TRƯỜNG ĐẠI HỌC SÀI GÒN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG NGHIÊN CỨU KHOA HỌC

Phương Pháp Nghiên Cứu Khoa Học

PHÂN TÍCH CẢM XÚC TRONG ĐÁNH GIÁ SẢN PHẨM SỬ DỤNG PHƯƠNG PHÁP HỌC MÁY

Giảng Viên Hướng Dẫn : ThS. Đỗ Như Tài

Nhóm Sinh Viên Thực Hiện

Diệp Thụy An	**3122410001**
Đỗ Mai Anh	**3122400006**
Cao Tiến Cường	**3122410043**
Võ Hoàng Phúc Hy	**3123410142**

Tháng 4/2025 - Thành Phố Hồ Chí Minh

MỤC LỤC

LỜI CẢM ƠN

Nhóm chúng em xin chân thành cảm ơn **Trường Đại học Sài Gòn** đã tạo điều kiện thuận lợi về cơ sở vật chất và môi trường học tập để chúng em có thể thực hiện và hoàn thành bài báo cáo này.

Đặc biệt, nhóm xin gửi lời cảm ơn sâu sắc đến **TS.Đỗ Như Tài** – giảng viên hướng dẫn, người đã tận tình chỉ bảo, định hướng và hỗ trợ chúng em trong suốt quá trình thực hiện đề tài. Những ý kiến đóng góp quý báu của Thầy là kim chỉ nam giúp nhóm hoàn thiện bài báo cáo một cách tốt nhất.

Nhóm đã nỗ lực hết mình trong quá trình học tập, nhưng do hạn chế về kiến thức và kinh nghiệm, bài báo cáo chắc chắn vẫn còn những thiếu sót. Nhóm rất mong nhận được sự góp ý từ Thầy để hoàn thiện hơn trong những lần sau.

Nhóm chúng em xin chân thành cảm ơn!

ĐỀ CƯƠNG NGHIÊN CỨU KHOA HỌC

1. Lý do chọn đề tài

Trong thời đại số hóa hiện nay, các nền tảng thương mại điện tử đang ngày một phát triển và nắm giữ một vai trò quan trọng trong việc kết nối các doanh nghiệp với khách hàng. Một trong những yếu tố ảnh hưởng trực tiếp đến quyết định mua hàng của khách hàng đó chính là những đánh giá, nhận xét từ những khách hàng trước đó.

Với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI) và học máy (Machine Learning - ML), các ứng dụng trong xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) ngày càng phổ biến, đặc biệt là trong lĩnh vực phân tích dữ liệu người dùng. Một trong những ứng dụng quan trọng của NLP là phân tích cảm xúc (Sentiment Analysis - SA), giúp đánh giá quan điểm của người dùng dựa trên các bài đánh giá sản phẩm trực tuyến [1].

Phân tích cảm xúc (Sentiment Analysis - SA) là một phương pháp khai thác thông tin phản hồi dưới dạng văn bản và là một lĩnh vực quan trọng của xử lý ngôn ngữ tự nhiên. Với nhiều ứng dụng, SA được thực hiện nhằm xác định cảm xúc (tích cực, tiêu cực hoặc trung tính) đối với một thực thể theo quan điểm con người [2]. Theo Liu (2022), SA còn được gọi là khai thác ý kiến (opinion mining), là lĩnh vực nghiên cứu phân tích ý kiến, được thể hiện với các thành phần chính gồm thực thể (sản phẩm, dịch vụ, sự kiện,...), khía cạnh của thực thể (chất lượng, giá cả, dịch vụ hỗ trợ,...), cảm xúc đối với khía cạnh được đề cập (tích cực, tiêu cực hoặc trung tính), người thể hiện ý kiến và thời gian thể hiện ý kiến [2].

Trong nghiên cứu này, nhóm sẽ tập trung vào việc xây dựng và so sánh các mô hình Machine Learning, thông qua các thí nghiệm so sánh giữa các mô hình để đánh giá độ hiểu quả của các mô hình trong việc phân loại cảm xúc của người dùng dựa trên văn bản đánh giá và điểm số đánh giá.

2. Tổng quan vấn đề nghiên cứu

2.1. Tình hình nghiên cứu hiện tại

Việc phân tích cảm xúc của người dùng hiện nay gặp phải nhiều thách thức. Một trong những vấn đề lớn là các mô hình học máy khác nhau sẽ cho ra những kết quả khác nhau và việc lựa chọn mô hình phù hợp nhất để đạt được hiệu quả là vô cùng quan trọng. Bên cạnh đó, quá trình tiền xử lý văn bản cũng đóng vai trò then chốt trong việc nâng cao chất lượng phân tích. Với số lượng đánh giá sản phẩm trực tuyến tăng cao, việc xử lý một khối lượng lớn các đánh giá theo thu công là bất khả thi. Trong những năm gần đây, nhiều công trình đã được công bố, đem lại sự đóng góp lớn trong việc cải thiện độ chính xác và hiệu quả cao.

Ở nước ngoài, các nghiên cứu tập trung vào việc cải thiện độ chính xác và hiệu quả của các mô hình phân tích cảm xúc. Năm 2022, nghiên cứu của Zhao và Sun đã sử dụng **BERT (Bidirectional Encoder Representations from Transformers)** với **fine-tuning BERT** để phân tích cảm xúc từ các đánh giá sản phẩm trực tuyến [3]. Hướng nghiên cứu của họ cho thấy việc tận dụng khả năng trích xuất đặc trưng có thể cải thiện độ chính xác cao trong việc phân loại cảm xúc.

Vào năm 2023, với hướng tiếp cận khác của Aravindan và nhóm tác giả của mình đã sử dụng **PySpark**, một công cụ xử lý dữ liệu lớn, để phân loại cảm xúc đánh giá người dùng [4]. Nghiên cứu này cho thấy việc sử dụng các phương pháp phân tích dữ liệu đem loại hiệu quả tốt nhưng vẫn còn tồn tại nhiều thách thức về hiệu suất và tối ưu. Các nghiên cứu quốc tế ứng dụng ML và NLP để phân tích cảm xúc, ứng dụng các phương pháp học sâu và xử lý dữ liệu lớn. Tuy nhiên các thách thức về độ chính xác, khả năng xử lý dữ liệu vẫn còn khá nhiều hạn chế.

Trong nước, lĩnh vực phân tích cảm xúc đang thu hút sự chú ý của các nhà nghiên cứu, đặc biệt trong bối cảnh thương mại điện tử đang phát triển mạnh mẽ. Các công trình nghiên cứu sử dụng các mô hình học máy để phân loại dữ liệu để ứng dụng **NLP vào Tiếng Việt [5][6]**. Đáng chú ý với nghiên cứu của Trần Quang Phúc và nhóm của mình đã cho thấy việc tập trung trích xuất các từ ngữ (Opinion words) trong các đánh giá giúp nâng cao hiệu quả phân tích cảm xúc [7]. Nghiên cứu sử dụng các thuật toán học máy theo **phương pháp tập hợp (ensemble methods)** như **Gradient**

Boosting Classifier (GBC) đem lại độ chính xác cao trong giai đoạn kiểm tra, giúp cải thiện khả năng đánh giá cảm xúc tích cực và tiêu cực. Lĩnh vực phân tích cảm xúc đang có những bước tiến đáng kể nhưng vẫn còn rất nhiều tiềm năng để phát triển, mở ra các cơ hội khai thác dữ liệu tốt hơn.

2.2. Hướng tiếp cận của đề tài

Bài nghiên cứu của nhóm tập trung vào nghiên cứu khả năng hiệu suất và độ chính xác của các mô hình học máy trong việc phân tích cảm xúc đánh giá đến từ người dùng. Nghiên cứu sử dụng các phương pháp trích xuất đặc trưng **Bag of Words (BoW)** và **TF-IDF** để phân tích nội dung bình luận cảm xúc của người dùng. Với các module học máy như **Logistic Regression**, **Naive Bayes**, **Linear Regression** và **SVM** được dùng để xử lý các thông tin từ đặc trưng giúp phân loại cảm xúc đánh giá sản phẩm của người dùng dựa trên thang điểm.

Thông qua tập dữ liệu Amazon Fine Food Reviews - một tập đánh giá uy tín, chứa hàng ngàn đánh giá thực phẩm của người dùng [8]. Bài nghiên cứu của nhóm thực hiện việc tiền xử lý dữ liệu loại bỏ **StopWord**, **Lemmatization**, sau đó thực hi

ện so sánh hiệu suất giữa các mô hình với kết quả kỳ vọng sẽ đánh giá được mô hình học máy nào có độ chính xác cao trong quá trình phân tích cảm xúc đánh giá của người dùng dựa trên các bình luận.

3. Mục tiêu và nhiệm vụ nghiên cứu

Mục tiêu nghiên cứu: Bài nghiên cứu của nhóm hướng đến triển khai và đánh giá các mô hình học máy nhằm so sánh hiệu suất và độ chính xác giữa các phương pháp trích xuất đặc trưng khác nhau. Nghiên cứu tập trung phân tích và so sánh hiệu suất các mô hình để xác định mô hình nào có hiệu suất tốt trong phương pháp trích xuất đặc trưng cụ thể.

Nhiệm vụ nghiên cứu:

- **Xây dựng và huấn luyện mô hình:** Áp dụng các mô hình học máy Logistic Regression, Naive Bayes, Linear Regression và SVM để thực hiện phân tích cảm xúc đánh giá sản phẩm.

- **Thiết lập kịch bản thực nghiệm:** Thiết kế kịch bản thực nghiệm và tiến hành thử nghiệm trên bộ dữ liệu các đánh giá sản phẩm của người dùng Amazon Fine Food Reviews.
- **Đánh giá hiệu suất mô hình:** Sử dụng các chỉ số đánh giá quan trọng như Accuracy, Precision, Recall, F1-Score để đo lường hiệu suất, độ chính xác của các mô hình.
- **So sánh và kết luận:** So sánh hiệu suất giữa các mô hình, đưa ra kết luận mô hình có độ chính xác, hiệu suất cao trong phương pháp trích xuất đặc trưng.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Bài nghiên cứu của nhóm tập trung vào các phương pháp trích xuất đặc trưng Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF) để thực hiện lấy các đặc trưng trong nội dung bình luận của người dùng. Các mô hình học máy Linear Regression, Logistic Regression, NB, SVM sẽ được sử dụng để phân tích các đánh giá xem đó là đánh giá tích cực, tiêu cực hay trung lập với các tham số độ đo để đánh giá khả năng phân tích của các mô hình được sử dụng. Bài nghiên cứu triển khai sử dụng trong tập dữ liệu Amazon Fine Food Reviews, một tập dữ liệu phổ biến trong bài toán phân tích cảm xúc trong đánh giá sản phẩm của người dùng.

Phạm vi nghiên cứu: Nghiên cứu của nhóm tập trung vào việc ứng dụng các mô hình học máy bao gồm Linear Regression, Logistic Regression, NB, SVM để phân loại cảm xúc của dữ liệu của bình luận người dùng là tích cực, tiêu cực hoặc trung lập. Tập trung vào triển khai các mô hình học máy cho bài toán phân loại cảm xúc đánh giá của người dùng thông qua các phương pháp trích xuất đặc trưng BoW và TF-IDF, trên tập dữ liệu Amazon Fine Food Reviews với nhãn cảm xúc dựa trên điểm số đánh giá từ 1 đến 5.

- **Mô hình học máy:** Ứng dụng các mô hình học máy như Linear Regression, Logistic Regression, NB, SVM để thực hiện việc phân tích dữ liệu từ văn bản.
- **Phương pháp tiền xử lý:** Thực hiện tiền xử lý tập dữ liệu được chọn với việc bỏ qua các kí tự đặc biệt, các từ dừng (StopWord), đưa từ về dạng nguyên bản (Lemmatization) để tăng khả năng trích xuất đặc trưng.

- **Phương pháp trích xuất đặc trưng:** Thực hiện dùng phương pháp trích xuất đặc trưng các đặc điểm từ dữ liệu văn bản với Bag of Words (BoW) và TF-IDF.
- **Tập dữ liệu:** Sử dụng tập dữ liệu Amazon Fine Food Reviews, bao gồm các nội dung bình luận và điểm đánh giá sản phẩm đến từ khách hàng. Tập dữ liệu sẽ được tiền xử lý để phù hợp với mục tiêu nghiên cứu, đảm bảo chất lượng đầu vào.
- **Đánh giá mô hình:** Các chỉ số đánh giá hiệu suất như Accuracy, Percision, Recall, F1-Score được sử dụng để đo lường và so sánh khả năng hiệu quả, chính xác của từng mô hình học máy được sử dụng trên tập dữ liệu.

Phạm vi nghiên cứu được giới hạn ở bài toán phân tích cảm xúc từ dữ liệu nội dung đánh giá và điểm số đánh giá của người dùng, với mục tiêu phân tích mô hình máy học nào có hiệu năng, độ chính xác cao trong việc phân tích cảm xúc văn bản.

5. Phương pháp nghiên cứu

Bài nghiên cứu của nhóm là sự kết hợp chặt chẽ giữa các phương pháp lý thuyết, phương pháp thực nghiệm và phương pháp chuyên gia nhằm đảm bảo tính khoa học, tính nhất quán và hệ thống trong quá trình nghiên cứu thực hiện đề tài.

5.1. Phương pháp lý thuyết

Tìm hiểu các bài báo có liên quan đến chủ đề phân tích cảm xúc bằng mô hình máy học thông qua các bài nghiên cứu, bài báo trên các nguồn uy tín, chính xác.

Tổng hợp và phân tích các cơ sở lý thuyết về kĩ thuật xử lí ngôn ngữ tự nhiên (NLP), trích xuất đặc trưng bao gồm BoW, TF-IDF và các khái niêm các mô hình máy học.

Lựa chọn các nguồn báo uy tín từ các tạp chí khoa học hàng đầu như IEEE, Springer và các hội nghị trí tuệ nhân tạo (NeurIPS, ACL, EMNLO) để đảm bảo sự uy tín cho bài nghiên cứu.

5.2. Phương pháp thực nghiệm

Phân tích, thiết kế triển khai kịch bản thực nghiệm, các mô hình học máy trên tập dữ liệu Amazon Fine Food Reviews.

Cài đặt các mô hình và thuật toán bằng python và sử dụng các thư viện học máy như NLTK, Scikit-learn, Pandas, Seaborn.

Thực hiện tiền xử lý dữ liệu văn bản sử dụng tập dữ liệu Amazon Fine Food Reviews để cho đề tài phân tích cảm xúc từ đánh giá của người dùng, đảm bảo dữ liệu đáp ứng yêu cầu đầu vào cảu mô hình.

Thực hiện huấn luyện và kiểm thử các mô hình, ghi nhận và đánh giá hiệu suất các mô hình bằng các chỉ số độ đo như Accuracy, Precision, Recall và F1-Score.

Thực hiện so sánh và phân tích các mô hình với các phương pháp trích xuất đặc trưng, từ đó rút ra kết luận hiệu quả của mô hình trong việc phân tích cảm xúc đánh giá sản phẩm của người dùng.

5.3. Phương pháp chuyên gia

Trao đổi với giảng viên hướng dẫn để làm rõ các khuất mắt trong qui trình thực hiện đồng thời tiếp nhận các đánh giá và phản hồi nhằm cải thiện bài nghiên cứu và đưa ra mục tiêu , nhiệm vụ phù hợp cho từng giai đoạn

Dựa trên ý kiến của giảng viên để điều chỉnh cách thực hiện đề tài, triển khai thực nghiệm và đảm bảo tiến độ thực hiện bài nghiên cứu.

6. Giả thuyết khoa học

Các mô hình học máy được sử dụng để phân loại cảm xúc (tích cực, tiêu cực, trung tính) từ các đánh giá sản phẩm với độ chính xác cao hơn so với phương pháp phân tích thủ công.

Việc áp dụng các kỹ thuật tiền xử lý văn bản như loại bỏ StopWord, Lemmatization sẽ giúp cải thiện độ chính xác của các mô hình học máy trong việc huấn luyện để phân loại cảm xúc từ đánh giá sản phẩm

Các phương pháp trích xuất đặc trưng khác nhau như BoW, TF-IDF sẽ mang lại những kết quả với sự thay đổi đáng kể trong việc cho ra kết quả sau thực hiện các mô hình.

7. Dự kiến kế hoạch nghiên cứu

STT	Nội dung	Dự kiến thời gian thực hiện

1	Nghiên cứu tổng quan và xác định đề tài	1 tuần
2	Đọc các bài báo liên quan	2 tuần
3	Nghiên cứu tài liệu các thuật toán học máy	1 tuần
4	Chuẩn bị và khảo sát dữ liệu	1 tuần
5	Tiền xử lý dữ liệu	1 tuần
6	Trích xuất đặc trưng	2 tuần
7	Chia tập dữ liệu và huấn luyện mô hình	2 tuần
8	Đánh giá, so sánh và phân tích kết quả	1 tuần
9	Viết và hoàn thiện báo cáo	1 tuần

8. Dự kiến nội dung tiêu luận nghiên cứu

Tóm tắt

- Lý do nghiên cứu.
- Mục tiêu nghiên cứu.
- Phương pháp nguyên cứu.
- Kết luận chính.

Chương 1: Tổng quan vấn đề

- Lý do chọn đề tài.
- Vấn đề nghiên cứu.
- Mục tiêu nghiên cứu.
- Câu hỏi nghiên cứu.
- Phạm vi nghiên cứu.

Chương 2: Lược khảo tài liệu

- Tổng hợp các tài liệu, nghiên cứu trước liên quan.
- Xây dựng cơ sở lý thuyết cho nghiên cứu.
    - Giới thiệu khái niệm về bài toán phân tích cảm xúc.
    - Các phương pháp phân tích cảm xúc, phương pháp tiền xử lý.
    - Tập dữ liệu sử dụng Amazon Fine Food Reviews.
    - Các chỉ số đánh giá hiệu suất: Accuracy, Precision, Recall, F1-Score.
- Điểm mạnh, điểm yếu của các nghiên cứu trước và hướng nghiên cứu của nhóm.

Chương 3: Phương pháp nghiên cứu

- Quy trình nghiên cứu và các bước thực hiện
- Đối tượng và mẫu nghiên cứu.
- Thiết kế kịch bản thực nghiệm:
    - Chuẩn bị dữ liệu: Tiền xử lý, chia tập huấn luyện và kiểm thử.
    - Cài đặt mô hình: Sử dụng Python với các thư viện học máy.

Chương 4: Thực nghiệm và thảo luận

- Thực hiện chạy chương trình trên tập dữ liệu Amazon Fine Food Reviews
    - Huấn luyện và kiểm thử các mô hình học máy.
    - Đánh giá hiệu suất các mô hình thông qua các chỉ số độ đo Accuracy, Precision, Recall, F1-Score.
- Sử dụng biểu đồ, bảng biểu, hình ảnh minh họa cho việc xử lý tập dữ liệu.
- Đánh giá và giải thích kết quả nghiên cứu.
- So sánh với các nghiên cứu trước.

Chương 5: Kết luận và hướng phát triển

- Tóm tắt những điểm chính đã đạt được.
- Đánh giá độ chính xác, hiệu quả của các mô hình học máy có độ chính xác cao với phương pháp sử dụng.
- Chỉ ra những hạn chế và đề xuất nghiên cứu tiếp theo.

9. Danh mục tài liệu tham khảo

[1] Shan, S., Sun, J., & Macawile, R. M. C. (2025). Examining Customer Satisfaction through Transformer-Based Sentiment Analysis for Improving Bilingual E-Commerce Experiences. IEEE Access.

[2] Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature.

[3] Zhao, X., & Sun, Y. (2022). Amazon fine food reviews with BERT model. Procedia Computer Science, 208, 401-406.

[4] Aravindan, T., Vigneshwar, C., & Ga, S. U. G. A. N. E. S. H. W. A. R. I. (2023, January). Sentiment Classification for Amazon Fine Foods Reviews Using Pyspark. In Recent Developments in Electronics and Communication Systems: Proceedings of the First International Conference on Recent Developments in Electronics and Communication Systems (RDECS-2022) (Vol. 32, p. 431). IOS Press.

[5] Bằng, N. Đ. L., & Thành, H. T. (2021). Mô hình khai phá ý kiến và phân tích cảm xúc khách hàng trực tuyến trong ngành thực phẩm. TẠP CHÍ KHOA HỌC ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH-KINH TẾ VÀ QUẢN TRỊ KINH DOANH, 16(1), 64-78.

[6] Trần, K. T., & Tiế, P. M. S. (2020). Một số khái niệm và hướng tiếp cận phân tích cảm xúc-Áp dụng cho tiếng Việt. Tạp chí Khoa học HUFLIT, 6(1), 82-82.

[7] Tran, P. Q., Trieu, N. T., Dao, N. V., Nguyen, H. T., & Huynh, H. X. (2020). Effective opinion words extraction for food reviews classification. International Journal of Advanced Computer Science and Applications, 11(7).

[8] Huang, H., Zavareh, A. A., & Mustafa, M. B. (2023). Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. IEEE Access, 11, 90367-90382.