



ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO TỔNG KẾT
ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN
NĂM 2021

MÔ HÌNH KHÁM PHÁ TRẢI NGHIỆM KHÁCH HÀNG DỰA
TRÊN PHƯƠNG PHÁP PHÂN TÍCH QUAN ĐIỂM VÀ MÁY HỌC
SV 2022 228

Lĩnh vực khoa học: Kinh tế

Chuyên ngành: Thương mại - quản trị kinh doanh và du lịch- marketing

Nhóm nghiên cứu:

TT	Họ tên	MSSV	Đơn vị	Nhiệm vụ	Điện thoại	Email
1.	Nguyễn Trần Thúy Quỳnh	K2040 60307	Khoa HTTT	NT	0867511291	quynhntt20406c@uel.edu.vn
2.	Bùi Nguyễn Bích Ngọc	K2040 60288	Khoa HTTT	TV	0779177368	ngocbnb20406@st.uel.edu.vn
3.	Nguyễn Thị Bảo Trâm	K2041 10588	Khoa HTTT	TV	0353345869	tramntb20411@st.uel.edu.vn
4.	Trần Nhật Nguyên	K2040 61440	Khoa HTTT	TV	0767510181	nguyentn20406c@st.uel.edu.vn
5.	Võ Bá Tùng	K2040 60299	Khoa HTTT	TV	0832914036	tungvb20406@st.uel.edu.vn

ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KINH TẾ - LUẬT

BÁO CÁO TỔNG KẾT
ĐỀ TÀI NGHIÊN CỨU KHOA HỌC SINH VIÊN
NĂM 2021

**MÔ HÌNH KHÁM PHÁ TRẢI NGHIỆM KHÁCH HÀNG DỰA
TRÊN PHƯƠNG PHÁP PHÂN TÍCH QUAN ĐIỂM VÀ MÁY HỌC**

Đại diện nhóm nghiên cứu

(Ký, họ tên)

Giảng viên hướng dẫn

(Ký, họ tên)

Chủ tịch Hội đồng

(Ký, họ tên)

Lãnh đạo Khoa/Bộ môn/Trung tâm

(Ký, họ tên)

TÓM TẮT ĐỀ TÀI

Với sự tăng trưởng mạnh mẽ của Internet, thương mại di động tại Việt Nam bùng nổ và có nhiều sự phát triển vượt bậc. Những năm gần đây, người dùng có xu hướng quan tâm và để lại nhiều tương tác cũng như bình luận trên các ứng dụng để thể hiện những ý kiến và nhận xét sau trải nghiệm sản phẩm và dịch vụ được cung cấp. Đây là những thông tin quý giá và quan trọng ảnh hưởng đến sự phát triển của doanh nghiệp. Tuy nhiên, với số lượng dữ liệu vô cùng lớn và phức tạp, việc phân tích và thấu hiểu khách hàng trở nên khó khăn. Mục tiêu của bài nghiên cứu này chính là phân tích quan điểm của khách hàng qua các ý kiến bình luận trên các ứng dụng trong lĩnh vực thương mại di động bằng phương pháp kết hợp phân tích quan điểm và máy học. Từ phương pháp đề xuất trên, những hành vi, tâm lý và thói quen khách hàng được khám phá giúp các doanh nghiệp có thêm những chiều thông tin và tri thức đáng tin cậy để xây dựng những chiến lược nhằm cải thiện và nâng cao chất lượng của sản phẩm và dịch vụ.

Một số phương pháp nghiên cứu chính được đề tài áp dụng:

- Phương pháp định tính để khảo sát nguồn dữ liệu thứ cấp và các công trình nghiên cứu liên quan.
- Phương pháp tổng hợp lý thuyết về các phương pháp máy học và mô hình đánh giá.
- Phương pháp thực nghiệm trên bộ dữ liệu bình luận của khách hàng, dựa trên sự kết hợp giữa phương pháp máy học có giám sát và kỹ thuật xử lý ngôn ngữ tự nhiên để phân loại bộ dữ liệu thành 2 lớp quan điểm: tích cực và tiêu cực.
- Phương pháp thống kê, trực quan hóa kết quả thu được.

Tóm lược về kết quả nghiên cứu đã đạt được và các nhận định chính:

- Sau quá trình tiền xử lý và trích xuất đặc trưng thực nghiệm và phân tích hơn 935.000 bình luận là tập dữ liệu được thu thập từ các ứng dụng thương mại di động, kết quả nghiên cứu đã phân loại bình luận thành hai lớp quan điểm: tích cực và tiêu cực.
- Quá trình huấn luyện và thực nghiệm trên tập dữ liệu, 04 phương pháp máy học được áp dụng. Trong đó phương pháp Hồi quy Logistic cho độ chính xác cao nhất (Tiki: 92%, Sendo: 90%, Shopee: 91%, Lazada: 92%), và là phương pháp phù hợp nhất với bộ dữ liệu thực tế trong nghiên cứu.
- Qua các biểu đồ, báo cáo trực quan thể hiện thống kê về các bình luận tích cực và tiêu cực, các từ ngữ phổ biến, độ dài bình luận, xu hướng quan điểm tích cực và tiêu cực theo thời gian của khách hàng, nghiên cứu đưa ra những

nhận định về thái độ, hành vi và những mối quan tâm của khách hàng khi trải nghiệm mua sắm trên các nền tảng kỹ thuật số.

Các kết luận và đề xuất chính:

- Nghiên cứu đem lại những kết quả khả quan cùng với tính ứng dụng cao trong lĩnh vực thương mại di động nói riêng và thương mại điện tử nói chung. Các doanh nghiệp có thể ứng dụng phương pháp đề xuất từ nghiên cứu vào trong phân tích xu hướng thị trường, khám phá tâm lý, hành vi khách hàng và đưa ra những quyết định phù hợp. Trong tương lai, nghiên cứu sẽ tiếp tục được tập trung vào phát triển mô hình phân tích nhằm cải thiện hiệu suất về thời gian, độ chính xác và nghiên cứu mở rộng sang các lĩnh vực khác để có thể mang lại những lợi ích tối đa cho doanh nghiệp và người dùng.

MỤC LỤC

TÓM TẮT ĐỀ TÀI	1
MỤC LỤC	3
DANH SÁCH HÌNH	5
DANH SÁCH BẢNG	6
DANH MỤC TỪ VIẾT TẮT	7
1. Giới thiệu.....	8
2. Mục tiêu đề tài.....	9
3. Ý nghĩa khoa học và thực tiễn.....	9
3.1. Ý nghĩa khoa học	9
3.2. Ý nghĩa thực tiễn.....	9
4. Đối tượng và phạm vi nghiên cứu.....	10
5. Phương pháp nghiên cứu.....	10
6. Công cụ sử dụng.....	11
7. Kết cấu báo cáo	11
CHƯƠNG 1: TỔNG QUAN VỀ TÌNH NGHIÊN CỨU	13
1.1. Tổng quan tình hình nghiên cứu	13
1.1.1. Một số nghiên cứu ở nước ngoài	14
1.1.2. Một số nghiên cứu ở Việt Nam	17
1.2. Nhận định các kết quả nghiên cứu liên quan và đề xuất mô hình	17
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT.....	22
2.1. Khai phá văn bản.....	22
2.2. Xử lý ngôn ngữ tự nhiên	22
2.3. Thư viện	23
2.3.1. Pandas	23
2.3.2. Matplotlib.....	23
2.3.3. Regex	23
2.3.4. Sklearn	24
2.3.5. Seaborn	24
2.4. Phân tích quan điểm.....	24
2.5. Phương pháp trích xuất đặc trưng TF_IDF	26

2.6. Các phương pháp máy học.....	26
2.6.1. Hồi quy logistic.....	26
2.6.2. SVM.....	27
2.6.3. Naive Bayes	30
2.6.4. Random Forest.....	31
2.7. Phương pháp đánh giá các phương pháp máy học.....	32
CHƯƠNG 3: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU.....	34
Giới thiệu chương	34
3.1. Thu thập dữ liệu	34
3.1.1. Thư viện sử dụng	35
3.1.2. Quá trình thu thập dữ liệu:.....	36
3.1.3. Môi trường thực nghiệm.....	36
3.2. Phân tích khám phá dữ liệu (EDA).....	37
3.3. Tiền xử lý dữ liệu	38
3.4. Dán nhãn dữ liệu	41
3.5. Trích xuất đặc trưng.....	42
CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH.....	44
Giới thiệu chương	44
4.1. Mô hình dự đoán	44
4.1.1. Hồi quy Logistics.....	44
4.1.2. SVM.....	45
4.1.3. Naive Bayes	45
4.1.4. Random Forest.....	46
4.2. So sánh và lựa chọn mô hình	46
4.3. Kết quả thực nghiệm và thảo luận.....	49
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	57
5.1. Kết quả đạt được	57
5.2. Hạn chế.....	57
5.3. Hướng phát triển	58
DANH MỤC CÔNG TRÌNH CÔNG BỐ	59
TÀI LIỆU THAM KHẢO	60

DANH SÁCH HÌNH

Hình 0-1: Quy trình thực hiện nghiên cứu	12
Hình 1-1: Mô hình nghiên cứu tổng quan	21
Hình 2-1: Hồi quy logistic áp dụng cho phạm vi từ -20 đến 20	28
Hình 2-2: Support Vector Machine cho bài toán phân lớp	29
Hình 2-3: Tác động của việc tăng C đối với lề trong SVM	30
Hình 2-4: Phân loại theo mô hình Random Forest	33
Hình 3-1: Quy trình xử lý dữ liệu trong bài toán phân tích quan điểm	35
Hình 3-2: Quy trình thu thập dữ liệu	37
Hình 3-3: Thiết lập các thư viện cần thiết	37
Hình 3-4: ID ứng dụng trên Cửa hàng Google Play	37
Hình 3-5: Xác định các ngoại lai của tập dữ liệu	39
Hình 3-6: Xử lý các ngoại lai của tập dữ liệu	39
Hình 4-1: Những từ mang quan điểm tích cực và tiêu cực thường xuất hiện	54
Hình 4-2: Phân bổ phần trăm đánh giá theo ứng dụng và theo năm	55
Hình 4-3: WordCloud từ Tích cực	56
Hình 4-4: WordCloud từ Tiêu cực	56
Hình 4-5: Số lượng đánh giá của khách hàng từ năm 2015 đến 2021	57
Hình 4-6: Phân bổ độ dài của các đánh giá khách hàng	58
Hình 4-7: Mối quan hệ giữa độ dài và điểm đánh giá của bình luận	60

DANH SÁCH BẢNG

Bảng 3-1: Một phần tập dữ liệu thu thập được từ Sendo	36
Bảng 3-2: Môi trường thực nghiệm của nghiên cứu	38
Bảng 3-3: Minh họa dữ liệu trước và sau khi tiền xử lý	41
Bảng 3-4: Nhãn phân loại dựa vào số sao đánh giá	43
Bảng 3-5: Một phần dữ liệu đã được gán nhãn dựa vào đánh giá của Sendo	43
Bảng 3-6: Minh họa ma trận tần xuất TF_IDF	43
Bảng 4-1: Kết quả đánh giá mô hình trên toàn bộ dữ liệu	49
Bảng 4-2: Kết quả so sánh các mô hình	49
Bảng 4-3: Kết quả dự đoán sau khi thực hiện mô hình	51

DANH MỤC TỪ VIẾT TẮT

BERT: Bidirectional Encoder Representations from Transformers

TF_IDF: Term Frequency – Inverse Document Frequency

NLP: Natural language processing

EDA: Exploratory Data Analysis

SVM: Support Vector Machine

B2C: Business to Customer

IQR: Interquartile Range

LR: Logistic Regression

RF: Random Forest

NB: Naive Bayes

Pos: Positive

Neg: Negative

TỔNG QUAN ĐỀ TÀI

1. Giới thiệu

Đến năm 2021, thương mại di động tại Việt Nam sẽ tiếp tục phát triển nhanh chóng và bền vững. Khẳng định này của Hiệp hội Thương mại điện tử Việt Nam (VECOM) được phân tích và dự báo dựa trên sự phát triển xu hướng của lĩnh vực này trong giai đoạn 2016 đến 2020 cũng như kết quả khảo sát từ hàng ngàn doanh nghiệp trên cả nước. Trong quá trình bùng nổ đầu tiên của đại dịch Covid-19, vào tháng 5 năm 2020, Thủ tướng Chính phủ đã ban hành Quyết định số 645/QĐ-TTg phê duyệt kế hoạch phát triển thương mại điện tử quốc gia Lập kế hoạch cho giai đoạn 2021 đến năm 2025. Quyết định này nói rằng các doanh nghiệp là lực lượng cốt lõi trong các ứng dụng thương mại điện tử và thiết lập mục tiêu thu hẹp khoảng cách giữa các thành phố lớn và các địa phương khác. Theo đó, đến năm 2025, các địa phương ngoại trừ Hà Nội và Thành phố Hồ Chí Minh sẽ chiếm 50% giá trị của các giao dịch thương mại điện tử B2C trên toàn quốc ("Vietnam E-commerce Index 2021 Report", 2022).

Những năm trở lại đây, với sự phát triển mạnh mẽ của công nghệ Internet thì việc mua sắm trực tuyến trở nên phổ biến và điều này tạo nên áp lực lên các nền tảng công nghệ phải liên tục cải thiện chất lượng dịch vụ để đáp ứng nhu cầu và sự hài lòng đến từ người dùng. Việc mua sắm trực tuyến làm thay đổi thói quen và hành vi mua sắm người tiêu dùng bởi vì sự tiện lợi và nhanh chóng từ nó, mặt khác vẫn còn tồn tại nhiều vấn đề khi gặp khó khăn, không thuận lợi khi trải nghiệm mua sắm như giao diện khó sử dụng, dịch vụ giao hàng còn nhiều bất cập và nhiều yếu tố đã trở thành rào cản khi ngành thương mại di động đạt được lòng tin từ khách hàng. Có nhiều giải pháp và nghiên cứu để có thể hiểu và nắm bắt được trải nghiệm của khách hàng. Trong đó, phân tích quan điểm được nhiều doanh nghiệp và nhà nghiên cứu quan tâm nhằm mục tiêu xác định thái độ, cảm xúc, quan điểm của khách hàng khi trải nghiệm mua sắm trên các nền tảng kỹ thuật số với mục đích nâng cao trải nghiệm khách hàng để từ đó gia tăng sự trung thành, hài lòng và nâng cao được chất lượng, phát triển ứng dụng tốt hơn trong môi trường cạnh tranh như hiện nay.

Khi người tiêu dùng sử dụng các thiết bị di động ưu tiên mua sắm trực tuyến để tìm kiếm thông tin, truy cập, so sánh và đánh giá các sản phẩm thì giải pháp thương mại di động là kênh đóng vai trò quan trọng đối với doanh nghiệp. Các doanh nghiệp, nhà bán hàng tăng cường các chiến lược kỹ thuật số nhằm tiếp cận người dùng hiệu quả và đây là cơ hội để thu hút khách hàng, hiện diện thương hiệu doanh nghiệp trên các nền tảng trực tuyến. Bên cạnh đó, thấu hiểu tâm lý khách hàng để đưa chiến lược tiếp thị

phù hợp, xây dựng lòng tin đối với họ càng trở nên quan trọng hơn bao giờ hết. Việc thu thập, phân tích các đánh giá của người dùng trên các nền tảng di động khi trải nghiệm mua sắm trên các nền tảng kỹ thuật số với mục đích nâng cao trải nghiệm khách hàng để từ đó gia tăng sự trung thành, hài lòng và nâng cao được chất lượng và phát triển ứng dụng tốt hơn trong môi trường cạnh tranh như hiện nay (Ritter & Pedersen, 2019). Chính vì vậy, phân tích quan điểm của khách hàng dựa trên những bình luận và phản hồi trên các kênh mua sắm trực tuyến là vô cùng quan trọng. Việc khai thác dữ liệu một cách hiệu quả giúp doanh nghiệp nhận biết được trải nghiệm của khách hàng, nắm bắt những vấn đề đang gặp phải nhằm cải thiện kết quả kinh doanh và đưa ra những chiến lược phục vụ thị trường tốt hơn, giúp cắt giảm chi phí, tăng doanh thu.

2. Mục tiêu đề tài

Nghiên cứu và đề xuất mô hình khám phá trải nghiệm khách hàng dựa trên sự kết hợp phương pháp máy học và phân tích quan điểm, cảm xúc của khách hàng qua những bình luận, tương tác và đánh giá trên các trang thương mại di động. Từ đó phân tích và tìm ra tâm lý, hành vi và thói quen khách hàng đưa ra những đề xuất để cải thiện chất lượng của ứng dụng, sản phẩm trên ứng dụng và hàm ý doanh nghiệp.

3. Ý nghĩa khoa học và thực tiễn

3.1. Ý nghĩa khoa học

Nghiên cứu tập trung vào xây dựng và tối ưu mô hình phân tích quan điểm. Qua việc áp dụng nhiều phương pháp tiền xử lý, trích xuất đặc trưng, các mô hình máy học, đánh giá mô hình,... chúng tôi đã thực nghiệm trên bộ dữ liệu thực, làm rõ các lý thuyết và đặc điểm của các mô hình. Mô hình đề xuất có độ chính xác cao và thời gian huấn luyện tương đối ngắn, có thể ứng dụng trong phân tích quan điểm trong các lĩnh vực khác.

3.2. Ý nghĩa thực tiễn

Với tốc độ phát triển nhanh chóng của công nghệ, hàng ngày, hàng giờ, con người tiếp cận với một lượng thông tin khổng lồ. Dữ liệu luôn là tài nguyên vô cùng quý giá trong thời đại ngày nay, đặc biệt hữu ích trong việc phân tích và đưa ra kết luận. Với vấn đề dữ liệu lớn như hiện nay, việc ứng dụng các phương pháp máy học ngày càng cần thiết, bởi chúng ta sẽ có thể tiếp cận và phân tích vấn đề một cách tổng quan, chính xác hơn, có thể liên kết các vấn đề nhỏ để mở ra những hướng giải quyết sáng tạo, các quyết định hiệu quả.

Bên cạnh đó, do ảnh hưởng của đại dịch Covid 19, thương mại điện tử Việt Nam có sự phát triển rõ rệt. Theo Sách trắng Thương mại điện tử Việt Nam, năm 2020, tốc độ tăng trưởng của thương mại điện tử đạt mức 18%, quy mô đạt 11,8 tỷ USD và là nước duy nhất ở Đông Nam Á có tăng trưởng thương mại điện tử hai con số. Con người ngày càng chú trọng đến những cảm xúc, trải nghiệm mua sắm trên các website, ứng dụng Thương mại điện tử, họ luôn mong muốn những trải nghiệm mua sắm tốt nhất và tiện lợi nhất.

Với xu thế mua sắm trực tuyến, việc khách hàng trao đổi, để lại bình luận và đánh giá về sản phẩm, dịch vụ cũng trở nên dễ dàng và minh bạch hơn. Từ đó, người dùng sẽ quan tâm nhiều hơn về các bình luận và trao đổi về sản phẩm để có thể đưa ra những quyết định mua sắm phù hợp nhất. Về phía nhà sản xuất và các doanh nghiệp, họ quan tâm đến những phản hồi của người dùng, tìm ra những điểm mạnh, điểm yếu của sản phẩm để nhanh chóng đổi mới và phát triển ứng dụng.

Từ đó, nhu cầu về phân tích trải nghiệm và cảm xúc khách hàng trên các ứng dụng Thương mại di động ngày càng cần thiết. Việc phân tích quan điểm dựa trên những phương pháp máy học sẽ mang đến tính chính xác cao hơn, giúp cho các nhà phát triển ứng dụng có thể xác định những ưu điểm hiện có của ứng dụng để tiếp tục phát huy, đồng thời nhận ra những hạn chế còn tồn tại để nhanh chóng khắc phục, nâng cao tối đa trải nghiệm mua sắm trực tuyến của người dùng.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu là quan điểm tích cực và tiêu cực thông qua trải nghiệm của khách hàng, xu hướng quan điểm theo từng giai đoạn thời gian và phương pháp kết hợp máy học trong phân tích quan điểm khách hàng trong lĩnh vực thương mại di động.

Phạm vi đề tài tập trung trong lĩnh vực Thương mại di động tại Việt Nam, thực hiện phân tích các bình luận, tương tác của khách hàng trên 04 ứng dụng thương mại di động tại Việt Nam, bao gồm Tiki, Shopee, Lazada và Sendo trong khoảng thời gian từ năm 2015 đến năm 2021.

5. Phương pháp nghiên cứu

Trong bài nghiên cứu này, phương pháp nghiên cứu định tính và thực nghiệm được áp dụng. Trong đó, phương pháp định tính được dùng để khảo sát các dữ liệu thứ cấp, các kết quả, các công trình đã được công bố về phân tích quan điểm khách hàng khi sử dụng các ứng dụng thương mại di động nói riêng và khách hàng trực tuyến nói chung. Từ đó, tìm ra được các khoảng trống nghiên cứu trong các nghiên cứu khác để tiến hành

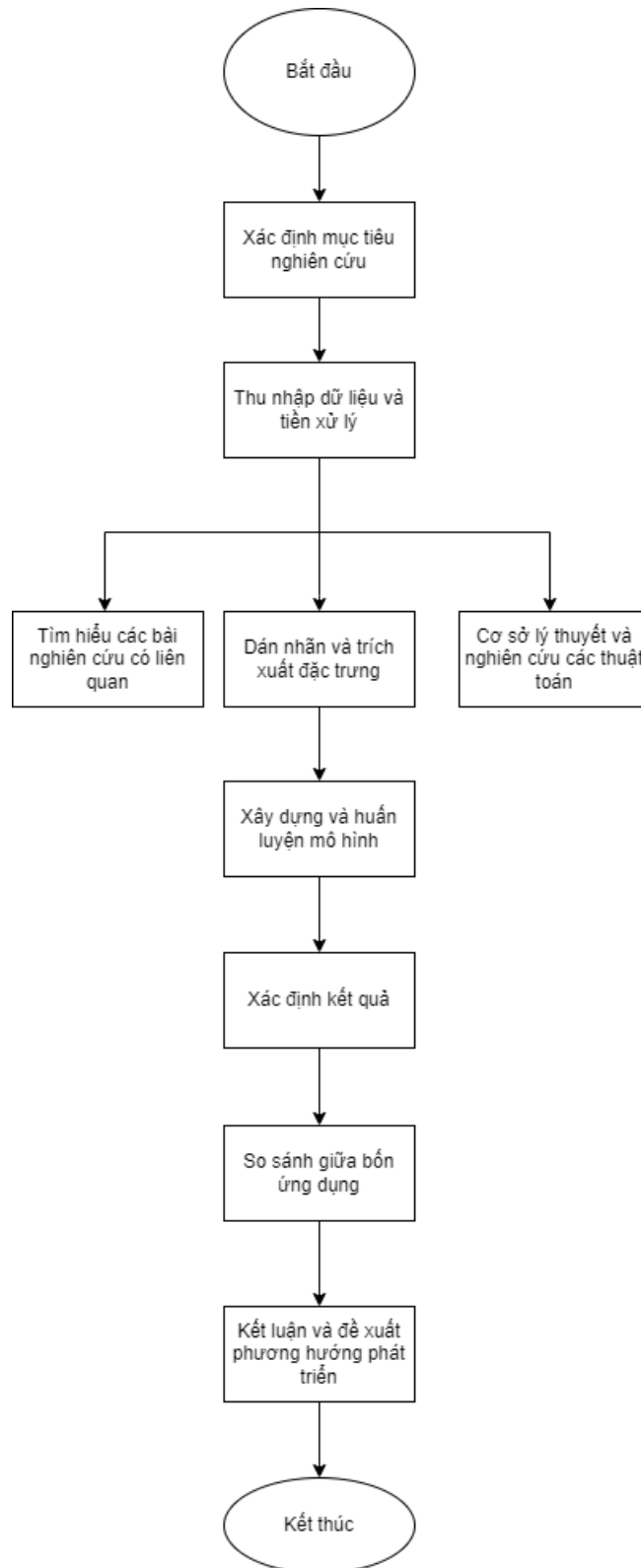
xây dựng mô hình và quy trình để thực hiện nghiên cứu thực nghiệm; Phương pháp thực nghiệm được áp dụng để tiến hành khảo sát và thu thập dữ liệu, phân tích xử lý các dữ liệu từ bộ dữ liệu thu thập được và thực nghiệm bằng các phương pháp máy học. Phương pháp này được áp dụng để đánh giá kết quả thực nghiệm và trực quan các kết quả phân tích quan điểm của khách hàng.

6. Công cụ sử dụng

- Jupyter Notebook, phiên bản 5.3.1
- Visual Studio Code, phiên bản 1.65
- Google Colab, phiên bản 3.7.12

7. Kết cấu báo cáo

Để thực hiện nghiên cứu, chúng tôi bắt đầu định hướng, xác định mục tiêu đề tài nghiên cứu. Sau đó tiến hành thu thập và áp dụng các kỹ thuật xử lý dữ liệu là bình luận và đánh giá sản phẩm và dịch vụ trên 04 nền tảng thương mại di động lớn ở Việt Nam. Để tăng tính thuyết phục cho bài nghiên cứu, chúng tôi đã tham khảo và tìm ra khoảng trống các nghiên cứu liên quan đã thực hiện trước đó. Tiếp theo nghiên cứu chuyên sâu để tìm hiểu các phương pháp cũng như lý thuyết được sử dụng trong bài nghiên cứu nhằm xây dựng mô hình thích hợp và trực quan hóa, sau đó so sánh kết quả nghiệm thu từ các tập dữ liệu. Cuối cùng, bài nghiên cứu sẽ đưa ra những giải pháp và hàm ý cho doanh nghiệp.



Hình 0-1: Quy trình thực hiện nghiên cứu (Nguồn: nhóm tác giả)

CHƯƠNG 1: TỔNG QUAN VỀ TÌNH NGHIÊN CỨU

1.1. Tổng quan tình hình nghiên cứu

Phân tích quan điểm còn được gọi là khai thác ý kiến, là lĩnh vực nghiên cứu về ý kiến, tình cảm, đánh giá, thẩm định, thái độ và quan điểm đối với các thực thể như sản phẩm và dịch vụ (Liu, B., 2012)(Sharma, R., Nigam, S. and Jain, R., 2014). Phân tích quan điểm thường được phân loại thành 03 khía cạnh: tích cực, tiêu cực và trung tính.

Phân tích quan điểm cũng được áp dụng để quản lý thông tin chính phủ (Bang, B. and Lee, L., 2008) cho phép chính phủ theo dõi được những ý kiến đóng góp hoặc phản ánh của người dân. Phân tích quan điểm cũng có thể ứng dụng vào phân tích những tin tức thời sự nhằm phân tích những nội dung tin tức hoặc xác định xu hướng tin tức được quan tâm nhiều nhất (Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D. and Keim, D., 2009). Ngoài ra, phân tích quan điểm cũng được áp dụng để cảm thiện hệ thống giáo dục dựa vào các đánh giá về khóa học, học kỳ hay thậm chí là giảng viên giảng dạy (Binali, H., Potdar, V. and Wu, C., 2009).

Đối với một doanh nghiệp, những nội dung do người dùng tạo trên các ứng dụng di động đã có thể cung cấp cho doanh nghiệp những thông tin về ý kiến tích cực, tiêu cực hay trung tính của người tiêu dùng về sản phẩm của doanh nghiệp và của đối thủ cạnh tranh (Liu, B., 2012). Các doanh nghiệp thường gặp khó khăn trong việc đo lường mức độ quan tâm của người tiêu dùng và xác định dữ liệu người dùng nào thực sự hữu ích để họ thu thập. Bằng cách sử dụng phân tích quan điểm được bổ sung với sự thông minh của con người, các doanh nghiệp có thể lọc ra dữ liệu nhiều, mơ hồ và với sự trợ giúp của công nghệ máy học có thể xác định dữ liệu quan trọng thúc đẩy hoạt động kinh doanh của họ (Al-Otaibi, S., Alnassar, A., Alshahrani, A., Al-Mubarak, A., Albugami, S., Almutiri, N. and Albugami, A., 2018).

Cho đến hiện nay, có nhiều phương pháp được nghiên cứu và áp dụng vào phân tích khách hàng ý kiến khách hàng trực tuyến. Trong đó, điển hình là hai phương pháp đó là phương pháp máy học (machine learning) (Li Z, Fan Y, Jiang B, Lei T, Liu W)(Pang B, Lee L) và phương pháp dựa trên từ vựng (lexicon-based method)(Taboada M, Brooke J, Tofiloski M, Voll K, Stede M)(Melville P, Gryc W, Lawrence R)(Ding X, Liu B, Yu P). Mỗi phương pháp có những ưu điểm và hạn chế khi được áp dụng. Chẳng hạn như, phương pháp dựa trên từ vựng là một cách tiếp cận không giám sát tương đối dễ thực hiện. Tuy nhiên, đối với các tập dữ liệu lớn với nhiều đặc trưng để phân tích thì chỉ phương pháp dựa trên từ vựng là không đủ để tiếp cận một cách có hiệu quả (Poria S, Chaturvedi I, Cambria E, Bisio F)(Ruder S, Ghaffari P, G. Breslin J). Trong khi đó

các phương pháp máy học có giám sát khi phân loại dữ liệu thành các lớp yêu cầu dữ liệu đầu vào phải sạch và được gán nhãn theo cấu trúc nhất định (Hutto C, Gilbert E).

Các phương pháp máy học được áp dụng cho phân tích quan điểm chủ yếu thuộc về phân loại có giám sát. Một số phương pháp máy học được sử dụng để phân loại các đánh giá: Naive Bayes (Domingos, P. and Pazzani, M., 1997), Hồi quy Logistic (Maalouf, M., 2011), Random Forest (Cutler, A., Cutler, D. and Stevens, J., 2012), Support Vector Machine (Barbosa, L. and Feng, J., 2010) và nhiều mô hình khác. Phương pháp máy học bắt đầu từ việc thu thập tập dữ liệu huấn luyện, sau đó huấn luyện một bộ phân loại trên dữ liệu huấn luyện. Khi một kỹ thuật phân loại được giám sát được chọn, một bước quan trọng: quyết định thực hiện là lựa chọn đặc trưng. Cuối cùng, phân loại có giám sát cho biết cách thức tập dữ liệu thể hiện (Nandi, A. and Sharma, P., 2021).

Mục tiêu của nghiên cứu này, nhằm để tăng tính hiệu quả của phân tích quan điểm, sẽ đề xuất kết hợp hai phương pháp học máy và dựa trên từ vựng (hybrid method) (Mudinas, A., Zhang, D. and Levene, M., 2012) và thực nghiệm phương pháp này trên bộ dữ liệu là ý kiến phản hồi của khách hàng trên 04 trang thương mại di động. Từ kết quả thực nghiệm, nghiên cứu sẽ áp dụng phương pháp ma trận nhằm lần để đánh giá hiệu suất của phương pháp đề xuất sau khi huấn luyện và lựa chọn mô hình hiệu quả nhất để áp dụng cho dữ liệu thực tế và đưa ra phương pháp phù hợp giúp cho các nhà phát triển ứng dụng có thể xác định những ưu điểm hiện có của ứng dụng để tiếp tục phát huy, đồng thời nhận ra những khuyết điểm còn tồn tại để nhanh chóng khắc phục, nâng cao tối đa trải nghiệm mua sắm trực tuyến của người dùng.

1.1.1. Một số nghiên cứu ở nước ngoài

1. Z. Singla, S. Randhawa & S. Jain, (2017). *Sentiment analysis of customer product reviews using machine learning*.

Tiến hành phân tích quan điểm của các bài đánh giá điện thoại di động và phân loại những đánh giá thành tích cực và tiêu cực bằng cách sử dụng 3 mô hình phân loại: Naive Bayes, SVM, Cây quyết định và kết quả dự đoán SVM là tốt nhất.

2. Ravi, K.S. & Dr. Kamalraj, R. (2021). *Amazon Product Review Sentiment Analysis with Machine Learning*.

Bài báo nhằm mục đích áp dụng và mở rộng xử lý ngôn ngữ tự nhiên và phân tích quan điểm hiện có đối với dữ liệu được thu thập từ Amazon. Họ sử dụng phương pháp học máy có giám sát (Hồi quy Logistic, Cây quyết định, SVM) để phân cực một tập dữ liệu khổng lồ của Amazon và mô hình SVM được đánh giá đạt kết quả dự đoán chính xác cao nhất.

3. Jansher, Rabnawaz. (2020). *Sentimental Analysis of Amazon Product Reviews Using Machine Learning Approach*

Dữ liệu đánh giá sản phẩm trên Amazon được chia thành hai lớp tích cực và tiêu cực và sử dụng 3 phương pháp máy học có giám sát (Naive Bayes, SVM). Kết quả SVM phân loại có độ chính xác và giá trị thu hồi tốt hơn Naive Bayes.

4. Wassan, Sobia & Chen, Xi & Shen, Tian & Waqar, Muhammad & Zaman, Noor. (2021). *Amazon Product Sentiment Analysis using Machine Learning Techniques*

Thu thập tập dữ liệu từ trung tâm dữ liệu thế giới nơi tỷ lệ các ý kiến, quan điểm được phát hiện đầu tiên trong phân tích. Sau khi thực hiện các thao tác tiền xử lý dữ liệu, túi từ, parameter tuning và kiểm chứng chéo, cuối cùng là gán nhãn dữ liệu thành 2 nhãn tiêu cực và tích cực.

5. Behrooz Noori. (2021). *Classification of Customer Reviews Using Machine Learning Algorithms*.

Bài viết này là về phân loại và dự đoán cảm xúc của khách hàng. Trong bài viết này, một khuôn khổ mới để phân loại và dự đoán tình cảm của khách hàng đã được đề xuất. Các đánh giá của khách hàng được thu thập từ một khách sạn quốc tế. Trong bước tiếp theo, các bài đánh giá của khách hàng được xử lý, sau đó đưa vào các phương pháp máy học khác nhau. Các phương pháp được sử dụng trong bài báo này là máy vectơ hỗ trợ (SVM), mạng nơ-ron nhân tạo (ANN), Naive Bayes (NB), cây quyết định (Decision Tree), C4.5 và k-nearest neighbor (K-NN). Trong số các phương pháp này, cây quyết định đưa ra kết quả tốt hơn.

6. R. Nagamanjula and A. Pethalakshmi. (2018). *A Machine Learning Based Sentiment Analysis by Selecting Features for Predicting Customer Reviews*

Bài báo này trình bày phương pháp phân loại mới gọi là SVM để cải thiện độ chính xác bằng cách chia dữ liệu thành hai lớp tích cực và tiêu cực. Ban đầu, các từ được thu thập từ Amazon và tiền xử lý bằng công cụ wordnet. Để phân loại đánh giá, nhận xét từ người dùng.

7. Hanhoon Kang; Seong Joon Yoo; Dongil Han (2012). *Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews*.

Khi phân loại tập dữ liệu thành tích cực và tiêu cực sử dụng các phương pháp máy học có giám sát (SVM, Naive Bayes) dựa vào senti-lexicon thì phân loại tích cực có khuynh hướng cho độ chính xác xuất hiện cao hơn so với phân loại tiêu cực.

8. Naz, Sheeba; Sharan, Aditi; Malik, Nidhi (2018). *Sentiment Classification on Twitter Data Using Support Vector Machine*.

Nghiên cứu này nhằm mục đích phân loại cảm xúc trên một thang đo hai điểm (phân loại nhị phân) với SVM sử dụng các đặc trưng n-gram cùng 3 trọng số khác nhau. Họ nhằm mục đích quan sát cách tiếp cận kết hợp của vector điểm số cảm xúc và N-grams ảnh hưởng đến hành vi của mô hình SVM về mặt chính xác của phân loại.

9. Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2022). Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches.

Trong bài báo này, họ trình bày nghiên cứu so sánh về cảm xúc văn bản mô hình phân loại sử dụng TF-IDF trong cả hai kỹ thuật học máy và kỹ thuật dựa trên từ vựng. Sau đó so sánh 6 phương pháp: Hồi quy logistic, SVM, Gradient Boosting và 3 phương pháp dựa trên từ vựng là: VADER, Pattern và SentiWordnet.

10. Yadav, Nikhil & Kudale, Omkar & Gupta, Srishti & Rao, Aditi & Shitole, Ajitkumar. (2020). *Twitter Sentiment Analysis Using Machine Learning For Product Evaluation*.

Bài báo này nhấn mạnh các kỹ thuật phân loại khác nhau (Naive Bayes, Cây quyết định, Random Forest, XGBoost, SVM) được sử dụng để phân loại sản phẩm phê bình theo các phê bình được thể hiện trong các tweet để phân tích liệu tích cực, tiêu cực, trung tính.

11. D'souza, Stephina Rodney; Sonawane, Kavita (2019). *Sentiment Analysis Based on Multiple Reviews by using Machine learning approaches*.

Nghiên cứu thực hiện về việc phân tích quan điểm kép xử lý cảm xúc với tất cả khía cạnh (tích cực, tiêu cực, trung tính) và phương pháp phân loại được sử dụng bao gồm: SVM, Naive Bayes.

12. Bayhaqy, Achmad; Sfenrianto, Sfenrianto; Nainggolan, Kaman; Kaburuan, Emil R. (2018). *Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes*.

Rapidminer được sử dụng để hỗ trợ đưa ra phân tích quan điểm bằng cách sử dụng ba phương pháp phân loại khác nhau dự đoán nhãn dán trong tập dữ liệu: Cây quyết định, K-NN và Naive Bayes.

13. Yiran, Ye; Srivastava, Sangeet (2019). *Aspect-based Sentiment Analysis on mobile phone reviews with LDA*.

Các mô hình LDA được sử dụng để phân cụm các từ chủ đề với các giá trị xác suất tương ứng của chúng. Dựa vào kết quả khung học máy, bằng cách sử dụng khung này để thực hiện dán nhãn chủ đề và phân tích quan điểm bằng cách sử dụng ma trận nhầm lẫn và F-measure.

1.1.2. Một số nghiên cứu ở Việt Nam

1. Bằng, N.Đ., Hồ, N.V., & Thành, H.T. (2021). *Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm*.

Nghiên cứu hoàn thành giải pháp ứng dụng phân tích ngôn ngữ tự nhiên cụ thể là phân tích cảm xúc khách hàng dựa trên bình luận được đăng tải trên trang web Foody.vn dựa vào phương pháp học máy có giám sát áp dụng phương pháp Hồi quy Logistic, Cây quyết định, Naive Bayes và kết quả nghiên cứu cho thấy hồi quy Logistic là phương pháp tốt hơn so với các phương pháp còn lại dựa vào thời dự đoán và độ chính xác.

2. Nguyễn, T. T., & Trần, G. T. C. (2019). *Một mô hình học máy trong phân tích ý kiến khách hàng dựa trên văn bản tiếng việt: Bài toán dịch vụ khách sạn*.

Nghiên cứu đề xuất phương pháp máy học trong phân tích ý kiến khách hàng trên văn bản tiếng Việt trường hợp bài toán dịch vụ khách sạn, phân thành 2 nhãn dữ liệu là tích cực và tiêu cực thu thập từ website booking.com với 7 mô hình huấn luyện: SGD, SVM, Hồi quy Logistic, Naive Bayes, Random Forest, K-Neighbors và Cây quyết định.

3. Trinh, S., Nguyen, L., Vo, M., & Do, P. (2016). *Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese Language. Studies in Computational Intelligence*

Trong nghiên cứu này, các tác giả đề xuất một phương pháp dựa vào từ vựng cho phân tích quan điểm với dữ liệu Facebook cho tiếng Việt theo hai trọng tâm thành phần cốt lõi trong hệ thống cảm xúc. Đó là xây dựng từ điển cảm xúc Việt Nam (VED) bao gồm 5 tiểu từ điển: danh từ, động từ, tính từ, trạng từ và đề xuất các tính năng dựa trên phương pháp phân tích quan điểm tiếng Anh và thích ứng với ngôn ngữ truyền thống của Việt Nam và sau đó SVM được sử dụng để xác định cảm xúc thông điệp người dùng.

1.2. Nhận định các kết quả nghiên cứu liên quan và đề xuất mô hình

Có nhiều ứng dụng và cải tiến trên các phương pháp phân tích quan điểm đã được đề xuất và sử dụng từ một số nhiều năm cho đến nay. Bài nghiên cứu này nhằm mục đích cung cấp một cái nhìn sâu hơn về các kỹ thuật phổ biến được sử dụng trong kinh

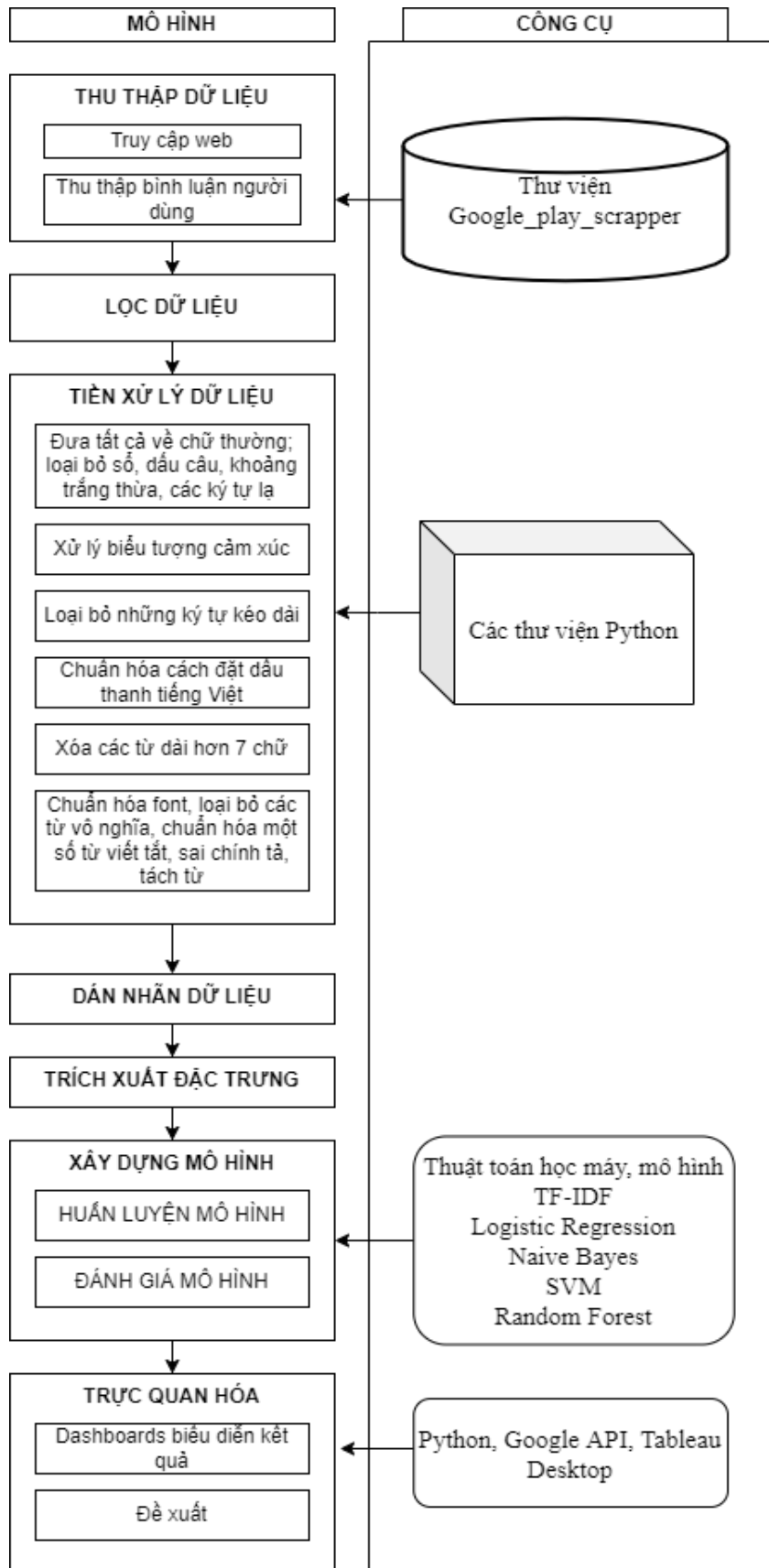
doanh bán lẻ, đặc biệt là trong lĩnh vực thương mại di động và đưa ra một đánh giá toàn diện về nó. Trong phân loại quan điểm, có hai lĩnh vực nghiên cứu chính như máy học và Lexicon, và trong mỗi lĩnh vực đều có sự chia nhỏ. Tuy nhiên, cũng có một số nghiên cứu kết hợp hai kỹ thuật này và đạt được một cách tương đối hiệu quả tốt hơn trong hoạt động phân tích quan điểm.

Phân tích quan điểm nhằm xác định quan điểm của các đánh giá dựa trên cơ sở mang các ý nghĩa tích cực, tiêu cực và trung tính. Các nghiên cứu dựa trên từ vựng - Lexicon Based sử dụng các cụm từ được xác định trước và các thành ngữ quan điểm trong đó mỗi cụm từ và thành ngữ được đánh giá là cảm xúc tích cực hoặc tiêu cực. Phần lớn các nhà nghiên cứu đã sử dụng các phương pháp tiếp cận tự động như từ điển và ngữ liệu để gán các từ quan điểm, nhưng chúng vẫn theo cách thủ công chỉ định các từ và câu trong các tuyên bố về quan điểm để đảm bảo sự phân công từ và câu đúng. Quy tắc này nghiêng về các từ gần gũi về mặt ngữ nghĩa, trong trường hợp của các ngành bán lẻ như các trang thương mại di động. Ở đây có rất nhiều khám phá các nhận xét và đánh giá bao gồm tiếng lóng và sai chính tả do các ngôn ngữ khác nhau. Tình huống này dẫn đến khó khăn cho việc thiết kế và phát triển hệ thống tự động. Thêm vào đó, để đánh giá quan điểm của nhận xét, sự hiểu biết về cách dùng từ ngữ hiện đại là cần thiết để phân loại các cực của ý kiến. Hai phương pháp tiếp cận có thể được sử dụng trong bộ phân loại từ vựng chẳng hạn như phương pháp dựa trên từ điển và phương pháp dựa trên Corpus để thu thập từ điển trực tuyến với số lượng lớn đa dạng các phát biểu quan điểm để tham chiếu cho các từ đồng nghĩa và trái nghĩa tương ứng. Các cụm từ mới được thêm vào danh sách các đề xuất sử dụng và tiếp tục thêm các cụm từ lặp đi lặp lại cho đến khi không tìm thấy cụm từ mới. Nó đã được nhân mạnh bằng cách sử dụng kiểm tra thủ công để làm sạch và cho ra danh sách cuối cùng. Phương pháp máy học – sử dụng các kỹ thuật khác nhau để phân tích quan điểm trong số đó một trong những kỹ thuật thường gặp SVM, Naive Bayes, Hồi quy Logistic, Cây quyết định, Random Forest, K-NN, ...

Đối với phương pháp dựa trên từ vựng thì kết quả ảnh hưởng phần lớn vào chất lượng từ miêu tả cảm xúc. Đối với các phương pháp học máy ví dụ như SVM, Hồi quy Logistic, Cây quyết định,... kết quả cách trích xuất đặc trưng phụ thuộc vào mô hình huấn luyện Bag of N - gram hoặc đặc trưng từ vựng (Lexicon-based features). Đối với các nghiên cứu trong nước về tiếng Việt, nghiên cứu khai phá ý kiến và phân tích quan điểm khách hàng thì trước khi đưa vào huấn luyện, nghiên cứu đã phân loại cảm xúc theo điểm số đánh giá hoặc một câu văn bản thành hai nhóm tích cực và tiêu cực.

Bên cạnh đó, lĩnh vực ứng dụng mô hình và phương pháp nghiên cứu của các nghiên cứu trước đó được khảo sát hầu như chưa tập trung khai thác ý kiến khách hàng trong lĩnh vực thương mại di động, một lĩnh vực tiềm năng đang phát triển mạnh trong hiện tại và tương lai. Điều này dẫn đến nhu cầu phân tích hành vi trải nghiệm khách hàng thông qua các bình luận để lại trên các ứng dụng di động ngày càng cần thiết và không chỉ trong lĩnh vực nghiên cứu mà trong cả doanh nghiệp.

Từ kết quả khảo sát các nghiên cứu liên quan, trong phạm vi nghiên cứu, chúng tôi đề xuất kết hợp phương pháp máy học và phương pháp dựa trên từ vựng - Hybrid (Sentiment và Lexicon) (Hình 1-1) để phân tích quan điểm khách hàng trong lĩnh vực thương mại di động. Quá trình gán nhãn dữ liệu trong phương pháp đề xuất chúng tôi đưa ra sẽ dựa vào điểm số đánh giá thành 2 bộ dữ liệu (tích cực và tiêu cực) trước khi đưa vào huấn luyện. Và phân loại quan điểm dựa trên phương pháp học máy với phương pháp biểu diễn văn bản sang dạng ma trận vector dựa vào mô hình TF-IDF để xây dựng vector đặc trưng trước khi đưa vào các phương pháp máy học. Sau cùng là dùng ma trận nhầm lẫn (Confusion Matrix) để xác định một câu nhận xét được nhập vào hay nhiều câu nhận xét trong tập kiểm thử.



Hình 1-1: Mô hình nghiên cứu tổng quan (Nguồn: nhóm tác giả)

Với mô hình nghiên cứu tại hình 1-1 được đề xuất, bắt đầu từ việc phân tích yêu cầu, thu thập dữ liệu thô các đánh giá mà khách hàng để lại từ 04 ứng dụng thương mại di động Tiki, Shopee, Sendo và Lazada. Tập dữ liệu này được tiền xử lý, chuẩn hóa và gán nhãn trước khi đưa vào trích xuất đặc trưng. Bộ dữ liệu của 04 ứng dụng thương mại di động được chia làm hai phần: tập dữ liệu huấn luyện (training data) và tập dữ liệu kiểm tra (test data). Tập dữ liệu huấn luyện sử dụng để thiết lập các phương pháp máy học và tập dữ liệu kiểm tra được dùng để đánh giá các phương pháp máy học, từ đó chọn ra mô hình phù hợp nhất với bộ dữ liệu thu thập được. Cuối cùng, sau khi có được mô hình phù hợp với bộ dữ liệu, dữ liệu được trực quan hóa để có thể so sánh, đánh giá giữa các ứng dụng thương mại di động và tìm ra hướng đi phù hợp cho các doanh nghiệp triển khai thương mại di động.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Khai phá văn bản

Khai phá văn bản (Text Mining) là một quá trình xử lý và trích xuất thông tin từ văn bản, quá trình này là một phần của việc phân tích văn bản trong khai phá dữ liệu. Khai phá văn bản chia thành các vấn đề nhỏ hơn bao gồm: phân loại tài liệu (text categorization, text classification), gom cụm văn bản (text clustering), trích xuất thực thể (concept/entity extraction), phân tích quan điểm (sentiment analysis), tóm tắt tài liệu (document summarization), và trích xuất quan hệ giữa các thực thể (entity relation modeling) (Hotho, Andreas & Nürnberger, Andreas & Paass, Gerhard, 2005).

Trong bài nghiên cứu sau đây, khai phá văn bản sẽ tập trung vào phân tích quan điểm. Phân tích quan điểm liên quan đến việc lấy thông tin, phân tích từ vựng để nghiên cứu sự phân bố của các tần số từ, nhận dạng mẫu, gắn thẻ/chú thích, trích xuất thông tin, kỹ thuật khai thác dữ liệu bao gồm phân tích liên kết, trực quan và phân tích dự đoán. Mục tiêu cơ bản là, để chuyển đổi văn bản thành dữ liệu dùng cho phân tích, thông qua kỹ thuật xử lý ngôn ngữ tự nhiên (NLP), các loại thuật toán và phương pháp phân tích khác nhau (Feldman, Ronen & Ronen, & Sanger, & James, 2007).

2.2. Xử lý ngôn ngữ tự nhiên

Ngôn ngữ tự nhiên là công cụ mà con người sử dụng như là phương tiện để giao tiếp, trao đổi, truyền đạt nhằm cung cấp một thông tin nào đó, ví dụ như qua giọng nói, văn bản (email, SMS, web pages...), gọi các thông tin ấy là dạng dữ liệu ngôn ngữ tự nhiên.

Xử lý ngôn ngữ tự nhiên (Natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người (Kesej, V., 2009). Các bước xử lý ngôn ngữ tự nhiên, gồm có:

- Phân tích hình thái: Trong bước này từng từ sẽ được phân tích và các ký tự không phải chữ (như các dấu câu) sẽ được tách ra khỏi các từ. Trong tiếng Anh và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. Tuy nhiên trong tiếng Việt, dấu cách được dùng để phân tách các tiếng (âm tiết) chứ không phải từ.
- Phân tích cú pháp: Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ này. Sẽ có những dãy từ bị loại do vi phạm các quy luật văn phạm.
- Phân tích ngữ nghĩa: Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ phân tích cú pháp.

- Tích hợp văn bản: Ngữ nghĩa của một câu riêng biệt có thể phụ thuộc vào những câu đứng trước, đồng thời nó cũng có thể ảnh hưởng đến các câu phía sau.
- Phân tích thực nghĩa: Cấu trúc thể hiện điều được phát ngôn sẽ được thông dịch lại để xác định nó thật sự có nghĩa là gì.

Các bài toán và ứng dụng của kỹ thuật xử lý ngôn ngữ tự nhiên: Nhận dạng chữ viết (đánh máy hoặc viết tay), Nhận dạng tiếng nói (chuyển giọng nói thành văn bản), Tổng hợp tiếng nói (chuyển văn bản thành giọng nói), Dịch tự động, Tìm kiếm thông tin, Tóm tắt văn bản, Khai phá dữ liệu và phát hiện ra tri thức.

Trong nghiên cứu này, phương pháp phân tích hình thái được áp dụng để phân tích từ, tách từ và gán nhãn từ loại để phục vụ cho phân tích khía cạnh của từ trong bình luận của khách hàng.

2.3. Thư viện

2.3.1. Pandas

Pandas là một gói thư viện viết bằng Python phổ biến cho khoa học dữ liệu và với lý do như: nó cung cấp các cấu trúc dữ liệu mạnh mẽ, linh hoạt giúp các thao tác và phân tích dữ liệu dễ dàng hơn. DataFrame là một trong những cấu trúc dữ liệu rất mạnh của Pandas. Pandas kết hợp các tính năng tính toán mảng hiệu suất cao của NumPy với khả năng thao tác dữ liệu linh hoạt của bảng tính và cơ sở dữ liệu quan hệ (như SQL). Nó cung cấp chức năng lập chỉ mục chính xác để giúp dễ dàng định hình lại, cắt và trộn, thực hiện tổng hợp và chọn tập hợp dữ liệu. Pandas là công cụ chính mà chúng ta sẽ sử dụng trong bài báo này để xử lý dữ liệu.

2.3.2. Matplotlib

Matplotlib là một thư viện vẽ đồ thị cho ngôn ngữ lập trình Python và phần mở rộng toán học số NumPy của nó. Nó cung cấp một API hướng đối tượng để nhúng các lô vào ứng dụng bằng cách sử dụng các bộ công cụ GUI có mục đích chung như Tkinter, wxPython, Qt hoặc GTK.

2.3.3. Regex

Regex là các các kí tự được kết hợp với nhau theo quy tắc để tạo nên một trình tự giúp chúng ta tìm kiếm và thay thế văn bản một cách thông minh, nhanh chóng, đơn giản và thuận tiện. Regex có thể dùng được trong hầu hết các ngôn ngữ lập trình bậc cao như Java, C#, Python, JS, PHP, ...

2.3.4. Sklearn

Scikit-learn (trước là scikits.learn, còn được gọi là sklearn) là một thư viện phần mềm máy học miễn phí dành cho ngôn ngữ lập trình Python. Các tính năng của thư viện áp dụng cho phương pháp phân lớp, đệ quy, gom cụm, bao gồm support vector machines, random forests, gradient boosting, k-means và DBSCAN, đồng thời thư viện này được thiết kế để phối hợp với thư viện số Python và các thư viện cụ thể như NumPy và SciPy.

Scikit-learn là một trong những thư viện máy học nổi tiếng nhất trên cộng đồng GitHub, phần lớn được viết bằng Python, và một số thuật toán viết bằng python để tăng hiệu suất. Support vector machines được thực thi từ LIBSVM; Logistic regression và linear support vector machines từ LIBLINEAR. Trong những trường hợp này thì người dùng không thể mở rộng phương pháp với bằng Python. Scikit-learn phối hợp tốt với rất nhiều thư viện Python khác như matplotlib và plotly để đánh dấu, đánh số cho các vector mảng, pandas dataframes, scipy, ...

2.3.5. Seaborn

Seaborn là một thư viện trực quan dữ liệu Python dựa trên matplotlib. Nó cung cấp một giao diện cấp cao để vẽ đồ họa thống kê hấp dẫn và thông tin. Để giới thiệu ngắn gọn về các ý tưởng đằng sau thư viện, có thể đọc các ghi chú giới thiệu hoặc giấy. Truy cập trang cài đặt để xem cách bạn có thể tải xuống gói và bắt đầu với nó. Có thể duyệt bộ sưu tập ví dụ để xem một số thứ mà có thể làm với Seaborn và sau đó kiểm tra tham chiếu hướng dẫn hoặc API để tìm hiểu làm thế nào.

2.4. Phân tích quan điểm

Là một lĩnh vực nghiên cứu, có liên quan chặt chẽ (hoặc có thể được coi là một phần) với ngôn ngữ học tính toán, xử lý ngôn ngữ tự nhiên và khai thác văn bản. Tiến hành từ nghiên cứu về trạng thái tình cảm (tâm lý học) và phán đoán (lý thuyết thẩm định), lĩnh vực này tìm cách trả lời các câu hỏi được nghiên cứu từ lâu trong các lĩnh vực diễn ngôn khác bằng cách sử dụng các công cụ mới được cung cấp bởi khai thác dữ liệu và ngôn ngữ học tính toán.

Sentiment Analysis (phân tích quan điểm) có nhiều tên gọi. Thường được gọi là phân tích chủ quan, khai thác ý kiến và khai thác thẩm định, với một số kết nối với điện toán cảm tính (nhận dạng máy tính và biểu hiện của cảm xúc) (Bang, B. and Lee, L., 2008). Lĩnh vực này thường nghiên cứu các yếu tố chủ quan, được định nghĩa bởi như

là “các diễn đạt ngôn ngữ của các trạng thái riêng trong ngữ cảnh” (Wiebe, Janyce & Wilson, Theresa & Bruce, Rebecca & Bell, Matthew & Martin, Melanie, 2004).

Do sự phức tạp của vấn đề (khái niệm cơ bản, biểu thức trong văn bản, v.v.), phân tích quan điểm bao gồm một số nhiệm vụ riêng biệt. Chúng thường được kết hợp để tạo ra một số kiến thức về các ý kiến được tìm thấy trong văn bản. Phần này cung cấp tổng quan về các nhiệm vụ này và phần tiếp theo sẽ thảo luận về một số công cụ được sử dụng cho từng công việc.

- Nhiệm vụ đầu tiên là phát hiện cảm xúc hoặc ý kiến, có thể được xem như phân loại văn bản là khách quan hoặc chủ quan. Thông thường, việc phát hiện ý kiến dựa trên việc kiểm tra các tính từ trong câu.
- Nhiệm vụ thứ hai là phân loại phân cực. Với một đoạn văn bản có nhiều ý kiến, mục tiêu là phân loại ý kiến thuộc một trong hai thái cực tình cảm đối lập, hoặc xác định vị trí của nó trên sự liên tục giữa hai thái cực này (Bang, B. & Lee, L., 2008).

Để phân biệt sự trộn lẫn khác nhau của hai đối cực, phân loại phân cực sử dụng thang điểm đa điểm (chẳng hạn như số lượng sao cho một bài đánh giá phim).

Hai nhiệm vụ trên có thể được thực hiện ở một số cấp độ: thuật ngữ (term level), cụm từ (phrase level), câu (sentence level), khía cạnh (aspect level) hoặc cấp tài liệu (document level). Người ta thường sử dụng đầu ra của một mức làm đầu vào cho các lớp cao hơn (Turney và Littman, 2003; Dave et al., 2003; Kanayama et al., 2004). Các kỹ thuật khác nhau phù hợp với các cấp độ khác nhau. Các kỹ thuật sử dụng phân loại n-gram hoặc từ vựng (lexicons) thường hoạt động ở cấp độ thuật ngữ, trong khi gắn thẻ Part-Of-Speech được sử dụng để phân tích cụm từ và câu. Heuristics thường được sử dụng để khái quát cảm xúc đến cấp độ tài liệu.

Phân tích quan điểm có nhiều hướng tiếp cận khác nhau: Phân tích quan điểm tiếp cận theo Xử lý ngôn ngữ tự nhiên (Natural Language Processing); Phân tích quan điểm tiếp cận theo phương pháp Học máy (Machine Learning); Phân tích quan điểm tiếp cận theo phương pháp Khai thác văn bản (Text Mining),... của Itisha Gupta và Nisheeth Joshi (2019) cho rằng hướng tiếp cận lai (Hybrid Approach) khai thác cả hai hướng tiếp cận học máy và Xử lý ngôn ngữ tự nhiên thể hiện độ chính xác cao và ổn định hơn hẳn.

Trong bài nghiên cứu này, nhóm chúng tôi tiếp cận theo hướng lai, kết hợp hướng tiếp cận Học máy và Xử lý ngôn ngữ tự nhiên, với các phương pháp máy học: Hồi quy Logistic, SVM, Naive Bayes và Random Forest.

2.5. Phương pháp trích xuất đặc trưng TF_IDF

Trong các nghiên cứu của (Ahmed, H., Awan, M., Khan, N., Yasin, A. and Faisal Shehzad, H., 2021)(Nasim, Z., Rajput, Q. and Haider, S., 2017) đã đưa ra một số kỹ thuật tiền xử lý tập dữ liệu dạng văn bản, sau đó gán nhãn và sử dụng véc-tơ trọng số TF-IDF để đánh giá mức độ quan trọng của 1 từ và tần suất xuất hiện của từ đó trong đoạn văn bản, khi áp dụng kỹ thuật trích xuất đặc trưng này thì chúng ta sẽ xếp hạng được các véc-tơ đặc trưng cùng với các phương pháp máy học phân cụm.

Sử dụng 2 bước xử lý để cho ra được mô hình trọng số:

- TF: Ước lượng tần suất xuất hiện của từng từ trong văn bản.

$$TF(t, d) = (\text{số lần từ } t \text{ xuất hiện trong văn bản } d) / (\text{tổng số từ trong văn bản } d) \quad (2-1)$$

- IDF : Nghịch đảo tần suất của văn bản, giúp đánh giá tầm quan trọng của một từ. Khi tính TF, thì giả định mức độ quan trọng của các từ là như nhau, tuy có một số từ được thể hiện nhiều nhưng lại không có quan trọng để thể hiện nghĩa của văn bản, ví dụ như: “là”, “của”, “đó”, “thế”, “nhỉ”,... Vì vậy ta cần giảm mức độ quan trọng của những từ đó đi.

$$IDF(t, D) = \log_e(\text{Tổng số văn bản trong tập mẫu } D / \text{Số văn bản có chứa từ } t) \quad (2-2)$$

- Chỉ số TF_IDF là tích của hai thông số này:

$$TF_IDF(t) = TF(t) * IDF(t) \quad (2-3)$$

2.6. Các phương pháp máy học

2.6.1. Hồi quy logistic

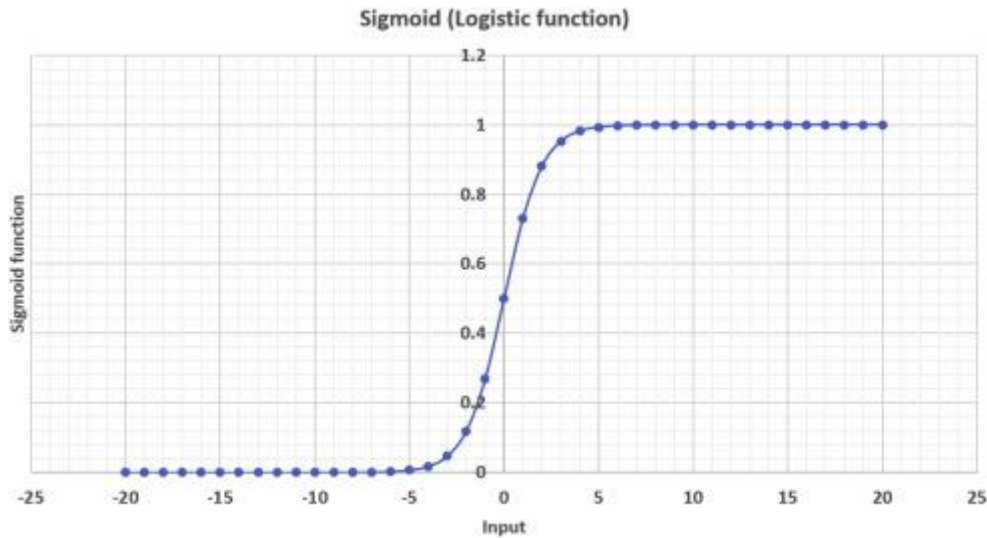
Hồi quy logistic (Edgar, T., & Manz, D. (2017) là một phương pháp máy học có giám sát mạnh mẽ khác được sử dụng cho các bài toán phân loại nhị phân (khi mục tiêu là phân loại). Cách tốt nhất để nghĩ về hồi quy logistic là nó là một hồi quy tuyến tính nhưng dành cho các bài toán phân loại. Hồi quy logistic về cơ bản sử dụng một hàm logistic được định nghĩa dưới đây để lập mô hình biến đầu ra nhị phân (Tolles & Meurer, 2016). Sự khác biệt cơ bản giữa hồi quy tuyến tính và hồi quy logistic là phạm vi của hồi quy logistic bị giới hạn từ 0 đến 1. Ngoài ra, trái ngược với hồi quy tuyến tính, hồi quy logistic không yêu cầu mối quan hệ tuyến tính giữa các biến đầu vào và đầu ra. Điều

này là do việc áp dụng một phép biến đổi log phi tuyến đối với tỷ lệ chênh lệch (sẽ được xác định ngay sau đây).

Phương trình hàm Logistic:

$$y = f(s) = \frac{1}{1 + e^{-x}} \quad (2-4)$$

Trong phương trình hàm logistic, x là biến đầu vào, đưa các giá trị -20 đến 20 vào hàm logistic. Như minh họa, các đầu vào đã được chuyển từ 0 đến 1 .



Hình 2-1: Hồi quy logistic áp dụng cho phạm vi từ -20 đến 20

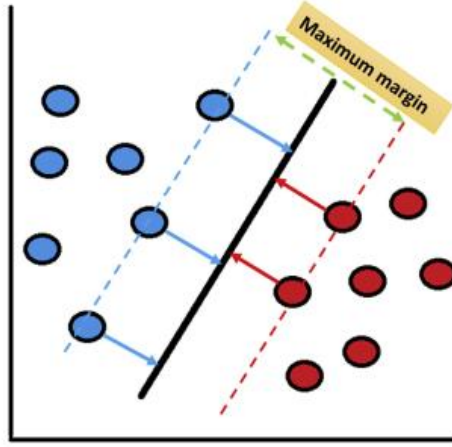
(Nguồn: Edgar, T., & Manz, D. (2017))

2.6.2. SVM

Support Vector Machine (SVM) là phương pháp máy học có giám sát (Supervised Machine Learning), mục tiêu của hàm là để giảm thiểu hàm chi phí (Cost function), mục tiêu trong SVM là tối ưu hóa lề (margin) giữa các véc-tơ hỗ trợ (support vector) qua một siêu phẳng (hyperplane) tách biệt (Cortes, Vapnik., 1995).

Trong hình 2-2, một siêu phẳng phân chia (đường liền nét màu đen) được vẽ để tách các chấm màu xanh khỏi những cái màu đỏ. Siêu phẳng này được vẽ theo cách để tối ưu hóa biên độ cả hai bên. Khoảng cách giữa đường nét đứt màu xanh và đường nét đứt màu đỏ được gọi là “Margin” (lề). Hình minh họa này dành cho không gian hai

chiều, rút ra được rằng siêu phẳng sẽ là $n-1$ nếu không gian là n chiều.



Hình 2-2: Support Vector Machine cho bài toán phân lớp

(Nguồn: towardsdatascience.com)

Giả sử tập dữ liệu huấn luyện Z bao gồm N điểm dữ liệu. Trong đó điểm dữ liệu thứ i là $Z_i = (x_i, y_i)$ với $x_i \in \mathbb{R}^d$ là vector đầu vào và y_i là biến mục tiêu; là một trong hai giá trị $\{-1, 1\}$. Tập dữ liệu này được giả định là *phân tuyến (linear separable)*.

Khoảng cách từ một điểm tới một siêu phẳng

- Trong trường hợp tổng quát, khoảng cách từ một điểm bất kỳ $Z_i = (x_i, y_i)$ tới đường biên là siêu phẳng H có phương trình $b + w^T x = 0$ sẽ là:

$$d(Z_i, H) = \frac{|b + w^T x_i|}{\|w\|_2} = \frac{y_i(b + w^T x_i)}{\|w\|_2} \quad (2-5)$$

- Trong công thức trên thì $|b + w^T x| = y_i(b + w^T x_i)$ là vì:
 - + Xét trường hợp nhãn $y_i = -1$ thì điểm Z_i nằm ở mặt âm và có $b + w^T x_i \leq 0$. Do đó, $y_i(b + w^T x_i) \geq 0$.
 - + Xét trường hợp nhãn $y_i = 1$ thì điểm Z_i nằm ở mặt dương và có $b + w^T x_i \geq 0$. Do đó, $y_i(b + w^T x_i) \geq 0$.
- Trong cả hai trường hợp, đẳng thức $|b + w^T x| = y_i(b + w^T x_i)$ đều xảy ra.

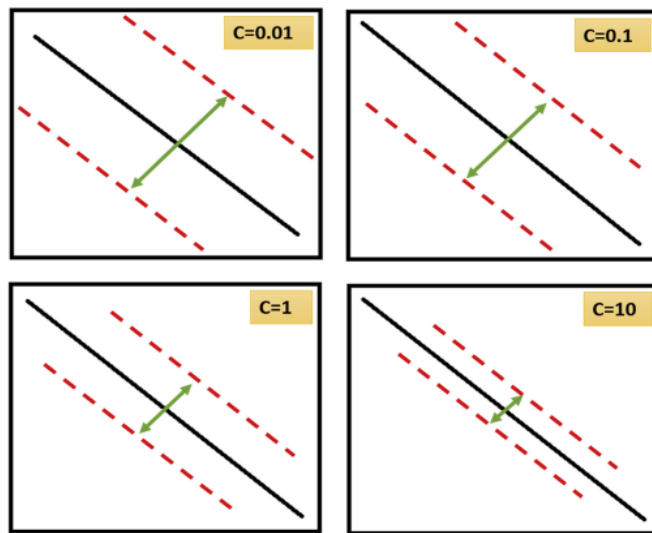
Bài toán dự báo nhãn:

- Nhãn của một quan sát trong mô hình SVM sẽ phụ thuộc vào dấu của đường biên:

$$\begin{aligned}
 h_{w,b}(x_i) &= b + w^T x_i \\
 &= b + \left(\sum_{(x_j, y_j) \in S} \lambda_j y_j x_j^T \right) x_i \\
 &= b + \sum_{(x_j, y_j) \in S} \lambda_j y_j x_j^T x_i
 \end{aligned}
 \tag{2-6}$$

Trong trường hợp $h_{w,b}(x_i) > 0$ thì điểm được dự báo có nhãn 1 và trái lại sẽ là nhãn -1.

Công thức trên cho thấy thay vì phải xác định nhãn dựa trên các hệ số của phương trình đường biên w thì chúng ta có thể thông qua các điểm thuộc tập hỗ trợ.



Hình 2-3: Tác động của việc tăng C đối với lề trong SVM

(Nguồn: Yunqian Ma, Guodong Guo., 2014)

Soft margin: được sử dụng để tìm một đường phân tách lớp nhưng đường này sẽ chấp nhận một hoặc một vài trường hợp bị phân loại sai. Mức độ chấp nhận được thể hiện bằng tham số “C” trong thuật toán SVM. Hình 2-3, C về cơ bản là một tham số đánh đổi được sử dụng giữa việc có biên độ rộng và phân loại chính xác dữ liệu đào tạo. C được xem như một tham số chính quy khi sử dụng SVM.

Kernel tricks: khi dữ liệu không thể phân tách tuyến tính, các thủ thuật hạt nhân (kernel) sử dụng các đặc trưng hiện có và áp dụng một số hàm chuyển đổi để tạo các đặc trưng mới.

2.6.3. Naive Bayes

Phương pháp máy học có phân loại Naive Bayes là một phân loại Bayesian (Berry, 1996) mà làm cho một giả định được đơn giản hóa về cách các đặc trưng tương tác với nhau. Naive Bayes là một bộ phân loại theo xác suất, có nghĩa là đối với dữ liệu văn bản d , vector x , thay vì tìm ra chính xác nhãn của mỗi điểm dữ liệu $x \in \mathbb{R}$, ta tìm xác suất để kết quả rơi vào mỗi nhãn: $p(y = c|x)$, từ đó xác định nhãn của mỗi điểm dữ liệu bằng cách chọn ra nhãn có xác suất cao nhất.

$$c = \operatorname{argmax} p(c|x) \quad \text{với } c \in (1, \dots, C) \quad (2-7)$$

Ý tưởng suy luận Bayesian này đã được biết đến từ công việc của Bayes (1763), và lần đầu tiên được áp dụng cho phân loại văn bản bởi Mosteller và Wallace (năm 1964). Vì khó có thể tính trực tiếp $p(c|x)$, chúng ta sử dụng quy tắc của Bayes để có thể tính xác suất có điều kiện.

$$c = \operatorname{argmax} p(c|x) = \operatorname{argmax} \frac{p(x|c)p(c)}{p(x)} = \operatorname{argmax} p(x|c)p(c) \quad (2-8)$$

Ở dấu bằng thứ hai, chúng ta đã đơn giản hóa một cách thuận tiện bằng cách bỏ mẫu số $p(x)$. Lý do là chúng ta sẽ tính $p(x|c)p(c)$ cho mỗi lớp khác nhau, tuy nhiên, $p(x)$ không hề thay đổi ở mỗi lớp.

Tiếp tục, chúng ta đơn giản hóa phép toán với hai giả định:

- Thứ nhất, là túi từ giả định: chúng ta giả định, vị trí không quan trọng, các từ ở vị trí đầu tiên, thứ 2, thứ 3 hay cuối cùng không ảnh hưởng đến kết quả cuối cùng. Các đặc trưng $x_1, x_2, x_3, \dots, x_d$ được mã hóa do nhận dạng từ vựng mà không phải vị trí trong văn bản.
- Thứ hai, là giả định Naive Bayes: giả thuyết về xác suất của các yếu tố khi chúng độc lập với nhau.

$$p(x|c) = p(x_1, x_2, x_3, \dots, x_d|c) = \prod p(x_i|c) \quad (2-9)$$

Như vậy, phép toán được đơn giản hóa thành:

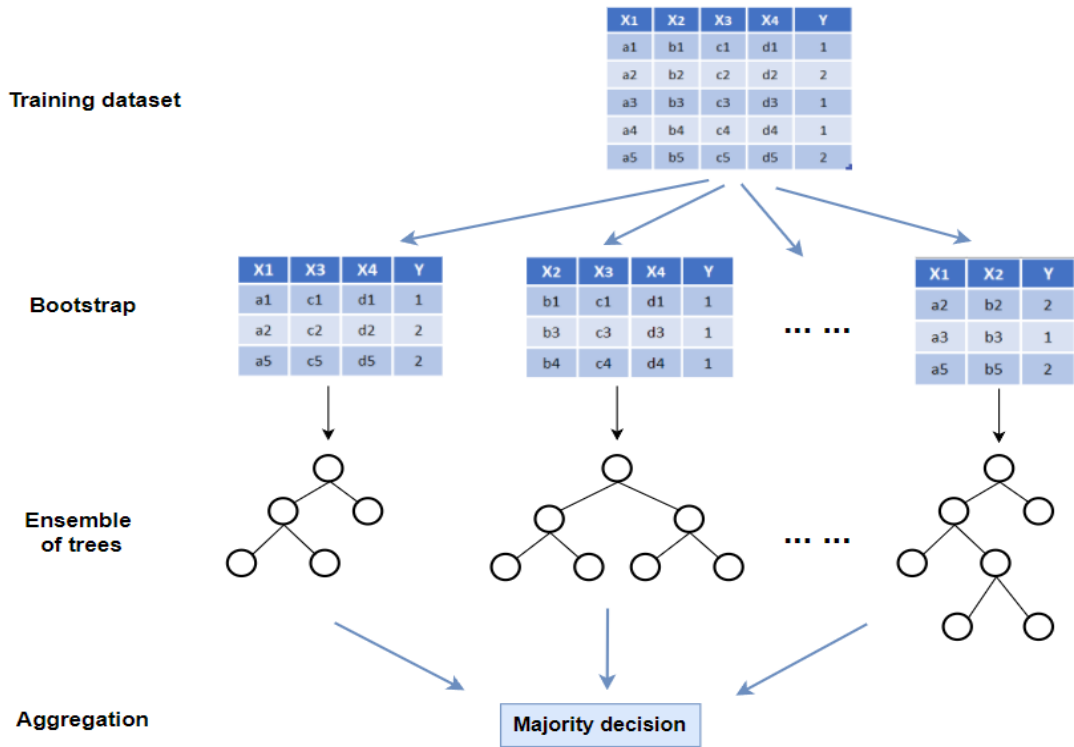
$$c = \operatorname{argmax} p(c) \prod p(x_i|c), \text{ với } c \in (1, \dots, C), i=1 \quad (2-10)$$

Khi d lớn và các xác suất nhỏ, biểu thức ở vế phải là một số rất nhỏ, dẫn đến khi tính toán chúng ta có thể gặp sai số. Để giải quyết việc này, chúng ta viết lại phép toán dưới dạng tương đương bằng cách lấy log của vế phải, việc này không ảnh hưởng tới kết quả vì log là một hàm đồng biến trên tập các số dương.

$$c = \operatorname{argmax} (\log(p(c)) + \sum \log(p(x_i|c))), \text{ với } c \in (1, \dots, C), i=1 \quad (2-11)$$

2.6.4. Random Forest

Phân loại Random Forest (Misra, S., & Li, H. (2020)) là một phương pháp hòa tấu mà huấn luyện một số cây quyết định song song với bootstrapping theo sau là tập hợp, cùng được đề cập đến khi đóng gói. Bootstrapping chỉ ra rằng một số cây quyết định cá nhân được huấn luyện song song với các tập con khác nhau của bộ dữ liệu huấn luyện bằng cách sử dụng các tập con khác nhau có các tính năng có sẵn. Bootstrapping đảm bảo rằng mỗi cây quyết định cá nhân trong Random Forest là duy nhất, làm giảm phương sai tổng thể của bộ phân loại Random Forest. Đối với quyết định cuối cùng, phân loại Random Forest tổng hợp các quyết định của từng cây riêng lẻ; Do đó, phân loại Random Forest thể hiện sự khái quát tốt. Trình phân loại Random Forest có xu hướng vượt trội so với hầu hết các phương thức phân loại khác về tính chính xác mà không có vấn đề quá mức. Giống như phân loại cây quyết định, bộ phân loại Random Forest không cần mở rộng tính năng. Không giống như phân loại cây quyết định, bộ phân loại Random Forest mạnh mẽ hơn với việc lựa chọn các mẫu huấn luyện và loại bỏ dữ liệu bị nhiễu trong tập dữ liệu huấn luyện. Trình phân loại Random Forest khó giải thích hơn nhưng dễ dàng hơn để điều chỉnh siêu tham số (hyperparameter) so với phân loại cây quyết định.



Hình 2-4: Phân loại theo mô hình Random Forest

(Nguồn: Misra & Li, 2020)

Triển khai bộ phân loại Random Forest trên tập dữ liệu có bốn tính năng (X1, X2, X3 và X4) và hai lớp (Y = 1 và 2). Bộ phân loại Random Forest là một phương pháp tập hợp để huấn luyện một số cây quyết định song song với việc khởi động và theo sau là tổng hợp. Mỗi cây được huấn luyện trên các tập hợp con khác nhau của mẫu và tính năng huấn luyện.

2.7. Phương pháp đánh giá các phương pháp máy học

Đối với bài nghiên cứu này, chúng tôi sử dụng Ma trận nhầm lẫn (Confusion Matrix) để đánh giá các phương pháp máy học:

Ma trận nhầm lẫn (Confusion Matrix) là một kỹ thuật giúp đo lường hiệu suất của một phương pháp máy học. Tính toán ma trận nhầm lẫn mang ý nghĩa so sánh kết quả dự đoán phân loại so với kết quả phân loại thực tế, cung cấp những thông tin hữu ích về điểm đúng và điểm bị lỗi về mô hình. (Xinyang Deng, Qi L. Yong D. Sankaran M., (2016))

Ma trận nhầm lẫn có cấu trúc dạng bảng, với 4 chỉ số đối với mỗi lớp phân loại. Trong khuôn khổ đề tài nghiên cứu, 4 chỉ số mang ý nghĩa như sau:

- TP (True Positive): Số lượng bình luận tích cực được dự đoán đúng
- TN (True Negative): Số lượng bình luận tiêu cực được dự đoán đúng
- FP (False Positive - Type 1 Error): Số lượng bình luận tích cực được dự đoán sai
- FN (False Negative - Type 2 Error): Số lượng bình luận tiêu cực được dự đoán sai

Từ 4 chỉ số trên, ta xác định được các chỉ số đánh giá quan trọng:

+ Accuracy: Tính chính xác của mô hình theo tổng thể

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2-12)$$

+ Precision: Độ chính xác là tỷ lệ giữa các dự đoán Tích cực Đúng trên tổng số dự đoán Tích cực. Precision cao đồng nghĩa độ chính xác của mô hình dự đoán số lượng bình luận Tích cực cao.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2-13)$$

+ Recall: Recall là tỷ lệ giữa các dự đoán Tích cực Đúng trên tổng số các bình luận Tích cực. Recall cao đồng nghĩa với việc bỏ sót các bình luận thực sự Tích cực thấp.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2-14)$$

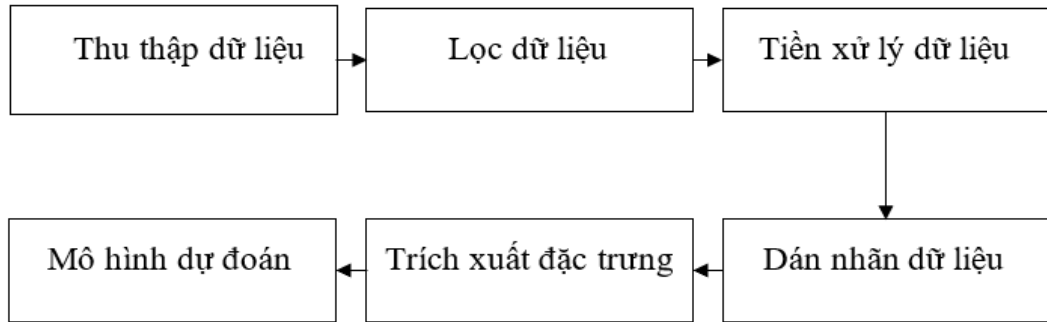
+ F_score: là trung bình điều hòa (harmonic mean) của Precision và Recall, giúp tối ưu hóa cân bằng cả hai chỉ số. F_score càng cao thể hiện Precision và Recall càng cao, mô hình phân lớp tốt.

$$\text{F_score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2-15)$$

CHƯƠNG 3: THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU

Giới thiệu chương

Chương này tập trung vào quá trình cần thực hiện để thu thập, phân tích khám phá dữ liệu để có nhìn tổng quan về dữ liệu để thực hiện các bước tiền xử lý những dữ liệu được thu thập từ các ứng dụng Shopee, Lazada, Tiki, Sendo. Bên cạnh đó, bước quan trọng trong quá trình tiền xử lý để đưa vào phương pháp máy học là dán nhãn dữ liệu. Trích xuất đặc trưng chuyển đổi tập dữ liệu sau khi tiền xử lý ban đầu thành tập các thuộc tính có thể giúp cải thiện độ chính xác của mô hình dự đoán hiện tại. Bước cuối cùng là huấn luyện phương pháp máy học cho tập dữ liệu huấn luyện sẽ được miêu tả chi tiết ở Chương 4.



Hình 3-1: Quy trình xử lý dữ liệu trong bài toán phân tích quan điểm

3.1. Thu thập dữ liệu

Nhóm nghiên cứu tiến hành trích xuất và thu thập các bài đánh giá ứng dụng từ Cửa hàng Google Play của 04 ứng dụng thương mại di động phổ biến: Shopee, Tiki, Lazada và Sendo. Đây là những ứng dụng thương mại di động phổ biến nhất tại Việt Nam với số lượng người sử dụng rất lớn. Nhóm đã thu thập dữ liệu các bài đánh giá của 04 ứng dụng gồm 935.000 bình luận trong khoảng thời gian từ năm 2015 đến nay. Tập dữ liệu khi được lấy về có dạng như bảng dưới đây, với cột “userName” là tên người đăng bình luận, cột “content” chứa nội dung bình luận, cột “at” hiển thị thời gian bình luận được đăng tải, cột “score” thể hiện điểm đánh giá từ 1 đến 5 của người bình luận, cột “thumbsUpCount” cho thấy số lượng người khác đồng tình với đánh giá của chủ bình luận, cột “address” là tên ứng dụng người đó để lại bình luận.

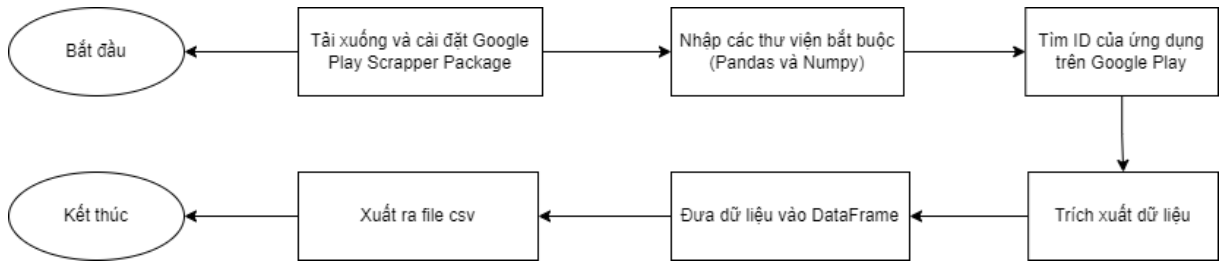
Bảng 3-1: Một phần tập dữ liệu thu thập được từ Sendo

	userName	content	at	score	thumbsUpCount	address
0	Bình Đăng	Tốt	12/14/2021	5	0	Sendo
1	Thu Thanh Hoàng	Tốt	12/14/2021	5	0	Sendo
2	Sinh Nguyen	Lập trình càng ngày càng hoàn thiện	12/14/2021	5	0	Sendo
3	Thành Trinh	Ok	12/14/2021	5	0	Sendo
4	Ky Huynh	Vãi	12/14/2021	5	0	Sendo
...
150249	Duy Dũng Hà	Tạm được	05/23/2015	3	0	Sendo
150250	the do em la	S cứ đứng hoài hà.....đầu trang hông sao kéo ...	05/12/2015	5	17	Sendo
150251	Thiện Minh An Phan	Trang mua sắm rất tốt và uy tín	05/12/2015	5	1	Sendo
150252	Nam Long Nguyen	Rất tốt, ủng hộ sendo	05/09/2015	5	0	Sendo
150253	Tony Doan	Sử dụng rất đơn giản mua hàng cực thích!	05/06/2015	5	0	Sendo

3.1.1. Thư viện sử dụng

Google Play Scraper: Google Play Scraper là một thư viện Python để trích xuất các đánh giá ứng dụng từ Cửa hàng Google Play.

3.1.2. Quá trình thu thập dữ liệu:



Hình 3-2: Quy trình thu thập dữ liệu

- Để bắt đầu quá trình thu thập dữ liệu, đầu tiên chúng tôi tải và cài đặt thư viện *google_play_scraper*.

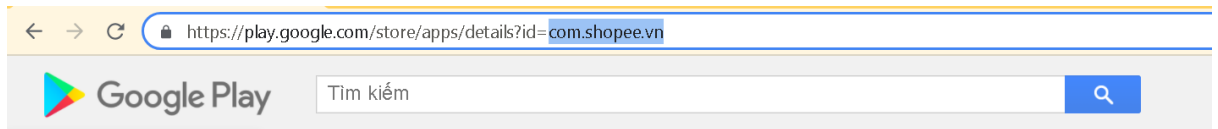
- Ngoài *google_play_scraper*, chúng tôi sẽ cần phải nhập các thư viện cần thiết là *pandas* và *numpy* để đưa dữ liệu vào một dataframe.

```

from google_play_scraper import app
import pandas as pd
import numpy as np
  
```

Hình 3-3: Thiết lập các thư viện cần thiết

- Tìm ID ứng dụng trên Cửa hàng Google Play: ID nằm ở cuối URL của ứng dụng trong Cửa hàng Google Play được biểu thị bằng lựa chọn màu xanh lam trong hình ảnh bên dưới.



Hình 3-4: ID ứng dụng trên Cửa hàng Google Play

- Trích xuất dữ liệu.
- Đưa các dữ liệu đã trích xuất vào DataFrame.
- Xuất ra file csv.

3.1.3. Môi trường thực nghiệm

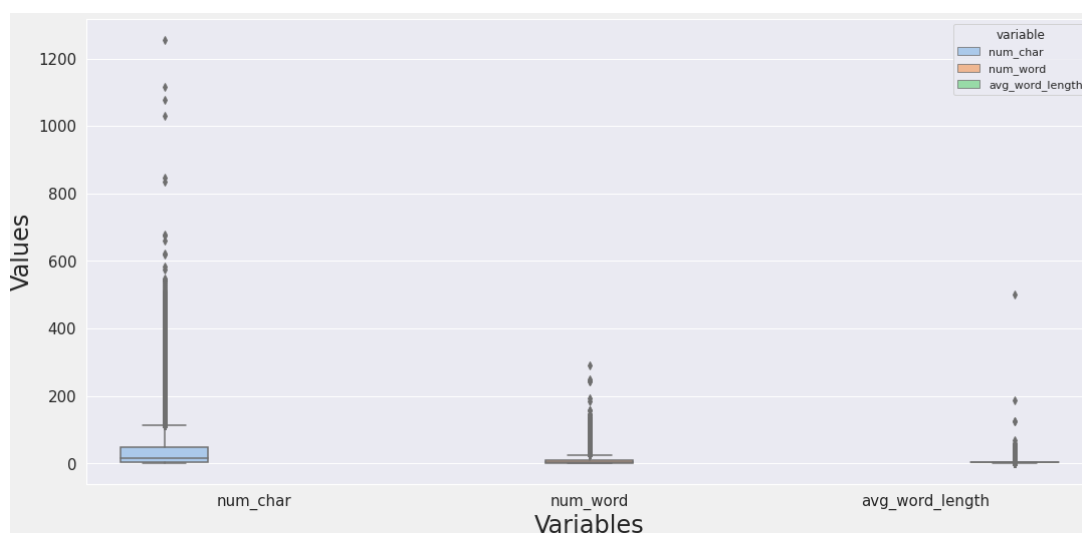
Trong nghiên cứu này, chúng tôi đã tiến hành thực nghiệm thu thập dữ liệu và xử lý các dữ liệu thu thập được với môi trường bao gồm hệ điều hành, vi xử lý, RAM, ngôn ngữ, miền dữ liệu, nguồn dữ liệu, tổng số bài đánh giá thu thập được được mô tả bằng bảng dưới đây:

Bảng 3-2: Môi trường thực nghiệm của nghiên cứu

Hệ điều hành	Windows 10 Education
Vi xử lý	2.40 GHz
RAM	8 Gb
Ngôn ngữ	Python
Miền dữ liệu	Bài đánh giá về 4 ứng dụng thương mại
Nguồn dữ liệu	play.google.com
Tổng số bài đánh giá thu thập được	1107758 bài đánh giá

3.2. Phân tích khám phá dữ liệu (EDA)

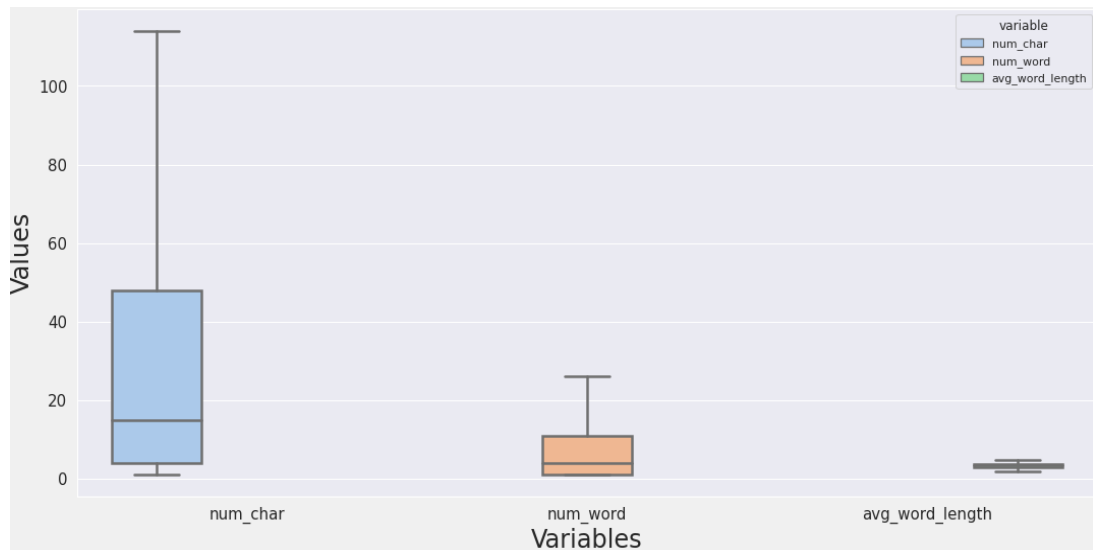
Trước tiên, dữ liệu thô sẽ được phân tích bởi phương pháp phân tích dữ liệu khám phá (Exploratory Data Analysis - EDA) để khái quát hóa dữ liệu một cách tổng quan, tìm ra những đặc điểm chính của dữ liệu, sau đó thông qua tiền xử lý và lấy mẫu, gán nhãn lần thứ nhất trước khi thực hiện các phương pháp máy học.



Hình 3-5: Xác định các ngoại lai của tập dữ liệu

Dựa vào hình 3-5, các biểu đồ hộp (Boxplot) cho thấy dữ liệu thô thu thập từ 04 ứng dụng thương mại di động có rất nhiều đánh giá có số lượng từ hoặc số lượng ký tự rất lớn, các giá trị ngoại lai này sẽ ảnh hưởng đến độ chính xác và gây nhiễu cho các phương pháp máy học. Vì vậy, chúng tôi đã lọc ra các giá trị ngoại lai nằm ngoài chặn trên (upper fence) và chặn dưới (lower fence) của các đánh giá thu thập được bằng cách loại bỏ các giá trị ngoại lai và thay thế các giá trị ngoại lai bằng IQR để cải thiện hiệu

suất và độ tin cậy của các mô hình (kết quả được thể hiện trên hình 3-6). Từ đó, đồng bộ được dữ liệu và tiến hành các bước xử lý tiếp theo.



Hình 3-6: Xử lý các ngoại lai của tập dữ liệu

3.3. Tiền xử lý dữ liệu

Trong lĩnh vực xử lý ngôn ngữ tự nhiên, pha tiền xử lý dữ liệu vô cùng quan trọng và ảnh hưởng đến kết quả phân tích. Khi thu thập bộ dữ liệu từ các ứng dụng thương mại di động, dữ liệu đang ở dạng thô (chưa được qua xử lý, có những trường bị rỗng, sai chính tả, chứa các ký tự đặc biệt hoặc các biểu tượng cảm xúc,...), dữ liệu này sẽ làm giảm độ chính xác của kết quả của các mô hình. Vì vậy, chúng tôi đã tiến hành các bước hiệu chỉnh góp phần tăng độ chính xác cho phân loại.

Xóa dữ liệu bị khuyết: các dữ liệu bị khuyết sẽ tạo nên các điểm dữ liệu bị mập mờ và những điểm dữ liệu gây nhiễu và làm tốn bộ nhớ lưu trữ nên phải xóa đi để tránh ảnh hưởng đến kết quả của các mô hình.

Đưa dữ liệu về chữ thường, loại bỏ số, dấu câu, khoảng trắng thừa và các ký tự lạ: các ký tự lạ, số và dấu câu không mang ý nghĩa phân loại và một phần sẽ gây nhiễu trong quá trình phân tích dữ liệu. Chuyển tất cả về chữ thường: mỗi số, ký tự đặc biệt, ký tự là đại diện cho một dãy nhị phân trong bộ nhớ máy tính. Chữ in hoa sẽ có mã Unicode khác chữ in thường, về mặt ngữ nghĩa là giống nhau tuy nhiên máy tính sẽ không thể phân biệt dữ liệu đầu vào, dẫn đến có thể kết quả dự đoán bị ảnh hưởng.

Hàm **re.sub()**: dùng để xóa các khoảng trắng, ký tự và số trong các đánh giá của khách hàng (.,*&^%\$,0123456789, ...)


Chuẩn hóa một số từ viết tắt, từ sai chính tả, từ kéo dài và xóa các từ dài hơn 7 chữ không có nghĩa: với mục đích là đưa văn bản từ các dạng không đồng nhất về cùng một dạng vì dữ liệu thu thập được là tiếng Việt nên về mặt ngữ nghĩa là rất cần thiết, nếu sai sẽ trở thành một ý nghĩa hoàn toàn khác. Tuy nhiên trong quá trình đánh giá của khách hàng sẽ có những từ viết tắt, viết teencode hoặc sai chính tả. Ví dụ: “k đc” (không được), “nch” (nói chuyện), “qc” (quảng cáo), “bh” (bây giờ),... các từ như vậy sẽ làm dữ liệu không được đồng bộ nên cần chuẩn hóa lại. Điều này sẽ ảnh hưởng tương đối đến kết quả của phương pháp máy học. Làm cho mức độ nhận diện cảm xúc của khách hàng bị ảnh hưởng và khó để dự đoán được kết quả khi chạy với bộ dữ liệu sau quá trình huấn luyện. Từ dài nhất trong tiếng Việt có độ dài là 7 chữ (nghiêng) nên các từ có độ dài hơn 7 chữ nhưng không có ý nghĩa về mặt ngôn ngữ trong hệ thống tiếng Việt sẽ bị xóa đi hoặc là các từ kéo dài hơn 7 chữ sẽ được chuẩn hóa lại. Vì vậy, chúng ta cũng cần chuẩn hóa lại để có thể nhận diện được cảm xúc của đánh giá. Ví dụ: “hayyyyyyy quáaaaaaa điiii” sẽ trở thành “hay quá đi” hoặc “Ghhjjiinyyhvcc” sẽ bị xóa đi.

Tách từ: trong Tiếng Việt, các từ nếu ghép lại sẽ có các ý nghĩa khác nhau so với các từ đứng riêng lẻ và dấu cách lúc này không được sử dụng như ký hiệu để tách từ mà nó có ý nghĩa tách các âm tiết trong một từ. Ví dụ: “thành” và “công” đều có ý nghĩa riêng khi đứng một mình nhưng khi ghép lại thành “thành công” nó sẽ mang một ý nghĩa khác chỉ kết quả tốt đẹp mà mọi người mong muốn đạt được. Vì vậy, tách từ là bước rất có ý nghĩa trong giai đoạn tiền xử lý ngôn ngữ tự nhiên trong bài toán phân loại cảm xúc.

Kết quả sau khi thực hiện tiền xử lý:

Bảng 3-3: Minh họa dữ liệu trước và sau khi tiền xử lý

Trước	Cho 5 sao ủng hộ chứ xài vẫn còn nhiều cái chán lắm:(nên cải thiện	Sau	cho sao ủng_hộ chứ xài vẫn còn nhiều cái chán lắm nên cải thiện
	Đúng nơi mua sắm cho mọi người		đúng nơi mua_sắm cho mọi người
	Hài lòng ,		hài_lòng
	JSC		jsc
	Làm việc chậm chạp		làm_việc chậm_chạp

 nha	ồn nha
Như đồ đánh	như đồ đánh
Ok	ồn
Sendo dùng rất tốt rất rẻ và tiện lợi khi mua qua online	sendo dùng rất tốt rất rẻ và tiện_lợi khi mua qua trực_tuyến
Ồn	ồn
OK! Rất tốt! Rất hài lòng về cách phục vụ của sendo!	ồn rất tốt rất hài_lòng về cách phục_vụ của sendo
Phí ship quá đắt	phí vận_chuyển quá đắt
rất hay	rất hay
Rất mua được nhiều đồ rẻ	rất mua được nhiều đồ rẻ
Tốt nhé ủng hộ ap của Vietnam	tốt nhé ủng_hộ ứng_dụng của việt nam
Tốt sản phẩm đa dạng giá hợp lý	tốt sản_phẩm đa_dạng giá hợp_lý
Càng ngày giá càng mắc so với các sản khác.	càng ngày giá càng mắc so với các sản khác
Cũng đc	cũng được
Dùng rất tốt	dùng rất tốt
Hài lòng	hài_lòng

Sau khi thực hiện tiền xử lý, dữ liệu từ dữ liệu thô đã được chuẩn hóa đưa về cùng một dạng hỗ trợ cho việc huấn luyện và phân tích quan điểm khách hàng thông qua đánh giá để có được một kết quả dự đoán tốt nhất.

3.4. Dán nhãn dữ liệu

Sau quá trình tiền xử lý dữ liệu, nhóm nghiên cứu tiến hành gán nhãn dữ liệu đầu vào. Về cơ bản, việc gán nhãn dữ liệu là một bước quan trọng trong quá trình tiền xử lý dữ liệu trước khi đưa vào các phương pháp máy học, đặc biệt đối với học máy có giám sát. Tập dữ liệu được gán nhãn khi áp dụng các phương pháp máy học nghĩa là máy tính sẽ được đào tạo để ghi nhớ các đặc trưng mỗi một ô dữ liệu của từng loại nhãn dán, khi đó hình thành lên các cơ sở để cung cấp đầu ra và xác thực các phương pháp máy học. Sau khi được huấn luyện bởi dữ liệu được gán nhãn, các mô hình học máy có thể bắt đầu nhận ra các mẫu giống nhau và có thể đưa ra phân loại đối với tập dữ liệu mới được đưa vào.

Để tiến hành phân loại dữ liệu, nhóm đã dựa vào số sao đánh giá của các bình luận theo khung phân loại như sau:

Bảng 3-4: Nhãn phân loại dựa vào số sao đánh giá

Nhãn	Ý nghĩa	Tiêu chí gán nhãn
Neg	Tích cực (Positive)	Đánh giá 4 hoặc 5 sao
Pos	Tiêu cực (Negative)	Đánh giá 1, 2 hoặc 3 sao

Bảng dưới đây, thể hiện các trường dữ liệu mà chúng tôi đã thu thập từ ứng dụng Sendo. Theo thứ tự từ trái qua phải, trường “userName” là cột dữ liệu về tên khách hàng/người bình luận, “content” là cột dữ liệu trình bày nội dung bình luận, “at” là cột dữ liệu về thời gian khách hàng đó bình luận và được đăng tải lên, “score” là cột dữ liệu thể hiện điểm đánh giá từ 1 đến 5 cho ứng dụng, “content_handle” là cột dữ liệu thể hiện những bình luận của khách hàng dành cho sản phẩm, dịch vụ, “sentiment” là cột dữ liệu thể hiện bình luận của khách hàng đó mang tính tích cực hay tiêu cực.

Bảng 3-5: Một phần dữ liệu đã được gán nhãn dựa vào đánh giá của Sendo

userName	content	at	score	address	content_handle	sentiment
Bình Đăng	Tốt	12/14/2021	5	Sendo	tốt	Pos
Thu Thanh Hoàng	Tốt	12/14/2021	5	Sendo	tốt	Pos
Sinh Nguyen	Lập trình càng ngày càng hoàn	12/14/2021	5	Sendo	lập trình càng ngày càng	Pos

Doc_9	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Doc_10	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

TF-IDF chuyển đổi tập dữ liệu sau khi tiền xử lý ban đầu thành tập các thuộc tính (features) có thể giúp biểu diễn tập dữ liệu ban đầu tốt hơn, giúp tương thích với từng mô hình dự đoán cụ thể, cũng như cải thiện độ chính xác của mô hình dự đoán hiện tại. Các thuộc tính trong tập dữ liệu ảnh hưởng trực tiếp đến mô hình dự đoán, do đó ta cần xác định cấu trúc của các thuộc tính sao cho diễn đạt hiệu quả nhất bản chất của tập dữ liệu. Vì dữ liệu của 04 ứng dụng thương mại di động thu thập là lớn nên nếu dùng mô hình Bag of Word có các vector đặc trưng dựa trên tần số tuyệt đối, có thể sẽ có một số từ xuất hiện thường xuyên trên tất cả các mẫu và chúng sẽ có xu hướng làm lu mờ các từ khác. Tuy nhiên, mô hình TF-IDF sẽ giải quyết vấn đề này bằng cách sử dụng hệ số tỷ lệ hoặc chuẩn hóa trong tính toán của nó. Trong nghiên cứu này, sau bước tiền xử lý dữ liệu, phương pháp TF-IDF được áp dụng để xây dựng véc-tơ trọng số thể hiện tần suất xuất hiện của từ trong các bình luận của khách hàng. Bảng 2 trình bày một số mẫu minh họa ma trận tần suất TF_IDF với các dòng thể hiện mỗi bình luận của khách hàng và các cột thể hiện trọng số của từ xuất hiện trong bình luận.

CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ MÔ HÌNH

Giới thiệu chương

Chương này trình bày các phương pháp máy học để thực nghiệm như Hồi quy Logistics, SVM, Naive Bayes và Random Forest tiếp đến là so sánh, đánh giá mô hình phù hợp cho tập dữ liệu. Nhìn vào các đánh giá được trực quan sau khi được áp dụng mô hình và xử lý, kết quả thực nghiệm đã cho thấy hiệu suất của việc phân loại dựa trên quan điểm từ tập dữ liệu về các đánh giá ứng dụng thương mại di động.

4.1. Mô hình dự đoán

Các cách học máy để phân tích quan điểm thường dựa trên các phương pháp phân loại có giám sát, ở bất kỳ đâu dữ liệu khi được gán nhãn thì được sử dụng cho phương pháp này. Phương pháp huấn luyện mô hình tự huấn luyện để thích ứng với một đầu vào cụ thể (văn bản) với đầu ra tương ứng (thẻ) dựa trên dữ liệu mẫu được cung cấp cho mục đích huấn luyện, dựa trên nguyên tắc 80:20. Ở đây 80% dữ liệu được đưa vào ứng dụng với mục đích huấn luyện nó, 20% còn lại tiếp theo cho giai đoạn dự đoán. Đây là tỷ lệ chia tập dữ liệu phổ biến mà nhiều bài nghiên cứu đã ứng dụng (Akshat Verma, Shivam W., Ishwar W., Ritesh W., Radha T. & Sanika Patankar. 2022), (Kiran S Raj & Priyanka Kumar. 2021), (Mahmoud Nabil, Mohamed Aly & Amir F. Atiya. 2015)... và mang lại kết quả chính xác cao (Afshin Gholamy, Vladik K., Olga K., 2018). Chức năng trích xuất đặc trưng là chuyển đầu vào văn bản từ bước trước thành một vectơ đặc trưng, trong đó ma trận nhãn văn bản là được xây dựng và sau đó các vectơ và thẻ tính năng này (ví dụ: tích cực (Pos) hoặc tiêu cực (Neg)) được đưa vào mã học máy hoặc phương pháp máy học sẽ tạo ra một mô hình.

Trong giai đoạn dự đoán, công việc trích xuất đặc trưng là biến đổi các đầu vào văn bản không nhìn thấy thành các vectơ đặc trưng. Các vectơ đặc trưng sau đó được đưa vào mô hình sẽ tạo ra các nhãn được dự đoán hoặc mong đợi (tức là tích cực, phủ định hoặc trung tính) mà mô hình đã huấn luyện cho mẫu dữ liệu 80% ở bước trước.

4.1.1. Hồi quy Logistics

Import mô hình **LogisticRegression()** từ thư viện sklearn của python. Sau đó, dùng hàm `.fit()` để mô hình học từ dữ liệu huấn luyện và kết quả dán nhãn dùng trong huấn luyện. Và hàm `.predict()` để dự đoán kết quả của tập dữ liệu kiểm tra.

Cấu trúc:

```
from sklearn.linear_model import LogisticRegression
```

```
model_lr = log_reg.fit(X_train, y_train)
```

```
predict_lr = model_lr.predict(X_test)
```

Phân tích kết quả:

- Kết quả dự đoán cho tập dữ liệu của 4 ứng dụng: Tiki: 92%, Sendo: 90%, Shopee: 91%, Lazada: 92%.

- Độ chính xác: dự đoán tích cực (Tiki: 93%, Sendo: 92%, Shopee: 93%, Lazada: 94%) và dự đoán tiêu cực (Tiki: 83%, Sendo: 76%, Shopee: 84%, Lazada: 82%)

4.1.2. SVM

Import mô hình SVC() từ thư viện sklearn của python. Sau đó, dùng hàm .fit() để mô hình học từ dữ liệu huấn luyện và kết quả dán nhãn dùng trong huấn luyện. Và hàm .predict() để dự đoán kết quả của tập dữ liệu kiểm tra.

Cấu trúc:

```
from sklearn.svm import SVC
```

```
model_svm = svm.fit(X_train, y_train)
```

```
predict_svm = svm.predict(X_test)
```

Phân tích kết quả:

- Kết quả dự đoán cho tập dữ liệu của 4 ứng dụng: Tiki: 91%, Sendo: 90%, Shopee: 91%, Lazada: 92%

- Độ chính xác: dự đoán tích cực (Tiki: 93%, Sendo: 92%, Shopee: 94%, Lazada: 94%) và dự đoán tiêu cực (Tiki: 82%, Sendo: 76%, Shopee: 84%, Lazada: 82%)

4.1.3. Naive Bayes

Import mô hình **MultinomialNB()** từ thư viện sklearn của python. Sau đó, dùng hàm .fit() để mô hình học từ dữ liệu huấn luyện và kết quả dán nhãn dùng trong huấn luyện. Và hàm .predict() để dự đoán kết quả của tập dữ liệu kiểm tra.

Cấu trúc:

```
from sklearn.naive_bayes import MultinomialNB
```

```
model_naive_bayes = naive_bayes.fit(X_train, y_train)
```

```
predict_naive_bayes = model_naive_bayes.predict(X_test)
```

Phân tích kết quả:

- Kết quả dự đoán cho tập dữ liệu của 4 ứng dụng: Tiki: 91%, Sendo: 89%, Shopee: 90%, Lazada: 91%

- Độ chính xác: dự đoán tích cực (Tiki: 93%, Sendo: 92%, Shopee: 94%, Lazada: 94%) và dự đoán tiêu cực (Tiki: 78%, Sendo: 70%, Shopee: 80%, Lazada: 85%)

4.1.4. Random Forest

Import mô hình **RandomForestClassifier()** từ thư viện sklearn của python. Sau đó, dùng hàm `.fit()` để mô hình học từ dữ liệu huấn luyện và kết quả dán nhãn dùng trong huấn luyện. Và hàm `.predict()` để dự đoán kết quả của tập dữ liệu kiểm tra.

Cấu trúc:

```
from sklearn.ensemble import RandomForestClassifier
model_random_forest = random_forest.fit(X_train, y_train)
predict_random_forest = model_random_forest.predict(X_test)
```

Phân tích kết quả:

- Kết quả dự đoán cho tập dữ liệu của 4 ứng dụng: Tiki: 92%, Sendo: 90%, Shopee: 91%, Lazada: 93%

- Độ chính xác: dự đoán tích cực (Tiki: 94%, Sendo: 92%, Shopee: 94%, Lazada: 94%) và dự đoán tiêu cực (Tiki: 81%, Sendo: 74%, Shopee: 83%, Lazada: 82%)

4.2. So sánh và lựa chọn mô hình

Sau khi trích xuất đặc trưng toàn bộ tập dữ liệu của 04 ứng dụng Thương mại di động, chúng tôi tiến hành huấn luyện phương pháp máy học cho tập dữ liệu huấn luyện. Dữ liệu được lấy mẫu được chia thành 2 nhóm: tập dữ liệu huấn luyện (80%) và tập dữ liệu kiểm tra (20%). Tập dữ liệu huấn luyện được sử dụng để thiết lập bởi phương pháp máy học bao gồm Hồi qui Logistic (LR), Support Vector Machine (SVM), Naïve Bayes (NB) và Random Forest (RF) và sau đó áp dụng ma trận nhầm lẫn (Confusion Matrix) với các độ đo Precision, Recall, F-score và Accuracy để đánh giá kết quả nhằm chọn ra mô hình phù hợp nhất, tiến hành gán nhãn lần thứ hai để áp dụng cho tập dữ liệu kiểm tra.

Kết quả thể hiện trên bảng 4.1 cho thấy rằng, độ chính xác của phương pháp máy học Random Forest là cao nhất (92%), lần lượt theo sau là Logistic Regression, Support Vector Machine (91%) và Naive Bayes (90%) là thấp nhất. Về thời gian huấn luyện mô hình thì phương pháp Naive Bayes là tốn ít thời gian nhất (9.04s), xếp theo sau là Logistic Regression (17.8s), Random Forest (41min 4s) và Support Vector Machine (17h 38min 47s). So về thời gian dự đoán, Logistic Regression là phương pháp sử dụng ít thời gian để dự đoán nhất (1.3s), tiếp theo là Naive Bayes (1.33s), Random Forest (24.3s) và Support Vector Machine (28min 16s). Với những chỉ số đánh giá trên, chúng tôi cho rằng với tập dữ liệu chúng tôi thu thập thì phương pháp máy học Logistic Regression là phù hợp nhất.

Bảng 4-1: Kết quả đánh giá mô hình trên toàn bộ dữ liệu

Phương pháp học máy	Precision		Recall		F_score		Accuracy	Training time	Prediction time
	Pos	Neg	Pos	Neg	Pos	Neg			
LR	0.93	0.83	0.96	0.73	0.95	0.78	0.91	17.8s	1.3s
SVM	0.93	0.83	0.96	0.74	0.95	0.78	0.91	17h 38min 47s	28min 16s
NB	0.94	0.77	0.94	0.75	0.94	0.76	0.90	9.04s	1.33s
RF	0.94	0.82	0.96	0.76	0.95	0.79	0.92	41min 4s	24.3s

Tập dữ liệu của 04 ứng dụng thương mại di động sau khi huấn luyện học máy đã cho ra được bảng kết quả hiệu suất của các mô hình phân tích quan điểm khác nhau (xem bảng 4.2):

Bảng 4-2: Kết quả so sánh các mô hình

Phương pháp máy học/Tập dữ liệu trên các ứng dụng		Precision		Recall		F_score		Accuracy	Training time	Prediction time
		Pos	Neg	Pos	Neg	Pos	Neg			
LR	Tiki	0.93	0.83	0.96	0.73	0.95	0.78	0.92	2.1s	104ms
	Sendo	0.92	0.76	0.97	0.55	0.94	0.63	0.90	2.93s	217ms

	Shopee	0.93	0.84	0.95	0.82	0.94	0.83	0.91	7.56s	603ms
	Lazada	0.94	0.82	0.97	0.67	0.96	0.74	0.92	6.88s	480ms
SVM	Tiki	0.93	0.82	0.96	0.73	0.95	0.77	0.91	1min	6.2s
	Sendo	0.92	0.76	0.97	0.54	0.94	0.63	0.90	15min 44s	34.5s
	Shopee	0.94	0.84	0.94	0.82	0.94	0.83	0.91	1h 21min 39s	4min 33s
	Lazada	0.94	0.82	0.97	0.67	0.96	0.74	0.92	2h 28min 28s	2min 21s
NB	Tiki	0.93	0.78	0.95	0.73	0.94	0.76	0.91	554ms	119ms
	Sendo	0.92	0.70	0.95	0.56	0.94	0.62	0.89	1.35s	218ms
	Shopee	0.94	0.80	0.93	0.83	0.93	0.81	0.90	3.67s	591ms
	Lazada	0.94	0.75	0.96	0.69	0.95	0.72	0.91	3.77s	478ms
RF	Tiki	0.94	0.81	0.95	0.76	0.95	0.78	0.92	32.6s	1.31s
	Sendo	0.92	0.74	0.96	0.55	0.94	0.63	0.90	2min 2s	2.97s
	Shopee	0.94	0.83	0.94	0.84	0.94	0.84	0.91	10min 39s	7.16s
	Lazada	0.94	0.82	0.97	0.70	0.96	0.75	0.93	7min 37s	8.07s

Sau khi thực nghiệm phương pháp máy học trên tập dữ liệu, kết quả cho thấy phương pháp máy học SVM có thời gian huấn luyện và thời gian dự đoán lâu hơn tương đối nhiều so với các phương pháp khác, bởi vì bộ dữ liệu của 04 ứng dụng tương đối lớn nên mất nhiều thời gian để ánh xạ dữ liệu vào một không gian nhiều chiều hơn. Thời gian huấn luyện nhanh nhất Naive Bayes bởi vì phương pháp này chạy dựa trên lý thuyết các biến dữ liệu độc lập với nhau, nhưng độ chính xác lại thấp hơn so với các phương pháp khác khi chạy trên các bộ dữ liệu của 4 ứng dụng thương mại di động thu thập được. Độ chính xác của phương pháp Logistic Regression cao hơn các phương pháp khác trong hầu hết các bộ dữ liệu (Tiki: 92%, Sendo: 90%, Shopee: 91%, and Lazada: 92%), chỉ thua độ chính xác của Random Forest trong bộ dữ liệu Lazada (93%). Tuy nhiên, thời gian huấn luyện và dự đoán lại nhanh hơn rất nhiều so với các phương pháp

Random Forest và SVM. Kết quả cho thấy, Logistic Regression là phương pháp tốt hơn so với các phương pháp khác khi xét về tổng thể thời gian dự đoán, huấn luyện cũng như là độ chính xác khi thực thi. Từ đó cho thấy rằng hồi quy Logistic phù hợp với bộ dữ liệu của 04 ứng dụng thương mại di động.

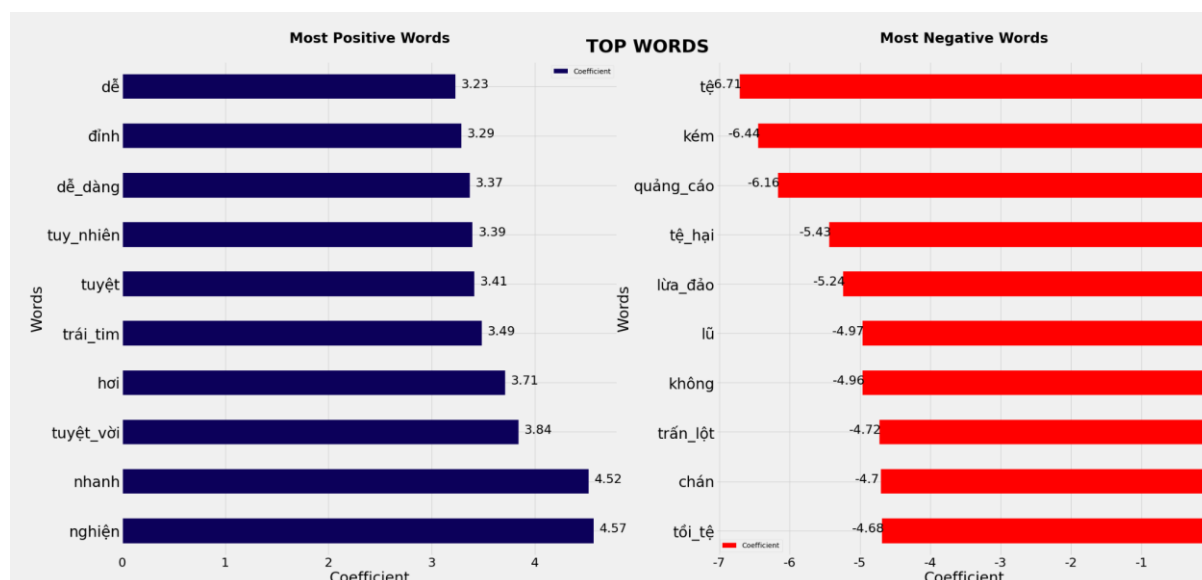
Sau khi mô hình dự đoán được thực nghiệm, các nhãn tích cực và tiêu cực vào các đánh giá của khách hàng được gán. Đánh giá được mô hình dự đoán là tích cực sẽ được gán nhãn “Pos”, ngược lại đánh giá được mô hình dự đoán là tiêu cực sẽ được gán nhãn là “Neg”. Kết quả dự đoán được lưu trong cột “predict”, kết quả dán nhãn theo nhận xét thực tế được lưu trong cột “sentiment”. Kết quả đã thu được và trình bày trên bảng 4.3:

Bảng 4-3: Kết quả dự đoán sau khi thực hiện mô hình

userName	score	address	sentiment	word_tokenize	predict
Công Nhân Trần	5	Sendo	Pos	đúng nơi mua_sắm cho mọi người	Pos
Ninh	5	Sendo	Pos	hài_lòng	Pos
Quang Tường Tô	1	Sendo	Neg	làm_việc chậm_chạp	Neg
Tuan Trancong	5	Sendo	Pos	ồn	Pos
Xuan Nguyen Nguyen	1	Sendo	Neg	phí vận_chuyển quá đắt	Neg

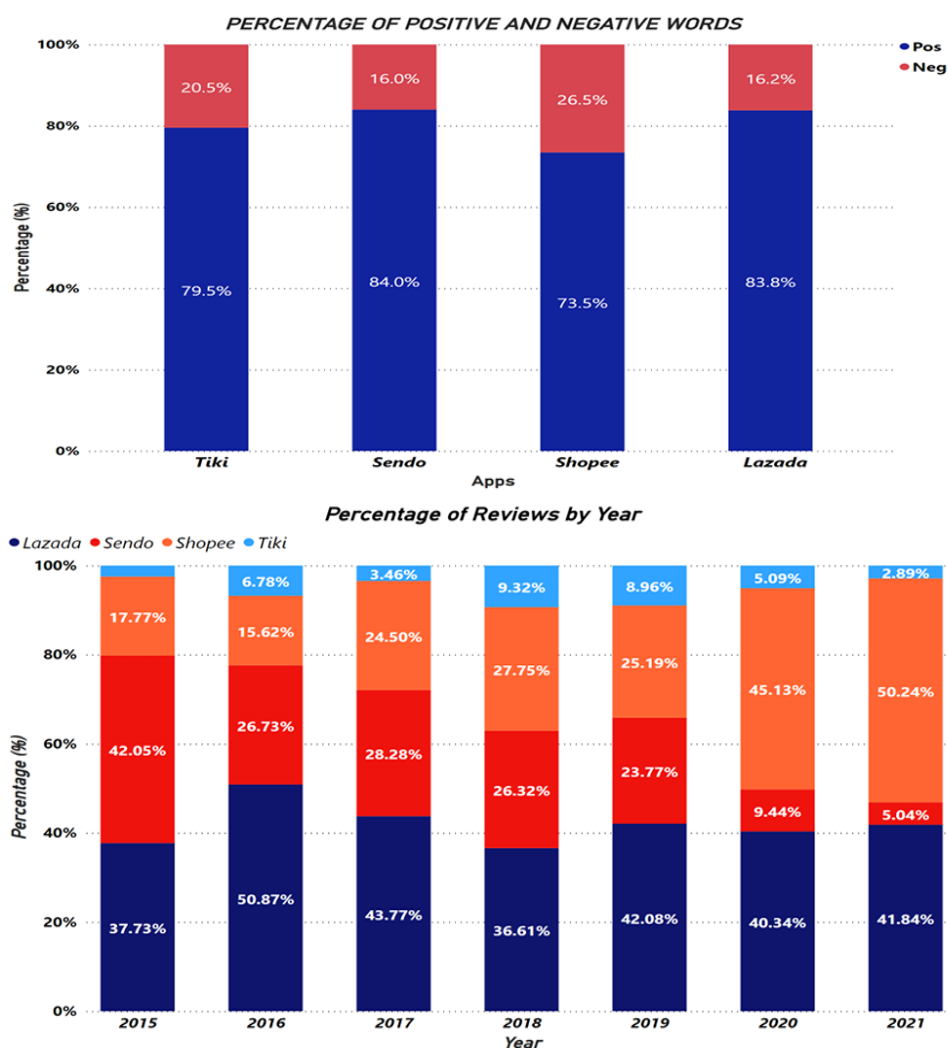
4.3. Kết quả thực nghiệm và thảo luận

Nhìn vào các đánh giá được trực quan sau khi được áp dụng mô hình và xử lý, kết quả thực nghiệm đã cho thấy hiệu suất của việc phân loại dựa trên cảm tính từ tập dữ liệu về các đánh giá ứng dụng thương mại di động. Theo kết quả phân tích, tỷ lệ đánh giá tiêu cực (3 sao trở xuống) và tích cực (4 sao trở lên) trên 04 bộ dữ liệu của 04 ứng dụng thương mại di động tại Việt Nam cho thấy, tỷ lệ đánh giá tích cực của 04 ứng dụng đều ở mức cao (Tiki: 79,54%, Sendo: 83,99%, Shopee: 73,46%, Lazada: 83,80%).



Hình 4-1: Những từ mang quan điểm tích cực và tiêu cực thường xuất hiện

Biểu đồ thanh như trong hình 4-1 được tạo ra để trực quan hóa những từ thường gặp nhất trong số tất cả các bình luận của khách hàng được phân tích. Biểu đồ đã cho thấy cách khách hàng nghĩ và cảm nhận về sản phẩm và dịch vụ trên các ứng dụng thương mại di động. Các từ “nghiện”, “nhanh”, “tuyệt_vời”, “tuyệt”,... là những từ thường gặp nhất trong các đánh giá mang tính tích cực, điều này cho thấy khách hàng có cảm nhận tốt đối với 04 ứng dụng thương mại di động được khảo sát. Từ “nhanh” và “nghiện” là hai trong số những từ thường gặp nhất trong số tất cả các bình luận mang tính tích cực, cho thấy rằng hầu hết khách hàng đều có cảm nhận rất tích cực đối với bốn ứng dụng. Các từ “quảng_cáo”, “tệ”, “chán”, “kém”,... là những từ thường gặp trong các bình luận tiêu cực, điều này cho thấy khách hàng vẫn có những cảm nhận chưa được tốt khi dùng các ứng dụng thương mại di động.



Hình 4-2: Phân bố phần trăm đánh giá theo ứng dụng và theo năm

Hình 4-2 chỉ ra rằng hiệu suất của các đánh giá tích cực vượt trội hơn các đánh giá tiêu cực và trung bình chiếm hơn 75% trong tổng số các đánh giá thu thập được, từ đó, cho thấy số lượng đánh giá tích cực và mức độ hài lòng của khách hàng đối với các hoạt động trên các ứng dụng thương mại di động được khảo sát là cao. Điều này thể hiện sự hài lòng và tin tưởng của người dùng trong lĩnh vực thương mại di động. Nhìn vào ứng dụng Shopee, chúng ta dễ dàng nhận thấy tỷ lệ lượt đánh giá trong năm 2015 đến nay ngày càng tăng lên, chiếm hơn nửa tổng số các lượt đánh giá và đang dẫn đầu trong số các ứng dụng thương mại di động còn lại. Mặt khác, tỷ lệ lượt đánh giá của Sendo từ năm 2015 chiếm vị trí đứng đầu so với các ứng dụng thương mại di động khác nhưng đến năm 2021 đã giảm xuống hơn 8 lần so với trước và giữ vị trí thứ 3 chỉ đứng trước Tiki.

Điều này cũng cho thấy được rằng, về tổng quan Shopee có sự phát triển mạnh mẽ nhất khi thu hút được thêm nhiều khách hàng qua số lượng bình luận được gia tăng

hàng năm. Tiếp theo là về Lazada, khi số lượng bình luận, đánh giá có thể nói khá ổn định qua từng năm, tuy không phát triển nhanh như Shopee nhưng vẫn giữ vững được thị phần của mình và tiếp tục phát triển.



Hình 4-3: WordCloud từ Tích cực

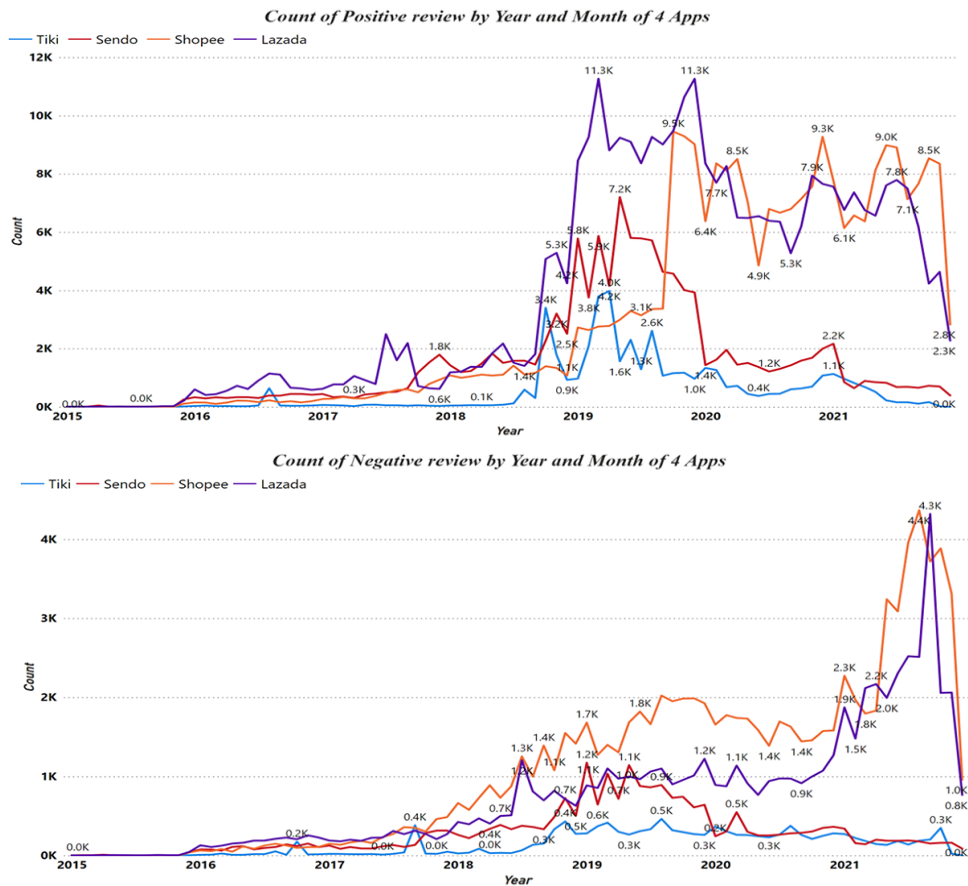


Hình 4-4: WordCloud từ Tiêu cực

Dựa vào hình 4-3 và 4-4, các từ thể hiện quan điểm được sử dụng trong các đánh giá của khách hàng được hình ảnh hóa thông qua WordCloud. Từ đây, có thể thấy rằng các từ được khách hàng dùng nhiều trong các bình luận được thể hiện qua kích thước từ lớn đến nhỏ dần. Cụ thể số lượng từ được dùng nhiều mang ý nghĩa tích cực (hình 4-3) và tiêu cực (hình 4-4), người quản lý có thể hình dung được khách hàng của mình đang quan tâm đến vấn đề gì.

Chẳng hạn như từ “ngon” thì phần nhiều sẽ được hiểu là đang nói về thực phẩm/thức ăn, điều này có nghĩa là thực phẩm được bán trên ứng dụng được đánh giá tốt, còn với từ “dễ”, có nghĩa là dễ sử dụng các ứng dụng. Mặt khác, khi nói đến dịch vụ giao hàng, từ “nhanh” đại diện giao hàng hay thanh toán nhanh sau giao dịch. Những quan điểm hay khía cạnh mà khách hàng quan tâm sẽ giúp người quản lý nắm bắt tâm lý khách hàng một cách nhanh chóng và có chiến lược phát triển hiệu quả cho dịch vụ của mình.

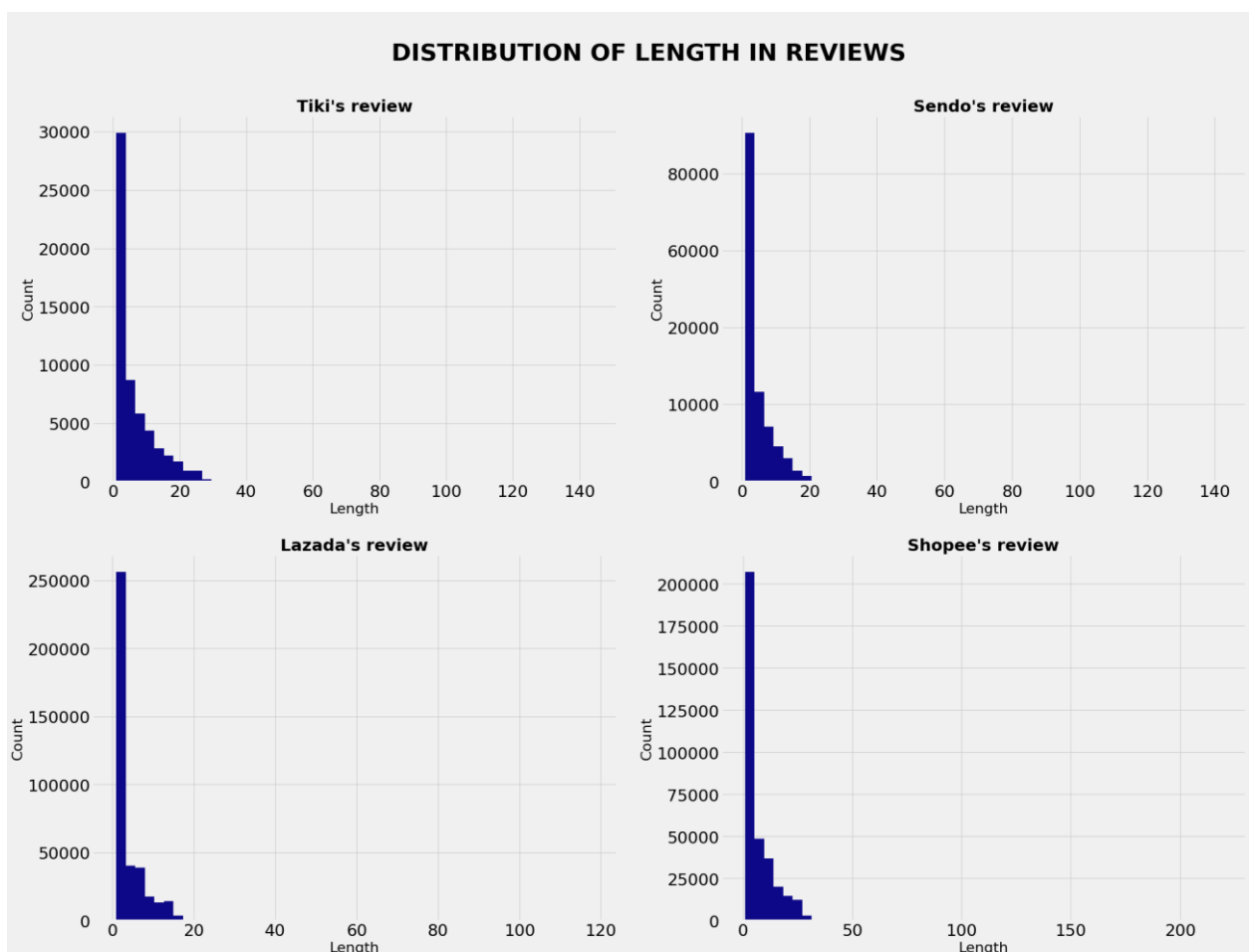
Ở kết quả phân tích từ hình 4-5 cho thấy kết quả về xu hướng đánh giá tiêu cực và tích cực theo từng giai đoạn thời gian, và thể hiện tổng quan về số lượng đánh giá tiêu cực và tích cực tại theo từng tháng từ năm 2015 đến nay. Thông qua hai biểu đồ thể hiện trên hình 4-5, ta có thể thấy năm 2015 đến 2017 thì thương mại di động mới bước đầu xuất hiện tại Việt Nam nên số lượng đánh giá không thu thập được nhiều, nhưng sau đó số lượng đánh giá có sự phát triển theo thời gian. Từ năm 2018, thương mại di động tại Việt Nam tiếp tục phát triển toàn diện với mức tăng trưởng cao 30% so với các năm trước. Shopee là ứng dụng thương mại di động có mức độ hài lòng thấp với số lượng đánh giá tiêu cực liên tục gia tăng, đỉnh điểm là vào tháng 8 năm 2021 với khoảng 38% đánh giá tiêu cực, tiếp theo là Lazada, Sendo, Tiki. Shopee và Lazada là hai ứng dụng có số lượt đánh giá cao và cũng đang phát triển rất mạnh mẽ về mức độ nhận diện cũng như thương hiệu so với Sendo và Tiki.



Hình 4-5: Số lượng đánh giá của khách hàng từ năm 2015 đến 2021

Kết quả trên hình 4-5 cũng cho thấy, với xu hướng thu hút khách hàng mua sắm trực tuyến trong những năm tới cũng như một phần lợi thế do đại dịch Covid-19 tạo ra làm nhu cầu mua sắm trực tuyến tăng lên, bên cạnh thời gian giao hàng và chất lượng

sản phẩm, trải nghiệm ứng dụng đóng vai trò quan trọng trong việc phát triển các ứng dụng thương mại di động đáp ứng nhu cầu mua sắm của khách hàng. Thêm vào đó, với các bình luận tiêu cực, doanh nghiệp có thể tìm ra được lỗ hổng trong các dịch vụ của mình, từ đó cải thiện lại các hoạt động nhằm đem lại trải nghiệm tốt cho khách hàng của mình, điều này giúp doanh nghiệp nhìn vào dữ liệu mà khách hàng để lại trực tiếp từ ứng dụng (lấy dữ liệu tức thời và xử lý) thay vì phải chủ động đi khảo sát thị trường từng nhóm khách hàng, điều này làm mất nhiều thời gian và kém hiệu quả trong việc bắt kịp xu hướng mua sắm của hiện tại.



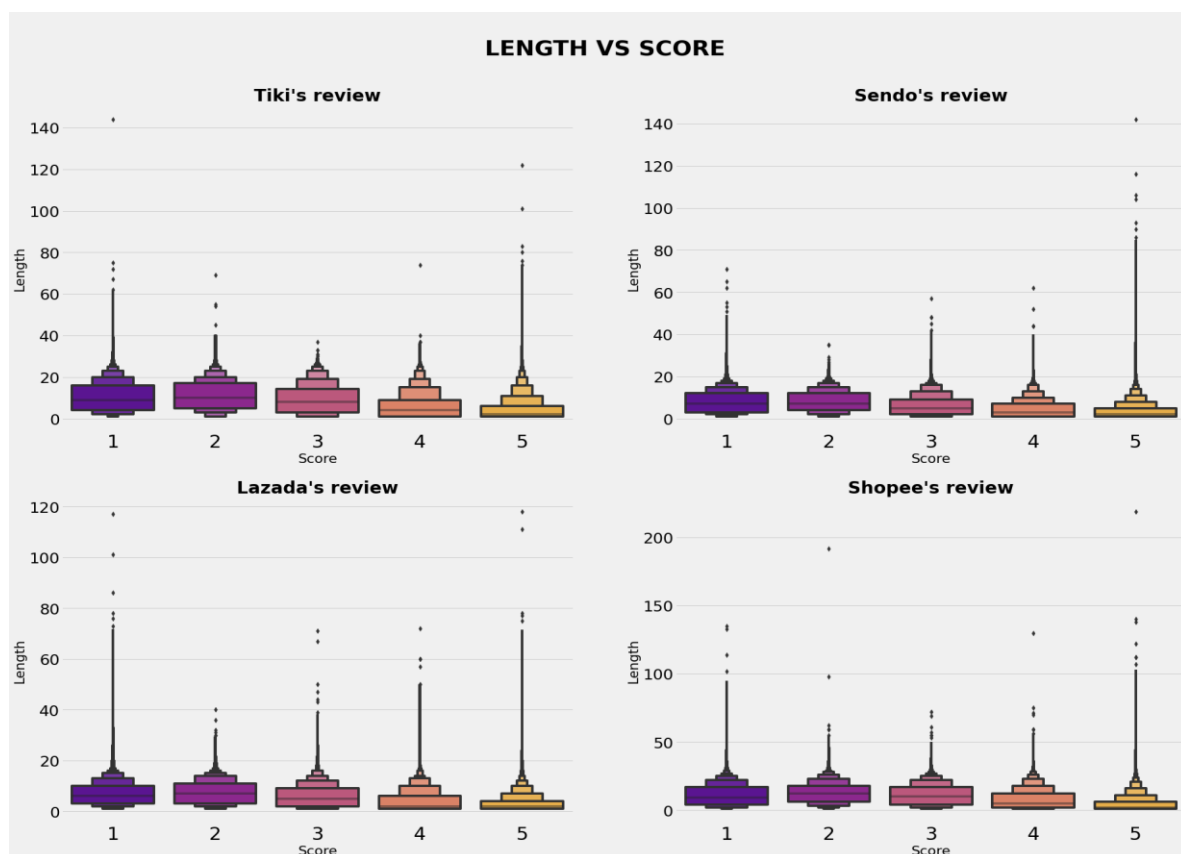
Hình 4-6: Phân bố độ dài của các đánh giá khách hàng

Nhờ vào biểu đồ Histogram (Hình 4-6), chúng ta quan sát được sự phân bố độ dài của các bài đánh giá do khách hàng bình luận. Hầu hết khách hàng viết đánh giá có độ dài nội dung từ 5 đến 30 từ, và chiếm phần lớn trên cả 04 ứng dụng thương mại điện tử là đánh giá trực tiếp từ khách hàng mà không có nội dung bình luận. Quan sát trên cả bốn biểu đồ, cho thấy được sự trải dài theo mức độ giảm dần của số lượng người dùng

khi tiến dần về các nội dung có nhiều từ hơn. Đặc biệt, ở Shopee và Tiki thì có số lượng từ trong một bình luận nhiều hơn Lazada và Sendo. Cùng với tổng tỷ lệ đánh giá có nội dung bình luận của khách hàng khi so sánh với các lượt đánh giá không có nội dung bình luận thì lần lượt là Shopee: 34.03%, Lazada: 30.04%, Tiki: 28.55%, Sendo: 14.17%. Điều này cho thấy được rằng, Shopee và Lazada sẽ nhận được nhiều phản hồi có ý nghĩa đóng góp hơn từ khách hàng.

Tuy nhiên, vẫn sẽ có những bình luận mang tính thương mại và không liên quan đến ứng dụng của doanh nghiệp, bởi vì thông qua quá trình tiền xử lý dữ liệu chúng tôi đã loại bỏ được các bình luận không liên quan đó. Vì thế, tổng quan lại rằng sự phát triển của Shopee và Lazada cũng vượt trội hơn so với 02 ứng dụng thương mại điện tử còn lại là Tiki và Sendo, thông qua lượng bình luận ngày càng tăng cũng như đón nhận được nhiều ý kiến từ khách hàng giúp công ty nhận ra được cần cải thiện hoặc đẩy mạnh phát triển những gì từ đó nâng cao dịch vụ, đáp ứng nhu cầu và làm tăng trải nghiệm của khách hàng.

Với biểu đồ hộp (BoxPlot) dưới đây, chúng ta có thể thấy được mối quan hệ giữa độ dài và điểm đánh giá. Điều đáng chú ý là tất cả các bài đánh giá có độ dài khá giống nhau với tất cả điểm đánh giá khi quan sát ở 04 ứng dụng. Tuy nhiên, có sự khác biệt rõ ràng giữa độ dài của các bài đánh giá xếp hạng thấp và các bài đánh giá xếp hạng cao.



Hình 4-7: Mối quan hệ giữa độ dài và điểm đánh giá của bình luận

Theo biểu đồ và như đã đề cập, các bài đánh giá có điểm xếp hạng thấp có xu hướng dài hơn các bài đánh giá được điểm đánh giá cao. Hầu hết các khách hàng đánh giá ở 04 ứng dụng thương mại di động 5 sao đã viết các bài đánh giá ngắn hơn so với những khách hàng đánh giá 1, 2 hoặc 3 sao. Đó có thể là do khách hàng cảm thấy không hài lòng nên cần phải giải thích lý do không thích sản phẩm, hoặc cần nói về trải nghiệm tệ khi mua hàng để người bán cải thiện. Mặt khác, những khách hàng hài lòng và vui vẻ nhận thấy không nhất thiết phải viết dài dòng, họ sẽ để lại một đánh giá ngắn gọn về chất lượng sản phẩm và dịch vụ của ứng dụng thương mại di động mà họ đã sử dụng để mua sắm.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết quả đạt được

Sau quá trình nghiên cứu này, mô hình phân tích quan điểm được đề xuất và thực nghiệm trên bộ dữ liệu trong lĩnh vực thương mại di động và kết quả đánh giá mô hình với tính chính xác cao. Việc tiến hành thực nghiệm và so sánh 04 phương pháp máy học (Naive Bayes, SVM, Hồi quy Logistic và Random Forest) đã phản ánh những đặc điểm, thuộc tính của từng phương pháp máy học đối với bộ dữ liệu nghiên cứu thông qua các chỉ số của ma trận nhầm lẫn như Accuracy, F_Score. Đối với bộ dữ liệu đã thu thập được, SVM có tổng thời gian huấn luyện, dự đoán và tính chính xác cao hơn, là phương pháp tối ưu và phù hợp với bộ dữ liệu nghiên cứu. Các kết quả từ việc phân loại quan điểm (tích cực và tiêu cực) qua bình luận khách hàng đã cho thấy nhiều thông tin hữu ích về tâm lý và hành vi khách hàng, giúp các doanh nghiệp xác định được nhu cầu khách hàng và đưa ra gợi ý sản phẩm phù hợp cho những khách hàng tiềm năng.

Bộ dữ liệu thực nghiệm được thu thập trong khoảng thời gian từ năm 2015 đến năm 2021, đảm bảo tính thực tế và phản ánh được sự biến động trong thời gian dài. Điều này giúp các nhà phân tích tìm ra được xu hướng thị trường và đưa ra các dự đoán cho tương lai, từ đó có những chiến lược đầu tư hợp lý giảm thiểu rủi ro ở mức thấp nhất. Có thể thấy, việc ứng dụng phân tích quan điểm qua những bình luận bằng các phương pháp học máy và xử lý ngôn ngữ tự nhiên trở thành một giải pháp phân tích mang tới kết quả tối ưu, giúp thấu hiểu nhu cầu của người dùng (customer insights) qua chính những dòng bình luận của họ về sản phẩm hay dịch vụ.

Qua những báo cáo, biểu đồ trực quan hóa dữ liệu, các doanh nghiệp có thể phân tích những mong muốn của khách hàng về nhiều khía cạnh trên các ứng dụng thương mại di động (chất lượng sản phẩm, đề xuất tìm kiếm, phản hồi, dịch vụ vận chuyển, ưu đãi giảm giá, vv...), từ đó đưa ra những chiến lược tối ưu nâng cao trải nghiệm khách hàng và tăng lợi thế cạnh tranh của thương hiệu. Bên cạnh đó, mô hình của nghiên cứu có thể được tích hợp thêm vào ứng dụng nhằm khảo sát quan điểm khách hàng đối với nhiều lĩnh vực, dịch vụ, sản phẩm khác nhau của các doanh nghiệp.

5.2. Hạn chế

Đạt được nhiều kết quả tích cực, nghiên cứu còn tồn tại một vài điểm hạn chế. Trong giai đoạn gán nhãn dữ liệu, việc dán nhãn được thực hiện khá đơn giản và mang tính chủ quan, xác định tính tích cực hoặc tiêu cực dựa trên số sao được khách hàng đánh giá kèm với bình luận đó. Do đó, bước này có thể bị ảnh hưởng dẫn đến sai lệch và bỏ qua một số dữ liệu.

Bài nghiên cứu tập trung phân loại quan điểm theo hai nhãn là “tích cực” và “tiêu cực”, điều này hạn chế phân tích sâu hơn về cảm xúc và tâm lí con người, bởi trên thực tế, còn rất nhiều cảm xúc của người dùng đối với dịch vụ và sản phẩm như: vui vẻ, hài lòng, sợ hãi, buồn chán....

Tuy đưa ra được nhiều kết quả phân tích và thảo luận về quan điểm khách hàng trên các ứng dụng thương mại di động, các nhà quản trị cần có thêm những góc nhìn về chiến lược doanh nghiệp, kinh doanh và phát triển sản phẩm để có thể đưa ra những kết luận và quyết định chính xác. Những kết quả bài nghiên cứu đưa ra chỉ là sự phân tích về mặt kỹ thuật. Bên cạnh đó, dữ liệu thực nghiệm được tập trung thu thập từ 04 ứng dụng thương mại di động phổ biến nhất tại Việt Nam, chưa phản ánh được đa dạng về trải nghiệm khách hàng trên các nền tảng thương mại di động. Hiện nay, có rất nhiều nền tảng khác đang không ngừng phát triển mạnh mẽ.

5.3. Hướng phát triển

Trong tương lai, nhóm đề xuất một số hướng phát triển cho đề tài như sau:

Thứ nhất, nâng cao hiệu suất và tính chính xác của mô hình bằng cách nghiên cứu và khai thác phương pháp BERT (Bidirectional Encoder Representation from Transformer), trích xuất các khía cạnh đặc trưng với kỹ thuật gắn thẻ từng phần (POS Tagging) và phát triển mô hình phân tích các biểu tượng cảm xúc trong phần bình luận.

Thứ hai, tiếp tục nghiên cứu tối ưu hóa mô hình để xây dựng một hệ thống đánh giá, phân loại đánh giá người dùng với cơ chế hoạt động liên tục, thu thập dữ liệu, áp dụng mô hình phân loại đưa ra các báo cáo trực quan hỗ trợ doanh nghiệp ra quyết định, định hướng rõ hơn về cơ cấu, cách thức tổ chức hoạt động vừa đạt doanh thu tối đa vừa xây dựng niềm tin, lòng trung thành của khách hàng đối với sản phẩm, dịch vụ.

Thứ ba, mở rộng mô hình nghiên cứu sang các lĩnh vực khác không chỉ dừng lại ở Thương mại di động. Khai thác mọi tương tác và bình luận của khách hàng trực tuyến với mục đích mang lại những lợi ích tối đa cho doanh nghiệp và người dùng.

DANH MỤC CÔNG TRÌNH CÔNG BỐ

Nguyễn Trần Thúy Quỳnh, Bùi Nguyễn Bích Ngọc, Nguyễn Thị Bảo Trâm, Trần Nhật Nguyên, Võ Bá Tùng và Hồ Trung Thành, *Mô hình khám phá trải nghiệm khách hàng dựa trên phương pháp phân tích quan điểm và máy học*, Tạp chí phát triển khoa học công nghệ Đại học Quốc Gia, đã gửi bài ngày 20/3/2022.

“Nghiên cứu này được Trường Đại học Kinh tế - Luật, ĐHQG-HCM tài trợ trong đề tài có mã số SV 2022 228”

TÀI LIỆU THAM KHẢO

- (1) Taboada, Maite & Brooke, Julian & Tofiloski, Milan & Voll, Kimberly & Stede, Manfred. (2011). *Lexicon-Based Methods for Sentiment Analysis*. Computational Linguistics. 37. 267-307. 10.1162/COLI_a_00049.
- (2) Singla, Zeenia & Randhawa, Sukhchandan & Jain, Sushma. (2017). *Sentiment analysis of customer product reviews using machine learning*. 1-5. 10.1109/I2C2.2017.8321910.
- (3) Kumar, R. (2021, July 10). *Amazon Product Review Sentiment Analysis with Machine Learning : Ravi Kumar Singh / Dr. Kamalraj Ramalingam : Free Download, Borrow, and Streaming*. Internet Archive. Retrieved March 19, 2022, from <https://archive.org/details/httpswww.ijtsrd.comcomputer-sciencedata-processing42372amazon-product-review-sen>
- (4) Jansher, Rabnawaz. (2020). *Sentimental Analysis of Amazon Product Reviews Using Machine Learning Approach*. 10.13140/RG.2.2.36392.80645.
- (5) Wassan, Sobia & Chen, Xi & Shen, Tian & Waqar, Muhammad & Zaman, Noor. (2021). *Amazon Product Sentiment Analysis using Machine Learning Techniques*. 30. 695-703. 10.24205/03276716.2020.2065.
- (6) Noori, Behrooz. (2021). *Classification of Customer Reviews Using Machine Learning Algorithms*. Applied Artificial Intelligence. 35. 1-22. 10.1080/08839514.2021.1922843.
- (7) R. Nagamanjula and A. Pethalakshmi, "A Machine Learning Based Sentiment Analysis by Selecting Features for Predicting Customer Reviews," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 1837-1843, doi: 10.1109/ICCONS.2018.8663153.
- (8) Kang, H., Yoo, S., & Han, D. (2012). *Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews*. *Expert Systems With Applications*, 39(5), 6000-6010. doi: 10.1016/j.eswa.2011.11.107
- (9) Sharma, Anuj & Dey, Shubhamoy. (2012). *A comparative study of selection and machine learning techniques for sentiment analysis*. Proceeding of the 2012 ACM Research in Applied Computation Symposium, RACS 2012. 1-7. 10.1145/2401603.2401605.

- (10) Zhang, K., Cheng, Y., Xie, Y., Honbo, D., Agrawal, A., & Palsetia, D. et al. (2011). *SES: Sentiment Elicitation System for Social Media Data*. 2011 IEEE 11Th International Conference On Data Mining Workshops. doi: 10.1109/icdmw.2011.153
- (11) Agarwal, Basant & Mittal, Namita. (2014). *Semantic Feature Clustering for Sentiment Analysis of English Reviews*. IETE Journal of Research. 60. 414-422. 10.1080/03772063.2014.963172.
- (12) Mitra, Ayushi. (2020). *Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)*. Journal of Ubiquitous Computing and Communication Technologies. 2. 145-152. 10.36548/jucct.2020.3.004.
- (13) Naz, S., Sharan, A., & Malik, N. (2018). *Sentiment Classification on Twitter Data Using Support Vector Machine*. 2018 IEEE/WIC/ACM International Conference On Web Intelligence (WI). <https://doi.org/10.1109/wi.2018.00-13>
- (14) Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2022). *Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches*. Retrieved 14 February 2022, from <https://scholar.smu.edu/datasciencereview/vol1/iss4/7/>
- (15) Yadav, Nikhil & Kudale, Omkar & Gupta, Srishti & Rao,, Aditi & Shitole, Ajitkumar. (2020). *Twitter Sentiment Analysis Using Machine Learning For Product Evaluation*. 10.1109/ICICT48043.2020.9112381.
- (16) D'souza, S., & Sonawane, K. (2019). *Sentiment Analysis Based on Multiple Reviews by using Machine learning approaches*. 2019 3Rd International Conference On Computing Methodologies And Communication (ICCMC). <https://doi.org/10.1109/iccmc.2019.8819813>
- (17) Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. (2018). *Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes*. 2018 International Conference On Orange Technologies (ICOT). <https://doi.org/10.1109/icot.2018.8705796>
- (18) Yiran, Y., & Srivastava, S. (2019). *Aspect-based Sentiment Analysis on mobile phone reviews with LDA*. Proceedings Of The 2019 4Th International Conference On Machine Learning Technologies. <https://doi.org/10.1145/3340997.3341012>

- (19) Bằng, N., Hồ, N., & Thành, H. (2021). *Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm*. *KINH TẾ VÀ QUẢN TRỊ KINH DOANH*, 16(1), 64-78. doi: 10.46223/hcmcoujs.econ.vi.16.1.1388.2021
- (20) Nguyễn, T., & Trần, T. (2019). *Một mô hình học máy trong phân tích ý kiến khách hàng dựa trên văn bản tiếng Việt: Bài toán dịch vụ Khách sạn*. Nhà Xuất Bản Đà Nẵng. Retrieved from <http://thuvien.vku.udn.vn/handle/123456789/975>
- (21) Le, B., & Nguyen, H. (2015). Twitter Sentiment Analysis Using Machine Learning Techniques. *Advanced Computational Methods For Knowledge Engineering*, 279-289. doi: 10.1007/978-3-319-17996-4_25
- (22) Trinh, Son & Nguyen, Luu & Vo, Minh & Do, Phuc. (2016). *Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese Language*. 10.1007/978-3-319-31277-4_23.
- (23) Edgar, T., & Manz, D. (2017). Exploratory Study. *Research Methods For Cyber Security*, 95-130. doi: 10.1016/b978-0-12-805349-2.00004-2
- (24) Misra, S., & Li, H. (2020). *Noninvasive fracture characterization based on the classification of sonic wave travel times*. *Machine Learning for Subsurface Characterization*, 243–287. doi:10.1016/b978-0-12-817736-5.00009-0
- (25) Liu, B. (2011). Web Data Mining. <https://doi.org/10.1007/978-3-642-19460-3>
- (26) Manu Konchady, Text Mining Application Programming (Programming Series), May 2006, ISBN 1584504609
- (27) Matthew A. Russell, Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, 2011, ISBN 1449388345.
- (28) Ian H. Witten, Eibe Frank, and Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems), 2011.
- (29) Le, Ngoc-Bao-Van, and Jun-Ho Huh. (2021). "Applying Sentiment Product Reviews and Visualization for BI Systems in Vietnamese E-Commerce Website: Focusing on Vietnamese Context" *Electronics* 10, no. 20: 2481. <https://doi.org/10.3390/electronics10202481>

(30) Vietnam E-commerce Index 2021 Report. VietNam E-Commerce Association. (2022). Retrieved 19 March 2022, from <http://en.vecom.vn/vietnam-e-commerce-index-2021-report>.

(31) Ahmed, H., Awan, M., Khan, N., Yasin, A. and Faisal Shehzad, H., 2021. Sentiment Analysis of Online Food Reviews using Big Data Analytics. *Elementary Education Online*, [online] 20(2), pp.827-836. Available at: <<https://ssrn.com/abstract=3827110>>.

(32) Nasim, Z., Rajput, Q. and Haider, S., 2017. Sentiment analysis of student feedback using machine learning and lexicon based approaches. *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*,

(33) The value and role of data in electronic commerce and the digital economy and its implications for inclusive trade and development [Internet]. UNCTAD; 2019. Available from: https://unctad.org/system/files/official-document/tdb_ed3d2_en.pdf

(34) Ritter, Thomas; Pedersen, Carsten Lund (2019). *Digitization capability and the digitalization of business models in business-to-business firms: Past, present, and future*. Industrial Marketing Management, (), S0019850119300999–. doi:10.1016/j.indmarman.2019.11.019

(35) Liu, B., 2012. *Sentiment analysis and opinion mining*. San Rafael, Calif.: Morgan & Claypool.

(36) Sharma, R., Nigam, S. and Jain, R., 2014. Opinion Mining In Hindi Language: A Survey. *International Journal in Foundations of Computer Science & Technology*, 4(2), pp.41-47.

(37) Bang, B. and Lee, L., 2008. *Opinion mining and sentiment analysis*. *Foundations and Trends in Information Retrieval*, 2(1-2), pp.1-135.

(38) Wanner, F., Rohrdantz, C., Mansmann, F., Oelke, D. and Keim, D., 2009. Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008. *Visual Interfaces to the Social and the Semantic Web (VISSW 2009)*,.

(39) Binali, H., Potdar, V. and Wu, C., 2009. A state of the art opinion mining and its application domains. *2009 IEEE International Conference on Industrial Technology*,.

- (40) Liu, B., 2012. *Sentiment analysis and opinion mining*. San Rafael, Calif.: Morgan & Claypool.
- (41) Al-Otaibi, S., Alnassar, A., Alshahrani, A., Al-Mubarak, A., Albugami, S., Almutiri, N. and Albugami, A., 2018. *Customer Satisfaction Measurement using Sentiment Analysis*. International Journal of Advanced Computer Science and Applications, 9(2).
- (42) Li Z, Fan Y, Jiang B, Lei T, Liu W. *A survey on sentiment analysis and opinion mining for social multimedia*. Multimedia Tools and Applications. 2018;78(6):6939-6967.
- (43) Pang B, Lee L, Vaithyanathan S. *Thumbs up?*. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02. 2002;.
- (44) Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. *Lexicon-Based Methods for Sentiment Analysis*. Computational Linguistics. 2011;37(2):267-307.
- (45) Melville P, Gryc W, Lawrence R. *Sentiment analysis of blogs by combining lexical knowledge with text classification*. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09. 2009;.
- (46) Ding X, Liu B, Yu P. *A holistic lexicon-based approach to opinion mining*. Proceedings of the international conference on Web search and web data mining - WSDM '08. 2008;.
- (47) Poria S, Chaturvedi I, Cambria E, Bisio F. *Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis*. 2016 International Joint Conference on Neural Networks (IJCNN). 2016;.
- (48) Ruder S, Ghaffari P, G. Breslin J. *A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis*. EMNLP. 2016;;7.
- (49) Hutto C, Gilbert E. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. ICWSM. 2015;.
- (50) Domingos, P. and Pazzani, M., 1997. *Machine Learning*, 29(2/3), pp.103-130.
- (51) Maalouf, M., 2011. *Logistic regression in data analysis: an overview*. International Journal of Data Analysis Techniques and Strategies, 3(3), p.281.

(52) Cutler, A., Cutler, D. and Stevens, J., 2012. *Random Forests. Ensemble Machine Learning*, pp.157-175.

(53) Barbosa, L. and Feng, J., 2010. *Robust sentiment detection on Twitter from biased and noisy data*. Proceedings of the 23rd International Conference on Computational Linguistics, pp.36-44.

(54) Nandi, A. and Sharma, P., 2021. Comparative Study of Sentiment Analysis Techniques. *Interdisciplinary Research in Technology and Management*, pp.456-460.

(55) Mudinas, A., Zhang, D. and Levene, M., 2012. *Combining lexicon and learning based approaches for concept-level sentiment analysis*. Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining - WISDOM '12,.

(56) Hotho, Andreas & Nürnberger, Andreas & Paass, Gerhard. (2005). *A Brief Survey of Text Mining*. LDV Forum - GLDV Journal for Computational Linguistics and Language Technology. 20. 19-62.

(57) Feldman, Ronen & Ronen, & Sanger, & James,. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*.

(58) Wiebe, Janyce & Wilson, Theresa & Bruce, Rebecca & Bell, Matthew & Martin, Melanie. (2004). *Learning Subjective Language*. Computational Linguistics. 30. 277-308. 10.1162/0891201041850885.

(59) Keselj, V. (2009). *Speech and Language Processing* (second edition) Daniel Jurafsky and James H. Martin (Stanford University and University of Colorado at Boulder) Pearson Prentice Hall, 2009, xxxi+988 pp; hardbound, ISBN 978-0-13-187321-6, \$115.00. *Computational Linguistics*, 35(3), 463-466. doi: 10.1162/coli.b09-001

(60) Edgar, T., & Manz, D. (2017). Exploratory Study. *Research Methods For Cyber Security*, 95-130. doi: 10.1016/b978-0-12-805349-2.00004-2

(61) Turney, Peter & D., Peter & Littman, & Littman, Michael. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*. 21. 315-. 10.1145/944012.944013.

(62) Dave, Kushal & Lawrence, Steve & Pennock, David. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. 775152. 10.1145/775152.775226.

(63) Kanayama A, Seth RB, Sun L, Ea CK, Hong M, Shaito A, Chiu YH, Deng L, Chen ZJ. TAB2 and TAB3 activate the NF-kappaB pathway through binding to polyubiquitin chains. Mol Cell. 2004 Aug 27;15(4):535-48. doi: 10.1016/j.molcel.2004.08.008. PMID: 15327770.

(64) Akshat Verma, Shivam W., Ishwar W., Ritesh W., Radha T. & Sanika Patankar. (2022). Sentiment Analysis using Transformer Based Pre-Trained Models for the Hindi Language.

(65) G.A.I.T. Wijewickrama. (2020) Analysis of social media feedback to gain profit for business organizations using sentiment analysis techniques.