

**TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT**  
**KHOA HỆ THỐNG THÔNG TIN**

---



**BÁO CÁO CUỐI KỲ MÔN HỌC**  
**TÊN MÔN HỌC: PHÂN TÍCH DỮ LIỆU WEB**  
**MÃ HỌC PHẦN: 231IS99**

**PHÂN TÍCH DỮ LIỆU WEB HASAKI VÀ TIKI ĐỂ DỰ ĐOÁN CẢM XÚC**  
**BÌNH LUẬN CỦA NGƯỜI DÙNG TRONG LĨNH VỰC MỸ PHẨM BẰNG**  
**MÔ HÌNH LSTM**






Giảng viên hướng dẫn: **TS Nguyễn Thôn Dã**

**Danh sách thành viên nhóm:**

1. K204110559, Hoàng Ngọc Thảo Duyên
2. K204110565, Phan Trịnh Kim Hạnh
3. K204110575, Phạm Thị Kim Ngân
4. K214130891, Nguyễn Đăng Tú Minh
5. K214131995, Huỳnh Hà Anh Thư

**Thành phố Hồ Chí Minh, 2023**

## Bảng tự đánh giá thành viên nhóm

STT	MSSV	Họ và tên	Điểm tự đánh giá (thang điểm 10)	Ký tên
1	K204110559	Hoàng Ngọc Thảo Duyên	10	
2	K204110565	Phan Trịnh Kim Hạnh	10	
3	K204110575	Phạm Thị Kim Ngân	10	
4	K214130891	Nguyễn Đăng Tú Minh	10	
5	K214131995	Huỳnh Hà Anh Thư	10	



# Lời cảm ơn của nhóm

Chúng em xin gửi lời cảm ơn chân thành đến giảng viên hướng dẫn, ***TS. Nguyễn Thôn Dã***. Trong suốt các buổi học, thầy đã luôn tận tình hướng dẫn những kiến thức cần thiết và hỗ trợ, giải đáp các thắc mắc của sinh viên. Thầy đã giúp chúng em định hướng được những điểm quan trọng của đề tài và đưa ra những nhận xét giúp chúng em cải thiện đề tài tốt hơn. Nhờ sự chỉ dẫn nhiệt tình từ thầy, chúng em đã tiếp thu được nhiều điều mới và là nền tảng cho các môn học sắp tới.

Chúng em xin chân thành cảm ơn các nhà khoa học và tác giả của các công trình công bố mà chúng em đã trích dẫn trong báo cáo. Đó là những nguồn tư liệu quý báu đã giúp nhóm nghiên cứu và hoàn thành đề tài này.

Chúng em xin chân thành cảm ơn.

***Nhóm thực hiện***

# Lời cam kết

Chúng tôi cam đoan kết quả nghiên cứu này là của riêng chúng tôi, chúng tôi khẳng định không sao chép kết quả nghiên cứu của những cá nhân hoặc nhóm nghiên cứu nào khác.

Thành phố Hồ Chí Minh, ngày 26 tháng 12 năm 2023

**Nhóm trưởng**

**Hoàng Ngọc Thảo Duyên**

# Mục lục

---

<b>Danh mục bảng .....</b>	<b>vi</b>
<b>Danh mục hình ảnh .....</b>	<b>vii</b>
<b>Danh mục từ viết tắt .....</b>	<b>viii</b>
<b>Tóm tắt báo cáo.....</b>	<b>1</b>
<b>1. Giới thiệu .....</b>	<b>1</b>
<b>2. Các nghiên cứu liên quan.....</b>	<b>4</b>
<i>2.1 Tác động của đánh giá trực tuyến đến ý định mua sắm của khách hàng .....</i>	<i>4</i>
<i>2.2 Các yếu tố ảnh hưởng đến ý định mua sắm mỹ phẩm trực tuyến.....</i>	<i>4</i>
<i>2.3 Phân tích bình luận của khách hàng .....</i>	<i>5</i>
<i>2.4 Phương pháp phân tích bình luận và gắn nhãn cảm xúc cho bình luận.....</i>	<i>7</i>
<b>3. Mô tả, phân tích tổng quan về dữ liệu .....</b>	<b>7</b>
<b>4. Phương pháp luận nghiên cứu .....</b>	<b>10</b>
<i>4.1 Tiền xử lý dữ liệu.....</i>	<i>11</i>
<i>4.2 Phân tích dữ liệu .....</i>	<i>12</i>
<i>4.3 Các chỉ số đo lường.....</i>	<i>13</i>
<b>5. Kết quả thử nghiệm và phân tích kết quả.....</b>	<b>13</b>
<i>5.1 Bình luận bằng tiếng Anh.....</i>	<i>14</i>
<i>5.2 Bình luận bằng tiếng Việt.....</i>	<i>16</i>
<i>5.3 Nhận xét.....</i>	<i>18</i>
<b>6. Kết luận .....</b>	<b>19</b>
<b>TÀI LIỆU THAM KHẢO.....</b>	<b>21</b>

# Danh mục bảng

Bảng 5.1: Các chỉ số đo lường trên tập test.....	14
Bảng 5.2:Các chỉ số đo lường trên tập Validation .....	14
Bảng 5.3: Độ lớn tập thư viện của 2 ngôn ngữ.....	19

# Danh mục hình ảnh

Hình 4.1: Quy trình phân tích dữ liệu.....	10
Hình 5.1: Hàm mất mát khi gán nhãn cho bình luận tiếng Anh.....	15
Hình 5.2: Độ chính xác của tập Validation khi gán nhãn cho bình luận tiếng Anh.....	16
Hình 5.3: Hàm mất mát khi gán nhãn cho bình luận tiếng Việt.....	17
Hình 5.4: Độ chính xác của tập Validation khi gán nhãn cho bình luận tiếng Việt.....	18



# Danh mục từ viết tắt

<b>Từ</b>	<b>Mô tả</b>
TMĐT	Thương mại điện tử
LSTM	Long Short-Term Memory
B2C	Business to Customer
KOL	Key Opinion Leader
API	Application Programming Interface

## Tóm tắt báo cáo

Thương mại điện tử Việt Nam đang trong giai đoạn phát triển mạnh, số lượng sản phẩm trên thị trường tăng lên nhanh chóng, kéo theo đó là các bài thảo luận, review về sản phẩm, đặc biệt là lĩnh vực mỹ phẩm từ những khách hàng đã trải nghiệm sản phẩm cũng tăng lên đáng kể. Thế nên những người bán hàng sẽ cần rất nhiều thời gian để đọc toàn bộ các nhận xét và rút ra được nhu cầu của khách hàng. Để quá trình phân tích suy nghĩ của người tiêu dùng diễn ra nhanh chóng hơn, trong đề tài này, nhóm đã sử dụng mô hình học sâu LSTM để hỗ trợ người bán hàng đọc và phân loại các bình luận của khách hàng theo 3 cấp độ: Tích cực, Tiêu cực và Trung tính. Nhờ những phân nhóm bình luận đó, người bán sẽ đưa ra được các chiến lược tiếp thị tới đúng nhóm đối tượng mục tiêu hơn. Các chỉ số đo lường hiệu suất mô hình đã cho ra kết quả rất khả quan. Nhóm tác giả đã so sánh hiệu quả của 2 mô hình khi phân tích 2 loại ngôn ngữ khác nhau là tiếng Việt và tiếng Anh. Kết quả cho thấy việc phân tích các bình luận tiếng Việt đã mang lại hiệu quả tốt hơn. Độ chính xác của mô hình LSTM này sẽ đạt được hiệu suất tốt hơn nếu được kết hợp với thư viện hỗ trợ ngôn ngữ tiếng Việt có đa dạng từ vựng hơn.

## 1. Giới thiệu

*“Mua sắm trực tuyến đang ngày càng phổ biến và tăng trưởng trên toàn cầu, chủ yếu là do sự mở rộng nhanh chóng của Internet”* [1]. Theo tổ chức Thương mại thế giới (WTO) *“Thương mại điện tử bao gồm việc sản xuất, quảng cáo, bán hàng và phân phối sản phẩm được mua bán và thanh toán trên mạng Internet, nhưng được giao nhận một cách hữu hình, tất cả các sản phẩm giao nhận cũng như những thông tin số hoá thông qua mạng Internet”* [2]. Theo Báo cáo thương mại điện tử Việt Nam năm 2023, thương mại điện tử đang phát triển rất nhanh chóng. Ước tính, số lượng người tiêu dùng mua sắm trực tuyến ở Việt Nam lần đầu tiên có thể sẽ chạm mốc 59 - 62 triệu người. Giá trị mua sắm trực tuyến của mỗi người dùng sẽ tiếp tục tăng, dự báo sẽ dao động từ 300-320 USD/người trong năm 2023. Tỷ trọng doanh thu thương mại điện tử B2C so với tổng mức bán lẻ hàng hóa và dịch vụ tiêu dùng cả nước sẽ vượt mốc 7,5% của năm 2022, đạt từ 7,8% - 8%. Nhiều sản phẩm quốc tế trong và ngoài nước gia nhập thị trường tại Việt Nam đã đạt được thành

công và chỗ đứng nhất định trên thị trường như Shopee, Amazon, Lazada, ... Ngoài ra, ngành hàng hóa gồm quần áo, giày dép và mỹ phẩm chiếm tỉ lệ cao nhất trong số các loại hàng hóa, dịch vụ được mua trên mạng (chiếm 76%). Kênh mua sắm được nhiều người dùng nhất là website TMĐT (chiếm 70%), mạng xã hội và các ứng dụng mua sắm có tỉ lệ tương đối bằng nhau. Những số liệu trên cho thấy được rằng thị trường mua bán mỹ phẩm trực tuyến ở nước ta có tiềm năng rất lớn, tuy nhiên khách hàng lại lo ngại về vấn đề chất lượng sản phẩm kém hơn so với quảng cáo, chiếm tỉ lệ vượt trội 68% so với các trở ngại mua sắm khác, và lý do lớn nhất mà khách hàng chưa mua sản phẩm trực tuyến vì họ không tin tưởng đơn vị bán hàng [3]. Do đó, các thương hiệu cần quan tâm đến cảm nhận, suy nghĩ của người tiêu dùng từ đó xây dựng lòng tin với họ.

Theo Thành.T.H (2021) “Để xây dựng chiến lược kinh doanh phù hợp cho từng sản phẩm, dịch vụ, đòi hỏi doanh nghiệp phải phân tích cảm xúc của khách hàng và hình thành quan điểm dựa trên những phản hồi đi trước, trong và sau khi mua hàng cho từng nhóm người tiêu dùng” [4]. 80% số người được hỏi trong cuộc thăm dò do Celebrity Intelligence thực hiện cho biết KOL và các bài review của những khách hàng trước có vai trò quan trọng trong việc ảnh hưởng đến quan điểm và quyết định của khách hàng. Hơn nữa, 83% số người được hỏi cho rằng KOL rất cần thiết trong việc phát triển mẫu mã, sản phẩm và xu hướng trong ngành mỹ phẩm. Cụ thể, trong thực tế nếu một sản phẩm có nhiều bình luận tốt từ 70% trở lên thì khả năng sẽ nhận được sự tin tưởng từ khách hàng cao hơn. Và theo báo cáo về TMĐT 2023, có đến 67% người tham gia khảo sát cho biết cách thức tìm kiếm thông tin khi mua sắm trực tuyến của họ là xem các bình luận, đánh giá trên mạng (gồm mạng xã hội, sàn TMĐT), tỉ lệ cao nhất trong danh sách [3]. Vậy nên, việc phân tích nhận xét của khách hàng: icon cảm xúc, comment, rating,... là các yếu tố rất quan trọng để cải thiện sản phẩm, dịch vụ chăm sóc khách hàng và tư vấn mua hàng, đưa ra được các chiến lược marketing đúng đối tượng và gia tăng tình yêu thương hiệu của khách hàng.

Với thực trạng đã nêu và nhận thấy tầm quan trọng của việc phân tích, khai phá phản hồi của khách hàng nên nhóm quyết định chọn đề tài: “***Dự đoán cảm xúc cho các bình luận của người dùng trong lĩnh vực mỹ phẩm bằng Mô hình LSTM***” để hiểu được

ý kiến của người bình luận một cách dễ dàng hơn. Bên cạnh đó việc gắn nhãn cảm xúc cho các bình luận của khách hàng còn giúp cho doanh nghiệp nhận diện cảm xúc, sự hài lòng của khách hàng đối với sản phẩm, từ đó theo dõi một cách hiệu quả phản hồi của khách hàng và tìm ra các giải pháp tối ưu hóa chất lượng sản phẩm cũng như các chiến lược tiếp thị.

Trong bài báo cáo này, nhóm đã xây dựng mô hình LSTM dự đoán cảm xúc bằng cách gắn nhãn vào những bình luận có ngôn ngữ là tiếng Việt bao gồm cả tích cực và tiêu cực của khách hàng thu thập trên website Hasaki ngành hàng chăm sóc da mặt và ngành hàng mỹ phẩm của Tiki. Ngoài ra, nhóm còn sử dụng API Azure Translator để tự động dịch các bình luận này sang tiếng Anh và tiến hành phân tích so sánh hiệu quả của hai mô hình với hai loại ngôn ngữ trên. Mô hình phân tích này giúp doanh nghiệp có thể nhận diện và đánh giá về sản phẩm một cách nhanh chóng, hiểu được nhu cầu của khách hàng khi mua sản phẩm và độ hài lòng về sản phẩm dựa trên những nhận xét mà khách hàng để lại. Hơn thế nữa, mô hình này sẽ giúp doanh nghiệp xây dựng chiến lược kinh doanh phù hợp, lên kế hoạch tiếp tục bán hay ngừng bán những sản phẩm có nhiều đánh giá tiêu cực, đánh giá hiệu suất chiến dịch tiếp thị và điều chỉnh chiến lược nếu cần thiết.

Đề tài đã tìm và tổng hợp được các công trình nghiên cứu trong và ngoài nước để so sánh và làm nền tảng cho bài nghiên cứu của mình. Bên cạnh đó, đề tài cũng chỉ ra được một số hạn chế trong phương pháp cũng như nội dung của các nghiên cứu trước đó. Đề tài đã xây dựng được mô hình LSTM dự đoán tính tích cực và tiêu cực của những bình luận về sản phẩm của người đã mua hàng với độ chính xác tương đối cao. Khi phân tích bình luận bằng tiếng Anh và tiếng Việt có sự khác nhau về kết quả đo lường hiệu suất do công cụ dịch thuật mà nhóm sử dụng còn bị hạn chế về từ ngữ và thư viện tiếng Việt underthesea còn ít các từ liên quan đến lĩnh vực mỹ phẩm.

Mô hình này có thể được phát triển thêm để phân tích các bình luận khác trên các trang mạng xã hội khi nội dung của bài đăng nói về mỹ phẩm (các bài viết, video review mỹ phẩm), tuy nhiên, mô hình chưa được phát triển hoàn thiện do hạn chế về mặt thời gian

và tập dữ liệu thu thập cho việc huấn luyện chưa đủ lớn. Dù vậy, kết quả phân tích của mô hình được xây dựng trong đề tài này là rất khả quan.

## **2. Các nghiên cứu liên quan**

### ***2.1 Tác động của đánh giá trực tuyến đến ý định mua sắm của khách hàng***

Mô hình khả năng thuyết phục (ELM) được sử dụng để đánh giá độ tin cậy của đánh giá trực tuyến và tác động của nó đến việc mua hàng vì đây là mô hình quy trình tích hợp với các yếu tố chất lượng lập luận như độ chính xác, sự đầy đủ và tính kịp thời và các dấu hiệu ngoại vi như: lượng đánh giá, tính nhất quán của đánh giá, chuyên gia đánh giá, đánh giá sản phẩm/dịch vụ, danh tiếng của website. Cả những yếu tố này đều có ảnh hưởng tích cực đến ý định mua sắm của người tiêu dùng [5]. Bên cạnh đó, hình ảnh tinh thần về sản phẩm tác động rất nhiều đến ý định mua hàng và công cụ tiếp thị qua Internet cũng được xác định là có thể tăng cường số lượng mua hàng và đóng góp vào sự phổ biến của sản phẩm và dịch vụ với giá cả cạnh tranh [6]. Tran, L. T. T. (2020) đã kết hợp lý thuyết sử dụng và hài lòng (U&G) cùng với lý thuyết văn hóa tiêu dùng (CCT) để hiểu về lý do người tiêu dùng sử dụng nền tảng truyền thông xã hội (SMP) và nghiên cứu chỉ ra rằng tương tác giữa chủ nghĩa quốc tế và niềm tin trực tuyến vào thương hiệu đóng vai trò quan trọng trong tăng ý định mua hàng [7]. Nghiên cứu này mang lại hiểu biết toàn diện về lý do sử dụng SMP và đặc biệt chú ý đến vai trò quan trọng của chủ nghĩa quốc tế trong tương tác xã hội đa quốc gia.

### ***2.2 Các yếu tố ảnh hưởng đến ý định mua sắm mỹ phẩm trực tuyến***

Ý định mua sắm trực tuyến của người tiêu dùng là quyết định tiền đề hoặc dự định mua sắm sản phẩm hoặc dịch vụ qua các nền tảng trực tuyến. Nó được ảnh hưởng bởi một loạt các yếu tố, và việc hiểu rõ ý định này có thể giúp các doanh nghiệp và nhà quảng cáo hiểu rõ hơn nhu cầu và hành vi của khách hàng. Theo Hà, N. T. (2016), bằng việc áp dụng thành công lý thuyết hành vi có hoạch định (Theory of Planned Behavior - TPB) như là một khung lý thuyết để dự đoán ý định và hành vi mua trực tuyến và đã nghiên cứu được rằng “*Ý định mua trực tuyến của người tiêu dùng bị ảnh hưởng bởi thái độ của người tiêu*

dùng đối với trang web, nhận thức kiểm soát hành vi của người tiêu dùng và rủi ro cảm nhận của chính họ”. Ngoài các yếu tố trên thì trong bài nghiên cứu của mình, Thanh. T. V và cộng sự (2021) đã thử sử dụng một mô hình khác đó là TAM - TECHNOLOGY ACCEPTANCE MODEL làm cơ sở lý thuyết để xây dựng và phát triển mô hình nghiên cứu sau đó kiểm định và điều tra các nhân tố ảnh hưởng đến ý định mua sắm trực tuyến của khách hàng. Kết quả nghiên cứu cho thấy có 4 nhân tố: nhận thức tính hữu ích, niềm tin, cảm nhận rủi ro, và tâm lý an toàn có ảnh hưởng đến ý định mua sắm trực tuyến của người tiêu dùng.

Cụ thể hơn, trong lĩnh vực mỹ phẩm có rất nhiều nghiên cứu đã phát hiện và chỉ ra rằng có nhiều hơn 3 yếu tố tác động mạnh mẽ đến ý định mua mỹ phẩm trực tuyến của khách hàng. Dựa trên Lý thuyết hành động hợp lý được đề xuất bởi Fishbein và Ajzen (1975) kết hợp với mô hình TAM phù hợp, *Tuyền. D. T. K & cộng sự (2021)* đã khẳng định các yếu tố như: Chất lượng thiết kế trang web, nhận thức hữu ích, nhận thức dễ sử dụng, chuẩn chủ quan có quan hệ cùng chiều với ý định mua sắm mỹ phẩm trực tuyến của người tiêu dùng. Hay theo nghiên cứu của *Nguyen. T. H. N & cộng sự (2021)*, bằng việc thu thập 300 mẫu khảo sát và khẳng định được rằng các yếu tố giá cả, sản phẩm, xúc tiến bán hàng, thương hiệu, thái độ và nhóm tham khảo đều có mối quan hệ cùng chiều với ý định mua mỹ phẩm của khách hàng nữ thuộc thế hệ Z sống ở khu vực Thành Phố Hồ Chí Minh. Trong số các yếu tố này, ảnh hưởng mạnh đến ý định mua mỹ phẩm là thương hiệu, nhóm tham khảo, giá cả và sản phẩm, hai yếu tố còn lại là thái độ và xúc tiến bán hàng có tác động không đáng kể.

### ***2.3 Phân tích bình luận của khách hàng***

Với sự cải thiện của thông tin xã hội và sự phổ biến của các thiết bị di động khác nhau, thương mại điện tử đang duy trì xu hướng phát triển nhanh chóng và sự bùng nổ tiếp theo của thông tin bình luận trực tuyến đang rất cần giải pháp hợp lý cho tất cả những người tham gia thương mại điện tử. Các nghiên cứu cho thấy có 90.0% người tiêu dùng mua sắm trực tuyến duyệt qua các bình luận đánh giá sản phẩm và 92.2% trong số họ nói rằng các đánh giá trực tuyến sẽ ảnh hưởng đến quyết định mua hàng của họ. Đồng thời, đối với

người bán trên các nền tảng, các bình luận - văn bản đánh giá trực tuyến, dưới dạng ngôn ngữ do người dùng xác định, chứa thông tin danh tiếng thực sự và thông tin ưu tiên người dùng về sản phẩm. So với điểm riêng của dự án do người dùng đánh giá, các bình luận đánh giá chứa sở thích của người dùng chính xác hơn. Thông qua phân tích toàn diện về đánh giá trực tuyến, người bán trên nền tảng sẽ dễ dàng nắm bắt phản hồi của người tiêu dùng về các sản phẩm có liên quan một cách nhanh chóng và hiệu quả hơn.

Trong các nghiên cứu liên quan, các nhà nghiên cứu - những người giải quyết các vấn đề ngôn ngữ học tính toán, bao gồm các vấn đề phân loại văn bản, thường chọn một phương pháp là học máy. Cách tiếp cận này mang lại kết quả tốt trong việc phân loại văn bản theo chủ đề. Tuy nhiên, cách tiếp cận này không phù hợp để phân tích tình cảm, cảm xúc trong bình luận của khách hàng vì các từ thể hiện trạng thái cảm xúc không phải là duy nhất cho các tài liệu văn bản và do đó thuật toán đào tạo máy móc không tính đến khi sử dụng công cụ vector hóa văn bản tiêu chuẩn, chẳng hạn như “bag-of” , “-words” và TF-TDF (thuật ngữ tần số tài liệu nghịch đảo tần số). Do những nhược điểm của việc sử dụng các phương pháp tiêu chuẩn để phân tích tình cảm, người ta đề xuất sử dụng một kỹ thuật dựa trên sự tổng hợp của ba phương pháp: dựa trên quy tắc, từ điển và học máy có giám sát. Ngoài ra, người ta đề xuất sử dụng một cách tiếp cận khác để vector hóa văn bản - vector hóa âm sắc, khác với các phương pháp tiêu chuẩn như “bag - of - words” và TF-IDF. Ý tưởng chính là tìm trong văn bản nhận xét các từ thể hiện đánh giá tích cực hoặc tiêu cực dựa trên phân tích này, đánh giá cảm xúc của nhận xét đó - tích cực hay tiêu cực.

Trong nghiên cứu của Jia Ke và các cộng sự (2024), người ta đã sử dụng mô hình phân tích tương quan cảm xúc và bản đồ tự tổ chức (SOM) vào ứng dụng để xây dựng vectơ cảm xúc người dùng chi tiết dựa trên bình luận và thực hiện phân tích cụm hình ảnh, giúp nhanh chóng khai thác các đặc điểm và phân cụm người dùng từ bình luận trực tuyến. Bản đồ tự tổ chức (SOM) là một mô hình mạng thần kinh đặc biệt, có thể phản ánh cấu trúc liên kết không gian của dữ liệu chiều cao (high-dimensional) sang dữ liệu chiều thấp (low-dimensional) để thực hiện trực quan hóa chiều ngược lại của dữ liệu có sẵn. Nó phù hợp để xử lý và phân tích dữ liệu vector điểm cảm xúc của người dùng trong nghiên cứu.

Cũng trong nghiên cứu này, thuật toán VC-SOM đã được nhóm tác giả sử dụng để phân tích các tập dữ liệu bằng cách tạo bản đồ tính năng phân bố cụm SOM bằng phương pháp ánh xạ màu. Phân phối topo của dữ liệu chiều cao được trực quan hóa và thu được các tính năng phân phối dữ liệu, theo đó số lượng danh mục phân loại cảm xúc được xác định.

#### ***2.4 Phương pháp phân tích bình luận và gắn nhãn cảm xúc cho bình luận***

Trong xử lý ngôn ngữ tự nhiên, phân tích tình cảm (SA) rất quan trọng và có nhiều cách để tiếp cận nó. [10] cho thấy rằng, NB và SVM là một trong những thuật toán học máy (machine learning) phổ biến được sử dụng trong SA. Ngược lại, các kỹ thuật dựa trên học sâu (deep learning) tận dụng mạng lưới thần kinh (neural) để tự động học cách trình bày văn bản và trong thời gian gần đây đã cho thấy kết quả đáng khích lệ trong SA, bao gồm các mô hình như: CNN, RNN, LSTM, Transformer, BERT, Generative Pre-Trained (GPT) và xu hướng hướng tới các phương pháp học sâu đang gia tăng [11]. Zahoor et.al [12] tiến hành phân tích và phân loại tình cảm của đánh giá nhà hàng bằng Machine Learning. Nghiên cứu đã so sánh kết quả của các thuật toán: Naive Bayes Classifier, Logistic Regression, Support Vector Machine (SVM), và Random Forest. Kết quả cho thấy thuật toán Random Forest cho độ chính xác cao nhất, với 95%. Bodapati et.al [13] cũng tiến hành phân loại cảm xúc bằng 5 thuật toán là Logistic Regression, SVM, MLP, DNN và LSTM. Kết quả cho thấy SVM mang lại hiệu suất thấp nhất, trong khi MLP và DNN cho kết quả gần bằng với mô hình dựa trên LSTM. Các nghiên cứu trước đó đã đạt được những kết quả tốt trong phân tích cảm xúc của bình luận.

### **3. Mô tả, phân tích tổng quan về dữ liệu**

Dữ liệu được thu thập từ trang web Hasaki kết hợp trang web Tiki bằng thư viện Selenium. Tiếp theo, nhóm đã xử lý các bình luận và dịch máy bằng API Azure Translator để tổng hợp thành bảng gồm 6416 dòng và 6 cột lần lượt là:

- **STT**: số thứ tự của sản phẩm
- **Rate**: (gồm 5 mức độ) Mức độ tốt tăng dần từ 1-5
- **Content comment**: chứa các bình luận thô của khách hàng lấy được từ trang web



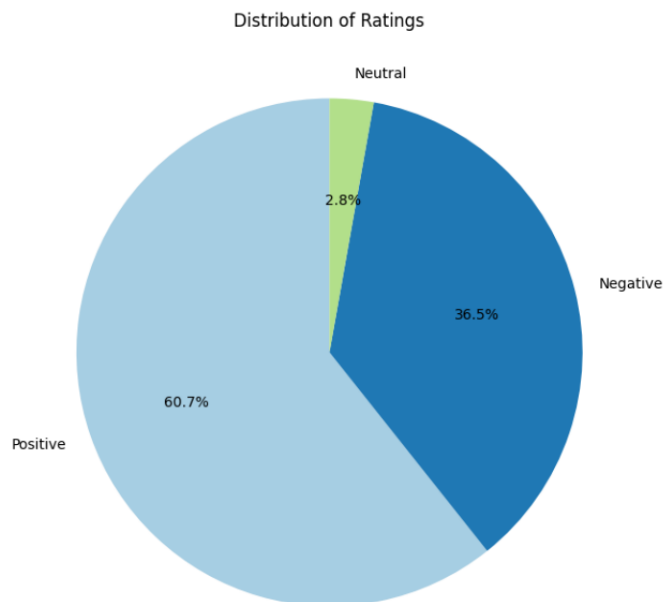
- **Sentiment:** gồm trạng thái positive - tích cực, negative - tiêu cực và neutral - trung tính
- **Content\_comment\_vi:** dữ liệu bình luận dành cho việc huấn luyện và kiểm tra ở ngôn ngữ tiếng Việt
- **Content\_comment\_en:** dữ liệu bình luận dành cho việc huấn luyện và kiểm tra ở ngôn ngữ tiếng Việt để sau khi áp dụng thuật toán LSTM sẽ dự đoán cảm xúc của comment thuộc positive hay negative.

Về phần đánh giá trạng thái sentiment, theo thang đo như sau:

- Nhóm thuộc từ 1-2 sao: negative
- Nhóm thuộc 3 sao: neutral
- Nhóm thuộc từ 4-5 sao: positive

	rate	content_comment	sentiment	content_comment_vi	content_comment_en
0	5	chân ái của tui xài chai thứ 3 rồi cấp ẩm nhẹ ...	Positive	chân ái của tui xài chai thứ 3 rồi cấp ẩm nhẹ ...	My Chan Ai used the 3rd bottle and moisturized...
1	5	ok	Positive	ok	Satisfactory
2	5	Dùng rất ok nha cấp ẩm tốt tonner chân ái dùng...	Positive	dùng rất ok nha cấp ẩm tốt tonner chân ái dùng...	Very ok dentist good moisturizer Toner Chan A...
3	5	okk nha	Positive	okk nha	okk home
4	4	Vừa nhận được chưa sd . Ko tặng bông tẩy trang...	Positive	vừa nhận được chưa sử dụng . không tặng bông t...	just received unused. Do not give cotton remov...
...	...	...	...	...	...

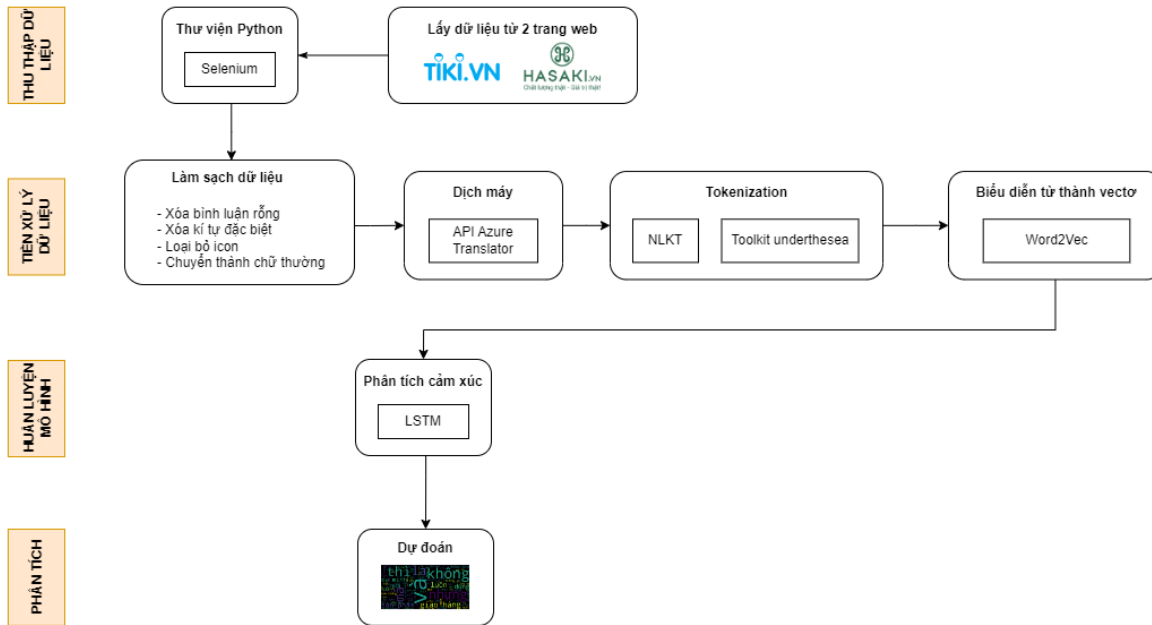
*Hình 3.1: Bảng dữ liệu đưa vào phân tích mô hình LSTM*



*Hình 3.2: Tỷ lệ phần trăm bình luận tích cực, tiêu cực và trung tính trong toàn bộ dữ liệu*

Dựa vào hình 3.2 ta thấy được trạng thái positive khá vượt trội với số lần xuất hiện lên đến 3894 lần chiếm 60,71%. Tiếp đến là trạng thái negative xuất hiện 2340 lần chiếm 36,48%. Và thấp hơn chỉ xuất hiện 180 lần chiếm một phần rất nhỏ (2,81%) là trạng thái neutral.

#### 4. Phương pháp luận nghiên cứu



Hình 4.1: Quy trình phân tích dữ liệu

Hình 4.1 trình bày mô hình nghiên cứu dự đoán cảm xúc cho bình luận của người dùng về mỹ phẩm trên các sàn thương mại điện tử dựa trên phương pháp học sâu. Mô hình bài toán được chia ra làm 4 phần: Thu thập dữ liệu, Tiền xử lý dữ liệu, Huấn luyện mô hình và Phân tích. Sử dụng thư viện Selenium trong ngôn ngữ lập trình Python để thu thập dữ liệu từ các sàn TMĐT như Tiki và website Hasaki.vn. Các dữ liệu đầu vào được xử lý sạch như xóa dòng trống, kí tự icon, kí tự đặc biệt, chuyển về ký tự thường,... trước khi được đưa vào huấn luyện mô hình thông qua các thư viện có sẵn trong ngôn ngữ lập trình Python, việc này làm cho dữ liệu thô được điều chỉnh lại phù hợp với các bước sau. Dữ liệu còn được dịch máy bằng API Azure Translator. Sử dụng bộ công cụ ngôn ngữ NLTK để hỗ trợ xử lý giá trị rỗng và ngoại lệ đối với bình luận đã dịch sang tiếng Anh và toolkit underthesea để chuẩn hóa từ và tách từ cho cột bình luận tiếng Việt. Thư viện Word2Vec để biểu diễn từ thành vector trong không gian vector. Thuật toán LSTM được dùng để dự đoán cảm xúc bình luận của người dùng.

#### ***4.1 Tiền xử lý dữ liệu***

Ban đầu nhóm tác giả đã sử dụng thư viện Selenium trong quá trình thu thập dữ liệu bình luận và lưu dưới dạng file CSV. Nguồn được lựa chọn để thu thập dữ liệu là 2 trang web Hasaki.com - trang thương mại điện tử về mỹ phẩm ở Việt Nam và Tiki.com - sàn thương mại điện tử đa sản phẩm tại Việt Nam. Với mục đích tập trung vào lĩnh vực mỹ phẩm, đối với sàn Tiki, nhóm tác giả lựa chọn hạng mục mỹ phẩm cho việc lấy dữ liệu. Dữ liệu thu thập những bình luận của người Việt nên dữ liệu sẽ là những ý kiến với ngôn ngữ Tiếng Việt.

Sau khi thu thập những ý kiến từ người dùng, dữ liệu cần được xử lý để loại bỏ những ý kiến mang tính trùng lặp hoặc seeding để lọc được những bình luận mang tính xây dựng nhất. Ngôn ngữ tiếng Việt đặc biệt là bình luận thì dữ liệu luôn có những từ viết tắt, teencode làm mất tính trọn vẹn của câu. Vì vậy dữ liệu sẽ cần xử lý qua một lần để thay thế những từ viết tắt, teencode và những emoji trong câu. Dựa vào kết quả điểm đánh giá (rating), gán nhãn cho các bình luận có điểm từ 1 đến 2 là Negative (tiêu cực), bình luận được đánh giá 3 điểm là Neutral (trung tính), và các bình luận có điểm 4 và 5 mang ý nghĩa Positive (tích cực). Cuối cùng sẽ chuyển toàn bộ các chữ in hoa trong câu thành chữ thường.

Tiếp đến, thay thế và xóa các giá trị rỗng hoặc bỏ những dòng chỉ chứa các giá trị ngoại lệ nhờ sự hỗ trợ của bộ công cụ ngôn ngữ NLTK.

Ở bước tiếp theo, nhóm tác giả từ dữ liệu của cột content để tách ra thành 2 cột tương ứng với content\_comment\_vi và content\_comment\_en lần lượt là dữ liệu bình luận dành cho việc huấn luyện và kiểm tra ở ngôn ngữ tiếng Việt và tiếng Anh, có thể dễ dàng so sánh:

Đối với dữ liệu cột “content\_comment\_en”: từ dữ liệu comment sẽ sử dụng api từ Azure Text Translation của Microsoft cho việc dịch thuật comment từ tiếng Việt sang tiếng Anh. Tiếp đến sẽ loại bỏ các Stop words (từ dừng), việc này có tác dụng là loại bỏ những từ xuất hiện nhiều trong từ ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa trong việc huấn luyện mô hình. Trong Tiếng Anh, stop words là những từ như: is, that, this...

Đối với dữ liệu cột “content\_comment\_vi”: từ dữ liệu comment sẽ sử dụng thư viện `underthesea` - là bộ NLP Vietnamese Toolkit, bộ dữ liệu và hướng dẫn mã nguồn mở hỗ trợ nghiên cứu và phát triển về Xử lý ngôn ngữ tự nhiên tiếng Việt. Bộ toolkit hỗ trợ việc chuẩn hóa từ (Normalization) và tách từ (Tokenization), chuyển một dãy các ký tự thành một dãy các token - một dãy các ký tự mang ý nghĩa cụ thể, biểu thị cho một đơn vị ngữ nghĩa trong xử lý ngôn ngữ. Không như tiếng Anh, tiếng Việt cần những đơn vị tiếng đơn hay từ đơn để ghép thành từ có nghĩa như: sản\_phẩm, sử\_dụng, kem\_chống\_nắng. Tiếng anh mỗi nghĩa được thể hiện ở 1 từ: product, using, Sunscreen,.. Vì vậy, việc tách từ cần phụ thuộc và độ chính xác và sự đa dạng tách từ trong bộ toolkit của `underthesea`.

Để biểu diễn giá trị là các từ dưới dạng các vector trong không gian vector, nhóm lựa chọn thư viện `Word2Vec` và truyền một số cấu hình tham số quan trọng được đặt ra để điều chỉnh quá trình huấn luyện như sau:

- **vector\_size= 300**: Tham số xác định kích thước của vector từ được tạo ra bởi `Word2Vec`. Trong trường hợp này, vector từ sẽ có 300 chiều.
- **window= 7**: Tham số xác định kích thước của cửa sổ trượt qua văn bản trong quá trình huấn luyện. Cụ thể, mỗi từ sẽ được xem xét và đánh giá trong ngữ cảnh của 7 từ xung quanh nó.
- **min\_count= 10**: Tham số này chỉ định số lượng tối thiểu của một từ trong toàn bộ văn bản để được thêm vào từ điển và tham gia quá trình huấn luyện. Trong trường hợp này, chỉ có các từ xuất hiện ít nhất 10 lần mới được xem xét.

## **4.2 Phân tích dữ liệu**

### **4.2.1 Word2Vec**

`Word2Vec` sử dụng mạng nơ-ron để tự động học mối quan hệ từ văn bản. Sau khi huấn luyện, nó có thể phát hiện từ đồng nghĩa và gợi ý từ cho câu không hoàn chỉnh. `Word2Vec` chuyển đổi mỗi từ thành vector số cụ thể với kích thước cố định. Các vector được chọn cẩn thận để nắm bắt ngữ nghĩa và cú pháp, giúp biểu diễn từng từ dưới dạng

danh sách số. Các vector này tổ chức thông minh để đo lường tương đồng ngữ nghĩa giữa từ thông qua cosine similarity.

#### **4.2.2 LSTM**

LSTM (Long Short-Term Memory) là một mô hình mạng nơ-ron mở rộng từ RNN, được thiết kế để giải quyết vấn đề phụ thuộc dài hạn trong chuỗi dữ liệu. RNN thường gặp khó khăn trong việc duy trì thông tin dài hạn, do đó, LSTM xuất hiện để giữ thông tin qua thời gian dài mà không mất đi tính hiệu quan trọng. Một đơn vị LSTM thông thường bao gồm một tế bào (cell), một cổng vào (input gate), một cổng ra (output gate) và một cổng quên (forget gate). Ba cổng này cho phép LSTM linh hoạt điều chỉnh luồng thông tin trong mô hình. LSTM được ứng dụng rộng rãi trong các lĩnh vực xử lý ngôn ngữ tự nhiên, dự đoán chuỗi thời gian, và phân loại dữ liệu chuỗi, nơi cần xử lý thông tin dài hạn và mối quan hệ phức tạp. Trong nhiều trường hợp, LSTM mang lại hiệu suất cao hơn so với RNN truyền thống.

#### **4.3 Các chỉ số đo lường**

Có 5 chỉ số đo lường được sử dụng để đánh giá hiệu suất của mô hình gồm: Accuracy, Loss Function, F1-score, Recall, Precision. Accuracy cho thấy tỉ lệ giữa số mẫu dự đoán đúng so với tổng số mẫu trong tập dữ liệu. Loss Function thể hiện sự chênh lệch giữa nhãn thật của mẫu và nhãn được dự đoán cho mẫu. Trong đề tài này, hàm mất mát được dùng để tính chênh lệch khi mô hình trải qua 7 lần huấn luyện, nhằm quan sát hiệu quả của mô hình sau mỗi lần học. Độ chính xác càng cao và giá trị mất mát càng thấp thì hiệu suất của mô hình càng tốt. F1-score, Precision và Recall nằm trong nửa đoạn  $(0;1]$ , các chỉ số này đồng thời càng gần với 1 thì kết quả dự đoán của mô hình càng chính xác.

### **5. Kết quả thử nghiệm và phân tích kết quả**

*Bảng 5.1: Các chỉ số đo lường trên tập test*

	<b>Accuracy</b>	<b>F1-score</b>	<b>Recall</b>	<b>Precision</b>
<b>Tiếng Anh</b>	0.9	0.92	0.93	0.92
<b>Tiếng Việt</b>	0.91	0.93	0.92	0.94

Ngoài ra, để tránh tình trạng overfitting khi huấn luyện, nhóm đã thực hiện chia 6415 bình luận thành 3 tập gồm: tập test ( $\frac{1}{3}$  tập dữ liệu - 2058 bình luận), tập train ( $\frac{2}{3}$  tập dữ liệu - 4178 bình luận) - trong đó có 10% dữ liệu được dùng làm tập validation

*Bảng 5.2: Các chỉ số đo lường trên tập Validation*

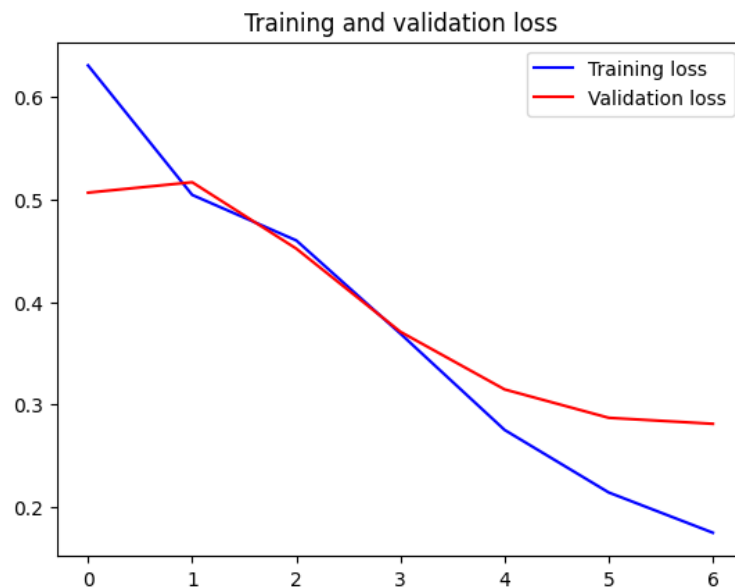
	<b>Độ chính xác của bình luận tích cực</b>	<b>Độ chính xác của bình luận tiêu cực</b>
<b>Tiếng Anh</b>	0.9252	0.8712
<b>Tiếng Việt</b>	0.9167	0.8936

### ***5.1 Bình luận bằng tiếng Anh***

Đối với tập test, các chỉ số đo lường cho thấy các bình luận bằng tiếng Anh được phân tích và gắn nhãn với tỉ lệ chính xác cao, tất cả chỉ số đều trên 90%. Precision càng cao cho thấy mức độ dự đoán nhãn chính xác của các điểm mà nó tìm thấy càng cao, và đồng thời kết hợp với chỉ số Recall cao cho thấy số điểm True Positive bị bỏ sót là rất ít. Tỷ lệ dự đoán các bình luận bằng tiếng Anh (92%), có độ chính xác của các điểm tìm được thấp hơn so với tiếng Việt (94%), nhưng số điểm True Positive mà nó tìm được lại nhiều hơn (93%).

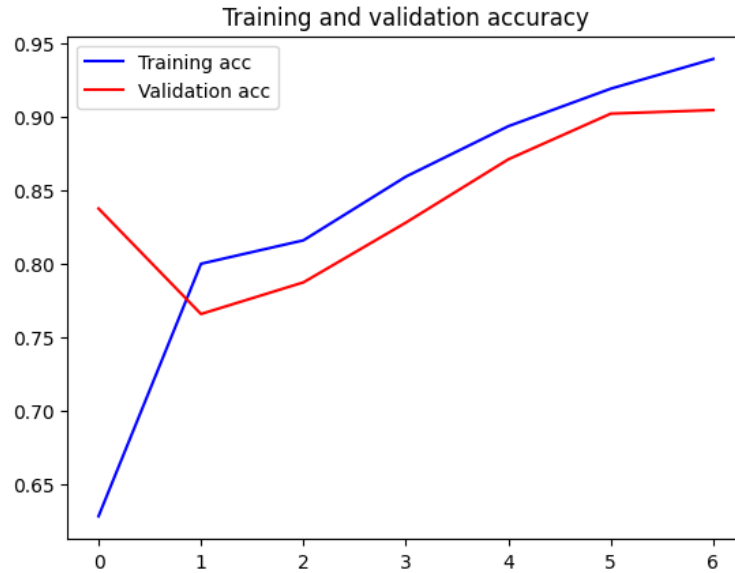
Đối với tập validation, độ chính xác của các bình luận tích cực rất cao xấp xỉ 92.52% và độ chính xác của các bình luận tiêu cực xấp xỉ 87.12%. Hai tỷ lệ có sự chênh lệch nhưng không quá lớn, điều này xảy ra có thể do số bình luận tiêu cực mà nhóm thu thập được ít hơn số bình luận tích cực. Tuy nhiên, tỷ lệ này cũng đã cho thấy được khả năng dự đoán của mô hình tương đối tốt.

Trong quá trình huấn luyện mô hình, nhóm đã cho mô hình LSTM trải qua 7 lần huấn luyện (từ lặp lại 0 lần đến 6 lần) với cùng một tập train và dùng tập validation để tính toán hàm mất mát cho ra kết quả như hình 5.1 và độ chính xác của tập validation trong khi huấn luyện như hình 5.2.



Hình 5.1: Hàm mất mát khi gán nhãn cho bình luận tiếng Anh





*Hình 5.2: Độ chính xác của tập Validation khi gán nhãn cho bình luận tiếng Anh*

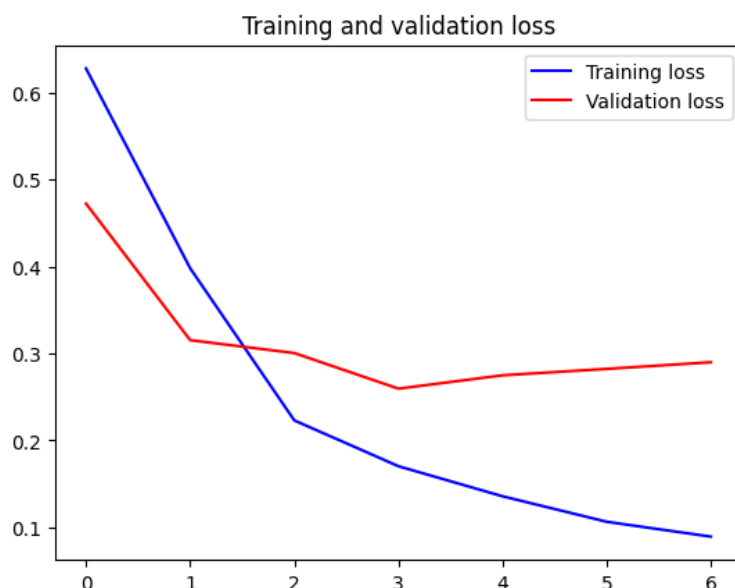
Tổng quan, mô hình LSTM đã cho kết quả dự đoán chính xác hơn sau mỗi lần lặp lại huấn luyện. Accuracy của tập validation giảm trong 2 lần huấn luyện đầu và tăng nhanh từ sau lần thứ 2. Ở lần 6 và 7, giá trị này không có sự thay đổi rõ rệt. Giá trị hàm mất mát của tập validation cũng giảm nhanh sau lần huấn luyện thứ 2 và giảm chậm lại từ sau lần huấn luyện thứ 6. Vì nhóm chưa tính được số lần huấn luyện lặp lại cho đến khi hội tụ nên chưa thể đưa ra kết luận về giá trị tốt nhất của tham số epochs, nhưng nếu chỉ xét trong kết quả 7 lần huấn luyện như hình 5.1, lần thứ 6 và thứ 7 cho ra kết quả phù hợp nhất khi có giá trị validation loss thấp nhất. Như vậy, mô hình LSTM phân tích bình luận tiếng Anh trong bài là mô hình được nhóm đánh giá tốt nhất.

## ***5.2 Bình luận bằng tiếng Việt***

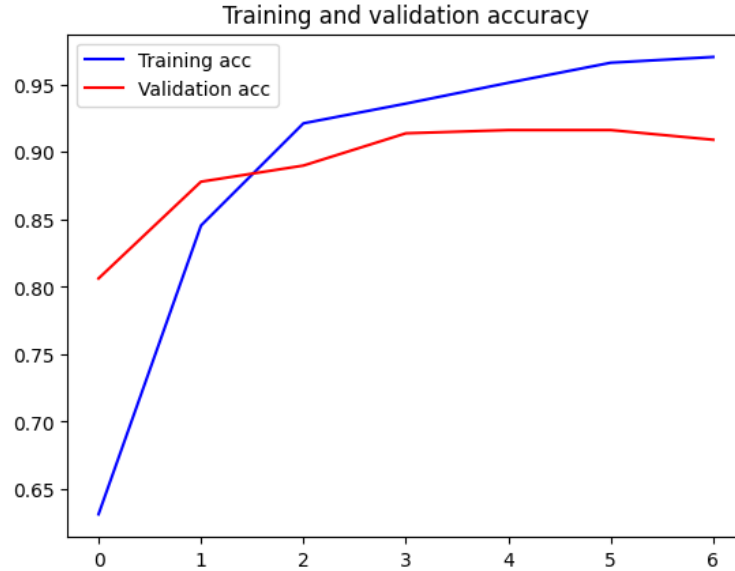
Đối với tập test, các chỉ số được tính cũng có kết quả khả quan giống với kết quả phân tích các bình luận tiếng Anh. Độ chính xác đạt 91%, chênh lệch rất ít so với tiếng Anh. Chỉ số F1-score cao hơn (93% so với 92%), và như đã được đề cập ở phần trên, chỉ số Recall là 92% và Precision là 94%. Dù bị hạn chế về công cụ dịch thuật và thư viện từ ngữ tiếng Việt về chủ đề mỹ phẩm, nhưng kết quả dự đoán của mô hình vẫn cho ra được

số liệu gần như tương tự với việc phân tích bình luận bằng tiếng Anh với thư viện từ ngữ được phát triển phong phú sẵn.

Đối với tập validation, độ chính xác của các bình luận tích cực (xấp xỉ 91.67%) thấp hơn so với tiếng Anh, nhưng độ chính xác của các bình luận tiêu cực (xấp xỉ 89.36%) cao hơn so với tiếng Anh. Sự chênh lệch này có thể là do quá trình dịch máy, Azure Translator chưa hiểu được hết các cách sử dụng từ ngữ tiếng Việt nên dẫn đến việc làm giảm tính tiêu cực trong các bình luận, và các câu bình luận không hoàn chỉnh (thiếu chủ ngữ hoặc vị ngữ, không có dấu tách câu) khiến công cụ này gặp khó khăn khi dịch thuật. Nếu được dịch bằng công cụ có hỗ trợ tốt hơn, độ chính xác của tập validation khi dịch bình luận tiếng Anh có thể sẽ tăng cao hơn.



Hình 5.3: Hàm mất mát khi gán nhãn cho bình luận tiếng Việt



*Hình 5.4: Độ chính xác của tập Validation khi gán nhãn cho bình luận tiếng Việt*

Nhìn tổng quan, hình 5.3 và 5.4 cho thấy mô hình LSTM đã học và dự đoán được tỷ lệ đúng cao hơn và độ lệch của giá trị dự đoán so với giá trị thực đã giảm. Điều này cho thấy mô hình được cải thiện hơn sau mỗi lần huấn luyện. Tuy nhiên, từ sau lần học thứ 4, accuracy của tập train vẫn tăng trong khi chỉ số này của tập validation có giảm nhẹ, có thể dễ quan sát hơn khi training loss và validation loss lại đi theo hai chiều ngược nhau, của tập train giảm xuống rất thấp nhưng tập validation lại tăng dần đều. Điều này cho thấy mô hình bắt đầu học hỏi các mẫu sai lệch trong tập dữ liệu huấn luyện và tình trạng overfitting có thể xảy ra vì mô hình đã rơi vào tình trạng “thuộc lòng” dữ liệu trong tập huấn luyện. Xét trong kết quả 7 lần huấn luyện như hình 5.3 thì mô hình ở lần huấn luyện thứ 4 cho kết quả tốt nhất vì giá trị của hàm mất mát thấp nhất.

### **5.3 Nhận xét**

Nhóm nhận thấy rằng mô hình LSTM cho tập tiếng Việt đã đạt được các chỉ số đánh giá cao và tích cực. Điều này có thể chỉ ra rằng mô hình này phản ánh tốt sự phức tạp và ngữ cảnh của ngôn ngữ tiếng Việt trong các bình luận. Việc đánh giá cao cũng có thể ánh bật sự linh hoạt và khả năng học mối quan hệ phức tạp của LSTM trong ngữ cảnh ngôn

ngữ Việt. Sự phức tạp và cấu trúc ngôn ngữ Việt thường đòi hỏi một mô hình có khả năng học các mối quan hệ phức tạp, và LSTM có thể đã đáp ứng được yêu cầu này.

So với tiếng Việt, hiệu suất của mô hình LSTM trong tiếng Anh có thể được so sánh để đưa ra một cái nhìn tổng thể về khả năng chung của thuật toán. Nếu mô hình cho tiếng Anh cũng đạt được kết quả tích cực, điều này có thể chứng minh tính chuyển giao của thuật toán khi áp dụng cho nhiều ngôn ngữ khác nhau.

Bên cạnh việc dựa vào các biểu đồ và chỉ số, nhóm đã quan sát thêm về tập thư viện đã tạo ra trong quá trình huấn luyện:

*Bảng 5.3: Độ lớn tập thư viện của 2 ngôn ngữ*

<b>Độ lớn tập thư viện ở tiếng Anh</b>	<b>Độ lớn tập thư viện ở tiếng Việt</b>
5971	6057

Nhóm cho rằng sự chênh lệch giữa 2 tập là không lớn vì tập sinh ra còn bị ảnh hưởng bởi quá trình xử lý dữ liệu trước khi đưa vào mô hình. Nhưng điều này cũng có thể nói rằng mô hình LSTM với ngôn ngữ tiếng Việt mang lại sự tối ưu hơn dù chỉ chênh lệch không quá lớn so với tiếng anh nếu biết lựa chọn tham số lần học là hợp lý.

Mặc dù LSTM đã cho thấy hiệu suất tích cực, việc triển khai mô hình trên quy mô lớn và trong các ứng dụng thực tế có thể đặt ra những thách thức khác nhau. Chi phí tính toán, tốc độ suy luận, và khả năng giải thích cũng cần được xem xét khi quyết định về việc triển khai mô hình này.

## **6. Kết luận**

Dựa vào kết quả nghiên cứu và phân tích từ tập dữ liệu bình luận thu thập từ Hasaki và Tiki, bài báo cáo đã thành công xây dựng một mô hình dự đoán cảm xúc của bình luận bằng mô hình LSTM, và kết quả thử nghiệm đã đạt được đánh giá cao. Mặc dù quá trình huấn luyện trên tập dữ liệu tiếng Việt vẫn đối mặt với khó khăn và thiếu sót, nhưng những

bước tiến này đã mở đường cho nhóm thực hiện nghiên cứu liên quan sâu sắc hơn trong tương lai. Mô hình này có thể ứng dụng hiệu quả vào việc phân tích ý kiến của người dùng mỹ phẩm trên các diễn đàn mạng xã hội, cung cấp nguồn thông tin quý giá giúp quản trị viên khắc phục hiệu quả những vấn đề mà khách hàng quan tâm.

Tuy nhiên, mô hình vẫn đối mặt với một số hạn chế. Giới hạn về thời gian đã làm cho tập dữ liệu chưa đạt đến kích thước lớn như mong muốn, bên cạnh đó là giới hạn về kinh phí nên nhóm chọn phương pháp dịch bình luận bằng Azure Translator thay vì sử dụng các công cụ dịch máy tích hợp AI khác như Chat GPT hay Google AI Translate. Bên cạnh đó việc không tìm đủ thư viện từ ngữ tiếng Việt về mỹ phẩm để tách từ trong quá trình huấn luyện là một thách thức. Nhóm sẽ tập trung vào việc khắc phục những trở ngại này trong các nghiên cứu tiếp theo.

Về tương lai của đề tài, nhóm đặt ra nhiều mục tiêu hứa hẹn. Thứ nhất, nhóm sẽ không chỉ dừng lại ở việc sử dụng mô hình LSTM mà còn thử nghiệm trên nhiều mô hình khác nhau để so sánh và lựa chọn ra mô hình phù hợp và tối ưu nhất. Thứ hai, nhóm nhận thức rằng ngoài thư viện underthesea, vẫn còn nhiều thư viện tiềm năng và toolkit hỗ trợ cho việc phân tích ngôn ngữ tiếng Việt, nhóm sẽ chọn lọc, hoặc kết hợp các thư viện để làm đa dạng vốn từ vựng hơn. Nhóm sẽ tiếp tục nghiên cứu, lựa chọn và kết hợp các công cụ này để làm phong phú hóa quy trình nghiên cứu. Cuối cùng, nhóm sẽ mở rộng phạm vi phân tích và dự đoán, không chỉ dừng lại ở hai yếu tố tích cực và tiêu cực, mà còn khám phá giải pháp phân loại cảm xúc thành nhiều trạng thái và mức độ khác nhau, nhằm đạt được dự đoán chính xác hơn về hành vi và sở thích của người tiêu dùng.

## TÀI LIỆU THAM KHẢO

- [1] Wu L., Cai Y. & Liu D. (2011). Online shopping among Chinese customers: An exploratory investigation of demographics and value orientation, *International Journal of Customer Studies*, 35, 458-469.
- [2] WTO. “Electronic commerce”.
- [3] Bộ Công thương - Cục thương mại điện tử và Kinh tế số (2023). Báo cáo thương mại điện tử Việt Nam 2023.
- [4] Thanh. H. T và cộng sự (2021). Phân tích ý kiến khách hàng trực tuyến dựa theo phương pháp học máy.
- [5] Thomas, M. J., Wirtz, B. W., & Weyerer, J. C. (2019). DETERMINANTS OF ONLINE REVIEW CREDIBILITY AND ITS IMPACT ON CONSUMERS' PURCHASE INTENTION. *Journal of Electronic Commerce Research*, 20(1), 1-20.
- [6] Al-Abbadi, L., Bader, D., Mohammad, A., Al-Quran, A., Aldaihani, F., Al-Hawary, S., & Alathamneh, F. (2022). The effect of online consumer reviews on purchasing intention through product mental image. *International Journal of Data and Network Science*, 6(4), 1519-1530.
- [7] Tran, L. T. T. (2020). Online reviews and purchase intention: A cosmopolitanism perspective. *Tourism Management Perspectives*, 35, 100722.
- [8] Hà, N. T. (2016). Các yếu tố ảnh hưởng đến ý định mua sắm trực tuyến của người tiêu dùng Việt Nam: Nghiên cứu mở rộng thuyết hành vi có hoạch định. *VNU JOURNAL OF ECONOMICS AND BUSINESS*, 32(4).
- [9] M. Lutfullaeva, M. Medvedeva, Candidate of Physic-Mathematical Sciences, Associate Professor, E. Komotskiy, K. Spasov, PhD (2018). Optimization of Sentiment Analysis Methods for classifying text comments of bank customers.

- [10] Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017* (pp. 639-647). Springer Singapore.
- [11] Huang, H., Asemi, A., & Mustafa, M. B. (2023). Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions. *IEEE Access*.
- [12] Zahoor, K., Bawany, N. Z., & Hamid, S. (2020, November). Sentiment analysis and classification of restaurant reviews using machine learning. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-6). IEEE.
- [13] Bodapati, J. D., Veeranjanyulu, N., & Shareef, S. N. (2019). Sentiment Analysis from Movie Reviews Using LSTMs. *Ingénierie des Systèmes d Inf.*, 24(1), 125-129.
- [14] Jia Ke, Ying Wang, Mingyue Fan, Xiaojun Chen, Wenlong Zhang, Jianping Gou (2024). “Discovering e-commerce user groups from online comments: An emotional correlation analysis-based clustering method”, *Computers and Electrical Engineering*, 109035.
- [15] Nguyen. T. N. H., Ngoc. N. T. B.(2021). Nghiên cứu các yếu tố đến ý định mua mỹ phẩm của khách hàng nữ tại thành phố Hồ Chí Minh