**ĐỀ CƯƠNG CHI TIẾT**

Ngày thực hiện đề cương:

| **TÊN ĐỀ TÀI** | Phân tích đánh gíá sản phẩm |
| --- | --- |
| **LĨNH VỰC CHUYÊN NGÀNH** | Khoa học dữ liệu và xử lý ngôn ngữ tự nhiên |
| **LOẠI HÌNH NGHIÊN CỨU** | Nghiên cứu sáng tạo |
| **NGƯỜI HƯỚNG DẪN** | Đỗ Như Tài |
| **SINH VIÊN THỰC HIỆN 1** | 3124720066 - Nguyễn Gia Thành |
| **SINH VIÊN THỰC HIỆN 2** | 3123410151 - Nguyễn Dương Khang |
| **SINH VIÊN THỰC HIỆN 3** | 3121560024 - Nguyễn Quang Dương |

**GIỚI THIỆU**

| **CÂU HỎI GỢI Ý** | Lĩnh vực và nghiên cứu liên quan đã và đang phát triển như thế nào ?  Các vấn đề, và bài toán đặt ra cần giải quyết là gì ?̀ |
| --- | --- |
| **HƯỚNG DẪN** | Giới thiệu tổng quan về đề tài – những vấn đề và lĩnh vực liên quan đến đề tài. |
| Trong đời sống ngày càng nâng cao, công nghệ phát triển, hệ thống truyền tải dữ liệu ngày càng tân tiến, việc con người dựa dẫm vào công nghệ cũng theo đó tăng lên, các hoạt động kinh doanh dần trở nên công nghệ hóa, các sàn thương mại điện bắt đầu mọc lên như nấm, cụ thể phải kể đến : amazon, alibaba, shopee, tiki, lazada,... chúng phát triển mạnh mẽ với phương thức hoạt động như một phiên chợ với nhiều sản phẩm thú vị, điều đặc biệt là người dùng không cần phải trực tiếp di chuyển ra các cửa hàng, khu chợ để tìm kiếm và mua những thứ thiết yếu, đáp ứng nhu cầu của bản thân. Mọi người chỉ cần truy cập lên các trang thương mại điện tử hoặc các ứng dụng của chúng và tìm kiếm các vật dụng, sản phẩm cần mua, chúng sẽ gợi ý hàng chục sản phẩm với nhiều mẫu mã mà bạn vừa tìm kiếm trên nhiều gian hàng khác nhau. Cũng chính vì quá nhiều gian hàng như thế, đã gây ra khó khăn trong việc nhận biết chất lượng sản phẩm khi thanh toán, khó chọn lọc ra các gian hàng uy tín, giúp hài lòng với các tiêu chuẩn của khách hàng. Biết được điều đó nhóm mình có thể đưa ra cho mọi người 1 giải pháp có thể hạn chế việc mua phải hàng đểu, kém chất lượng. Đó chính là chương trình phân tích đánh giá sản phẩm khách hàng, chương trình này giúp mọi người tìm ra những nhận xét hài lòng, không hài lòng, tạm chấp nhận từ lời bình phẩm mà các người tiêu dùng để lại và sau đó suy xét đến việc có nên mua sản phẩm hay không. Chương trình trên của nhóm chúng mình giúp mọi người có thể giảm thiểu rủi ro mua phải hàng lậu, hàng nhái, bên cạnh đó rút ngắn thời gian suy nghĩ, và đưa quyết định.  Chương trình phát triển xây dựng dựa trên các dữ liệu có sẵn từ các nhận xét sản phẩm trên các trang thương mại điện tử ( Tiki, Shopee, Lazada, Sendo, Amazon ) triển khai bằng ngôn ngữ python với sự hỗ trợ của các mô hình Xử lý ngôn ngữ tự nhiên, Machine learning. | |
|  | |
| **Ghi chú** | |
|  | |

**TÌNH HÌNH NGHIÊN CỨU TRONG VÀ NGOÀI NƯỚC.**

| **Câu hỏi gợi ý** | Lĩnh vực và nghiên cứu liên quan đã và đang phát triển như thế nào?  Các vấn đề, và bài toán đặt ra cần giải quyết là gì? |
| --- | --- |
| **Hướng dẫn** | Tìm hiểu các nghiên cứu đã công bố gần nhất (5 năm trở lại) về lĩnh vực liên quan trong và ngoài nước. Phân tích các kết quả đóng góp, nhận xét các hạn chế còn tồn tại. |
| Phân tích đánh giá sản phẩm trên các nền tảng thương mại điện tử là một lĩnh vực nghiên cứu quan trọng, thu hút sự quan tâm của nhiều nhà khoa học trong và ngoài nước. Dưới đây là tổng quan về tình hình nghiên cứu liên quan:  Một số bài báo nghiên cứu trong nước   1. Phân tích cảm xúc trong đánh giá sản phẩm tiếng Việt: Một nghiên cứu đã đề xuất mô hình đánh giá cảm xúc của khách hàng bằng cách trích xuất đối tượng và từ ngữ cảm xúc, sử dụng các thuật toán học máy. Nghiên cứu này tập trung vào ngôn ngữ tiếng Việt với cấu trúc ngữ nghĩa đặc thù. [[1]](https://www.zotero.org/google-docs/?Pry4xV) 2. Phân tích chủ đề và cảm xúc trong đánh giá khách hàng: Nghiên cứu này sử dụng mô hình LDA kết hợp với phân tích cảm xúc để hiểu rõ hơn về quan điểm và ý kiến của khách hàng trên các nền tảng thương mại điện tử tại Việt Nam. 3. Phân tích so sánh trong đánh giá sản phẩm tiếng Việt: Một nghiên cứu khác đã tập trung vào khai thác và phân tích các đánh giá so sánh sản phẩm bằng tiếng Việt, nhằm trích xuất thông tin hữu ích cho người tiêu dùng và doanh nghiệp.   Nghiên cứu quốc tế   1. Cải thiện phân tích đánh giá khách hàng trên thương mại điện tử: Nghiên cứu này đã kiểm tra các thuật toán học máy khác nhau để dự đoán cảm xúc của đánh giá trên các trang web thương mại điện tử, nhằm nâng cao độ chính xác trong việc phân loại cảm xúc. ​[Nature](https://www.nature.com/articles/s41598-022-26432-3?utm_source=chatgpt.com) 2. Phân tích cảm xúc trên đánh giá sản phẩm thương mại điện tử sử dụng học sâu: Nghiên cứu cung cấp cái nhìn tổng quan về việc sử dụng các thuật toán học máy và học sâu để phân tích cảm xúc trong đánh giá sản phẩm trên các nền tảng thương mại điện tử. ​ 3. Tác động của đánh giá trực tuyến đến quyết định mua hàng của người tiêu dùng: Nghiên cứu này sử dụng công nghệ theo dõi mắt để điều tra tác động của đánh giá sản phẩm trực tuyến đến quyết định mua hàng của người tiêu dùng, cung cấp cái nhìn sâu sắc về cách người tiêu dùng tương tác với đánh giá trực tuyến. ​[Frontiers](https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.865702/full?utm_source=chatgpt.com)  Xu hướng và hướng nghiên cứu tương lai :  * Tích hợp AI: sử dụng AI trong việc phân tích để đưa ra những thông số trực quan, chính xác nhất một cách tự động ( AI tự thu thập dữ liệu, tự kiểm tra, xử lý và phân tích dữ liệu ) và việc của chúng ta là kiểm tra huấn luyện AI nhiều lần bằng các mô hình như ( Transformer (BERT, RoBERTa, T5, GPT, v.v. ) để có thể hạn chế rủi ro ở mức thấp nhất, mà mang lại thông tin chính xác nhất. * Tích hợp đa phương thức: Kết hợp phân tích văn bản, hình ảnh và video trong đánh giá sản phẩm để cung cấp cái nhìn toàn diện hơn về phản hồi của khách hàng.​ * Phân tích cảm xúc nâng cao: Sử dụng các mô hình học sâu tiên tiến để cải thiện độ chính xác trong việc phân loại cảm xúc và phát hiện cảm xúc tiềm ẩn trong đánh giá.​ * Phát hiện đánh giá giả mạo: Phát triển các hệ thống tự động để nhận diện và loại bỏ các đánh giá không chân thực, đảm bảo tính minh bạch và tin cậy của thông tin.​ * Phân tích ngôn ngữ tự nhiên đa ngôn ngữ: Mở rộng nghiên cứu sang nhiều ngôn ngữ khác nhau để phục vụ thị trường toàn cầu và hiểu rõ hơn về sự khác biệt văn hóa trong đánh giá sản phẩm.​   Nhìn chung, lĩnh vực phân tích đánh giá sản phẩm trên các nền tảng thương mại điện tử đang phát triển mạnh mẽ, với nhiều nghiên cứu tập trung vào việc cải thiện độ chính xác và hiệu quả của các phương pháp phân tích, cũng như ứng dụng thực tiễn trong kinh doanh và marketing. | |
|  | |
| **Ghi chú** | |
|  | |

**MỤC TIÊU VÀ PHẠM VI NGHIÊN CỨU.**

| **CÂU HỎI NGHIÊN CỨU** | **Mục tiêu nghiên cứu chính của đề tài là gì?**  **Phạm vi nghiên cứu là gì?** |
| --- | --- |
| **HƯỚNG DẪN** | Đặt bài toán giải quyết và trình bày mục tiêu nghiên cứu chính của đề tài.  Nêu phạm vi nghiên cứu của đề tài, bao gồm việc giới hạn phạm vi nghiên cứu và triển khai, các giả định ban đầu đối với nghiên cứu. |
| Với vấn đề nêu ra ở trên mục tiêu nghiên cứu của đề tài chúng tôi là phát triển một hệ thống phân tích đánh giá sản phẩm hướng đến các chức năng chính như sau:   * Phân tích cảm xúc và nhận diện chủ đề * Tóm tắt nội dung và nhấn mạnh ý chính. * Tổng hợp các nhận xét có chủ đề tương đồng và hướng đến cùng một mục tiêu.   Mục tiêu nghiên cứu chính : xây dựng chương trình thu thập các bình đưa, nhận xét và đưa ra mức độ hài lòng của người dùng đối với sản phẩm từ các nguồn dữ liệu đã thu thập.  Hướng phát triển và hoàn thiện tối ưu sản phẩm : Ứng dụng được mô hình AI vào trong việc thu thập dữ liệu và phân tích cảm xúc (giúp mọi thứ trở nên tự động hóa mà không cần quá nhiều tác động tự con người)  Về phạm vi nghiên cứu: Nguồn dữ liệu để xử lý được lấy từ kaggle có thể cũ và chưa được cập nhật, ngoài ra còn bị hạn chế về giới hạn dữ liệu. Nếu lấy trực tiếp từ các trang thì có nguy cơ bị chặn IP hoặc phải xác nhận captcha, nhưng vẫn có thể khắc phục bằng cách sử dụng proxy hay IP xoay vòng để hạn chế rủi ro, hay đơn giản hơn lại cài đặt thời gian tiếp nhận thu thập dữ liệu (khoảng cách thời gian lấy dữ liệu nới rộng hơn). | |
|  | |
| **Ghi chú** | |
|  | |

**CÁC GIẢ THUYẾT VÀ CÁCH TIẾP CẬN**

| **CÂU HỎI NGHIÊN CỨU** | **Các giả thuyết đặt ra để giải quyết bài toán chính?**  **Các cách tiếp cận để giải quyết bài toán đặt ra?** |
| --- | --- |
| **HƯỚNG DẪN** | Đặt ra những giả thuyết, hay vấn đề - bài toán con cần phải giải quyết để đạt được mục tiêu nghiên cứu đề tài.  Các cách tiếp cận (dự kiến) để giải quyết các giả thuyết, bài toán con đã đặt. |
| **Giả thuyết:**  Có thể xây dựng một hệ thống phân tích đánh giá sản phẩm đáp ứng các yêu cầu sau:  o Thu thập và tích hợp dữ liệu đa nguồn: thu thập đánh giá sản phẩm từ các nguồn trực tuyến như website thương mại điện tử, diễn đàn, mạng xã hội, qua đó tạo thành một kho dữ liệu thống nhất phục vụ cho quá trình phân tích. o Xử lý và tổ chức dữ liệu (logfile): dữ liệu thu thập được sẽ được tiền xử lý (làm sạch, chuẩn hóa văn bản, loại bỏ stopwords, …) và tổ chức lại thành cấu trúc logfile phù hợp, đảm bảo chứa đủ thông tin cần thiết để khai thác các đặc trưng của đánh giá. o Phân tích cảm xúc và nhận diện chủ đề: dựa trên dữ liệu đã xử lý, hệ thống áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) và học máy để tự động phân loại cảm xúc (không hài lòng, bình thường, hài lòng) và nhận diện các chủ đề chính mà người dùng quan tâm.  **Từ mục tiêu nghiên cứu, chúng tôi đặt ra hai bài toán chính của đề tài:**  Bài toán thứ nhất: Xây dựng hệ thống thu thập và xử lý dữ liệu đánh giá sản phẩm.  Bài toán thứ hai: Xây dựng mô hình phân tích cảm xúc và nhận diện chủ đề tự động.  Với bài toán thứ nhất, chúng tôi sẽ tổ chức lại cấu trúc logfile từ dữ liệu đánh giá thu thập được, thực hiện tiền xử lý dữ liệu (làm sạch, chuẩn hóa, trích xuất thông tin cần thiết). Đồng thới cố gắng phát triển chức năng tách dữ liệu đánh giá thành các nhóm có thể phân tích riêng biệt, ví dụ: đánh giá theo sản phẩm, theo danh mục, hoặc theo thời gian.  Với bài toán thứ hai, chúng tôi sẽ phát triển bộ thuật toán (sử dụng các thư viện Python như scikit-learn, TensorFlow, hoặc NLTK, spaCy) để phân tích cảm xúc của từng đánh giá và trích xuất các chủ đề chính từ nội dung đánh giá. So sánh kết quả phân tích giữa các nhóm đánh giá khác nhau, từ đó đưa ra các thông tin hỗ trợ cho doanh nghiệp.  Cách tiếp cận:   1. Thiết kế và triển khai hệ thống thu thập dữ liệu để thu thập dữ liệu đánh giá từ nhiều nguồn trực tuyến. 2. Tiền xử lý và tổ chức dữ liệu thành logfile, phân chia rõ ràng các loại thông tin như đánh giá cá nhân, thông tin sản phẩm và các thuộc tính liên quan khác. 3. Xây dựng cấu trúc “profile sản phẩm” dựa trên dữ liệu thu thập được, kết hợp thông tin tĩnh (đặc điểm sản phẩm, danh mục, giá cả) và thông tin động (đánh giá, phản hồi của người tiêu dùng). 4. Phát triển bộ thuật toán phân tích, xây dựng và tinh chỉnh các mô hình phân tích cảm xúc (sử dụng thuật toán như SVM, Naive Bayes, hoặc các mô hình deep learning) để phân loại cảm xúc của từng đánh giá. 5. Xây dựng chức năng thống kê và trực quan hóa kết quả, phát triển các báo cáo thống kê tổng hợp kết quả phân tích: tỷ lệ cảm xúc, chủ đề nổi bật theo thời gian, phân tích so sánh giữa các sản phẩm hoặc danh mục. | |
|  | |
| **Ghi chú** | |
|  | |

**TÀI LIỆU THAM KHẢO**

| **CÂU HỎI NGHIÊN CỨU** | Mục tiêu nghiên cứu chính của đề tài là gì?  Phạm vi nghiên cứu là gì? |
| --- | --- |
| **HƯỚNG DẪN** |  |
| [**[1] N.-B.-V. Le and J.-H. Huh, “Applying Sentiment Product Reviews and Visualization for BI Systems in Vietnamese E-Commerce Website: Focusing on Vietnamese Context,” *Electronics*, vol. 10, no. 20, Art. no. 20, Jan. 2021, doi: 10.3390/electronics10202481.**](https://www.zotero.org/google-docs/?broken=TWwuzj) | |
|  | |
| **Ghi chú** | |
|  | |

| **Người thực hiện đề cương** | | |
| --- | --- | --- |
| **Người thực hiện 1** | Nguyễn Gia Thành | 3124720066 |
| **Người thực hiện 2** | Nguyễn Dương Khang | 3123410151 |
| **Người thực hiện 3** | Nguyễn Quang Dương | 3121560024 |

| **XÁC NHẬN CỦA GIÁO VIÊN HƯỚNG DẪN** | | |
| --- | --- | --- |
| **NGƯỜI HƯỚNG DẪN** | **Đỗ Như Tài** | |
| **Ý KIẾN** |  | |
|  | □ Đồng ý hướng dẫn | □ Không đồng ý hướng dẫn |