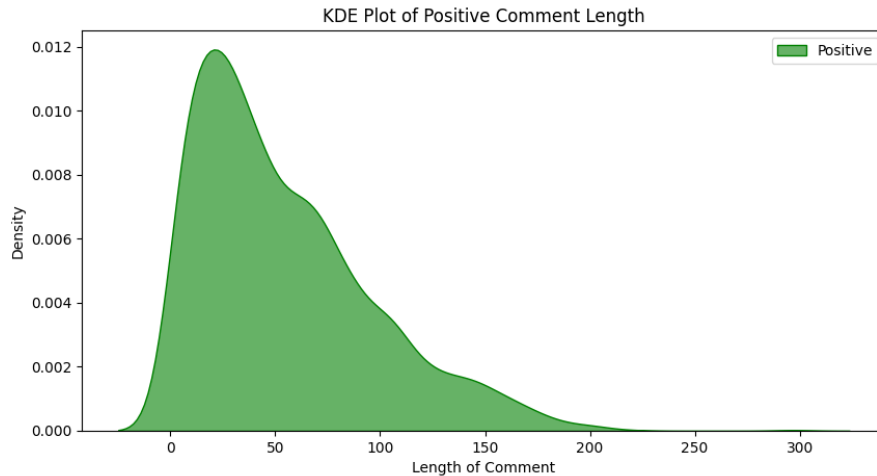
- Lợi ích của việc cân đối dữ liệu:

- **Hạn chế kỹ thuật phức tạp:** Do sự cân bằng tương đối giữa hai nhãn, bài toán không yêu cầu sử dụng các kỹ thuật xử lý mất cân bằng dữ liệu như **oversampling**, **undersampling**, hoặc **weighted loss functions**. Điều này giúp tiết kiệm tài nguyên xử lý, giảm thời gian phát triển và huấn luyện mô hình.

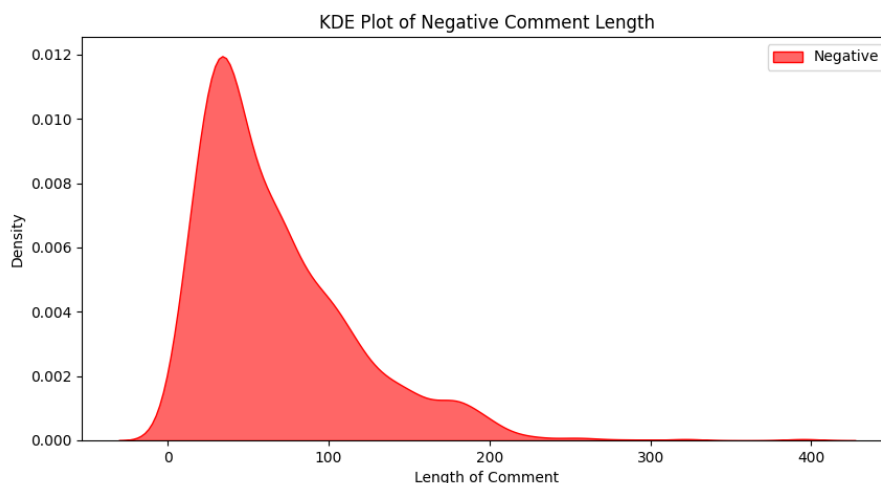
- **Hiệu suất cao:** Sự phong phú và cân đối dữ liệu cho phép mô hình đạt hiệu suất cao trong cả hai nhiệm vụ: phân tích cảm xúc chính xác và tổng quát hóa trên các tập dữ liệu thực tế. Các kết quả kiểm tra cho thấy mô hình có độ chính xác dao động từ **85% đến 90%**, phản ánh rõ ràng hiệu quả của việc sử dụng tập dữ liệu cân đối và chất lượng.

3.2.2 Khám phá dữ liệu

- Biểu đồ về độ dài bình luận:



Hình 7. Biểu đồ KDE độ dài bình luận tích cực



Hình 8. Biểu đồ KDE độ dài bình luận tiêu cực

- Sơ lược về biểu đồ hai KDE trên:

Biểu đồ KDE (**Kernel Density Estimation**) là một công cụ thống kê phổ biến, được sử dụng để ước lượng hàm mật độ xác suất (**PDF - Probability Density Function**) của một biến ngẫu nhiên dựa trên dữ liệu quan sát. Không giống như biểu đồ tần suất (histogram) yêu cầu chia dữ liệu thành các khoảng cố định (bins), KDE cho phép hiển thị một cách trực quan và mượt mà hơn phân phối liên tục của dữ liệu mà không cần giả định trước về dạng phân phối cụ thể, chẳng hạn như phân phối chuẩn. Điều này làm cho KDE trở thành một phương pháp mạnh mẽ để hiểu rõ hơn về các đặc trưng tiềm ẩn trong dữ liệu.

- **Biểu đồ Tích cực:**

Biểu đồ này cho thấy phần lớn các bình luận tích cực tập trung trong khoảng độ dài từ **20 đến 100 từ**. Đây là vùng có mật độ cao nhất, phản ánh rằng người dùng có xu hướng sử dụng số lượng từ vừa phải để bày tỏ sự hài lòng hoặc khen ngợi về sản phẩm và dịch vụ.

Sau khoảng 100 từ, mật độ giảm dần, cho thấy rằng các bình luận tích cực dài hơn là ít phổ biến hơn. Điều này có thể liên quan đến thực tế rằng khách hàng thường dùng các bình luận ngắn gọn để khen ngợi sản phẩm, thay vì diễn giải chi tiết.

- **Biểu đồ Tiêu cực:**

Tương tự bình luận tích cực, bình luận tiêu cực cũng có vùng tập trung mật độ cao nhất trong khoảng **20 đến 100 từ**. Tuy nhiên, một điểm khác biệt quan trọng là biểu đồ này có một phần **đuôi dài hơn** (long tail), kéo dài đến các bình luận có độ dài lớn hơn.

Sự hiện diện của phần đuôi dài cho thấy rằng khi người dùng không hài lòng với sản phẩm hoặc dịch vụ, họ thường dành thời gian để mô tả chi tiết vấn đề mà họ gặp phải, từ đó tạo ra các bình luận tiêu cực có độ dài lớn hơn. Điều này phản ánh tâm lý chung rằng người dùng có xu hướng diễn giải nhiều hơn khi họ cảm thấy thất vọng hoặc không hài lòng.

- Mặc dù thông tin về độ dài bình luận không được sử dụng trực tiếp trong việc huấn luyện các mô hình phân tích cảm xúc như **LSTM** và **BERT**, nhưng việc hiểu rõ về phân bố độ dài bình luận mang lại nhiều lợi ích quan trọng:

Hiểu đặc điểm của dữ liệu: Việc nhận biết sự khác biệt về độ dài giữa bình luận tích cực và tiêu cực giúp nhóm dự án hiểu sâu hơn về đặc điểm của dữ liệu thu thập được. Thông tin này có thể được sử dụng để điều chỉnh các bước tiền xử lý dữ liệu (data preprocessing), chẳng hạn như cắt ngắn hoặc lọc bỏ các bình luận quá ngắn hoặc quá dài, để cải thiện hiệu quả của mô hình.

Gợi ý chiến lược xử lý dữ liệu: Sự khác biệt về phân bố độ dài có thể dẫn đến các chiến lược khác nhau khi xây dựng các tập huấn luyện hoặc đánh giá mô hình. Ví dụ, nhóm có thể cân nhắc áp dụng các kỹ thuật bổ sung để xử lý các bình luận dài hơn nhằm đảm bảo rằng mô hình học được các đặc trưng đầy đủ từ những dữ liệu này.

Ứng dụng trong phân tích sâu hơn: Việc nắm bắt xu hướng độ dài bình luận cũng mở ra cơ hội cho các nghiên cứu bổ sung, chẳng hạn như phân tích mối

quan hệ giữa độ dài bình luận và cường độ cảm xúc, hoặc khám phá các đặc trưng ngôn ngữ khác biệt trong các bình luận dài và ngắn.

- Biểu đồ wordcloud:

Word Cloud cho label 0



Hình 9. Word cloud cho bình luận tiêu cực (label 0)

Word Cloud cho label 1



Hình 10. Word cloud cho bình luận tích cực (label 1)

- Phân tích hai biểu đồ word cloud trên:

▪ Word Cloud cho Label 0 (Bình luận tiêu cực)

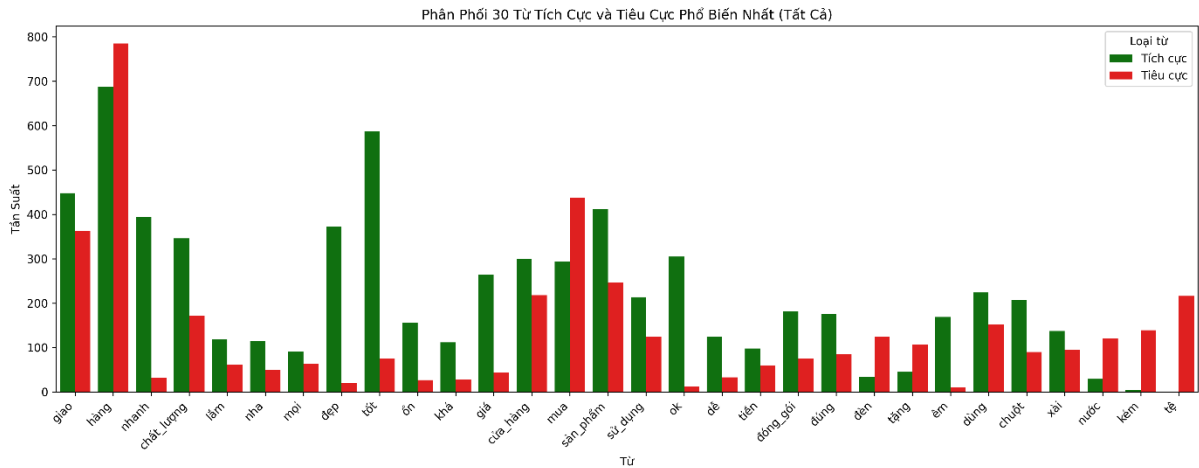
- Từ khóa nổi bật:
 - Các từ khóa như *"mua", "hàng", "sản phẩm", "chất lượng", "tệ", "thất vọng", "quảng cáo"* xuất hiện với tần suất cao.
 - Nhiều từ mang sắc thái tiêu cực như *"lỗi", "hư", "kém", "xấu", "đắt"* được thể hiện rõ.
- Ý nghĩa:
 - Bình luận tiêu cực tập trung vào những trải nghiệm không tốt liên quan đến chất lượng sản phẩm (*"chất lượng kém", "hư"*) hoặc dịch vụ giao hàng (*"giao hàng lâu"*) và sự không đúng kỳ vọng (*"quảng cáo sai sự thật", "thất vọng"*).
 - Điều này phản ánh các vấn đề cụ thể mà khách hàng không hài lòng.

▪ Word Cloud cho Label 1 (Bình luận tích cực)

- Từ khóa nổi bật:
 - Các từ như *"đẹp", "nhanh", "giao hàng", "chất lượng", "ok", "sản phẩm tốt"* xuất hiện với mật độ cao.
 - Ngoài ra, các từ khen ngợi như *"tuyệt vời", "phù hợp", "êm", "giá tốt"* thường xuyên xuất hiện.
- Ý nghĩa:
 - Bình luận tích cực tập trung vào sự hài lòng với chất lượng sản phẩm (*"đẹp", "chất lượng tốt"*) và dịch vụ nhanh chóng (*"giao hàng nhanh", "phục vụ tốt"*).

▪ Ảnh hưởng đến bài toán NLP: Word Cloud giúp xác định các từ khóa phổ biến và sự khác biệt ngữ nghĩa giữa bình luận tích cực và tiêu cực, hỗ trợ trong việc xây dựng tập từ vựng và tối ưu hóa vector từ (embedding) cho mô hình AI. Các từ cảm xúc rõ ràng như *"thất vọng", "tuyệt vời"* có thể cải thiện khả năng phân tách nhãn, trong khi các từ trung lập như *"mua", "sản phẩm"* cần được xử lý để giảm nhiễu. Trực quan hóa này còn định hướng chiến lược huấn luyện, như ưu tiên trọng số cho từ đặc trưng và tối ưu hóa độ dài chuỗi đầu vào cho mô hình BERT hoặc LSTM.

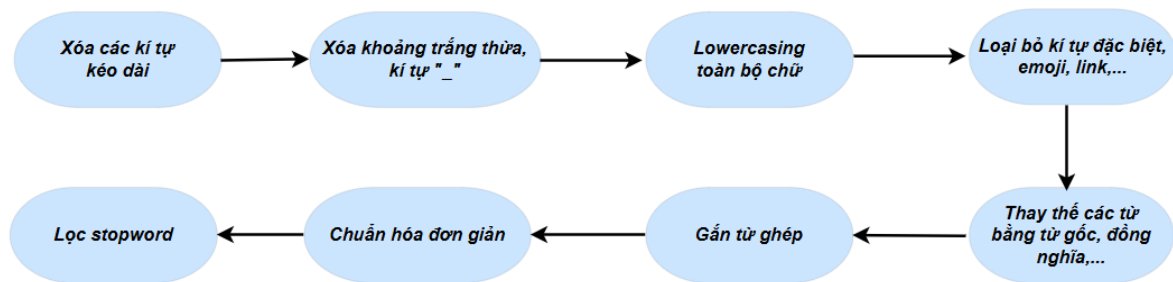
- Biểu đồ liên quan khác về dữ liệu thu thập:



Hình 11. Biểu đồ phân phối 30 từ tích cực và tiêu cực phổ biến

Biểu đồ trên thể hiện **phân phối 30 từ tích cực và tiêu cực phổ biến nhất** trong tập dữ liệu bình luận đánh giá sản phẩm. Các từ tích cực (cột màu xanh) như "đẹp," "tốt," "chất_lượng" xuất hiện với tần suất cao, phản ánh sự hài lòng của khách hàng về sản phẩm. Trong khi đó, các từ tiêu cực (cột màu đỏ) như "lỗi," "kém," "không" thường được sử dụng để diễn đạt sự không hài lòng.

3.3 Tiền xử lý dữ liệu



Hình 12. Quy trình tiền xử lý dữ liệu

Trước khi sử dụng cho quá trình huấn luyện model, dữ liệu được tiến hành tiền xử lí lần lượt qua các bước sau:

- Xóa các kí tự kéo dài (repeated characters): nhằm loại bỏ các kí tự lặp lại, chẳng hạn như hayyy -> hay,...
- Loại bỏ những khoảng trắng thừa đầu và cuối câu.
- Chuyển toàn bộ chữ thành chữ thường (lowercasing).
- Loại bỏ các kí tự đặc biệt, emoji, mã liên kết, link, ...
- Thay thế các từ trong **replace_list** để đưa các từ, cụm từ về từ gốc, từ đồng nghĩa hoặc các từ thay thế đã định trước.

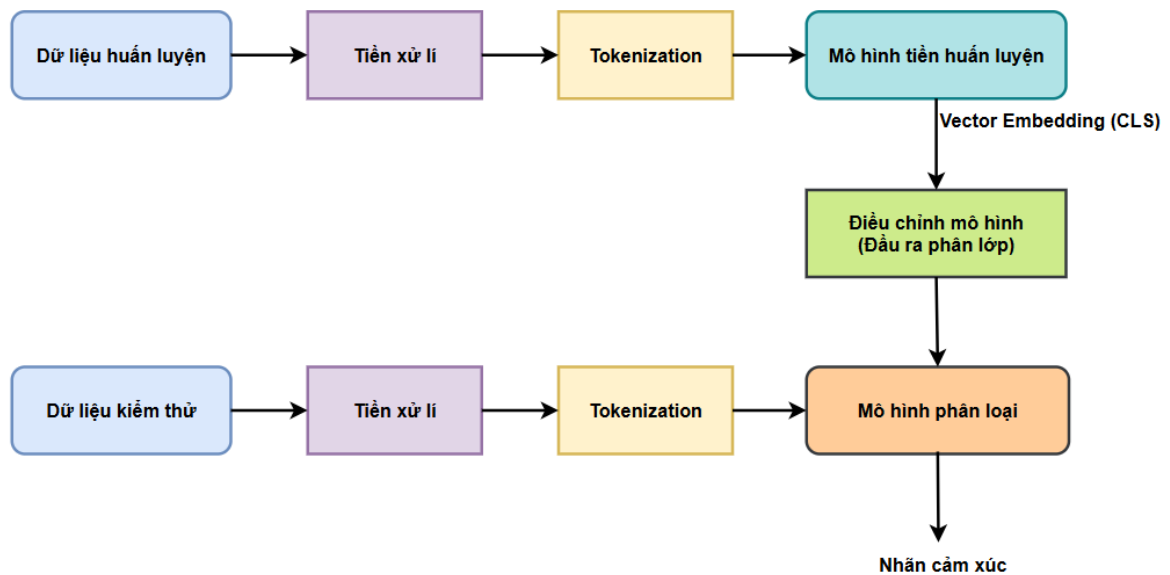
- Loại bỏ các kí tự “_”.
- Gắn từ ghép: tách hoặc gắn các từ lại dựa trên tokenizer của VnCoreNLP
- Chuẩn hóa đơn giản (simple_preprocess): Loại bỏ các ký tự không mong muốn và chuẩn hóa văn bản (sử dụng hàm simple_preprocess của module thư viện gensim.utils).
- Lọc các stopwords: loại bỏ các từ không ảnh hưởng đến ý nghĩa ngữ nghĩa trong ngữ cảnh của một câu.

Sau đó dữ liệu được phân chia thành các tập train/test/validation với tỉ lệ 8/1/1.

4 CÁC THUẬT TOÁN TRIỂN KHAI

4.1. Tổng quan về hai thuật toán

a. Sơ đồ thuật toán sử dụng mô hình BERT



Hình 13. Sơ đồ tổng quát sử dụng thuật toán mô hình BERT

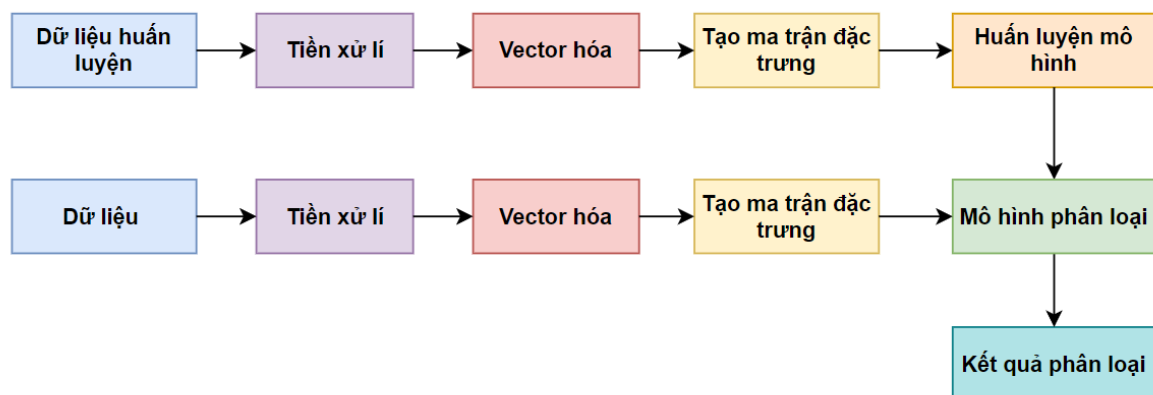
– Mô tả:

- **Dữ liệu huấn luyện:** Tập dữ liệu huấn luyện bao gồm các câu văn bản đã được gắn nhãn cảm xúc (tích cực/tiêu cực) được sử dụng để huấn luyện mô hình.
- **Tiền xử lý:** Xóa kí tự đặc biệt, biểu tượng cảm xúc, số, stopword, tách từ đơn, từ ghép, viết thường,
- **Tokenization (Tách từ):** Văn bản sau khi tiền xử lý được chia nhỏ thành các token bằng tokenizer của mô hình BERT. Các token này sau đó được ánh xạ thành **token IDs**.
- **Mô hình tiền huấn luyện (BERT-BASE):** Token IDs được đưa vào mô hình BERT đã tiền huấn luyện để trích xuất vector biểu diễn ngữ nghĩa thông qua:
 - **Embedding Layer:** Token IDs được chuyển đổi thành vector đầu vào bằng cách tổng hợp **Token Embeddings (TE)**, **Segment Embeddings (SE)**, và **Position Embeddings (PE)**. Kết quả là mỗi token được biểu diễn dưới dạng một vector 768 chiều (Hidden Size của BERT-BASE).
 - **Lớp Encoder:** Gồm 12 lớp encoder, mỗi lớp có **Multi-Head Self-Attention** (12 đầu attention song song) giúp học mối quan hệ giữa các token trong câu,

và **FFNN (3072 units)**, áp dụng hàm kích hoạt **GELU** để tinh chỉnh biểu diễn ngữ nghĩa.

- **Vector Embedding (CLS):** Sau 12 lớp encoder, vector tương ứng với token [CLS] (token đã được thêm vào đầu câu) được trích xuất làm đại diện tổng hợp nội dung cho toàn bộ câu.
- **Điều chỉnh mô hình:** Vector Embedding CLS được đưa ra vào một đầu ra phân lớp (được thêm vào sau mô hình BERT-BASE) bao gồm một lớp FFNN cộng với lớp Sigmoid để thực hiện phân lớp và đưa ra nhãn cảm xúc.
- **Dữ liệu kiểm thử:** Các đánh giá của người dùng trên một sản phẩm.
- **Nhãn cảm xúc:** Kết quả phân loại cảm xúc các đánh giá là tiêu cực hoặc tích cực.

b. Sơ đồ thuật toán sử dụng mô hình CNN – BiLSTM



Hình 14. Sơ đồ tổng quát sử dụng thuật toán CNN-BiLSTM

– **Mô tả:**

- **Tiền xử lý:** Xóa kí tự đặc biệt, biểu tượng cảm xúc, số, stopword, tách từ đơn, từ ghép, viết thường,
- **Vector hóa:** Chuẩn hóa độ dài của mỗi đánh giá và vector hóa từ dựa trên bộ từ vựng.
- **Tạo ma trận đặc trưng:** Sử dụng word embedding để tạo ma trận đặc trưng cho các từ.
- **Huấn luyện mô hình:** Huấn luyện model bằng các thuật toán học sâu (CNN và Bi-LSTM).
- **Dữ liệu:** Các đánh giá của một người dùng trên một sản phẩm.

- **Kết quả:** Kết quả phân loại cảm xúc các đánh giá là tiêu cực hoặc tích cực..

4.1 Thuật toán

4.1.1 Thuật toán sử dụng mô hình BERT

4.1.1.1 Tiền xử lí dữ liệu cho mô hình BERT

- **Dữ Liệu Đã Tiền Xử Lí Chung**

- + Xóa kí tự đặc biệt, stopword, biểu tượng cảm xúc, link, ...
- + Viết thường, tách từ đơn và từ ghép,...

- **[CLS]:**

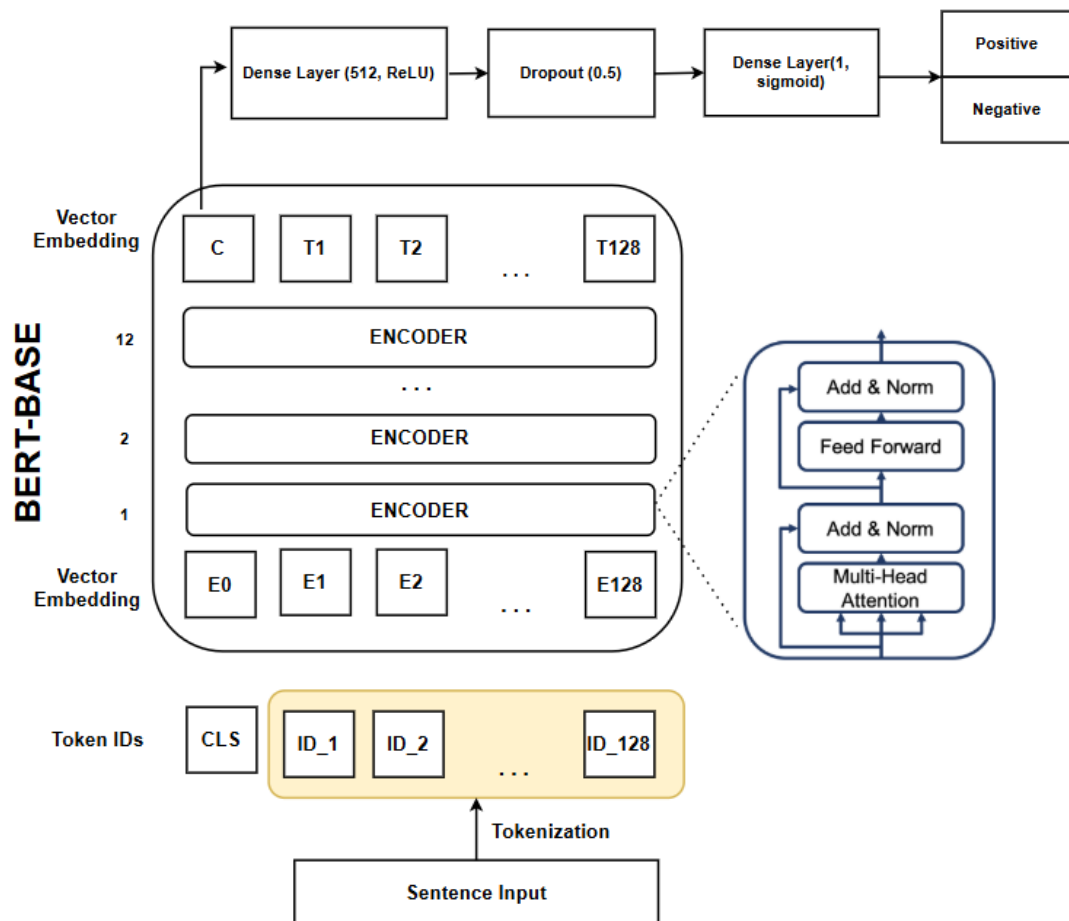
- + Token này được thêm vào đầu mỗi chuỗi đầu vào.
- + Nó đại diện cho ngữ nghĩa tổng thể của toàn bộ câu, và thường được sử dụng trong các tác vụ phân loại.

- **Tokenization:**

- + Toàn bộ chuỗi (bao gồm [CLS], token của các từ trong câu) được chuyển thành token IDs dựa trên bộ từ điển của BERT, để đưa vào mô hình BERT.
- + Mỗi chuỗi được cắt hoặc thêm token <PAD> để đạt chiều dài cố định $SEQ_LEN = 128$.

4.1.1.2 Kiến trúc mô hình đề xuất

Sơ Đồ Kiến Trúc Mô Hình BERT



Hình 15. Sơ đồ kiến trúc mô hình BERT-BASE

- **Lớp Embedding:**

- Chức năng: Ánh xạ chuỗi token ID thành Vector Embedding, giúp mô hình hiểu được ngữ nghĩa và ngữ cảnh.
- Cấu trúc Embedding:
 - + **Token Embeddings (TE):**
 - Biểu diễn ngữ nghĩa cơ bản của từng token.
 - Kích thước: 768 chiều (đối với BERT-BASE).
 - + **Segment Embeddings (SE):** Dùng để phân biệt câu thứ nhất và câu thứ hai trong ngữ cảnh có nhiều câu.
 - + **Position Embeddings (PE):** Thêm thông tin về vị trí thứ tự của các token trong câu, đảm bảo mô hình hiểu được quan hệ thứ tự giữa các từ.
- Tổng hợp: TE+SE+PE tạo thành vector embedding đầu vào cho mỗi token.

- **Bộ mã hóa (BERT Encoder)**

- Chức năng: Học biểu diễn ngữ nghĩa toàn cục của chuỗi đầu vào thông qua cơ chế **Self-Attention** và các lớp encoder transformer.
- Cấu trúc Encoder:
 - + Số lớp (Layers): Mô hình BERT-BASE gồm 12 lớp encoder, mỗi lớp có cấu trúc tương tự nhau.
 - + Cơ chế **Multi-Head Self-Attention**:
 - Chia các vector embedding thành 12 đầu (attention heads).
 - Học mối quan hệ ngữ nghĩa giữa các token, bất kể khoảng cách trong chuỗi đầu vào.
 - Các trọng số attention xác định mức độ quan trọng của từng token trong ngữ cảnh.
 - + Feed Forward Neural Network (FFNN):
 - Gồm 2 lớp Dense:
 - Lớp thứ nhất có 3072 units, kích hoạt bằng hàm GELU (Gaussian Error Linear Unit).
 - Lớp thứ hai giảm về kích thước 768 units (để phù hợp với đầu ra của attention).
 - Mục đích: Biến đổi và tinh chỉnh thông tin sau bước attention.
 - + Add & Norm: Chuẩn hóa và kết hợp thông tin đầu ra từ các thành phần attention và FFNN.
- Đầu ra:
 - + Tại mỗi lớp, các vector embedding của các token được cập nhật dựa trên mối quan hệ ngữ cảnh.
 - + Vector của token [CLS] từ lớp cuối cùng chứa thông tin tổng hợp toàn bộ chuỗi đầu vào, kích thước 768 chiều.

- **Đầu ra phân lớp (Classification Head)**

- Cấu trúc:
 - + Dense Layer:
 - Chức năng:
 - Chuyển đổi vector [CLS] từ không gian 768 chiều sang không gian biểu diễn nhỏ hơn (512 chiều).
 - Học các đặc trưng phi tuyến phục vụ cho bài toán phân loại.
 - Thông số:
 - Số units: 512.

- Hàm kích hoạt: **ReLU** để thêm tính phi tuyến.
- Dropout: Tỷ lệ dropout 0.5 để giảm thiểu hiện tượng overfitting.
- + Output Layer:
 - Chức năng:
 - Dự đoán xác suất cảm xúc của câu dựa trên đặc trưng từ vector.
 - Phân loại đầu ra thành hai nhãn: **Positive** và **Negative**.
 - Thông số:
 - Số units: 1 (đầu ra nhị phân).
 - Hàm kích hoạt: **Sigmoid**, chuẩn hóa đầu ra thành xác suất trong khoảng [0, 1].

4.1.1.3 Triển khai Mô Hình

- **Chuẩn Bị Môi Trường Làm Việc**
 - **Ngôn Ngữ Lập Trình: Python**
 - **Thư Viện:**
 - + Tensorflow và keras: Xây dựng và huấn luyện mô hình học sâu.
 - + keras-bert: Sử dụng mô hình BERT đã được huấn luyện trước.
 - + sklearn: Xử lý dữ liệu và đánh giá hiệu năng mô hình.
 - + numpy, pandas: Thao tác và xử lý dữ liệu.
 - + google.colab: Hỗ trợ quản lý dữ liệu từ Google Drive và môi trường Colab.
 - + matplotlib, seaborn: Trực quan hóa dữ liệu.
- **Tải dữ liệu**
 - Dữ liệu huấn luyện, kiểm tra, và kiểm định được lưu trữ dưới dạng các tệp CSV (đã được tiền xử lý).
 - Mỗi tệp bao gồm hai cột: **review** chứa nội dung văn bản và **label** chứa nhãn tương ứng (0 hoặc 1).
- **Tải mô hình BERT đã được huấn luyện trước**
 - Sử dụng phiên bản đa ngôn ngữ **multi_cased_L-12_H-768_A-12** của Google, đã được huấn luyện trước trên nhiều ngôn ngữ khác nhau.
 - Các thành phần tải về bao gồm:
 - + bert_config.json: Cấu hình mô hình.
 - + bert_model.ckpt: Trọng số đã được huấn luyện.
 - + vocab.txt: Từ điển để token hóa dữ liệu đầu vào.

- Cấu hình mô hình: Mô hình được thiết lập với độ dài chuỗi tối đa (SEQ_LEN) là 128 và sử dụng 4 lớp BERT đầu ra (output_layer_num=4).
- **Token hóa và vector hóa dữ liệu**
 - **Token hóa dữ liệu:**
 - + Sử dụng Tokenizer của keras-bert để chuyển đổi các câu văn bản thành các chuỗi mã hóa.
 - + Quá trình mã hóa bao gồm:
 - Thêm các token đặc biệt ([CLS], [SEP]) vào đầu và cuối mỗi câu.
 - Chuyển đổi từng từ thành ID dựa trên từ điển (vocab.txt).
 - Đệm (padding) các câu ngắn hơn độ dài tối đa (SEQ_LEN) bằng token <PAD>.
 - **Vector hóa dữ liệu:**
 - + Chuỗi đầu vào được chuyển thành hai thành phần:
 - input_ids: Dữ liệu mã hóa của câu văn bản.
 - segment_ids: Thể hiện ranh giới giữa các câu trong mô hình BERT.
- **Thiết kế kiến trúc mô hình**
 - **Tích hợp đầu ra từ BERT:**
 - + Tầng đầu ra của BERT được cấu hình với 4 lớp cuối cùng, sau đó tích hợp thông qua lớp Extract để lấy vector đại diện từ vị trí [CLS].
 - **Thêm các tầng xử lý:**
 - + Dense Layer: Một lớp dày đặc với 512 đơn vị và hàm kích hoạt ReLU để giảm chiều dữ liệu và trích xuất đặc trưng.
 - + Dropout Layer: Lớp Dropout với tỷ lệ 0.5 để giảm thiểu overfitting.
 - + Output Layer: Lớp đầu ra với một đơn vị và hàm kích hoạt sigmoid để tính toán xác suất nhãn dự đoán.
- **Huấn luyện mô hình**
 - **Cấu hình huấn luyện:**
 - + Hàm mất mát: binary_crossentropy để đo lường độ sai lệch giữa nhãn dự đoán và nhãn thực tế.
 - + Trình tối ưu hóa: Sử dụng thuật toán Adam với learning rate 1e-5.
 - + Chỉ số đánh giá: Độ chính xác (accuracy) được sử dụng để theo dõi hiệu năng mô hình trong quá trình huấn luyện.
 - **Callbacks:**

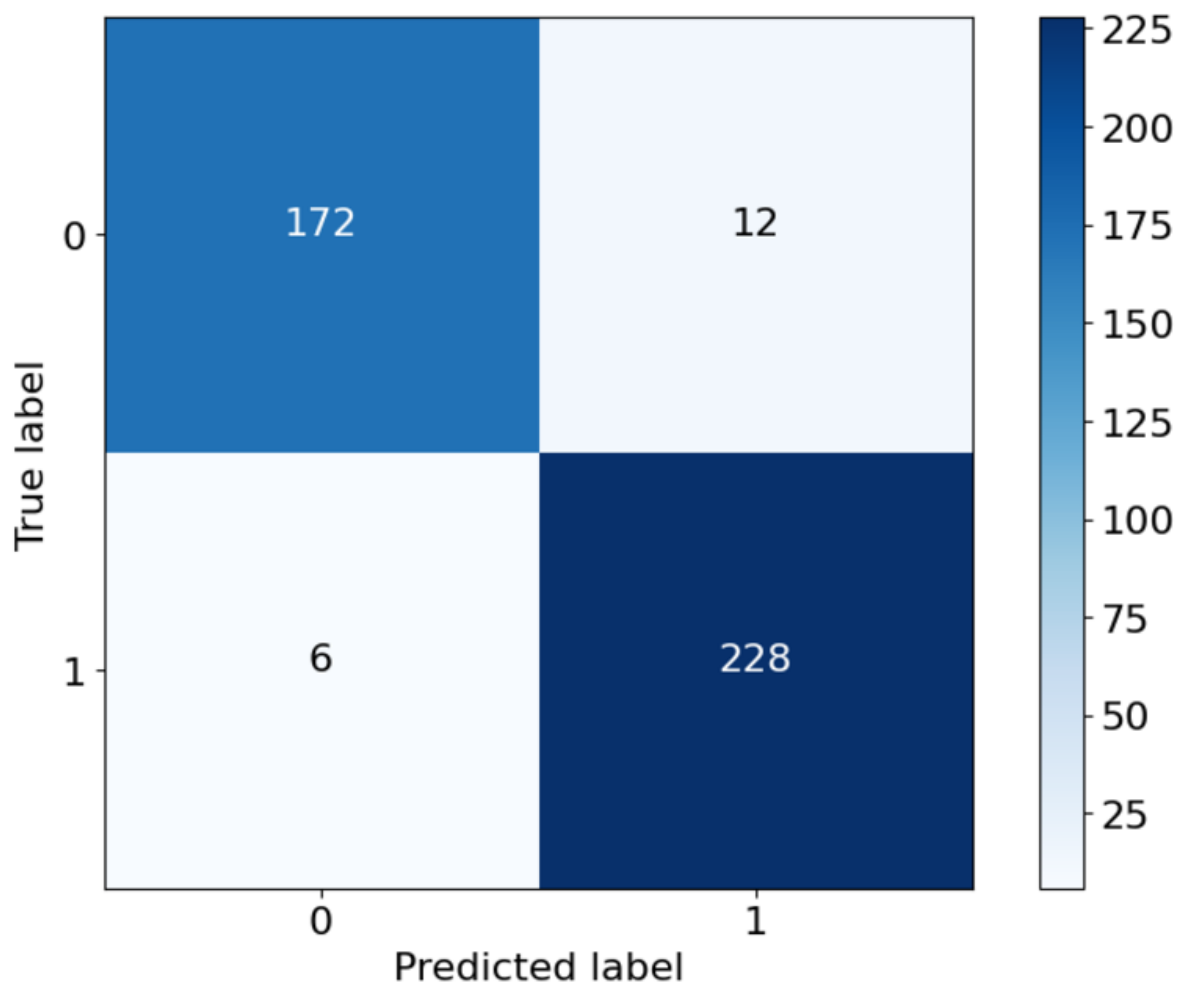
- + Checkpoint: Lưu trọng số mô hình tại điểm có giá trị mất mát nhỏ nhất (val_loss) và độ chính xác cao nhất (val_accuracy) trên tập kiểm định.
- **Thông số huấn luyện:**
 - + Epochs: 45.
 - + Batch size: 16.

4.1.1.4 Đánh giá Mô Hình

- **Kết quả trên tập thử nghiệm**

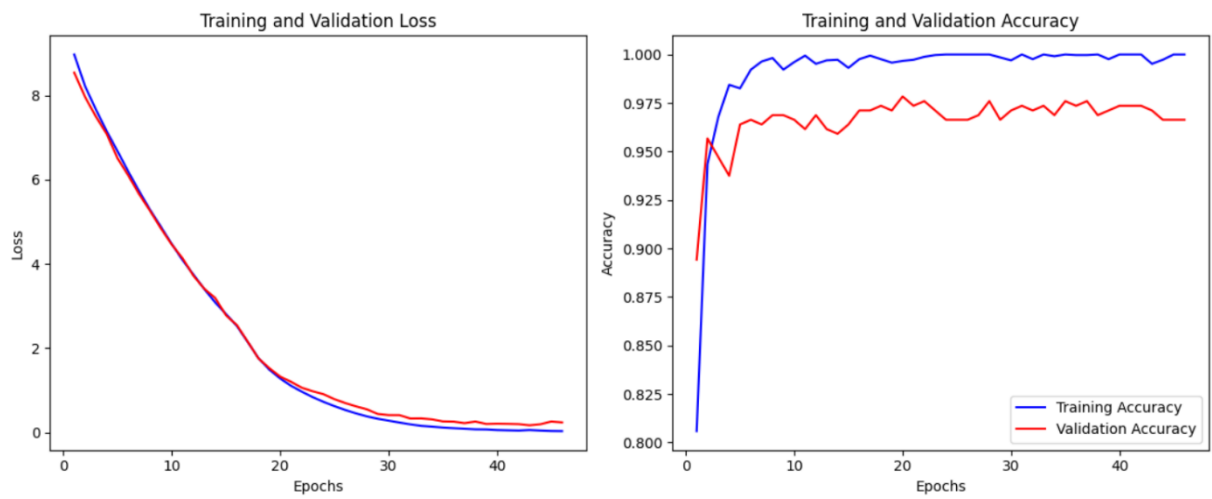
Accuracy	Loss	Precision	F1 Score	Recall
0.9569	0.22	0.9571	0.9568	0.9569

Bảng 1. Đánh giá BERT trên tập test



Hình 16. Confusion matrix trên tập test với mô hình BERT

- **Biểu đồ trực quan Loss và Accuracy**



Hình 17. Biểu đồ Training của mô hình BERT

4.1.2 Thuật toán sử dụng mô hình CNN-BiLSTM

4.1.2.1 Tiền xử lý dữ liệu cho mô hình CNN-BiLSTM

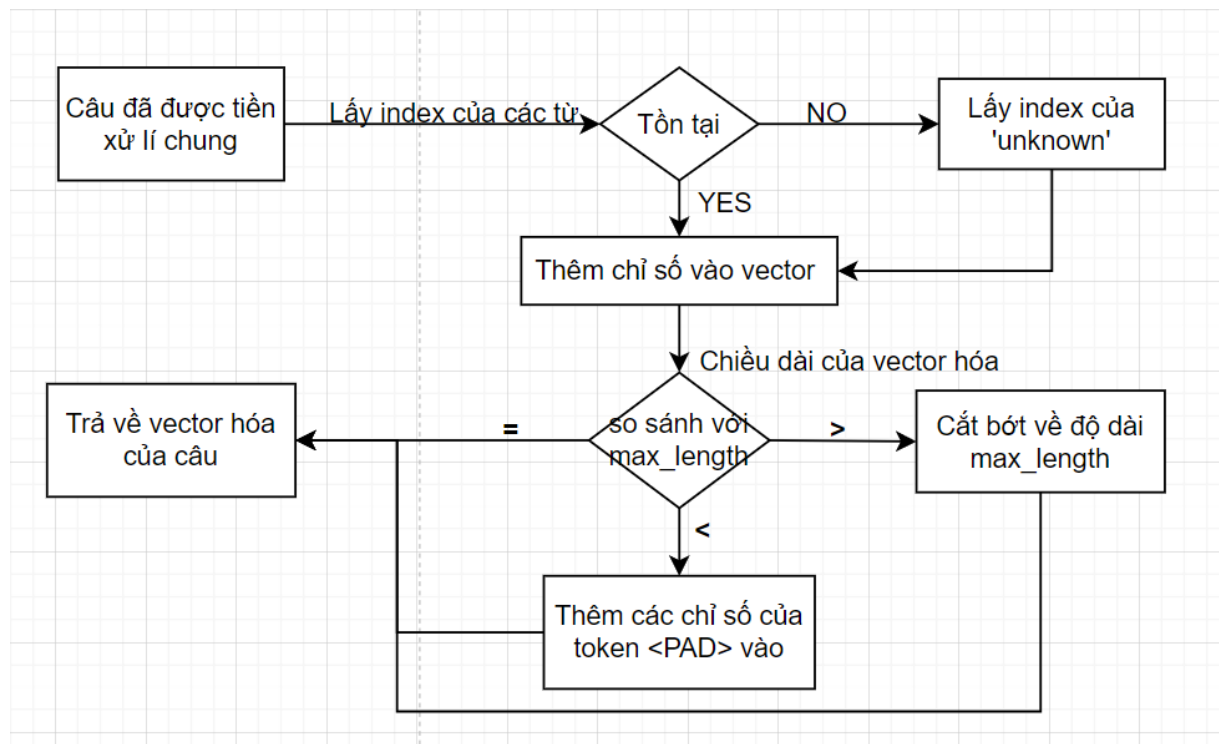
- **Dữ Liệu Đã Tiền Xử Lý Chung**

- + Xóa kí tự đặc biệt, stopword, biểu tượng cảm xúc, link, ...
- + Viết thường, tách từ đơn và từ ghép

- **Vector Hóa và Padding:**

+ Chuyển đổi văn bản thành chuỗi số dựa trên từ điển đã tạo và áp dụng padding để đảm bảo tất cả các chuỗi có cùng độ dài ($\text{max_length} = 25$)[4].

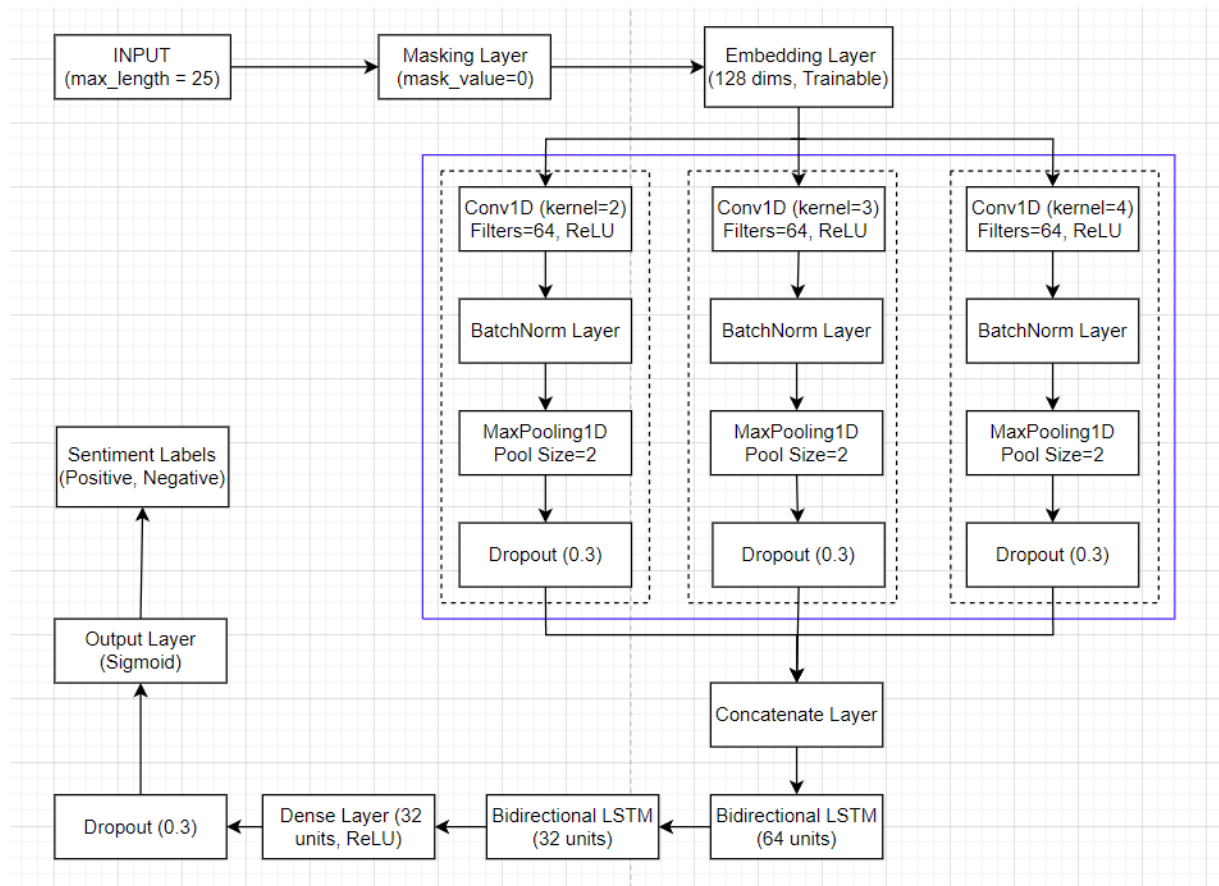
- **Sơ Đồ Tiền Xử Lý**



Hình 18. Sơ đồ tiền xử lý

4.1.2.2 Kiến trúc mô hình đề xuất

Sơ Đồ Kiến Trúc Mô Hình CNN-BiLSTM



Hình 19. Sơ đồ kiến trúc mô hình CNN-BiLSTM

- **Lớp Nhúng (Embedding Layer)**
 - Chức năng: Chuyển đổi các từ trong văn bản thành các vector số học, đại diện cho ý nghĩa ngữ nghĩa của từ.
 - Kỹ thuật: Sử dụng Word2Vec để tạo các vector nhúng từ.
 - Thông số:
 - + Kích thước vector nhúng: 128 chiều.
 - + Trainable: Cho phép tinh chỉnh vector embedding trong quá trình huấn luyện để phù hợp hơn với nhiệm vụ phân loại.
- **Lớp Masking (Masking Layer)**
 - Chức năng: Bỏ qua các vị trí padding (<PAD>) trong chuỗi đầu vào, giúp mô hình không bị ảnh hưởng bởi các giá trị padding không mang thông tin.
 - Thông số:
 - + Mask Value: Giá trị của token <PAD> (0 trong trường hợp này).

- **Các Lớp CNN với Kích Thước Bộ Lọc Khác Nhau (Convolutional Layers)**
 - Chức năng: Trích xuất các đặc trưng cục bộ từ chuỗi từ trong văn bản bằng cách sử dụng các bộ lọc (filters) với kích thước khác nhau.
 - Thông số:
 - + Kernel Sizes: 2, 3, 4 từ.
 - + Số lượng bộ lọc: 64 bộ lọc cho mỗi kích thước kernel.
 - + Kích hoạt: ReLU.
 - + Regularization: L2 regularization với hệ số 0.001 để giảm thiểu overfitting.
 - **Các bước xử lý sau lớp CNN:**
 - + Batch Normalization: Chuẩn hóa các đặc trưng đầu ra từ lớp CNN để tăng tốc độ huấn luyện và ổn định mô hình.
 - + Max Pooling: Giảm kích thước không gian của các đặc trưng đã trích xuất, giúp giảm số lượng tham số và tránh overfitting.
 - + Dropout: Áp dụng dropout với tỷ lệ 0.3 để ngăn chặn overfitting bằng cách ngẫu nhiên bỏ qua một tỷ lệ phần trăm các nơ-ron trong quá trình huấn luyện.
- **Lớp Kết Hợp Các Đặc Trưng CNN (Concatenate Layer)**
 - Chức năng: Kết hợp các đặc trưng đã trích xuất từ các lớp CNN với các kích thước kernel khác nhau thành một tensor duy nhất, giúp mô hình tận dụng được đa dạng các đặc trưng cục bộ từ văn bản.
- **Các Lớp BiLSTM (Bidirectional LSTM Layers)**
 - Chức năng: Xử lý các đặc trưng đã được trích xuất bởi CNN theo cả hai hướng của chuỗi từ, giúp mô hình nắm bắt được ngữ cảnh từ cả phía trước và phía sau của mỗi từ trong câu.
 - Thông số:
 - + Lớp BiLSTM đầu tiên:
 - Số đơn vị: 64
 - Dropout: 0.3
 - Recurrent Dropout: 0.3
 - Return Sequences: True để truyền dữ liệu tới lớp BiLSTM tiếp theo.
 - + Lớp BiLSTM thứ hai:
 - Số đơn vị: 32
 - Dropout: 0.3
 - Recurrent Dropout: 0.3
 - Return Sequences: False để trả về chỉ trạng thái cuối cùng.
- **Lớp Kết Nối Đầy Đủ (Dense Layer)**

- Chức năng: Kết nối các đặc trưng từ lớp BiLSTM để tạo ra các đầu ra cuối cùng cho nhiệm vụ phân loại.
- Thông số:
 - + Số đơn vị: 32
 - + Kích hoạt: ReLU
 - + Regularization: L2 regularization với hệ số 0.001
 - + Dropout: Áp dụng dropout với tỷ lệ 0.3 để giảm thiểu overfitting.
- **Lớp Đầu Ra (Output Layer)**
 - Chức năng: Phân loại cảm xúc vào hai lớp tương ứng (tích cực và tiêu cực).
 - Thông số:
 - + Số đơn vị: 1
 - + Kích hoạt: sigmoid để chuyển đổi đầu ra thành xác suất thuộc về lớp tích cực.

4.1.2.3 Triển Khai Mô Hình

- **Chuẩn Bị Môi Trường Làm Việc**
 - **Ngôn Ngữ Lập Trình: Python**
 - **Thư Viện:**
 - + underthesea: Xử lý ngôn ngữ tự nhiên tiếng Việt.
 - + gensim: Xây dựng mô hình Word2Vec.
 - + h5py: Làm việc với tệp HDF5.
 - + numpy, pandas: Xử lý dữ liệu.
 - + sklearn: Tính toán các chỉ số đánh giá.
 - + matplotlib, seaborn: Trực quan hóa dữ liệu.
 - + tensorflow, keras: Xây dựng và huấn luyện mô hình CNN-BiLSTM.
- **Kết Nối Google Drive và Tải Dữ Liệu**
 - **Tải dữ liệu:** Tải các bộ dữ liệu train, test và validation cho quá trình huấn luyện và kiểm thử mô hình.
- **Tạo Từ Điển với Token <PAD> Riêng Biệt**
 - **Xây dựng từ điển:** Đếm tần suất xuất hiện của mỗi từ trong tập huấn luyện và lọc các từ xuất hiện ít nhất 2 lần.
 - **Khởi tạo từ điển:** Bắt đầu với các token đặc biệt (<PAD>, digit, unknown) và thêm các từ còn lại.
 - **Lưu từ điển:** Ghi từ điển vào tệp voca.txt để sử dụng trong các bước sau.
- **Vector Hóa**

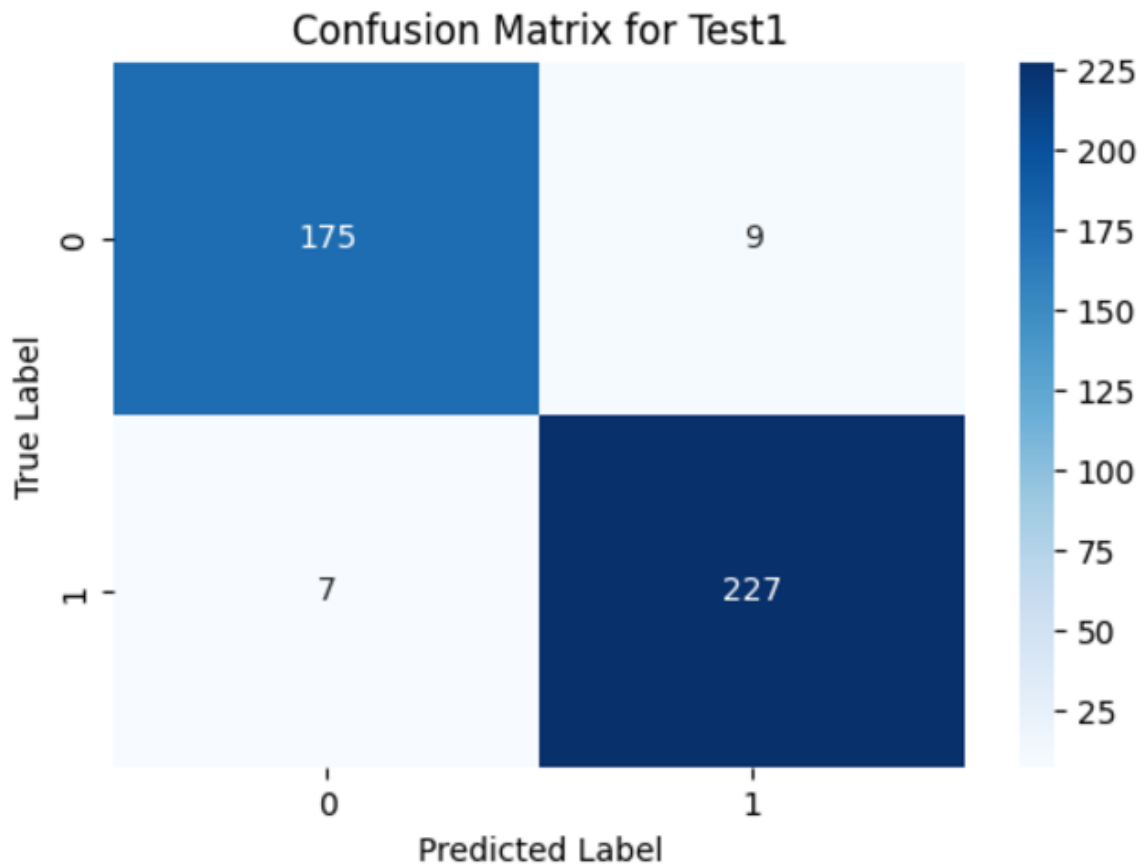
- **Chuyển đổi văn bản thành chuỗi số:** Sử dụng từ điển đã tạo để chuyển đổi từng từ trong văn bản thành số tương ứng.
- **Padding:** Đảm bảo tất cả các chuỗi có cùng độ dài ($\text{max_length} = 25$) bằng cách thêm các giá trị <PAD> ở cuối chuỗi nếu cần thiết.
- **Tạo Ma Trận Embedding với Token <PAD>**
 - **Xây dựng mô hình Word2Vec:** Tạo các vector nhúng từ dữ liệu văn bản đã được tách từ.
 - **Tạo ma trận embedding:** Khởi tạo ma trận embedding với các vector ngẫu nhiên và cập nhật các vector từ điển từ mô hình Word2Vec.
 - **Khởi tạo vector cho các token đặc biệt:** Đặt vector cho <PAD> là toàn 0 và các vector cho digit và unknown là các giá trị ngẫu nhiên từ phân phối chuẩn.
- **Xây Dựng Mô Hình với Masking và Các Lớp CNN và LSTM**
 - **Input Layer:** Nhận chuỗi số đại diện cho văn bản với độ dài tối đa là 25 từ.
 - **Masking Layer:** Bỏ qua các vị trí padding (<PAD>) trong chuỗi đầu vào.
 - **Embedding Layer:** Chuyển đổi các từ thành vector số học với kích thước 128 chiều, sử dụng ma trận embedding đã tạo.
 - **Convolutional Layers:** Ba lớp Conv1D với kích thước kernel 2, 3, 4 và 64 bộ lọc mỗi lớp, theo sau bởi Batch Normalization, Max Pooling và Dropout.
 - **Concatenate Layer:** Kết hợp các đặc trưng từ ba lớp Conv1D khác nhau thành một tensor duy nhất.
 - **Bidirectional LSTM Layers:** Hai lớp BiLSTM với số đơn vị giảm dần (64 và 32), mỗi lớp đều có dropout 0.3 để ngăn chặn overfitting.
 - **Dense Layer:** Lớp kết nối đầy đủ với 32 đơn vị và kích hoạt ReLU, theo sau là Dropout để giảm thiểu overfitting.
 - **Output Layer:** Lớp đầu ra với 1 đơn vị và kích hoạt sigmoid, chuyển đổi đầu ra thành xác suất thuộc về lớp tích cực.

4.1.2.4 Đánh Giá Mô Hình

- **Kết quả trên tập thử nghiệm**

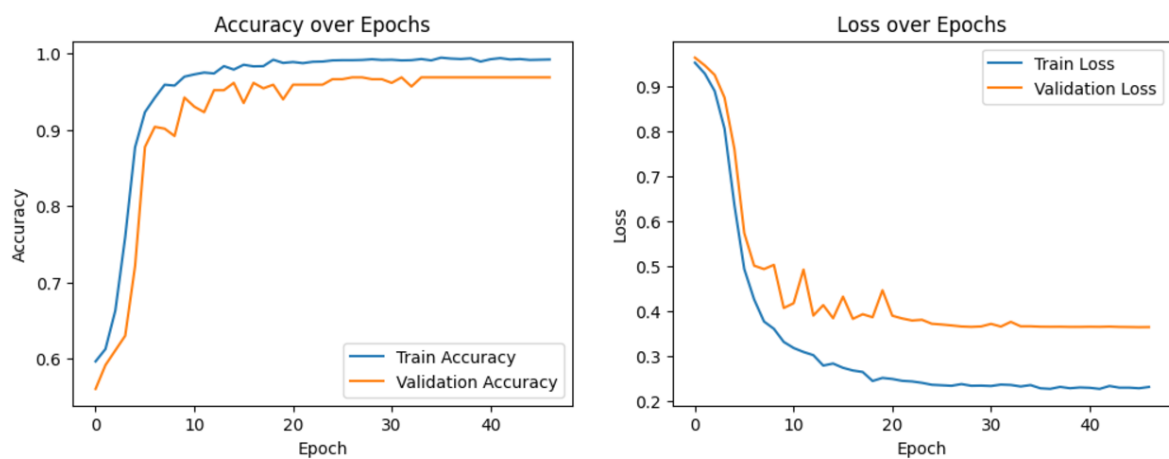
Accuracy	Precision	F1 Score	Recall
0.9617	0.9617	0.9617	0.962

Bảng 2. Kết quả phân loại của CNN-BiLSTM trên tập test



Hình 20. Confusion matrix của tập test với model CNN-BiLSTM

- **Trực Quan Biểu Đồ Loss & Accuracy**



Hình 21. Biểu đồ training của model CNN-BiLSTM

4.2 So sánh và đánh giá hai thuật toán

4.2.1. So sánh kết quả

Sau khi triển khai, chúng tôi đã tiến hành đánh giá hiệu suất của hai mô hình BERT và CNN-BiLSTM dựa trên bộ dữ liệu kiểm thử gồm 418 đánh giá sản phẩm. Dưới đây là các kết quả đánh giá chi tiết:

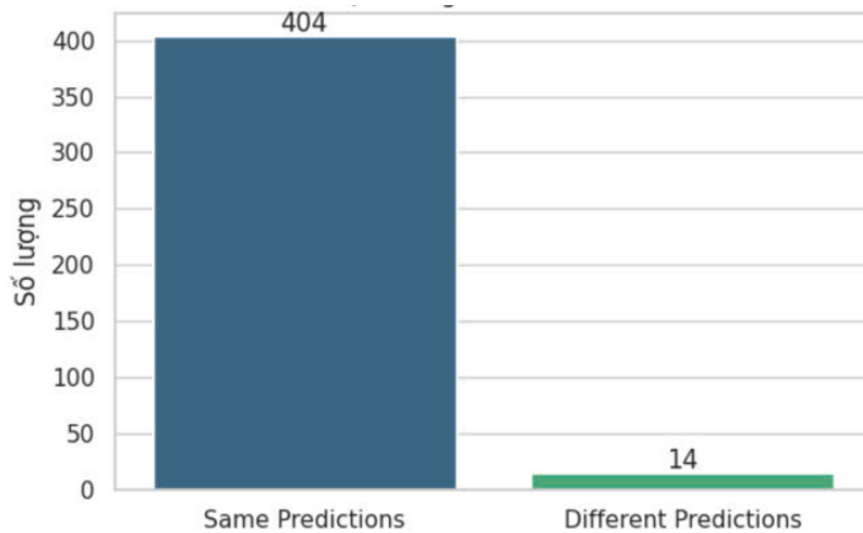
Các chỉ số đánh giá của 2 model trên tập test:

Model	Accuracy	Precision	Recall	F1-Score
BERT	0.956938	0.957172	0.956938	0.956853
CNN-BiLSTM	0.961722	0.961721	0.961722	0.961699

Bảng 3. Chỉ số đánh giá của hai model trên tập test

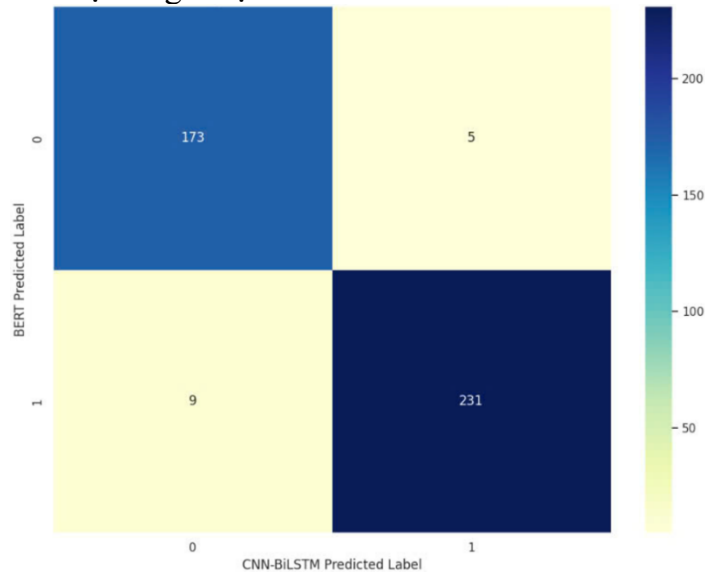
- **Đánh giá:** Cả hai mô hình đều đạt độ chính xác cao, CNN-BiLSTM cao hơn một chút. Điều này cho thấy cả hai mô hình đều rất hiệu quả trong việc phân loại đánh giá sản phẩm.
- **Precision, Recall và F1-Score:**
 - Precision: Cả hai mô hình đều đạt mức Precision trên 95%, chứng tỏ khả năng giảm thiểu các dự đoán sai lầm.
 - Recall: Độ bao phủ (Recall) của cả hai mô hình cũng vượt trội, trên 95%, cho thấy khả năng nhận diện chính xác các đánh giá tích cực và tiêu cực.
 - F1-Score: Mô hình CNN-BiLSTM đạt F1-Score cao hơn một chút so với BERT, phản ánh sự cân bằng tốt giữa Precision và Recall.

Biểu đồ so sánh số lượng dự đoán giống và khác:



Hình 22. Biểu đồ so sánh số lượng dự đoán giống và khác nhau

Heatmap so sánh mức độ đồng thuận:



Hình 23. Biểu đồ Heatmap hai model

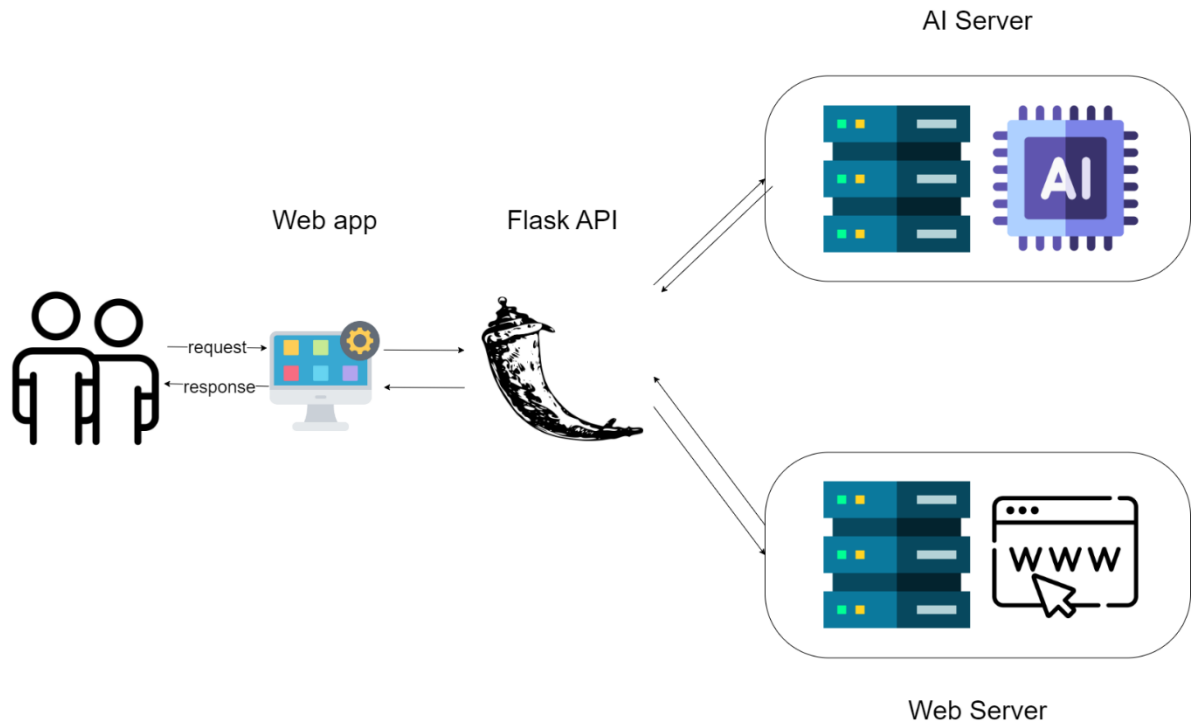
- Mức độ đồng thuận giữa hai mô hình:
 - 96.65% các đánh giá được dự đoán giống nhau bởi cả hai mô hình.
 - 3.35% các đánh giá được dự đoán khác nhau.
- Đánh giá: Mức độ đồng thuận cao giữa hai mô hình cho thấy sự ổn định và đáng tin cậy trong việc phân loại đánh giá. Những sự khác biệt nhỏ có thể là do đặc trưng riêng của từng mô hình trong việc xử lý dữ liệu.

4.2.2. Đánh giá và kết luận

- **Hiệu suất cao:** Cả hai mô hình BERT và CNN-BiLSTM đều thể hiện hiệu suất phân loại rất tốt với độ chính xác và các chỉ số đánh giá trên 95%. Điều này chứng tỏ rằng hệ thống của chúng tôi rất đáng tin cậy trong việc phân loại đánh giá sản phẩm.
- **CNN-BiLSTM vượt trội hơn BERT:** Mô hình CNN-BiLSTM đạt được các chỉ số cao hơn một chút so với BERT, cho thấy mô hình này có thể xử lý tốt hơn trong việc trích xuất đặc trưng và phân loại đánh giá.
- **Đồng thuận cao giữa hai mô hình:** Với 96.65% đánh giá được dự đoán giống nhau, việc sử dụng hai mô hình song song giúp tăng cường độ tin cậy và cung cấp các kết quả dự đoán đa dạng hơn cho người dùng.
- **Sự khác biệt nhỏ:** Mặc dù 3.35% đánh giá được dự đoán khác nhau, mức độ này là khá thấp và không ảnh hưởng đáng kể đến tổng thể hệ thống. Những sự khác biệt này có thể được sử dụng để phân tích thêm và cải thiện hệ thống trong tương lai.
- **Lợi ích tổng thể:** Việc triển khai đồng thời hai mô hình giúp hệ thống không chỉ đạt được hiệu suất cao mà còn cung cấp sự linh hoạt và ổn định, đảm bảo rằng người dùng luôn nhận được những đánh giá chính xác và hữu ích.

5 THIẾT KẾ ỨNG DỤNG

5.1 Tổ chức chương trình



Hình 24. Kiến trúc web

Hệ thống phân tích cảm xúc comment trên bài viết đánh giá sản phẩm được tổ chức thành các thành phần chính, hoạt động một cách phối hợp nhằm đảm bảo chức năng phân loại cảm xúc và hiển thị kết quả trực quan. Các thành phần này bao gồm:

a. Web Application (Frontend)

- Đây là giao diện chính mà người dùng tương tác. Người dùng nhập dữ liệu đầu vào (URL của bài viết hoặc sản phẩm) thông qua giao diện web.
- Web app gửi yêu cầu (request) đến Flask API và hiển thị kết quả phản hồi (response) dưới dạng biểu đồ trực quan và danh sách bình luận phân loại theo nhãn tích cực, tiêu cực.
- Giao diện được thiết kế tối giản, thân thiện, đảm bảo người dùng dễ dàng thao tác.

b. Flask API (Middleware)

- Đây là trung gian xử lý giữa Web App và các máy chủ phía sau (AI Server, Web Server).

- Flask API nhận yêu cầu từ Web App, xử lý dữ liệu đầu vào, và chuyển tiếp đến các mô hình AI để thực hiện phân tích cảm xúc. Sau khi nhận được kết quả từ các máy chủ, API gửi phản hồi lại cho Web App.
- Flask đảm bảo tốc độ xử lý nhanh và khả năng mở rộng khi cần tích hợp thêm các tính năng mới.

c. AI Server (Backend)

- Thành phần này chịu trách nhiệm thực hiện nhiệm vụ phân tích cảm xúc dựa trên dữ liệu đầu vào.
- Hệ thống sử dụng hai mô hình học sâu, BERT và LSTM, để xử lý dữ liệu ngôn ngữ tự nhiên (NLP).
- AI Server nhận dữ liệu từ Flask API, thực hiện phân tích cảm xúc (sentiment analysis) trên các bình luận, và trả về kết quả dưới dạng nhãn cảm xúc (tích cực hoặc tiêu cực).

d. Web Server

- Đây là thành phần hỗ trợ, cung cấp các thông tin bổ sung cần thiết như truy xuất dữ liệu liên quan hoặc lưu trữ kết quả phân tích.
- Web Server kết nối với Flask API để thực hiện các tác vụ xử lý bổ sung, đảm bảo hệ thống hoạt động ổn định và hiệu quả.

Hệ thống được thiết kế linh hoạt, các thành phần độc lập nhưng phối hợp chặt chẽ, dễ dàng bảo trì và nâng cấp. Sự kết hợp giữa Flask API và các mô hình AI đảm bảo tính chính xác và hiệu quả trong phân tích cảm xúc, trong khi giao diện web tạo trải nghiệm người dùng mượt mà.

5.2 Ứng dụng học máy trong chương trình

Trong hệ thống này, hai mô hình học máy **BERT** và **CNN-BiLSTM** được triển khai đồng thời để xử lý và phân tích dữ liệu đánh giá sản phẩm thu thập từ các liên kết mà người dùng nhập vào. Cả hai mô hình này hoạt động độc lập trên cùng một bộ dữ liệu, cung cấp kết quả dự đoán song song để hiển thị trên giao diện web, từ đó mang lại những đánh giá phong phú và chính xác hơn cho người dùng.

5.2.1 Cách thức hoạt động của hai mô hình

Dữ liệu (danh sách các đánh giá) sau khi được thu thập dưới dạng thô sẽ được tiền xử lý và được biến đổi, tokenize hóa để phù hợp với đầu vào của hai mô hình phân lớp đã được training. Sau đó, hai mô hình sẽ thực hiện dự đoán nhãn cho các đánh giá.

Kết quả dự đoán từ cả hai mô hình (tích cực/tiêu cực) sẽ được đẩy lên giao diện website, cho phép người dùng so sánh và hiểu được cảm xúc, mức độ đánh giá của sản phẩm đó.

5.2.2 Lợi ích khi sử dụng hai mô hình cùng nhau

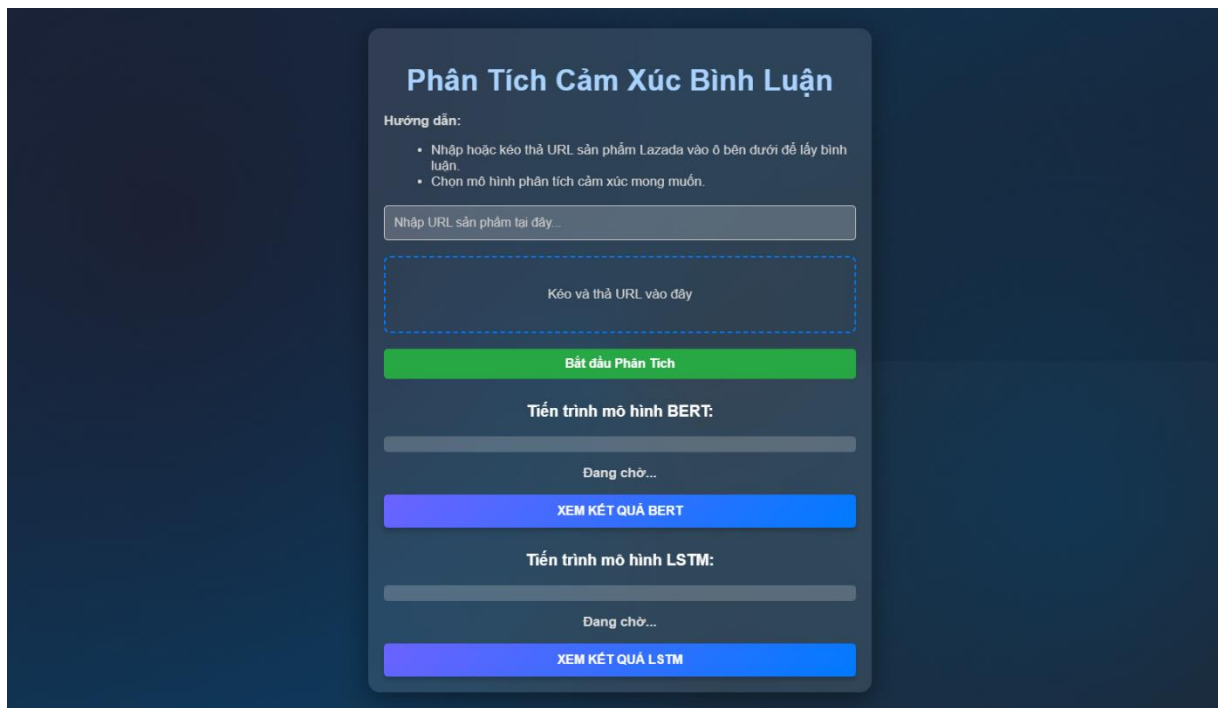
Việc triển khai đồng thời cả hai mô hình **BERT** và **CNN-BiLSTM** mang lại nhiều lợi ích đáng kể cho hệ thống:

- **Tăng độ chính xác và độ tin cậy:** **BERT** với khả năng hiểu ngữ nghĩa sâu sắc và **CNN-BiLSTM** với khả năng trích xuất đặc trưng mạnh mẽ giúp cải thiện đáng kể độ chính xác trong việc phân loại đánh giá.
- **Đa dạng hóa kết quả:** Cung cấp hai loại kết quả dự đoán khác nhau giúp người dùng có thêm thông tin để đưa ra quyết định mua sắm chính xác hơn, từ đó nâng cao trải nghiệm sử dụng web.
- **Ổn định hệ thống:** Nếu một mô hình gặp sự cố hoặc không chính xác trong một số trường hợp, mô hình còn lại vẫn đảm bảo cung cấp kết quả đáng tin cậy, tăng tính ổn định cho hệ thống tổng thể.
- **Cải thiện hiệu suất xử lý:** Hai mô hình hoạt động song song giúp giảm thời gian xử lý dữ liệu và trả kết quả nhanh chóng hơn, đáp ứng yêu cầu về thời gian thực của người dùng.
- **Khả năng mở rộng:** Hệ thống có thể dễ dàng mở rộng bằng cách thêm các mô hình học máy khác trong tương lai để tăng cường khả năng phân tích và dự đoán.

5.3 Kết quả

5.3.1 Giao diện chính của chương trình

- Hình ảnh giao diện chính của chương trình:



Hình 25. Hình ảnh giao diện chính của chương trình

- Mô tả về giao diện chính:

+ **Tiêu đề chính**

○ Tiêu đề của giao diện: "*Phân Tích Cảm Xúc Bình Luận*". Nó được hiển thị ở phần đầu trang để người dùng biết mục đích của chương trình.

+ **Hướng dẫn sử dụng**

Một khung hiển thị hướng dẫn chi tiết cho người dùng gồm các bước:

- **Nhập hoặc kéo thả** URL sản phẩm từ Lazada vào khu vực được chỉ định.
- **Chọn mô hình phân tích cảm xúc** mà người dùng muốn sử dụng.

+ **Khu vực nhập URL**

- Gồm một **ô nhập liệu** với nhãn hướng dẫn người dùng nhập URL sản phẩm trực tiếp vào.
- Một **khu vực kéo và thả** URL, nơi người dùng có thể thả đường dẫn mà không cần nhập tay.

+ **Nút thao tác chính**

- Nút "*Bắt đầu Phân Tích*" với màu sắc nổi bật (màu xanh lá) để kích hoạt quá trình phân tích cảm xúc.

+ **Tiến trình phân tích**

- Mô hình BERT:
 - Một thanh tiến trình cập nhật trạng thái phân tích.
 - Trạng thái ban đầu: "Đang chờ..."

- Các trạng thái tiếp theo: “Đang lấy dữ liệu từ URL...”, “Phân tích bằng BERT”, và “Hoàn tất phân tích bằng BERT”

Phân Tích Cảm Xúc Bình Luận

Hướng dẫn:

- Nhập hoặc kéo thả URL sản phẩm Lazada vào ô bên dưới để lấy bình luận.
- Chọn mô hình phân tích cảm xúc mong muốn.

<https://www.lazada.vn/products/but-mau-acrylic-marker-60-mau-cao-cap-mau-sac>

Kéo và thả URL vào đây

Bắt đầu Phân Tích

Tiến trình mô hình BERT:

Phân tích bằng BERT...

XEM KẾT QUẢ BERT

Tiến trình mô hình LSTM:

Phân tích bằng LSTM...

XEM KẾT QUẢ LSTM

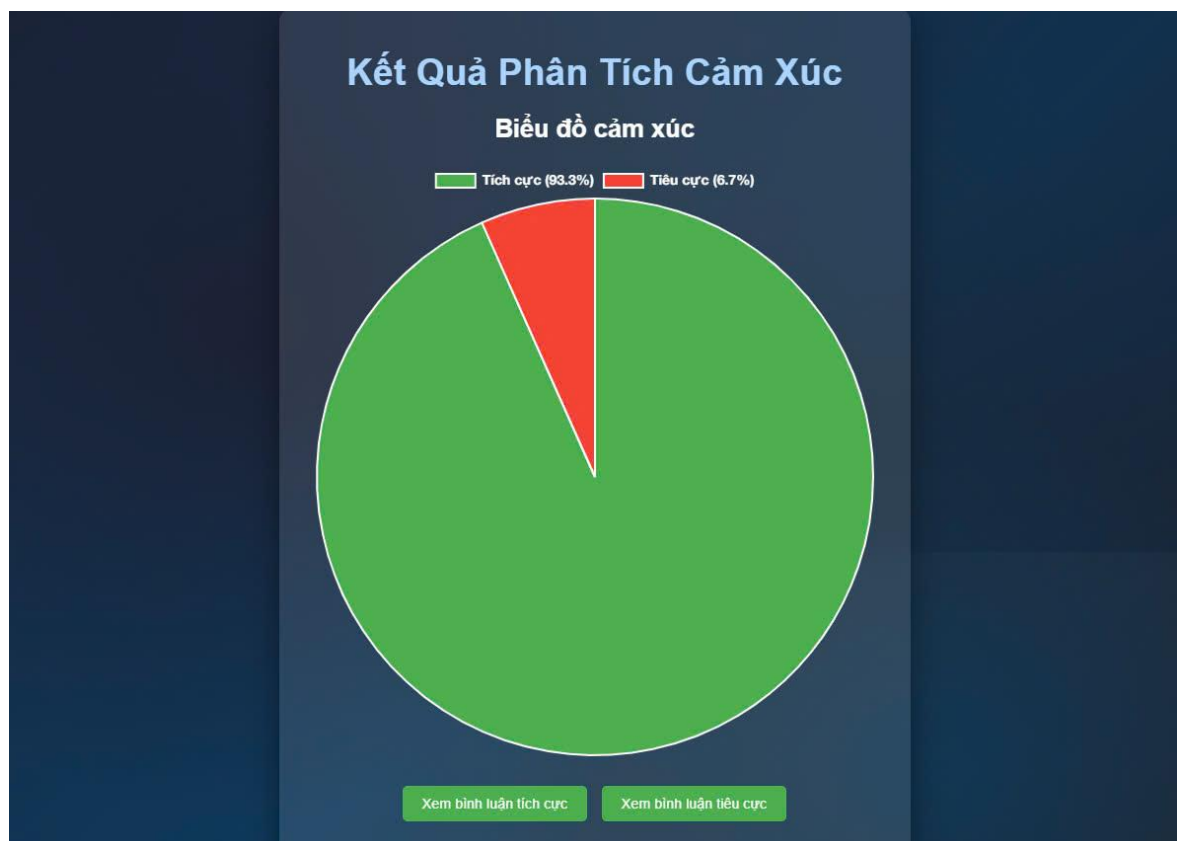
Hình 26. Hình ảnh giao diện đang chạy tiến trình

- Thanh tiến trình giúp người dùng theo dõi trực quan trạng thái và tiến độ của quá trình phân tích. Các trạng thái như "Đang lấy dữ liệu từ URL", "Phân tích bằng BERT", và "Hoàn tất" cung cấp thông tin rõ ràng, giúp người dùng hiểu được chương trình đang hoạt động và khi nào hoàn thành, từ đó nâng cao trải nghiệm sử dụng và giảm sự nóng lòng chờ đợi khi chương trình đang được xử lý.

- Một nút để xem kết quả ("*Xem Kết Quả BERT*").
- Mô hình LSTM:
 - Tương tự với BERT, bao gồm thanh tiến trình, mô tả trạng thái, và nút xem kết quả ("*Xem Kết Quả LSTM*").

+ Thiết kế tổng thể

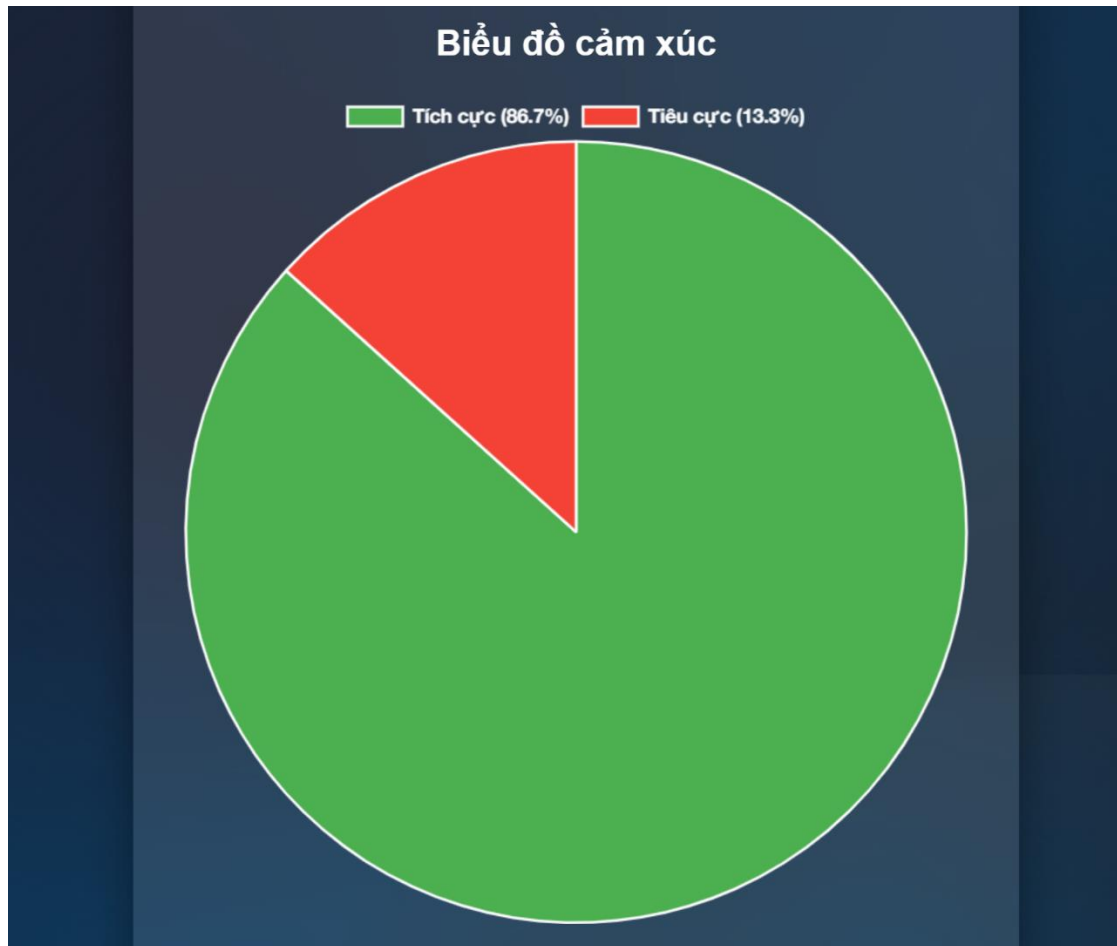
- Giao diện được thiết kế đơn giản, hiện đại với nền màu tối, các phần tử hiển thị rõ ràng và dễ thao tác.
 - Cho phép nhập liệu hoặc kéo thả linh hoạt, phù hợp với các loại người dùng khác nhau.
 - Các nút và thanh tiến trình được phân biệt bằng màu sắc, giúp người dùng dễ nhận diện.
- Hình ảnh phần đầu giao diện kết quả của chương trình:



Hình 27. Hình ảnh phần đầu giao diện kết quả

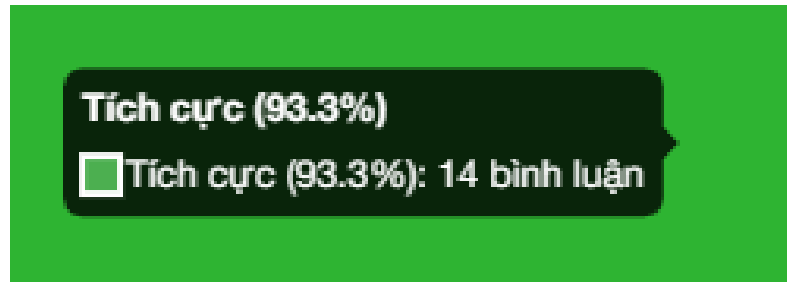
- Mô tả phần đầu hình ảnh kết quả:
 - **Tiêu đề chính:** "*Kết Quả Phân Tích Cảm Xúc*", được đặt ở vị trí trung tâm, dễ nhìn, giúp người dùng xác định nội dung của kết quả.
 - **Biểu đồ cảm xúc:**

- Dạng biểu đồ tròn (Pie Chart) được sử dụng để biểu diễn tỷ lệ cảm xúc.
- Gồm hai màu chính:
 - **Màu xanh lá cây:** Đại diện cho tỷ lệ các bình luận tích cực (xx.x%).
 - **Màu đỏ:** Đại diện cho tỷ lệ các bình luận tiêu cực (xx.x%).



Hình 28. Biểu đồ tỷ lệ tích cực tiêu cực

- Số liệu tỷ lệ được hiển thị rõ ràng ngay trên biểu đồ.
- Ngoài ra, người dùng có thể tương tác với biểu đồ theo những cách:
 - Tất phần tỉ lệ của tiêu cực hoặc tích cực để biểu đồ chỉ hiện phần còn lại.
 - Đưa con trỏ chuột vào phần biểu đồ tích cực hoặc tiêu cực để xem nhưng thông tin: Số lượng bình luận tích/tiêu cực, phần trăm bình luận tích/tiêu cực.



Hình 29. Hình ảnh thông tin hiện ra khi đưa trỏ chuột vào biểu đồ

- Hình ảnh phần sau của giao diện kết quả chương trình:

Xem bình luận tích cực

Xem bình luận tiêu cực

Danh sách bình luận

Bình luận	Cảm xúc
phần nhựa thừa ở mũi giày ko có fom nhưng giá u quá rẻ thì vậy thôi 5 sao	Tiêu cực
đc có điều hơi <u>mỏng</u> manh	Tiêu cực
Quay về Trang Chủ	

Hình 30. Hình ảnh phần sau của giao diện kết quả chương trình

- Mô tả phần sau giao diện kết quả chương trình:
 - + Hai nút "**Xem bình luận tích cực**" và "**Xem bình luận tiêu cực**" ở đầu bảng giúp người dùng dễ dàng chọn loại bình luận muốn xem.
 - + Bảng hiển thị bình luận: Được chia làm hai cột chính:
 - **Cột "Bình luận"**: Hiển thị nội dung chi tiết của từng bình luận.
 - **Cột "Cảm xúc"**: Hiển thị nhãn cảm xúc tương ứng (*Tích cực* hoặc *Tiêu cực*).
 - Các từ khóa quan trọng liên quan đến cảm xúc (*tích cực* hoặc *tiêu cực*) được **highlight** để làm nổi bật ý chính của bình luận.
 - + Nút điều hướng "Quay về Trang Chủ": Được đặt bên dưới bảng để dễ dàng trở lại giao diện chính.
- Lợi ích của giao diện này:
 - **Phân loại trực quan**: Giúp người dùng nhanh chóng nhận diện và tập trung vào các bình luận tích cực hoặc tiêu cực.

- **Hỗ trợ phân tích chi tiết:** Highlight từ khóa giúp dễ dàng hiểu nội dung cảm xúc mà bình luận thể hiện.

- **Tăng trải nghiệm người dùng:** Các nút điều hướng và bố cục rõ ràng giúp thao tác dễ dàng, nâng cao sự thuận tiện khi sử dụng chương trình.

5.3.2 Kết quả thực thi của chương trình

a. Hiệu suất xử lý

- Chương trình hoạt động mượt mà và hiệu quả, với thời gian phân tích một URL sản phẩm dao động từ **30 đến 60 giây**. Khoảng thời gian này cho phép người dùng nhận kết quả nhanh chóng mà vẫn đảm bảo độ chính xác của quá trình xử lý dữ liệu và phân tích cảm xúc.

b. Độ chính xác của mô hình

- Các mô hình phân tích cảm xúc được sử dụng, bao gồm **BERT** và **LSTM**, đạt mức độ chính xác cao, từ **85% đến 90%**. Điều này đảm bảo rằng chương trình có thể phân loại cảm xúc của bình luận một cách đáng tin cậy và hữu ích cho các mục đích nghiên cứu hoặc kinh doanh.
- Mô hình **BERT** cho kết quả phân tích mạnh mẽ nhờ khả năng xử lý ngữ cảnh sâu sắc, phù hợp với những bình luận phức tạp.
- Mô hình **LSTM** cũng hoạt động tốt trong việc nắm bắt mối quan hệ tuần tự trong dữ liệu văn bản, giúp phân tích các bình luận dài mạch lạc hơn.

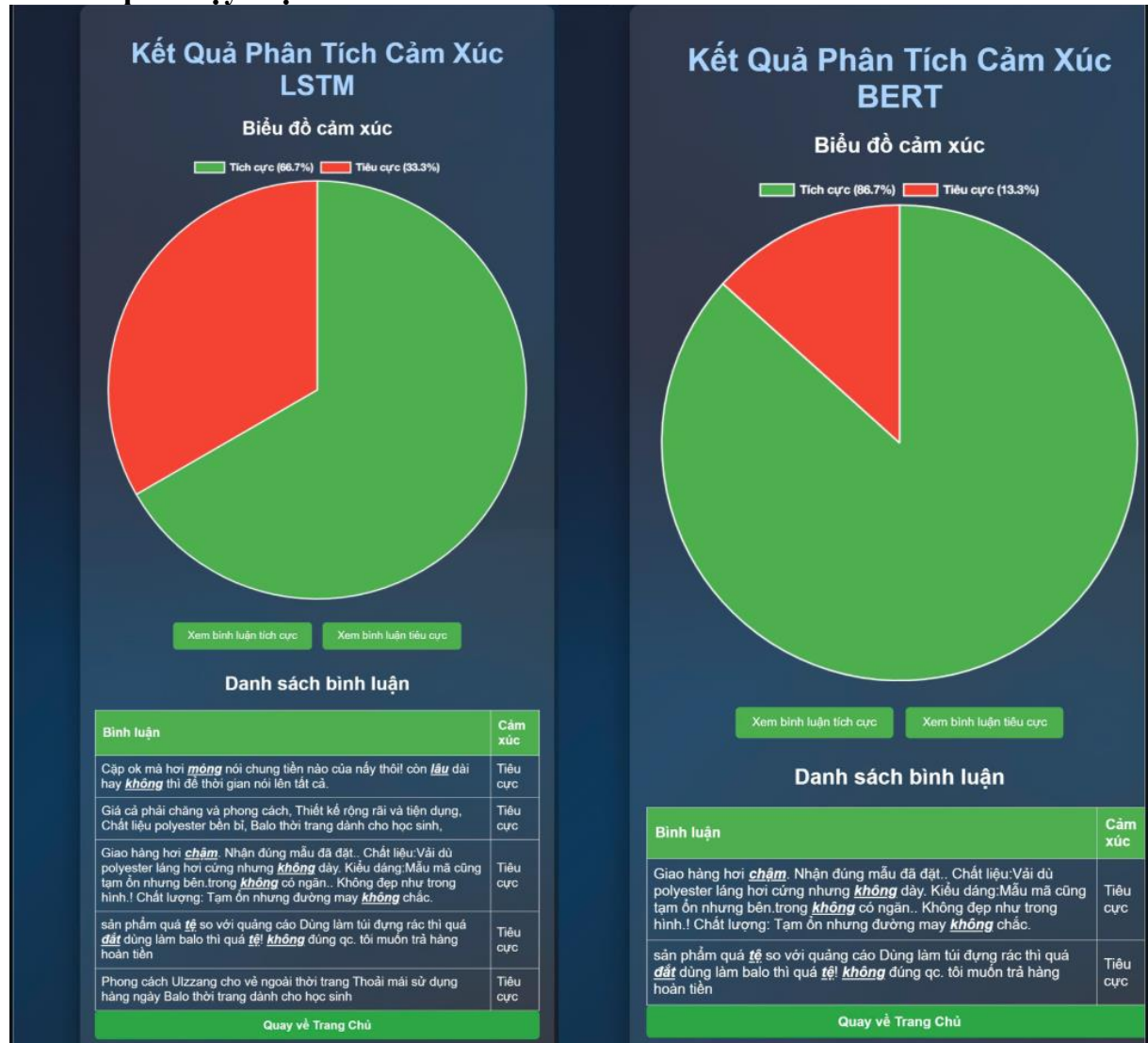
c. Đánh giá trải nghiệm người dùng

- **Giao diện thân thiện:** Người dùng dễ dàng thao tác với giao diện trực quan, từ việc nhập URL đến việc theo dõi tiến trình và xem kết quả phân tích.
- **Trực quan hóa kết quả:** Biểu đồ cảm xúc cùng bảng phân loại bình luận chi tiết cung cấp cái nhìn rõ ràng về cảm xúc của người dùng đối với sản phẩm, giúp đưa ra những nhận định nhanh chóng và chính xác.
- **Tính tương tác cao:** Các tính năng như tương tác với biểu đồ, xem bình luận tích cực hoặc tiêu cực, và quay lại trang chính giúp người dùng sử dụng chương trình một cách linh hoạt và hiệu quả.

d. Một số kết quả nổi bật

- **Trích xuất dữ liệu:** Chương trình đã thu thập thành công hàng trăm bình luận từ các sản phẩm khác nhau trên trang thương mại điện tử. Dữ liệu được lưu trữ dưới dạng tệp CSV, giúp dễ dàng phân tích và chia sẻ.
- **Phân tích cảm xúc:** Biểu đồ cảm xúc đã chỉ ra tỷ lệ bình luận tích cực và tiêu cực một cách chính xác.

- Kết quả chạy thực tế



Hình 31. Kết quả chạy thực tế

6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 Kết luận

Trong bối cảnh kinh doanh trực tuyến không ngừng mở rộng, đề án “Phân tích cảm xúc comment trên một bài viết đánh giá sản phẩm” đã đề xuất và triển khai thành công một hệ thống tự động phân loại cảm xúc (tích cực, tiêu cực) từ các bình luận đánh giá sản phẩm. Hệ thống dựa trên hai mô hình học sâu tiên tiến là BERT và LSTM, kết hợp với ứng dụng web trên nền Flask API, giúp người dùng dễ dàng truy cập và sử dụng.

Kết quả nghiên cứu cho thấy cả hai mô hình đều đạt được hiệu suất tốt khi xử lý bình luận tiếng Việt, trong đó BERT vượt trội hơn về độ chính xác nhờ khả năng học ngữ cảnh hiệu quả. Đề tài đã góp phần giải quyết bài toán xử lý ngôn ngữ tự nhiên trong phân tích cảm xúc, mang lại công cụ hữu ích cho doanh nghiệp trong việc thấu hiểu khách hàng và cải thiện sản phẩm.

Mặc dù đạt được những kết quả đáng khích lệ, hệ thống vẫn còn hạn chế ở một số khía cạnh như độ chính xác trong những trường hợp dữ liệu phức tạp hoặc đa nghĩa, và khả năng mở rộng sang các bối cảnh khác.

6.2 Hướng phát triển

Trong tương lai, đề tài có thể được mở rộng và cải thiện theo các hướng sau:

a. **Nâng cao chất lượng mô hình:**

- Sử dụng các mô hình ngôn ngữ hiện đại hơn, chẳng hạn như GPT hoặc các phiên bản cải tiến của BERT, nhằm tăng độ chính xác và khả năng xử lý dữ liệu phức tạp.
- Bổ sung thêm các nhãn cảm xúc khác như trung lập hoặc các cảm xúc chi tiết (vui vẻ, thất vọng) để phân tích cảm xúc đa chiều.

b. **Mở rộng nguồn dữ liệu:**

- Thu thập dữ liệu từ nhiều nền tảng khác ngoài Lazada và Tiki, như Shopee, Facebook hoặc Zalo, nhằm tăng tính đa dạng và khả năng ứng dụng của mô hình.
- Xử lý dữ liệu chứa tiếng Việt không dấu hoặc các dạng ngôn ngữ không chuẩn thường gặp trong bình luận trực tuyến.

c. **Cải thiện ứng dụng web:**

- Xây dựng giao diện thân thiện hơn với người dùng, tích hợp các tính năng như hiển thị biểu đồ thống kê trực quan, lọc và tìm kiếm cảm xúc theo từ khóa hoặc danh mục sản phẩm.

- Phát triển ứng dụng thành công cụ trực tuyến có thể phân tích cảm xúc theo thời gian thực.

d. Ứng dụng vào các lĩnh vực khác:

- Áp dụng hệ thống vào các lĩnh vực khác như phân tích dư luận xã hội, khảo sát ý kiến khách hàng, hoặc theo dõi xu hướng thị trường.

Thông qua các hướng phát triển này, hệ thống không chỉ hoàn thiện hơn trong việc phân tích cảm xúc mà còn có thể trở thành một giải pháp toàn diện, hỗ trợ các doanh nghiệp và cá nhân trong nhiều bối cảnh khác nhau.

TÀI LIỆU THAM KHẢO

- [1] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python* (2019). Xem tại: <https://www.nltk.org/book/>. [Truy cập ngày: 28.09.2024].
- [2] Quoc Thai Nguyen, Thoai Linh Nguyen, Ngoc Hoang Luong and Quoc Hung Ngo, *Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews* (2020). Xem tại: <https://arxiv.org/pdf/2011.10426.pdf>. [Truy cập ngày: 21.09.2024].
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (2019). Xem tại: <https://arxiv.org/pdf/1810.04805.pdf>. [Truy cập ngày: 21.09.2024].
- [4] Ranjan Kumar Behera, Monalisa Jena, Santanu Kumar Rath, Sanjay Misra, *Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data* (2021). Xem tại: https://www.researchgate.net/publication/347379205_Co-LSTM_Convolutional_LSTM_model_for_sentiment_analysis_in_social_big_data [Truy cập ngày: 21.09.2024].