

ĐẠI HỌC ĐÀ NẴNG
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CÔNG NGHỆ THÔNG TIN

Tel. (84-511) 3736949, Fax. (84-511) 3842771
Website: itf.dut.edu.vn, E-mail: cntt@dut.udn.vn



ĐỒ ÁN TỐT NGHIỆP CỬ NHÂN
NGÀNH CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI :

XÂY DỰNG WEB SITE DỰA VÀO CÁC BÌNH LUẬN
ĐỂ ĐÁNH GIÁ SẢN PHẨM ĐIỆN TỬ

SINH VIÊN : Trần Văn Hải
MÃ SINH VIÊN : 102200167
LỚP : 20TCLC-DT4
CBHD : ThS. Nguyễn Văn Nguyên

Đà Nẵng, 06/2024

NHẬN XÉT PHẢN BIỆN ĐỒ ÁN TỐT NGHIỆP

I. Thông tin chung:

- Họ và tên sinh viên: Trần Văn Hải
- Lớp: 20TCLC-DT4
- Tên đề tài: Xây dựng website dựa vào các bình luận để đánh giá sản phẩm điện tử
- Người phản biện:..... Học hàm/ học vị:

Số thẻ SV: 102200167

II. Nhận xét, đánh giá đồ án tốt nghiệp:

TT	Các tiêu chí đánh giá	Điểm tối đa	Điểm đánh giá
	Sinh viên có phương pháp nghiên cứu phù hợp, giải quyết đủ nhiệm vụ đồ án được giao	80	
a	- Tính mới (nội dung chính của ĐATN có những phần mới so với các ĐATN trước đây). - Đề tài có giá trị khoa học, công nghệ; có thể ứng dụng thực tiễn.	15	
b	- Kỹ năng giải quyết vấn đề; hiểu, vận dụng được kiến thức cơ bản, cơ sở, chuyên ngành trong vấn đề nghiên cứu. - Chất lượng nội dung ĐATN (thuyết minh, bản vẽ, chương trình, mô hình,...).	50	
c	- Có kỹ năng vận dụng thành thạo các phần mềm ứng dụng trong vấn đề nghiên cứu; - Có kỹ năng đọc, hiểu tài liệu bằng tiếng nước ngoài ứng dụng trong vấn đề nghiên cứu; - Có kỹ năng làm việc nhóm;	15	
2	Kỹ năng viết:	20	
2a	Bố cục hợp lý, lập luận rõ ràng, chặt chẽ, lời văn súc tích	15	
2b	Thuyết minh đồ án không có lỗi chính tả, in ấn, định dạng	5	
3	Tổng điểm đánh giá theo thang 100:		
	Quy về thang 10 (lấy đến 1 số lẻ)		

- Các tồn tại, thiếu sót cần bổ sung, chỉnh sửa:

.....
.....

- Câu hỏi đề nghị sinh viên trả lời trong buổi bảo vệ:

.....
.....

- Đề nghị: ☐ Được bảo vệ đồ án ☐ Bổ sung để bảo vệ ☐ Không được bảo vệ

Đà Nẵng, ngày tháng năm 2023

Người phản biện

TÓM TẮT

1. Thông tin chung

Tên đề tài: Xây dựng website dựa vào các bình luận để đánh giá sản phẩm điện tử
Sinh viên thực hiện: Trần Văn Hải

Số thẻ SV: 102200167

Lớp: 20TCLC-DT4

Người hướng dẫn: ThS. Nguyễn Văn Nguyên

2. Tóm tắt đề án

Ngày nay, tin học ngày càng phát triển nhanh chóng và được ứng dụng rộng rãi trong mọi lĩnh vực đời sống xã hội, việc học và nắm bắt công nghệ mới đặc biệt là công nghệ thông tin ngày càng trở nên bức thiết. Đối với sinh viên trong ngành càng phải tích cực học tập, nắm vững kiến thức về công nghệ thông tin.

Với sự phát triển không ngừng của công nghệ, các sản phẩm liên quan đến công nghệ đang ngày một phát triển, nhu cầu sử dụng các sản phẩm điện tử trong đời sống ngày càng tăng. Nhu cầu tìm hiểu và mua các sản phẩm từ đó cũng tăng theo. Với xu hướng đó việc lựa chọn, đánh giá sản phẩm là một bài toán được đặt ra

NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

Họ tên sinh viên: Trần Văn Hải Số thẻ sinh viên: 102200167
Lớp: 20TCLC-DT4 Khoa: Công nghệ thông tin Ngành: Hệ thống thông tin

- Tên đề tài đồ án:
- Xây dựng website dựa vào các bình luận để đánh giá sản phẩm điện tử
Đề tài thuộc diện: ☐ Có ký kết thỏa thuận sở hữu trí tuệ đối với kết quả thực hiện
- Các số liệu và dữ liệu ban đầu:

Không có

- Nội dung các phần thuyết minh và tính toán:

Nội dung của thuyết minh gồm:

Mở đầu: Phần mở đầu của luận văn, giới thiệu về nhu cầu thực tế và lý do thực hiện đề tài, đồng thời giới thiệu sơ lược về đề tài và mục tiêu phải đạt được.

Chương 1. Tổng quan về đề tài: trình bày tổng quan về đề tài những lý thuyết học được và đã áp dụng vào hệ thống.

Chương 2. Thu thập và xử lý dữ liệu và đầu vào: trình bày việc crawl dữ liệu, tiền xử lý dữ liệu

Chương 3. Triển khai xây dựng mô hình và đánh giá kết quả: mô tả cách cài đặt, kịch bản chương trình và đánh giá kết quả đạt được.

Kết luận và hướng phát triển

- Các bản vẽ, đồ thị (ghi rõ các loại và kích thước bản vẽ):

Không có

- Họ tên người hướng dẫn: ThS Nguyễn Văn Nguyên
- Ngày giao nhiệm vụ đồ án: 4/4/2024
- Ngày hoàn thành đồ án: 29/6/2024

Đà Nẵng, ngày tháng năm 2024

Trưởng Bộ môn Công Nghệ Thông Tin

Người hướng dẫn

ThS Nguyễn Văn Nguyên

➤ LỜI NÓI ĐẦU

Sau một thời gian học tập và nghiên cứu, tôi đã quyết định thực hiện đề tài “**Xây dựng website dựa vào các bình luận để đánh giá sản phẩm điện tử**”. Dựa trên việc xử lý dữ liệu là các bình luận của các sản phẩm trên trang web thegioididong.com để đánh giá các bình luận của sản phẩm đó là tích cực hay tiêu cực từ đó đem lại cái nhìn khách quan về sản phẩm đó cho người dùng.

Đầu tiên, tôi xin gửi lời cảm ơn chân thành đến ThS Nguyễn Văn Nguyên, người đã dành thời gian quý báu giúp tôi thực hiện đồ án tốt nghiệp này. Bên cạnh đó, tôi cũng xin cảm ơn tập thể thầy cô khoa Công nghệ Thông tin thuộc trường Đại học Bách Khoa đã tận tâm dạy dỗ, bồi dưỡng kiến thức cho tôi, cũng như các bạn sinh viên khác trong suốt những năm học trên ghế giảng đường.

Do kiến thức và thời gian có hạn nên đồ án chắc chắn không tránh khỏi thiếu sót, mong các thầy cô chân thành góp ý để tôi có thể hoàn thiện hơn.

Tôi xin chân thành cảm ơn.

➤ CAM ĐOAN

Tôi xin cam đoan :

1. Nội dung trong đồ án này là do tôi thực hiện dưới sự hướng dẫn trực tiếp của Thầy Nguyễn Văn Nguyên
2. Các tham khảo dùng trong báo cáo đều được trích dẫn rõ ràng tên tác giả, tên công trình, thời gian, địa điểm công bố.
3. Nếu có những sao chép không hợp lệ, vi phạm, tôi xin chịu hoàn toàn trách nhiệm
Sinh viên thực hiện

Trần Văn Hải

➤ **MỤC LỤC**

➤ LỜI NÓI ĐẦU	5
➤ CAM ĐOAN	7
➤ MỤC LỤC	8
➤ DANH MỤC HÌNH ẢNH	9
➤ DANH MỤC BẢNG BIỂU	10
➤ MỞ ĐẦU	12
CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI	13
1.1 Giới thiệu đề tài:.....	13
1.2 Giới thiệu RNN:	13
1.3 Giới thiệu LSTM :	14
1.4 Giới thiệu Bi-LSTM:.....	14
1.5 Giới thiệu BERT:	14
CHƯƠNG 2. XỬ LÝ DỮ LIỆU VÀ HUẤN LUYỆN MÔ HÌNH	16
2.1 Thu thập dữ liệu	16
2.2 Xử lý dữ liệu:	20
2.3 Trực quan dữ liệu:	20
CHƯƠNG 3. TRIỂN KHAI VÀ ĐÁNH GIÁ KẾT QUẢ:	23
3.1 Phương pháp trích xuất đặc trưng	23
3.2 Xử lý vấn đề mất cân bằng dữ liệu.....	24
3.3 Xây dựng mô hình LSTM, Bi-LSTM và RNN	25
3.4 Xây dựng mô hình với BERT:	26
3.5 Đánh giá kết quả:.....	29
3.6 Kết quả thực nghiệm:	29
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	32
➤ TÀI LIỆU THAM KHẢO	34

➤ DANH MỤC HÌNH ẢNH

Hình 1. Mô hình RNN.....	13
Hình 2: Mô tả kiến trúc LSTM.....	14
Hình 3: Mô tả kiến trúc BiLSTM.....	14
Hình 4: Mô hình BERT	15
Hình 5: Web crawl data.....	16
Hình 6: Kết quả crawl data.....	17
Hình 7: Hình mô tả trigger load sản phẩm	18
Hình 8: Kết quả crawl data.....	18
Hình 9: Mô tả cách xem trigger đánh giá.....	19
Hình 10: Mô tả crawl data text.....	19
Hình 11: Đám mây dữ liệu data	20
Hình 12: Đám mây dữ liệu data text	21
Hình 13: Mô tả token count text positive	21
Hình 14: Mô tả token count text negative	22
Hình 15: Dữ liệu trước khi xử lý cân bằng mẫu.....	24
Hình 16: Dữ liệu sau khi xử lý cân bằng mẫu.....	24
Hình 17: Mô tả kiến trúc mạng phân loại sử dụng LSTM	25
Hình 18: Mô tả kiến trúc mạng phân loại sử dụng Bi-LSTM	25
Hình 19: Mô tả kiến trúc mạng phân loại sử dụng RNN.....	26
Hình 20: Transformer architecture	27
Hình 21: BERT base cho bài toán phân loại	28
Hình 22: Giao diện chính của trang web đánh giá sản phẩm	29
Hình 23: Giao diện đánh giá sản phẩm sau khi kiểm tra.....	30
Hình 24: Giao diện đánh giá sản phẩm sau khi lọc bình luận tiêu cực	30
Hình 25: Giao diện đánh giá sản phẩm sau khi lọc bình luận tích cực.....	31

➤ **DANH MỤC BẢNG BIỂU**

Bảng 1: Bảng trực quan hóa dữ liệu bình luận.....	20
Bảng 2: So sánh kết quả khi áp dụng kiến trúc LSTM, Bi-LSTM và RNN.....	26
Bảng 3: Bảng kết quả So sánh kết quả khi áp dụng BERT và PhoBert của VinAI.....	29

DANH SÁCH CỤM TỪ VIẾT TẮT

Từ viết tắt	Ý nghĩa
NLP	NLP, viết tắt của "Natural Language Processing" (Xử lý ngôn ngữ tự nhiên), là một lĩnh vực của trí tuệ nhân tạo (AI) liên quan đến việc tương tác giữa máy tính và ngôn ngữ tự nhiên của con người
LSTM	Là một loại đặc biệt của mạng nơ-ron hồi quy (RNN) được thiết kế để xử lý và dự đoán dữ liệu chuỗi thời gian với các phụ thuộc dài hạn
biLSTM	Là sự kết hợp hai mạng LSTM (Long Short-Term Memory) chạy song song nhưng theo hướng ngược nhau
RNN	Là một loại mạng nơ-ron được thiết kế để xử lý dữ liệu tuần tự hoặc chuỗi thời gian
BERT	Là một mô hình học sâu được phát triển bởi Google, được thiết kế để xử lý ngôn ngữ tự nhiên (NLP)
phoBERT	PhoBERT là một mô hình xử lý ngôn ngữ tự nhiên (NLP) dựa trên BERT, được thiết kế và tối ưu hóa cho tiếng Việt. Được phát triển bởi nhóm nghiên cứu VinAI Research

➤ MỞ ĐẦU

Trong thời đại công nghệ số ngày nay, việc mua sắm trực tuyến đã trở thành một phần không thể thiếu trong cuộc sống của chúng ta. Khách hàng không chỉ tìm kiếm sản phẩm mà họ cần mà còn dựa vào các đánh giá và bình luận từ người dùng khác để đưa ra quyết định mua hàng. Đặc biệt, đối với các sản phẩm điện tử - những mặt hàng có tính chất kỹ thuật cao và thường đi kèm với chi phí không nhỏ - việc tham khảo ý kiến và trải nghiệm của người dùng trước đó càng trở nên quan trọng.

Đề tài "Xây dựng website dựa vào các bình luận để đánh giá sản phẩm điện tử" ra đời nhằm giải quyết nhu cầu thiết yếu này của người tiêu dùng. Mục tiêu chính của đề tài là phát triển một nền tảng trực tuyến cho phép người dùng dễ dàng tiếp cận và đánh giá chất lượng của các sản phẩm điện tử thông qua việc tổng hợp và phân tích các bình luận, đánh giá từ cộng đồng. Bằng cách này, người tiêu dùng có thể có cái nhìn toàn diện hơn về sản phẩm, từ đó đưa ra quyết định mua sắm thông minh và hợp lý hơn.

Trong phần tiếp theo của đề tài, tôi sẽ trình bày chi tiết về phương pháp nghiên cứu, các công cụ và công nghệ được sử dụng, cũng như cách thức triển khai và đánh giá hiệu quả của hệ thống website. Tôi tin rằng, với sự phát triển của đề tài này, việc mua sắm các sản phẩm điện tử sẽ trở nên dễ dàng và đáng tin cậy hơn bao giờ hết.

CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI

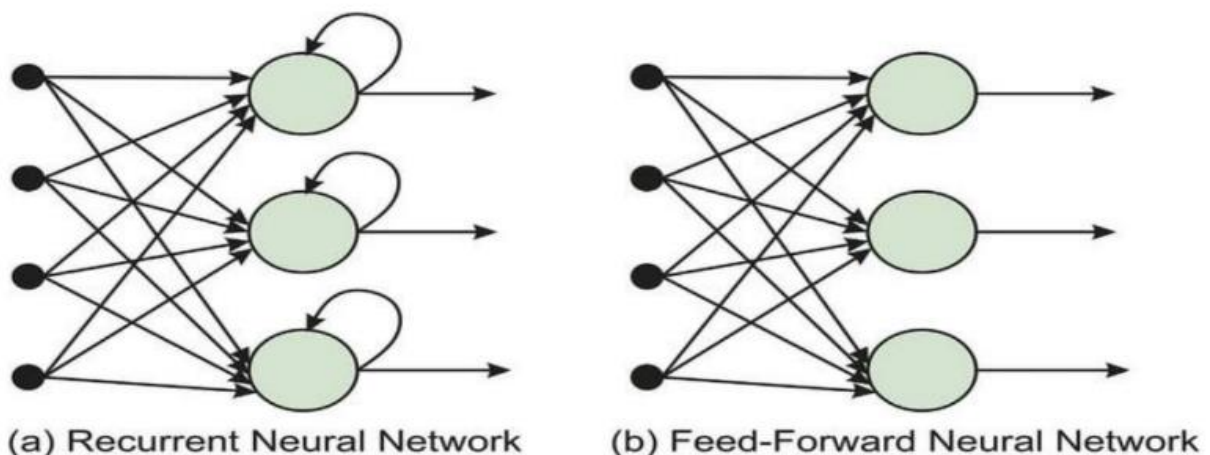
1.1 Giới thiệu đề tài

Với đề tài này vấn đề chính cần phải được giải quyết đó là việc xây dựng model để đánh giá sản phẩm qua bình luận để đạt được kết quả chính xác nhất, đó là một dạng của Natural Language Processing (NLP) : Xử lý ngôn ngữ tự nhiên là một lĩnh vực con của ngôn ngữ học, khoa học máy tính và trí tuệ nhân tạo liên quan đến các tương tác giữa máy tính và ngôn ngữ của con người, đặc biệt là cách lập trình máy tính để xử lý và phân tích lượng lớn dữ liệu ngôn ngữ tự nhiên. Mục tiêu là một máy tính có khả năng « hiểu » nội dung của các tài liệu, bao gồm các sắc thái ngữ cảnh của ngôn ngữ bên trong chúng.

Với đề tài này tôi sẽ phân loại cảm xúc người dùng dựa trên đánh giá của người dùng (Sử dụng 4 mô hình là LSTM+ BiLSTM + RNN + BERT)

1.2 Giới thiệu RNN

RNN là viết tắt của “Recurrent Neural Network” có thể dịch là “Mạng Nơ-ron Hồi Quy”. đây là một loại kiến trúc mạng nơ-ron trong lĩnh vực trí tuệ nhân tạo AI và học máy. RNN được thiết kế để xử lý dữ liệu chuỗi, nơi thông tin từ các bước thời gian trước đó được giữ lại để ảnh hưởng đến các bước thời gian sau.



Hình 1. Mô hình RNN

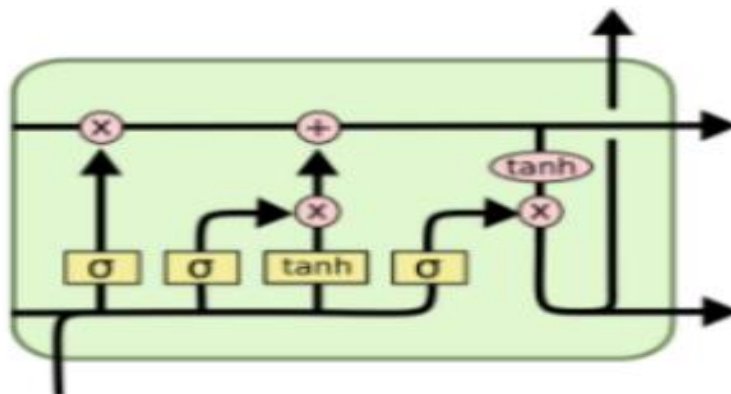
Mô hình RNN giữ một trạng thái ẩn (hidden state) và sử dụng nó để lưu trữ thông tin từ các bước thời gian trước đó. Điều này cho phép RNN hiểu và xử lý các mẫu trong dữ liệu chuỗi như trong các ứng dụng như dịch máy, nhận dạng giọng nói hoặc dự đoán chuỗi thời gian.

RNN được tạo thành từ các nơ-ron: các nút xử lý dữ liệu kết hợp cùng nhau để thực hiện các tác vụ phức tạp. Các nơ-ron được tổ chức dưới dạng lớp đầu vào, đầu ra và ẩn. Lớp đầu vào nhận thông tin để xử lý và lớp đầu ra cung cấp kết quả. Quá trình xử lý dữ liệu, phân tích và dự đoán diễn ra trong lớp ẩn.

Lớp ẩn của RNN chứa các trọng số (weights) và các tham số có thể được điều chỉnh trong quá trình huấn luyện để mô hình có thể học và bảo toàn thông tin quan trọng từ các bước thời gian trước đó.

1.3 Giới thiệu LSTM

Như đã đề cập ở trên, nhược điểm của mạng RNN là bị mất mát dữ liệu khi qua càng nhiều step. Từ đó mạng Long Short-term memory (LSTM) ra đời. Giống như tên gọi của nó, mạng LSTM có khả năng nhớ lâu hơn trong quá trình duyệt qua từng step.



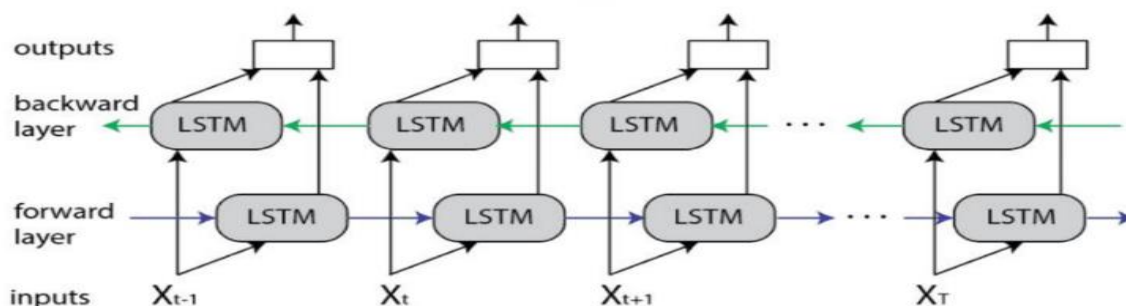
Hình 2: Mô tả kiến trúc LSTM

Để làm được việc đó, mỗi cell LSTM trang bị thêm cho mình cổng input giúp quyết định bao nhiêu thông tin đi vào, cổng forget giúp quyết định bao nhiêu thông tin sẽ cần quên đi và cổng output quyết định bao nhiêu thông tin đi ra. Từ việc thông tin được chọn lọc này giúp nó được lưu giữ lâu hơn so với kiến trúc RNN cũ.

1.4 Giới thiệu Bi-LSTM

LSTM hai chiều (BiLSTM) là mạng thần kinh tái phát được sử dụng chủ yếu trong xử lý ngôn ngữ tự nhiên. Không giống như LSTM tiêu chuẩn, đầu vào chảy theo cả hai hướng và có khả năng sử dụng thông tin từ cả hai phía. Nó cũng là một công cụ mạnh mẽ mô hình hóa sự phụ thuộc tuần tự giữa các từ và cụm từ theo hai hướng của chuỗi.

Tóm lại, BiLSTM bổ sung thêm một lớp LSTM, lớp này đảo ngược hướng của luồng thông tin. Tóm lại, điều đó có nghĩa là chuỗi đầu vào chảy ngược lại trong lớp LSTM bổ sung. Sau đó, tôi kết hợp các đầu ra từ cả hai lớp LSTM theo nhiều cách, chẳng hạn như trung bình, tổng, nhân hoặc ghép.



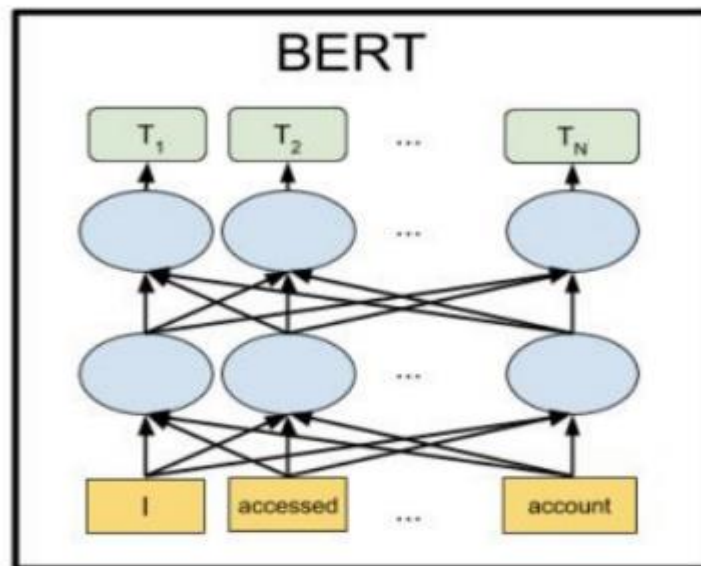
Hình 3: Mô tả kiến trúc BiLSTM

1.5 Giới thiệu BERT

BERT là mô hình được thiết kế để đào tạo ra các vector đại diện cho ngôn ngữ

văn bản thông qua ngữ cảnh 2 chiều(trái và phải) của chúng. Kết quả là, vector đại diện được sinh ra từ mô hình BERT được tính chỉnh với các lớp đầu ra bổ sung đã tạo ra nhiều kiến trúc cải tiến đáng kể phục vụ cho nhiều tác vụ phức tạp hơn

Đầu tiên, ta cùng xem xét tới kiến trúc của mô hình BERT.3.4



Hình 4: Mô hình BERT

Nói một cách dễ hiểu thì BERT chính là lấy ra phần Encoder từ kiến trúc Transformer. Như hình trên, input là các câu văn bản và output là các vector từ embedding[2]

Trong bài báo cáo này, tôi tập trung làm việc với bài toán phân loại nên chỉ tập trung vào kiến trúc BERT dành cho bài toán phân loại bằng phương pháp fine-tuning lại kiến trúc BERT sử dụng phiên bản BERT base.

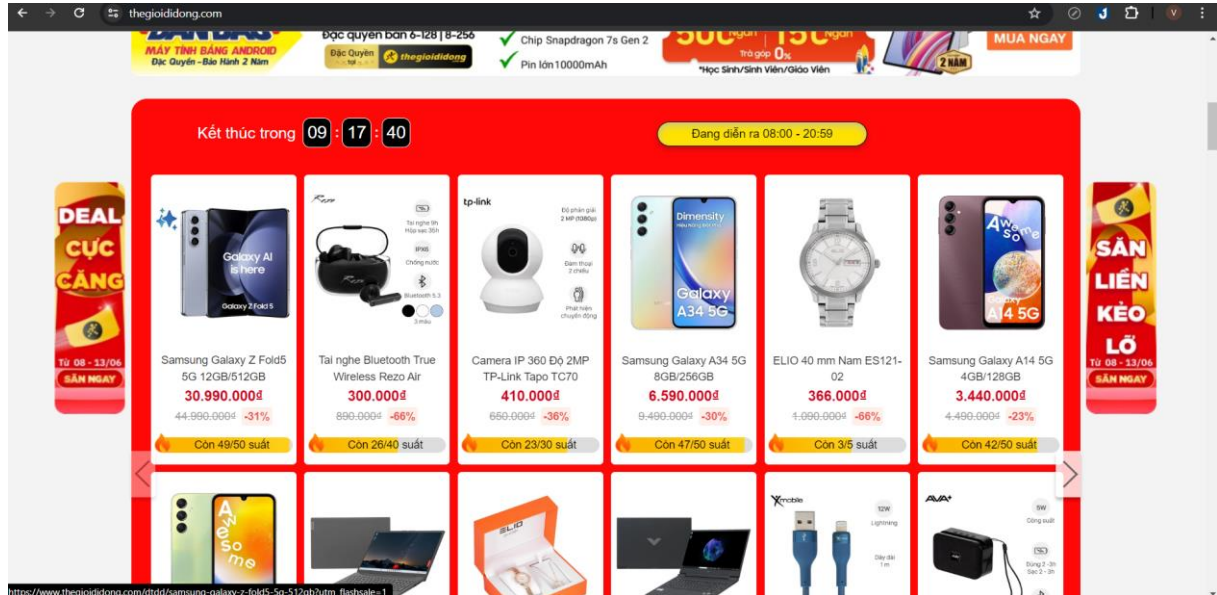
Để fine-tuning lại kiến trúc của BERT cho tác vụ phân loại văn bản. Chúng ta giữ nguyên kiến trúc encoder của transformer và sau đó trích xuất ra biểu diễn vector của token CLS đánh dấu vị trí đầu tiên. Vector này sẽ được sử dụng làm đầu vào cho thuật toán phân loại bằng cách thêm một linear projection layer (cũng chính là fully connected layer) ở cuối có kích thước bằng với số classes cần phân loại.

CHƯƠNG 2. XỬ LÝ DỮ LIỆU VÀ HUẤN LUYỆN MÔ HÌNH

2.1 Thu thập dữ liệu

Thu thập dữ liệu từ trang web : Thegioididong

Dựa vào nhu cầu bài toán ta cần thu thập bộ dữ liệu bình luận của khách hàng theo từng sản phẩm



Hình 5: Web crawl data

Đánh giá sơ bộ về trang web ta bắt đầu truy cập từ base url của web, ta có thể tìm ra được quy trình crawl dữ liệu được theo những bước sau

Bước 1: Sử dụng Selenium lấy toàn bộ source về, tìm tới thẻ div có class ‘main-menu’. Bằng cách này chúng ta có thể crawl được toàn bộ category của các sản phẩm và hơn nữa việc nếu trang web có cập nhật thêm category mới thì vẫn có thể tìm được category đó.

```

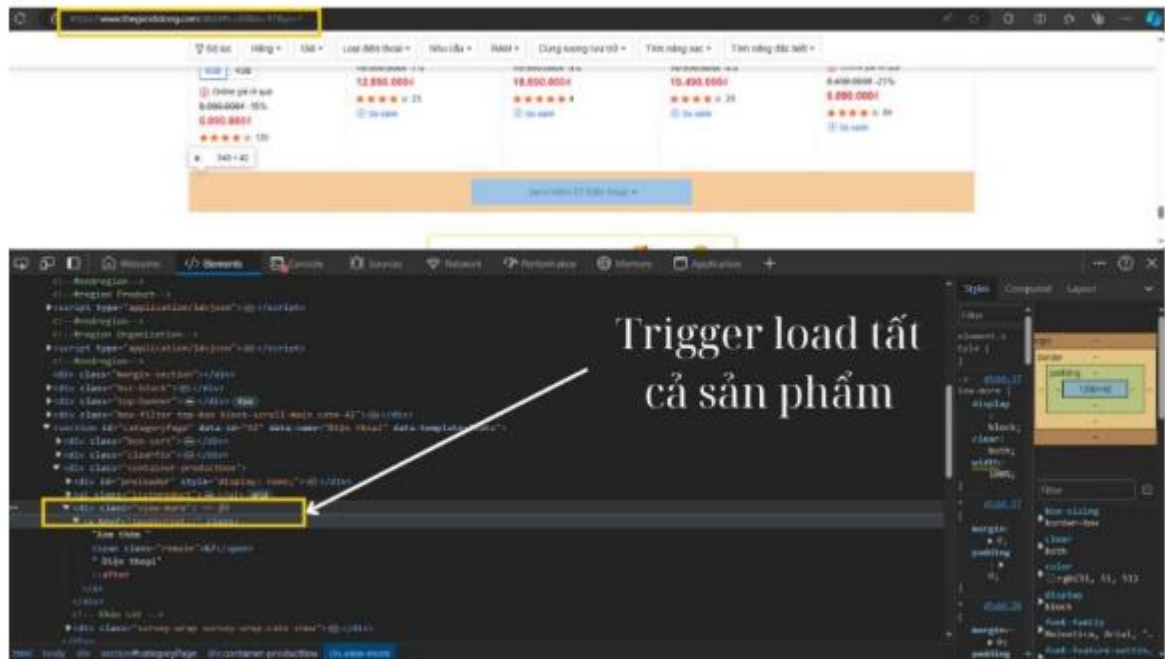
1  category, full link
2  dtdd, https://www.thegioididong.com/dtdd
3  may-tinh-bang, https://www.thegioididong.com/may-tinh-bang
4  phu-kien, https://www.thegioididong.com/phu-kien
5  sac-dtdd, https://www.thegioididong.com/sac-dtdd
6  sac-cap, https://www.thegioididong.com/sac-cap
7  hub-chuyen-doi, https://www.thegioididong.com/hub-chuyen-doi
8  airtag, https://www.thegioididong.com/airtag
9  phu-kien-thong-minh, https://www.thegioididong.com/phu-kien-thong-minh
10 tai-nghe-bluetooth, https://www.thegioididong.com/tai-nghe-bluetooth
11 tai-nghe-co-day, https://www.thegioididong.com/tai-nghe-co-day
12 micro-cac-loai, https://www.thegioididong.com/micro-cac-loai
13 camera-giam-sat?g=camera-giam-sat, https://www.thegioididong.com/camera-giam-sat?g=camera-giam-sat
14 camera-giam-sat-camera-ngoai-troi, https://www.thegioididong.com/camera-giam-sat-camera-ngoai-troi
15 may-chieu, https://www.thegioididong.com/may-chieu
16 phu-kien/apple, https://www.thegioididong.com/phu-kien/apple
17 phu-kien/samsung, https://www.thegioididong.com/phu-kien/samsung
18 phu-kien/imou, https://www.thegioididong.com/phu-kien/imou
19 phu-kien/baseus, https://www.thegioididong.com/phu-kien/baseus
20 phu-kien/jbl, https://www.thegioididong.com/phu-kien/jbl
21 phu-kien/anker, https://www.thegioididong.com/phu-kien/anker
22 phu-kien/xmobile, https://www.thegioididong.com/phu-kien/xmobile
23 chuot-may-tinh, https://www.thegioididong.com/chuot-may-tinh
24 ban-phim, https://www.thegioididong.com/ban-phim
25 thiet-bi-mang, https://www.thegioididong.com/thiet-bi-mang
26 phan-mem, https://www.thegioididong.com/phan-mem

```

Hình 6: Kết quả crawl data

Bước 2: Tiếp tục truy cập vào toàn bộ category, chúng ta có thể nhận thấy được rằng để có thể tìm ra tất cả sản phẩm nó phục thuộc vào 2 điều kiện: 1 là mục đó phải có sản phẩm, 2 là không có quảng cáo. Sau đó chúng ta sử dụng action để có thể trigger load tất cả sản phẩm.

Sau mỗi lần trigger chúng ta sẽ có được query của API như hình dưới. Tiếp tục trigger cho đến khi không thể trigger được nữa ta sẽ có được end_page_link của tất cả sản phẩm nằm trong category đó.

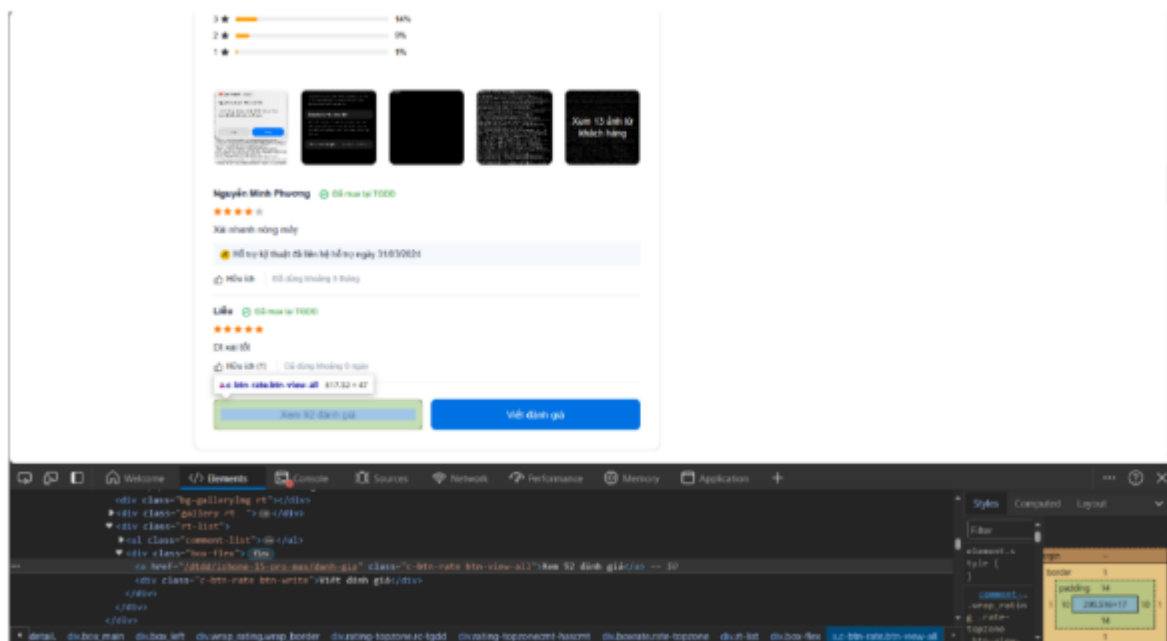


Hình 7: Hình mô tả trigger load sản phẩm



Hình 8: Kết quả crawl data

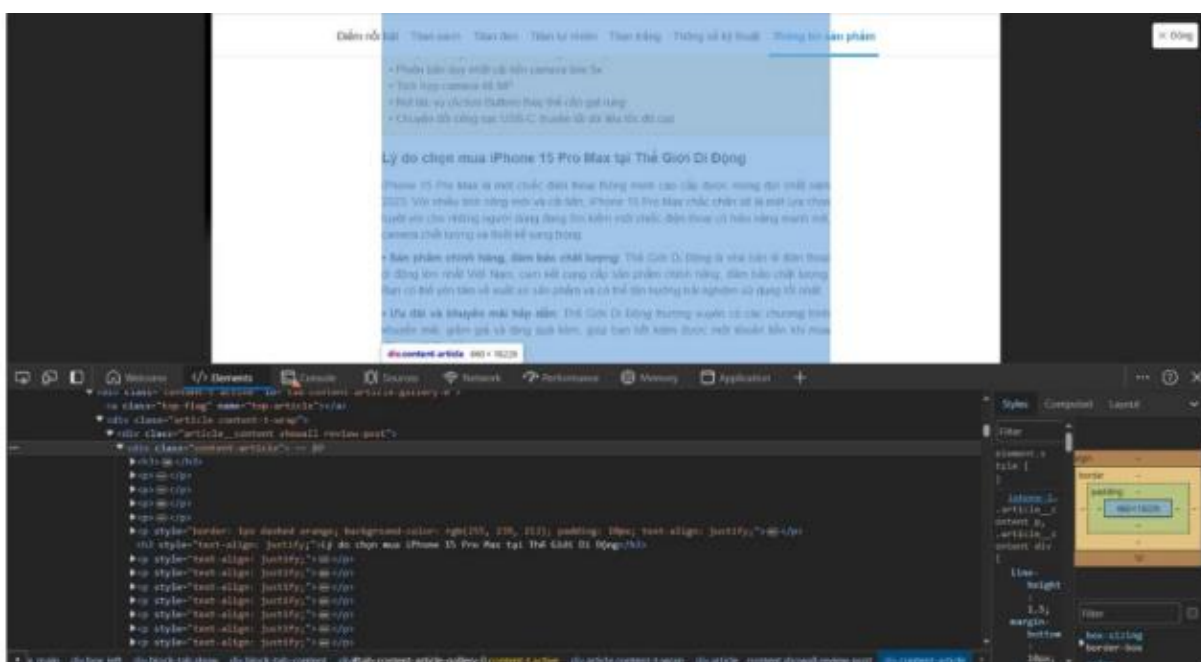
Bước 4: Toàn bộ sản phẩm được lưu tại 1 csv để phục vụ cho việc crawl mô tả và bình luận. Sử dụng các tương tự crawl end page link, ta trigger nút ‘Xem tất cả đánh giá’ để đi tới trang danh sách các đánh giá và tiếp tục cộng thêm trang tại url của trang đánh giá. Nếu tồn tại link thì url sẽ direct tới trang tiếp theo ngược lại nó sẽ được throw ra Exception và chúng ta có thể hoàn tất crawl data comment 1 sản phẩm.



Hình 9: Mô tả cách xem trigger đánh giá

Đối với mô tả sản phẩm, tìm tới thẻ ‘div’ có class ‘content-article’, sau đó tìm tất cả các thẻ p và thẻ span tùy vào nội dung của thẻ, mỗi thẻ sẽ có thông số hoặc mô tả khác nhau vì vậy chúng ta chỉ nên lấy những thẻ có số lượng từ trên 50 ký tự.

Hình 10: Mô tả crawl data text



2.2 Xử lý dữ liệu

Các bước tiền xử lý dữ liệu như sau:

Bước 1: Loại bỏ các kí tự đặc biệt và Lowercase.

- Loại bỏ các ký tự đặc biệt như !,\$,%,*,),... và chuyển tất cả chữ thành chữ thường

Bước 2: Loại bỏ số và các chữ dính số:

- Loại bỏ các số và các chữ dính số như (2e, 10as,...)

Bước 3: Thay thế các từ viết tắt thành các từ có nghĩa:

- Thay thế các từ viết tắt thành từ có nghĩa

Bước 4: Sử dụng model để Generate text và sửa lỗi chính tả:

- Sử dụng model bmd1905/vietnamese-correction để sửa lỗi chính tả

Bước 5: Tiếp tục xoá dấu câu, xoá từ dừng và Lowercase lại và segment:

- Xóa các dấu câu như .,(,)" và xóa các từ dừng trong bộ stop_word_vn

2.3 Trục quan dữ liệu

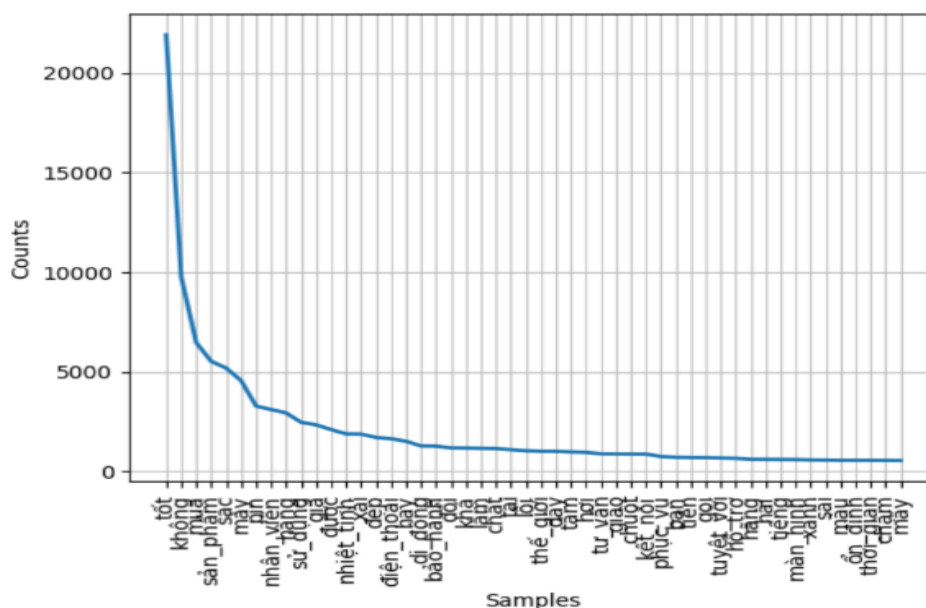
Số star đếm được sau khi xử lý dữ liệu:

Star	Count
1	1362
2	1574
3	2143
4	2905
5	7305

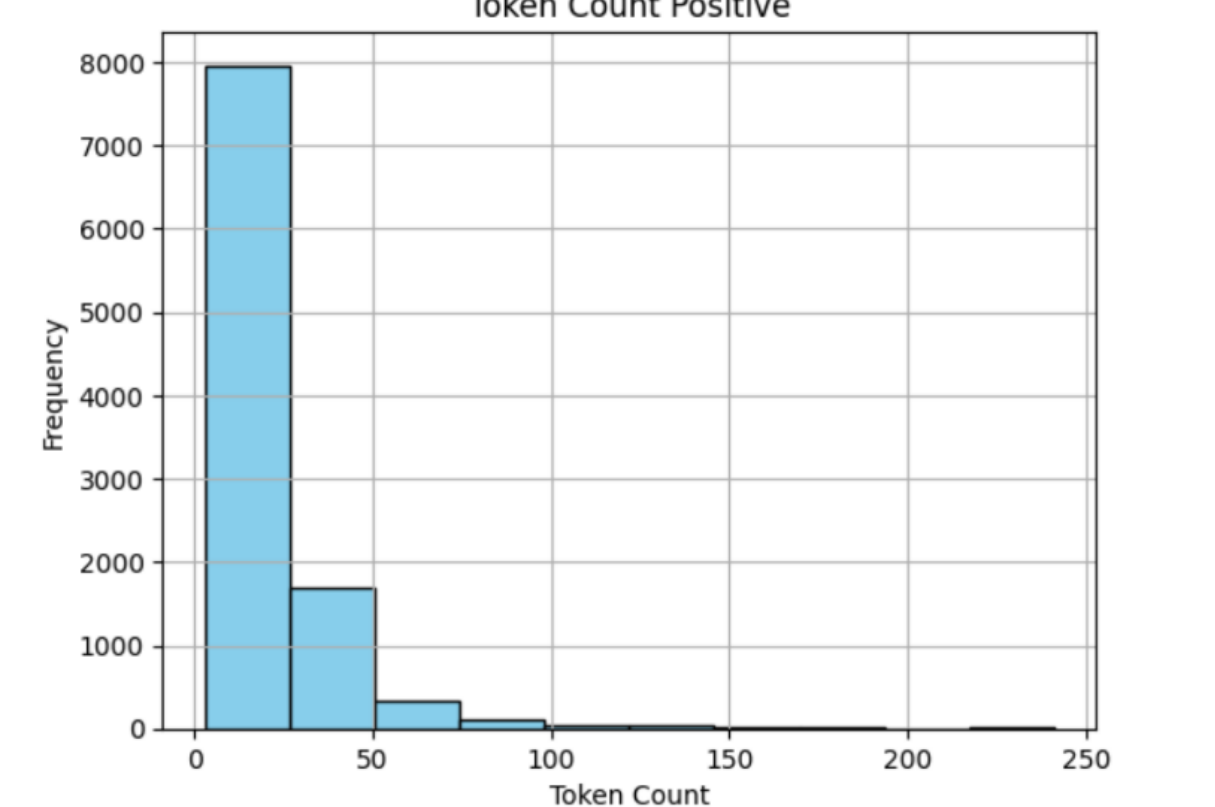
Bảng 1: Bảng trục quan hóa dữ liệu bình luận

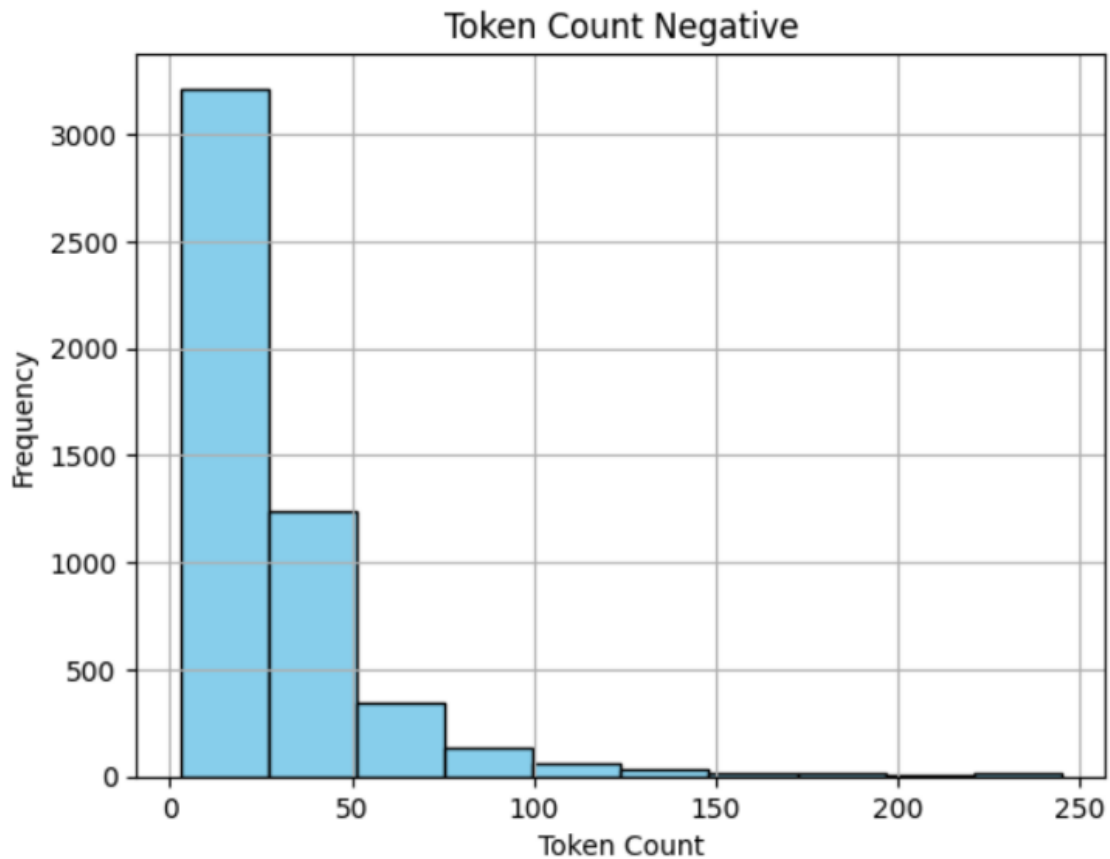
Đánh giá sơ bộ ta thấy rằng số sao biểu thị cho từng loại cảm xúc đang chênh lệch rất lớn dẫn tới khi mà train model cần cân bằng mẫu.

Dưới đây là đám mây từ biểu thị số từ xuất hiện của tập dữ liệu:



Hình 11: Đám mây dữ liệu data





Hình 14: Mô tả token count text negative

Đánh giá sơ bộ về tập dữ liệu sau khi tiền xử lý ta thấy rằng, cả Pos, Neg đều có số từ ngữ rơi vào khoảng 0 – 25 từ. Đối với Neg thì ta thấy có 1 chút nhỉnh hơn về số lượng từ trong câu từ 25 – 50 từ/ câu. Ta rút ra được đánh giá rằng các Neg này thể hiện được chính xác phản ánh của khách hàng về sự không hài lòng và tác trức của sản phẩm. Còn đối với Pos thì hoàn toàn nghiêng về sự hài lòng “Tốt, được, sản phẩm ổn...” chính vì thế các Pos này có giá trị với tập dữ liệu.

CHƯƠNG 3. TRIỂN KHAI VÀ ĐÁNH GIÁ KẾT QUẢ:

3.1 Phương pháp trích xuất đặc trưng

Trước khi đi đến phương pháp trích xuất đặc trưng, chúng ta cần hiểu rõ một số khái niệm:

- **Nhiệm 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. Mỗi bộ dữ liệu tương ứng với một loại tác NLP vụ như: Phân tích cảm xúc (Sentiment Analysis), hỏi đáp (Question and Answering), dự báo câu tiếp theo (NSP - Next Sentence Prediction), nhận diện thực thể trong câu (NER - Name Entity Recognition), suy luận ngôn ngữ tự nhiên (NLI - Natural Language Inference).

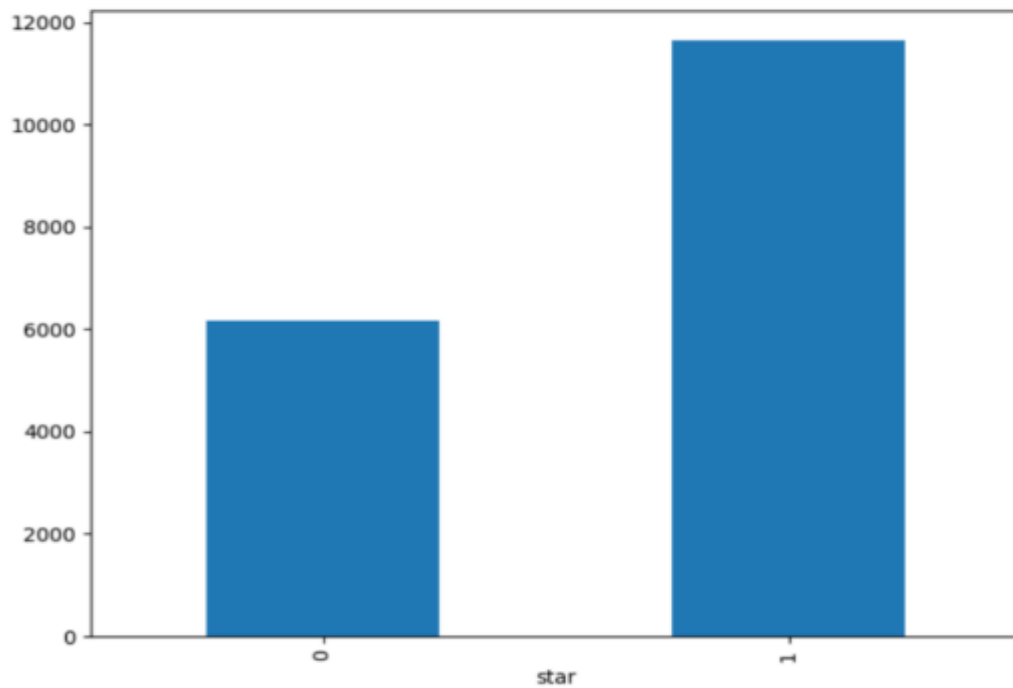
Ở đề tài này chúng em sử dụng BERT để biểu diễn ngôn ngữ. Lí do:

- Tăng GLUE score (General Language Understanding Evaluation score), một chỉ số tổng quát đánh giá mức độ hiểu ngôn ngữ lên 80.5%
- Tăng accuracy trên bộ dữ liệu MultiNLI đánh giá tác vụ quan hệ văn bản (text entailment) lên 86.7%.
- Tăng accuracy F1 score trên bộ dữ liệu SQuAD v1.1 đánh giá tác vụ question and answering lên 93.2%.
- Bi-directional (hai chiều): Ngữ nghĩa của một từ không chỉ được biểu diễn bởi những từ liền trước mà còn được giải thích bởi toàn bộ các từ xung quanh. Luồng giải thích tuân theo đồng thời từ trái qua phải và từ phải qua trái cùng một lúc. Đại diện cho các phép biểu diễn từ này là những mô hình sử dụng kỹ thuật transformer.
- Mô hình BERT ứng dụng kỹ thuật Transformer. Mà các comment của khách hàng sẽ cần biểu diễn để hiểu ngữ cảnh của câu.

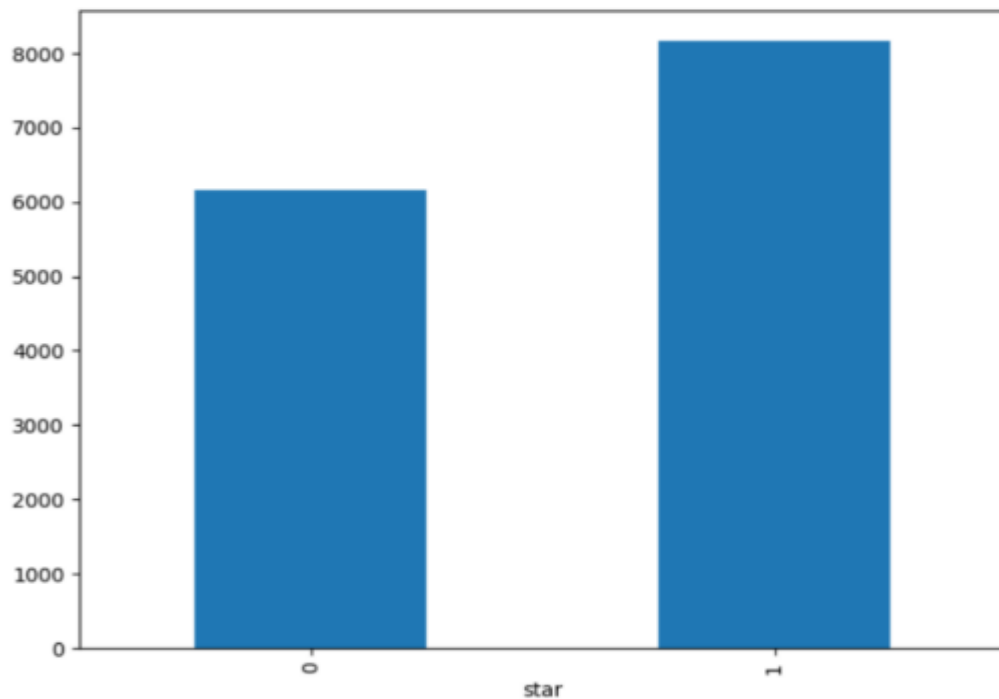
Ở thời điểm hiện tại, BERT đã được ứng dụng cho Tiếng Việt. Tôi sẽ sử dụng PhoBERT của VinAI về huấn luyện trước biểu diễn từ (pre-train word embedding) sử dụng model BERT.

3.2 Xử lý vấn đề mất cân bằng dữ liệu

Trong quá trình thu thập dữ liệu do số lượng nhãn Tích cực chiếm nhiều so với Tiêu cực nên tôi thực hiện loại bỏ 4000 nhãn Tích cực



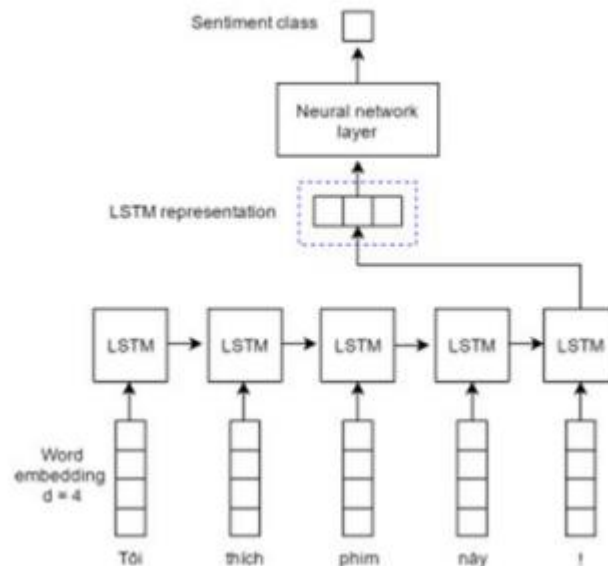
Hình 15: Dữ liệu trước khi xử lý cân bằng mẫu



Hình 16: Dữ liệu sau khi xử lý cân bằng mẫu

3.3 Xây dựng mô hình LSTM, Bi-LSTM và RNN

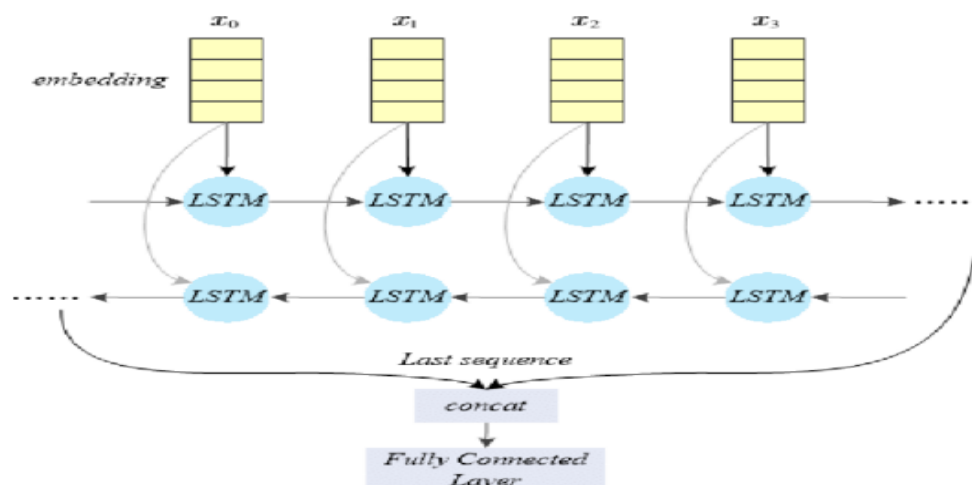
Chúng em đề xuất thử nghiệm với kiến trúc LSTM và các kiến trúc liên quan nhằm xây dựng mô hình cơ bản có khả năng phân loại 2 lớp cho các review.



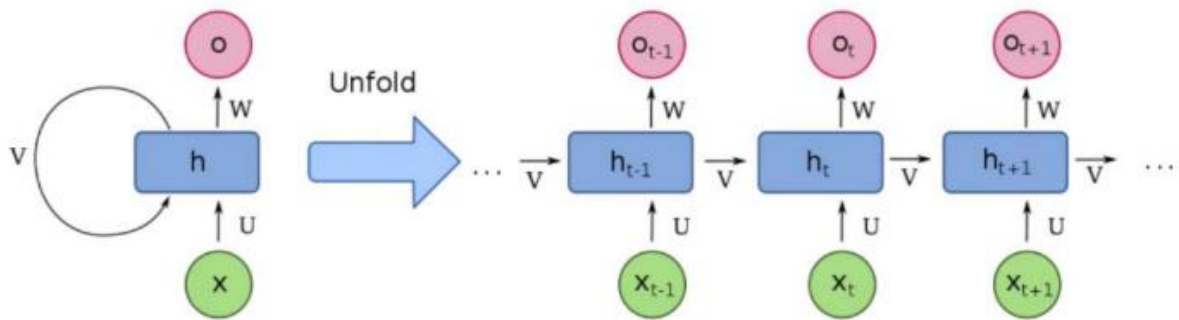
Hình 17: Mô tả kiến trúc mạng phân loại sử dụng LSTM

Dữ liệu văn bản là các từ sau khi mô hình hóa thành các vector, được đưa vào cell LSTM lần lượt và lấy output của cell cuối cùng được dùng để cho qua lớp fully connected, biến đổi thành đầu ra gồm 2 nút thể hiện cho xác suất của 2 lớp cần phân loại.

Chúng em thử nghiệm mạng Bi-LSTM với kiến trúc tương tự. Tuy nhiên ở đây chúng ta có 2 lớp LSTM chạy ngược nhau và feature dùng để cho qua lớp fullyconnect là sự kết hợp giữa 2 output cuối cùng mỗi lớp LSTM. Mạng Bi-LSTM giúp mô hình không chỉ học theo trình tự tự nhiên mà còn học văn bản theo trình tự ngược lại.



Hình 18: Mô tả kiến trúc mạng phân loại sử dụng Bi-LSTM



Hình 19: Mô tả kiến trúc mạng phân loại sử dụng RNN

Mạng RNN lấy nguyên tắc chung của mạng thần kinh chuyển tiếp nguồn cấp dữ liệu và cho phép chúng xử lý dữ liệu tuần tự bằng cách cung cấp cho mô hình một bộ nhớ trong. Phần “Recurrent” của tên RNN xuất phát từ thực tế là vòng lặp đầu vào và đầu ra. Sau khi đầu ra của mạng được tạo ra, đầu ra sẽ được sao chép và trả về mạng làm đầu vào. Khi đưa ra quyết định, không chỉ đầu vào và đầu ra hiện tại được phân tích mà đầu vào trước đó cũng được xem xét. Nói cách khác, nếu đầu vào ban đầu của mạng là X và đầu ra là H, thì cả H và X₁ (đầu vào tiếp theo trong chuỗi dữ liệu) đều được đưa vào mạng cho vòng học tiếp theo. Bằng cách này, bối cảnh của dữ liệu (đầu vào trước đó) được giữ nguyên khi mạng huấn luyện.

Kết quả của kiến trúc này là RNN có khả năng xử lý dữ liệu tuần tự. Tuy nhiên, RNN gặp phải một số vấn đề. RNN bị ảnh hưởng bởi biến mất gradient và bùng nổ các vấn đề về gradient. Độ dài của các chuỗi mà RNN có thể diễn giải khá hạn chế, đặc biệt là so với LSTM.

Quá trình huấn luyện sử dụng thuật toán tối ưu hóa AdamW, learning-rate khởi tạo 2e-5 và số lượng epoch là 5

Hàm mất mát trong lần thử nghiệm này là hàm loss Cross Entropy với công thức:

$$E = - \sum_{c=1}^M y_{o,c} \log(p_{o,c})$$

Dưới đây là kết quả các lần thực nghiệm trên mô hình LSTM, Bi-LSTM, RNN

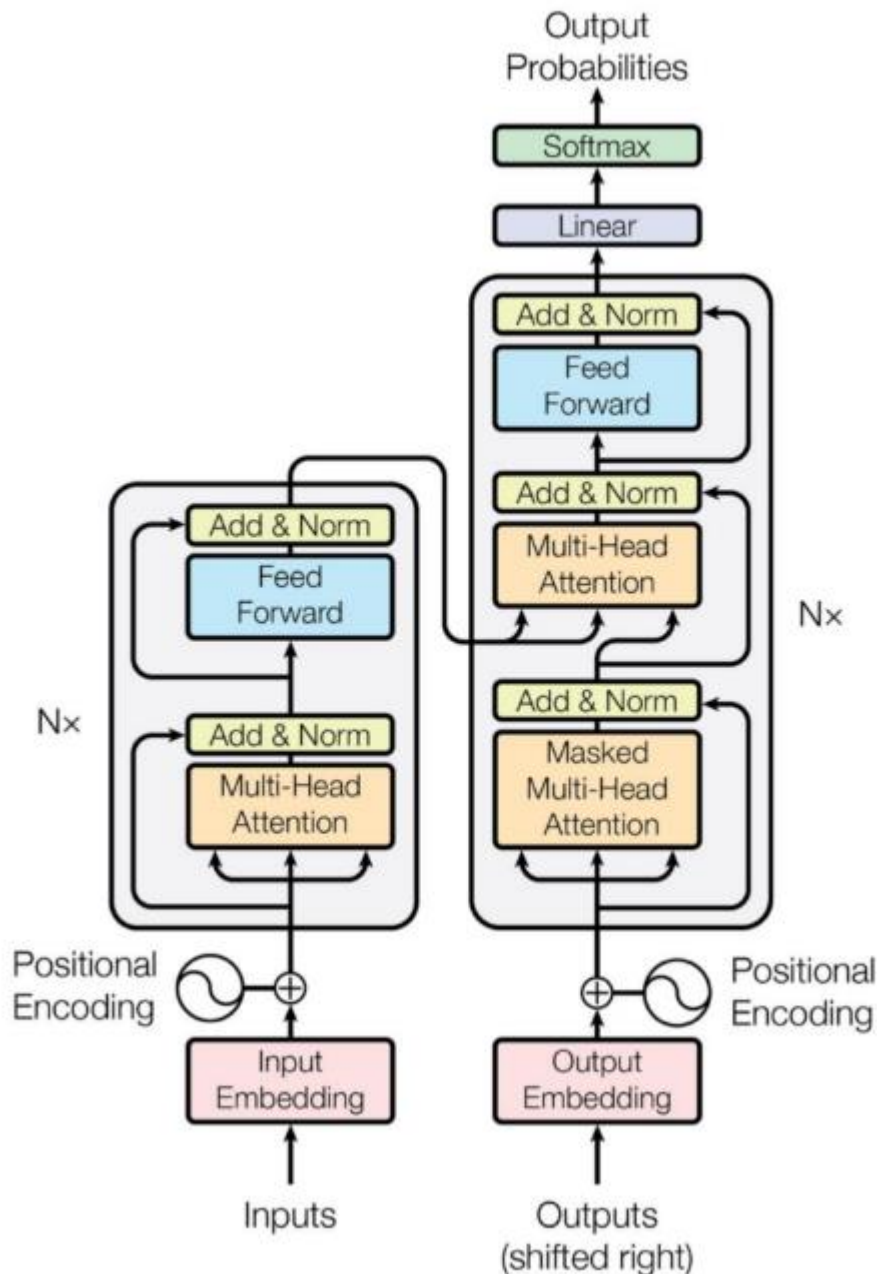
	LSTM	Bi-LSTM	RNN
F1 score with Label = Tiêu cực	0.86	0.86	0.85
F1 score with Label = Tích cực	0.88	0.88	0.89
Accuracy	0.86	0.86	0.86
F1 score macro avg	0.86	0.86	0.86

Bảng 2: So sánh kết quả khi áp dụng kiến trúc LSTM, Bi-LSTM và RNN

3.4 Xây dựng mô hình với BERT

Tiếp theo tôi thử nghiệm với kiến trúc Transformer mà cụ thể là phoBERT với BERT là viết tắt của Bidirectional Encoder Representations of Transformer và được huấn luyện riêng cho ngôn ngữ tiếng việt. Để hiểu kiến trúc của phoBERT thì trước tiên ta cần hiểu về BERT.

BERT là mô hình được thiết kế để đào tạo ra các vector đại diện cho ngôn ngữ văn bản thông qua ngữ cảnh 2 chiều(trái và phải) của chúng. Kết quả là, vector đại diện được sinh ra từ mô hình BERT được tính chnh với các lớp đầu ra bổ sung đã tạo ra nhiều kiến trúc cải tiến đáng kể phục vụ cho nhiều tác vụ phức tạp hơn.

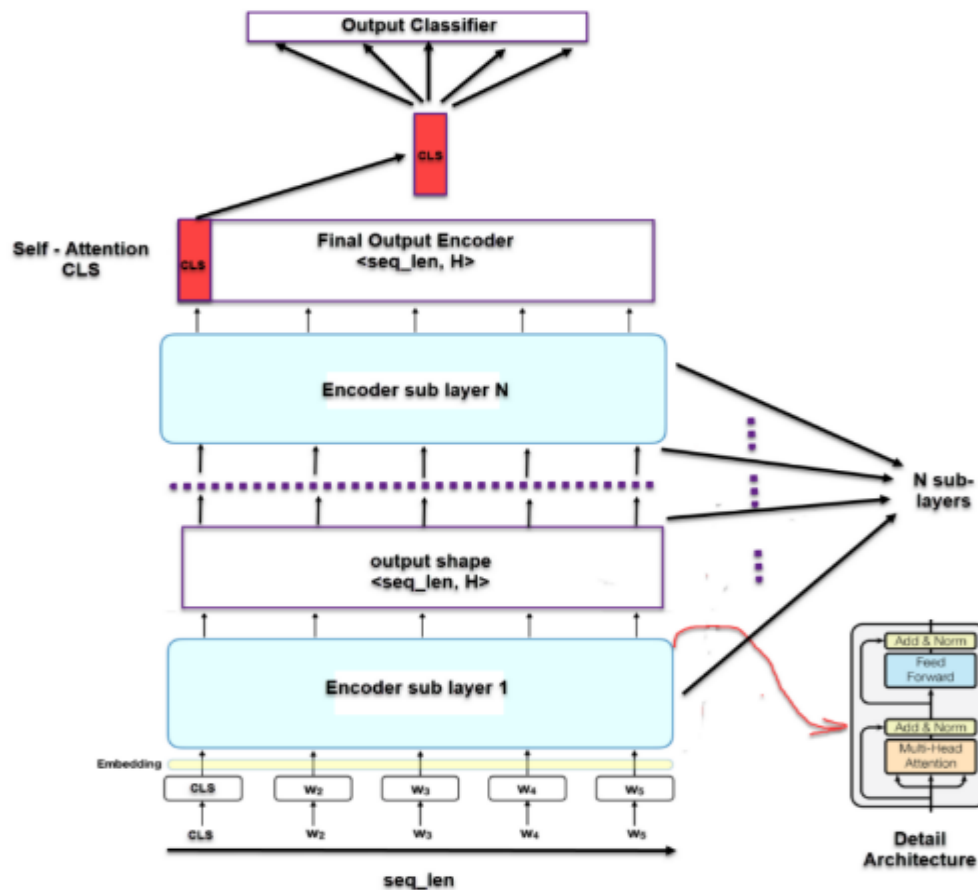


Hình 20: Transformer architecture

BERT chính là lấy ra phần Encoder từ kiến trúc Transformer. Input là các câu văn bản và output là các vector từ embedding

Trong bài báo cáo này, tôi tập trung làm việc với bài toán phân loại nên chỉ tập trung vào kiến trúc BERT dành cho bài toán phân loại bằng phương pháp fine-tuning lại kiến trúc BERT sử dụng phiên bản BERT base.

Để fine-tuning lại kiến trúc của BERT cho tác vụ phân loại văn bản. Chúng ta giữ nguyên kiến trúc encoder của transformer và sau đó trích xuất ra biểu diễn vector của token CLS đánh dấu vị trí đầu tiên. Vector này sẽ được sử dụng làm đầu vào cho thuật toán phân loại bằng cách thêm một linear projection layer (cũng chính là fully connected layer) ở cuối có kích thước bằng với số classes cần phân loại. Ngoài ra, để tránh overfitting và chi phí tính toán quá nhiều, tôi đã sử dụng thêm 1 lớp dropout và chuẩn hóa bộ trọng số.



Hình 21: BERT base cho bài toán phân loại

- sub-layers: 12
- hidden layers: 762
- trainable parameters: 110M
- n-classes: 2

Hàm mất mát trong lần thử nghiệm này là hàm loss Cross Entropy với công thức:

$$E = - \sum_{c=1}^M y_{o,c} \log(p_{o,c})$$

Quá trình huấn luyện sử dụng thuật toán tối ưu AdamW, learning-rate khởi tạo là 2e-5, chọn từ những giá trị được khuyến dùng từ BERT: 5e-5, 2e-5, 3e-5 và batch-size = 16, chọn từ những giá trị được khuyến dùng từ BERT: 16, 32. Quá trình huấn luyện thực hiện trên 5 epochs.

Dưới đây là kết quả thực nghiệm trên mô hình phoBERT.

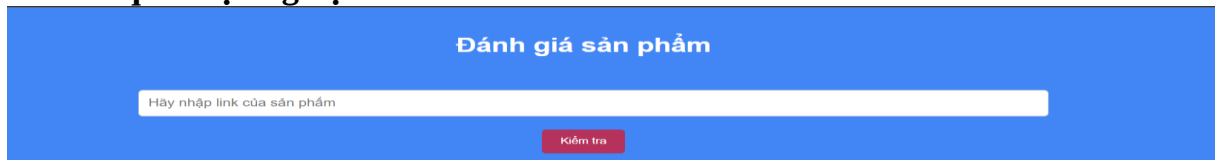
	PhoBert
F1 score with Label = Tiêu cực	0.87
F1 score with Label = Tích cực	0.89
Accuracy	0.88
F1 score macro avg	0.88

Bảng 3: Bảng kết quả So sánh kết quả khi áp dụng BERT và PhoBert của VinAI

3.5 Đánh giá kết quả

Qua các thực nghiệm trên, tôi đề xuất sử dụng mô hình phoBert, ứng dụng kiến trúc transformer cho bài toán phân tích cảm xúc review với kết quả thực nghiệm đạt accuracy 0.88 trên 2 lớp. Tuy nhiên có thể thấy độ chính xác vẫn chưa cao, lí do vì lượng dataset còn nhỏ vì phải loại bỏ bớt các lớp chiếm nhiều hơn, comment còn nhiều kí tự viết tắt chưa chuẩn hóa hết, đánh giá người dùng chỉ mang tính tham khảo. Vì vậy cần thu thập thêm dữ liệu từ nhiều nguồn để đảm bảo tích khách quan.

3.6 Kết quả thực nghiệm



Hình 22: Giao diện chính của trang web đánh giá sản phẩm

Đánh giá sản phẩm

https://www.thegioididong.com/dtdd/iphone-15-pro-max

Kiểm tra



Điện thoại iPhone 15 Pro Max

Giá Thành: 27.999.000 VND

Tổng số bình luận: 10

Điểm đánh giá trên web: 3/5 ★

Tất cả đánh giá

Tắt cả bình luận ▼

Người bình luận	Bình luận	Điểm đánh giá trên web	Đánh giá
Nguyễn Đình Hới	Ok	5 ★	Tích Cực
Minh	sản phẩm ok	5 ★	Tích Cực
Trương An	Em mới mua hồi 26/04 giờ mún bán lại thì còn được bao nhiêu ạ	5 ★	Tích Cực

Hình 23: Giao diện đánh giá sản phẩm sau khi kiểm tra

https://www.thegioididong.com/dtdd/iphone-15-pro-max

Kiểm tra



Điện thoại iPhone 15 Pro Max

Giá Thành: 27.999.000 VND

Tổng số bình luận: 10

Điểm đánh giá trên web: 3/5 ★

Bình Luận Tiêu Cực

Bình luận Tiêu Cực ▼

Người bình luận	Bình luận	Điểm đánh giá trên web	Đánh giá
Anh Khanh	Máy nóng rất tệ ,call cideo như muốn nổ ,đang nt zalo cái tắt nguồn,nhanh hết pin nữa đem ra thì bảo gửi về hãng ktra bảo không thấy nóng hay gì còn ktra bán hàng kiểu này chắc không có lại lần t2	1 ★	Tiêu Cực
Nhung	Mới mua 3 ngày. Chụp hình thời mà máy nóng đã man. Sạc 80 % đứng lại ko sạc nữa. Pin tụt nhanh quá	3 ★	Tiêu Cực
Võ Thị Diệu My	Lúc mua thì bảo lỗi 1 doi 1 tới lúc máy lỗi thì đi tham địnhh 10 ngày chưa xong. Mua thì nhanh còn doi thì rất lâu cũng k biết có doi dc k 15pro max xài chưa dc 30 ngày máy thì nóng muốn nổ tung.	1 ★	Tiêu Cực

Hình 24: Giao diện đánh giá sản phẩm sau khi lọc bình luận tiêu cực

https://www.thegioididong.com/dtdd/iphone-15-pro-max

Kiểm tra



Điện thoại iPhone 15 Pro Max

Giá Thành: 27.999.000 VND

Tổng số bình luận: 10

Điểm đánh giá trên web: 3/5 ★

Bình luận tích cực

Bình luận tích cực ▾

Người bình luận	Bình luận	Điểm đánh giá trên web	Đánh giá
Nguyễn Đình Hời	Ok	5 ★	Tích Cực
Minh	sản phẩm ok	5 ★	Tích Cực
Trường An	Em mới mua hồi 26/04 giờ mún bán lại thì còn được bao nhiêu ạ	5 ★	Tích Cực
Anh Nguyễn Thanh Dạm	Sản phẩm chất lượng đạt rất tốt. Mua voi phiếu Giảm giá càng tuyệt vời	5 ★	Tích Cực
Hưởng	okee	5 ★	Tích Cực

Hình 25 Giao diện đánh giá sản phẩm sau khi lọc bình luận tích cực

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1 . Kết quả đạt được

Dự án " Xây dựng hệ thống Web đánh giá sản phẩm " đã hoàn thành với các chức năng cơ bản là đánh giá được tính tích cực hay tiêu cực của sản phẩm điện tử dựa vào các link sản phẩm từ web thế giới di động. Những điểm chính có thể rút ra từ dự án bao gồm:

- Tính khả thi và hiệu quả: Website đã đáp ứng được nhu cầu trong việc đánh giá sản phẩm qua bình luận đưa ra được tính tích cực hay tiêu cực của sản phẩm giúp người sử dụng có thể dựa vào đó để đánh giá sản phẩm 1 cách khách quan hơn
- Trải nghiệm người dùng: Giao diện người dùng thân thiện, dễ sử dụng, giúp người dùng có thể dễ dàng nhận được sự đánh giá của sản phẩm dựa trên link của sản phẩm được đưa vào.
- Tính chính xác: Hệ thống đạt độ chính xác khá cao, từ đó đem lại độ tin cậy cho người dùng trong việc lựa chọn sản phẩm .
- Khả năng mở rộng và tích hợp: Website được xây dựng với kiến trúc mở, dễ dàng tích hợp thêm các tính năng mới hoặc liên kết với các hệ thống khác, đáp ứng nhu cầu phát triển trong tương lai.

2. Hạn chế

Sau khi hoàn thành sản phẩm, em nhận thấy hệ thống em tạo ra còn hạn chế như sau:

- Hệ thống hiện tại chỉ cung cấp các tính năng cơ bản như: xem các đánh giá của sản phẩm dựa vào đường dẫn. Một số tính năng nâng cao như gợi ý sản phẩm tương tự, đưa ra đánh giá chung vẫn chưa được triển khai
- Số lượng trang web có thể đánh giá còn hạn chế, chưa hỗ trợ được tất cả các trang web bán hàng điện tử lớn hiện nay
- Việc đánh giá sản phẩm đánh giá còn chưa sâu, mới chỉ đánh giá được bình luận có tính tích cực hay tiêu cực
- Độ chính xác còn chưa được hoàn thiện với một số những trường hợp có ngữ cảnh đặt biệt
- Hệ thống hiện tại có thể thiếu các tính năng hỗ trợ khách hàng như chat trực tuyến, FAQ chi tiết, hoặc hệ thống quản lý phản hồi và khiếu nại. Điều này có thể làm giảm trải nghiệm người dùng và gây khó khăn khi có vấn đề phát sinh.

3. Hướng phát triển

Mặc dù đề tài đã hoàn thiện các chức năng cơ bản nhưng vẫn còn một số thiếu sót, đề tài cần cải thiện các nội dung sau:

- Cần thêm nhiều dữ liệu để cho việc huấn luyện mô hình đạt hiệu quả cao để việc đánh giá bình luận một cách chuẩn xác hơn
- Phát triển thêm tính năng gợi ý sản phẩm tốt hơn(điểm đánh giá cao hơn từ hệ thống) để người tiêu dùng có nhiều lựa chọn hơn
- Phát triển thêm trong việc đánh giá các tiêu chí khác như số điểm trên thang 5 để người dùng có cái nhìn trực quan hơn

➤ TÀI LIỆU THAM KHẢO

- [1] BERT model. (<https://phamdinhhkhanh.github.io/2020/05/23/BERTModel.html>)
- [2] BERT, RoBERTa, PhoBERT, BERTweet. (https://viblo.asia/p/bert-roberta-phobert-bertweet-ung-dung-state-of-the-art-pre-trained-model-cho-bai-toan-phan-loai-van-ban-4P856PEWZY3#_phan-tich-sac-thai-binh-luan-voi-phobert-2)
- [3] Sentiment analysis using RoBERTa + LSTM(<https://www.kaggle.com/code/minhcng/sentiment-analysis-using-roberta-lstm>)
- [4] Imbalanced dataset (<https://phamdinhhkhanh.github.io/2020/02/17/ImbalancedData.html#43-under-sampling>)