TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP.HCM

Khoa Công Nghệ Thông Tin

----------ꕥ----------

BÁO CÁO CUỐI KÌ

Đề tài: Phân loại bình luận

***Giảng viên hướng dẫn*:** ThS. Trương Văn Thông

Môn **:** Công nghệ mới trong phát triển ứng dụng CNTT

Danh sách nhóm **:**

STT	MSSV	Tên
1	19501521	Phạm Thanh Hải
2	19495611	Võ Thị Thùy Trang
3	19499601	Võ Giang Khang

TP.HCM, ngày 11 tháng 12 năm 2022

LỜI MỞ ĐẦU

Thời đại kết nối và sức mạnh của hiệu ứng cộng đồng, truyền miệng luôn là một trong những phương thức marketing hiệu quả nhất. Ngày nay, các bình luận đánh giá của khách hàng về trải nghiệm của họ đối với một hàng hóa – dịch vụ trên các phương tiện truyền thông xã hội rất được chú trọng. Chúng là một nguồn tham khảo quan trọng, mang lại quyết định cho sự lựa chọn của khách hàng mới, và là cơ sở để xây dựng và cải tiến chất lượng dịch vụ nhằm gia tăng sự hài lòng và trung thành của khách hàng đối với doanh nghiệp. Trong nghiên cứu này, chúng tôi sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên trong việc thu thập và trích xuất thông tin bình luận trên văn bản tiếng Việt, thực nghiệm trên tập dữ liệu của bài toán dịch vụ khách sạn. Ứng dụng Deep Learning với các mô hình mạng Neural DNN, CNN, Bi-LSTM để phân lớp sắc thái bình luận là tích cực hay tiêu cực. Việc phân tích thống kê lại xem những nội dung bình luận, nhận xét đó là đánh giá tích cực hay tiêu cực về sản phẩm hoặc về thái độ phục vụ sẽ giúp cho doanh nghiệp biết được chất lượng sản phẩm, thái độ phục vụ và từ đó đưa ra những thay đổi hợp lý trong kinh doanh. Vì có rất nhiều lượt bình luận, trên các trang thương mại điện tử lớn có thể lên tới hàng chục triệu lượt bình luận trong một ngày nên việc phân tích bằng tay truyền thống là điều không thể. Đây là lúc các model machine learning thể hiện sức mạnh của mình.

Chúng em xin chân thành cảm ơn thầy Trương Văn Thông – giảng viên bộ môn Công nghệ mới trong phát triển ứng dụng CNTT lớp DHKHMT15A đã tận tình giảng dạy chúng em trong suốt thời gian vừa qua. Nhờ sự hướng dẫn, chỉ bảo và sửa lỗi của thầy để chúng em có thể hoàn thành được đề tài một cách tốt nhất. Một lần nữa nhóm em xin chân thành cảm ơn thầy!

MỤC LỤC

LỜI MỞ ĐẦU ..... - 2 -
CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN ĐỀ TÀI ..... - 4 -
1.1. Giới thiệu ..... - 4 -
1.2. Hướng tiếp cận và giải quyết bài toán ..... - 5 -
1.2.1. Cơ sở dữ liệu ..... - 5 -
1.2.2. Ý tưởng ..... - 6 -
1.3. Tổng kết chương ..... - 7 -
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT ..... - 8 -
2.1. Khái niệm Deep Learning ..... - 8 -
2.2. Xử lí ngôn ngữ tự nhiên - ngôn ngữ Tiếng Việt ..... - 9 -
2.3. Học máy vector hỗ trợ (Support Vector Machine, SVM) ..... - 9 -
2.4. Điểm khác biệt giữa Machine learning và Deep Learning ..... - 10 -
2.5. Azure App Service ..... - 12 -
2.6. Azure SQL ..... - 13 -
2.7. Bài toán phân lớp dữ liệu (Classification) ..... - 14 -
CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT ..... - 17 -
3.1. Chuẩn bị tập dữ liệu ..... - 17 -
3.2. Phân loại dữ liệu vào từng lớp ..... - 18 -
3.3. Làm sạch, tạo từ điển và trích xuất đặc trưng ..... - 19 -
3.4. Xây dựng mô hình phân loại ..... - 20 -
CHƯƠNG 4: TỔNG QUAN VỀ ỨNG DỤNG ..... - 21 -
4.1. Deploy ứng dụng lên Aure App Service ..... - 21 -
4.1.1 Chuẩn bị ứng dụng lên Aure ..... - 21 -
4.1.2 . Create a web app in Azure ..... - 22 -
4.1.3 . Deploy application code lên Azure ..... - 24 -
4.1.4 . Actions jobs trên github ..... - 24 -
4.2. Phân loại bình luận với SQL Azure ..... - 24 -
4.3. Giao diện ứng dụng ..... - 25 -
CHƯƠNG 5: KẾT QUẢ ĐẠT ĐƯỢC ..... - 26 -
5.1. Dữ liệu ..... - 26 -
5.2. Mô hình huấn luyện ..... - 26 -
5.3. Kết quả huấn luyện và đánh giá mô hình phân loại ..... - 29 -
5.4. Deploy web app lên Azure ..... - 29 -
5.5. Ưu điểm, nhược điểm và hướng phát triển ..... - 29 -
Tài liệu kham khảo ..... - 31 -
Bảng kế hoạch, phân công việc nhóm ..... - 32 -

CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN ĐỀ TÀI

1.1. Giới thiệu

Các doanh nghiệp trong lĩnh vực kinh doanh hằng năm chi ra một khoản chi phí vô cùng lớn trong việc hoàn thiện và nâng cao chất lượng dịch vụ. Nhưng để việc làm này thực sự hiệu quả với số tiền doanh nghiệp bỏ ra, họ phải cân nhắc đến cảm nhận thực của khách hàng đã trải nghiệm, vì điều đó phản ánh chính xác nhất tình trạng hiện tại của doanh nghiệp: làm tốt những mặt nào và còn hạn chế những điều gì?

Ứng dụng các mô hình Deep Learning vào việc xây dựng hệ thống phần mềm hỗ trợ trích xuất thông tin, phân loại và phân tích một cách tự động những dữ liệu nhận xét, đánh giá (review) trực tuyến của khách hàng ở dạng văn bản (ngôn ngữ tiếng Việt) về mức độ hài lòng.

Bài nghiên cứu sẽ tập trung giải quyết các mục tiêu cụ thể như sau:

- Xác định một review tích cực (Positive) hay tiêu cực (Negative)

- Tách đối tượng (được gom thành 2 nhóm đối tượng: nhân viên, sản phẩm) có xuất hiện trong một review

Đối tượng nghiên cứu:

- Lý thuyết xử lý ngôn ngữ tự nhiên với ngôn ngữ tiếng Việt, đánh giá mặt ngữ nghĩa trong câu

- Lý thiết về học máy (Deep Learning), lý thuyết khai phá dữ liệu (Data Mining) giúp tối ưu trí tuệ nhân tạo của việc phân tích và đưa ra kết quả.

Bài nghiên cứu thực hiện trong phạm vi các trang thương mại điện tử shopee có hỗ trợ tiếp nhận ý kiến phản hồi từ khách hàng bằng ngôn ngữ tiếng Việt.

Về mặt ý nghĩa khoa học, bài nghiên cứu thực nghiệm các giả thuyết về các kỹ thuật xử lý trích xuất dữ liệu tự động, xử lý ngôn ngữ tự nhiên, ... cũng như các kỹ thuật học máy, khai phá dữ liệu trong huấn luyện và trang bị khả năng tự học cho hệ thống.

1.2. Hướng tiếp cận và giải quyết bài toán

1.2.1. Cơ sở dữ liệu

Dữ liệu được thu thập bằng chương trình tự động, dữ liệu lấy từ trang thương mại điện tử shopee. Đây là phương pháp thu thập nội dung tự động từ các trang HTML của bất kỳ tài nguyên Internet bằng các chương trình hoặc mã lệnh đặc biệt. Với đối tượng và phạm vi nghiên cứu hướng đến là ngôn ngữ tiếng Việt.

Dữ liệu được cung cấp bởi shopee.com về các câu bình luận trong tiếng Việt trong đó bộ training dataset gồm 27694 câu bình luận, bộ testing dataset gồm 6924 câu bình luận.

Nhận thấy khi bình luận trên shopee thì người ta thường chỉ xoay quanh 2 vấn đề đó là bình luận về sản phẩm hoặc thái độ phục vụ của cửa hàng nên nhóm đã phân loại bình luận thành 5 label:

- Các bình luận chê sản phẩm như sản phẩm xấu, mỏng, không giống ảnh, chất lượng quá tệ sẽ cho vào label 0 - sản phẩm xấu chất lượng kém.
- Các bình luận vừa chê vừa khen sản phẩm hoặc không chê cũng không khen sản phẩm như sản phẩm ổn sẽ được phân loại thành label 1 – sản phẩm tạm chấp nhận.
- Các bình luận khen sản phẩm như sản phẩm đẹp, xịn, chất dày dặn, đúng mô tả, chất lượng tốt sẽ phân loại thành label 2- chất lượng sản phẩm tuyệt vời.
- Các bình luận gửi nhầm sản phẩm, đóng gói tệ, dịch vụ chăm sóc khách hàng kém, giao hàng chậm thì sẽ cho vào label 3 - cửa hàng phục vụ quá tệ.
- Các bình luận phục vụ nhiệt tình, giao hàng nhanh, thân thiện, đóng gói cẩn thận sẽ cho vào label 4 - cửa hàng phục vụ tốt chăm sóc khách hàng tuyệt vời.

	Y	X
0	0	sản_phẩm chỉ thừa nhiều
1	0	màu bên ngoài nhạt hơn trong ảnh đó áo thì nỉ ...
2	0	sản_phẩm mỏng vải xấu
3	0	màu quá xấu so với hình
4	0	tiền nào của nấy khi mặc vào ngực mình kiểu nó...
...	...	...
34613	4	giao hàng nhanh đóng_gói kỹ
34614	4	cửa_hàng giao hàng nhanh có thư cảm_ơn kèm the...
34615	4	săn giảm_giá cửa_hàng giao hàng siêu nhanh
34616	4	lần thứ 3 mua hàng ở cửa_hàng ưng cửa_hàng ạ l...
34617	4	vòng lưng chật cửa_hàng đã hỗ_trợ đổi hàng rất...

34618 rows × 2 columns

Hình 1.1. Dữ liệu của bài toán phân loại bình luận

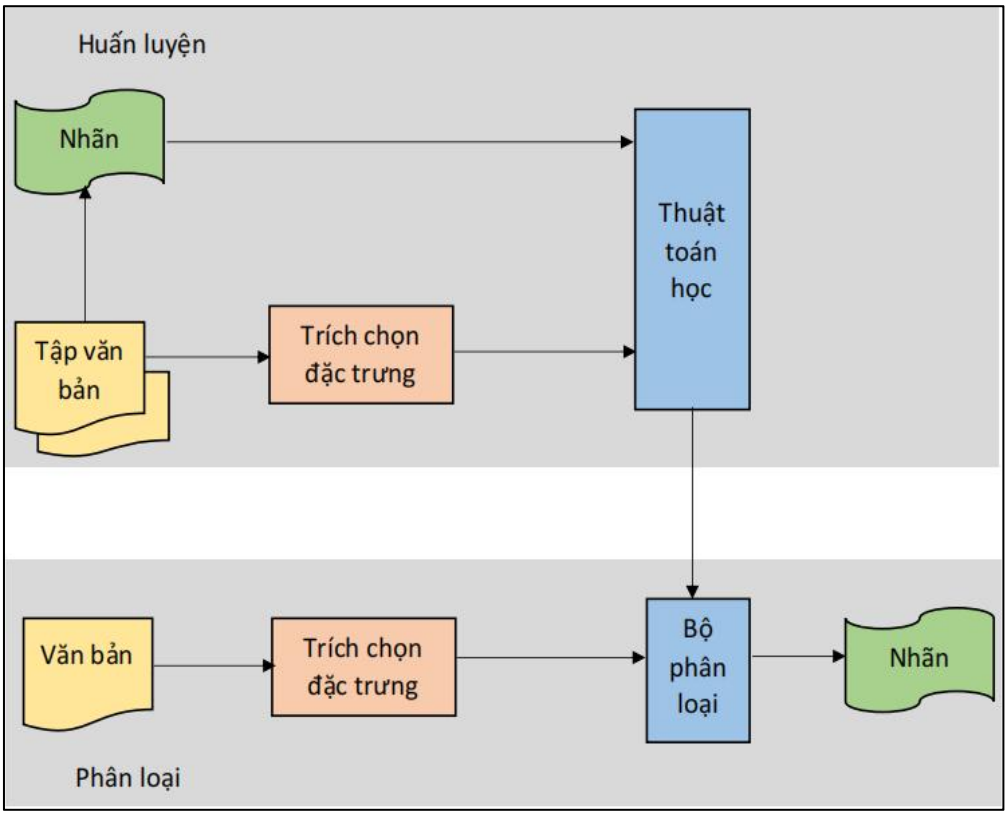
1.2.2. Ý tưởng

Sử dụng tập dữ liệu các câu bình luận được gán nhãn thủ công để huấn luyện mô hình máy học thuật toán cổ điển (Naïve bayes, KNN…), hay deep learning (Recurrent Neural Networks, Convolutional Neural Networks) để phân loại cảm xúc hay mức độ hài lòng của khách hàng.

Thuật toán cổ điển: đối với bài toán phân loại văn bản, có thể sử dụng mô hình Multinomial Naive Bayes và Bernoulli Naive Bayes.

Deep learning: sử dụng word embeddings giúp tìm ra mô hình không gian vector cho các từ. Hầu hết các nền tảng truyền thông mạng xã hội đều triển khai các hệ thống

phân tích dựa trên CNN và RNN để phân tích sắc thái văn bản và dự đoán chủ đề của văn bản.



Hình 1.2. Quá trình học phân loại văn bản

1.3. Tổng kết chương

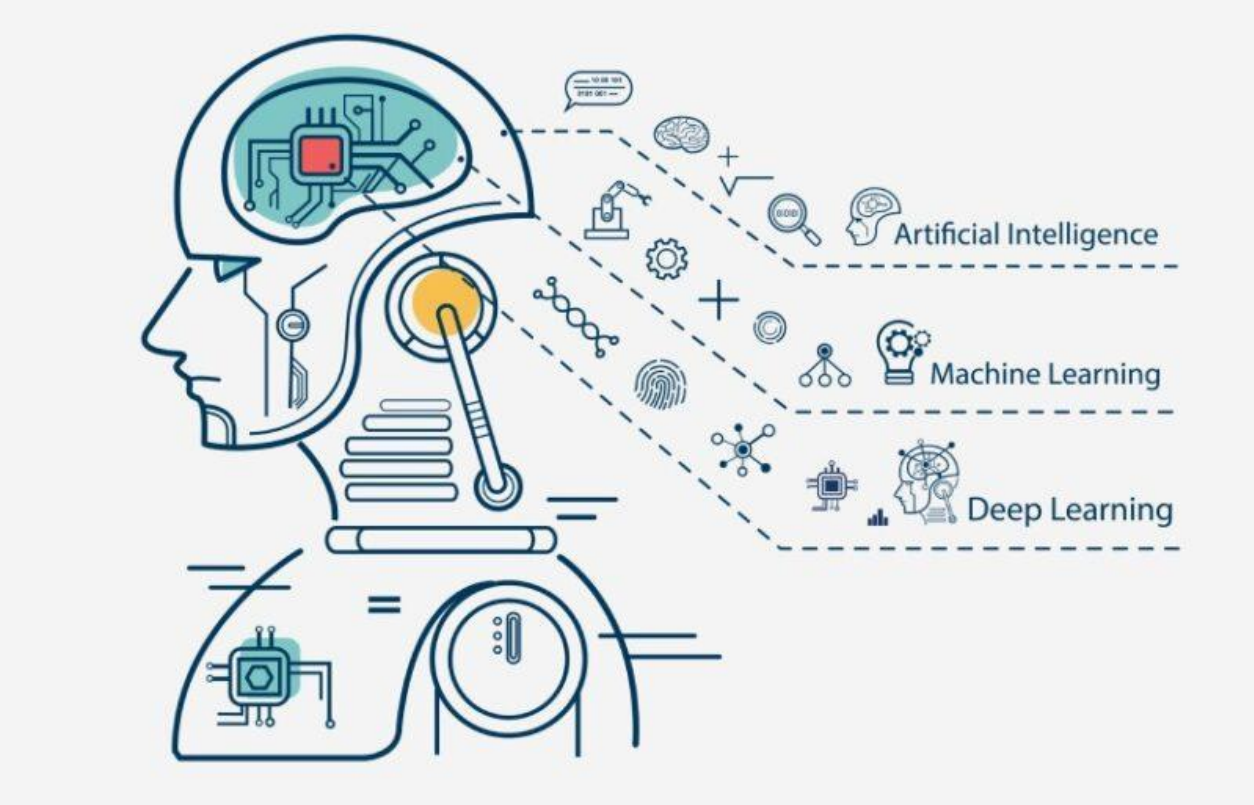
Qua chương này nhóm đã hiểu rõ hơn về bài toán phân loại bình luận, lên ý tưởng để giải bài toán. Để hiểu sâu hơn về bài toán nhóm sẽ qua chương Cơ sở lý thuyết để làm rõ về các mô hình Deep Learning được áp dụng vào để phân loại bình luận sản phẩm.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Khái niệm Deep Learning

Deep learning là một tập con của machine learning, về cơ bản là một mạng nơ-ron có ba lớp trở lên. Những mạng lưới thần kinh này cố gắng mô phỏng hành vi của não người, mặc dù không phù hợp với khả năng của nó cho phép nó "học" từ một lượng lớn dữ liệu.

Mặc dù mạng nơ-ron với một lớp duy nhất vẫn có thể đưa ra các dự đoán gần đúng, nhưng các lớp ẩn bổ sung có thể giúp tối ưu hóa và tinh chỉnh để có độ chính xác.



Deep learning thúc đẩy nhiều ứng dụng và dịch vụ trí tuệ nhân tạo (AI) nhằm cải thiện tự động hóa, thực hiện các tác vụ phân tích và vật lý mà không cần sự can thiệp của con người. Công nghệ deep learning nằm đằng sau các sản phẩm và dịch vụ hàng ngày (chẳng hạn như trợ lý kỹ thuật số, điều khiển TV hỗ trợ giọng nói và phát hiện gian lận thẻ tín dụng) cũng như các công nghệ mới nổi (chẳng hạn như ô tô tự lái).

2.2. Xử lí ngôn ngữ tự nhiên - ngôn ngữ Tiếng Việt

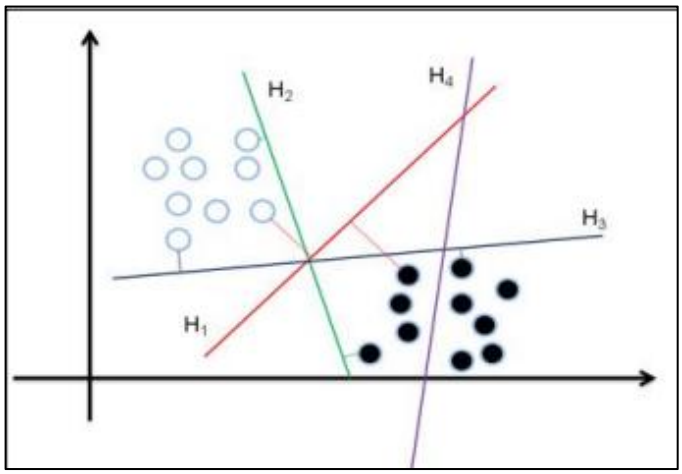
Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ - công cụ hoàn hảo nhất của tư duy và giao tiếp. [6] Xử lý ngôn ngữ là một kỹ thuật quan trọng nhằm giúp máy tính hiểu được ngôn ngữ của con người, qua đó hướng dẫn máy tính thực hiện và giúp đỡ con người trong những công việc có liên quan đến ngôn ngữ như: dịch thuật, phân tích dữ liệu văn bản, nhận dạng tiếng nói, tìm kiếm thông tin, ... [1].

2.3. Học máy vector hỗ trợ (Support Vector Machine, SVM)

SVM là phương pháp phân loại sẽ nhận một tập dữ liệu đầu vào và xác định từng phần tử thuộc về lớp nào trong hai lớp cho sẵn dựa trên tổ hợp tuyến tính của các giá trị đặc trưng. Cho trước một tập huấn luyện được biểu diễn trong không gian vector, SVM sẽ tìm ra một siêu mặt phẳng tối ưu nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng. Ví dụ trong Hình 2.1, ta có thể thấy sẽ có rất nhiều mặt phẳng có khả năng phân tách dữ liệu thành hai lớp (H1, H2 và H3) và cũng có những mặt phẳng không thể phân tách (H4). SVM sẽ tìm trong các mặt phẳng phân tách đó và chọn ra mặt phẳng tối ưu nhất. Với phương trình mặt phẳng trong không gian có dạng:

$$w.x - b = 0 \qquad (1)$$

Trong đó, w là vector pháp tuyến của mặt phẳng, x là vector các giá trị đặc trưng, b là hệ số tự do. Mục tiêu của SVM là tìm w và b sao cho đạt được cực đại khoảng cách từ mặt phẳng phân tách này đến các điểm gần nhất của mỗi lớp.



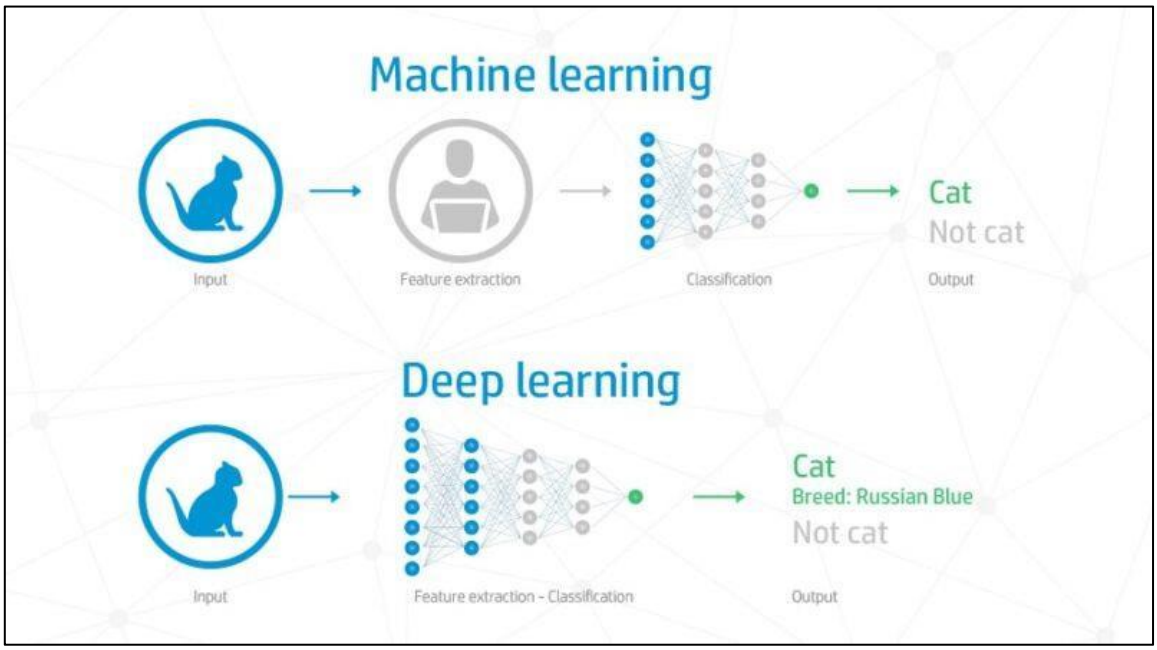
Hình 2.1. Một số mặt siêu phẳng phân tách SVM

2.4. Điểm khác biệt giữa Machine learning và Deep Learning

Các thuật toán machine learning tận dụng dữ liệu có cấu trúc, được gắn nhãn để đưa ra dự đoán – nghĩa là các tính năng cụ thể được xác định từ dữ liệu đầu vào cho mô hình và được tổ chức thành các bảng. Điều này không nhất thiết có nghĩa là nó không sử dụng dữ liệu phi cấu trúc, nó chỉ có nghĩa là nếu có, nó thường trải qua một số xử lý trước để tổ chức nó thành một định dạng có cấu trúc.

Deep learning loại bỏ một số quá trình xử lý trước dữ liệu thường liên quan đến machine learning. Các thuật toán này có thể nhập và xử lý dữ liệu phi cấu trúc, như văn bản và hình ảnh, đồng thời nó tự động hóa việc trích xuất tính năng, loại bỏ một số phụ thuộc vào các chuyên gia con người.

Ví dụ: giả sử có một bộ ảnh về các vật nuôi khác nhau và muốn phân loại theo “mèo”, “chó”, “hamster”, v.v. Các thuật toán deep learning có thể xác định đặc điểm nào (ví dụ: tai) là quan trọng nhất để phân biệt từng loài động vật với loài khác. Trong machine learning, hệ thống phân cấp các tính năng này được thiết lập thủ công bởi một chuyên gia là con người.



Sau đó, thông qua các quá trình giảm độ dốc và lan truyền ngược, thuật toán deep learning sẽ tự điều chỉnh và phù hợp với độ chính xác, cho phép nó đưa ra dự đoán về một bức ảnh động vật mới với độ chính xác cao hơn.

Các mô hình machine learning và deep learning cũng có khả năng thực hiện các kiểu học khác nhau, thường được phân loại là học có giám sát, học không giám sát và học tăng cường. Học có giám sát sử dụng các tập dữ liệu được gắn nhãn để phân loại hoặc đưa ra dự đoán, điều này đòi hỏi sự can thiệp của con người để gắn nhãn dữ liệu đầu vào một cách chính xác.

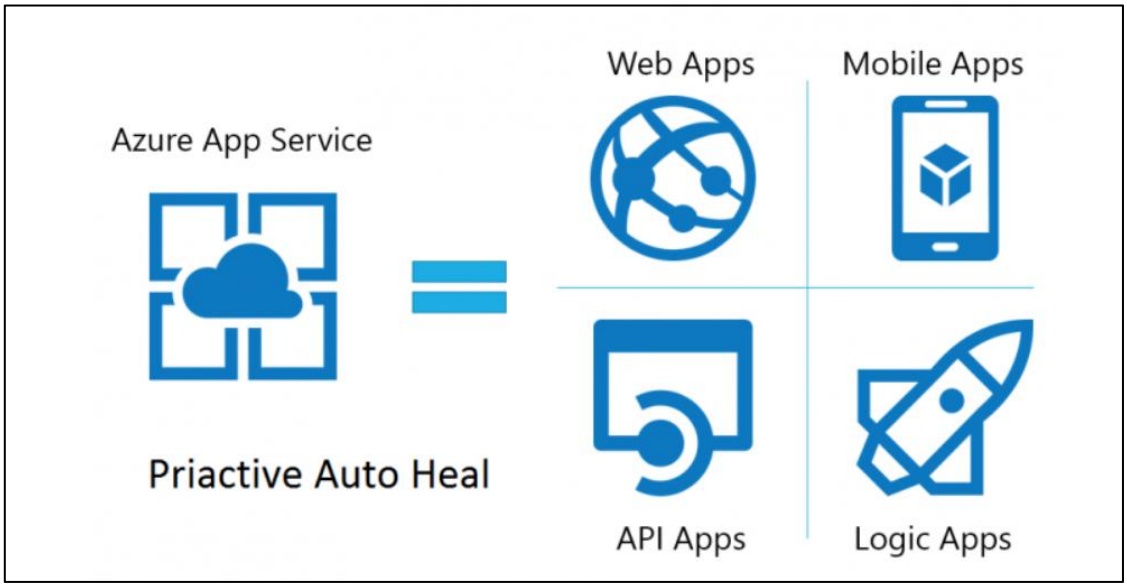
Ngược lại, học không giám sát không yêu cầu tập dữ liệu được gắn nhãn và thay vào đó, nó phát hiện các mẫu trong dữ liệu, nhóm chúng theo bất kỳ đặc điểm phân biệt nào. Học tăng cường là một quá trình trong đó một mô hình học cách trở nên chính xác hơn để thực hiện một hành động trong môi trường dựa trên phản hồi nhằm tối đa hóa dữ liệu.

2.5. Azure App Service

Azure App Service là một dịch vụ dựa trên HTTP để lưu trữ các ứng dụng web, API REST và mobile backends. Bạn có thể phát triển các ứng dụng bằng ngôn ngữ lập trình yêu thích của mình, chẳng hạn như .NET, .NET Core, Java, Ruby, Node.js, PHP hoặc Python. Các ứng dụng có thể chạy và mở rộng một cách dễ dàng trên cả 2 môi trường là Windows và Linux. Riêng với môi trường Linux, bạn có thể tìm hiểu rõ hơn về App Service trên Linux.

App Service không chỉ bổ sung sức mạnh của Microsoft Azure cho ứng dụng của bạn, chẳng hạn như bảo mật, cân bằng tải, tự động cân bằng và quản lý tự động, mà bạn còn có thể tận dụng các khả năng DevOps của nó chẳng hạn như triển khai liên tục từ Azure DevOps, GitHub, Docker Hub và các nguồn khác, hệ thống quản lý gói, môi trường dàn dựng, tên miền tùy chỉnh và chứng chỉ SSL.

Với Azure App Service, bạn sẽ chỉ phải chi trả cho lượng tài nguyên của Azure mà bạn sử dụng. Lượng tài nguyên Azure mà bạn sử dụng được xác định bởi gói App Service mà bạn lựa chọn để sử dụng cho ứng dụng của mình.



2.6. Azure SQL

Azure SQL là một nhóm các sản phẩm được quản lý, bảo mật và thông minh sử dụng công cụ cơ sở dữ liệu SQL Server trong đám mây Azure.

- **Cơ sở dữ liệu Azure SQL**: Hỗ trợ các ứng dụng đám mây hiện đại trên dịch vụ cơ sở dữ liệu được quản lý, thông minh, bao gồm tính toán không cần máy chủ.
- **Phiên bản Azure SQL được quản lý**: Hiện đại hóa các ứng dụng SQL Server hiện tại của bạn ở quy mô lớn với phiên bản được quản lý hoàn toàn thông minh dưới dạng dịch vụ, với gần như 100% tính năng tương đương với công cụ cơ sở dữ liệu SQL Server. Tốt nhất cho hầu hết các lần di chuyển lên đám mây.
- **Máy chủ SQL trên máy ảo Azure**: Nâng và chuyển khối lượng công việc Máy chủ SQL của bạn một cách dễ dàng và duy trì khả năng tương thích 100% của Máy chủ SQL và quyền truy cập cấp hệ điều hành.

Azure SQL được xây dựng dựa trên công cụ SQL Server quen thuộc, vì vậy bạn có thể di chuyển các ứng dụng một cách dễ dàng và tiếp tục sử dụng các công cụ, ngôn ngữ và tài nguyên mà bạn quen thuộc. Kỹ năng và kinh nghiệm của bạn được chuyển sang đám mây, vì vậy bạn có thể làm được nhiều hơn nữa với những gì bạn đã có.

2.7. Bài toán phân lớp dữ liệu (Classification)

Bài toán phân lớp (classification) là một trong những bài toán lớn trong lĩnh vực Machine Learnig (ML). Classification là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp (model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu. Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phần lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào.

Có nhiều bài toán phân lớp dữ liệu như phân lớp nhị phân (binary), phân lớp đa lớp (multiclass), phân lớp đa trị.

Bài toán phân lớp nhị phân là bài toán gắn nhãn dữ liệu cho đối tượng vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không có các đặc trưng (feature) của bộ phân lớp

Bài toán phân lớp đa lớp là quá trình phân lớp dữ liệu với số lượng lớp lớn hơn hai. Như vậy với từng dữ liệu chúng ta phải xem xét và phân lớp chúng vào những lớp khác nhau chứ không phải là hai lớp như bài toán phân lớp nhị phân.

Và thực chất bài toán phân lớp nhị phân là một bài toán đặt biệt của phân lớp đa lớp.

Ứng dụng của bài toán này được sử dụng rất nhiều và rộng rãi trong thực tế ví dụ như bài toán nhận dạng khuôn mặt, nhận diện giọng nói, phát hiện email spam…

Để xây dựng được mô hình phân lớp và đánh giá được mô hình chúng ta phải trải quá các quá trình như dưới đây:

Bước 1: Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction)

Công đoạn này được xem là công đoạn quan trọng trong các bài toán về Machine Learning. Vì đây là input cho việc học dể tìm ra mô hình của bài toán.

Chúng ta phải biết cần chọn ra những đặt trưng tốt (good feature) của dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu (noise). Ước lượng số chiều của dữ liệu bao nhiêu là tốt hay nói cách khác là chọn bao nhiêu feature. Nếu số chiều quá lớn gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn giữ được độ chính xác của dữ liệu (reduce demension).

Ở bước này chúng ta cũng chuẩn bị bộ dữ liệu để test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập datasets thành hai phần, một phần phục vụ cho training (training datasets) và phần còn lại phục vụ cho mục đích testing trên mô hình (testing dataset). Có hai cách thường sử dụng trong cross-validation là splitting và k-fold.

Bước 2: Xây dựng mô hình phân lớp (classifier model)

Mục đích của mô hình huấn luyện là tìm ra hàm f(x) và thông qua hàm f tìm được để chúng ta gán nhãn cho dữ liệu. Bước này thường được gọi là học hay training.

$$f(x) = y$$

Trong đó: x là các feature hay input đầu vào của dữ liệu, y là nhãn lớp hay output đầu ra.

Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát (supervised learning) như KNN, Neural Network, SVM, Decision Tree, Navie Bayers.

Bước 3: Kiểm tra dữ liệu với mô hình (make prediction)

Sau khi đã tìm được mô hình phân lớp ở bước 2, thì ở bước này chúng ta sẽ đưa vào các dữ liệu mới để kiểm tra trên mô hình phân lớp.

Bước 4: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất

Bước cuối cùng chúng ta sẽ đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu testing và dữ liệu traning thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn của chúng ta thì phải thay đổi các tham số (turning parameter) của các thuật toán học để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. Và cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán của chúng ta.

CHƯƠNG 3: PHƯƠNG PHÁP ĐỀ XUẤT

Bài toán phân loại bình luận sản phẩm được đánh giá như bài toán phân loại và chúng sẽ được giải quyết bằng máy học

3.1. Chuẩn bị tập dữ liệu

Chúng tôi đã sử dụng tập dữ liệu là những đánh giá về sản phẩm từ trang web bán hàng trực tuyến của Việt Nam. Cụ thể là trang thương mại điện tử shopee.com. Tập dữ liệu chúng tôi sử dụng gồm 34.617 đánh giá. Tuy nhiên, những đánh giá từ các trang web thường được viết dưới dạng ngôn ngữ không phải là tiếng Việt chuẩn ("wá", "hok", "ko"). Chính vì vậy, chúng tôi đã thực hiện bước tiền xử lý là chuẩn hóa lại những từ như trên ("wá"→ "quá", "hok" và "ko" →"không").

2| 10 điểm cho chất lượng hàng rẻ và rất xinh chất vải mềm giao hàng nhanh mua đúng đợt giảm giá nên giá cũng mềm lắm nhé sẽ ủ
0, 3| giao hàng chậm áo hai dây trắng không có nút điều chỉnh dây như áo đen và áo đỏ muốn mặc được phải mang ra thợ sửa
1| mình cao 1m5 mặc trùm mông luôn nhé áo siu siu đẹp luôn phần tay bồng dễ thương hột me phần màu trắng dưới áo mọi người kh
1 lớp bông mỏng mặc vào rất êm **mn => mọi người** nên thử ủng hộ cửa hàng nha
2| mua lúc **flash => sự kiện** giảm giá nên rất rẻ nhận được mã nghìn sau khi thanh toán sử dụng không gặp tư vấn để gì
0, 3| **shop => cửa hàng** xác nhận đơn nhưng ok không soạn hàng để giao rồi lại báo là hết màu hết **size => kích cỡ** mình không hủy đ
hàng thì mặc bị rộng dù chung kích cỡ 38
1| hình ảnh mang tính chất minh **hoạ => họa** vải hơi mỏng nhưng ok tiền nào của đấy hình in không đẹp nhìn lộm cộm lắm
1| hình ảnh và phim chỉ mang tính chất lấy lâu xu chất lượng tạ m **đc => được**
0, 4| hình ảnh mang tính chất minh họa mình luôn tin tưởng và ưng đồ của **gu => GUMAC** nhưng ok áo đợt này xinh làm mình thật s
giao lại áo không bị lỗi lại bị xù với sờn nhiều có chỗ sờn tưởng chừng không biết chất lượng cửa hàng giảm hay sao nữa
2| quên không chụp nên để tạm ảnh chất mượt khá được mặc không bị ngứa phù hợp giá tiền
1| áo có khá nhiều chỉ thừa ở cổ áo hình in đẹp ổn **so => sơ** với giá tiền nên mua
5| đóng gói chắc chắn thời gian giao hàng rất nhanh **shop => cửa hàng** thân thiện
0| hàng kiểu bị thiếu vải nối thêm rất xấu đã thế còn thừa nhiều chỉ nói chung rất xấu ống quá giọng nói đúng là tiền nào của đấy
2| hàng thì phù hợp với số tiền ổn định áo đẹp không lỗi ủng hộ **shop => cửa hàng**
0, 3| vải hình in giống hình mẫu **form => dáng** áo xấu cục ngủn giao hàng
0, 3| vải hình in giống hình mẫu **form => dáng** áo xấu cục ngủn giao hàng
1| vải hơi mỏng q tí nhưng ok **so => sơ** với giá tiền thì trên cả tuyệt dời
2| hình ảnh mang tính chất kiếm xu hàng này xinh rất đẹp nha giá lại còn rẻ nữa nha vãi cũng rất mềm yêu cửa hàng lắm cơ sẽ quay l
0| **0 => không** giống hình nha m n hơi thất vọng ạ
0| 1 áo thì bị rách 1 lỗ 1 áo thì bị rách tay áo

Hình 3.1: Tiền xử lí dữ liệu

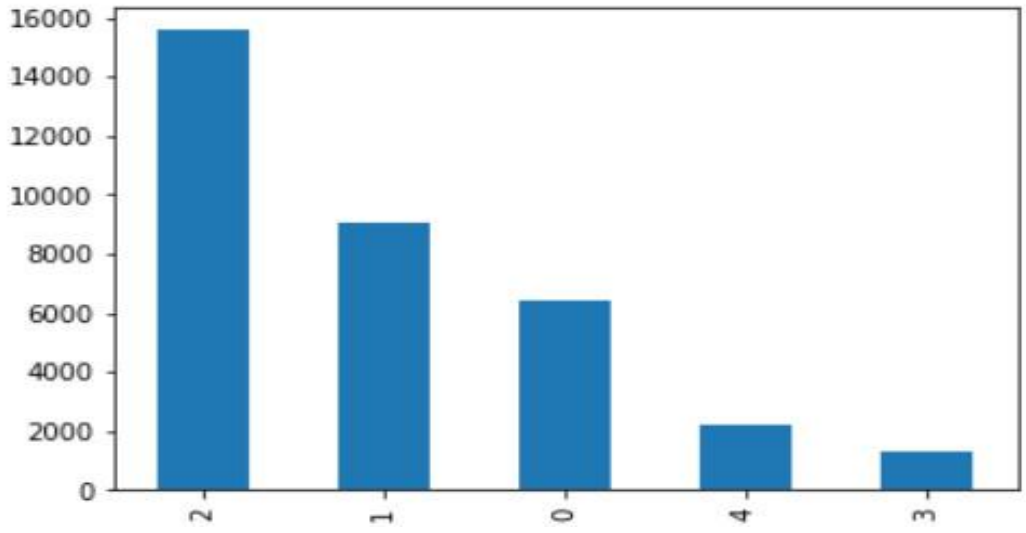
được	được
gia	giá
up	lên
ái	cái
mụt	một
ao	áo
so	_
az	a đến z
đay	đây
dung	đúng
hon	hơn
tam	tạm
nham	nhận
san	sản
khaby	khá
very much	rất nhiều
qtqd	
ròi	rồi
nhung	nhưng
ma	mà
vua	vừa

Hình 3.2: Tiền xử lí dữ liệu

3.2. Phân loại dữ liệu vào từng lớp

Sau khi thu thập các câu bình luận và xem qua tất cả bình luận thì thấy được đa số các câu bình luận nhận xét về chất lượng sản phẩm (class 0-1-2) và đánh giá dịch vụ giao hàng và chất lượng phục vụ của shop (class 3-4) sau đó tự gán nhãn dữ liệu một cách thủ công cho từng câu bình luận đó.

→Sau khi phân loại thủ công thì thu được: có 6425 bình luận label 0, 9073 bình luận label 1, 15562 bình luận label 2, 1322 bình luận label 3, 2519 bình luận label 4.



Hình 3.3: Biểu diễn số lượng của dữ liệu ở từng lớp

3.3. Làm sạch, tạo từ điển và trích xuất đặc trưng

Làm sạch dữ liệu: Bước này tiến hành làm sạch dữ liệu trước khi bắt đầu huấn luyện trên tập dữ liệu, bao gồm một số công đoạn:

- Xử lý ngôn ngữ tự nhiên như kiểm tra chính tả, xóa các kí tự đặc biệt
- Tách từ dùng thư viện underthesea.
- Loại bỏ từ dừng (Stop words): Từ dừng (stop words) là những từ xuất hiện nhiều trong tất cả các văn bản thuộc mọi thể loại trong tập dữ liệu, hay những từ chỉ xuất hiện trong một và một vài văn bản. Nghĩa là stop word là những từ xuất hiện quá nhiều lần và quá ít lần, vì thế nó không có ý nghĩa và không chứa thông tin đáng giá để sử dụng. Trong phân loại văn bản, sự xuất hiện của stop words không những không giúp gì trong việc đánh giá phân loại mà còn nhiễu và giảm độ chính xác của quá trình phân loại (như các từ: thì, là, mà, và, hoặc, bởi…). Ví dụ: câu "xử_lý ngôn_ngữ tự_nhiên (nlp) là 1 nhánh của trí_tuệ nhân_tạo" khi loại bỏ từ dừng thành "xử_lý ngôn_ngữ tự_nhiên (nlp) nhánh trí_tuệ nhân_tạo". Trong nghiên cứu, nhóm

tác giả sử dụng phương pháp loại bỏ từ dừng bằng cách tiến hành xóa các từ ít xuất hiện (< 20 lần).

- Tiếp theo tạo từ điển

Trích xuất đặc trưng: bước này sẽ chọn ra các đặc trưng tiêu biểu (chính là các từ khóa - Keywords) có tính đại diện cho tập dữ liệu để làm đầu vào (Input) cho thuật toán phân loại. Dùng Word Embedding:

- Chuyển mỗi từ trong từ điển về một vector N chiều, với N = 300.
- Sử dụng thuật toán Skip-Gram.

3.4. Xây dựng mô hình phân loại

Huấn luyện mô hình phân loại bình luận: giai đoạn này nhằm mục đích xác định một bình luận, nhận xét của khách hàng là sản phẩm xấu chất lượng kém, sản phẩm tạm chấp nhận, chất lượng sản phẩm tuyệt vời, cửa hàng phục vụ quá tệ và cửa hàng phục vụ tốt chăm sóc khách hàng tuyệt vời.

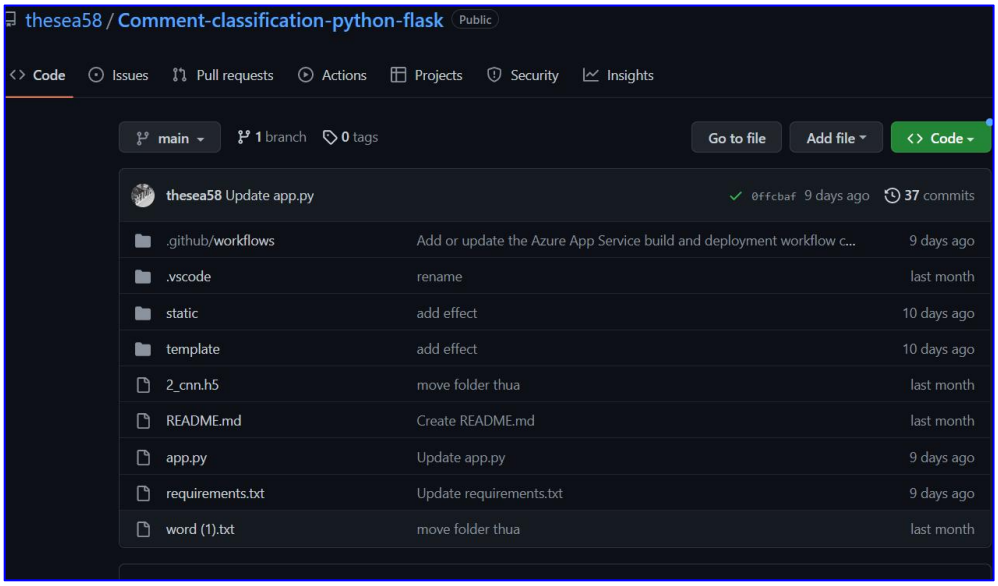
Chúng tôi lần lượt áp dụng 7 giải thuật: SVM, CNN, Bi-LSTM, RNN, CNN-BiLSTM, BahdanauAttention, TransformerAttention để xây dựng mô hình phân loại. Qua đó chúng tôi sẽ so sánh kết quả nhận được từ các giải thuật khác nhau.

CHƯƠNG 4: TỔNG QUAN VỀ ỨNG DỤNG

4.1. Deploy ứng dụng lên Aure App Service

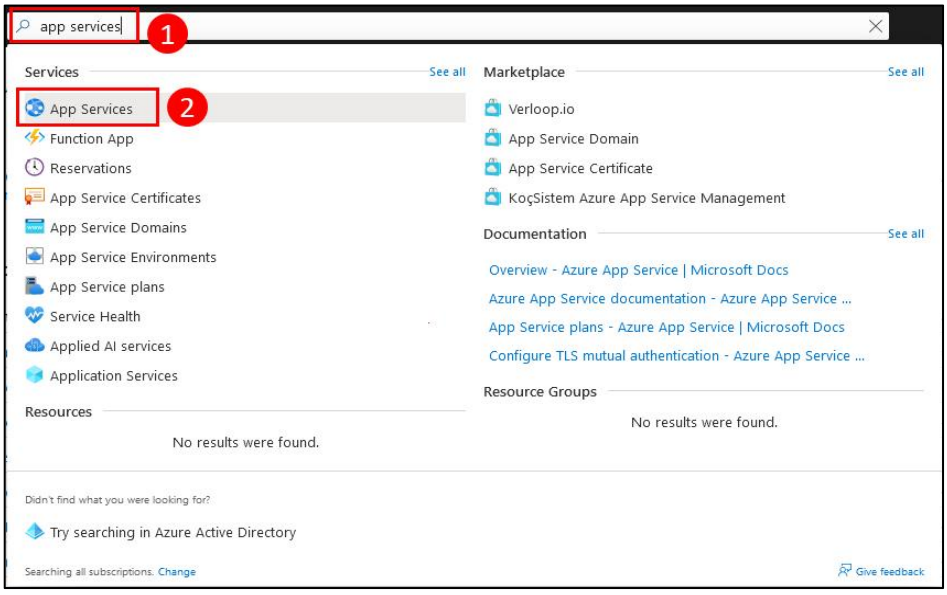
4.1.1 . Chuẩn bị ứng dụng lên Aure

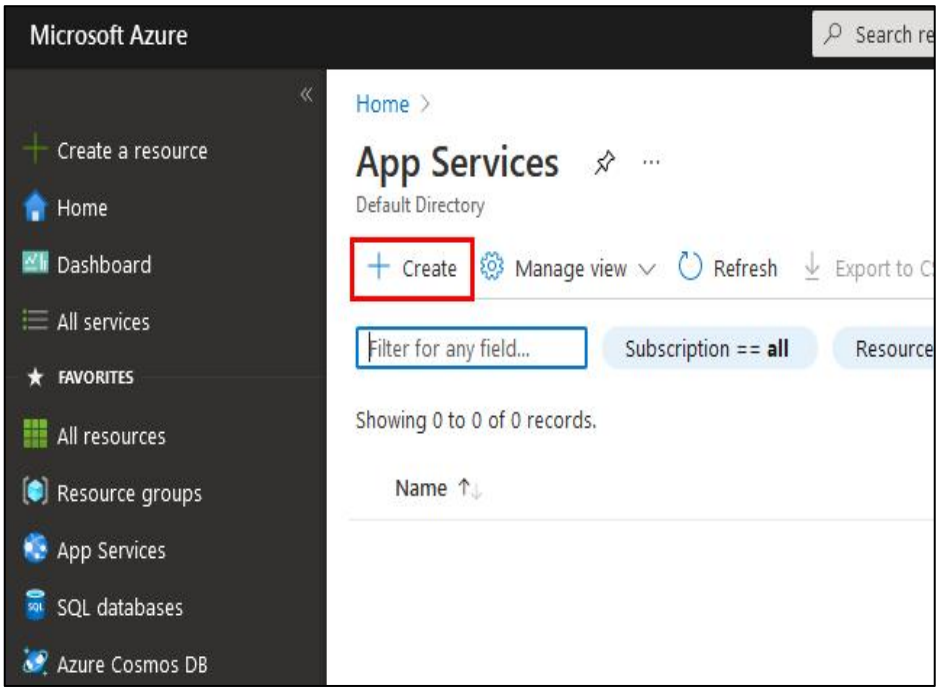
Web python được nhóm viết sẽ đưa source lên Github để tiến hành deploy web app lên Aure:

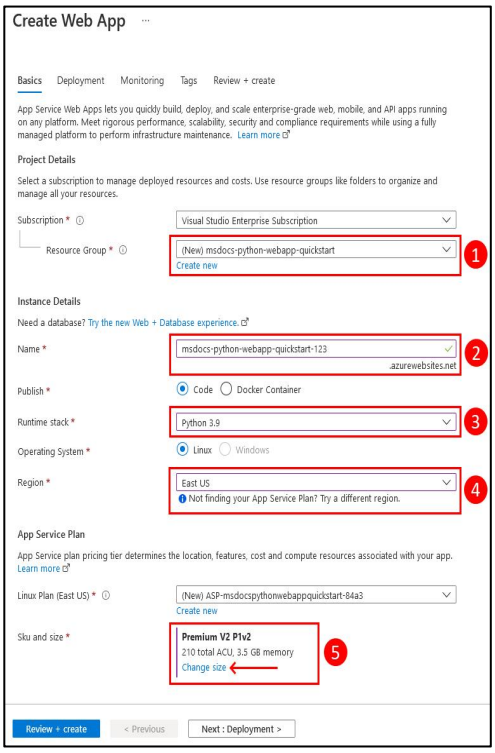


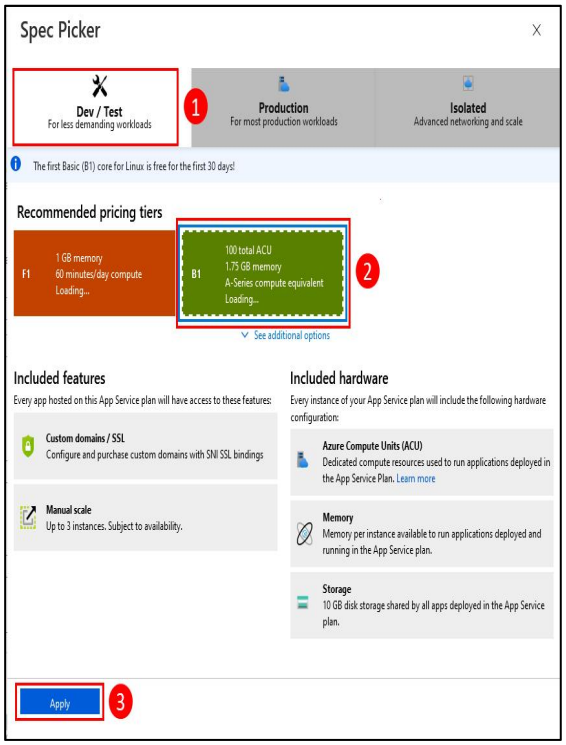
Link Github của web app: https://github.com/thesea58/Comment-classification-python-flask

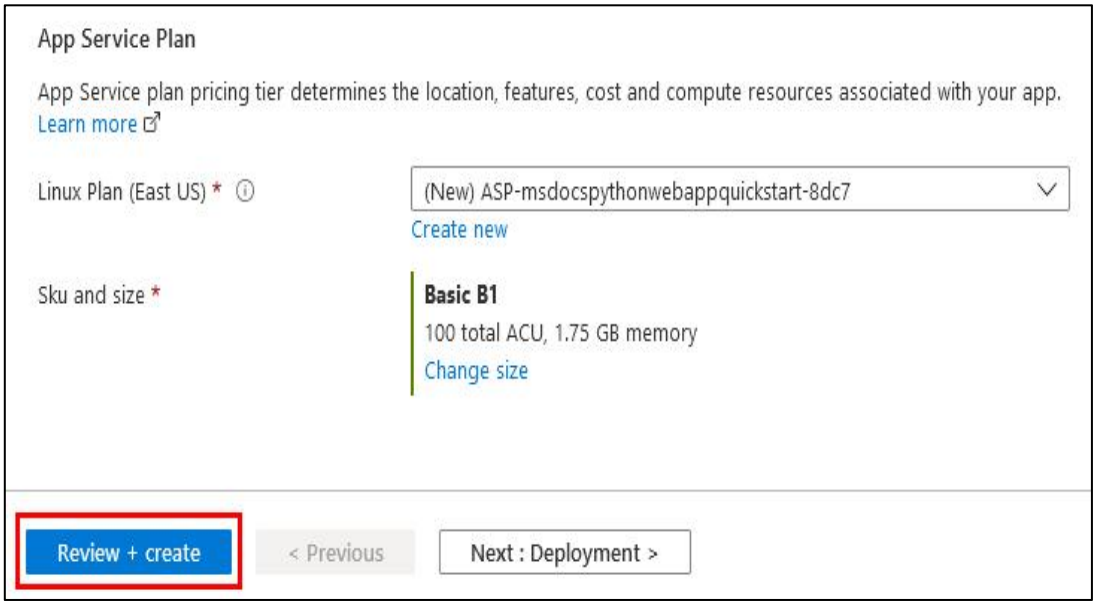
4.1.2. Create a web app in Azure





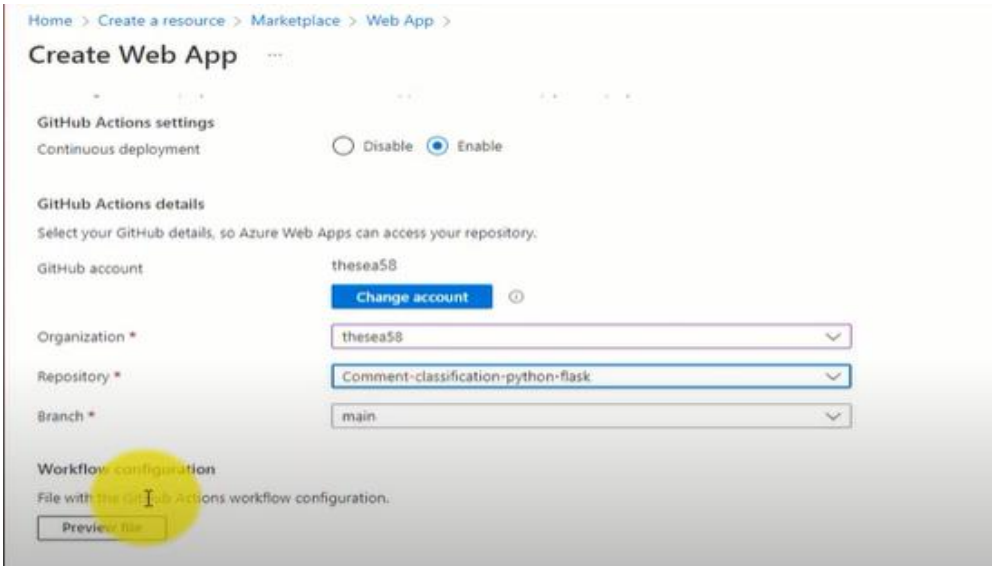

Create Web App
Basics
Deployment
Monitoring
Tags
Review + create
App Service Web Apps lets you quickly build, deploy, and scale enterprise-grade web, mobile, and API apps running on any platform. Meet rigorous performance, scalability, security and compliance requirements while using a fully managed platform to perform infrastructure maintenance. Learn more
Project Details
Select a subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.
Subscription *
Visual Studio Enterprise Subscription
Resource Group *
(New) msdocs-python-webapp-quickstart
Create new
Instance Details
Need a database? Try the new Web + Database experience.
Name *
msdocs-python-webapp-quickstart-123
.azurewebsites.net
Publish *
Code
Docker Container
Runtime stack *
Python 3.9
Operating System *
Linux
Windows
Region *
East US
Not finding your App Service Plan? Try a different region.
App Service Plan
App Service plan pricing tier determines the location, features, cost and compute resources associated with your app.
Learn more
Linux Plan (East US) *
(New) ASP-msdocspythonwebappquickstart-b4a3
Create new
Sku and size *
Premium V2 P1v2
210 total ACU, 3.5 GB memory
Change size
Review + create
< Previous
Next : Deployment >
1
2
3
4
5


Spec Picker
Dev / Test
For less demanding workloads
Production
For most production workloads
Isolated
Advanced networking and scale
The first Basic (B1) core for Linux is free for the first 30 days!
Recommended pricing tiers
F1
1 GB memory
60 minutes/day compute
Loading...
B1
100 total ACU
1.75 GB memory
A-Series compute equivalent
Loading...
See additional options
Included features
Every app hosted on this App Service plan will have access to these features:
Custom domains / SSL
Configure and purchase custom domains with SNI SSL bindings
Manual scale
Up to 3 instances. Subject to availability.
Included hardware
Every instance of your App Service plan will include the following hardware configuration:
Azure Compute Units (ACU)
Dedicated compute resources used to run applications deployed in the App Service Plan. Learn more
Memory
Memory per instance available to run applications deployed and running in the App Service plan.
Storage
10 GB disk storage shared by all apps deployed in the App Service plan.
Apply
1
2
3


App Service Plan
App Service plan pricing tier determines the location, features, cost and compute resources associated with your app.
Learn more
Linux Plan (East US) *
(New) ASP-msdocspythonwebappquickstart-8dc7
Create new
Sku and size *
Basic B1
100 total ACU, 1.75 GB memory
Change size
Review + create
< Previous
Next : Deployment >

4.1.3. Deploy application code lên Azure

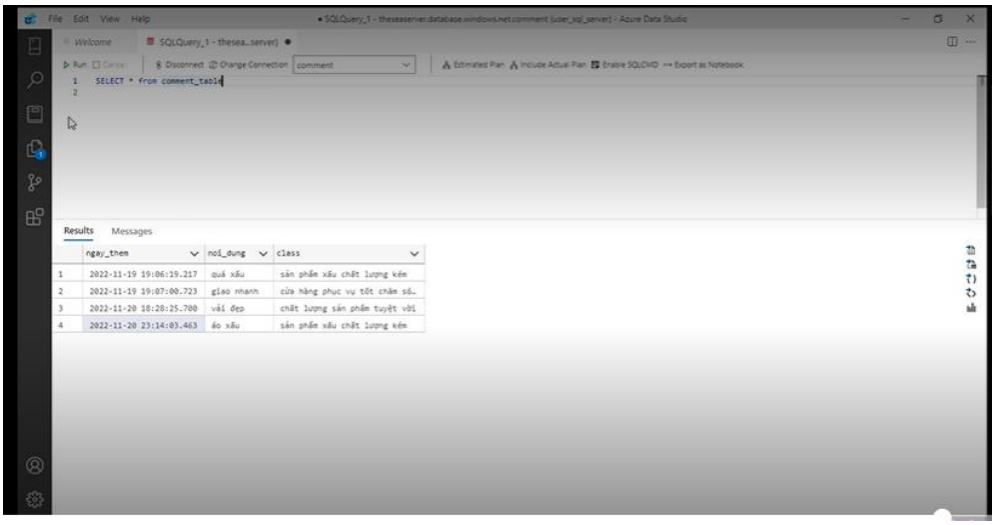
Sử dụng continuous deploy để tạo WorkFlows để install và deploy lên web app



4.1.4. Actions jobs trên github

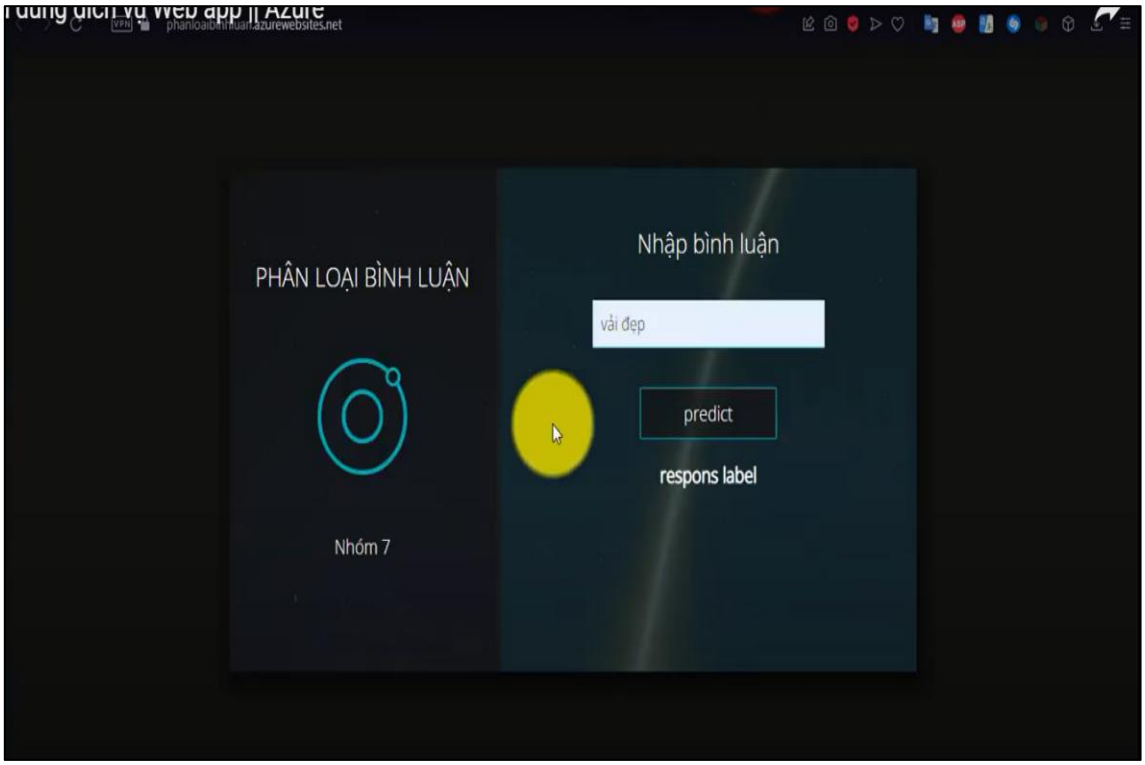
Chờ đợi deploy hoàn thành và truy cập vào web app để kiểm tra.

4.2. Phân loại bình luận với SQL Azure



4.3. Giao diện ứng dụng

Dưới đây là giao diện đề tài phân loại bình luận, gồm : tên đề tài, tên nhóm, tiêu đề, ô text để gõ bình luận vào và ô predict để phân loại bình luận.



CHƯƠNG 5: KẾT QUẢ ĐẠT ĐƯỢC

5.1. Dữ liệu

Nghiên cứu đã tiến hành thu thập dữ liệu bằng chương trình tự động, dữ liệu lấy từ trang thương mại điện tử shopee. Đây là phương pháp thu thập nội dung tự động từ các trang HTML của bất kỳ tài nguyên Internet bằng các chương trình hoặc mã lệnh đặc biệt. Với đối tượng và phạm vi nghiên cứu hướng đến là ngôn ngữ tiếng Việt. Do đó, dữ liệu chỉ sử dụng những bình luận của khách hàng bằng tiếng Việt. Tiếp đến, nghiên cứu đã tiến hành tiền xử lý dữ liệu bằng cách loại bỏ những dữ liệu khuyết, những nhận xét không chứa đựng thông tin cần thiết để tiến hành bước xử lý tiếp theo.

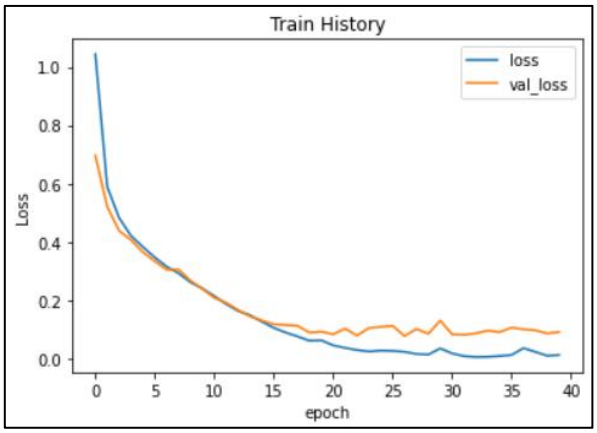
Dữ liệu thu thập đạt được 80% so với dự định ban đầu gồm 34.618 đánh giá

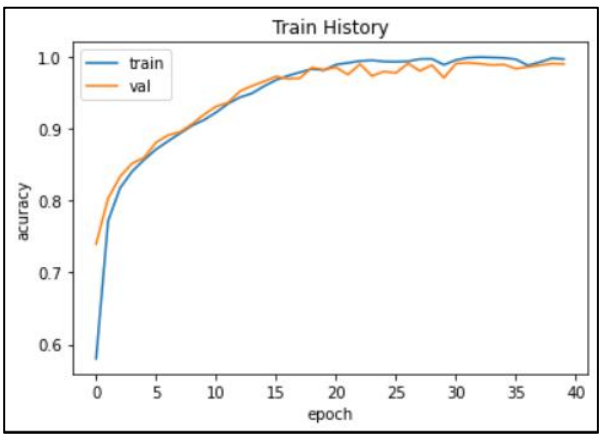
Được lưu trữ trên Google Drive theo link sau:
https://docs.google.com/spreadsheets/d/1mmnjXccKId6iKYxGx_lkdNWU0HqYuDR6/edit?usp=sharing&ouid=101988529137233457570&rtpof=true&sd=true

5.2. Mô hình huấn luyện

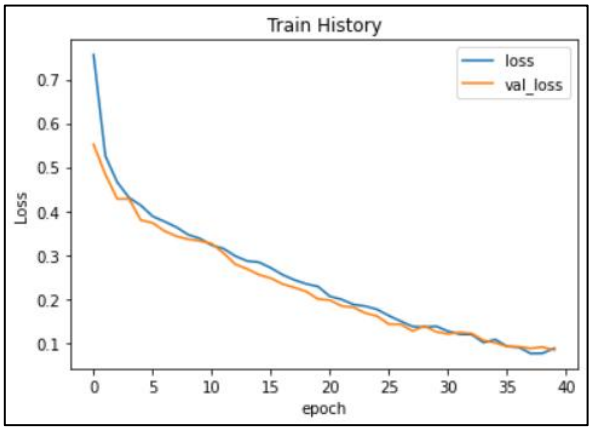
Xây dựng thành công các mô hình phân loại bình luận sản phẩm từ dataset thu thập được. Với loss và accuracy lần lượt như các hình sau:

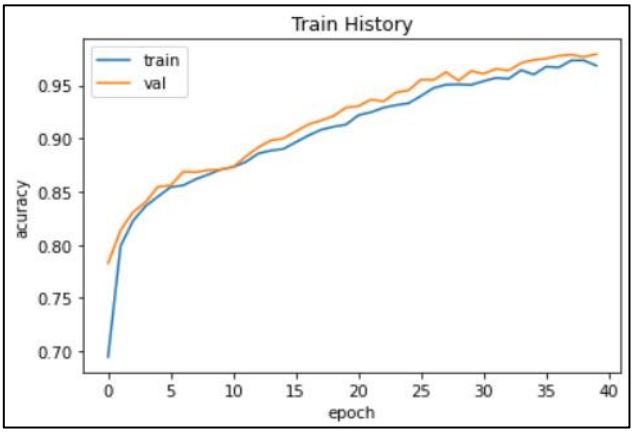
* CNN



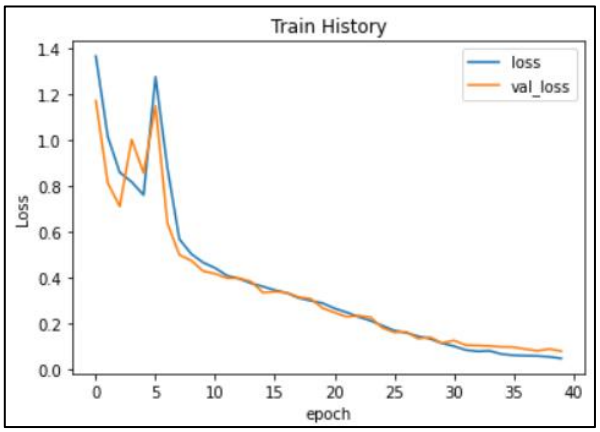


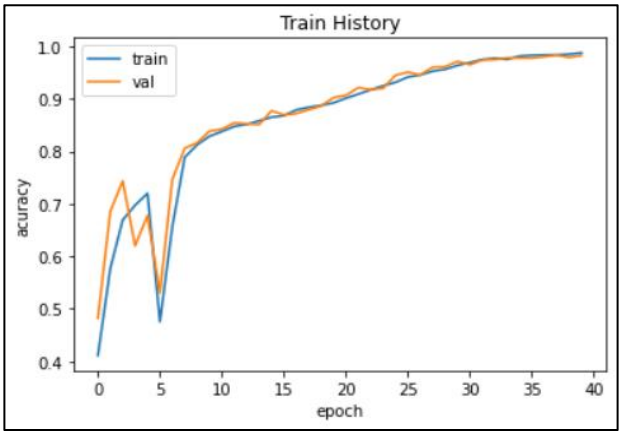
* Bi-LSTM



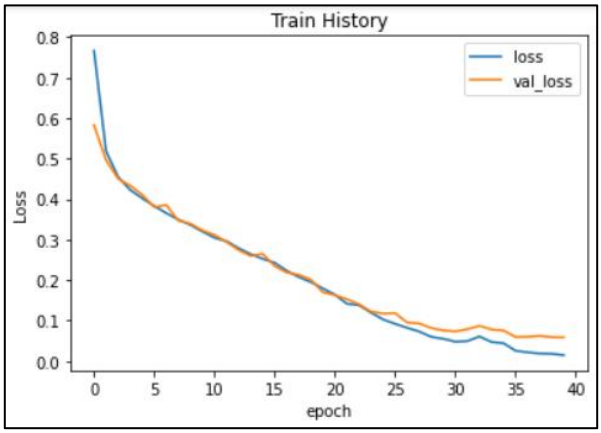


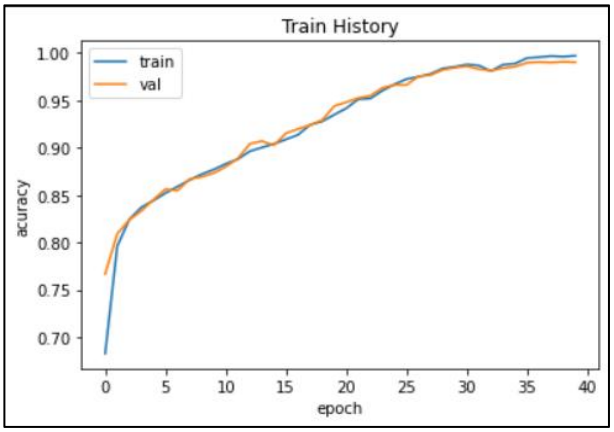
* RNN



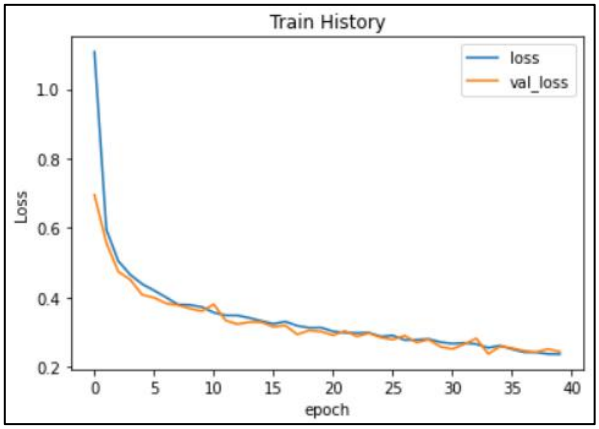


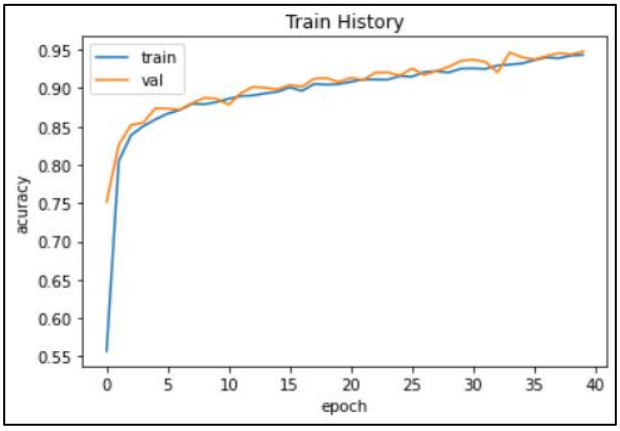
* CNN + BiLSTM



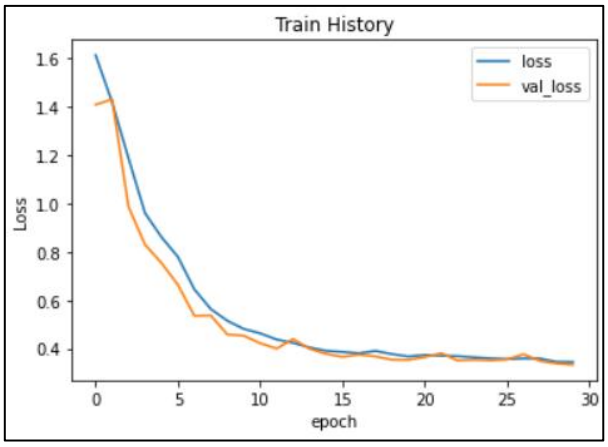


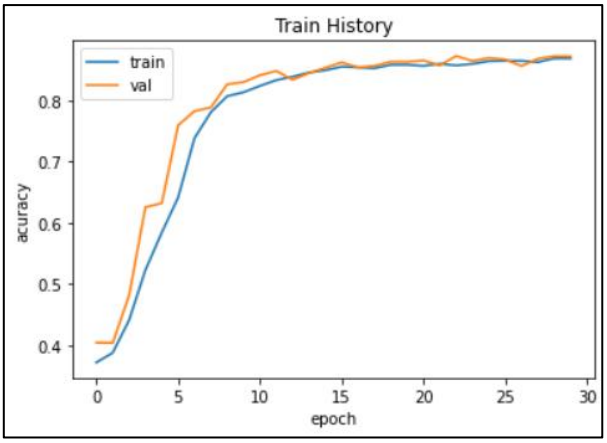
* LSTM + Attention





* TransformerAttention





Link Google Colab code bài làm:

https://colab.research.google.com/drive/1bpUt7685rAo6DvM6dhACliWqExhR2FAs?usp=sharing

5.3. Kết quả huấn luyện và đánh giá mô hình phân loại

STT	Mô hình	Accuracy (%)	Time - Pre	Time - Step	Size (MB)
1	SVM	94.27	0.055s	0 ms	2.67
2	CNN	98.96	0.435s	5 ms	5.24
3	Bi-LSTM	97.94	1.03s	18 ms	8.06
4	RNN	98.23	0.845s	90 ms	8.42
5	CNN + BiLSTM	98.23	1.022s	10 ms	4.92
6	BahdanauAttention	94.79	0.4s	11 ms	10.27
7	TransformerAttention	87.22	0.44s	820 ms	485.78

Bảng 5.1 So sánh các mô hình

Nhìn vào bảng trên ta thấy được mô hình CNN cho độ chính xác cao nhất 98.96%

5.4. Deploy web app lên Azure

Triển khai thành công web app dùng để phân loại bình luận sản phẩm từ model đã huấn luyện

Link Youtube demo ứng dụng đã Deploy thành công trên Azure: https://youtu.be/nyAcTX9-rcc

5.5. Ưu điểm, nhược điểm và hướng phát triển

* Ưu điểm:

- Dataset thu thập được gần đúng ý muốn của nhóm.

- Làm việc nhóm, có tài khoản Google Drive chung cho nhóm, dễ dàng quan sát hoạt động làm việc của các thành viên.

* Nhược điểm:

- Về thu thập dữ liệu, nghiên cứu này tuy thu thập được dữ liệu mong muốn nhưng thu thập dữ liệu là các bình luận, nhận xét của khách hàng về mặt hàng quần áo trên trang thương mại điện tử shopee chứ chưa thu thập được trên hầu hết các trang

- Về thang đo, nghiên cứu này chỉ phân loại phản hồi khách hàng theo thang đo 5 mức: sản phẩm tệ, sản phẩm tốt, bình thường, phục vụ tệ và phục vụ tốt.

* Hướng phát triển

Thu thập dữ liệu thêm từ các trang thương mại điện tử khác, có thể sử dụng thang đo nhiều mức hơn.

Nghiên cứu tiếp các giải pháp triển khai vào ứng dụng thực tế, nhằm cung cấp cho người dùng công cụ để thống kê được mức độ hài lòng của người dùng cho doanh nghiệp. Vấn đề gia tăng độ chính xác có thể giải quyết bằng việc kết hợp thuật toán trong các mạng neural đã trình bày trong đề tài với các thuật toán Deep Learning dạng 24 phân lớp để gia tăng độ chính xác đồng thời không ảnh hưởng quá nhiều tốc độ xử lý thông tin.

Tài liệu kham khảo

[1] Vũ Hữu Tiệp, 2018, Bải giảng Học Máy (Machine Learning), Nhà xuất bản Khoa học kỹ thuật.

[2] P.Đ.Thắng,"Các kiến trúc CNN hiện đại," 2020.

https://phamdinhkhanh.github.io/2020/05/31/CNNHistory.html.

[3] A. Wasicek, "Artificial Intelligence vs. Machine Learning vs. Deep Learning: What's the Difference?," 2018. https://www.sumologic.com/blog/machinelearning-deep-learning/

[4] https://machinelearningcoban.com/2016/12/27/categories/

[5] https://viblo.asia/p/danh-gia-cac-mo-hinh-hoc-may-RnB5pp4D5PG

[6] https://mastery.vn/cong-nghe-deep-learning-o-cac-san-pham-nhan-dang-khuon-mat-hikvision/

[7] S. Marsland, Machine Learning: An Algorithmic Perspective, Second Edition. 86 Stephen Marsland, 2014.

[8] S. A. Bleha and M. S. Obaidat, "Dimensionality reduction and feature extraction applications in identifying computer users," IEEE Trans. Syst. Man. Cybern., vol. 21, no. 2, pp. 452–456, 1991, doi: 10.1109/21.87093.

[9] G. Hinton, "Deep Belief Nets," IEEE Trans. Neural Networks, vol. 17, no. 6, pp. 1623–1629, 2006, doi: 10.1109/TNN.2006.880582.

Bảng kế hoạch, phân công việc nhóm

	Công việc	**Thành viên thực hiện**	**Ghi chú**
Tuần 1	1. Lên kế hoạch ban đầu thực hiện (Giáo viên tư vấn kế hoạch chung, Sinh viên thay đổi theo nhu cầu thực tế của nhóm)	Hải, Khang, Trang	
	2. Phân chia công việc nhóm (chi tiết các công việc cần làm)	Hải, Khang, Trang	
Tuần 2	1. Phân tích yêu cầu của đề tài. Làm rõ các yêu cầu của đề tài	Hải	
	2. Giới hạn mục tiêu của đề tài.	Khang	
	3. Lên kế hoạch dự kiến thực hiện	Trang	
Tuần 3	Xác định số lượng và thu thập data về các đánh giá sản phẩm trên trang shopee	Trang, Khang	

Tuần 4	1. Đưa ra các mô hình cho đề tài	Trang	
	2. Thử nghiệm từng loại mô hình	Hải	
Tuần 5	Xác định mô hình phân lớp tốt nhất và phù hợp nhất với bộ dữ liệu	Khang	
Tuần 6	Bổ sung data và retrain lại mô hình phân lớp với bộ dữ liệu mới	Khang, Trang, Hải	
Tuần 7	Báo cáo đề cương với giáo viên hướng dẫn	Khang	
Tuần 8	Tiếp tục bổ sung và tinh chỉnh lại dataset để retrain mô hình phân lớp	Hải, Trang	
Tuần 9	Xây dựng ứng dụng phân loại bình luận	Hải	

Tuần 10	1. Hoàn thiện ứng dụng	Hải	
	2. Kiểm thử ứng dụng	Trang	
Tuần 11	Làm file word báo cáo và ppt thuyết trình	Khang, Trang	
Tuần 12	Báo cáo cuối kỳ	Hải	