ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KĨ THUẬT THÔNG TIN

BÁO CÁO ĐỒ ÁN
MÔN HỌC MÁY THỐNG KÊ

Đề tài:

Phân loại cảm xúc người dân khi thăm khám tại bệnh viện qua review trên Google maps

GVHD: ThS. Nguyễn Văn Kiệt
CN. Trần Quốc Khánh
CN. Nguyễn Hiếu Nghĩa

Nhóm sinh viên thực hiện: Nhóm 7

1.	Trần Tuyết Minh	MSSV: 21521144
2.	Lê Đào Xuân Thành	MSSV: 21522595
3.	Trần Tấn Thịnh	MSSV: 21522639
4.	Mai Hiếu Hiền	MSSV: 20521305

□□ Tp. Hồ Chí Minh, 1/2024 □□

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

……, ngày…….....tháng……năm 2020

Người nhận xét

(Ký tên và ghi rõ họ tên)

LỜI CẢM ƠN

Đầu tiên, nhóm chúng em xin được gửi lời cảm ơn chân thành đến Trường Đại Học Công Nghệ Thông Tin, Khoa Khoa Học và Kỹ Thuật Thông Tin đã tạo điều kiện cho chúng em có thể học tập và tiếp cận với môn học Học Máy Thống Kê, một môn học đầy tính thiết thực và ứng dụng cao. Chúng em xin gửi lời cảm ơn sâu sắc tới thầy Nguyễn Văn Kiệt cũng như sự hỗ trợ của anh Trần Quốc Khánh, Nguyễn Hiếu nghĩa đã chỉ dạy tận tình, trang bị cho chúng em những kiến thức, kỹ năng cần thiết để chúng em có thể học tập và đặc biệt là có thể hoàn thành được công trình nghiên cứu này.

Tuy nhiên, do kiến thức chuyên môn và kỹ năng còn hạn chế nên trong quá trình nghiên cứu còn có nhiều sai sót. Rất mong nhận được sự quan tâm, đánh giá và góp ý của thầy cô để chúng em có thể rút được kinh nghiệm, học hỏi thêm để hoàn thiện hơn bản thân, đáp ứng kỳ vọng của quý thầy cô.

Chúng em xin chân thành cảm ơn!

DANH MỤC CÁC BẢNG, HÌNH ẢNH

Danh mục các bảng:

Danh mục hình ảnh:

MỤC LỤC

Chương 1: Giới Thiệu

1.1 Giới thiệu tổng quan:

Với sự ứng dụng mạnh mẽ của công nghệ thông tin vào đời sống, việc áp dụng nó trong y tế là một việc cấp thiết hơn bao giờ hết. Ngày nay khi mà cảm xúc của con người đang được chú trọng và càng mạnh mẽ hơn nhất là đối với môi trường y tế thì việc thấu hiểu nó sẽ là một bước tiến lớn góp phần đổi mới cách hoạt động cũng như nâng cao chất lượng của các bệnh viện nói riêng và y tế nói chung. Vì thế mà bài toán phân loại cảm xúc người bệnh tại các bệnh viện là một bài toán quan trọng của lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và có thể đóng góp rất nhiều vào việc cải thiện chăm sóc sức khỏe và tương tác giữa bệnh nhân và nhân viên y tế. Nó thường được áp dụng trong các tình huống như đánh giá tâm trạng của bệnh nhân dựa trên các bản ghi lịch sử bệnh, ghi chú y tế, cuộc họp giữa bác sĩ và bệnh nhân, hoặc thậm chí qua các phản hồi từ các nền tảng mạng như website, google maps,.. . Bài toán này đặt ra mục tiêu giúp phân loại và đánh giá các cảm xúc của người bệnh đối với dịch vụ và thái độ dựa trên thông tin thu thập từ các nguồn khác nhau từ các bệnh viện trong hệ thống y tế của nước ta. Với các mô hình máy học hiện đại ngày này, việc phân loại cảm xúc đang dần trở nên dễ dàng hơn và cũng như là một đề tài được nghiên cứu nhiều trên các tạp chí và hội nghị vì tính tiềm năng của nó , song vẫn còn hạn chế về số lượng mô hình thử nghiệm, các cách trích xuất đặc trưng và áp dụng mô hình học máy để so sánh.

1.2 Mục tiêu đề tài

Mục tiêu đặt ra của bài toán này là xác định được một mô hình có khả năng phân loại cảm xúc người thăm khám tốt nhất sử dụng nhiều loại mô hình học máy. Qua đó chúng tôi muốn truyền tải một cái nhìn tổng quan, khái quát và có sự so sánh cụ thể giữa các mô hình với nhau.

Chương 2: CƠ SỞ LÝ THUYẾT

2.1 Giới thiệu mô hình

Bước đầu tiên trong phân loại bình luận về các bệnh viện trên google maps, chúng tôi tiến hành thu thập các bình luận trực tiếp từ google maps, tiếp theo chúng tôi có dùng kỹ thuật trích xuất đặc trưng TF-IDF cho các mô hình máy học và sử dụng nhiều mô hình máy học cũng như các mô hình transformer để đánh giá kết quả nhằm có thể đưa ra mô hình tối ưu nhất cho bài toán của chúng tôi

2.2 Cơ sở lý thuyết

2.2.1 Trích xuất đặc trưng TF-IDF

2.2.1.1. Khái niệm:

TF-IDF là một phương pháp trích xuất đặc trưng được biết đến rộng rãi trong việc xác định tầm quan trọng của một từ trong đoạn văn bản, nó thường được sử dụng như một trọng số trong việc khai phá dữ liệu dạng văn bản.

2.2.1.2. Công Thức Tính Toán:

TF: Là tần suất suất xuất hiện của một từ trong đoạn văn bản. Nó được tính bởi công thức:

$$Tf\ (t) = \frac{f(t,d)}{T}$$

là một từ trong đoạn văn bản $f(t,d)$ là tần xuất xuất hiện của t trong đoạn văn bản T là tổng số từ trong văn bản đó

IDF: Tính toán tầm quan trọng của một từ. Công thức tính IDF như sau:

$$idf(t) = \log\frac{N}{(t\epsilon D: t\epsilon d)}$$

N là tổng số đoạn văn bản. Tập $|(t\epsilon D: t\epsilon d)|$ là số văn bản chứa từ t

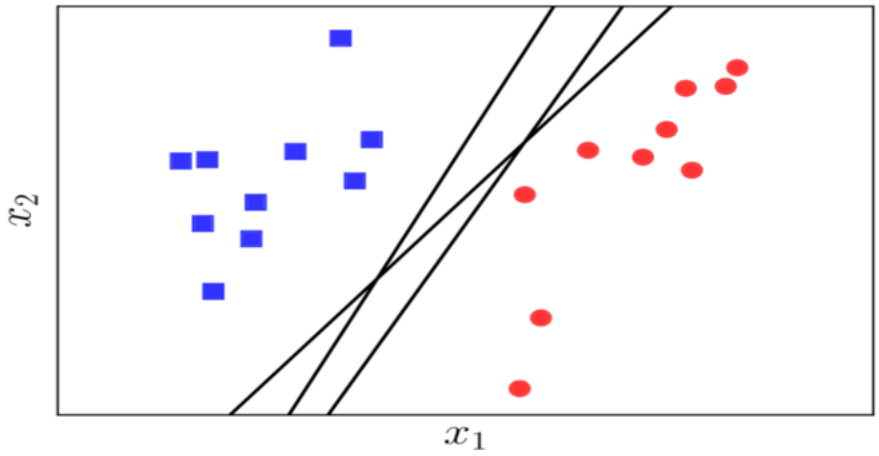
Từ đó, TF-IDF được tính bởi công thức: $\mathrm{tf} - \mathrm{idf(t)} = \mathrm{tf(t)} \times \mathrm{idf(t)}$

2.2.2. Mô hình huấn luyện:

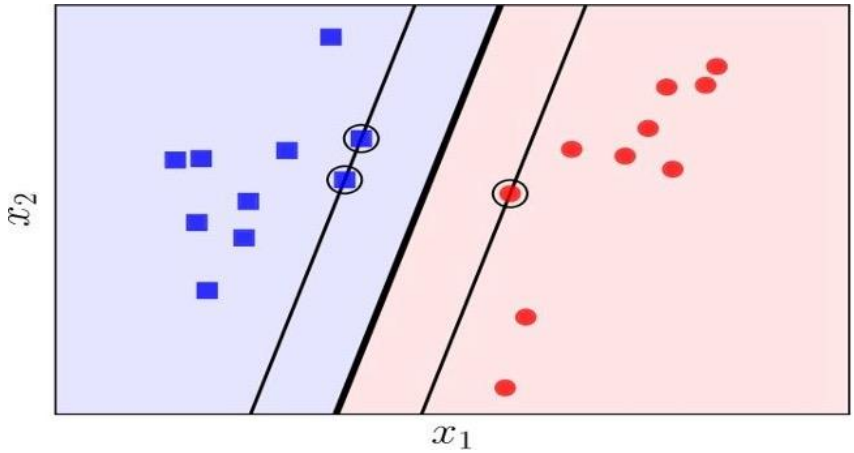
2.2.2.1. Support Vector Machine (SVM):

SVM là một trong những thuật toán phân lớp phổ biến và hiệu quả, nó được sử dụng rất nhiều trong Machine Learning và còn được biết đến như là một thuật toán tiệm cận nhất với các mô hình học sâu. Huấn luyện mô hình này là việc đi tìm một siêu phẳng để phân chia các nhãn với nhau, việc này gần giống như thuật toán Logistic Regression. Những điểm khiến thuật toán này vượt trội hơn là siêu phẳng được xác định còn thỏa mãn được khoảng cách tới các điểm dữ liệu gần nhất giữa các lớp là lớn nhất (maximum margin), đồng nghĩa với việc

các điểm dữ liệu sẽ có một khoảng cách an toàn tới mặt phân cách. Mô hình sẽ dự đoán các điểm dữ liệu mới dựa trên siêu phẳng tìm được (hay còn gọi là support vectors).



Hình 2.1 Mô hình phân loại tuyến tính thông thường



Hình 2.2 Mô hình phân loại với SVM

Đặc điểm đặc biệt của SVM:

Thay vì dựa trên các đặc trưng nhất của các điểm dữ liệu ứng với từng lớp để phân loại điểm dữ liệu mới như hầu hết các mô hình Machine Learning khác. SVM lại dựa vào các điểm dữ liệu dễ gây nhầm lẫn nhất của các lớp để phân loại

2.2.2.2. Softmax Regression:

Hàm softmax tính toán xác suất xảy ra của một sự kiện. Nói một cách khái quát, hàm softmax sẽ tính khả năng xuất hiện của một class trong tổng số tất cả các class có thể xuất hiện. Sau đó, xác suất này sẽ được sử dụng để xác định class mục tiêu cho các input.

Nguyên lý đằng sau hàm Softmax khá đơn giản. Với một vài số cho trước

- Tính hàm lũy thừa số e, với số mũ là những số đã cho
- Tính tổng các lũy thừa đó. Đó sẽ là mẫu số.

- Sử dụng lũy thừa của mỗi số là tử số
- Xác suất sẽ là tử số/mẫu số

Viết ngắn gọn lại, ta được công thức hàm Softmax:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_i}}$$

Ưu điểm của Softmax:

- Hàm softmax là tối ưu khi tính toán xác suất tối đa trong tham số mô hình.
- Tính chất của hàm softmax khiến hàm phù hợp với sự thông dịch xác suất, rất hữu ích trong Machine Learning (Học máy).
- Chuẩn hóa softmax là một cách để giảm thiểu ảnh hưởng của những giá trị cực trị hay dữ liệu ngoại lai trong dữ liệu mà không phải chỉnh sửa dữ liệu ban đầu.

2.2.2.3. VisoBERT

Là một mô hình transformer được giới thiệu vào năm 2023 bởi nhóm nghiên cứu sinh viên của trường Đại học Công Nghệ Thông Tin. Mô hình được huấn luyện trên dữ liệu thu thập trên các nền tảng mạng xã hội lớn như FaceBook, YouTube, TikiTok từ các người dùng sử dụng ngôn ngữ tiếng Việt. Model transformer có thể sử lý các task cần xử lý trên các bình luận ở nền tảng google maps.

Ưu điểm:

- Mô hình được huấn luyện trên dữ liệu tiếng việt
- Cấu trúc, đặc điểm dữ liệu gần như tương tự với dữ liệu được chúng tôi thu thập trên google maps (**viết tắt, icon**)

2.2.2.4. TwHIN-BERT

Là mô hình transformer đa ngôn ngữ được giới thiệu vào năm 2023 bởi một nhóm nghiên cứu ở Mỹ. Mô hình được huấn luyện dựa trên dữ liệu được thu thập từ nền tảng mạng xã hội twitter (7 tỉ bài đăng) bao gồm hơn 100 ngôn ngữ khác nhau. TwHIN-BERT khác biệt so với các mô hình ngôn ngữ tiền huấn luyện trước đó vì nó được đào tạo không chỉ bằng cách tự giám sát dựa trên văn bản mà còn với một mục

tiêu xã hội dựa trên các tương tác xã hội phong phú trong mạng thông tin không đồng nhất Twitter (TwHIN)

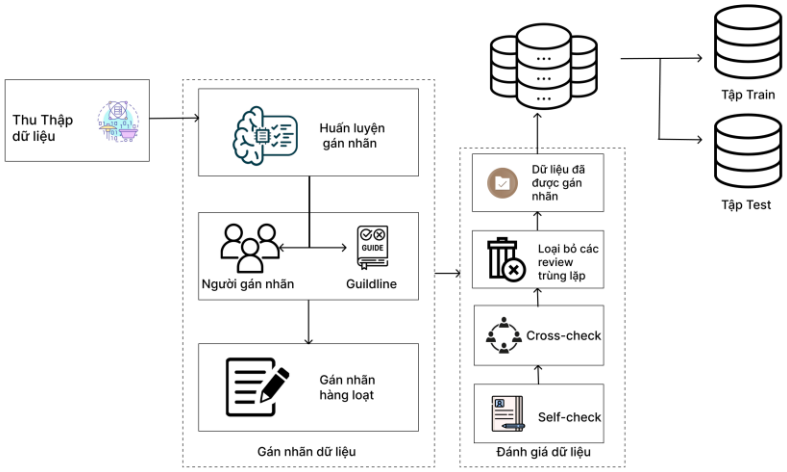
Ưu điểm:

- Mô hình được huấn luyện trên dữ liệu đa ngôn ngữ được thu thập từ nền tảng mạng xã hội
- Vì là dữ liệu được thu thập trên mạng xã hội vì vậy cấu trúc dữ liệu có thể giống với cấu trúc dữ liệu của chúng tôi (xử lý tiếng Anh và tiếng Việt)

3. THU THẬP VÀ XỬ LÝ DỮ LIỆU

3.2. Quy trình thu thập và xử lý dữ liệu

Sau đây là biểu đồ miêu tả quy trình thu thập và sử lý dữ liệu của chúng tôi, đối với bài toán phân loại review của các bệnh viện trên Google Maps.

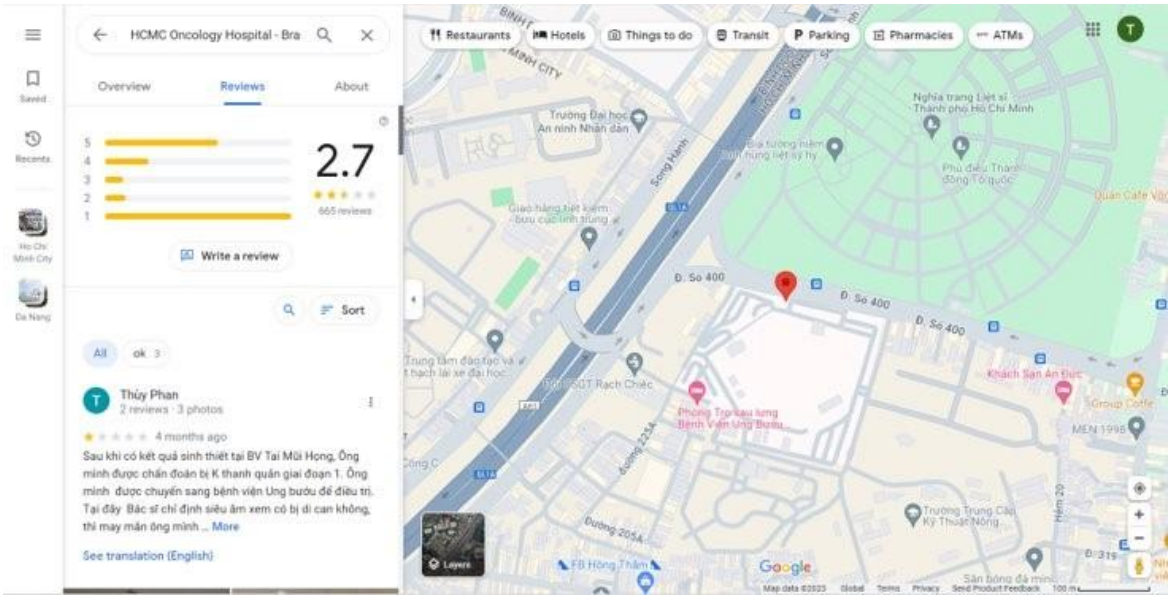


Hình 3.1 Quy trình thu thập và xử lý dữ liệu

3.3. Nguồn và phương pháp thu thập dữ liệu:

Các reviews của những bệnh viện được lấy từ google maps.

Data được crawl từ những bệnh viện ở các tỉnh, thành phố lớn như Đà Nẵng , Cần Thơ, TPHCM,... bao gồm nhiều loại bệnh viện: bệnh viện quốc tế, bệnh viện công và bệnh viện tư nhân. 3 thành phố trên là nơi tập trung nhiều bệnh viện lớn và được người dân ưu tiên tới thăm khám.



Hình 3.2 Hình ảnh miêu tả Google Maps

3.4. Quy cách gán nhãn

Quy tắc chung :

- Mỗi bình luận sẽ có 4 nhãn: positive, negative, neutral, other
- Nhãn positive: thể hiện sự đánh giá tốt về bệnh viện

- Nhãn negative: thể hiện sự đánh giá không tốt về bệnh viện
- Nhãn neutral: thể hiện sự đánh giá tốt lẫn không tốt về bệnh viện
- Nhãn other: không thể hiện sự đánh giá tốt hoặc không tốt về bệnh viện
- Nếu có sự nhập nhằng giữa các nhãn, sử dụng thông tin về rating để gán nhãn

Nhãn: POSITIVE

- Cơ sở vật chất, dịch vụ: bệnh viện không bị quá tải, đáp ứng nhu cầu của người bệnh

  Ví dụ : 'tiện nghi, khang trang, hiện đại.'
- Nhân viên: nhân viên có thái độ thân thiện với bệnh nhân'

  Ví dụ : 'nhân viên nhiệt tình'
- Giá cả: phù hợp với chất lượng dịch vụ

Nhãn: NEGATIVE

- Cơ sở vật chất, dịch vụ: quá tải, chật chội

  Ví dụ : 'Bv này đông đúc và khá chật. '
- Nhân viên: nhân viên không thân thiện

  Ví dụ : 'Bệnh nhân đau, sốt cao ói mửa, gọi điều dưỡng còn bị la mắng và không chịu gọi bác sĩ, khi chịu gọi thì củng phải …'
- Giá cả: chặt chém, không phù hợp với chất lượng dịch vụ

  Ví dụ : 'Giá giữ xe trái quy định của Nhà nước.'

Nhãn: NEUTRAL

- Chứa cả cảm xúc tiêu cực và tích cực về bệnh viện:

  Ví dụ : "Vị trí mà mặt bằng hơi nhỏ. Có chổ để xe máy thu phí. Xe oto đậu ngoài đường có phí. Đi khám buổi sáng hơi đông. Buổi chiều thưa hơn. Ngày thường khám đến 17h thôi, không có khám ngoài giờ. Trang thiết bị tạm ổn. Bs, y tá ok. Phí khám hơi cao.."

Nhãn: OTHER

- Không thể hiện rõ thái độ phàn nàn hoặc đánh giá tốt về bệnh viện

  Ví dụ : 'Tôi có một em trai bị ruọt thừa.Nửa đêm nhập viên đau đến tái cả mặt mài'

RATING :

Rating	**Nội dung**
1 sao	Rất tệ

2 sao	Tệ
3 sao	Bình thường
4 sao	Tốt
5 sao	Rất tốt

Bảng 1 Mô tả nội dung của rating

Sử dụng rating khi có sự nhập nhằng giữa nhãn positive và negative

Ví dụ : **'Bệnh viện quốc tế lớn, khang trang có khoa cấp cứu 24/7. Tuy nhiên giá cũng rất cao. Phòng bệnh 4tr/ng. Khám cấp cứu 1.2tr/lần',** rating **4 sao**

Giá rất cao sẽ mang 2 nghĩa: (1) giá rất cao nhưng chất lượng dịch vụ không tốt (mang yếu tố tiêu cực), (2) giá rất cao và chất lượng dịch vụ tương xứng (mang yếu tố tích cực). Do đó sẽ dùng rating và dựa vào bảng mô tả nội dung sẽ quyết định review này là nhãn **positive**

3.5. Đặc điểm dữ liệu

Do tính chất của một review nhanh gọn trên nền tang google maps, bài bình luận có tính chất gần như tương tự với một bình luận trên nền tang mạng xã hội khác như teencode, viết tắt, không dấu,… Sau đây là một số đặc trưng dữ liệu mà chúng tôi phân tích:

3.5.2. Có tiếng Anh và tiếng Việt

- **24Mar2020: Doctors have attitude badly. They dont respect the old-man. We hope they will read comments and have adapted**.\n1- Chỗ nghỉ: Bv có 1 khu vực gần Khoa Cấp Cứu có cây xanh, tán mát và hàng ghế

3.5.3. Viết tắt, dùng teencode

- Nếu **BV** được đánh giá thì tôi không cho sao nào lun á chứ. **Bv** gì mà **NV** thì không **bik** cười, mặt như đưa đám , quy trình làm việc thì lâu,lấy **KQ** thôi mà người thì chỉ lên tầng 3, người chỉ lên tầng 4. Không có thống nhất thông tin

- **Không có ý nghĩa, bình luận spam**

- Oke oke 123 123
- 12 :') :') :') :') :') :') :') :')

- **Từ ngữ địa phương**

- Không ưng được cái **chi** ở bệnh viện này cả, gọi bác sĩ, y tá miết để hỏi khoa nội ở **mô** mà không ai đoái hoài **chi** hết.

- Thái độ của BS & ĐD vô cùng khó chịu, nói chuyện cọc cằn. Nói chuyện với NGƯỜI ĐI KHÁM BỆNH còn hơn gì nữa. Người ta dô khám bệnh thì thắc mắc là chuyện bình thường, trả lời người ta như **mắc nợ từ kiếp nào**. :)))
- Dịch vụ KCB như thế nào thì chưa biết, **phí mãi lộ của quan triều đình** thì, biển báo phí gởi xe qua đêm 5.000 VNĐ, gởi từ 2hơi cao nha0h đến 6h là 8.000 VNĐ, được quan triều đình thông báo từ 5h đến 6h là 3.000 VNĐ
- Chưa hài lòng lắm khi cách ứng sử còn quá kém thiếu tôn trong bên nhân ,chờ đợi đến lược thì số **nhảy loạn cào cào** lên .thắc mắt vấn đề thì hỏi nhân viên nhân viên bệnh viện đáp lạy bằng ánh mắt rất khó chiệu .....
- Nếu mà có 0 sao t sẽ k cho 1 sao, bởi 1 sao cũng k xứng đáng,Khám dịch vụ mà rất tệ hại, thua cả khám bảo hiểm, vừa lâu mà nv cứ **cà rề cà rề**, hok có khám gì mà **chờ dài cổ**, vậy khám dich vụ lm con mẹ gì, ngta đồng ý bỏ tiền, bỏ chi phí cao …
- Bệnh viện từ bác sĩ tới thực tập sinh nói chuyện rất khó nghe 1 vài người thì rất vui vẻ còn 1 số người nói chuyện rất là láo đều dưỡng thì coi thường bệnh nhân **nói trên đầu trên cổ mình** dô đây mình có trả tiền chứ không phải ở không bệnh nhân hỏi thì la làng TÔI MONG BỆNH VIỆN NÊN XEM LẠI câu LƯƠNG Y NHƯ TỪ MẪU

3.6. Huấn luyện annotator:

Chúng tôi sử dụng độ đo Fleiss Kappa để đo độ đồng thuận giữa bốn annotator là các thành viên trong nhóm của chúng tôi. Chúng tôi chia mỗi set là 30 câu reviews, khi độ đồng thuận lớn hơn 80 thì dừng huấn luyện annotator và tiến hành gán nhãn đồng loạt. Sau khi tiến hành thực hiện huấn luyện annotator đến set thứ 3 thì chúng tôi nhận được kết quả là cả 2 set 2 và 3 đều lớn hơn 80, vì vậy chúng tôi tiến hành hoàn thiện guidline và gán nhãn đồng loạt.

	Set 1	Set 2	Set 3
Fleiss Kappa	68.03	85.91	85.33

Bảng 2 Độ đồng thuận giữa các annotators

3.7. Đặt trưng về ngôn ngữ của dữ liệu

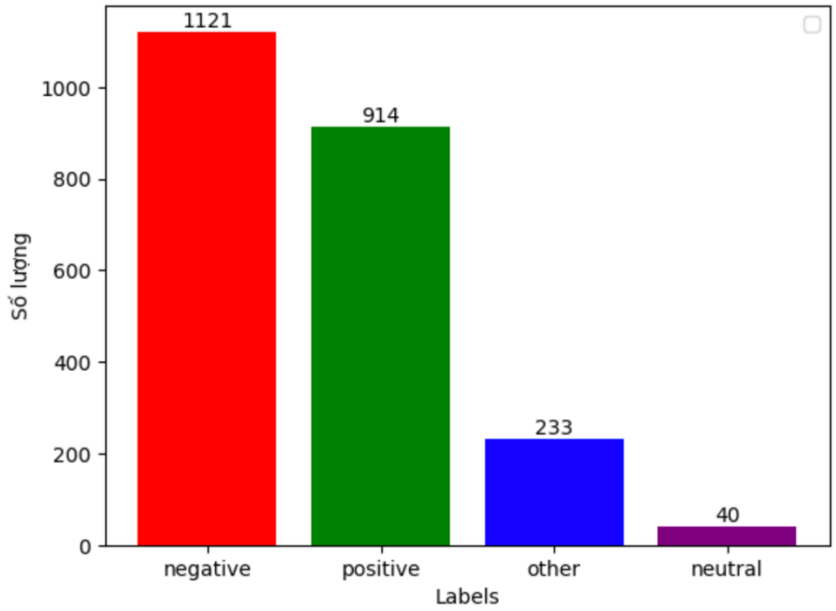
3.8. Thống kê dữ liệu

3.7.1. Thống kê dữ liệu trên toàn tập dữ liệu

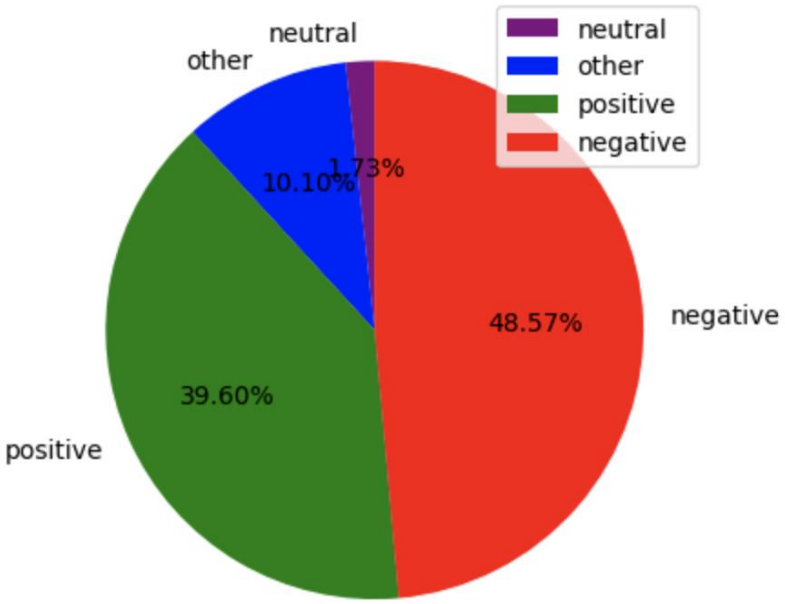
3.7.1.1. Thống kê số lượng

STT	Nhãn	Số Lượng	Tỉ lệ
1	Negative	1121	48.57%
2	Positive	914	39.60%
3	Neutral	40	10.10%
4	Other	233	1.73%
Tổng		693	100%

Bảng 3 Bảng thống kê dữ liệu trên toàn tập dữ liệu



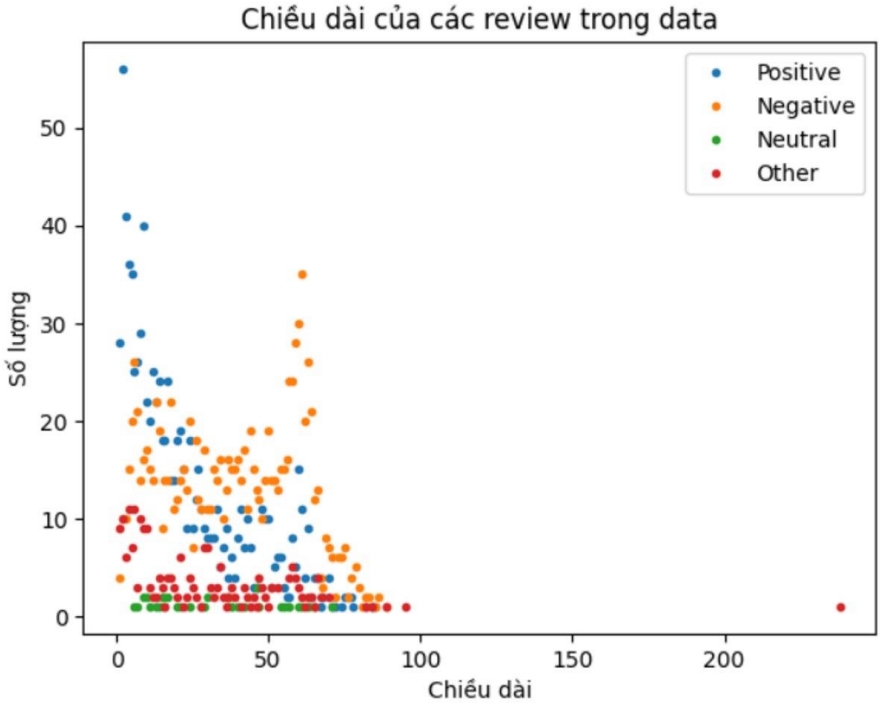
Hình 3.3 Biểu đồ cột thống kê số lượng từng nhãn trên toàn tập dữ liệu



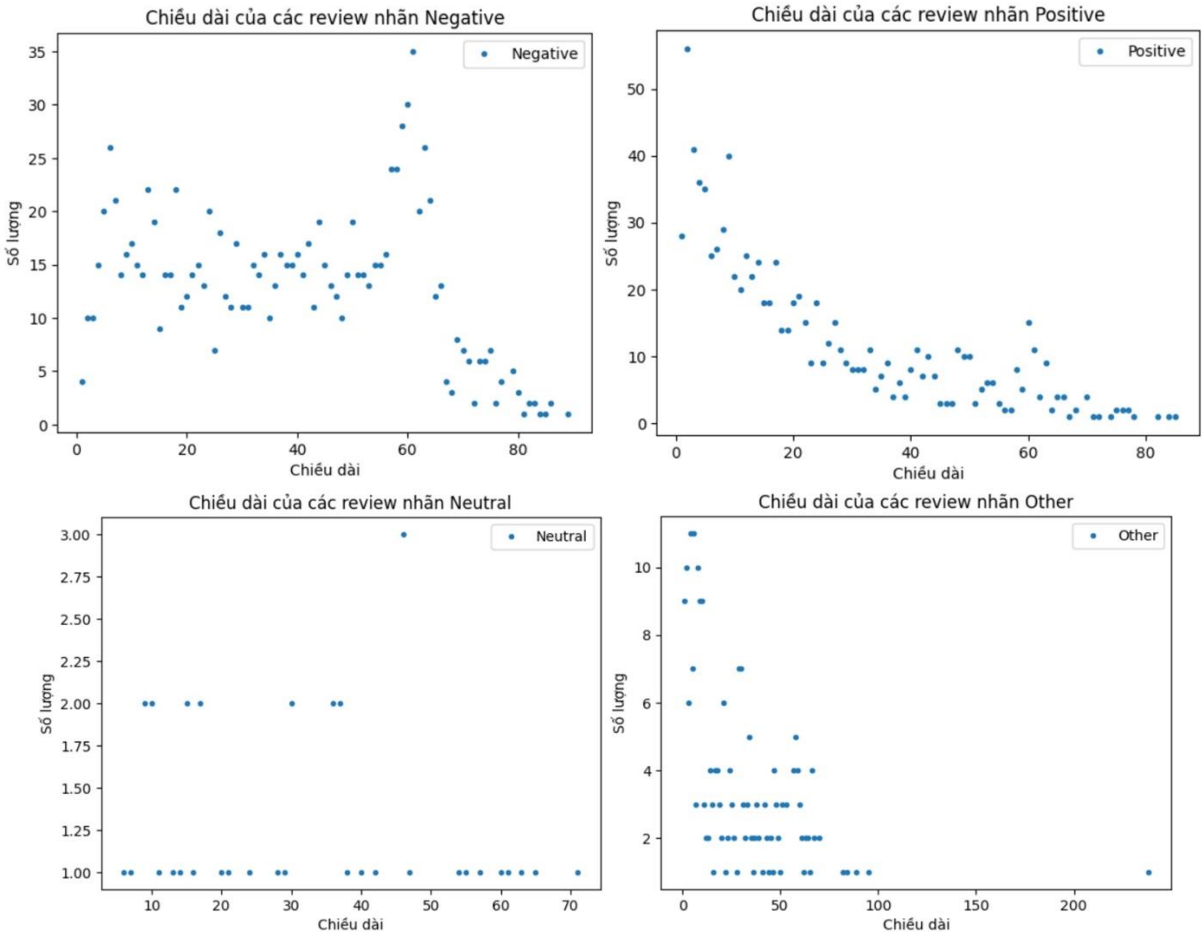
Hình 3.4 Biểu đồ tròn thống kê số lượng từng nhãn trên toàn tập dữ liệu

Từ dữ liệu dã được gắn nhãn, có thể thấy số lượng nhãn negative chiếm nhiều nhất (1121 mẫu với tỉ lệ là 48.57%) và nhãn neutral chiếm tỉ lệ ít nhất (40 mẫu với tỉ lệ 1.73%). Nhãn positive và negative chiếm 88.17%, do đó, phần lớn người dân có xu hướng viết review chỉ mang tính chất tích cực và tiêu cực, ít mang tính trung hòa. và không liên quan đến cảm xúc khi thăm khám tại bệnh viện. Vì vậy, dữ liệu này không cân bằng giữa các nhãn, và các mô hình phân loại sẽ có khả năng cao dự đoán không chính xác các review thuộc nhãn neutral và other

3.7.1.2. Thống kê chiều dài của các review trên tập dữ liệu



Hình 3.5 Thống kê chiều dài của các review trên toàn tập dữ liệu



Hình 3.6 Chiều dài của các review theo từng nhãn

Độ dài của review được tính bằng số lượng từ, các kí hiệu dấu, số, đặc biệt trong câu. Từ hình 3.3, nhìn chung đa số các review đều có ít hơn 100 từ, chỉ có 1 review nhiều hơn 200 từ. Từ hình 3.4, có thể thấy rằng số lượng review ngắn nhiều hơn số lượng review dài ở nhãn positive, và chiều dài của review tăng dần thì số lượng review có chiều dài đó giảm dần. Còn với nhãn negative, số lượng nhiều tập trung vào review có chiều dài nhỏ hơn 70, và số lượng review có chiều dài lớn hơn 70 có xu hướng giảm dần. Với nhãn neutral, chỉ có 1 hoặc 2 review có chiều dài giống nhau. Trong nhãn other, số lượng các review giống nhau về chiều dài có xu hướng giảm khi chiều dài tăng lên. Đặt biệt là có 1 review có độ dài lớn hơn 200 trong nhãn other.

3.7.1.3. Tần suất xuất hiện của từ

Từ	Số lượng	Tỉ lệ
bệnh	344	1.68%
viện	319	1.56%
và	279	1.36%
rất	269	1.32%
khám	232	1.13%

Bảng 4 Các từ có tần suất cao nhất trong nhãn positive

Từ	Số lượng	Tỉ lệ
bệnh	750	1.76%
khám	630	1.48%
không	499	1.17%
có	462	1.08%
viện	456	1.07%

Bảng 5Các từ có tần suất cao nhất trong nhãn negative

Từ	Số lượng	Tỉ lệ

có	22	1.70%
nhưng	20	1.55%
bệnh	19	1.47%
viện	17	1.32%
rất	15	1.16%

Bảng 6 Các từ có tần suất cao nhất trong nhãn neutral

Từ	Số lượng	Tỉ lệ
viện	110	1.73%
bệnh	96	1.51%
có	89	1.40%
khám	85	1.34%
Bệnh	73	1.15%

Bảng 7 Các từ có tần suất xuất hiện cao nhất trong nhãn other

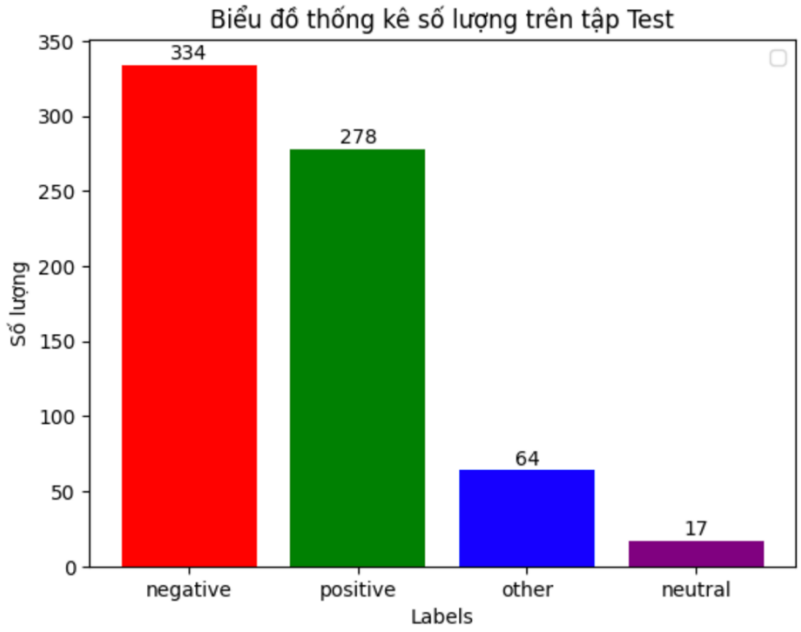
Nhìn chung, các từ xuất hiện nhiều nhất trong cả bốn nhãn rất nhiều là "bệnh", "viện", "có", "khám", do đó giữa các nhãn positive, neutral, other không bị thiêng lệch khi được dự đoán bởi mô hình phân loại. Khác với các nhãn positive, neutral, other, từ "không" xuất hiện rất nhiều trong nhãn negative và điều nãy sẽ khiến mô hình có xu hướng phân loại nhãn "negative" khi review xuất hiện từ "không".

3.7.2. Thống kê dữ liệu trên tập Test

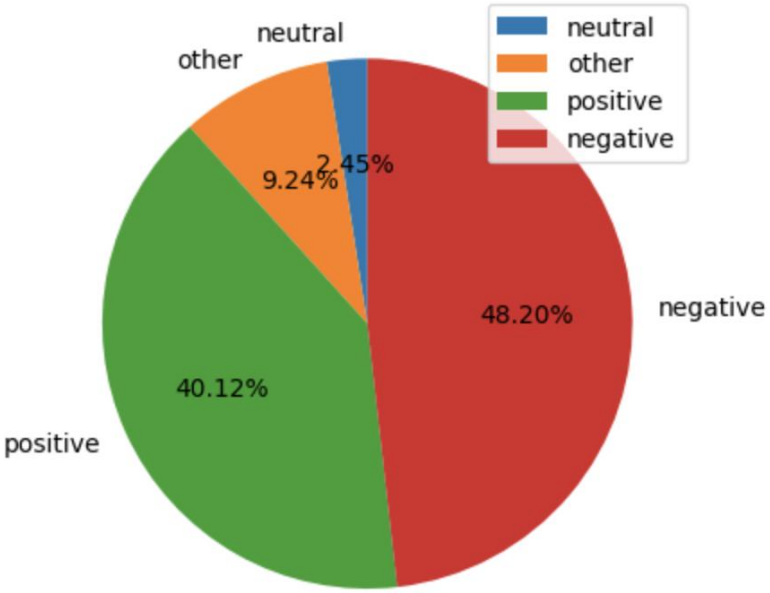
3.7.2.1. Thống kê số lượng nhãn trên tập Test

STT	Nhãn	Số Lượng	Tỉ lệ
1	Negative	334	48.20%
2	Positive	278	40.12%
3	Neutral	17	2.45%
4	Other	64	9.23%
Tổng		693	100%

Bảng 8 Bảng thống kê dữ liệu trên tập Test



Hình 3.7 Biểu đồ cột thống kê số lượng từng nhãn trên tập Test



Hình 3.8 Biểu đồ tròn thống kê tỉ lệ từng nhãn trên tập Test

Phân phối nhãn của dữ liệu test gần giống như dữ liệu ban đầu, số lượng nhãn negative chiếm nhiều nhất với 48.20% và số lượng nhãn neutral ít nhất với 2.45%.

3.7.2.2. Thống kê dữ liệu tiếng Anh và tiếng Việt trên tập Test

STT	Ngôn ngữ	Tỉ lệ
1	Tiếng Việt	93.4%
2	Tiếng Anh	6.6%

Tổng	100%

Bảng 9 Bảng thống kê tỉ lệ Tiếng Anh và Tiếng Việt

Về phần ngô ngữ, số lượng review là tiếng Anh chiếm 6.6%, số lượng review là tiếng Việt chiếm đa số với 93.4%.

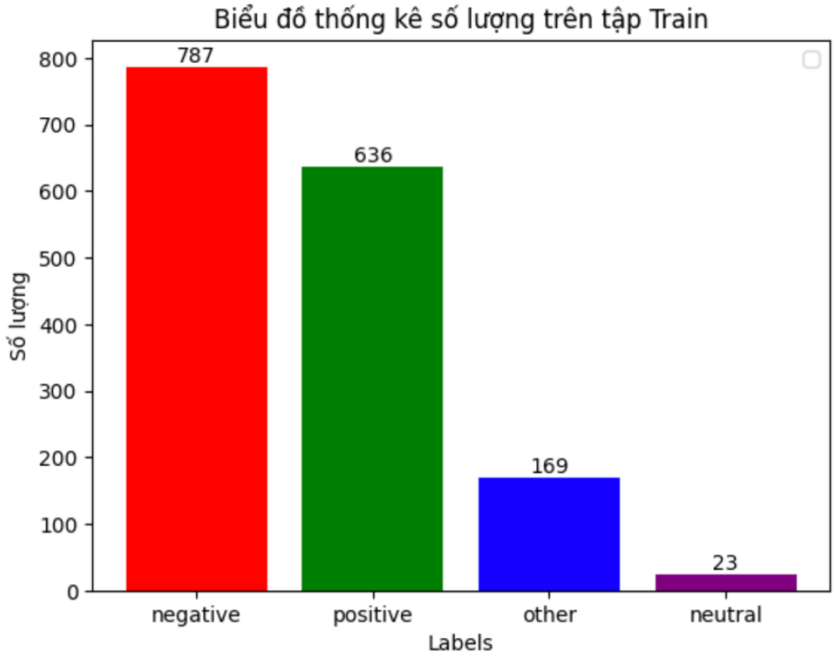
Hình 3.9 Biểu đồ tròn phân chia tỉ lệ tiếng Anh và tiếng Việt

3.7.3. Thống kê dữ liệu trên tập Train

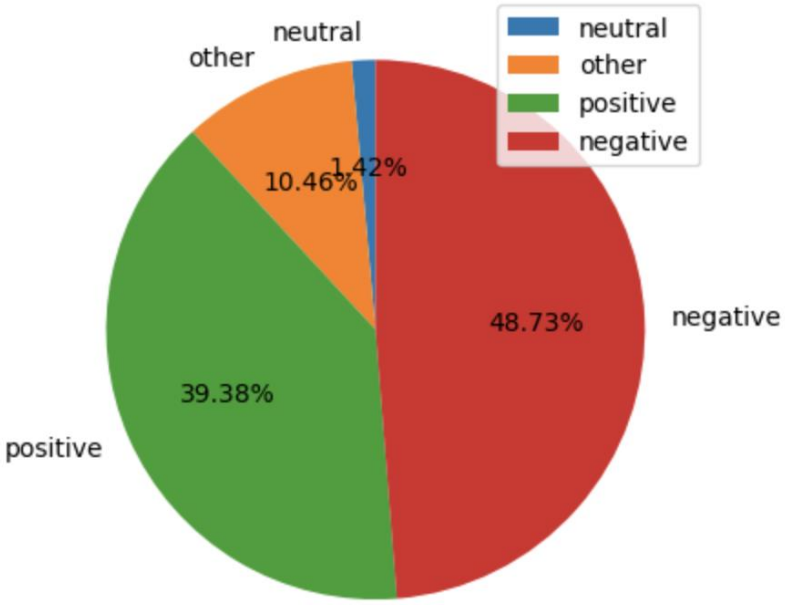
3.7.3.1. Thống kê số lượng nhãn trên tập Test

STT	Nhãn	Số Lượng	Tỉ lệ
1	Negative	787	48.73%
2	Positive	636	39.38%
3	Neutral	23	1.42%
4	Other	169	10.46%
Tổng		1615	

Bảng 10 Bảng thống kê trên tập train



Hình 3.10 Biểu đồ cột thống kê số lượng từng nhãn trên tập Train



Hình 3.11 Biểu đồ tròn thống kê số lượng từng nhãn trên tập Train

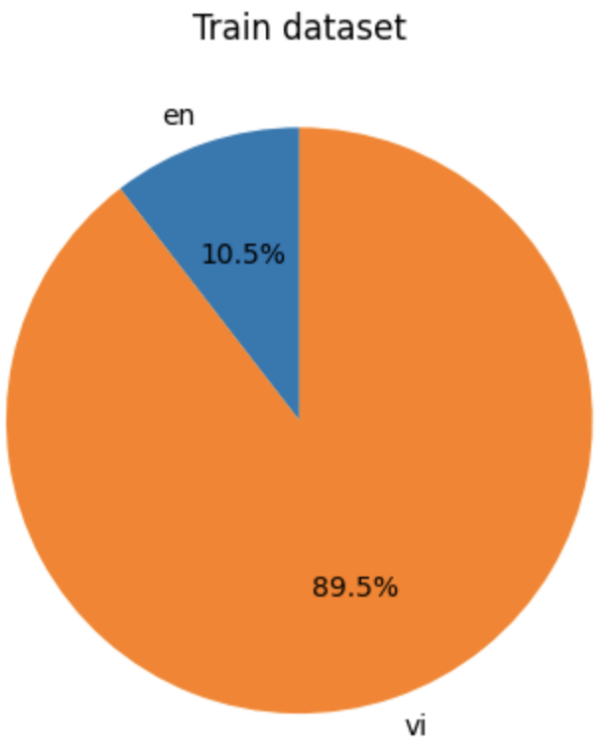
Phân phối nhãn của dữ liệu train gần giống như dữ liệu test, số lượng nhãn negative chiếm nhiều nhất với 48.73% và số lượng nhãn neutral ít nhất với 1.42%.

3.7.3.2. Thống kê dữ liệu tiếng Anh và tiếng Việt trên tập train

STT	Ngôn ngữ	Tỉ lệ

1	Tiếng Việt	93.4%
2	Tiếng Anh	6.6%
Tổng		100%

Bảng 11 Bảng thống kê tỉ lệ tiếng Anh và tiếng Việt trên tập Train



Hình 3.12 Biểu đồ tròn phân chia tỉ lệ tiếng Anh và tiếng Việt trên tập Train

Về phần ngô ngữ, số lượng review là tiếng Anh chiếm 10.5%, số lượng review là tiếng Việt chiếm đa số với 89.5%.

4. ĐÁNH GIÁ HIỆU XUẤT MÔ HÌNH

Chúng tôi sử dụng các số Accu chỉ racy, F1-score (marco) trên các mô hình mà chúng tôi sử dụng để thực nghiệm là TF-IDF SVM, Softmax, và hai model transformer là VisoBERT và TwHIN-BERT, chúng tôi thu được các số liệu sau

4.7. Accuracy, F1-score (marco)

Model	**Accuracy**	**F1 score (Macro)**
TF-IDF + SVM	0.80	0.49
TF-IDF + Softmax Regression	0.80	0.50
TwHIN-BERT	0.82	0.58
ViSoBERT	**0.83**	**0.60**

Bảng 12 Bảng điểm Accuracy và f1-score trên từng mô hình

Nhìn chung các mô hình đạt độ chính xác (accuracy) tốt, mô hình SVM và Softmax Regression đạt độ chính xác thấp nhất là 0.8 và mô hình ViSoBERT đạt độ chính xác cao nhất là 0.83. Với độ đo F1 score (macro), SVM đạt F1 thấp nhất (0.49) và mô hình ViSoBERT đạt F1 cao nhất (0.60). Điều này chứng tỏ mô hình pre-trained transformer tốt hơn so với mô hình máy học truyền thống (SVM và Softmax Regression).

Nhãn	**Số lượng**	**SVM**	**Softmax Regression**	**TwHIN-BERT**	**ViSoBERT**
Negative	334	0.85	0.86	0.87	**0.88**
Neutral	17	0.00	0.00	0.00	**0.09**
Positive	278	0.83	0.83	0.86	**0.86**
Other	64	0.28	0.30	0.58	**0.59**

Bảng 13 F1-score của các mô hình theo từng nhãn

Phân tích theo từng nhãn, có thể thấy nhãn positive và negative chiếm nhiều trong dữ liệu train, do đó nhãn positive và negative đạt F1 score cao nhất trên tập test. Nhãn neutral xuất hiện rất ít (23 reviews trên tập train và 17 reviews trên tập test), do đó mô hình SVM, Softmax Regression và TwHIN-BERT dự đoán sai toàn bộ các review thuộc nhãn neutral trên tập test, trong khi ViSoBERT đạt F1 là 0.09. Điều này là do các review thuộc nhãn neutral chứa cả 2 yếu tố tích cực và tiêu cực, và khiến cho các model gặp khó khi phân loại.

4.7.1. Ảnh hưởng của ngôn ngữ

	Số lượng	**SVM**	**Softmax Regression**	**TwHIN-BERT**	**ViSoBERT**
Tiếng Việt	647	0.50	0.50	0.59	**0.62**
Tiếng Anh	46	0.33	0.36	**0.42**	0.31

Bảng 14 F1-score (marco) của mô hình transformer trên dữ liệu tiếng Anh và tiếng Việt

Phân chia theo ngôn ngữ, mô hình SVM và Softmax Regression đạt F1 score thấp hơn so với các mô hình transformer trên cả tiếng Việt và tiếng Anh. Với mô hình TwHIN-BERT, F1 score trên tiếng anh cao hơn so với ViSoBERT, vì TwHIN-BERT là mô hình đa ngôn ngữ (bao gồm tiếng anh) nên sẽ xử lý tiếng Anh tốt hơn. Với mô hình ViSoBERT, mô hình đạt F1 cao hơn so với TwHIN-BERT trên tiếng Việt, vì mô hình

được huấn luyện trên dữ liệu mạng xã hội tiếng Việt, đo đó xử lý tốt hơn một số đặt thù của dữ liệu như viết tắt, viết sai chính tả, teencode.

4.7.2. Ảnh hưởng của độ dài review

Độ dài	Số lượng	TwHIN-BERT	ViSoBERT
1 – 10	204	0.57	**0.72**
11 – 40	277	**0.57**	**0.57**
41 – 80	218	0.54	**0.55**
> 80	5	**1.0**	0.55

Bảng 15 13 F1-score (marco) của mô hình transformer theo độ dài của bài review

Có thể thấy, với review có độ dài thấp thì mô hình ViSoBERT xử lý tốt hơn so với TwiHINB-BERT, cụ thể với review từ 1 đến 10 chữ thì ViSoBERT đạt F1 là 0.72, trong khi TwHIN-BERT là 0.57. Tương tự với các review từ độ dài từ 11 đến 40 và từ 41 đến 80, ViSoBERT đạt cao hơn TwHIN-BERT với F1 lần lượt là 0.57 và 0.55. Khi độ dài review lớn hơn 80, TwHIN-BERT đạt F1 cao nhất là 1.0, trong khi ViSoBERT đạt 0.55.

5. KẾT LUẬN

5.7. Ưu điểm

- Đây là một đề tài thiết thực, có thể đóng góp cho nghiên cứu ngôn ngữ tự nhiên
- Các mô hình được lựa chọn đều đưa ra được kết quả khá khả quan.

5.8. Nhược điểm

- Dữ liệu thu thập được chưa khái quát hết các tỉnh thành trên nước Việt Nam, vì thế mà chưa bao quát được cái từ ngữ địa phương, vùng miền.
- Chưa thực nghiệm trên nhiều mô hình khác nhau.

5.9. Hướng phát triển

- Thu thập thêm dữ liệu từ các tỉnh thành trên đất nước, từ đó có thể tăng vốn từ cho mô hình.
- Thực nghiệm trên nhiều mô hình khác nhau và các phương pháp thực nghiệm khác.
- Thu thập thêm hình ảnh từ các reviẹw, nhằm có thể nghiên cứu phát triển thành multi-modal.

- Có thể phát triển thành bài toán khuyến nghị, cho người có nhu cầu khám chữa bệnh.

TÀI LIỆU THAM KHẢO

[1] Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, Kiet Van Nguyen. ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing,2023

[2] Aswathisa, " Support Vector Machine (SVM) Algorithm," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[3] Wenhao Lu, Jian Jiao, Ruofei Zhang. ViSoBERT: TwinBERT: Distilling Knowledge to Twin-Structured BERT Models for Efficient Retrieval,2020

[4] Riturajsaha, " Understanding TF-IDF (Term Frequency-Inverse Document Frequency)," GeeksforGeeks, [Online]. Available: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

[5] Baohua Su,Jun Peng. Sentiment Analysis of Comment Texts on Online Courses Based on Hierarchical Attention Mechanism,2023

[6] Abdulrahman Alrumaih, Ruaa Alsabah, Hiba J Aleqabie, Ahmed Yaseen Mjhool,Ali Al-Sabbagh, and James Baldwin. Analyzing User Behavior and Sentimental in Computer Mediated Communication,2020

[7] Ritika Singh,Ayushka Tiwari. YOUTUBE COMMENTS SENTIMENT ANALYSIS,2021

[8] Rawan Fahad Alhujaili, Wael M.S. Yafooz. Sentiment Analysis for Youtube Videos with User Comments: Review,2021

PHỤ LỤC

Link source code: https://github.com/TuyetMinh14/DS102-Classification_review_on_googlemaps.git