Báo cáo ck xử lí ngôn ngữ tự nhiên

National Language Processing (University of Information Technology)

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
VNUHCM - UIT

KHOA KHOA HỌC MÁY TÍNH

ĐỒ ÁN MÔN HỌC

CS221_XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Đề tài:

PHÂN LOẠI CẢM XÚC VĂN BẢN

Giảng viên: Nguyễn Trọng Chỉnh

Sinh viên thực hiện: **Ngô Quang Vinh_19522523**

Nguyễn Văn Đức Ngọc_20521666

Trần Nguyễn Phúc Minh Quân_20521797

Tp. Hồ Chí Minh, tháng 07 năm 2024

Mục lục

CHƯƠNG 1: Giới thiệu bài toán

Phân loại cảm xúc văn bản (hay Sentiment Analysis) là một nhiệm vụ quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), nhằm xác định cảm xúc hoặc ý kiến của người viết trong một đoạn văn bản. Bài toán này có nhiều ứng dụng thực tiễn, bao gồm:

- Phân tích đánh giá sản phẩm: Xác định ý kiến của khách hàng về sản phẩm hoặc dịch vụ.
- Phân tích xu hướng thị trường: Hiểu được cảm nhận chung của cộng đồng về một chủ đề cụ thể.
- Phát hiện ngôn từ kích động thù địch: Xác định các bình luận có chứa ngôn ngữ tiêu cực hoặc xúc phạm.
- Cá nhân hóa trải nghiệm người dùng: Đề xuất nội dung phù hợp với sở thích và cảm xúc của người dùng.

Để giải quyết bài toán phân loại cảm xúc văn bản, các phương pháp NLP thường được sử dụng bao gồm:

- Xử lý ngôn ngữ tự nhiên (NLP): Các kỹ thuật NLP được sử dụng để tiền xử lý văn bản, trích xuất đặc trưng và biểu diễn văn bản dưới dạng vectơ.
- Học máy: Các thuật toán học máy như SVM, Naive Bayes, mạng nơ-ron nhân tạo được sử dụng để phân loại văn bản thành các lớp cảm xúc khác nhau (ví dụ: tích cực, tiêu cực, trung lập).

Bài toán phân loại cảm xúc văn bản là một lĩnh vực nghiên cứu đang phát triển mạnh mẽ với nhiều tiềm năng ứng dụng thực tiễn. Các nghiên cứu trong lĩnh vực này đang tập trung vào việc phát triển các mô hình hiệu quả hơn, có thể xử lý được các loại văn bản phức tạp và đa dạng hơn.

CHƯƠNG 2: Ngữ liệu

- Bộ ngữ liệu nhóm sử dụng cho đồ án là bộ ngữ liệu ViHSD. Đây là bộ ngữ liệu bằng tiếng Việt, được thu thập từ các bình luận trên các mạng xã hội như FaceBook và Youtube. Các nhãn trong bộ ngữ liệu được chú thích thủ công nhằm mục đích phục vụ cho các nghiên cứu phát hiện lời nói chán ghét (hate speech) một cách tự động trên các mạng xã hội.

- Bộ ngữ liệu bao gồm 33400 bình luận đã được gán nhãn với tổng số lượng từ vựng là 21239 từ. Các bình luận trong bộ ngữ liệu được chia làm 3 nhóm tương ứng với 3 nhãn sau: HATE (2), OFFENSIVE (1) và CLEAN (0). Trong bộ ngữ liệu ViHSD, có hai nhãn thể hiện bình luận chứa lời nói chán ghét và một nhãn chỉ đến bình luận bình thường. Các ý nghĩa chi tiết về các nhãn được mô tả trong bảng dưới đây:

Label	**Description**
CLEAN	Các bình luận không có bất kỳ hành vi quấy rối nào
OFFENSIVE	Các bình luận chứa nội dung quấy rối, kể cả những từ thô tục, nhưng không tấn công một đối tượng cụ thể.
HATE	Các ý kiến có nội dung quấy rối và thô tục, trực tiếp nhắm vào một cá nhân hoặc một nhóm người dựa trên đặc điểm cá nhân, tôn giáo và quốc tịch. Một số trường hợp xảy ra với nhãn HATE: ● Trường hợp 1: Các ý kiến chứa những từ thô tục và tấn công một đối tượng cụ thể như một cá nhân, cộng đồng, quốc gia hoặc tôn giáo. Trường hợp này dễ nhận biết là lời nói chán ghét. ● Trường hợp 2: Các ý kiến có tính chất phân biệt chủng tộc, quấy rối và có ý nghĩa chứa đựng sự căm ghét, tuy nhiên, không chứa các từ ngữ rõ ràng. ● Trường hợp 3: Các ý kiến có tính chất phân biệt chủng tộc, quấy rối và có ý nghĩa chứa đựng sự căm ghét, nhưng được thể hiện dưới hình thức ẩn dụ. Để nhận biết ý kiến này, người dùng cần có kiến thức đặc biệt về xã hội.

Bảng 1: Bảng ý nghĩa các nhãn trong bộ ngữ liệu ViHSD

Một số mẫu bình luận và nhãn tương ứng của nó trong tập ngữ liệu như sau đây:

text,nhãn thực,nhãn dự đoán
không giống ai,1,0
má,1,0
abe mua áo như này là chet con sela rồi,1,0
thoát tội nếu hiếp lại dm,1,2
hết mie nó rừng vàng biển bạc rồi,1,0
lông là lông dương là có nghĩa là lông cu,1,0
dùng tiền cu ơi từ thành thôi 1 người chống cả được,1,0
để mầm chi,1,0
đồ chai sting vô đầu nha,1,0
đấu đá nhau hơn cả xh đen vì có quyền đc qua,1,0
ôi cát mẹt của bả,1,0
lại sống như,1,0
làm kh thích thích làm con cặt,1,0
chửi mà đéo biết báo từ đời nào rồi và chỉ phạt các mua điện nhưng lại k để tránh thôi chứ dân dùng bằng nào thu bằng đấy toàn những con bò,1,2
óc chó,1,2
abe ochco,2,0
cá chết vẫn chưa rõ biển thì vẫn đầy vẫn những chuyện và những chuyện cứ ta sẽ sống đây,2,0
con noi doc giong con 3 vang ma,2,0
ns vậy lại cái đầu b i,2,0
tuong ao giac ba dien than kinh,2,1
lại hợp sâm clm đến v thì ff,2,1
mày nhìn lại mày đi đell đc 5 like à phải rồi mới bươi móc kiếm để kiểu ganh tỵ,2,0
trần ủa tao nhai lai của ai m thế đmm lướt lại cmt xem toàn mấy thánh kêu sống chậm thôi sống nhanh quá ak hay m cx nằm trong số cmt đấy r m quay ra ns tao,2,0
người j đâu mà chưởi người mình trưởi người ta quá trời đúng là bã chưa thấy chưa,2,0
giờ m mới ra và tai đã nghe đc rùi phải k,2,0
đã nói bao lần rồi các anh các chị có tiền đi vẫn đ hiểu à,2,0
a3 3 no j dan vay qua troi dan og muon nu vay s k bo tien ra mat ngta o do ng hay k o do no sua nhu cho sua,2,1
phobolsatv bien me đi may về mày đi nhin may nguoi con ko tha phobolsa rất sam,2,0
kệ mẹ thầy,2,0
em đẹp quá nhìn chị là mà xấu quá,2,0

ジェイ ホープ nói hay thì làm diêu tút bơ kiếm tiền ở đây nghĩ ra lên chém gió ko ra tiền lấp đầy cái bụng kêu đâu ku,0,1
cú lội ngược dòng của nhung,0,1
đang họp 1 ông người tàu bước vào phòng búng tay cái póc 10 thằng tàu 10 cái vali tiền vào phát cho 1 1 cái mấy lỉ ăn đi và dùm bọn ngộ,0,2
thế cong trải thảm tiếp cẩu tập,0,2
xạo có tiếng haha,0,1
giống abe vailon,0,1
chắc đéo bán dc hàng đâu,0,1
lại thêm một con khoác lác,0,2
các bác nhé lụt ở chỗ nào chứ ở tp có xây bờ kè có lũ vào mắt cho dân vùng thôi,0,2
đừng nghe cs nói mà hãy nhìn cs làm,0,2
nhin o tran dan dot nay ong ko dong tram toi dan vn trong nuoc pham tot ong da nhin nhan ra nghia,0,2
đậu nhìn như thằng nghiện ghê bỏ mẹ ra,0,1
má ba ơi giúp con chén,0,2
hôm bữa tôi nói con này khóc là thôi nhung anh này giọng nói đâu phải miền bắc đâu mà chửi bắc này bắc kia đến người ta,0,2
vâng là dân v là có thật v,0,2

Label	**Comment**	**Explain**
CLEAN	Mọi người ơi, cho mik hỏi mik theo dõi cô ấy mà mik hk pít cô ấy là con gái thiệt hả m.n	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
	Trước khi cắm lên đầu em 4 cái sừng nó cũng bảo y như thế :((((	Bình luận này xét về nghĩa, nó đang mô tả một tình huống xấu nhưng bình luận không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
	1 ng thì giết 1 mạng . 1 ng thì feed tụt lol	Bình luận này có chứa từ thô tục ‘lol’ nhưng viết theo kiểu tiếng lóng. Theo định nghĩa để phân biệt các nhãn, nội dung đầu tiên cần xét tới là có chứa nội dung quấy rối hay

	không thì bình luận này không chứa nội dung quấy rối nên được gán nhãn sạch.
Được anh ưi, lâu rồi không nghe phúc du rap 🤭	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
Cực kì thuyế...hác xì....á chết mẹ..	Bình luận này có chứa từ ngữ thô tục "Chết mẹ" nhưng nội dung không quấy rối, tấn công bất kì ai nên được gán nhãn sạch
Nó giống vãi lol a ưi!	Bình luận này có chứa từ ngữ thô tục "vãi lol" nhưng chỉ mang ý nghĩa nhấn mạnh, nội dung không quấy rối, tấn công bất kì ai nên được gán nhãn sạch
chất vlon a ưii	Bình luận này có chứa từ ngữ thô tục "vlon" nhưng chỉ mang ý nghĩa nhấn mạnh, nội dung không quấy rối, tấn công bất kì ai nên được gán nhãn sạch
Đinh Phước sống nhanh chậm t chửi hết 👌	Bình luận này mang ý nghĩa tiêu cực nhưng bình luận không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai nên được gán nhãn sạch
Thuyết phục vl	Bình luận này có chứa từ ngữ thô tục "vl" nhưng chỉ mang ý nghĩa nhấn mạnh, nội dung không quấy rối, tấn công bất kì ai nên được gán nhãn sạch
hợp lý a ưi 🤭🤭	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai

Xin phép cả nhà nhaaaa!!!	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
Tự hào quá VN ơi 😂	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
MV khá chất lượng	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
Tào Hưng đm khô máu :)))	Bình luận này có chứa từ ngữ thô tục "đm" nhưng chỉ mang ý nghĩa biểu cảm, nội dung không quấy rối, tấn công bất kì ai nên được gán nhãn sạch
Vc	Bình luận này có chứa từ ngữ thô tục "Vc" nhưng chỉ mang ý nghĩa biểu cảm, nội dung không quấy rối, tấn công bất kì ai nên được gán nhãn sạch
T đã đi ngủ từ lúc 10r nhưng đ thể nào ngủ nổi 😳	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
Phan Quỳnh Anh bóp d' là gì m	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
Bé bú trên để bố bú dưới, =))	**Bình luận này có từ ngữ thô tục 'bú' và cũng mang ý nghĩa không phù hợp, nên gán nhãn sạch là chưa hợp lý**

	Oiiiii nàng công chúa	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai
	Đoàn Quang Vinh dm quảng cáo	**Bình luận này chứa từ ngữ thô tục "dm" và cũng mang ý nghĩa tiêu cực nên gán nhãn sạch là chưa hợp lý**
OFFENSIVE	Đồ khùng	Bình luận này chứa từ xúc phạm "khùng". Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
	Bảo dan tri thap thì tự ái :)	Bình luận này chứa từ xúc phạm "dan tri thap". Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
	Bố lạy mày đừng "hiếp dâm" thị giác của bố nữa ad à 😂😂😂	Bình luận này chứa từ thô tục "hiếp dâm". Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
	Lưu Bảo Toàn đọc hại não vcl	Bình luận này chứa từ ngữ thô tục "vcl". Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
	lolzzzz	Bình luận này chứa từ ngữ thô tục "lolzzzz" mang ý nghĩa biểu thị cảm xúc tiêu cực. Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
	Ncc	Bình luận này chứa từ ngữ thô tục "Ncc" mang ý nghĩa biểu thị cảm xúc tiêu cực. Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.

đọc bài báo xong buồn ĩa đéo chịu đc	Bình luận này chứa từ ngữ thô tục "ĩa" mang ý nghĩa biểu thị thái độ tiêu cực. Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
mặc cái quần giống thằng biến thái	Bình luận này mang ý nghĩa miệt thị, tiêu cực nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
Im mẹ i	Bình luận này mang ý nghĩa miệt thị, tiêu cực nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
Về vn ăn mau lol à.	Bình luận này chứa từ ngữ thô tục "mau lol" và mang ý nghĩa quấy rối nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
Phúc Hưng clm :)))	**Bình luận này chứa từ ngữ thô tục "clm" nhưng về ý nghĩa lại thể hiện cảm xúc vui vẻ, tích cực nên gán nhãn offensive là chưa hợp lý**
đéo ai thèm làm ,vì họ biết khả năng của họ đến đâu	Bình luận này chứa từ ngữ thô tục "đéo" và mang ý nghĩa quấy rối, tiêu cực nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
Toàn ngu muội. Đi ra đó mà đánh đi.	Bình luận này mang ý nghĩa miệt thị, tiêu cực nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
Đéo ai bắt xem ??=}}	Bình luận này chứa từ ngữ thô tục "đéo" và mang ý nghĩa quấy rối, tiêu cực nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể

	Lê Thị Trang đắt vcd	Bình luận này chứa từ ngữ thô tục "vcl" nhưng không nhắm vào một người hoặc một nhóm cụ thể
	bắn bỏ	Bình luận này mang ý nghĩa quấy rối, tiêu cực nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
	tự hào vì có nhiều admin fanpage ngu nhất thế giới nữa :))	**Bình luận này mang ý nghĩa quấy rối, tiêu cực và chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể "admin fanpage"**
	Kkkk...iu nhau là phải có đụ..	Bình luận này có từ ngữ thô tục "đụ" và mang ý nghĩa quấy rối nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể.
	Mu này tởm thật	**Bình luận này mang ý nghĩa quấy rối hướng đến một đối tượng cụ thể nên gán nhãn offensive là chưa hợp lý**
	Người ta gọi là đú mà.	Bình luận này mang ý nghĩa quấy rối, miệt thị nhưng không chứa từ ngữ nhắm vào một người hoặc một nhóm cụ thể
	Tạ Thị Ngọc Duyên mày cũng v chứ khác đéo đâu xớ	**Bình luận này có từ ngữ thô tục "đéo" và mang ý nghĩa quấy rối, tiêu cực, đồng thời nhắm tới một đối tượng cụ thể nên gán nhãn 1 là chưa lý**
HATE	Dành cho lũ quan ngại	Bình luận này chứa một cụm từ "lũ quan ngại" đề cập đến một nhóm người với ý nghĩa tiêu cực

Dm Có a mới không ổn. Mày rình mày chịch riết ổn cái lol	Bình luận này chứa nhiều từ thô tục "Dm", "chịch", "cái lol". Ngoài ra nó còn chứa đại từ nhân xưng "mày" có ý nghĩa nhắm đến một người cụ thể
Ở đấy ngột ngạt quá thì đưa nó qua vũ hán cho nó thoáng mát	Bình luận này chứa cụm từ vũ hán đề cập đến một địa điểm cụ thể, mang ý nghĩa phân biệt chủng tộc. Tuy nhiên, bình luận này không chứa bất kỳ từ ngữ xấu nào.
Xem cái này chỉ muốn ỉa vào mặt mấy thằng cho xây tượng đài...	Bình luận này chứa từ ngữ thô tục 'ỉa' đồng thời mang nội dung tiêu cực nhắm đến một đối tượng cụ thể "mấy thằng"
Tủnn Tủnn trẩu	Bình luận này chứa nội dung quấy rối, tiêu cực và nhắm đến một người cụ thể "Tủnn Tủnn"
Tui nghĩ là xe Merc sẽ có camera hành trình phía sau lưu lại xe nào quay được cảnh này. Tội cắt ghép clip này là bôi nhọ lãnh đạo, vu khống, xuyên tạc chứ ít gì.	**Bình luận này không chứa từ ngữ thô tục và không mang nội dung quấy rối nhắm đến một người hay nhóm nào nên gán nhãn hate là chưa hợp lý**
xuất cai L, đm quần áo Vn toàn nhập siêu từ TQ hết 70% chứ xuất xuất clm chúng mày	Bình luận này chứa nhiều từ ngữ thô tục "cái L", "đm", "clm" và có từ "chúng mày" nhắm đến một nhóm đối tượng cụ thể
Hoàng Ngọc Hòa sũa cc	Bình luận này chứa từ ngữ thô tục "sũa cc" và có nội dung quấy rối nhắm dến một người cụ thể "Hoàng Ngọc Hoà"
Nguyễn Việt Cường ơ thằng này ra tù rồi à	Bình luận này không chứa từ thô tục nào nhưng có nội dung chế giễu,

	quấy rối nhắm đến một người cụ thể "thằng này" nên được gán nhãn hate
Trần dần bệnh tâm thần chuyện có nói ko chuyện ko nói có làm cho cả cộng đồng người việt ở Mỹ quá bức xúc truyền thông pho bolsa tv giống Trần dần bịa đặt chuyện có nói ko chuyện ko nói có	Bình luận này không chứa từ thô tục nào nhưng có nội dung chế giễu, quấy rối nhắm đến một người cụ thể "Trần dần" và cả nhóm người "người Việt ở Mỹ" nên được gán nhãn hate
Bài này như lồn	Bình luận này chứa từ ngữ thô tục "lồn" và có nội dung quấy rối nhắm đến một đối tượng cụ thể "bài này"
Những chú dâm loạn viên đâu tui muốn thấy cái mõm chó của mấy bạn	Bình luận này chứa từ ngữ thô tục "dâm loạn", "mõm chó" và có nội dung quấy rối nhắm đến một nhóm đối tượng cụ thể "mấy bạn"
Nhìn nhà tiên tri vủ trụ trần dần giống con chó quá trời luôn ta ơi .	Bình luận này chứa từ ngữ thô tục "chó" và có nội dung quấy rối nhắm đến một người cụ thể "trần dần"
loại đàn bà độc ác	**Bình luận này chứa từ ngữ tiêu cực "độc ác" và có nội dung quấy rối, miệt thị nhưng không chứa từ ngữ nhắm đến một người hay nhóm cụ thể nên gán nhãn hate là chưa hợp lý**
Một số lớn người VN phải xếp vào một trong những dân tộc tàn độc nhất thế giới ,mục tiêu là chủ nghĩa cá nhân là trên hết.	Bình luận này chứa từ ngữ tiêu cực "tàn độc" và có nội dung quấy rối nhắm đến một nhóm đối tượng cụ thể "người VN"
Lũ chó chết. Đkm loạn mẹ nó rồi. Xh như lồn	Bình luận này chứa nhiều từ ngữ tiêu cực "chó chết", "Đkm", "mẹ", "lồn" và có nội dung quấy rối, miệt thị và chứa từ ngữ nhắm đến một

	người hay nhóm cụ thể nên gán nhãn hate là hợp lý
Ca sĩ k-icm đéo gì 😳 ca sĩ câm	Bình luận này chứa từ ngữ tiêu cực "đéo", "câm" và có nội dung quấy rối nhắm đến một người cụ thể "Ca sĩ k-icm"
Vân Thuỳ xàm vl :))	Bình luận này chứa từ ngữ thô tục "vl" và có nội dung quấy rối nhắm đến một người cụ thể "Vân Thuỷ"
Con này là bê đê bong lộ xấu gớm... xấu xúc phạm ng nhìn...	Bình luận này mang nội dung quấy rối, miệt thị, phân biệt và nhắm đến một người cụ thể "con này" nên được gán nhãn hate
Chúng bắt đầu cho báo chí đổ tại người dân dọn nhà đón Tết gây ra ô nhiễm kìa	Bình luận này mang ý nghĩa tiêu cực và nhắm tới một đối tượng cụ thể

Bảng 2: Ví dụ một số trường hợp cụ thể của các nhãn

Thực tế, nhiều ý kiến trong bộ dữ liệu được viết dưới hình thức không chính thức. Các ý kiến thường chứa viết tắt như M.n (mọi người), mik (mình) trong và Dm, cùng với những ngôn ngữ lóng như "chịch", "cái lol". Ngoài ra, các ý kiến thường mang ý nghĩa ẩn dụ thay vì ý nghĩa rõ ràng. Ví dụ, từ "lũ quan ngại" thường được sử dụng bởi nhiều người dùng Facebook Việt Nam trên nền tảng truyền thông xã hội để đề cập đến một nhóm người luôn suy nghĩ tiêu cực và đăng nội dung tiêu cực.

Mỗi bình luận trong tập ngữ liệu bao gồm 1 hoặc 2 câu ngắn, có thể chứa các biểu tượng cảm xúc. Tuy nhiên các biểu tượng cảm xúc này không mang ý nghĩa quá nhiều cho việc xác định nhãn của bình luận. Ví dụ một số mẫu câu có biểu tượng cảm xúc như sau:

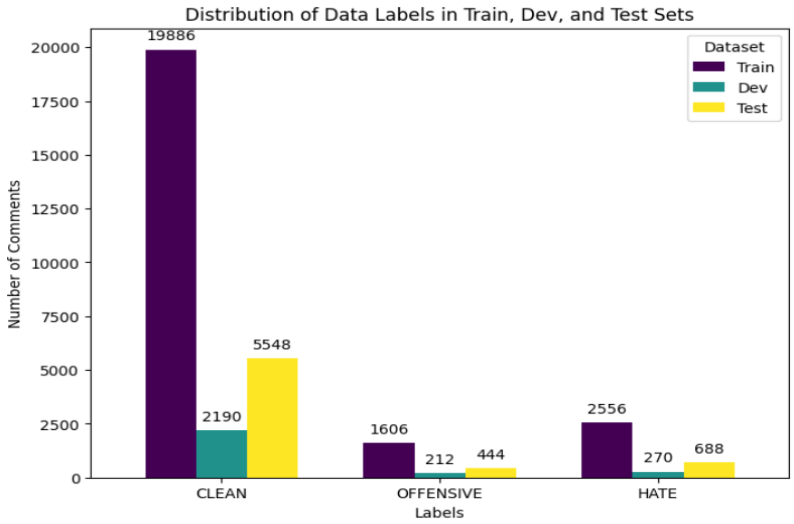
Comments	**Label**	**Description**
Cũng được đấy chứ a Độ ộ ộ ộ ộ ộ....😂 😂 😂	CLEAN	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai

Mang tiếng fan cứng đăng nhưng đéo đc duyệt 😳	OFFENSIVE	Bình luận này chứa từ thô tục "đéo". Tuy nhiên, nó không chứa bất kỳ từ ngữ nào nhằm vào một người hoặc một nhóm.
Ca sĩ k-icm đéo gì 😳 ca sĩ câm	HATE	Bình luận này chứa cụm từ thô tục như "đéo" và cụm từ "ca sĩ câm" mang ý nghĩa chê bai, tấn công một cá nhân.
🤣🤣🤣🤣🤣	CLEAN	Bình luận này hoàn toàn sạch, không chứa nội dung quấy rối hoặc ngôn ngữ thô tục, và không tấn công bất kỳ ai.

Bảng 3: Một số mẫu câu có biểu tượng cảm xúc

Từ bảng 3 ta thấy các nhãn được gán cho bình luận liên quan chủ yếu đến các từ hoặc các cụm từ có trong bình luận. Các biểu tượng cảm xúc không có giá trị quá lớn trong việc phân loại nhãn.

Bộ ngữ liệu bộ dữ liệu thành ba phần: tập huấn luyện (train), phát triển (dev) và kiểm thử (test), tương ứng với tỷ lệ 7-1-2. Phân phối ngữ liệu trên ba nhãn của các tập dữ liệu đó như sau:



Hình 1: Phân phối ngữ liệu trên ba nhãn của các tập dữ liệu train, dev và test

Theo biểu đồ, ta thấy phân phối nhãn dữ liệu trên các tập huấn luyện, phát triển và kiểm thử là giống nhau, và dữ liệu nghiêng về nhãn CLEAN.

CHƯƠNG 3: Phương pháp

3.1. Sử dụng các mô hình máy học

3.1.1 Phương pháp rút trích đặc trưng

3.1.1.1 TF-IDF Embedding

Trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), việc chuyển đổi văn bản thô thành các biểu diễn số học có ý nghĩa là một bước quan trọng. Kỹ thuật TF-IDF (Term Frequency-Inverse Document Frequency) là một trong những phương pháp phổ biến và hiệu quả để thực hiện điều này. TF-IDF là một chỉ số thống kê phản ánh mức độ quan trọng của một từ trong một văn bản cụ thể so với toàn bộ tập văn bản. Nó được tính bằng cách nhân hai thành phần:

- TF (Term Frequency): Tần suất xuất hiện của từ trong văn bản.
  TF(t, d) = số lần từ t xuất hiện trong văn bản d / tổng số từ trong văn bản d.
- IDF (Inverse Document Frequency): Nghịch đảo tần suất xuất hiện của từ trong toàn bộ tập văn bản.
  IDF(t) = log(Tổng số văn bản trong tập D / Số văn bản chứa từ t).

Công thức tính TF-IDF của một từ t trong văn bản d (hay câu bình luận đề tài) :

TF-IDF(t, d) = TF(t, d) * IDF(t)

Ưu điểm:

- Đơn giản và dễ hiểu: TF-IDF có khái niệm và cách tính toán đơn giản, dễ dàng triển khai và giải thích.
- Hiệu quả: TF-IDF đã được chứng minh là hiệu quả trong nhiều tác vụ NLP như phân loại văn bản, tìm kiếm thông tin, và phân cụm văn bản.
- Khả năng mở rộng: TF-IDF có thể được áp dụng cho các tập dữ liệu lớn mà không gặp nhiều khó khăn về tính toán.

Nhược điểm:

- Không tính đến ngữ cảnh: TF-IDF chỉ xem xét tần suất xuất hiện của từ mà không quan tâm đến vị trí của từ trong câu hoặc ngữ cảnh xung quanh.
- Không phân biệt được các từ đồng nghĩa: TF-IDF coi các từ khác nhau là độc lập với nhau, không thể phân biệt được các từ có nghĩa tương tự nhau.

VD : 'người ta bị lường gạt 1 tỷ mấy còn mà đằng này hơn trăm triệu thôi mà'

->(0, 6691) 0.27523577745916666

(0, 6769) 0.31121089074446756

(0, 8217) 0.34135503849605897

(0, 7087) 0.29835338683610957

(0, 2432) 0.33715974511868263

(0, 3791) 0.37267673417560787

(0, 1481) 0.17903629192168138

(0, 6192) 0.20916111262565576

(0, 4533) 0.17111157197273213

(0, 782) 0.19907901592365987

(0, 6424) 0.18601382102598102

(0, 4131) 0.291538591745277

(0, 2885) 0.2107938623122215

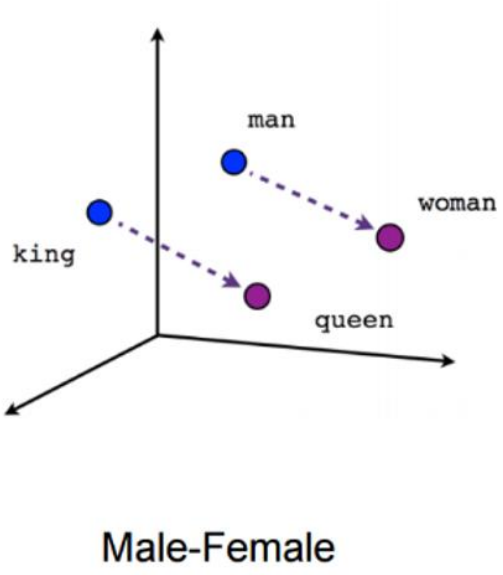
(0, 4930) 0.1537244304062

(0, 4241) 0.19213345917686675

Bộ từ điển có 8428 từ nên vector sẽ biểu diễn số chiều là 8428 và dưới dạng kết quả ma trân thưa như trên biểu diễn các giá trị các mỗi index các tử tương ứng của chúng

3.1.1.2 Word2vec Embedding

Word2Vec là một mô hình nhúng từ (word embedding) được giới thiệu bởi Google vào năm 2013. Mục tiêu của Word2Vec là chuyển đổi các từ thành các vector số có ý nghĩa, sao cho các từ có ngữ nghĩa tương tự nhau sẽ có các vector gần nhau trong không gian vector. Điều này cho phép máy tính hiểu và xử lý ngôn ngữ một cách hiệu quả hơn.



Hình 3. 1 Biểu diễn trực quan từ được nhúng Word2vec

Word2Vec sử dụng một trong hai kiến trúc mạng nơ-ron nông (shallow neural network) để học các vector biểu diễn từ.mô hình sẽ học cách ánh xạ mỗi từ thành một vector sao cho các từ có ngữ cảnh tương tự nhau sẽ có các vector gần nhau.

Word2Vec có thể học được các vector biểu diễn từ có ý nghĩa từ một lượng lớn dữ liệu văn bản không cần gán nhãn (unsupervised learning). Nắm bắt được ngữ nghĩa ví dụ như từ đồng nghĩa, trái nghĩa, hoặc các từ có quan hệ về mặt khái niệm (ví dụ: "Hà Nội" - "Việt Nam" + "Nhật Bản" = "Tokyo") nhưng mà Word2Vec gán một vector duy nhất cho mỗi từ, không phân biệt được các nghĩa khác nhau của từ trong các ngữ cảnh khác nhau.Word2Vec embeddings phụ thuộc rất nhiều vào lượng và chất lượng của dữ liệu huấn luyện.

3.1.2 Các mô hình máy học sử dụng

3.1.2.1 Logistic regression

Regression logistic là một phương pháp phân tích thống kê khác được Machine learning mượn. Nó được sử dụng khi biến phụ thuộc của chúng ta là lưỡng phân

hoặc nhị phân. Nó chỉ có nghĩa là một biến chỉ có 2 đầu ra, ví dụ: Một người có sống sót sau tai nạn này hay không, Học sinh có vượt qua kỳ thi này hay không. Kết quả có thể là có hoặc không (2 đầu ra 0,1). Kỹ thuật Regression này tương tự như Regression Linear và có thể được sử dụng để dự đoán Xác suất cho các bài toán phân loại.

Ta có bộ số X có D mẫu dữ liệu,và bộ số y là các giá trị dự đoán có giá trị 0,1

$X^{(i)} = (x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, .., x_n^{(i)})$ (Bộ giá trị ở dòng thứ i ,Có n thuộc tính x)

Ta có hàm biểu diễn quan hệ giữa X và y theo bộ tham số w $= (w_0, w_1, w_2, .., w_n)$

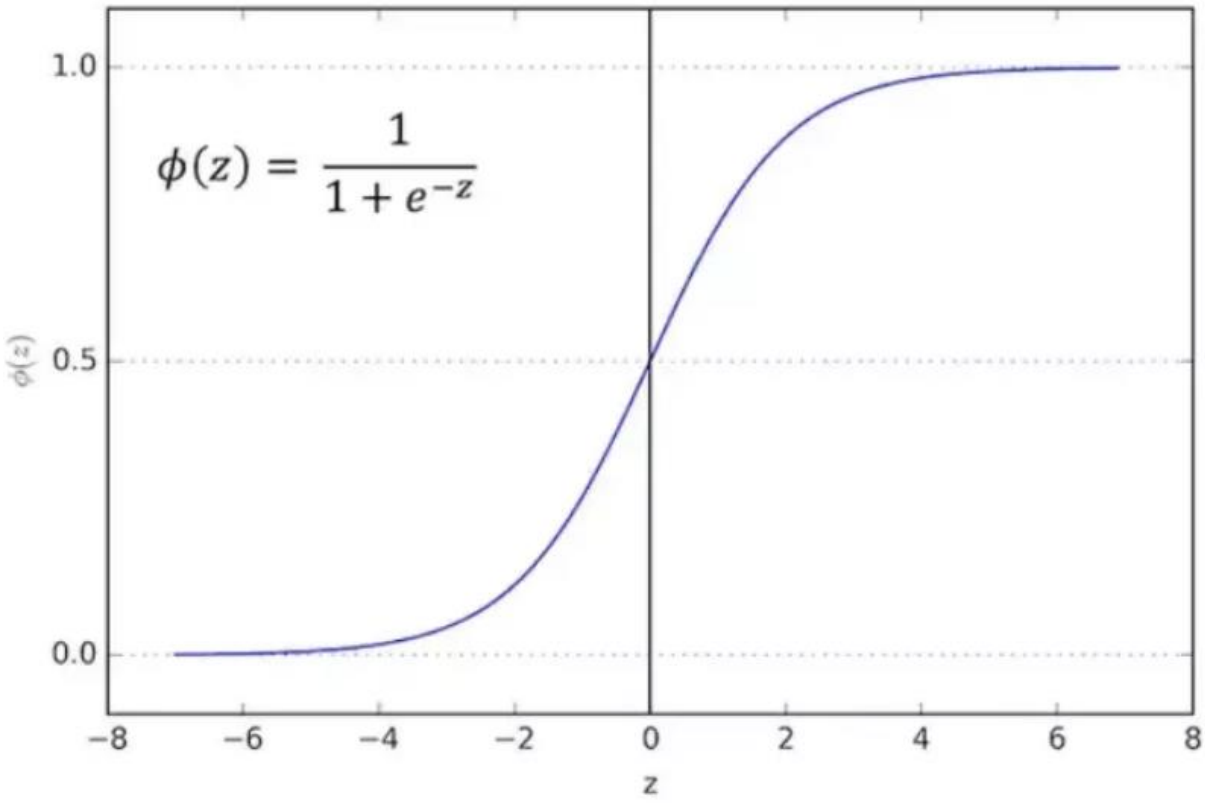
Khác với linear regression ,giá trị đầu ra y là 1 số thực bất kì nên ta có thể biểu diễn mối quan hệ giữa X,y trong linear regression là:

$Y = w^T X$

$Y^{(i)} = w_0 + w_1 * x_1^{(i)} + w_2 * x_2^{(i)} + .. + w_n * x_n^{(i)}$

Trở lại với logicstic ,giá trị đầu ra của y là có giá trị 0 ,hoặc 1.Mặt khác ta cũng phải biểu diễn sự liên quan giữa các giá trị của X để suy ra được giá trị y (ta dùng hàm tuyến tính lấy trong linear regression).

Hàm mà chuyển các giá trị liên tục trên tập R về khoảng [0,1] mà luôn có đạo hàm với mọi giá trị trên tập R và đơn giản chính là hàm Sigmoid :



Hàm Sigmoid

Như vậy y sẽ được biểu diễn bằng hàm sigmoid ,kết quả cuối cùng của y sẽ là 1 số trong khoảng [0,1] ,ta sẽ làm tròn giá trị này để đưa ra kết quả cuối cùng .

$z = w^T X$

$z^{(i)} = w_0 + w_1 * x_1^{(i)} + w_2 * x_2^{(i)} + .. + w_n * x_n^{(i)}$

(Dòng thứ i)

Hàm z ta lấy như trên bởi vì ta còn cần phải biểu diễn mối quan hệ giữa y với các đại lượng trong X ,điều này khá giống trong linear regression

Như vậy ta sẽ có hàm dự đoán là:

$$\hat{y}_i = \Phi\left(w_0 + w_1 * x_1^{(i)} + .. + w_n * x_n^{(i)}\right) = \frac{1}{1 + e^{w_0 + w_1 * x_1^{(i)} + .. + w_n * x_n^{(i)}}}$$

Ký hiệu $z^{(i)} = z_i = f(w^T x_i) = \hat{y}_i$

Ta có thể giả sử rằng xác suất để một điểm dữ liệu X rơi vào class 1 là $\boldsymbol{f(w^T X)}$(**f** là hàm sigmoid) và rơi vào class 0 là $1 - (\boldsymbol{w^T X})$. Với mô hình được giả sử như vậy, với các điểm dữ liệu training (đã biết đầu ra y), ta có thể viết như sau:

$$P(y_i = 1 | x_i; w) = f(\boldsymbol{w^T x_i})$$

$$P(y_i = 0 | x_i; w) = 1 - f(\boldsymbol{w^T x_i})$$

Viết gộp lại hai biểu thức bên trên ta có:

$$P(y_i | x_i; w) = z_i^{y_i}(1 - z_i)^{1 - y_i}$$

Chúng ta muốn mô hình gần với dữ liệu đã cho nhất, tức xác suất này đạt giá trị cao nhất hay là cần tìm max phương trình dưới đây :

$$P(y | X; w) = \prod_{i=1}^{D} P(y_i | x_i; w) = \prod_{i=1}^{D} z_i^{y_i} (1 - z_i)^{1 - y_i}$$

Ta lấy log và thêm dấu – trước phương trình ta sẽ cần tìm giá trị nhỏ nhất cho hàm mất mát như dưới đây, thay $z_i = \hat{y}_i$:

$$L = -(y_i * \log(\hat{y_i}) + (1 - y_i) * \log(1 - \hat{y_i}))$$

Đến đây ta dùng gradient descent hay các biến thể của nó để tìm ra bộ tham số W tối ưu làm cho L nhỏ nhất có thể

Đấy là trên 1 điểm dữ liệu, còn trên toàn bộ dữ liệu:

$$\frac{dL}{dw_1} = \sum_{i=1}^{D} x_1^{(i)} * (\hat{y}_i - y_i)$$

$\frac{dL}{dw_2} = \sum_{i=1}^{D} x_2^{(i)} * (\hat{y}_i - y_i)$

……

$\frac{dL}{dw_n} = \sum_{i=1}^{D} x_n^{(i)} * (\hat{y}_i - y_i)$

Tổng quát hóa lên ta được :

$\frac{dL}{dw_j} = \sum_{i=1}^{D} \frac{dL}{dw_i} = x_j^{(i)} * (\hat{y}_i - y_i)$ | $x_0^0 = 1$,0<=j<=n,D là số dòng dữ liệu

($x_0^0 = 1$, 1<=i<=D)

Đạo hàm thì ta cũng đã tính được ,đến đây ta chỉ cần cập nhập giá trị bộ tham số W theo thuật toán gradient descent đến khi nào hàm độ lớn giữa L của 2 lần cập nhập liên tiếp giảm không đáng kể thì dừng ,ngoài ra còn có thế có các điều kiện dừng khác .

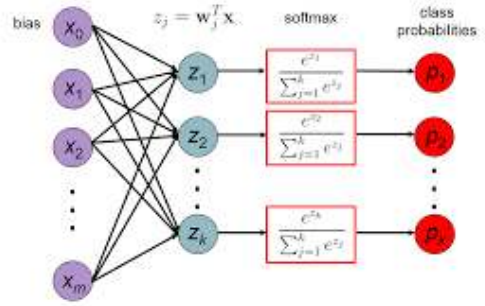
$$w_{t+1} = w_t - learning\_rate * L'(w_t)$$

Sau khi tìm được bộ tham số w tối ưu ta tìm phương trình phần cách các điểm dữ liệu thành 2 lớp bằng 1 tham số t,thường là t=0.5

Trong trường hợp tổng quát t bất kì,$\hat{y}_i > t$ thì :

$$w_0 + w_1 * x_1^{(i)} + .. + w_n * x_n^{(i)} > -Ln(\frac{1}{t} - 1)$$

Vậy trong bài toán phân ra nhiều lớp khác nhau hay 3 lớp như trong đề tài này thì ta thay thế hàm kích hoạt là sigmoid sang hàm softmax để chuyển thành Multinomial Logistic Regression (Softmax regression) để biểu diễn kết quả tuyến tính đầu ra thành phân phối xác suất các lớp khác nhau lớp nào có xác suất cao nhất thì ta phân loại về lớp đó.



Hình 3. 2 Softmax regression

Trực quan kết quả biểu diễn đầu vào và đầu ra:

Câu "vâng là dân v là có thật v " biểu diễn là ma trận thưa qua tf-idf như sau

(0, 7366) 0.6003659705019697

(0, 6475) 0.4003382021509419

(0, 3686) 0.4935028246339071

(0, 1887) 0.41424337826622154

(0, 1483) 0.25327339181746417

Đầu vào có kích thước là 1 x 8428 là kích thước vector đặc trưng biểu diễn câu dùng tfidf, qua lớp softmax cho bài toán phân thành 3 nhãn câu ta được kết quả:

Phân phối xác suất: [[0.35755791 0.15523211 0.48720998]]

Nhãn dự đoán: 2 do 0.487 là phân phối xác xuất cao nhất trong 3 lớp

Các tham số đầu vào mô hình Logistic regression sklearn :

+ penalty: Kiểu regularlization (điều chuẩn) được sử dụng. Các giá trị có thể là:

- 'l1': Lasso (L1) regularization
- 'l2': Ridge (L2) regularization
- 'elasticnet': Kết hợp cả L1 và L2
- 'none': Không sử dụng regularization

+ C: Nghịch đảo của độ mạnh regularlization (strength of regularization). Giá trị C càng nhỏ thì regularlization càng mạnh.

+ solver: Thuật toán tối ưu được sử dụng để tìm nghiệm của bài toán Logistic Regression.

+ max_iter: Số lần lặp tối đa cho thuật toán tối ưu…..

+ multi_class: Loại bài toán phân loại:

- 'auto': Tự động chọn 'ovr' nếu số lượng lớp là 2, và 'multinomial' nếu số lượng lớp lớn hơn 2.
- 'ovr': One-vs-Rest (OvR)
- 'multinomial': Multinomial Logistic Regression (Softmax Regression)

3.1.2.2 Random forest

Random Forest là một thuật toán học máy tập hợp (ensemble learning) thuộc nhóm các phương pháp bagging. Nó xây dựng nhiều cây quyết định (decision trees) trên các tập con mẫu ngẫu nhiên của dữ liệu huấn luyện và kết hợp dự đoán của chúng để đưa ra quyết định cuối cùng. Random Forest được sử dụng rộng rãi trong các bài toán phân loại và hồi quy nhờ tính hiệu quả và khả năng chống overfitting (quá khớp) tốt.

Nguyên lý hoạt động

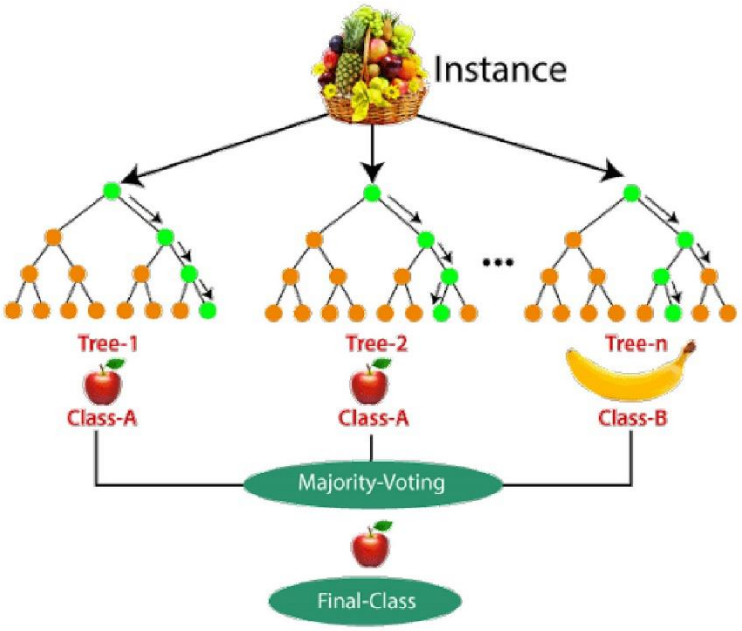
- Bagging (Bootstrap Aggregating): Tạo ra nhiều tập con mẫu ngẫu nhiên (bootstrap samples) từ tập dữ liệu huấn luyện. Mỗi tập con mẫu có thể chứa các điểm dữ liệu trùng lặp.
- Xây dựng cây quyết định: Với mỗi tập con mẫu, xây dựng một cây quyết định. Trong quá trình xây dựng cây, tại mỗi nút, chỉ một tập hợp ngẫu nhiên các đặc trưng (features) được xem xét để tìm ra phép chia tốt nhất.
- Tổng hợp dự đoán: Khi dự đoán một mẫu mới, mỗi cây quyết định sẽ đưa ra một dự đoán. Random Forest sẽ tổng hợp các dự đoán này bằng cách lấy đa số phiếu (majority voting) trong trường hợp phân loại

Random Forest là một thuật toán học máy tập hợp (ensemble learning) thuộc nhóm các phương pháp bagging. Nó xây dựng nhiều cây quyết định (decision trees) trên các tập con mẫu ngẫu nhiên của dữ liệu huấn luyện và kết hợp dự đoán của chúng để đưa ra quyết định cuối cùng. Random Forest được sử dụng rộng rãi trong các bài toán phân loại và hồi quy nhờ tính hiệu quả và khả năng chống overfitting (quá khớp) tốt.

Bagging (Bootstrap Aggregating): Tạo ra nhiều tập con mẫu ngẫu nhiên (bootstrap samples) từ tập dữ liệu huấn luyện. Mỗi tập con mẫu có thể chứa các điểm dữ liệu trùng lặp.

Xây dựng cây quyết định: Với mỗi tập con mẫu, xây dựng một cây quyết định. Trong quá trình xây dựng cây, tại mỗi nút, chỉ một tập hợp ngẫu nhiên các đặc trưng (features) được xem xét để tìm ra phép chia tốt nhất.

Tổng hợp dự đoán: Khi dự đoán một mẫu mới, mỗi cây quyết định sẽ đưa ra một dự đoán. Random Forest sẽ tổng hợp các dự đoán này bằng cách lấy đa số phiếu (majority voting) trong trường hợp phân loại, hoặc trung bình cộng trong trường hợp hồi quy.



Hình 3. 3 Minh họa random forest

Ưu điểm:

- Hiệu suất cao: Thường cho kết quả tốt hơn so với các cây quyết định đơn lẻ.
- Chống overfitting: Nhờ tính ngẫu nhiên trong việc chọn mẫu và đặc trưng, Random Forest có khả năng chống overfitting tốt.
- Xử lý được dữ liệu có nhiều đặc trưng: Có thể lựa chọn các đặc trưng quan trọng và loại bỏ các đặc trưng không cần thiết.
- Ước lượng được tầm quan trọng của đặc trưng: Giúp hiểu rõ đặc trưng nào ảnh hưởng nhiều nhất đến kết quả dự đoán.

Nhược điểm:

- Khó giải thích: Mô hình phức tạp hơn so với cây quyết định đơn lẻ, do đó khó giải thích kết quả dự đoán.
- Tốn kém về mặt tính toán: Việc xây dựng nhiều cây quyết định có thể tốn nhiều thời gian và tài nguyên.

Các Tham số Tuning (Siêu tham số) Quan trọng

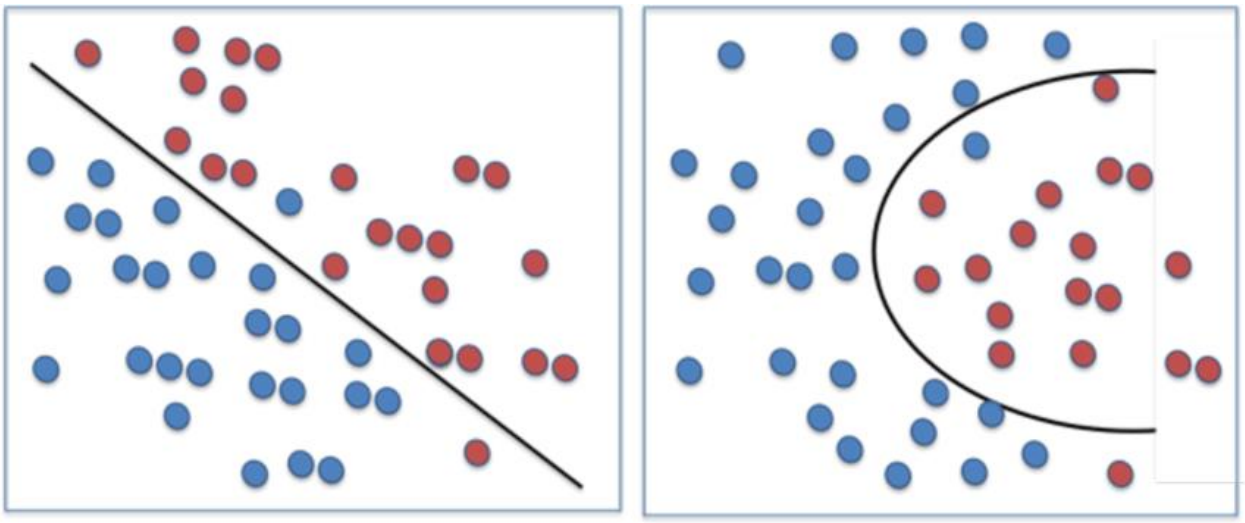
- n_estimators: Số lượng cây quyết định trong rừng. Tăng giá trị này thường làm tăng hiệu suất, nhưng cũng làm tăng thời gian huấn luyện và có thể dẫn đến overfitting.
- max_depth: Độ sâu tối đa của mỗi cây quyết định. Giá trị lớn hơn cho phép cây học các mẫu phức tạp hơn, nhưng cũng dễ dẫn đến overfitting.
- min_samples_split: Số lượng mẫu tối thiểu cần có để chia một nút trong cây. Giá trị lớn hơn giúp tránh overfitting, nhưng có thể làm giảm hiệu suất nếu quá lớn.
- min_samples_leaf: Số lượng mẫu tối thiểu cần có ở một lá. Tương tự như min_samples_split.
- max_features: Số lượng đặc trưng được xem xét tại mỗi nút khi chia cây. Giá trị nhỏ hơn giúp tăng tính ngẫu nhiên và giảm overfitting.

3.1.2.3 Support vector machine (SVM)

Support Vector Machine (SVM), hay còn gọi là Máy Vectơ Hỗ Trợ, là một thuật toán học máy mạnh mẽ và phổ biến được sử dụng rộng rãi trong các bài toán phân loại và hồi quy. SVM hoạt động dựa trên nguyên tắc tìm kiếm siêu phẳng (hyperplane) tối ưu để phân chia các lớp dữ liệu với margin (khoảng cách) lớn nhất.

SVM hoạt động theo các bước sau:

1. Biểu diễn dữ liệu: Mỗi điểm dữ liệu được biểu diễn dưới dạng một vector trong không gian nhiều chiều.
2. Tìm kiếm siêu phẳng tối ưu: SVM tìm kiếm siêu phẳng có thể phân chia các lớp dữ liệu với margin lớn nhất. Margin là khoảng cách giữa siêu phẳng và các điểm dữ liệu gần nhất (gọi là support vectors).
3. Phân loại: Đối với một điểm dữ liệu mới, SVM sẽ dự đoán lớp của nó dựa trên vị trí của điểm đó so với siêu phẳng.



Hình 3. 4 Ví dụ về sử dụng SMV trong bài toán phân loại

Ưu điểm:

- Hiệu suất cao: SVM thường cho kết quả tốt trên nhiều bài toán phân loại khác nhau.
- Hiệu quả với dữ liệu có số chiều cao: SVM có thể hoạt động tốt với dữ liệu có số lượng đặc trưng lớn.
- Chống quá khớp (Overfitting): Nhờ vào việc tối đa hóa margin, SVM có khả năng chống overfitting tốt.

Nhược điểm

- Nhạy cảm với nhiễu: SVM có thể nhạy cảm với nhiễu trong dữ liệu.
- Khó giải thích: Mô hình SVM có thể khó giải thích, đặc biệt khi sử dụng kernel trick.
- Tốn kém về mặt tính toán: Việc huấn luyện SVM có thể tốn nhiều thời gian và tài nguyên, đặc biệt với dữ liệu lớn.

Các tham số quan trọng

- C: Tham số điều chuẩn (regularization parameter). Giá trị C lớn hơn sẽ dẫn đến mô hình phức tạp hơn, dễ bị overfitting.
- kernel: Loại hàm kernel được sử dụng ('linear', 'poly', 'rbf', 'sigmoid').
- gamma: Tham số của hàm kernel RBF, kiểm soát ảnh hưởng của mỗi điểm dữ liệu.

3.2. Sử dụng mạng noron học sâu

FastText là một mô hình học máy được phát triển bởi Facebook AI Research (FAIR) cho các bài toán phân loại văn bản và học từ biểu diễn (word representation). FastText nổi bật nhờ khả năng xử lý nhanh và hiệu quả, đặc biệt với các ngôn ngữ có cấu trúc phức tạp và từ vựng phong phú như tiếng Việt.

Nguyên lý hoạt động:

1. Biểu diễn từ (Word Representation): Thay vì chỉ sử dụng các từ riêng lẻ, FastText chia nhỏ từ thành các n-grams. Điều này giúp mô hình hiểu rõ hơn về cấu trúc từ và các từ chưa xuất hiện trong tập huấn luyện.
2. **Mô hình hóa văn bản (Text Representation)**: Mỗi văn bản được biểu diễn bằng cách lấy trung bình các vector từ của các từ (bao gồm cả n-grams) trong văn bản đó. Quá trình này giúp giảm chiều dữ liệu và tăng tốc độ huấn luyện.
3. Huấn luyện mô hình (Model Training): FastText sử dụng mạng neural tuyến tính (linear neural network) để huấn luyện trên các vector từ. Điều này giúp tăng tốc độ huấn luyện so với các mô hình phức tạp hơn như RNN hoặc CNN.

Ưu điểm:

- Tốc độ cao: FastText có khả năng huấn luyện rất nhanh, đặc biệt hữu ích cho các tập dữ liệu lớn.
- Xử lý từ mới: Nhờ việc sử dụng n-grams, FastText có thể xử lý các từ mới hoặc hiếm mà không cần tái huấn luyện mô hình.
- Hiệu quả với các ngôn ngữ phức tạp: FastText hoạt động tốt với các ngôn ngữ có từ vựng phong phú và cấu trúc ngữ pháp phức tạp.
- Đơn giản và dễ triển khai: Mô hình đơn giản hơn nhiều so với các mô hình neural network phức tạp, dễ dàng triển khai và tích hợp vào các hệ thống hiện có.

Nhược điểm:

- Hiệu suất có thể thấp hơn trên các nhiệm vụ phức tạp: Với các nhiệm vụ yêu cầu hiểu ngữ cảnh sâu rộng, các mô hình phức tạp hơn như BERT có thể cho kết quả tốt hơn.
- Thiếu khả năng giải thích: Mặc dù FastText nhanh và hiệu quả, mô hình này có thể thiếu khả năng giải thích so với các phương pháp khác.

Các tham số tuning (siêu tham số) quan trọng:

- learning_rate: Tốc độ học của mô hình, ảnh hưởng đến việc cập nhật trọng số trong quá trình huấn luyện.
- epoch: Số lần huấn luyện qua toàn bộ tập dữ liệu. Tăng số epoch có thể cải thiện hiệu suất nhưng cũng có thể dẫn đến overfitting.
- word_ngrams: Số lượng n-grams từ được sử dụng. Sử dụng nhiều n-grams có thể cải thiện khả năng hiểu từ ngữ cảnh nhưng cũng làm tăng độ phức tạp tính toán.
- vector_dim: Kích thước của vector từ. Kích thước lớn hơn có thể giúp mô hình học được các biểu diễn từ tốt hơn nhưng cũng tốn kém về mặt tính toán.

3.3. Đánh giá mô hình

Precision

Precision đo lường khả năng mô hình dự đoán đúng nhãn positive trong số tất cả các mẫu được dự đoán là positive. Công thức tính precision:

Precision = TP / (TP + FP)

Recall Recall đo lường khả năng mô hình dự đoán đúng nhãn positive trong số tất cả các mẫu positive thực sự. Công thức tính recall :

Recall = TP / (TP + FN)

F1-score F1-score là trung bình điều hòa (harmonic mean) của precision và recall. Nó là một chỉ số tổng hợp hữu ích khi bạn muốn cân bằng giữa precision và recall. Công thức tính F1-score:

F1-score = 2 * (Precision * Recall) / (Precision + Recall)

Confusion matrix là một bảng tóm tắt số lượng dự đoán đúng và sai của mô hình cho từng lớp. Nó cung cấp cái nhìn tổng quan về hiệu suất của mô hình và giúp xác định các loại lỗi mà mô hình thường mắc phải.

	Dự đoán postive	Dự đoán Negative
Thực tế True	TP	TN
Thực tế False	FP	FN

CHƯƠNG 4: Cài đặt và thử nghiệm

4.1. Cài đặt

4.1.1. Sử dụng các mô hình máy học

Trong để tài này, chúng em sử dụng hai phương pháp trích xuất đặc trưng văn bản khác nhau để đánh giá hiệu suất của các mô hình học máy trong bài toán phân loại:

- TF-IDF (Term Frequency-Inverse Document Frequency): Phương pháp này đánh giá mức độ quan trọng của một từ trong một văn bản dựa trên tần suất xuất hiện của từ đó trong văn bản và trong toàn bộ tập dữ liệu. Các từ xuất hiện nhiều trong một văn bản nhưng hiếm trong toàn bộ tập dữ liệu sẽ được đánh giá cao hơn.
- Word2vec Embeddings: Mô hình này học cách biểu diễn mỗi từ dưới dạng một vector số, sao cho các từ có ý nghĩa tương tự nhau sẽ có các vector gần nhau trong không gian vector. Chúng em sử dụng Word2Vec để chuyển đổi các từ trong văn bản thành các vector số, sau đó tính trung bình các vector này để đại diện cho cả câu

Mô hình máy học và tối ưu siêu tham số:

- Logistic Regression (LR): Một mô hình tuyến tính đơn giản nhưng hiệu quả, thường được sử dụng làm baseline cho các bài toán phân loại.
- Random Forest (RF): Một mô hình tập hợp (ensemble) gồm nhiều cây quyết định, có khả năng học các quy tắc phức tạp và chống quá khớp (overfitting) tốt.
- Support Vector Machine (SVM): Một mô hình mạnh mẽ có thể tìm ra các siêu phẳng tối ưu để phân chia các lớp dữ liệu.

Để tối ưu hóa hiệu suất của từng mô hình, chúng em sử dụng kỹ thuật Grid Search + k fold validation để tìm kiếm các siêu tham số tốt nhất. Các siêu tham số được thử nghiệm cho từng mô hình như sau:

LR:

- C: Độ mạnh của regularization (0.001, 0.01, 0.1, 1, 10, 50)

- penalty: Loại regularization ('l2')

RF:

- n_estimators: Số lượng cây quyết định (10, 50, 100)
- max_depth: Độ sâu tối đa của cây (None, 5, 10, 15, 20)
- min_samples_split: Số lượng mẫu tối thiểu cần có để chia một nút (2, 5, 10)

SVM:

- C: Độ mạnh của regularization (0.1, 0.5, 1, 5, 10)
- kernel: Loại hàm kernel ('linear', 'rbf')
- gamma: Hệ số kernel cho 'rbf' ('scale', 'auto')

4.1.2. Sử dụng mô hình mạng học sâu

FastText là một công cụ huấn luyện và phân loại văn bản nhanh chóng và hiệu quả. Đây là mô hình tuyến tính nhưng lại có khả năng học các đặc trưng từ n-grams, giúp cải thiện độ chính xác cho các bài toán phân loại văn bản.

Để tối ưu hóa hiệu suất của mô hình FastText, chúng em sử dụng kỹ thuật Grid Search và k-fold validation để tìm kiếm các siêu tham số tốt nhất. Các siêu tham số được thử nghiệm bao gồm:

- `learning_rate`: Tốc độ học của mô hình (0.01, 0.05, 0.1, 0.5, 1.0)
- `epoch`: Số lần lặp lại qua toàn bộ tập dữ liệu (5, 10, 20, 25)
- `wordNgrams`: Số lượng n-grams từ được sử dụng (1, 2, 3)
- `minCount`: Số lần xuất hiện tối thiểu của một từ để nó được xem xét trong mô hình (1, 2, 5)
- `vector_dim`: Kích thước của vector từ (50, 100, 150, 200)

4.2. Kết quả đạt được

4.2.1. Sử dụng mô hình máy học dùng tf-idf làm vector đặc trưng

```
Mô hình lr:
  - Best params: {'clf__C': 10, 'clf__penalty': 'l2'}
  - Best score: 0.8633563886838995
Mô hình rf:
  - Best params: {'clf__max_depth': None, 'clf__min_samples_split': 5, 'clf__n_estimators': 100}
  - Best score: 0.8602792753884456
Mô hình svm:
  - Best params: {'clf__C': 1, 'clf__gamma': 'scale', 'clf__kernel': 'linear'}
  - Best score: 0.8685961310415458
```

Hình 4. 1 Các siêu tham số tốt nhất sau khi dùng gridsearch các mô hình

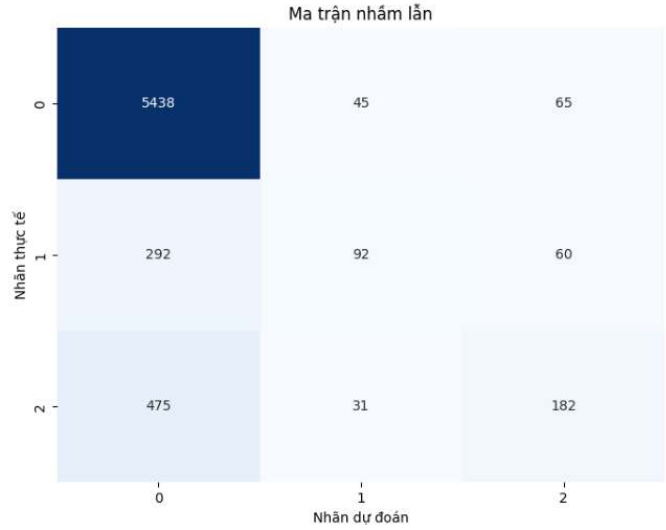
Báo cáo phân loại lr:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	5548
1	0.47	0.25	0.33	444
2	0.57	0.41	0.48	688
accuracy			0.86	6680
macro avg	0.64	0.54	0.58	6680
weighted avg	0.83	0.86	0.84	6680

Hình 4. 2 Bảng kết quả tổng hợp mô hình logistic regression

Báo cáo phân loại rf:

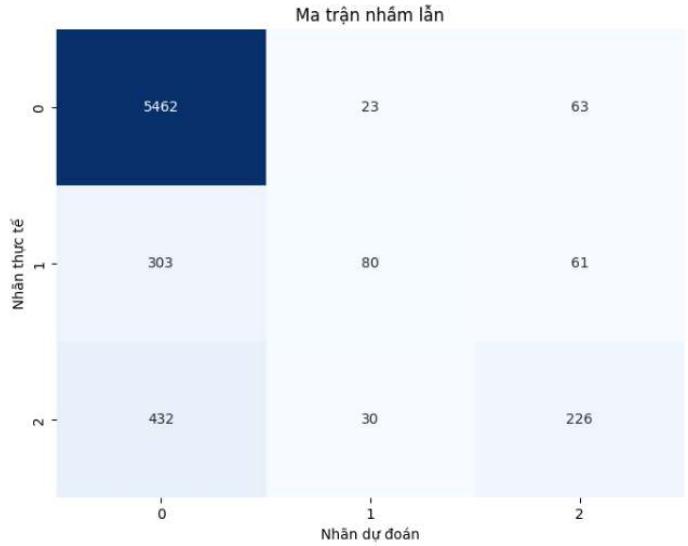
	precision	recall	f1-score	support
0	0.88	0.98	0.93	5548
1	0.55	0.21	0.30	444
2	0.59	0.26	0.37	688
accuracy			0.86	6680
macro avg	0.67	0.48	0.53	6680
weighted avg	0.83	0.86	0.83	6680



Hình 4. 3 Bảng kết quả tổng hợp mô hình random forest

Báo cáo phân loại svm:

	precision	recall	f1-score	support
0	0.88	0.98	0.93	5548
1	0.60	0.18	0.28	444
2	0.65	0.33	0.44	688
accuracy			0.86	6680
macro avg	0.71	0.50	0.55	6680
weighted avg	0.84	0.86	0.84	6680



Hình 4. 4 Bảng kết quả tổng hợp mô hình SVM

Nhận xét: Báo cáo phân loại cho thấy các mô hình trên đều đạt được độ chính xác tổng thể (accuracy) là khoảng 0.86 trên tập dữ liệu kiểm thử gồm 6680 mẫu. Tuy nhiên, hiệu suất của các mô hình trên các lớp khác nhau có sự chênh lệch đáng kể.

Đối với lớp 0 thì kết quả khá tốt cả recall lẫn precision đều cao nhưng ngược lại với lớp 1,2 thì nó precision mỗi lớp khoảng tầm 0.5 trở lên và recall của nó cực thấp ở lớp thứ 1 điều này 1 phần bởi vì tập train có sự mất cân bằng dữ liệu với các lớp 1,2. Để chọn lựa mô hình ổn nhất trong những cái trên ta nên đánh giá theo marco recall và marco precision bởi vì sử chênh lệch dữ liệu các lớp lớn nên khi xét precision tổng thì tuy là 0.86 khá cao nhưng nó không được trực quan nếu ta chú trọng vào macro recall thì mô hình logistic regression cho kết quả tốt nhất ,còn theo precision thì dùng mô hình svm.

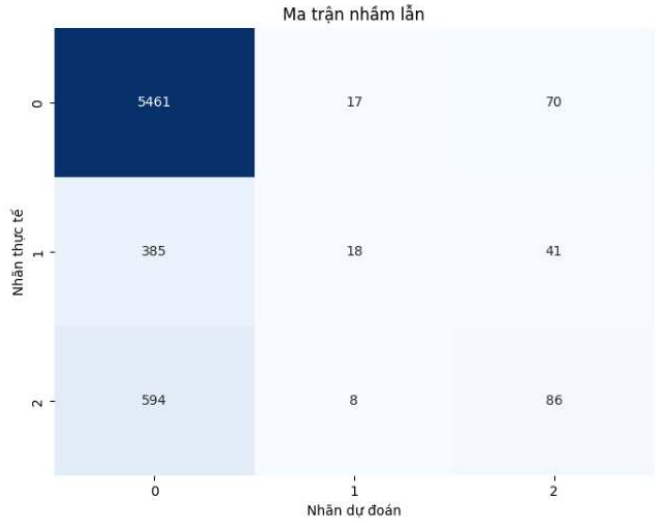
4.2.2. Sử dụng mô hình máy học dùng word2vec embedding

```
Mô hình lr:
  - Best params: {'clf__C': 100, 'clf__penalty': 'l2'}
  - Best score: 0.8322105684551099
Mô hình rf:
  - Best params: {'clf__max_depth': None, 'clf__min_samples_split': 5, 'clf__n_estimators': 50}
  - Best score: 0.8384480415921465
Mô hình svm:
  - Best params: {'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel': 'linear'}
  - Best score: 0.8269294795058988
```

Hình 4. 5 Các siêu tham số tốt nhất sau khi dùng gridsearch các mô hình

Báo cáo phân loại lr:

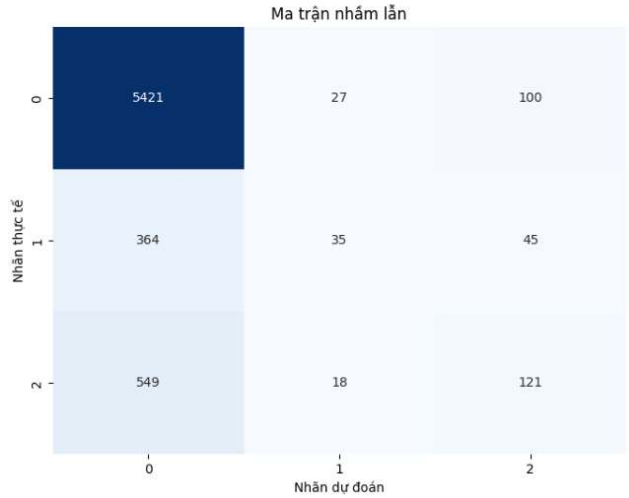
	precision	recall	f1-score	support
0	0.85	0.98	0.91	5548
1	0.42	0.04	0.07	444
2	0.44	0.12	0.19	688
accuracy			0.83	6680
macro avg	0.57	0.38	0.39	6680
weighted avg	0.78	0.83	0.78	6680



Hình 4. 6 Bảng kết quả tổng hợp mô hình logistic regression

Báo cáo phân loại rf:

	precision	recall	f1-score	support
0	0.86	0.98	0.91	5548
1	0.44	0.08	0.13	444
2	0.45	0.18	0.25	688
accuracy			0.83	6680
macro avg	0.58	0.41	0.43	6680
weighted avg	0.79	0.83	0.79	6680



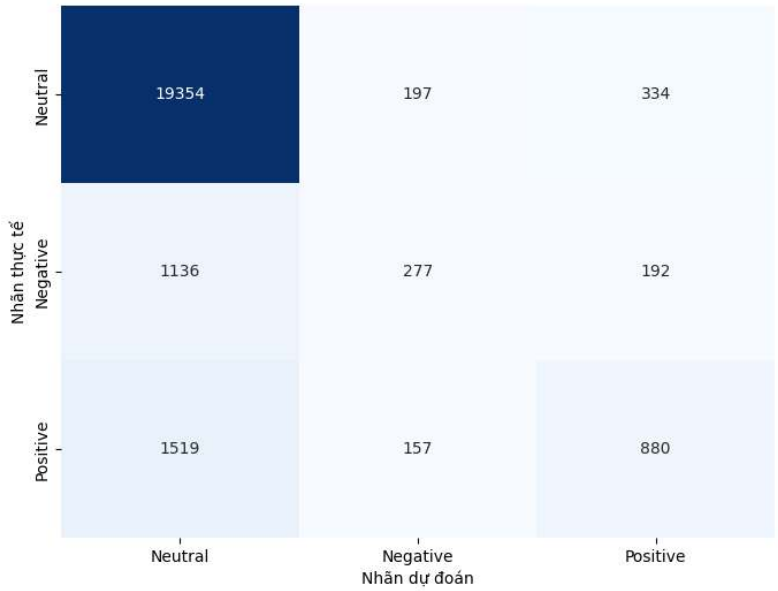
Hình 4. 7 Bảng kết quả tổng hợp mô hình random forest

Nhận xét: Các mô hình máy học sử dụng các câu đầu vào tổng hợp trung bình các từ được nhúng thành vector word2vec đạt kết quả kém hơn nhiều so với khi sử dụng đầu vào là dùng tf-idf trích xuất ra vector đặc trưng.

4.2.3. Sử dụng mạng học sâu

Báo cáo phân loại:

	precision	recall	f1-score	support
0	0.88	0.97	0.92	19885
1	0.44	0.17	0.25	1605
2	0.63	0.34	0.44	2556
accuracy			0.85	24046
macro avg	0.65	0.50	0.54	24046
weighted avg	0.82	0.85	0.83	24046



Hình 4. 8 Bảng kết quả tổng hợp mô hình FastText

Nhận xét:

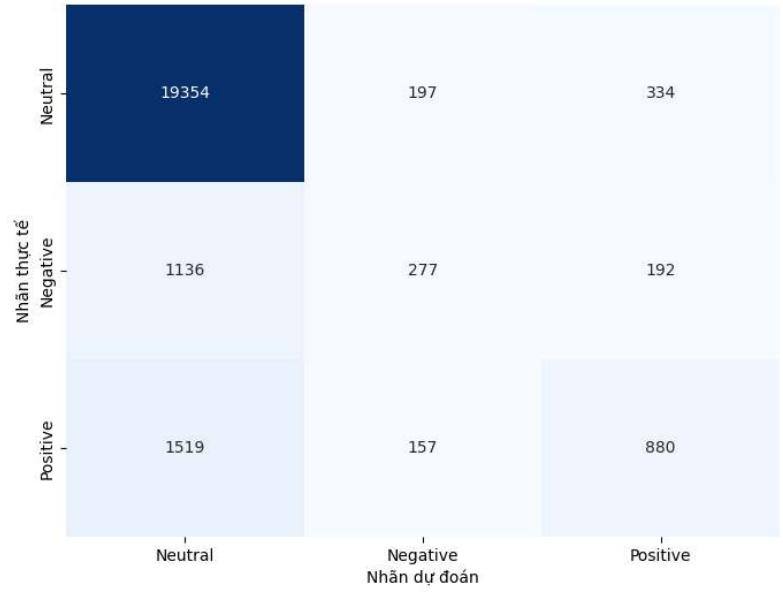
- Lớp 0 (Neutral): Mô hình đạt được độ chính xác (precision) là 0.88, độ nhạy (recall) là 0.97, và điểm F1 là 0.92. Điều này cho thấy mô hình hoạt động rất tốt trong việc phân loại các văn bản trung tính, với rất ít nhầm lẫn.
- Lớp 1 (Negative): Mô hình đạt được độ chính xác là 0.44, độ nhạy là 0.17, và điểm F1 là 0.25. Đây là lớp có hiệu suất kém nhất, cho thấy mô hình gặp khó khăn trong việc nhận diện các văn bản tiêu cực.
- Lớp 2 (Positive): Mô hình đạt được độ chính xác là 0.63, độ nhạy là 0.34, và điểm F1 là 0.44. Mặc dù mô hình hoạt động tốt hơn so với lớp tiêu cực, nhưng vẫn có nhiều nhầm lẫn khi phân loại các văn bản tích cực.
- Độ chính xác tổng thể: Mô hình đạt được độ chính xác tổng thể là 0.85, tức là mô hình dự đoán đúng 85% các văn bản trong tập kiểm tra.
- Trung bình có trọng số (weighted avg): Trung bình có trọng số của độ chính xác là 0.82, độ nhạy là 0.85, và điểm F1 là 0.83. Điều này cho thấy, khi xem xét tất cả các lớp và tầm quan trọng của chúng, mô hình có hiệu suất tổng thể khá tốt.

4.2.4. Đánh giá các trường hợp sai

Mỗi lớp sẽ trích ra 15 mẫu để phân tích tính đúng sai của nó

Báo cáo phân loại:

	precision	recall	f1-score	support
0	0.88	0.97	0.92	19885
1	0.44	0.17	0.25	1605
2	0.63	0.34	0.44	2556
accuracy			0.85	24046
macro avg	0.65	0.50	0.54	24046
weighted avg	0.82	0.85	0.83	24046



Hình 4. 8 Bảng kết quả tổng hợp mô hình FastText

Nhận xét:

- Lớp 0 (Neutral): Mô hình đạt được độ chính xác (precision) là 0.88, độ nhạy (recall) là 0.97, và điểm F1 là 0.92. Điều này cho thấy mô hình hoạt động rất tốt trong việc phân loại các văn bản trung tính, với rất ít nhầm lẫn.
- Lớp 1 (Negative): Mô hình đạt được độ chính xác là 0.44, độ nhạy là 0.17, và điểm F1 là 0.25. Đây là lớp có hiệu suất kém nhất, cho thấy mô hình gặp khó khăn trong việc nhận diện các văn bản tiêu cực.
- Lớp 2 (Positive): Mô hình đạt được độ chính xác là 0.63, độ nhạy là 0.34, và điểm F1 là 0.44. Mặc dù mô hình hoạt động tốt hơn so với lớp tiêu cực, nhưng vẫn có nhiều nhầm lẫn khi phân loại các văn bản tích cực.

- Độ chính xác tổng thể: Mô hình đạt được độ chính xác tổng thể là 0.85, tức là mô hình dự đoán đúng 85% các văn bản trong tập kiểm tra.
- Trung bình có trọng số (weighted avg): Trung bình có trọng số của độ chính xác là 0.82, độ nhạy là 0.85, và điểm F1 là 0.83. Điều này cho thấy, khi xem xét tất cả các lớp và tầm quan trọng của chúng, mô hình có hiệu suất tổng thể khá tốt.

CHƯƠNG 5: Kết luận

Phân loại cảm xúc văn bản trong NLP

Mục đích: Phân loại cảm xúc trong văn bản là một nhiệm vụ quan trọng trong Xử lý ngôn ngữ tự nhiên (NLP), giúp xác định thái độ hoặc cảm xúc được thể hiện trong một đoạn văn bản. Nó có nhiều ứng dụng thực tế như:

- Phân tích đánh giá sản phẩm, nhận xét khách hàng
- Lọc spam, bình luận tiêu cực trên mạng xã hội
- Phát triển chatbot, hệ thống hỗ trợ khách hàng tự động
- Hiểu rõ hơn về tâm lý người dùng, xu hướng thị trường

Cách tiếp cận: Có hai phương pháp chính để phân loại cảm xúc văn bản:

- Dựa trên quy tắc: Sử dụng các quy tắc thủ công được thiết lập dựa trên kiến thức ngôn ngữ và tâm lý học để xác định các từ ngữ, cụm từ liên quan đến cảm xúc.
- Học máy: Sử dụng các thuật toán học máy để tự động học hỏi từ dữ liệu mẫu, phân biệt các kiểu cảm xúc khác nhau. Các mô hình phổ biến bao gồm SVM, Naive Bayes, mạng nơ-ron nhân tạo (ANN).

Đánh giá hiệu quả: Hiệu quả của mô hình phân loại cảm xúc được đánh giá dựa trên các chỉ số như độ chính xác, độ chính xác, độ thu hồi, điểm F1.

Xu hướng phát triển:

- Phân loại cảm xúc đa nhãn: Xác định nhiều loại cảm xúc khác nhau trong cùng một văn bản.
- Phân loại cảm xúc theo ngữ cảnh: Xem xét ngữ cảnh của văn bản để xác định cảm xúc chính xác hơn.

- Phân loại cảm xúc phi ngôn ngữ: Phân tích cảm xúc từ giọng nói, biểu tượng cảm xúc, v.v.

Kết luận: Phân loại cảm xúc văn bản là một lĩnh vực nghiên cứu đang phát triển mạnh mẽ trong NLP với nhiều tiềm năng ứng dụng. Việc phát triển các mô hình hiệu quả và chính xác hơn sẽ góp phần mang lại nhiều lợi ích cho các lĩnh vực khác nhau như marketing, dịch vụ khách hàng, v.v.

Tài liệu tham khảo:

1. **Sonlam1102/vihsd (**
   https://github.com/sonlam1102/vihsd?fbclid=IwZXh0bgNhZW0CMTAAAR3iwpgfe53ZB1P2QHzLTvQj4h7AveNFJEt16S_nbXHgKTKH_m4WJIrxeGU_aem_OHtvG-AQ3rd-hQy6xYrCW)
2. Cài đặt phân loại cảm xúc tiếng Việt_Lê Minh Tú
   https://viblo.asia/p/cai-dat-mo-hinh-phan-loai-cam-xuc-tieng-viet-018J2vdRJYK