BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH

TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ

ĐỒ ÁN MÔN HỌC

ĐỀ TÀI:

XÂY DỰNG MÔ HÌNH PHÂN TÍCH CẢM XÚC
TRONG PHẢN HỒI CỦA SINH VIÊN VIỆT NAM
BẰNG CÁC PHƯƠNG PHÁP XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Học phần: XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Chuyên ngành: Khoa học dữ liệu – Khóa 47

Nhóm Sinh Viên:

Họ và tên	MSSV
Phan Dương Hoàng Vũ	31211022533
Đỗ Quang Thiên Phú	31211024191
Phạm Dương Thị Thúy Truyền	31211027682

Giảng Viên: TS. Đặng Ngọc Hoàng Thành

TP. Hồ Chí Minh, ngày 26 tháng 11 năm 2023

LỜI CẢM ƠN

Để hoàn thành tốt đồ án này trước tiên, nhóm chúng em xin gửi đến quý Thầy, Cô Khoa Công nghệ thông tin kinh doanh - Trường Công nghệ và Thiết kế UEH lời cảm ơn chân thành và sâu sắc nhất vì đã đưa bộ môn Xử lý ngôn ngữ tự nhiên vào chương trình giảng dạy, giúp chúng em học, khai thác và thực hành được nhiều kiến thức, kỹ năng bổ ích từ bộ môn này.

Đặc biệt, nhóm chúng em xin gửi lời cảm ơn sâu sắc nhất đến giảng viên giảng dạy TS. Đặng Ngọc Hoàng Thành đã nhiệt tình giảng dạy truyền tải, hướng dẫn cho chúng em nhiều kiến thức bổ ích trong suốt thời gian học. Cảm ơn sự hỗ trợ, giúp đỡ tận tình của Thầy để nhóm chúng em có thể hoàn thành đồ án này một cách tốt nhất.

Nhận thức rằng nhóm chúng em còn nhiều hạn chế về mặt kiến thức kiến thức, nên không tránh khỏi sai sót trong quá trình làm bài. Nhóm chúng em mong nhận được sự đánh giá, nhận xét và góp ý xây dựng từ phía Thầy để đồ án của nhóm trở nên hoàn thiện hơn.

Chúng em xin chân thành cảm ơn Thầy và kính chúc Thầy nhiều sức khỏe, thành công và hạnh phúc trong công việc giảng dạy và trong sự nghiệp của mình.

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Từ gốc
SVM	Support Vector Machine
Maxent	Maximum Entrophy
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
TF-IDF	Term Frequency - Inverse Document Frequency
CBOW	Continuous bag of words

MỤC LỤC

CHƯƠNG 1: MỞ ĐẦU

1. Tính cấp thiết của đề tài

Ngày nay, với sự phát triển vượt bậc của khoa học và công nghệ, việc đánh giá chất lượng giáo dục và trải nghiệm học tập thông qua phân tích phản hồi của sinh viên trở nên cực kỳ quan trọng, thu hút sự quan tâm trong lĩnh vực giáo dục.

Một hệ thống giáo dục xuất sắc không chỉ tập trung vào việc truyền đạt kiến thức mà còn chú trọng vào việc xây dựng một môi trường học tập tích cực, thúc đẩy sự sáng tạo và tăng cường hứng thú của sinh viên. Việc hiểu rõ cảm xúc của họ thông qua phân tích phản hồi là cần thiết để điều chỉnh phương pháp giảng dạy, tạo điều kiện thuận lợi cho sự phát triển cá nhân và chuyển giao kiến thức hiệu quả.

Phân tích cảm xúc không chỉ giúp định rõ các yếu tố tiêu cực và tích cực trong quá trình học tập mà còn mở ra cơ hội để tạo ra các biện pháp cải thiện đột phá. Các nhóm nghiên cứu và giảng viên có thể điều chỉnh các chiến lược giảng dạy linh hoạt hơn, đáp ứng đa dạng nhu cầu của sinh viên và kích thích sự tò mò và sáng tạo.

Đặc biệt, lắng nghe phản hồi từ sinh viên đang trở thành một quy trình quan trọng của quá trình quản lý đào tạo. Việc thường xuyên thu thập và phân tích dữ liệu phản hồi giúp các trường hiểu rõ hơn về đặc điểm tạo nên động lực, những mong đợi và khó khăn của sinh viên. Từ đó, họ có thể áp dụng những biện pháp cải tiến linh hoạt để tối ưu hóa trải nghiệm học tập và định hình hướng phát triển cho ngôi trường của mình.

Nhìn chung, sự kết hợp giữa phân tích cảm xúc và lắng nghe phản hồi sinh viên không chỉ đáp ứng nhu cầu ngày càng cao về chất lượng giáo dục mà còn thúc đẩy sự phát triển bền vững của hệ thống giáo dục, hướng tới một tương lai với những thế hệ sinh viên tự tin, sáng tạo và sẵn sàng đối mặt với thách thức toàn cầu.

2. Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của đề tài là phân tích cảm xúc của sinh viên thông qua các phản hồi, xây dựng chương trình minh họa cho phép người dùng nhập vào phản hồi đưa ra kết luận về phản hồi đó. Phân tích phản hồi theo cảm xúc: Phân loại mỗi phản hồi sẽ thuộc loại nào trong số các loại đây: tích cực, tiêu cực hay trung lập.

3. Đối tượng, phạm vi nghiên cứu

a. Đối tượng nghiên cứu:

- Tất cả những phản hồi trong kho phản hồi của sinh viên Việt Nam (UIT - VSFC).

Với UIT - VSFC là nguồn tài liệu gồm hơn 16.000 câu về các phản hồi của sinh viên Việt Nam.

b. Phạm vi nghiên cứu:

- *Về nội dung*: Nghiên cứu đề cập đến việc phân tích cảm xúc sinh viên thông qua phản hồi để xác định xem đó là phản hồi tích cực, tiêu cực, hay trung lập.
- *Về không gian*: Nghiên cứu được tiến hành dựa trên dữ liệu phản hồi trong kho phản hồi của sinh viên Việt Nam (UIT - VSFC).
- *Về thời gian*: cuối học kỳ năm 2013 và năm 2016.

4. Phương pháp nghiên cứu

- *Thu thập dữ liệu*: Tiến hành thu thập dữ liệu phản hồi sinh của sinh viên từ UIT - VSFC.
- *Tiền xử lý và trích xuất đặc trưng*: Thực hiện quá trình tiền xử lý chuẩn bị dữ liệu cho các phương pháp phân tích. Bước này bao gồm loại bỏ stopwords, từ ngoại lai, và biểu tượng, cũng như gắn thẻ POS xác định loại từ cho câu. Quá trình này không chỉ giúp làm sạch dữ liệu mà còn nhấn mạnh các đặc trưng quan trọng để sử dụng trong các bước phân tích tiếp theo.
- *Quy trình chú thích*: Xây dựng các hướng dẫn đầy đủ các chú thích để định rõ các quy tắc và tiêu chí.
- *Phân tích cảm xúc*: Đánh giá câu có cảm xúc tích cực, tiêu cực hay trung lập dựa trên các đặc trưng đã được chọn. Xây dựng hệ thống phân loại, điều chỉnh ngôn ngữ, từ điển cho tiếng Việt dựa trên các mô hình: SVM, Maximum-Entropy Markov Model, Deep Learning và Vectorizer.

5. Kết cấu đồ án

Kết cấu đồ án bao gồm các chương sau:

- ***Chương 1*** *- Mở đầu*: Trình bày lý do chọn đề tài, đối tượng, phạm vi nghiên cứu, mục tiêu và tổng quan các nghiên cứu trước.
- ***Chương 2*** *– Vietnamese Student's Feedback Dataset*: Thực hiện bài nghiên cứu phân tích cảm xúc dựa trên Student's Feedback Dataset
- ***Chương 3*** *- Phương pháp luận*: Trình bày cơ sở lý thuyết, của các mô hình được sử dụng trong bài nghiên cứu.
- ***Chương 4*** *- Phân tích kết quả thực nghiệm*: Trình bày quá trình thực nghiệm, và phân tích, đánh giá quá trình thực nghiệm.

- ***Chương 5*** - *Ứng dụng*: Xây dựng chương trình ứng dụng vào phân tích cảm xúc.
- ***Chương 6*** - *Kết luận*: Tổng kết các kết quả đạt được trong bài nghiên cứu, những hạn chế chưa được giải quyết và hướng phát triển trong tương lai.

6. Ứng dụng

Việc phân tích cảm xúc từ phản hồi từ sinh viên mang lại nhiều lợi ích và có tính ứng dụng cao trong lĩnh vực giáo dục và quản lý đào tạo. Dưới đây là một số ứng dụng quan trọng của việc phân tích phản hồi sinh viên:

- *Cải thiện chất lượng giảng dạy*: Thông qua các phản hồi, giảng viên có thể hiểu rõ hơn về sự hài lòng và nhu cầu học tập của sinh viên. Cung cấp các thông tin chi tiết về cảm nhận sinh viên qua quá trình dạy, từ đó giảng viên linh hoạt điều chỉnh hoạt động giảng dạy để đáp ứng nhu cầu, mong muốn của sinh viên, cải thiện quá trình truyền tải.
- *Nâng cao chất lượng chương trình học*: Phân tích cảm nhận của sinh viên qua các phản hồi giúp phòng đào tạo đánh giá được chương trình đào tạo. Giúp phòng đào tạo hiểu rõ được các phần nào là trọng tâm, được sinh viên đánh giá cao, góp phần tối ưu hóa nội dung và cấu trúc của chương trình đào tạo.
- *Tích hợp cảm xúc vào giảng dạy*: Phản hồi của sinh viên là công cụ quan trọng để cải thiện mối quan hệ và tương tác giữa sinh viên và giảng viên, giúp sinh viên và giảng viên điều chỉnh linh hoạt cảm xúc để tạo ra một môi trường học thoải mái, thiết lập môi trường học tập tích cực, thoải mái sáng tạo.

7. Tổng quan các công trình nghiên cứu có liên quan

- Công trình nghiên cứu: *"Opinion Mining and Sentiment Analysis"* của Bo Pang và Lillian Lee. Nghiên cứu này tập trung vào việc phát triển các phương pháp để phân tích ý kiến, cảm xúc của người dùng đối với miền dữ liệu điện ảnh. Trong công trình này, Pang và Lee đã đề xuất mô hình: Naive bayes, maximum entropy và SVM để phân loại xem phản hồi đó là tích cực hay tiêu cực. Bài nghiên cứu đã cho thấy kết quả có thể so sánh được với các giải pháp khác, dao động từ 71% đến 85% tùy thuộc vào phương pháp và bộ dữ liệu kiểm thử.

- Công trình nghiên cứu: *"Improving e-learning with sentiment analysis of users' opinions"* của Zied Kechaou, Mohamed Ben Ammar và Adel.M Alimi tập trung vào việc phân tích phản hồi của người học trên các diễn đàn e-learning. Trong nghiên cứu này, ba đặc trưng chính đã được áp dụng, đó là IG, MI và CHI. Kết quả của nghiên cứu chỉ ra rằng

IG vượt trội so với MI và CHI trong quá trình phân tích. Đồng thời, mô hình kết hợp HMM và SVM theo quy tắc Sum đã mang lại kết quả tốt hơn, làm nổi bật tính hiệu quả của phương pháp này trong việc cải thiện e-learning thông qua việc đánh giá ý kiến của người học.

- Công trình nghiên cứu: *"Statistical and Sentiment Analysis of Consumer Product Reviews"* của Zeenia Singla, Sukhchandan Randhawa và Sushma Jain. Nghiên cứu này tập trung vào việc phân tích cảm xúc đối với các đánh giá từ người tiêu dùng, để hiểu rõ hơn về cách mà người tiêu dùng thể hiện ý kiến và cảm xúc của họ đối với các sản phẩm thông qua các đánh giá trực tuyến bằng cách thực hiện các phân tích thống kê, biểu diễn trực quan và đánh giá hiệu quả thông qua việc áp dụng mô hình SVM và kết luận thông qua chỉ số: Accuracy qua các lần chạy mô hình. Bài nghiên cứu đã phân các đánh giá thành các phân cực bao gồm: tích cực, tiêu cực, tức giận, mong đợi, sợ hãi, vui vẻ, buồn bã, tin tưởng, ngạc nhiên, ghê tởm. Kết quả phân loại với Accuracy là 84,85%

- Công trình nghiên cứu: *"Integrating Grammatical Features into CNN Model for Emotion Classification"* của Huỳnh Thị Thanh Thủy và Lê Anh Cường. Nghiên cứu này đã đề xuất một mô hình cải tiến trong đó các đặc trưng ngữ pháp quan trọng có thể cải thiện ý nghĩa của văn bản nhúng như một đặc trưng hỗ trợ mô hình CNN cơ bản. Mô hình đề xuất của đã đạt được độ chính xác cao hơn đáng kể trong việc phân loại cảm xúc trên cả bộ dữ liệu ISEAR và bộ dữ liệu cảm xúc tiếng Việt so với mô hình CNN cơ bản. Điều này chỉ ra rằng việc tích hợp những đặc trưng ngữ pháp là quan trọng có thể cải thiện khả năng phân loại cảm xúc và tăng hiệu suất của mô hình trong việc hiểu và xử lý ngôn ngữ tự nhiên.

CHƯƠNG 2: VIETNAMESE STUDENT'S FEEDBACK DATASET

1. Mô tả bộ dữ liệu

UIT-VSFC là bộ dữ liệu đầu tiên được thực hiện trong lĩnh vực giáo dục. Cụ thể, bộ dữ liệu được thu thập dựa trên những phản hồi của các bạn sinh viên trong quá trình tham gia học tập tại trường. Bộ dữ liệu UIT-VSFC được sử dụng cho hai nhiệm vụ: (1) phân loại dựa trên cảm xúc (sentiment analysis) và phân loại dựa theo chủ đề (text classification). Người ta thu thập phản hồi của sinh viên thông qua các cuộc khảo sát sinh viên vào cuối học kỳ 2013 và 2016, với hơn 16.000 phản hồi. Có hai loại phản hồi chính: (1) phản hồi từ giảng viên cho sinh viên để giúp đỡ sinh viên nhận thức được điểm yếu và điểm mạnh của mình để cải thiện nghiên cứu và (2) phản hồi từ sinh viên cho giảng viên để phản ánh và cải thiện giảng dạy của họ. Đặc biệt, sinh viên đưa ra ý kiến về một loạt các vấn đề khác nhau. Ví dụ: phản hồi của sinh viên thể hiện những gì sinh viên thích hoặc không thích về bài giảng cũng như những gì giảng viên giảng dạy là xuất sắc hoặc tệ.

Mỗi câu trong corpus được gán nhãn một trong ba cực cảm xúc bao gồm: ***Tích cực (Positive), Tiêu cực (Negative) và Trung lập (Neutral)***. Chi tiết chú thích được mô tả ngắn gọn như sau:

- *Tích cực*: Những câu được sinh viên dùng để bày tỏ sự hài lòng, khen ngợi về yếu tố trong việc học tập tại trường đại học như giảng viên, môn học, bài tập… Ví dụ, câu "Giảng viên rất nhiệt tình và tận tâm" được gán nhãn là tích cực.
- *Tiêu cực*: Các câu thể hiện sự không hài lòng, kiến nghị và phàn nàn của sinh viên liên quan đến giảng viên, chương trình giảng dạy, cơ sở vật chất… Ví dụ, "Nội dung môn học này quá nhiều" là một câu thể hiện cảm xúc tiêu cực.
- *Trung lập*: Câu chưa đầy đủ, không rõ ràng về ý nghĩa hoặc không chứa đựng ý kiến. Ví dụ như câu "Giờ giấc và cách giảng dạy" là một câu trung lập, chỉ là một cụm danh từ không đầy đủ cú pháp của một câu và không chứa đựng bất kỳ cảm xúc nào. Hoặc câu đã hoàn chỉnh nhưng không thể hiện được cảm xúc nào của sinh viên. Ví dụ như "Em cảm ơn thầy", không thể hiện bất kỳ cảm xúc nào, vì vậy đây là một câu trung lập.

Có những câu thể hiện cả cảm xúc tiêu cực và tích cực rất khó để đánh giá. Chúng thường là những câu phức tạp với các liên từ như "nhưng", "tuy nhiên"... Trong trường hợp này, người gán nhãn sẽ chọn cảm xúc được thể hiện mạnh hơn tùy theo cảm nhận của họ. Có thể thấy câu "Thầy dạy dễ hiểu nhưng nhiều lúc bay nhanh" được gán nhãn Tiêu

cực dù có chứa mệnh đề Tích cực “Thầy dạy dễ hiểu”.

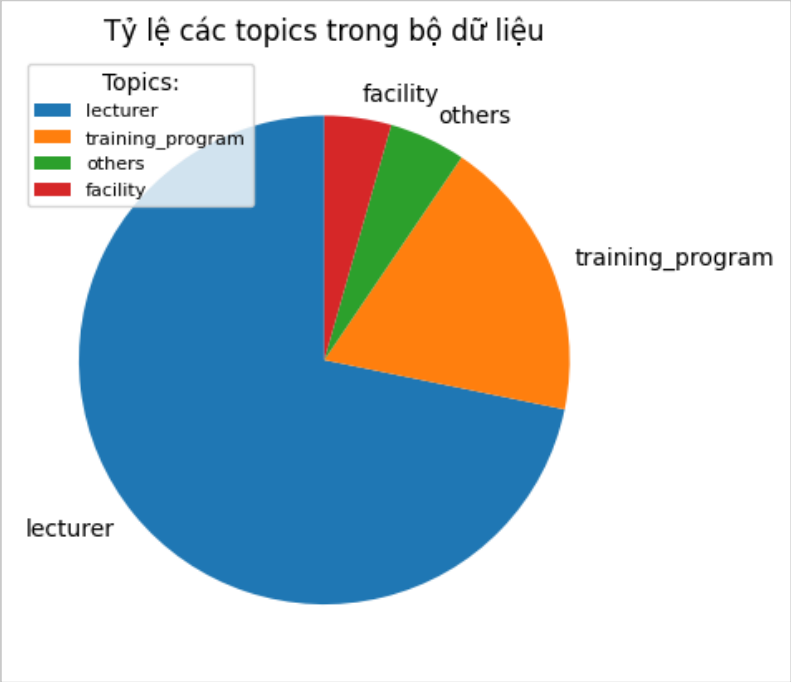
STT	*Sentence*	*Sentiment*	*Topic*
1	nhiệt tình giảng dạy , gần gũi với sinh viên .	positive	lecturer
2	chưa áp dụng công nghệ thông tin và các thiết bị hỗ trợ cho việc giảng dạy .	negative	lecturer
3	em sẽ nợ môn này , nhưng em sẽ học lại ở các học kỳ kế tiếp .	neutral	others
4	thầy rất tận tình và đi dạy rất đúng giờ .	positive	lecturer
5	dễ bị áp lực .	negative	others
6	đang dạy thầy wzjwz208 đi qua nước ngoài giữa chừng , thầy wzjwz209 dạy thay .	neutral	lecturer

Bảng 1. Ví dụ gán nhãn sentiment (cảm xúc) và topic (chủ đề) trong bộ UIT-VSFC

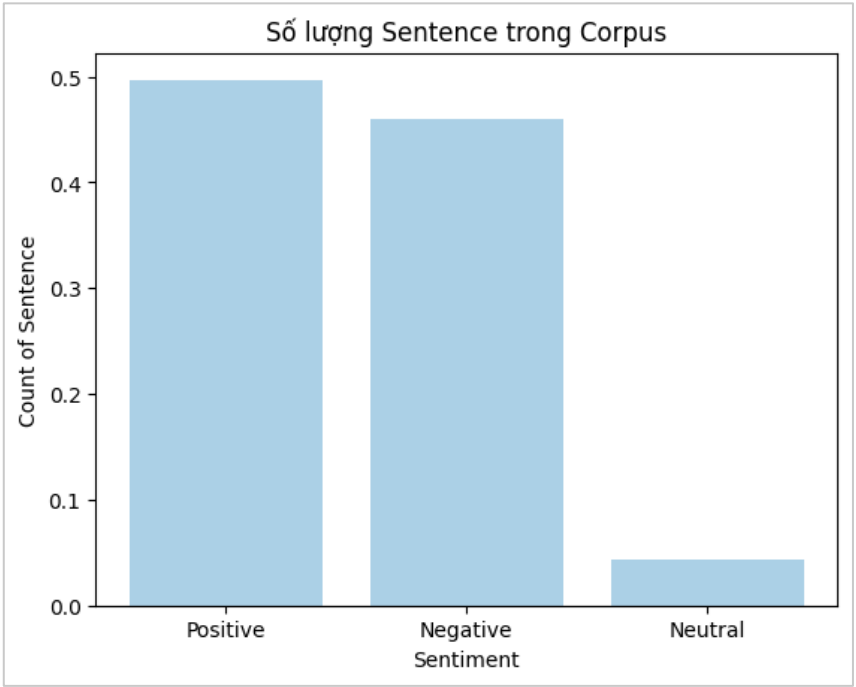
Trong bộ dữ liệu, feedback của sinh viên Việt Nam có những đặc điểm đặc biệt. Những feedback của sinh viên là những tin nhắn ngắn gọn được viết tự do. Chúng chứa nhiều từ viết tắt, lỗi chính tả, emotion, icon và những ký tự khác thể hiện những ý nghĩa đặc biệt. Ví dụ, trong phần phản hồi có rất nhiều từ viết tắt: gv - giảng viên, sv - sinh viên, kh1 - học kỳ một… Những từ viết tắt này có ích cho việc xây dựng từ điển từ viết tắt. Phản hồi của sinh viên còn chứa đựng những emotion có cảm xúc tích cực: “<3”, “:-)”, “:D”, “=)”... và những cảm xúc tiêu cực: “:(“, “:-(“, “:p”, “:@”... Tuy nhiên, có những cảm xúc xuất hiện ở cả hai phân lớp, ví dụ như một câu tích cực “Chưa bao giờ có ý định nghỉ học lớp của cô :v” hay với một câu phủ định “Đôi lúc thầy khó tính quá :v”.

2. Phân tích khám phá dữ liệu (EDA)

Bộ dữ liệu UIT-VSFC được chia thành ba tập như training set, validation set và testing set. Training set chiếm khoảng 70% tổng bộ dữ liệu, trong khi validation và test set lần lượt chiếm khoảng 10% và 20% dữ liệu. Quan sát phân bổ các nhãn được thể hiện trong bộ dữ liệu được thể hiện trong hai biểu đồ dưới đây về topics và về sentiment. Tuy nhiên, đối với nội dung của đồ án, ta tập trung vào phân bổ sentiment trong bộ dữ liệu và các tập được chia ra.



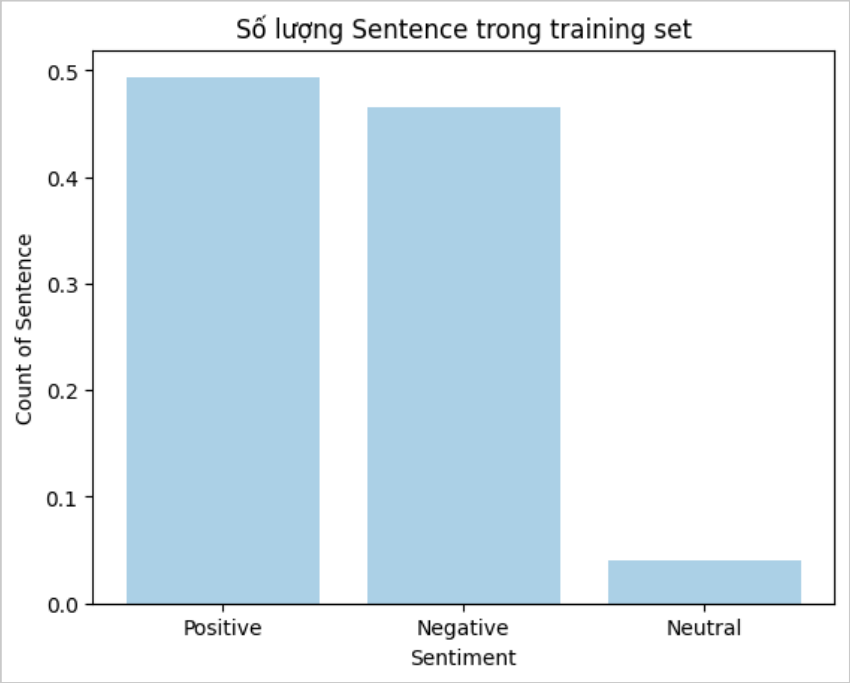
Hình 1. Biểu đồ thể hiện tỷ lệ các topics trong bộ dữ liệu



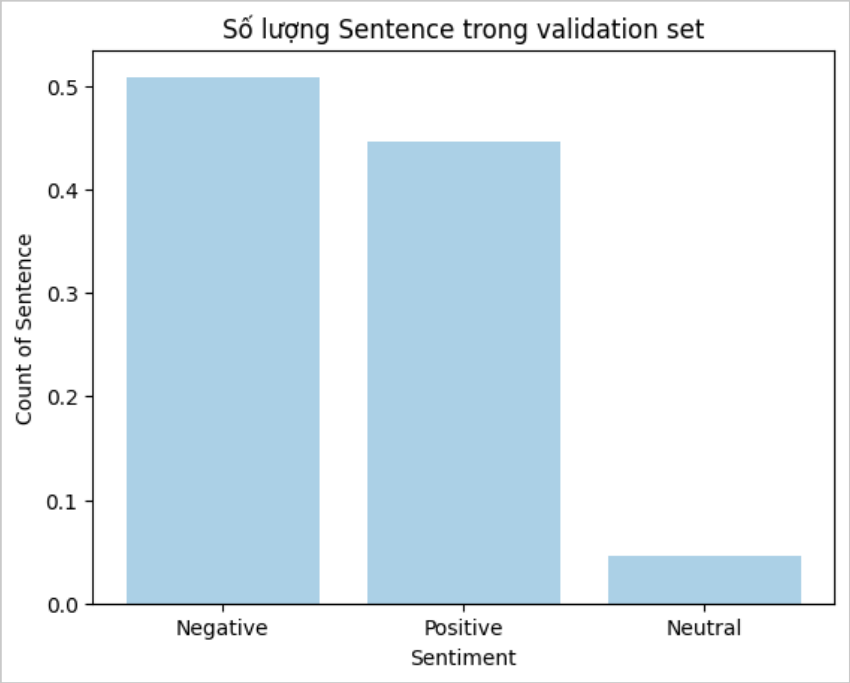
Hình 2. Số lượng feedback theo mỗi sentiment trong bộ dữ liệu

Nhận xét: Xét biểu đồ tròn về tỷ lệ các topic trong bộ dữ liệu, khoảng 70% chủ đề nói về Giảng viên (lecturer), khoảng 20% feedback đề cập đến chương trình đào tạo (training program) và 10% còn lại là về cơ sở vật chất (facility) và các yếu tố khác (others). Đối với biểu đồ ở trên về phân bổ các sentiment: Positive, Negative và Neutral trong bộ dữ liệu. Ta thấy, trong bộ dữ liệu UIT-VSFC các câu được gán nhãn nhiều nhất là nhãn Positive với 8.038 câu (chiếm khoảng 49.7% bộ dữ liệu), ít hơn một chút là các câu được gán nhãn

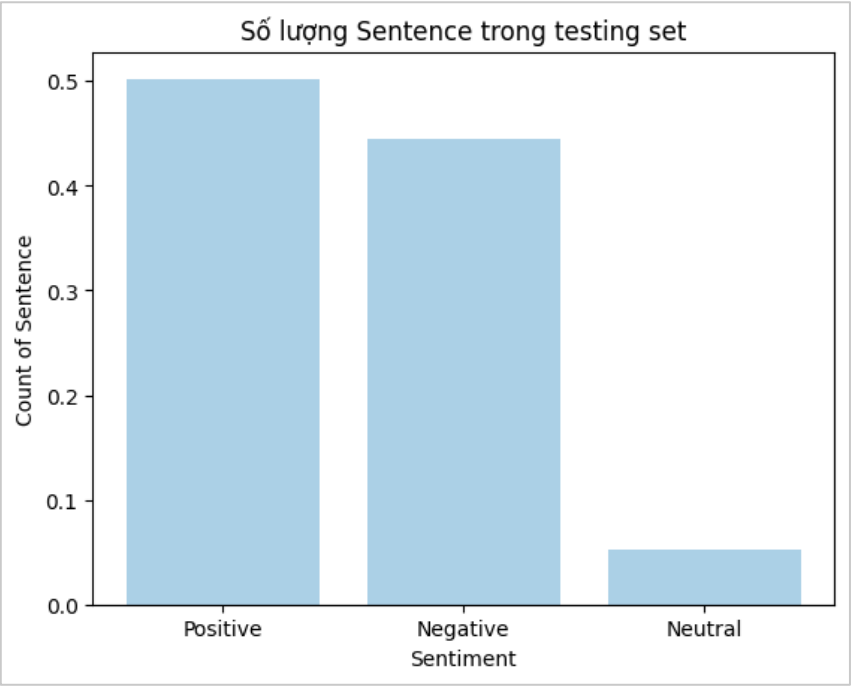
Negative với 7.439 câu (chiếm khoảng 46%). Và chiếm ít nhất trong bộ dữ liệu là nhãn Neutral với 698 câu (chiếm khoảng 4.3%). Nhãn Neutral chiếm rất ít so với hai nhãn còn lại trong bộ dữ liệu, cho thấy rằng bộ dữ liệu được phân lớp khá rõ ràng ở hai phía Tích cực và Tiêu cực.



Hình 3. Số lượng feedback theo sentiment trong tập dữ liệu training set



Hình 4. Số lượng feedback theo sentiment trong tập dữ liệu validation set



Hình 5. Số lượng feedback theo sentiment trong tập dữ liệu testing set

Nhận xét: Phân bố sentiment trong ba tập Training set, Validation set và Testing set với tỷ lệ tương tự như trong bộ dữ liệu lần lượt khoảng 50%, 45% và 5%. Điều này sẽ giúp các công việc việc huấn luyện, đánh giá và kiểm tra sẽ được diễn ra hiệu quả hơn. Huấn luyện mô hình sẽ được học qua ba nhãn, đánh giá và kiểm tra sẽ cho ra kết quả khách quan hơn.

Số lượng feedback chi tiết về ba nhãn Negative, Neutral và Positive trong ba tập Train, Test và Validation được thể hiện trong bảng dưới đây:

	Negative	***Neutral***	***Positive***	***Overall***
Training Set	5.325	458	5.643	11.426
Validation Set	705	73	805	1.583
Testing Set	1.409	167	1.590	3.166
Overall	7.439	698	8.038	16.175

Bảng 2. Số lượng sentiment trong từng tập dữ liệu

CHƯƠNG 3: CƠ SỞ LÝ THUYẾT

1. Phương pháp học

a. Support Vector Machines (SVM)

Support Vector Machines (SVM) là một thuật toán học máy có giám sát, ra đời từ lý thuyết học thống kê do Vapnik và Chervonenkis giới thiệu vào năm 1995 (Cortes and Vapnik, 1995). Đặc tính đáng chú ý của SVM là khả năng học của chúng có thể độc lập với chiều của không gian đặc trưng. SVM xem dữ liệu đầu vào như các vector trong không gian và phân loại chúng vào các lớp khác nhau bằng cách xây dựng một siêu phẳng trong không gian nhiều chiều làm mặt phân cách các lớp dữ liệu. Để tối ưu kết quả phân lớp thì phải xác định siêu phẳng (hyperplane) có khoảng cách đến các điểm dữ liệu (margin) của tất cả các lớp xa nhất có thể. Kích thước của siêu phẳng phụ thuộc vào số lượng đặc trưng. Nếu số lượng đặc trưng đầu vào là hai thì siêu phẳng chỉ là một đường thẳng. Nếu số lượng đặc điểm đầu vào là ba thì siêu phẳng sẽ trở thành mặt phẳng 2-D. Tổng quát hóa khoảng cách từ một điểm có tọa độ x0 đến một siêu phẳng được xác định theo công thức:

$$\frac{|\mathbf{w}^T\mathbf{x}_0 + b|}{||\mathbf{w}||_2}$$

Để tối ưu hóa SVM, ta thực hiện tối đa hóa giá trị margin, tìm ra siêu phẳng đẹp nhất để phân 2 lớp dữ liệu. Margin là khoảng cách giữa siêu phẳng đến 2 điểm dữ liệu gần nhất tương ứng với 2 phân lớp, trong trường hợp không gian 2 chiều thì nó là đường thẳng. Nhờ việc tối ưu này, sự phân chia giữa 2 lớp trở nên rạch ròi hơn và SVM có thể giảm thiểu việc phân lớp sai đối với điểm dữ liệu mới đưa vào.

Tổng quát hóa việc tối đa hóa giá trị margin cho không gian nhiều chiều được mô tả với phương trình biểu diễn siêu phẳng cần tìm (hyperlane) trong không gian đa chiều là: $w^T x+b=0$ và giá trị margin = $2|w^T x+b|/\|w\| = 2/\|w\|$. Khi đó giá trị margin cực đại đồng nghĩa với việc $\|w\|$ đạt cực tiểu với điều kiện:

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \forall n = 1, 2, \ldots, N$$

SVM có một số ưu điểm như: Tiết kiệm bộ nhớ, phân lớp nhanh, vì quá trình test chỉ cần so điểm dữ liệu mới với mặt siêu phẳng tìm được mà không cần tính toán lại, độ chính xác phân lớp cao, ít bị overfitting. Ngoài ra, SVM còn xử lý tốt trong không gian nhiều chiều và hiệu quả trong trường hợp có chiều cao. Vừa có thể phân lớp tuyến tính và

phi tuyến (sử dụng các kernel khác nhau). Tuy nhiên, SVM cũng có một số nhược điểm sau: Kém hiệu quả với kho dữ liệu lớn (thời gian huấn luyện). Trong trường hợp số chiều dữ liệu lớn hơn số dòng dữ liệu thì SVM cho ra kết quả không tốt, và chưa thể hiện tính xác suất trong phân lớp. Đồng thời SVM cũng khá nhạy cảm với nhiễu.

b. Maximum Entropy (Maxent)

Mô hình Entropy cực đại là mô hình dựa trên xác suất có điều kiện, được sử dụng để ước lượng phân phối xác suất của một biến ngẫu nhiên. Mục tiêu là tìm một phân phối xác suất có entropy lớn nhất. Phương pháp này được áp dụng rộng rãi trong các lĩnh vực như xử lý ngôn ngữ tự nhiên như: ngôn ngữ mô hình hóa (của Chen và Rosenfeld năm 1999), gán nhãn từ loại (của Ratnaparkhi năm 1996), phân loại văn bản (của Beeferman năm 1999), nhận dạng hình ảnh, và thị giác máy tính.

Một số các mô hình trong nhóm Maxent như như Logistic Regression, và các dẫn xuất Regularized của Logistic Regression, Ridge Regression, Conditional Random Fields (CRF model), Maximum-Entropy Markov Model (MEMM model). Trong bài nghiên cứu này nhóm chọn Ridge Regression đưa vào mô hình nghiên cứu.

Ridge Regression được giới thiệu bởi Hoerl and Kennard (1970), là một quá trình ước lượng có thể xử lý đa cộng tuyến mà không loại bỏ các biến ra khỏi mô hình hồi quy. Hồi quy Ridge có sự thay đổi hơn so với hồi quy tuyến tính bởi thêm vào một thành phần điều chỉnh vào hàm mất mát. Về bản chất, hồi quy Ridge tối ưu song song hai thành phần bao gồm tổng bình phương phần dư và thành phần điều chỉnh. Có thể giảm phương sai của công cụ ước tính OLS thông qua việc đưa ra một số sai lệch. Ridge Regression được tối ưu theo công thức:

*Mục tiêu = RSS + α * (tổng bình phương của các hệ số)*

Trong đó:

- RSS là tổng bình phương sai số (Residual Sum of Squares), đo lường sự chênh lệch giữa giá trị dự đoán và giá trị thực tế.
- α (alpha) là tham số cân bằng giữa việc tối thiểu hóa RSS và tối thiểu hóa tổng bình phương của các hệ số R
    - α = 0: Hàm mục tiêu, các hệ số trở thành như hồi quy tuyến tính đơn giản.
    - α = ∞: Hệ số sẽ bằng 0. Do phần α * *(tổng bình phương của các hệ số)* sẽ lớn, trở nên quan trọng hơn RSS, mục tiêu tối ưu hóa sẽ trở nên rất lớn. Vì

vậy để giảm mục tiêu thì các hệ số phải rất nhỏ thậm chí là bằng 0.

- $0 < \alpha < \infty$: Mô hình cố gắng cân bằng giữa việc giảm RSS và giảm độ lớn của các hệ số. Các hệ số sẽ nằm trong khoảng từ 0 đến 1 đối với hồi quy tuyến tính đơn giản.

Về ưu điểm của hồi quy Ridge: Ridge Regression thường được sử dụng để kiểm soát hiện tượng quá mức (overfitting), đặc biệt là khi có nhiều biến độc lập hoặc có sự đồng biến mạnh giữa các biến độc lập. Trong trường hợp có đa cộng tuyến giữa các biến độc lập, hồi quy Ridge giúp ổn định các ước lượng và ngăn chặn sự đột biến lớn của các hệ số.

Tuy nhiên, nó cũng gặp một số hạn chế sau: Hồi quy Ridge giữ lại tất cả các biến trong mô hình và giảm độ lớn của chúng, nhưng không thực sự loại bỏ chúng khỏi mô hình. Các biến với tác động thấp vẫn được giữ lại, làm cho khả năng diễn giải giảm. Tăng mức độ phức tạp đối với mô hình. Ngoài ra, Việc chọn giá trị của α cũng là một thách thức và cần phải được thực hiện thông qua các kỹ thuật như cross-validation.

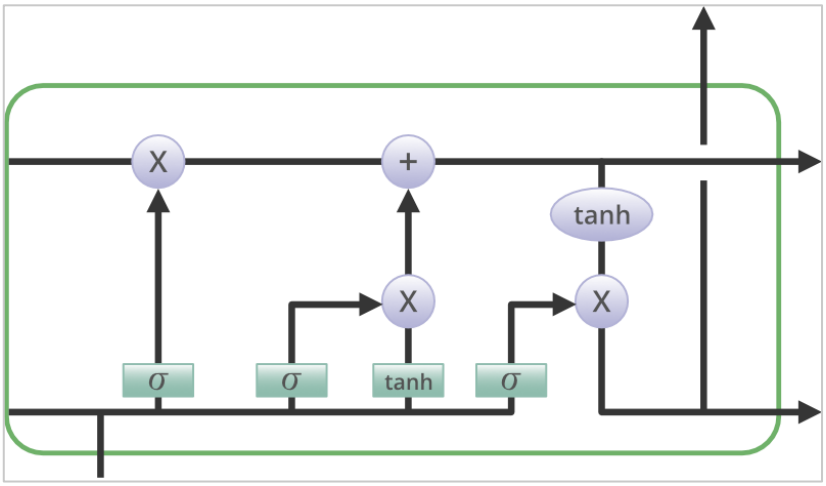
c. Long Short Term Memory (LSTM)

Mạng thần kinh hồi quy (Recurrent Neural Network - RNN) có thể xử lý thông tin dạng chuỗi (sequence/ time-series). Như ở bài dự đoán hành động trong video ở bài trước, RNN có thể mang thông tin của frame (ảnh) từ state trước tới các state sau, rồi ở state cuối là sự kết hợp của tất cả các ảnh để dự đoán hành động trong video. Tuy nhiên, các state càng xa ở trước đó thì càng bị vanishing gradient và các hệ số không được update với các frame ở xa. Hay nói cách khác là RNN không học được các thông tin ở trước đó xa do vanishing gradient. Như vậy về lý thuyết, RNN có thể mang thông tin từ các layer trước đến các layer sau, nhưng thực tế là thông tin chỉ mang được qua một số lượng state nhất định, sau đó thì sẽ bị vanishing gradient, hay nói cách khác là model chỉ học được từ các state gần đó (short term memory).

Long Short-Term Memory (LSTM) là phiên bản cải tiến của RNN được thiết kế bởi Hochreiter & Schmidhuber (1997). LSTM rất phù hợp với các công việc xử lý thông tin ở dạng chuỗi. Ứng dụng của nó mở rộng đến các tác vụ liên quan đến chuỗi thời gian và trình tự và vượt trội trong nắm bắt các phụ thuộc xa (long-term dependency), điều này làm nó phù hợp với các công việc như dịch ngôn ngữ, dự báo chuỗi thời gian. LSTM cũng có thể được sử dụng kết hợp với các kiến trúc mạng thần kinh khác như Convolutional Neural Network (CNN) để phân tích hình ảnh và video.

LSTM có cấu trúc dạng chuỗi các nút mạng như RNN, nhưng cấu trúc bên trong thì

lại phức tạp hơn, bao gồm 4 tầng tương tác với nhau. Điểm đặc biệt của mạng LSTM nằm ở trạng thái ô C (cell state), nơi lưu trữ các trọng số dài hạn của mô hình. Các thông số trạng thái ô C, trạng thái ẩn h (hidden state), đầu vào tại thời điểm t xt được đưa vào nút mạng. Sau khi được xử lý qua các hàm kích hoạt sigmoid σ, tanh và các phép toán véc–tơ, kết quả đầu ra là trạng thái ô C và trạng thái ẩn h tại thời điểm t sẽ được sử dụng cho nút mạng t+1 tiếp theo.



Hình 6. Cấu trúc bên trong của mạng LSTM.

LSTM có một số ưu điểm như sau:

- Phụ thuộc xa có thể được nhắm bắt bởi các mạng LSTM. Một cell state có khả năng lưu trữ thông tin lâu dài.
- LSTM giải quyết được vấn đề vanishing gradient của mạng thần kinh hồi quy RNNs truyền thống.
- LSTM cho phép mô hình nắm bắt và ghi nhớ bối cảnh quan trọng, ngay khi có khoảng cách thời gian đáng kể giữa các sự kiện có liên quan trong chuỗi. Vì vậy, khi ngữ cảnh đóng vai trò quan trọng thì LSTM được sử dụng, ví dụ như máy dịch.

Bên cạnh đó, LSTM cũng có những nhược điểm như:

- So sánh với các cấu trúc đơn giản hơn như mạng thần kinh truyền thẳng, LSTM tiêu tốn hơn về mặt tính toán. Điều này có thể hạn chế khả năng mở rộng với các bộ dữ liệu có quy mô lớn học môi trường bị hạn chế.
- Huấn luyện LSTM tốn nhiều thời gian hơn so với các mô hình đơn giản do độ phức tạp tính toán của chúng. Vì vậy, huấn luyện LSTM thường đòi hỏi nhiều dữ liệu và thời gian lâu hơn để đạt hiệu suất cao.
- LSTM xử lý từng từ một các tuần tự nên rất khó để song song hóa công việc

xử lý các câu.

2. Phương pháp biểu diễn văn bản

Trong học máy, máy tính không thể hiểu trực tiếp ngôn ngữ tự nhiên mà chỉ hiểu được ngôn ngữ khi chúng được biểu diễn dưới dạng không gian vector. Các chiều thuộc tính đầu vào sẽ được biểu diễn dưới dạng ma trận vector, có nhiều phương pháp để biểu diễn văn bản sang dạng ma trận vector chẳng hạn: cách truyền thống như mô hình Bag of N-grams, mô hình TF-IDF, mô hình chủ đề hay các cách cải tiến như các mô hình Word2Vec, GloVe, FastTex (Sarkar, 2019). Trong nghiên cứu này, chúng tôi áp dụng hai phương pháp là TF-IDF và Word2Vec để thử nghiệm mô hình và biểu diễn dữ liệu.

a. TF-IDF:

TF-IDF là viết tắt của Term Frequency - Inverse Document Frequency. TF-IDF là trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản (Robertson & Stephen, 2004).

Thuật toán này được sử dụng vì: Trong ngôn ngữ luôn có những từ xảy ra thường xuyên với các từ khác. Và một trong những phát biểu nổi tiếng nhất Zipf's Law phát biểu về vấn đề này như sau: "The nth most common word in a human language text occurs with a frequency inversely proportional to n".

Có nghĩa là luôn có một tập các từ mà tần số xuất hiện, sử dụng nhiều hơn các từ khác, điều này đúng trong bất kỳ ngôn ngữ nào. Chính vì vậy ta cần có một phương pháp để cân bằng mức độ quan trọng giữa các từ.

- *Term Frequency*: Dùng để ước lượng tần suất xuất hiện của từ trong văn bản. Tuy nhiên với mỗi văn bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn . Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó).

$$TF(t,d) = \frac{\text{Số từ t trong d}}{\text{Số lượng từ trong d}}$$

- *Inverse Document Frequency*: dùng để ước lượng mức độ quan trọng của từ đó như thế nào . Khi tính tần số xuất hiện tf thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn, ví dụ:
    - Từ nối: và, nhưng, tuy nhiên, vì thế, vì vậy…

- Giới từ: ở, trong, trên…
- Từ chỉ định: ấy, đó, nhỉ…

Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF:

$$IDF(t, D) = log(\frac{\text{Số văn bản trong D}}{\text{Số văn bản chứa t trong D}})$$

- *TF-IDF* được tính như sau:

$$TF - IDF(t, d, D) = TF(t, d) \times DF(t, D)$$

Những từ có có giá trị TD-IDF cao là những từ:

- Xuất hiện nhiều trong văn bản d (TF cao)
- Xuất hiện ít trong các văn bản (IDF cao)

Chính điều này giúp lọc ra những từ phổ biến và giữ lại những giá trị cao (coi là từ hóa của văn bản). Phương pháp này có thể được sử dụng trong tìm kiếm văn bản.

TF-IDF vectorization: Tương tự như Bag of Word chúng ta dùng TF-IDF vectorization để biểu diễn mỗi document dưới dạng vector. Ở đây ta sẽ gán giá trị TF-IDF của mỗi từ tương ứng với vị trí của nó trong document (Bag of Words sẽ thay số lần xuất hiện của từ đó trong document).

b. Word2vec:

Mô hình word2vec có 2 phương pháp chính (Mikolov et al., 2013) là:

- Sử dụng ngữ cảnh để dự đoán mục tiêu (CBOW)
- Sử dụng một từ để dự đoán ngữ cảnh mục tiêu (skip-gram)

**** Skip-gram Model:***

Mục tiêu: Học trọng số các lớp ẩn, các trọng số này là các words vector.

Cách thức: Cho một từ cụ thể ở giữa câu(input word), nhìn vào những từ ở gần và chọn ngẫu nhiên. Mạng neural sẽ cho chúng ta biết xác suất của mỗi từ trong từ vựng về việc trở thành từ gần đó mà chúng ta chọn.

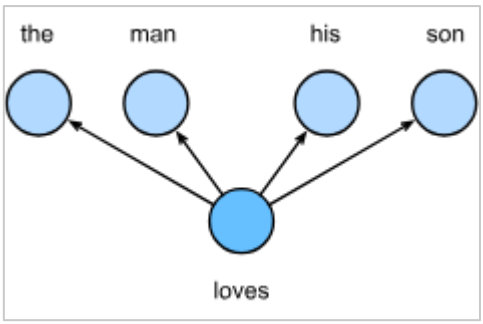
Ví dụ, giả sử chuỗi văn bản là “the”, “man”, “loves”, “his” và “son”. Ta sử dụng “loves” làm từ đích trung tâm và đặt kích thước cửa sổ ngữ cảnh bằng 2. Như mô tả trong hình, với từ đích trung tâm “loves”, mô hình skip-gram quan tâm đến xác suất có điều kiện sinh ra các từ ngữ cảnh (“the”, “man”, “his” và “son”) nằm trong khoảng cách không quá 2 từ:

$$P("the", "man", "his", "son" \mid "loves")$$

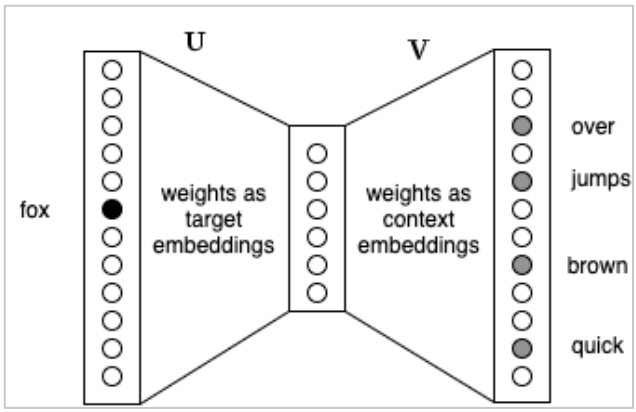
Ta giả định rằng, với từ đích trung tâm cho trước, các từ ngữ cảnh được sinh ra độc

lập với nhau. Trong trường hợp này, công thức trên có thể được viết lại thành:

$$P(\text{"the"} \mid \text{"loves"}) \cdot P(\text{"man"} \mid \text{"loves"}) \cdot P(\text{"his"} \mid \text{"loves"}) \cdot P(\text{"son"} \mid \text{"loves"})$$



Skip-gram word2vec là một mạng neural vô cùng đơn giản với chỉ một tầng ẩn không có hàm kích hoạt:

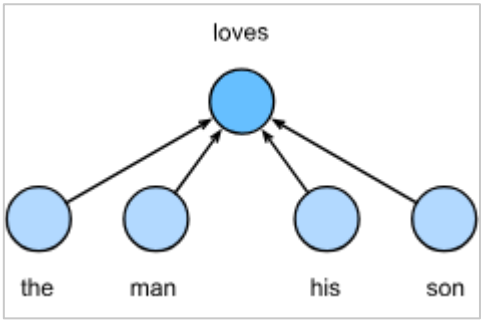


Hình 7. Minh họa Skip-gram dưới dạng mạng neural

**** CBOW***

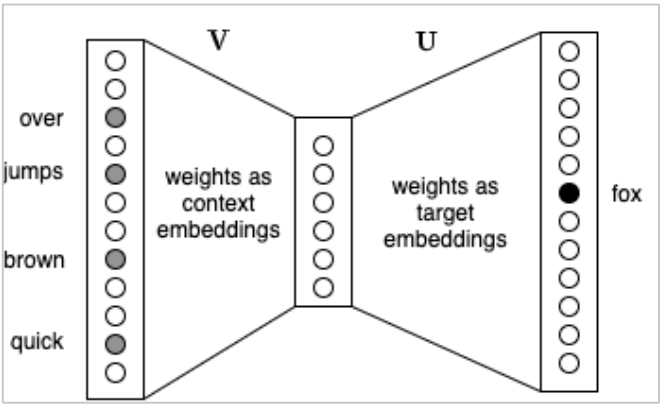
Chúng ta nhận thấy rằng mô hình skip-grams sẽ rất tốn chi phí để tính toán vì mẫu số xác suất là tổng của rất nhiều số mũ cơ số tự nhiên. Để hạn chế chi phí tính toán mô hình túi từ liên tục (Continuous bag of words - CBOW) được áp dụng. CBOW tương tự như mô hình skip-gram. Khác biệt lớn nhất là mô hình CBOW giả định rằng từ đích trung tâm được tạo ra dựa trên các từ ngữ cảnh phía trước và sau nó trong một chuỗi văn bản. Với cùng một chuỗi văn bản gồm các từ “the”, “man”, “loves”, “his” và “son”, trong đó “love” là từ đích trung tâm, với kích thước cửa sổ ngữ cảnh bằng 2, mô hình CBOW quan tâm đến xác suất có điều kiện để sinh ra từ đích “love” dựa trên các từ ngữ cảnh “the”, “man”, “his” và “son”.

$$P(\text{"loves"} \mid \text{"the"}, \text{"man"}, \text{"his"}, \text{"son"})$$



Vì có quá nhiều từ ngữ cảnh trong mô hình CBOW, ta sẽ lấy trung bình các vector từ của chúng làm "đại diện" và sau đó sử dụng phương pháp tương tự như trong mô hình skip-gram để tính xác suất có điều kiện.

Biểu diễn mạng neural cho CBOW được thể hiện như trong hình dưới đây:



Hình 8. Minh họa CBOW dưới dạng mạng neural

CHƯƠNG 4: KẾT QUẢ PHÂN TÍCH THỰC NGHIỆM

1. Tiền xử lý dữ liệu

Nhóm đã thực hiện xử lý và chạy thực nghiệm bài toán Sentiment Analysis để có thể hiểu rõ hơn về cơ chế hoạt động của ba thuật toán chính: Support Vector Machine, Maximum Entropy cụ thể là Ridge Regression, và mô hình học sâu Long Short-Term Memory. Như đã nêu ở trên, nhóm sử dụng một corpus tiếng việt UIT-VSFC (Vietnamese Student Feedback Corpus) chứa hơn 16.000 câu bình luận của học sinh, sinh viên đã được gán nhãn về cảm xúc. Tất cả đều được triển khai trên ngôn ngữ lập trình Python.

Dĩ nhiên trước khi thực hiện bài toán Sentiment Analysis, nhóm vẫn phải thực hiện các phương pháp tiền xử lý quen thuộc như viết thường, kiểm tra lỗi chính tả, xóa số và xóa các nội dung không liên quan trong corpus. Tuy nhiên, khác với thông thường, nhóm sẽ không thực hiện xóa đi các từ không mang ý nghĩa hay gọi là stopwords. Vì qua một số thực nghiệm thì nhóm nhận thấy được việc xóa đi stopwords sẽ làm mất đi một số ý nghĩa trong câu. Ví dụ: “Thầy chưa bao giờ làm học sinh thất vọng”. Thì việc xóa stopwords sẽ thành: “Thầy làm học sinh thất vọng”. Điều này là không tốt vì nó làm sai mặt ý nghĩa của câu sẽ ảnh hưởng đến kết quả.

Vì bộ dữ liệu sử dụng là bộ dữ liệu tiếng việt nên nhóm đã sử dụng thư viện underthesea để thực hiện tách từ (tokenizer). Các từ trong corpus sau khi tách ra thì được nhóm biểu diễn trên biểu đồ wordcloud để tiến hành quan sát các từ trong corpus:

Hình 9. Wordcloud các từ trong bộ dữ liệu

2. Thiết lập thực nghiệm

Sau khi đã thực hiện đầy đủ các bước tiền xử lý Corpus, việc tiếp theo đó chính là lựa chọn phương pháp vector hóa để có thể thực hiện huấn luyện.Tại đây nhóm sẽ thực hiện 2 phương pháp vector hóa chính là TF-IDF dựa trên Uni-gram và Hệ thống nhúng từ (Word Embeddings hay Word2Vec).

a. TF-IDF Vectorizer:

Phương pháp TF-IDF (Term Frequency-Inverse Document Frequency) là một kỹ thuật quan trọng trong xử lý ngôn ngữ tự nhiên và truy xuất thông tin.

Tần số xuất hiện của từ (TF – Term Frequency): Tính toán số lần xuất của mỗi từ trong một tài liệu cụ thể.

Tần số ngược của tài liệu (IDF - Inverse Document Frequency): Tính toán sức quan trọng của từng từ trong toàn bộ tập hợp các tài liệu.

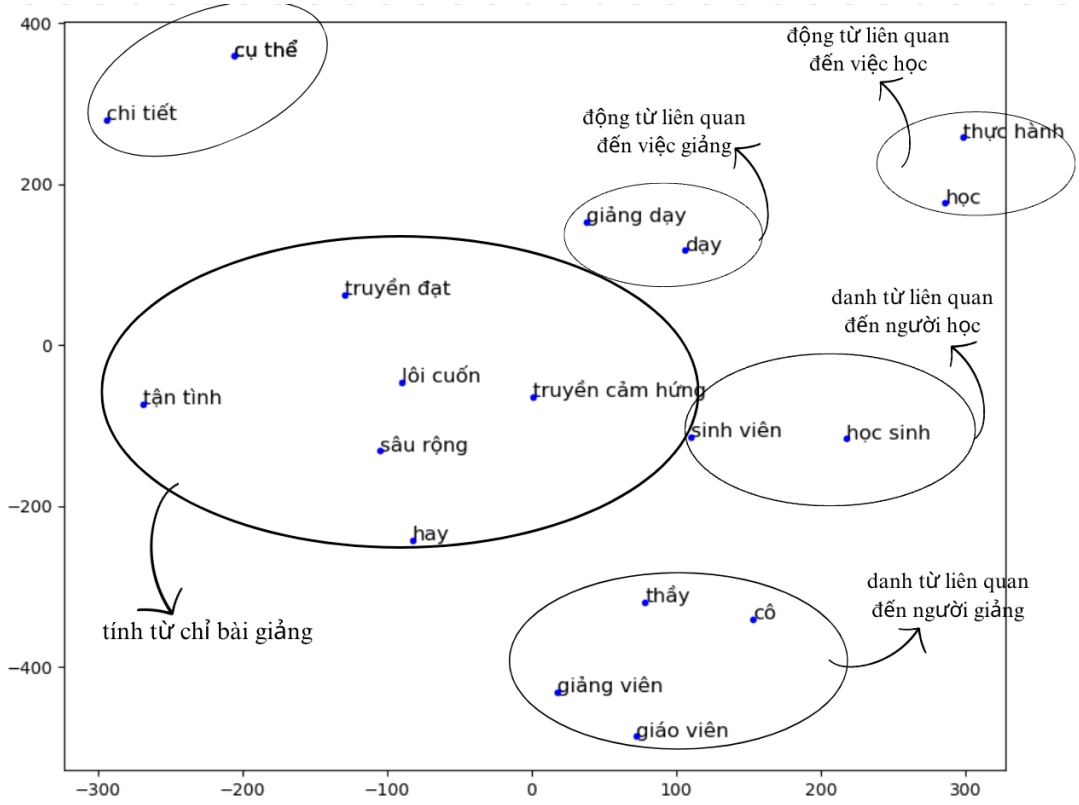
Nhóm đã xây dựng một lớp Vietnamese TF-IDF Vectorizer có thể được sử dụng cho ngôn ngữ Việt Nam dựa trên mô hình N-grams. Và sau khi thống kê với TF-IDF với mô hình Uni-gram, thì kết quả cho ra 4820 từ vựng Tiếng Việt khác nhau trong Corpus. Các câu được đưa vào vẫn sẽ được chuyển thành vector như bình thường.

b. Word2Vec:

Word2Vec là một phương pháp trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Mục tiêu của Word2Vec là chuyển đổi từ ngôn ngữ tự nhiên thành không gian vector, nơi các từ tương tự về ngữ nghĩa sẽ có biểu diễn gần nhau. Mô hình Word2Vec mà nhóm sẽ sử dụng chính là Skipgram.

Mô hình Skipgram được thiết kế để học biểu diễn vector cho từng từ thông qua việc dự đoán các từ xung quanh nó trong văn bản. Mục tiêu chính của Skipgram là tạo ra các vector từ sao cho các từ có ngữ nghĩa tương tự sẽ có vector gần nhau trong không gian vector. Nhóm quyết định sử dụng mô hình Skipgram vì nó có hiệu suất cao với dữ liệu lớn, cùng với đó khả năng chi tiết hóa từ vựng tốt hơn.

Nhóm đã thực hiện tiền huấn luyện (pre-trained) mô hình Skipgram trên chính Corpus sử dụng. Các từ sẽ được chuyển hóa thành các vector với kích thước 200 chiều, các từ mang ý nghĩa khác nhau sẽ có khoảng cách xa, độ tương đồng thấp. Dưới đây là một số kết quả của mô hình:



Hình 10. Vector các từ được biểu diễn trên không gian hai chiều

Đây là biểu đồ các từ trên không gian 2 chiều. Mặc dù, đã giảm chiều dữ liệu nhưng ta vẫn có thể thấy rõ có sự phân hoạch. Ví dụ các từ: "cô", "thầy", "giảng viên", "giáo viên" sẽ rất gần nhau, vì mang cùng ý nghĩa là về người giảng dạy. Tương tự, "sinh viên" và "học sinh" cũng rất gần nhau so với các từ khác. Các từ về cách giảng bài như: "hay", "sâu rộng", "lôi cuốn" nằm cùng về một cụm. Các từ về vấn đề học như học và thực hành.Có thể nó mô hình Word2Vec Skip-gram đã tạo ra các vector khá tốt.

3. Tiến hành thực nghiệm và kết quả

Đối với phương pháp TF-IDF thì nhóm chỉ thực hiện trên hai thuật toán SVM và Ridge Regression, lý do là vì phương pháp TF-IDF vector sẽ chuyển cả một câu thành một vector, không thể chuyển thành một chuỗi (Sequence) các vector như phương pháp bình thường để sử dụng cho LSTM. Ngoài ra, đối với phương pháp Ridge Regression, nhóm đã thực hiện Grid Search để tìm được giá trị alpha là 1,0. Dưới đây là kết quả khi sử dụng hai phương pháp SVM và Ridge Regression.

a. Ridge Regression với phương pháp TF-IDF:

	Precision	**Recall**	**F1-Score**
Tiêu cực	0.88	0.85	0.86
Trung Lập	0.32	0.10	0.15

Tích cực	0.84	0.92	0.88
Macro Average	0.68	0.62	0.63
Weighted Average	0.83	0.85	0.83

Bảng 3. Chỉ số kết quả mô hình Ridge Regression với pp TF-IDF

Kết luận: Độ chính xác (Accuracy) của Ridge Regression với pp TF-IDF là 0.85

b. Support Vector Machine với phương pháp TF-IDF:

	Precision	**Recall**	**F1-Score**
Tiêu cực	0.86	0.95	0.91
Trung Lập	0.62	0.05	0.09
Tích cực	0.90	0.91	0.90
Macro Average	0.79	0.64	0.63
Weighted Average	0.87	0.88	0.86

Bảng 4. Chỉ số kết quả mô hình SVM với pp TF-IDF

Kết luận: Độ chính xác (Accuracy) của SVM với pp TF-IDF là 0.88

Đối với phương pháp Word2Vec, thì nhóm có hai phương pháp tiếp cận để có thể phù hợp với các thuật toán mà nhóm đang sử dụng. Vì yêu cầu đầu vào để huấn luyện cho hai phương pháp SVM và Ridge Regression là các vector có cùng kích thước, nên nhóm đã xây dựng một phương thức để có thể chuyển các câu thành vector dựa trên các vector có sẵn trong mô hình Word2Vec bằng cách ghép lại các vector đó, và thêm vào các vector giá trị 0 để đảm bảo kích thước. Việc làm này sẽ giúp cho câu không bị mất nhiều ý nghĩa và phù hợp với hai phương pháp SVM và Ridge Regression. Dưới đây là kết quả 2 mô hình Ridge Regression và SVM với phương pháp Word2Vec:

c. Ridge Regression với phương pháp Word2Vec:

	Precision	**Recall**	**F1-Score**
Tiêu cực	0.87	0.91	0.89
Trung Lập	0.30	0.02	0.03
Tích cực	0.87	0.92	0.90
Macro Average	0.68	0.62	0.61
Weighted Average	0.84	0.87	0.85

Bảng 5. Chỉ số kết quả mô hình Ridge Regression với pp Word2Vec

Kết luận: Độ chính xác (Accuracy) của Ridge Regression với pp Word2Vec là 0.87

d. SVM với phương pháp Word2Vec:

	Precision	Recall	F1-Score
Tiêu cực	0.85	0.95	0.90
Trung Lập	0.67	0.06	0.11
Tích cực	0.91	0.91	0.91
Macro Average	0.81	0.64	0.64
Weighted Average	0.87	0.88	0.86

Bảng 6. Chỉ số kết quả mô hình SVM với pp Word2Vec

Kết luận: Độ chính xác (Accuracy) của SVM với pp Word2Vec là 0.88.

e. LSTM với phương pháp Word2Vec:

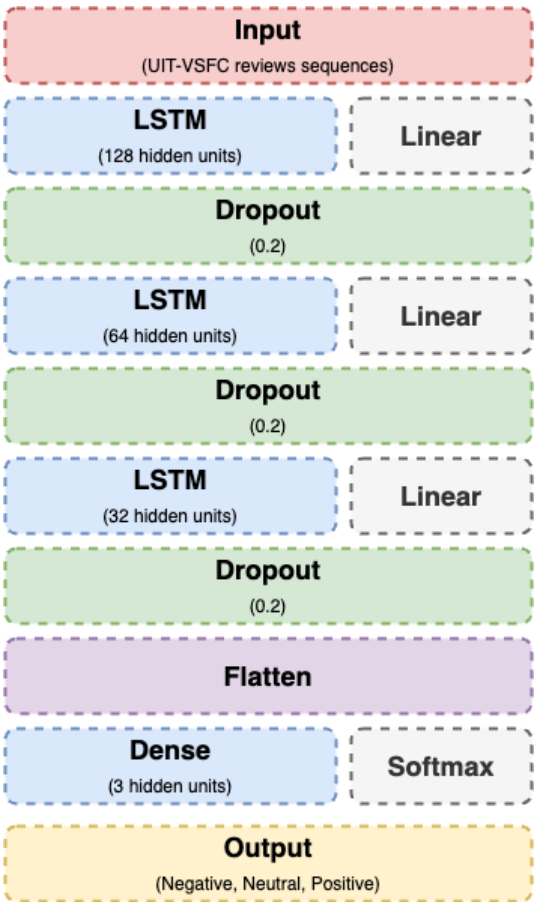
Đầu vào của mô hình LSTM, sẽ là một chuỗi các vector từ, nên nhóm đã xây dựng một phương thức chuyển đổi một câu thành ma trận với với số lượng vector ứng với số từ có trong câu và mỗi từ là vector. Kiến trúc của mô hình bắt đầu bằng một lớp đầu vào (Input Layer) với là các chuỗi các vector. Tiếp theo là ba lớp LSTM liên tiếp, mỗi lớp với số đơn vị tương ứng là 128, 64, và 32. Các lớp LSTM được thiết lập để trả về chuỗi, giúp giữ lại thông tin tại mỗi bước thời gian. Sau mỗi lớp LSTM là một lớp Dropout với tỷ lệ là 0.2, được áp dụng để ngăn chặn việc quá mức hóa (overfitting) bằng cách loại bỏ ngẫu nhiên một phần của các đơn vị đầu vào trong quá trình huấn luyện. Sau cùng, một lớp Flatten được sử dụng để làm phẳng đầu ra từ các lớp LSTM thành một mảng một chiều. Cuối cùng, một lớp Output với 3 đơn vị và hàm kích hoạt Softmax được thêm vào để thực hiện phân loại đa lớp, chuyển đổi đầu ra thành xác suất của ba lớp. Ngoài ra, nhóm đã áp dụng hàm Adam để tối ưu hóa cho việc huấn luyện mô hình với gia tốc học – learning rate là 0.01

Dưới đây là kết quả chạy mô hình LSTM:

	Precision	Recall	F1-Score
Tiêu cực	0.92	0.92	0.92
Trung Lập	0.42	0.48	0.45
Tích cực	0.93	0.92	0.92
Macro Average	0.76	0.77	0.77
Weighted Average	0.90	0.90	0.90

Bảng 7. Chỉ số kết quả mô hình LSTM với pp Word2Vec

Kết luận: Độ chính xác (Accuracy) của LSTM với phương pháp Word2Vec là 0.90



Hình 11. Cấu trúc mô hình LSTM được sử dụng thực nghiệm

Bảng tổng hợp so sánh các mô hình thực nghiệm (%):

Algorithms	Features	Precision	Recall	F1-Score	Accuracy
SVM	TF-IDF (Uni-gram)	86.9	88.1	86.1	88.1
	Word2Vec	87.2	88.0	86.1	88.0
MaxEnt	TF-IDF (Uni-gram)	82.8	84.6	83.3	84.6
	Word2Vec	84.0	86.8	84.7	86.8
LSTM	Word2Vec	**90.0**	**89.7**	**89.8**	**89.7**

Bảng 8. Bảng tổng hợp kết quả các mô hình thực nghiệm

4. Phân tích, đánh giá

Qua bảng tổng kết kết quả chạy các mô hình, nhóm đưa ra một số đánh giá sau:

- Đối với mô hình SVM, các kết quả đều khả quan, SVM thể hiện hiệu suất ổn định cho cả 2 loại đặc trưng Word2Vec và TF - IDP. Các chỉ số đánh giá đều đạt ở mức cao, cả hai đặc trưng đều trên mức 80%. Word2Vec dường như mang lại một chút cải thiện hơn so với TF - IDP, đặc biệt là ở chỉ số Precision.

- Đối với mô hình Maxent sử dụng Ridge Regression, kết quả cho ra cũng cho thấy hiệu suất tốt, và Word2Vec cũng cho ra một kết quả tốt hơn so với TF - IDP.
- Đối với mô hình LSTM với Word2Vec: cho ra kết quả tốt nhất trong 3 mô hình được sử dụng, với precision, accuracy lên tới 90%. Điều này cho thấy tính khả thi của việc áp dụng mô hình này vào việc phân tích cảm xúc thông qua các phản hồi. Làm tăng tính toàn diện và đáng tin cậy của mô hình.

Tổng kết: Ba thuật toán đã được đánh giá đều cho thấy hiệu suất tích cực trong nghiên cứu này. Word2Vec thường mang lại sự cải thiện so với TF-IDF trong việc phân loại cảm xúc từ phản hồi sinh viên, nhấn mạnh sức mạnh của việc sử dụng các biểu diễn từ vựng phong phú hơn.

Đặc biệt, mô hình LSTM với Word2Vec đã đạt hiệu suất cao nhất trong tất cả các chỉ số đánh giá. Tuy nhiên, để xây dựng được mô hình này đòi hỏi sự phức tạp cao hơn về mặt thời gian và xây dựng cao hơn so với 2 mô hình còn lại. Điều này có thể làm tăng chi phí và yêu cầu nguồn lực lớn hơn.

Với những ứng dụng đòi hỏi độ chính xác cao và cần thời gian đào tạo ngắn, thi SVM hoặc MaxEnt có thể là sự lựa chọn hợp lý. Bởi cả hai thuật toán này đều đạt được hiệu suất đáng kể và có thể được triển khai một cách linh hoạt. Còn đối với các yêu cầu đòi hỏi độ chính xác cao, sẵn sàng về mặt thời gian và nguồn lực thì LSTM là lựa chọn tốt nhất. Đối với ứng dụng website mà nhóm xây dựng trong phần sau, nhóm ưu tiên thuật toán có độ chính xác cao hơn nên sẽ sử dụng mô hình LSTM sử dụng Word2Vec (Skip-gram) để làm mô hình phân tích cảm xúc cho ứng dụng

CHƯƠNG 5: THIẾT KẾ ỨNG DỤNG

1. Bối cảnh ứng dụng

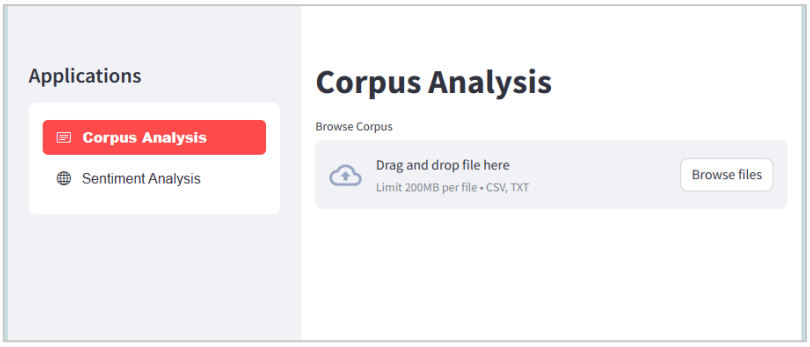
Các trường đại học ở Việt Nam hiện nay luôn không ngừng cải thiện chất lượng dịch vụ, cơ sở vật chất của mình để sinh viên người học có trải nghiệm học tập tốt nhất tại môi trường đại học. Chính vì thế, ý kiến của sinh viên, người học chính là những ý kiến liên quan trực tiếp nhất đến chất lượng dịch vụ (chương trình đào tạo, giảng dạy…) mà trường học cần phải cải thiện. Vào mỗi học kỳ, các trường đại học đều thực hiện khảo sát cho sinh viên về các vấn đề như giảng viên, học phần, khối lượng học tập… để có thể cải thiện vào các học kỳ sau. Nhưng ngoài những lựa chọn trắc nghiệm thể hiện sự hài lòng thì vẫn có những câu hỏi mở để sinh viên đóng góp ý kiến. Đối với loại trả lời này cần có một ứng dụng giúp những người quản lý phân loại ý kiến tích cực, tiêu cực và trung lập để nhận ra những vấn đề mà dịch vụ giảng dạy cung cấp cho người học đang gặp phải.

Vì lý do trên, nhóm đã áp dụng thuật toán LSTM sử dụng Word2Vec để thiết kế một website có thể giúp phân loại một feedback hoặc một bộ các feedback là tích cực hay tiêu cực. Để người quản lý có thể sử dụng thuận tiện và trực quan hơn.

2. Thiết kế giao diện

a. Corpus Analysis:

Ứng dụng "Corpus Analysis" giúp người dùng đưa vào một file gồm nhiều feedback thông qua thu thập từ người học dưới định dạng file csv hoặc txt với kích cỡ dưới 200MB. Kết quả trả về sẽ phân loại cảm xúc cho từng feedback trong corpus và vẽ biểu đồ tỷ lệ các cảm xúc trong corpus. Giao diện chung của phần này trong hình dưới:

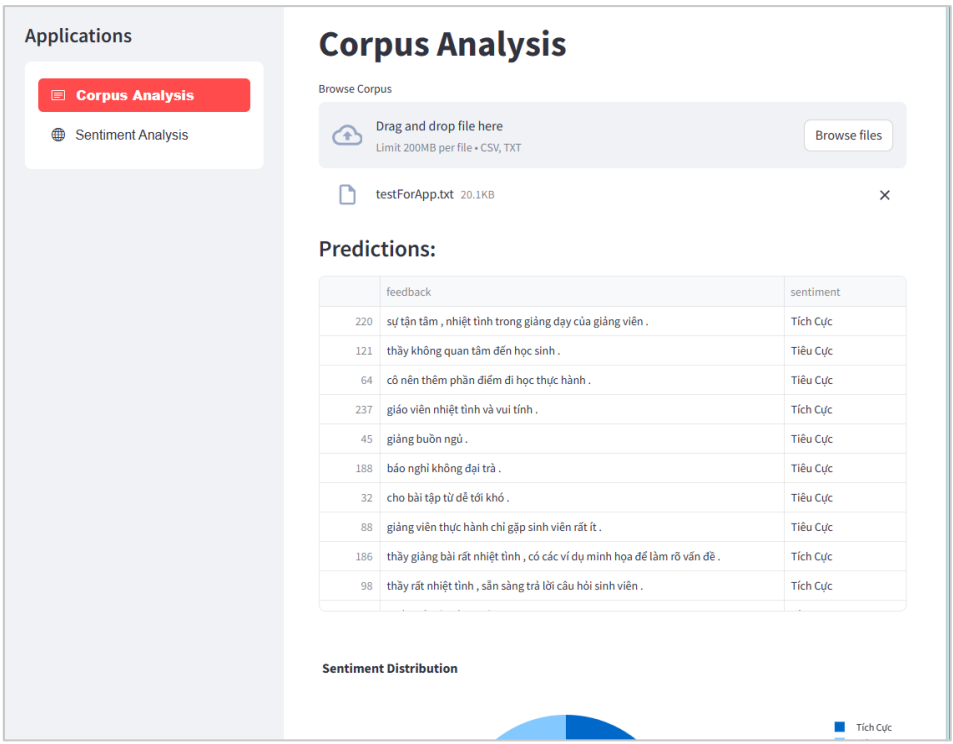


Hình 12. Giao diện chính phần Corpus Analysis

Để sử dụng ta thực hiện các bước:

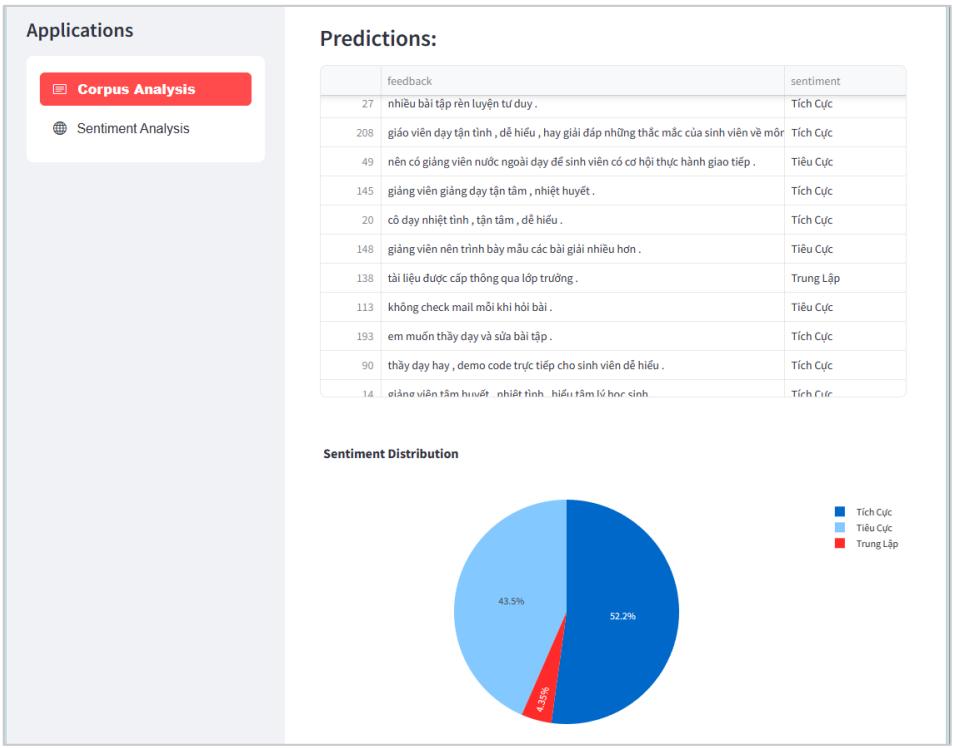
- Chọn Browse File và đưa vào file muốn thực hiện phân tích cảm xúc:
- Các cảm xúc được mô hình học sâu phân lớp sẽ được thể hiện ở bảng trong

phần “Predictions”



Hình 13. Kết quả thực hiện Corpus Analysis (Predictions)

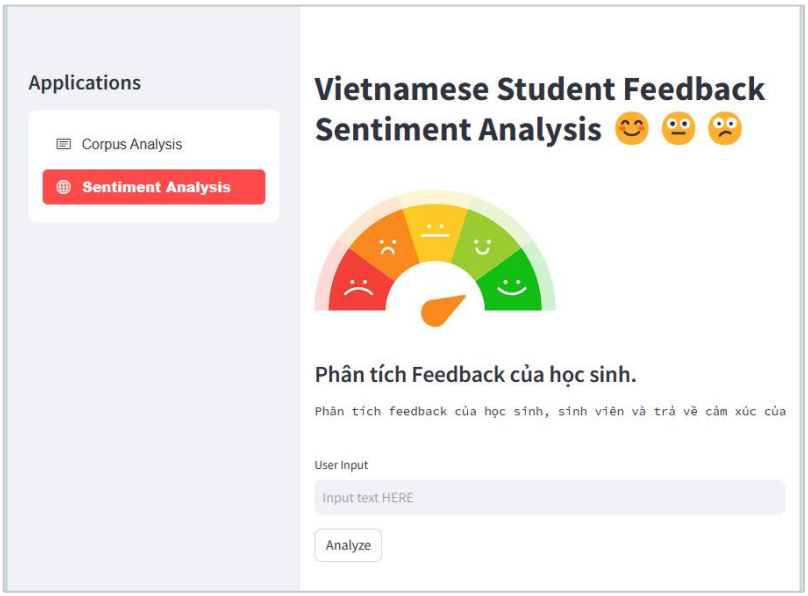
- Số lượng cảm xúc sẽ được trực quan hóa thành biểu đồ tròn trong phần “Sentiment Distribution”



Hình 14. Kết quả thực hiện Corpus Analysis (Sentiment Distribution)

b. Sentiment Analysis:

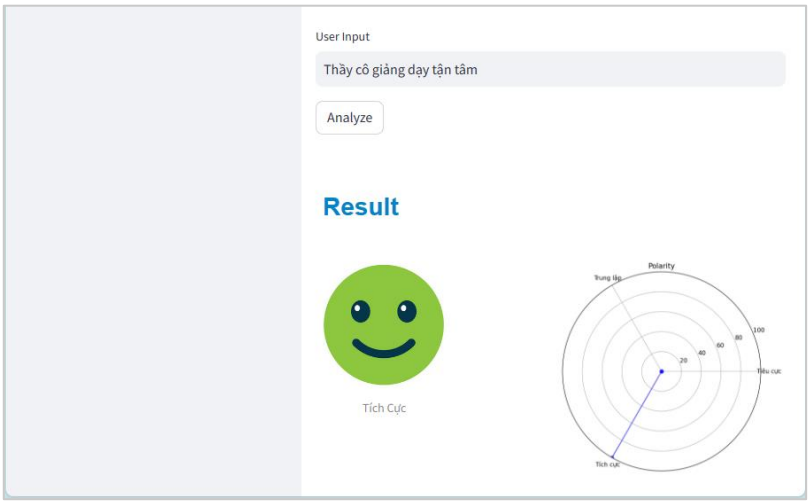
Ứng dụng "Sentiment Analysis" dùng để mô phỏng kết quả hoạt động của thuật toán hơn là dùng trong thực tế. Khi người dùng nhập một câu vào phần User Input, website sẽ trả về kết quả phân tích cảm xúc của câu đó là Tích cực, Tiêu cực hay Trung lập. Giao diện chi tiết của ứng dụng này trong hình dưới đây.



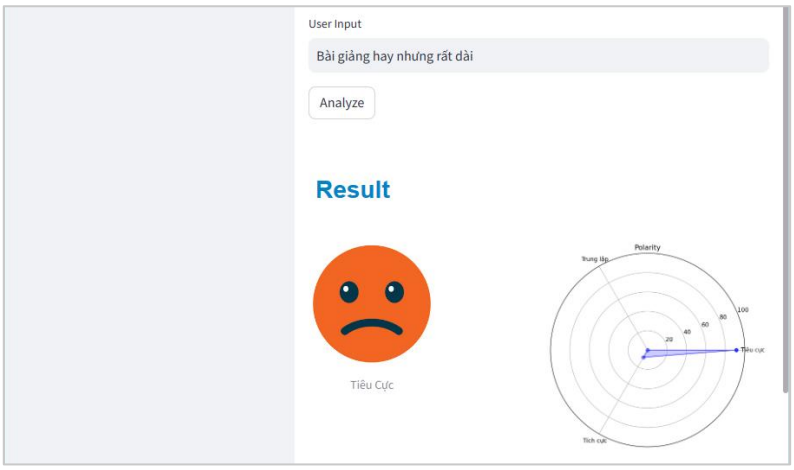
Hình 15. Giao diện chính phần Sentiment Analysis

Để sử dụng ta thực hiện các bước sau:

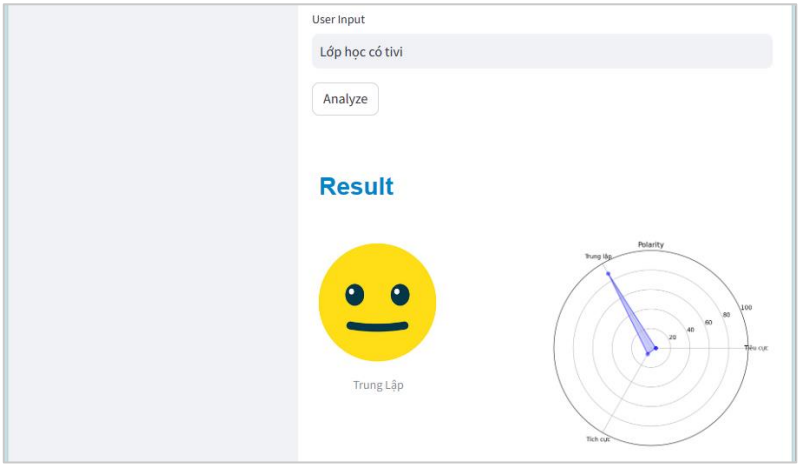
- Nhập feedback của người dùng vào "User Input" sau đó nhấn Analyze;
- Kết quả trả về sẽ bao gồm nhãn lớp và một biểu đồ radar thể hiện mức độ tích cực, tiêu cực hay trung lập của câu input đó;
- Phía dưới là các kết quả trả về ứng với mỗi nhãn lớp.



Hình 16. Giao diện kết quả cho label Tích cực



Hình 17. Giao diện kết quả cho label Tiêu cực



Hình 18. Giao diện kết quả cho label Trung lập

3. Đánh giá giao diện

Nhìn chung, giao diện website đã đáp ứng được những yêu cầu mà nhóm đặt ra cho một ứng dụng phân tích cảm xúc feedback của sinh viên với phần giao diện trực quan rõ ràng, thân thiện với người sử dụng. Ứng dụng website cho phép người dùng dễ dàng truy cập hơn so với ứng dụng desktop. Ứng dụng chia làm hai phần đáp ứng nhu cầu của nhóm. Tuy nhiên, ứng dụng website vẫn có những hạn chế như thời gian truy cập vào giao diện lâu, phần "Corpus Analysis" mất nhiều thời gian thực hiện cho phần phân tích cảm xúc đối với những bộ dữ liệu lớn. Phần trực quan chỉ mới tập trung vào tỷ lệ các cảm xúc trong corpus mà chưa phân tích đối với từng chủ đề, điều này sẽ làm các nhà quản lý khó để nhận định vấn đề nhận được những ý kiến tiêu cực hoặc tích cực thuộc về lĩnh vực nào. Điều này đang là hạn chế của giao diện nói riêng, và đồ án nói chung, đồng thời cũng sẽ là một hướng phát triển cho bài toán sau này.

CHƯƠNG 6: KẾT LUẬN

1. Kết quả đạt được

Trong bài nghiên cứu này, nhóm nghiên cứu đã thực hiện xây dựng một mô hình phân tích cảm xúc dựa trên phản hồi của sinh viên, sử dụng các thuật toán SVM, MaxEnt, và LSTM. Đồng thời, bài nghiên cứu đã áp dụng hai đặc trưng quan trọng là TF-IDF và Word2Vec để đánh giá hiệu suất của mô hình. Kết quả cho thấy rằng, trong bối cảnh này, mô hình Long Short-Term Memory sử dụng skipgram của Word2Vec thể hiện kết quả tốt hơn so với những mô hình còn lại.

Ngoài ra, nhóm đã tiến hành thiết kế một giao diện website hữu ích cho phép người dùng nhập vào phản hồi và tự động phân tích xem đó là phản hồi tích cực, tiêu cực, hay trung lập. Giao diện người dùng này không chỉ giúp đơn giản hóa quá trình phân tích cảm xúc mà còn mang lại tính ứng dụng cao trong thực tế, giúp các tổ chức và giáo viên nhanh chóng đánh giá sự hài lòng và ý kiến của sinh viên đối với các khía cạnh khác nhau của học tập và giảng dạy.

Sự kết hợp giữa mô hình phân tích cảm xúc và giao diện người dùng thân thiện đã tạo nên một chương trình có tính ứng dụng cao trong thực tế, mang lại giá trị lớn cho cộng đồng giáo dục và quản lý học thuật. Các kết quả từ nghiên cứu có thể đóng góp quan trọng cho việc phát triển các hệ thống tự động đánh giá và theo dõi cảm xúc của sinh viên trong môi trường giáo dục

2. Hạn chế đề tài

Bài toán đặt ra tuy đã được giải quyết nhưng vẫn còn những hạn chế mà đề tài của nhóm chưa đáp ứng được hay là thực hiện một cách tốt nhất. Các hạn chế trong đồ án được nhóm phát hiện ra:

- Mặc dù nhóm đã sử dụng bộ dữ liệu lớn, nhưng còn một số chức năng của bộ dữ liệu chưa được tận dụng hết. Mô hình chỉ đơn giản thực hiện phân loại cảm xúc, mà không phân loại chi tiết theo từng chủ đề hay lĩnh vực cụ thể. Điều này gây thiếu sót trong việc hiểu rõ hơn về cảm xúc và ý kiến của sinh viên đối với từng khía cạnh của dịch vụ giáo dục, bộ dữ liệu có phần topics sẽ giải quyết được nhưng chưa được nhóm sử dụng;
- Trong quá trình xử lý, nhóm không thực hiện loại bỏ các stopwords, điều này có thể ảnh hưởng đến kết quả mô hình do có một số từ xuất hiện nhiều mà không mang ý nghĩa. Ví dụ, các từ như "và", "rất",... không được loại bỏ;
- Mặc dù độ chính xác của mô hình đạt 89,7% đây không phải là một chỉ số thấp. Đối

với bộ dữ liệu lớn, tỷ lệ sai sót này cho ra một lượng lớn kết quả được phân loại sai nhãn.

- Một vấn đề lớn mà nhóm đang đối mặt là thời gian khởi động của trang web khá lâu, đặc biệt là khi nhóm thực hiện trên một bộ dữ liệu lớn. Điều này làm giảm trải nghiệm người dùng khi sử dụng website.
- Bài nghiên cứu hiện chỉ tập trung vào việc huấn luyện mô hình với bộ dữ liệu phản hồi của sinh viên, điều này hạn chế phạm vi của mô hình chỉ trong lĩnh vực giáo dục.

3. Hướng phát triển

Đề tài mở ra được nhiều hướng phát triển mới và đầy tiền năng. Sau đây là một số hướng phát triển mà nhóm xem xét đưa ra để bài toán được giải quyết tối ưu nhất:

- ***Tận dụng hết các chức năng của bộ dữ liệu***: Phát triển mô hình để có khả năng phân lớp theo từng khía cạnh cụ thể trong lĩnh vực giáo dục. Điều này sẽ giúp các bộ phận quản lý có cái nhìn chi tiết về những điều tích cực và tiêu cực của từng khía cạnh, từ đó dễ dàng đề xuất và thực hiện các cải tiến phù hợp.
- ***Loại bỏ Stopwords hiệu quả***: Thực hiện loại bỏ stopword với tập stopword hiệu quả. Điều này giúp tăng hiệu suất của mô hình bằng cách giảm ảnh hưởng của các từ không mang ý nghĩa và tăng chất lượng của dữ liệu đầu vào.
- ***Tối ưu thuật toán***: Tiếp tục tối ưu hóa thuật toán để đạt được độ chính xác cao hơn và giảm tỷ lệ sai sót. Quá trình này sẽ làm tăng tính chính xác của mô hình trong việc phân loại cảm xúc. Ngoài ra, nhóm sẽ tiếp tục nghiên cứu một số thuật toán khác như: FastText, CNN để đa dạng hóa phương pháp nghiên cứu và tăng cường khả năng của mô hình.
- ***Tối ưu giao diện và giảm thời gian khởi động***: Tối ưu hóa giao diện người dùng để tạo ra một trải nghiệm người dùng tốt hơn và giảm thời gian chạy của trang web. Điều này giúp người dùng trải nghiệm dịch vụ một cách mượt mà và hiệu quả hơn.
- ***Mở rộng phạm vi nghiên cứu***: Mở rộng phạm vi nghiên cứu sang phân tích cảm xúc dựa trên phản hồi và đánh giá từ nhiều lĩnh vực khác nhau. Ví dụ, trong lĩnh vực bán hàng trực tuyến, áp dụng mô hình này vào phân tích cảm nhận của khách thông qua các phản hồi của khách hàng. Cải tiến dịch vụ, thu hút nhiều khách hàng hơn.

TÀI LIỆU THAM KHẢO

[1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1-135.

[2] Kechaou, Z., Ammar, M. B., & Alimi, A. M. (2011, April). Improving e-learning with sentiment analysis of users' opinions. In *2011 IEEE global engineering education conference (EDUCON)* (pp. 1032-1038). IEEE.

[3] Singla, Z., Randhawa, S., & Jain, S. (2017, July). Statistical and sentiment analysis of consumer product reviews. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[4] Le, A. C. (2018, November). Integrating grammatical features into CNN model for emotion classification. In *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 243-249). IEEE.

[5] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[6] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

[7] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[8] Robertson, Stephen. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. Journal of Documentation - J DOC. 60. 503-520. 10.1108/00220410410560582.

[9] Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

PHỤ LỤC

1. **Link Github:** *(Source Code)*

2. **Lưu ý chạy chương trình:**

- Clone the Project: `git clone https://github.com/VuBacktracking/`
- Cài đặt các thư viện yêu cầu: `pip install -r requirements.txt`
- Cách chạy chương trình: `streamlit run app.py`

3. **Phân công công việc:**

STT	Họ và tên	MSSV	Nhiệm vụ
1	Phan Dương Hoàng Vũ	**31211022533**	1. Code thuật toán phân tích cảm xúc 2. Thiết kế giao diện 3. Thực nghiệm mô hình 4. Đưa source code lên Github
2	Đỗ Quang Thiên Phú	**31211024191**	1. Tìm kiếm bộ dữ liệu 2. Vietnamese Student's Feedback Dataset 3. Deep Learning và Vectorizer 4. Phân tích thiết kế giao diện 5. Tổng hợp word
3	Phạm Dương Thị Thúy Truyền	**31211027682**	1. Mở đầu 2. SVM và Maximum Entropy 3. Phân tích đánh giá kết quả thực nghiệm 4. Kết quả, hạn chế và hướng phát triển 5. Slide thuyết trình