ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN

PHÂN TÍCH CẢM XÚC ĐÁNH GIÁ SẢN PHẨM THƯƠNG MẠI ĐIỆN TỬ

Sentiment Analysis For E-Commerce Product Review

Giảng viên hướng dẫn: **TS. NGUYỄN TRỌNG CHỈNH**
ThS. NGUYỄN ĐỨC VŨ

Sinh viên thực hiện:

LƯƠNG VĨNH HƯNG	**21522116**
LÝ QUỐC HƯNG	**21522117**
NGUYỄN SỸ HÙNG	**21522119**

Lớp: **XỬ LÝ NGÔN NGỮ TỰ NHIÊN - CS221.O11**

NHẬN XÉT CỦA GIẢNG VIÊN

TP.HCM, Ngày … Tháng … Năm …

Người Nhận Xét

(Ký tên)

MỤC LỤC

MỞ ĐẦU

Thương mại điện tử (e-commerce) là một hình thức kinh doanh mà giao dịch mua bán hàng hóa và dịch vụ được thực hiện qua mạng Internet. Nó giúp người tiêu dùng và doanh nghiệp thực hiện các giao dịch mua bán, thanh toán trực tuyến và trao đổi thông tin sản phẩm mà không cần sự giao tiếp trực tiếp.

Hiện nay, mua hàng trực tuyến là hình thức mua sắm đã trở nên phổ biến khi mà chỉ cần 1 chiếc smartphone có kết nối với internet là đã có thể tham gia vào thị trường thương mại điện tử rộng lớn 1 cách dễ dàng.

Các trang thương mại điện tử thường cung cấp một nền tảng trực tuyến cho việc chọn lựa, xem thông tin sản phẩm, đặt hàng, và thanh toán bằng các phương tiện thanh toán. Nhưng đặc biệt nhất là dịch vụ " Rating & Reviews" cho phép khách hàng có thể để lại nhận xét và đánh giá của mình đối với 1 sản phẩm của 1 cửa hàng nào đó, dịch vụ này giúp cho những khách hàng sau này có cái nhìn khách quan về sản phẩm cũng như giúp cho cửa hàng cải thiện và nâng cao chất lượng sản phẩm, dịch vụ của mình để mang lại cho khách hàng trải nghiệm mua sắm tốt nhất.

Việc phân tích cảm xúc trong các bài nhận xét của khách hàng là một phần quan trọng, nó không chỉ cung cấp thông tin quý báu về ý kiến của khách hàng mà còn góp phần trực tiếp vào quyết định kinh doanh, cải thiện sản phẩm và nâng cao sự hài lòng của khách hàng

CHƯƠNG 1: TỔNG QUAN ĐỒ ÁN

1.1. Lý do chọn đồ án

Trong thị trường truyền thống người ta thường nói rằng "Khách hàng là thượng đế". Đúng vậy, khi mà khách hàng là mục tiêu của mọi doanh nghiệp, mọi hoạt động của doanh nghiệp đều xoay quanh khách hàng. Doanh nghiệp nào không có khách hàng thì doanh nghiệp đó chắc chắn phá sản. Đối với thị trường thương mại điện tử cũng vậy, khách hàng không chỉ là "Thượng đế" mà còn là một nguồn cung cấp thông tin quan trọng khi mà dịch vụ "Rating & Reviews" cung cấp ý kiến và đánh giá đối với sản phẩm, dịch vụ của doanh nghiệp một cách khách quan, nhanh chóng và dễ dàng.

Chúng tôi lựa chọn đề tài này từ nhiệm vụ quan trọng trong mỗi doanh nghiệp cần phải có. Phân tích cảm xúc trong bài đánh giá của khách hàng, việc này tập trung vào các đánh giá của khách hàng đối với sản phẩm để từ đó xác định cảm xúc, thái độ của khách từ dữ liệu mà chúng tôi thu thập được.

1.2. Đặt vấn đề

Bài toán sentiment analysis (phân tích cảm xúc) là một nhiệm vụ của trí tuệ nhân tạo nhằm đánh giá và phân loại cảm xúc trong văn bản. Đối với bài toán này, đầu vào là một đoạn văn bản (chẳng hạn như một đánh giá sản phẩm, bình luận trên mạng xã hội, hoặc tin tức), và đầu ra là một phân loại về cảm xúc của văn bản đó, thường là một trong các nhóm sau:

- Tích cực (Positive - 1): Nếu văn bản thể hiện sự hài lòng, tán dương, hoặc tích cực.
- Tiêu cực (Negative - 0): Nếu văn bản thể hiện sự không hài lòng, phê phán, hoặc tiêu cực.
- Trung tính (Neutral - 2): Nếu văn bản không chứa cảm xúc rõ ràng hoặc không thể xác định được cảm xúc nào.

Mục tiêu của bài toán sentiment analysis là để tự động phân loại văn bản vào các nhóm cảm xúc tương ứng, giúp tổng hợp và hiểu được ý kiến và tư duy của người viết. Điều này có ứng dụng rộng rãi trong lĩnh vực đánh giá sản phẩm,

quản lý dư luận trực tuyến, và nhiều ngữ cảnh khác liên quan đến phân tích ý kiến cộng đồng.

1.3. Mục tiêu và phạm vi đồ án

1.2.1. Mục tiêu chung

Xây dựng mô hình: xây dựng mô hình có khả năng phân loại cảm xúc của các nhận xét và đánh giá của khách hàng có thể là tích cực (positive), tiêu cực (negative) hay neutral(trung lập).

Tối ưu hóa hiệu suất: nghiên cứu và áp dụng các kỹ thuật tối ưu để tối ưu hóa hiệu suất của mô hình.

1.2.2. Phạm vi đồ án

Loại dữ liệu: Tập trung vào phân tích cảm xúc của các câu nhận xét và đánh giá thành các nhãn 0 (negative), 1 (positive) và 2 *neutral).

Khía cạnh kỹ thuật: Bao gồm quá trình tiền xử lý dữ liệu, trích xuất đặc trưng, lựa chọn và huấn luyện mô hình, đánh giá hiệu xuất, tối ưu hóa và triển khai mô hình

Phương pháp nghiên cứu: sử dụng các phương pháp vector hóa dữ liệu, các mô hình máy học phổ biến.

CHƯƠNG 2: NGỮ LIỆU

2.1. Thu thập dữ liệu

Dữ liệu được chúng tôi thu thập từ 2 sàn Lazada và Shopee, là 2 sàn thương mại điện tử lớn và uy tín nhất Việt Nam.

Dữ liệu được sử dụng gồm có 2 cột:

- cột 'review': là nhận xét của khách hàng đối với sản phẩm
- cột 'label': là nhãn của nhận xét đó

Chủ đề của dữ liệu được thu thập liên quan đến quần áo, giày dép, đồ ăn và đồ gia dụng. Mỗi chủ đề chúng tôi chọn ra 1 vài sản phẩm nổi bật nhiều lượt mua và đánh giá để đảm bảo tính đa dạng và khách quan.

Dữ liệu sẽ trông như sau:

	A	B
1	review	label
2	shop quá tệ đặt màu trắng giao màu xanh	0
3	mình rất hài lòng với sản phẩm của shop hàng đẹp đường chỉ may chắn chắn shop tư vấn nhiệt tình và chu đáo với khách hàng ship nhanh5 sao	1
4	hình ảnh chỉ mang tính chất nhận xu hàng về nhanh sản phẩm chất lượng shipper thân thiện sản phẩm ổn giá cả hợp lí ship hàng hơi lâu đánh giá mức độ trung bình giao hàng nhanh với giá này thì khá	1
5	sản phẩm thì như hình mình họa giao hàng nhanh đóng gói cẩn thận anh shipper thân thiện cám ơn shop sản phẩm y hệt hình ngon bổ rẻ quá ok nhé quên k chụp hình để nhận xu thôi ạ	1
6	sản phẩm chất lương đã mua cái thứ 2 rẻ hơn các shop khác mọi người nên mua hình video chủ yếu nhận xu	1
7	rất hài lòng y hình 1 ngày đã giao hàng shop cũng nhắn tin rất nhanh	1
8	đồ đẹp nhé mọi người shop giao đúng size rồi siêu nhanh luôn mới đặt ngày trước ngày sau có đồ	1
9	hàng đẹp lắm nha mọi người giống như hình nên mua nha mọi người đóng gói sản phẩm chắc chắn hàng y như hình không còn gì để miêu tả nữaaaaaaaa chất lượng tuyệt vời đúng mẫu mã shop làm ă	1
10	sản phẩm phù hợp với giá tiền mang đến phong cách trưởng thành cho người dùng các bạn nam mặc áo này vào sẽ trông như người lớn không còn trẻ trâu trẻ nghé như xưa gái thì khen sếp thì quý bố	1
11	sản phẩm rất ok một số đt à em đi đi nhá nhá em đang ở đâu nhỉ anh nhỉ thế giới này có gì đâuuu em có biết đâu em ơi à à anh à à cảm ơn shop nhiều sẽ ủng hộ shop nhiều em mk ưng lắm vừa zin luôn	1
12	omg quá đẹp quá xuất sắc đẹp hơn mình nghĩ khá nhiều không bao giờ lên review nhưng mà lần này không viết feedback không được vì thực sự là quá rẻ cho một sản phẩm tốt theo mình thấy đây khôn	1
13	ok nhận xuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuuu tốt so với giá tiền tôi cảm thấy hàng rất tốt vãi thun ko quá mỏng áo hoàn thiện 90 hợp với giá tiền phối đồ nào cũng hợp ngon lừa tình yêu của mình lor	1
14	chất siêu xịn hơn cả mong đợi sẽ ủng hộ shop lâu dài rẻ đẹp mát đúng size đúng màu like lần sau còn ủng hộ shop nhiều hơn nữa	1
15	hàng rất đẹp shipper rất thân thiện giao hàng rất nhanh ủng hộ dài dài phân loại hàng trắngl 60 67kg đã nhận được hàng chất lượng ổn vải này dễ mặc nha hình ảnh minh hoạ	1
16	công nhận giao nhanh túi đóng gói cẩn thận chất lượng cũng tốt so với giá cả thì quá ổn shop phục vụ nhiệt tình lưu shop lại lần sau mua riêng mình thay thi ok lam nha ca nha	1
17	chất liệu tốt màu sắc đẹp sản phẩm giá phải chăng mà mặc đẹp mình thích quá gói hàng kỹ lưỡng sạch sẽ tôi rất hài lòng hẹn gặp lại shop sớm chất liệu vải màu sắc đen trắng giao hàng nhanh hàng mu	1
18	hàng giao nhanh oki lắm giống ảnh sản phẩm tốt 10 điểm nha nên mua	1
19	đẹp cực kì giao hàng nhanh hình ảnh mang tính chất nhận xuu vvv	1
20	đặc áo khoác mà giao áo thun shop lừa đảo	0
21	mặc nhìn rất oce	1

Hình 2.1: Tổng quan dữ liệu sau khi thu thập

Khai phá dữ liệu:

2.2. Quy tắc gán nhãn

Sau khi thu thập dữ liệu, chúng tôi dựa vào cảm nhận của bản thân về con người thực hiện gán nhãn cho dữ liệu bằng cách thủ công.

Quy tắc gán nhãn đối với 3 nhãn như sau:

- Nhãn 0 (negative): những lời chê, phàn đối với sản phẩm hoặc dịch vụ của sản phẩm, doanh nghiệp liên quan.
- Nhãn 1 (positive): những lời khen ngợi hoặc khuyến khích mua hàng.

- Nhãn 2 (neutral): những nhận xét không chê cũng không khen ngợi hoặc cảm xúc khen bằng với cảm xúc chê hoặc những nhận xét có nội dung không liên quan.

Ví dụ với 60 câu nhận xét được chia đều có 20 câu thuộc nhãn 0, 20 câu thuộc nhãn 1 và 20 câu thuộc nhãn 2:

review	label	lý do	Các từ
rất hài lòng với sản phẩm đóng gói chắc chắn thời gian giao hàng rất nhanh cảm ơn shop nhiều	1	Thái độ hài lòng	hài lòng, chắc chắn
hàng giao nhanh ok lắm giống ảnh sản phẩm tốt 10 điểm nha nên mua	1	Khen ngợi, khuyên mua	nhanh, ok, lắm, giống ảnh, tốt
hàng nhận về chưa ăn nhưng sản phẩm nhìn rất sạch sẽ và ngon mắt hy vọng là ăn cũng sẽ rất ngon	1	Khen ngợi	rất, sạch sẽ, ngon
quần đẹp vải cũng đẹp size chuẩn luôn mặc đẹp lắm nha giá lại rẻ mua 3 quần mà gần 300k quá rẻ luôn giao hàng nhanh mà giao đúng màu đúng size luôn mọi người nên mua nhé	1	Khen ngợi, khuyên mua	đẹp, lắm, quá, rẻ, nhanh
đúng màu quần đúng size giao hàng nhanh show uy tín	1	Khen ngợi	đúng, nhanh, uy tín
mẹ em khen khô cá ngon lắm ạ shop đóng gói nhanh gói hàng chắc chắn vì là lần đầu em mua cho mẹ nhưng mà đến cả mẹ em cũng khen thì chất lượng không có gì để chê nhen	1	Khen ngợi	khen, ngon, lắm, nhanh, chắc chắn
giá trị tuyệt vời so với giá tiền kiểu dáng regular fit rất phù hợp chất liệu cao cấp bền đẹp	1	Khen ngợi	tuyệt vời, bền, đẹp
khô ngon mua nhiều lần rồi giao hàng nhanh chất lượng nên mua nha mn	1	Khen ngợi, khuyên mua	ngon, nhanh, chất lượng
rất đáng đồng tiền thiết kế đẹp và phong cách đa năng và phù hợp với mọi lứa tuổi	1	Khen ngợi	đẹp
bất kể màu sắc hay số lượng nó đều đáp ứng yêu cầu của đơn hàng tôi đã	1	Khen ngợi	tốt

đặt và sản phẩm trong tình trạng tốt			
tai nghe chụp tai dễ thương nghe nhạc hay âm thanh sống động to rõ chụp mềm không bị đau tai có thể điều chỉnh để vừa với đầu từng người rất đáng tiền shop gói hàng cẩn thận giao nhanh trả lời nhiệt tình	1	Khen ngợi	dễ thương, hay, đáng tiền, cẩn thận, nhanh, nhiệt tình
quá tốt ưng cực kì	1	Khen ngợi	quá, tốt, cực kì
sản phẩm rất tốt và đáng giá tiền vận chuyển nhanh chóng và đóng gói cẩn thận tôi sẽ giới thiệu cho bạn bè mua sản phẩm tại shop	1	Khen ngợi	tốt, đáng, nhanh, cận thận
shop tư vấn nhiệt tình giáp hàng nhanh sp lúc về giống như hình khi đeo ko cảm thấy nặng tai rất đáng để tham khảo khi mua	1	Khen ngợi	nhiệt tình, nhanh, giống, rất, đáng
tôi rất hài lòng với sản phẩm cảm ơn bạn vì đã giao hàng nhanh chóng	1	Khen ngợi	rất, hài lòng, nhanh
sản phẩm giao nhanh đúng size đúng màu form đẹp hài lòng về sản phẩm	1	Khen ngợi	đúng, đẹp, hài lòng
màu sắc đẹp hoa văn bắt mắt	1	Khen ngợi	đẹp
đồ rất đẹp nha mọi người	1	Khen ngợi	rất, đẹp
sản phẩm chất lượng tốt sẽ ủng hộ lần nữa nếu cần	1	Khen ngợi	chất lượng, tốt
hài lòng khi nhận hàng chất lượng sử dụng sẽ kiểm chứng khi sử dụng cảm ơn shop rất nhiều	1	Khen ngợi	hài lòng, chất lượng, rất
shop quá tệ đặc màu trắng giao màu xanh	0	Chê bai	tệ
đặt này giao kia không đúng màu đúng mẫu mua 3 cái không nào được shop lừa đảo	0	Chê bai	không đúng, lừa đảo
khác hoàng toàn với ảnh nha	0	Chê bai	khác
vải hơi dỏm nón thì hơi nhỏ giao hàng bị nhăn khá nhiều nói chung là áo hơi nhỏ mặc dù đặt xl	0	Chê bai	dỏm, nhăn, nhiều, nhỏ
hàng tạm chân bàn yếu vừa mua về bị	0	Chê bai	tạm, yếu, hư

hư ngay			
thất vọng nha không như mong đợi	0	Chê bai	thất vọng, không như mong đợi
chất liệu nhựa mũ xấu mũ có mùi nhựa rất khó chịu	0	Chê bai	xấu, khó chịu
hàng kém chất lượng	0	Chê bai	kém
giao hàng lỗi cho mình ạ mũ to đội lỏng lẻo kính không bấm được vào đều vậy	0	Chê bai	lỗi, không, đều
không giống đặt hàng này giao hàng khác	0	Chê bai	không giống, khác
không đúng như sản phẩm đặt	0	Chê bai	không đúng
mùi không được thơm	0	Chê bai	không được
mùi nước hoa khó chịu như mùi cồn gây buồn nôn	0	Chê bai	khó chịu, buồn
áo bé quá vải cũng xấu dư chỉ nhiều không ưng ý được cái nào	0	Chê bai	bé, quá, xấu, không ưng ý
thất vọng lắm luôn í áo này nó trễ vai kinh khủng không mặc nổi ai mà vai nhỏ cần cân nhắc nha không có phần chun cố định đâu tuy rẻ nhưng mà không hài lòng với chất lượng tí nào cả	0	Chê bai	thất vọng, kinh khủng, không, hài lòng
không đúng hàng cũng không giống như trên ảnh	0	Chê bai	không đúng, không giống
không hài lòng lắm vì ảnh in lệch và không được đẹp shop rep tin nhắn lâu	0	Chê bai	không hài lòng, không được đẹp, lâu
pin chất lượng kém dung 3 ngày hỏng luôn	0	Chê bai	kém, hỏng
hàng không đúng mô tả có dấu hiệu của đã sử dụng hàng không có tem mác thông số	0	Chê bai	không đúng, không có
hình của mình thiều nhiều lắm á hình cắt sơ sài nữa nhắn với shop thì không thấy rep lại	0	Chê bai	thiếu, sơ sài, không
thông báo để phòng ngừa đuối nước	2	Nội dung không liên quan	

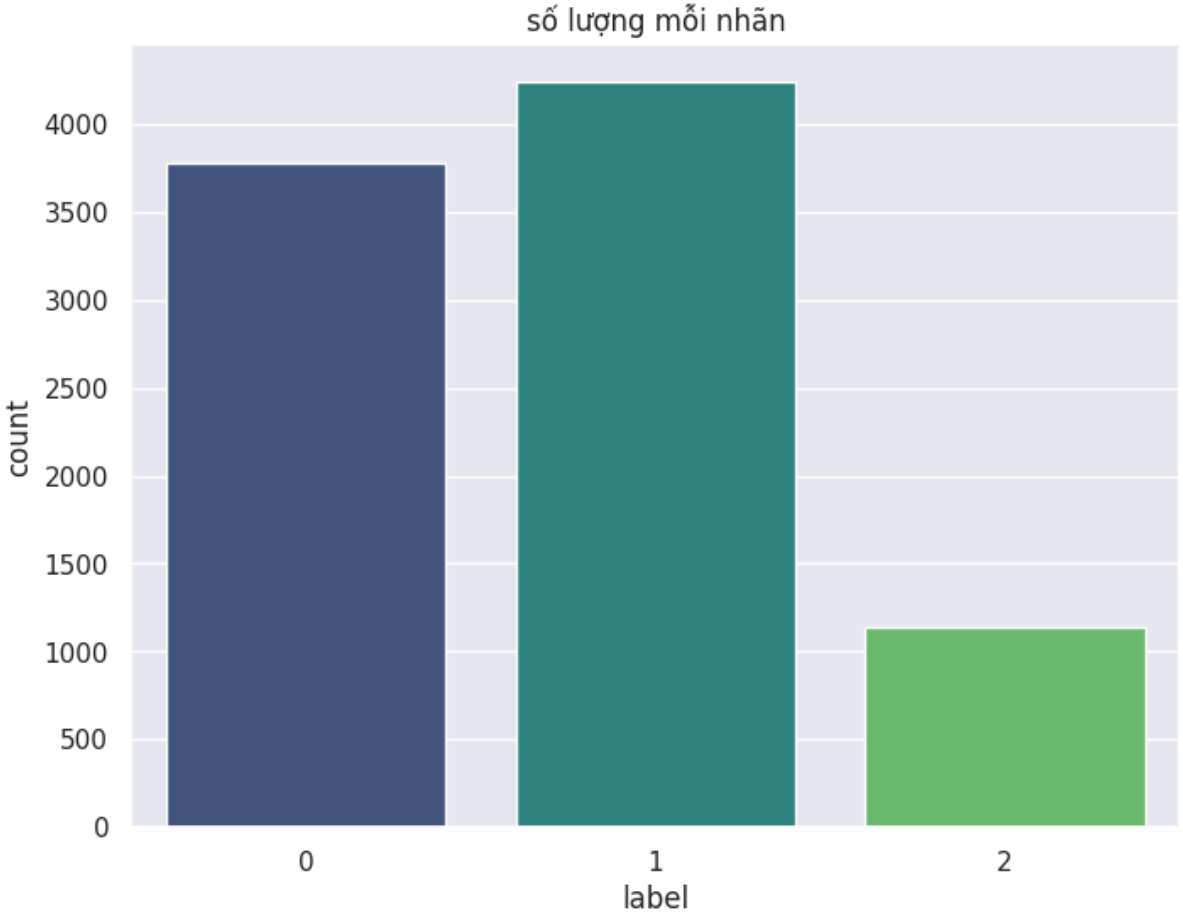
cho trẻ em bộ công an khuyến cáo trang bị kỹ năng bơi cho trẻ khi trẻ tắm vui chơi cần có người lớn giám sát mặc áo phao khi đi phương tiện thủy lấp các hố sâu giếng nước không cần thiết khi phát hiện người đuối nước phải hô hoán và dùng các vật dụng cây sào phao dây để kịp thời cứu nạn tuyệt đối không nhảy xuống nước cứu người khi không biết bơi người lớn cần chủ động trang bị kiến thức về sơ cứu người bị đuối nước			
chưa sử dụng nên chưa biết cho shop 5 sao hình ảnh không liên quan	2	Không chê cũng không khen	chưa
tạm được nhưng không biết có bền không	2	Không chê cũng không khen	tạm được, không biết
chưa dùng thử nên chưa biết chất lượng và hiệu quả thế nào nhưng nhìn kết cấu giống hàng xách tay	2	Không chê cũng không khen	chưa, chưa biết
nên tăng 1 size mới vừa nha mọi người	2	Không chê cũng không khen	
ảnh hơi cong xíu nhưng không sao	2	Không chê cũng không khen	không sao
khô này ăn rất ít hao nhe 1 con thôi là ăn được 1 dĩa do vấn đề dề bảo quản nên người ta ướp mặn mới để lâu được ai muốn ăn nhạt hơn thì ngâm nước muối nhe shop hướng dẫn ngâm nước ấm nhưng mình ngâm nước muối muốn khô giòn thì chiên 2 lần nhe lần đầu chiên lửa nhỏ cho khô dốt dốt thôi rồi để nguội cho bay hơi nước sau đó để dầu nóng bỏ khô vô chiên vàng bao giòn nhe mọi	2	Không chê cũng không khen	
phương pháp này áp dụng dựa trên giả định là hàng hóa nào nhập trước thì được xuất trước và hàng tồn cuối kỳ là hàng được nhập gần thời điểm cuối kỳ theo phương pháp này giá trị hàng xuất kho được tính theo giá thực tế của hàng nhập kho ở thời điểm đầu kỳ hoặc gần đầu kỳ và do vậy giá trị	2	Nội dung không liên quan	

của hàng tồn kho sẽ là giá của hàng nhập kho ở thời điểm cuối kỳ hoặc gần cuối kỳ còn tồn kho hàng tốt nên mua			
sử dụng biểu tượng chỉnh sửa để ghim thêm hoặc xóa đoạn sử dụng biểu tượng chỉnh sửa để ghim thêm hoặc xóa đoạn	2	Nội dung không liên quan	
để dùng xem có bền không sẽ đánh giá tiếp	2	Không khen cũng không chê	xem
em cảm thấy mình không có tiền mua đồ cho anh đi ăn sáng khum khum sao	2	Nội dung không liên quan	
tuyệt vời tôi biết tôi đã nhận được tin nhắn của bạn và tin nhắn của bạn vừa gửi cho tôi một tin nhắn từ người bạn của tôi người đang cố gắng và không trả lời tin nhắn của tôi nên vâng ireyio tôi chắc rằng đây sẽ là thời điểm tốt nếu	2	Nội dung không liên quan	
ok chơi game trúng xế hộp click để thu thập khóa vàng cùng voucher 250k ngay	2	Nội dung không liên quan	
không biết như vậy là có bị mốc không nữa có vài cái bị trắng trắng	2	Không khen cũng không chê	không biết
tạm được thôi ạ	2	Không khen cũng không chê	tạm
tạm ok sẽ ủng hộ thêm	2	Không khen cũng không chê	tạm
hơi bị ngọt tạm được	2	Không khen cũng không chê	tạm
ngon nhưng mà không có cay nếu cay thêm thì sẽ ngon hơn gà không bị vụn sấy khô vị vừa ăn vì không cay lắm nên vị ngọt đậm	2	Tính khen tính chê được cân bằng	ngon, không
đã nhận được hàng chưa ăn thử	2	Không khen cũng không chê	chưa
giao hàng nhanh đã ăn thử nhiều yến mạch ít hạt và không có hạt óc chó	2	Tính khen tính chê được cân bằng	nhanh, ít, không có

như quảng cáo			
granola siêu hạt ngũ cốc	2	Nội dung không liên quan	
khá ổn	2	Không khen cũng không chê	khá
ngọt	2	Không khen cũng không chê	
hộp có nhiều yến mạch hạ bí dừa khô nhưng cực kỳ ít các hạt óc chó hạnh nhân không có việt quất	2	Tính khen tính chê được cân bằng	nhiều, nhưng, ít, không có
được mỗi áo đen ổn nhất còn lại tiền nào của nấy thôi nói thật	2	Không khen cũng không chê	ổn

Bảng 2.1: Quy tắc gán nhãn

Số lượng mẫu của môi nhãn sau khi thu thập và gán nhãn như sau:



Hình 2.2: Biểu đồ cột thể hiện tổng số mẫu của mỗi nhãn

Nhận xét:

- Số lượng mẫu của nhãn 0 và 1 khá cân bằng
- Số lượng mẫu của nhãn 2 ít hơn nhiều so với 2 mẫu còn lại.

2.3. Các bước xử lý dữ liệu

2.3.1. Chuyển đổi chữ viết hoa thành chữ thường

Trong xử lý ngôn ngữ tự nhiên, việc chuyển đổi chữ hoa thành chữ thường là việc cần thiết. Việc này nhằm mục đích là đảm bảo tính thống nhất và tránh trường hợp trùng chữ do viết hoa.

Ví dụ:

Chưa xử lý: “ Đẹp wa OK”

Đã xử lý: “đẹp wa ok”

2.3.2. Loại bỏ ký tự đặc biệt

Các ký tự đặc biệt không có vai trò quá quan trọng trong quá trình phân tích. Vì thế nên loại bỏ chúng để giảm kích thước tập dữ liệu.

Ví dụ:

Chưa xử lý: “tuyệt vời!!!”

Đã xử lý: “tuyệt vời”

2.3.3. Loại bỏ chữ cái lặp

giới trẻ thường có văn hóa kéo dài chữ cái cuối cùng mỗi khi gõ chữ hoặc những lúc vô tình khiến cho 1 số chữ cái bị lặp nhiều hơn. Cần phải xử lý vấn đề này để dữ liệu tốt nhất có thể.

Ví dụ:

Chưa xử lý: “ngonnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn”

Đã xử lý: “ngon”

2.3.4. Xử lý chữ viết tắt

Thói quen viết tắt đã trở nên rất phổ biến khi mà việc gõ những chữ dài mà xuất hiện nhiều lần là một việc khiến con người cảm thấy khó chịu. Việc xử lý những chữ viết tắt là điều tối quan trọng để dữ liệu thống nhất và giảm kích thước dữ liệu.

để quá trình xử lý diễn ra nhanh chóng thì chúng tôi đã tự tạo ra một dictionary với key là các từ viết tắt và value của key tương ứng là từ gốc.

Ví dụ:

Chưa xử lý: “kh giống vs mô tả nha mk đặt màu nâu nhưng giao là màu xám và còn nhầm size cho mình”

Đã xử lý: “không giống với mô tả nha mình đặt màu nâu nhưng giao là màu xám và còn nhầm size cho mình”

Các từ viết tắt ở đây là:

- kh = không
- vs = với
- mk = mình

2.3.5. Tách từ

Chia các câu nhận xét thành các từ đơn, từ ghép có nghĩa thành các tokens. Là bước quan trọng không thể bỏ qua, bước này nhằm tạo ra kho từ vựng cho mô hình.

Ví dụ:

Chưa xử lý: “hàng chất lượng lắm hài lòng nha”

Đã xử lý: “hàng”, “chất_lượng”, “lắm”, “hài_lòng”, “nha”

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN

3.1. Trích xuất đặc trưng

Để mô hình có thể hoạt động trên bộ dữ liệu, chúng tôi sử dụng kỹ thuật TF-IDF (Term Frequency - Inverse Document Frequency) để vector hóa dữ liệu, nhưng sẽ có một số vấn đề khi một từ không xuất hiện trong văn bản dẫn đến công thức toán học gặp lỗi như $log(0)$. Để tránh trường hợp này chúng tôi sử dụng biện pháp additive smoothing để tránh trường hợp $log(0)$.

TF-IDF là một kỹ thuật được sử dụng để đánh giá tầm quan trọng của 1 một từ trong tập dữ liệu (hoặc 1 tokens sau khi tách từ). TF-IDF là viết tắt của "Term Frequency-Inverse Document Frequency", trong đó:

- TF-Term Frequency: TF đo lường tần suất xuất hiện của 1 tokens trong văn bản với công thức như sau: $TF(t,d) = \frac{số\ lần\ t\ xuất\ hiện\ trong\ d}{tổng\ số\ từ\ trong\ d}$

  với:
  - t: một tokens trong văn bản
  - d: văn bản trong tập dữ liệu
- IDF-Inverse Document Frequency: IDF ước lượng mức độ quan trọng của t với công thức như sau: $TF(t,D) = log(\frac{1+n}{DF(t)+1}) + 1$

  với:
  - n: tổng số văn bản trong tập dữ liệu
  - D: tập dữ liệu
  - DF(t): số văn bản chứa t
- TF-IDF: $TF - IDF(t,d,D) = TF(t,d) * IDF(t,D)$

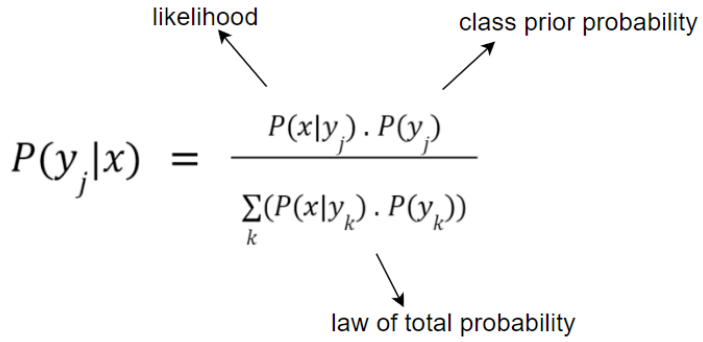
Kết quả là vector số hóa cho mỗi văn bản trong tập dữ liệu, trong đó mỗi phần tử của vector đại diện cho giá trị TF-IDF của một tokens cụ thể trong văn bản. Các giá trị cao thường đại diện cho sự quan trọng của tokens đó trong văn bản cụ thể, còn những giá trị thấp đại diện cho những tokens phổ biến.

3.2. Mô hình Multinomial Naive Bayes

Mô hình Multinomial Naive Bayes là một mô hình phân loại mạnh mẽ và linh hoạt dựa trên phương pháp học xác suất(Probabilistic Learning Method), đặc biệt là trong xử lý ngôn ngữ tự nhiên và các nhiệm vụ phân loại văn bản.

Xây dựng các hàm của mô hình theo công thức sau:

Xác suất của đánh giá cần xác định nhãn của từng nhãn:



Trong đó:

- Hàm likelihood: $log(P(x_i|y_j)) = log(\frac{count(x_i,y_j)+\alpha}{\sum_k (count(x_k,y_k)+\alpha}$)
  - count(x_i, y_i) là số lần xuất hiện của từ x_i trong các văn bản thuộc lớp y_i.
  - α là hằng số trọng số smooth để tránh giá trị xác suất bằng 0.
  - k duyệt qua tất cả các từ trong từ điển
- Hàm Class Prior: $log(P(y_j)) = log(\frac{count(y_j)+\alpha}{\sum_i (count(y_i)+\alpha)}$)
  - count(y_i) là số lần xuất hiện của lớp y_i. trong tập dữ liệu.
  - $\sum_i (count(y_i) + \alpha)$ là tổng các mẫu dữ liệu trong tập dữ liệu cộng thêm α.

3.3. Ví dụ minh họa

Cho ví dụ minh họa như sau

	review	label
doc1	áo hơi đắt áo đẹp	2
doc2	ác khác hình áo quá xấu	0
doc3	áo tốt	1
doc4	áo rất đẹp	1
doc5	áo cũng đẹp vừa người	?

Bảng 3.2.1: Ví dụ minh họa

Cho 5 mẫu dữ liệu trong đó có 4 mẫu đã được gán nhãn, mẫu còn lại được dùng để dự đoán. Với 0 là nhãn negative, 1 là nhãn positive và 2 là nhãn neutral.

TF*IDF										
	hình	hơi	khác	quá	rất	tốt	xấu	áo	đắt	đẹp
doc1	0	0.519	0		0	0	0	0.541	0.519	0.409
doc2	0.443	0	0.443	0.443	0	0	0.443	0.541	0	0
doc3	0	0	0	0	0	0.886	0	0.541	0	0
doc4	0	0	0	0	0.726	0	0	0.541	0	0.572
doc5	0	0	0	0	0	0	0	0.541	0	0.833

Bảng 3.2.2: Giá trị TF*IDF của các từ vựng trong 5 mẫu

Gán X_train bằng cột review và y_train là cột label. X_test sẽ là doc5.
Chuyển đổi y_train thành vector nhị phân.

0	1	2
0	0	1
1	0	0
0	1	0
0	1	0

Bảng 3.2.3: Vector nhị phân của 3 nhãn 0, 1 và 2

Bằng cách thực hiện phép nhân ma trận giữa ma trận y_train.T và X_train. Điều này giúp tính toán tổng của số lần xuất hiện của mỗi từ cho từng lớp.

$count(x_i, y_j)$										
y	hình	hơi	khác	quá	rất	tốt	xấu	áo	đắt	đẹp

0	0.443	0	0.443	0.443	0	0	0.443	0.462	0	0
1	0	0	0	0	0.726	0.886	0	0.841	0	0.572
2	0	0.519	0	0	0	0	0	0.541	0.519	0.409

Bảng 3.2.4: Sau khi nhân 2 ma trận y_train.T và X_train

Thực hiện +1 smoothing để tránh các giá trị xác suất bằng 0.

$count(x_i, y_j)$										
y	hình	hơi	khác	quá	rất	tốt	xấu	áo	đắt	đẹp
0	1.443	1	1.443	1.443	1	1	1.443	1.462	1	1
1	1	1	1	1	1.726	1.886	1	1.841	1	1.572
2	1	1.519	1	1	1	1	1	1.541	1.519	1.409

Bảng 3.2.5: Sau khi smoothing

Tính $\sum_k (count(x_k, y_k) + \alpha)$:

y	$log(\sum_k (count(x_k, y_k) + \alpha)$
0	12.234
1	13.025
2	11.988

Bảng 3.2.6: giá trị $log(\sum_k (count(x_k, y_k) + \alpha)$ của từng nhãn

Tính likelihood:

$log(P(x_i \mid y_j))$

y	hình	hơi	khác	quá	rất	tốt	xấu	áo	đắt	đẹp
0	-2.137	-2.504	-2.137	-2.137	-2.504	-2.504	-2.137	-2.214	-2.504	-2.504
1	-2.567	-2.567	-2.567	-2.567	-2.02	-1.932	-2.567	–1.956	-2.567	-2.114
2	-2.484	-2.065	-2.484	-2.484	-2.484	-2.484	-2.484	-2.501	-2.065	-2.14

Bảng 3.2.7: giá trị likelihood của từng từ vựng trong 3 nhãn

Tính class prior:

y	$log(P(y_i))$
0	-1.386
1	-0.693
2	-0.1386

Bảng 3.2.8: giá trị class prior của từng mẫu

Dự đoán cho doc5:

Tính: $log(\sum_k (P(x|y_k) \,.\, P(y_k)))$ = -2.951

y	$P(x\|y_j) \,.\, P(y_j)$	$log(P(x\|y_j))$	$P(x\|y_j)$
0	-4.647	-1.696	0.183
1	-3.535	-0.584	0.557
2	-4.3	-1.35	0.26

Bảng 3.2.9: Xác suất của từng mẫu

Vậy doc5 có nhãn là 1 do có xác suất cao nhất là 0.557 thuộc lớp 1.

CHƯƠNG 4: CÀI ĐẶT MÔ HÌNH

4.1. Demo

Để tìm hiểu chi tiết về tập dữ liệu và chương trình chúng tôi cài đặt hãy truy cập vào link : Hungtofu/NLP_Sentiment_Analysis (github.com)

CHƯƠNG 5: KẾT QUẢ VÀ ĐÁNH GIÁ

5.1. Đánh giá hệ thống

Trong đồ án này, để đánh giá kết quả phân loại của hệ thống chúng tôi sử dụng những độ đo như: Precision, Recall, F1-Score, Accuracy.

Giả sử các dự đoán có thể có như sau:

		Thực tế	
		Positive	Negative
Dự đoán	Positive	True Positive (TP)	False Positive
	Negative	False Negative (FN)	True Negative

Bảng 5.1.1: giả định kết quả dự đoán

- Precision : là có bao nhiêu trường hợp được dự đoán chính xác thực sự là Positive. Accuracy rất hữu ích trong trường hợp False Positive được quan tâm nhiều hơn False Negative giả. Công thức tính Precision như sau: $Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$
- Recall : là có bao nhiêu trường hợp True Positive mà chúng tôi có thể dự đoán chính xác bằng mô hình của mình. Recall là thước đo hữu ích trong trường hợp False Negative được quan tâm nhiều hơn False Positive. Công thức tính Recall như sau: $Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$
- F1-Score : là kết hợp giữa Precision và Recall. Công thức tính F1-Score như sau: $F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
- Accuracy : là tỷ lệ giữa số lượng các dự đoán đúng và tổng số lượng tất cả dự đoán. Công thức tính Accuracy như sau: $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

5.2. Kết quả

Ở đồ án này, chúng tôi sẽ chia dữ liệu ra làm 2 tập, tập train sẽ chiếm 80% tập dữ liệu, còn lại tập test sẽ là 20%. Sau khi thực hiện huấn luyện mô hình và tiết hành dự đoán và tính toán các độ đo thì chúng tôi thu được như sau:

	Precision	Recall	F1-Score
0	0.85	0.92	0.88

1	0.80	0.92	0.86
2	0.89	0.11	0.20
Accuracy			0.83

Bảng 5.2.1: Kết quả dựa trên tập test

Thông qua kết quả trên, chúng tôi có một số nhận xét như sau:

- Hai nhãn 0 và 1 cho ra 3 kết quả Precision, Recall và F1-Score khá tốt.
- Nhãn 2 cho kết cả 3 kết quả đều không tốt và không đáng tin cậy (khi Precision = 0.89 nhưng Recall = 0.11), lý do là vì tập dữ liệu bị mất cân bằng khi các nhận xét có nhãn là 2 quá ít.
- Tổng quan kết quả Accuracy cho ra tốt.

Một số trường hợp dự đoán sai:

Review	nhãn	dự đoán
đóng_gói không được cẩn_thận cho lắm nhưng giao hàng cũng rất nhanh	2	1
em có_thể làm gì đâu ý ý gì đâu ý nhỉ không có gì ăn_không có tiền ăn	2	0
hàng rất chất_lượng nên mua né mọi người mỗi tội đóng_gói hơi xơ xài	2	1
được không ngon	0	1
đúng như hình	1	0

Bảng 5.2.2: Một số mẫu dự đoán sai

Chúng tôi có một số nhận xét như sau:

- Mô hình không thể phân biệt được vị trí của những từ trong câu. Ví dụ như "được không "ngon", ở đây là câu nhận xét negative nhưng mô hình lại dự đoán là positive là vì "được" và "ngon" là 2 từ mang tính positive chỉ có "không" là mang tính neutral, nhưng khi "không" kết hợp với "ngon" lại mang tính negative và mô hình không phân biệt được điểm này để đưa ra dự đoán chính xác.
- Những câu mang nhãn 2 thường bị dự đoán sai là do số lượng mẫu mang nhãn 2 dùng để huấn luyện quá ít nên mô hình không thể dự đoán tốt cho nhãn 2.

TÀI LIỆU THAM KHẢO

[1] “Python Vietnamese Toolkit - Viet-Trung Tran” 30 thg 6, 2021. [Online]. Available: https://pypi.org/project/pyvi/?fbclid=IwAR3iLvf-uoSoOoypQlkxKFz2Lr-8zsK94fXra6u7Y1ArRMUD0RbawMCi9O4

[2] “vietnamese-stopwords - stopwords” 26 thg7, 2022. [Online]. Available: https://github.com/stopwords/vietnamese-stopwords?fbclid=IwAR1db_21ISF8NnuMP291Sa4-ucZT8MSERvjDgtn9IBTLDWntGsxzj_YBzbM

[3] “Naive Bayes classifier for multinomial models - scikit learn”. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn-naive-bayes-multinomialnb

[4] “tf-idf - wikipedia”. [Online]. Available: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

[5] “Accuracy and precision - wikipedia”. [Online]. Available: https://en.wikipedia.org/wiki/Accuracy_and_precision