BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

TẠ THỊ THIÊN THANH

PHÂN TÍCH XU HƯỚNG MUA HÀNG CỦA KHÁCH HÀNG TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ

ĐỒ ÁN NGÀNH

NGÀNH KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH, 2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

TẠ THỊ THIÊN THANH

PHÂN TÍCH XU HƯỚNG MUA HÀNG CỦA KHÁCH HÀNG TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ

Mã số sinh viên: 2151013088

ĐỒ ÁN NGÀNH

NGÀNH KHOA HỌC MÁY TÍNH

Giảng viên hướng dẫn: ThS. PHẠM CHÍ CÔNG

TP. HỒ CHÍ MINH, 2024

LỜI CẢM ƠN

Trước hết, em xin gửi lời cám ơn chân thành và sâu sắc nhất đến thầy ThS. Phạm Chí Công đã luôn đồng hành và hỗ trợ tận tình, tâm huyết trong suốt quá trình em thực hiện báo cáo. Cám ơn thầy vì những phản hồi mang tính xây dựng và cả những chia sẻ chân thành về hướng phát triển trong từng vấn đề mà em thắc mắc. Những gợi ý và lời chỉ bảo của thầy không những giúp em học hỏi thêm được những cái mới mà còn giúp em nâng cao tư duy nghiên cứu, thuận lợi hoàn thành tốt đồ án. Em tin rằng những kiến thức và kinh nghiệm học được từ thầy sẽ là nền tảng vững chắc cho sự phát triển của em trong tương lai khi em tiếp xúc với môi trường làm việc thực tế sau này.

Đồng thời, em cũng xin gửi lời cảm ơn chân thành đến Khoa Công nghệ thông tin vì đã luôn tạo điều kiện tốt nhất cho em và các bạn sinh viên khác trong quá trình học tập và nghiên cứu tại trường.

Một lần nữa, em xin bày tỏ lòng biết ơn sâu sắc đến thầy ThS. Phạm Chí Công vì sự tận tụy và sự đóng góp không ngừng nghỉ trong suốt thời gian qua và Khoa Công nghệ thông tin đã nỗ lực hỗ trợ tạo môi trường tốt nhất giúp em hoàn thành đề tài.

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

TÓM TẮT ĐỒ ÁN NGÀNH

Cùng với sự phát triển của thương mại điện tử, nhu cầu mua sắm các mặt hàng thời trang trực tuyến ngày càng cao, đặc biệt là thời trang nữ. Lượng thông tin cũng dần trở nên nhiều gây ra hiện tượng nhiễu loạn thông tin gây bối rối trong quá trình mua sắm. Mục tiêu của đề tài là giải quyết các vấn đề khó khăn trong đưa ra quyết định của khách hàng trong quá trình sử dụng các sàn thương mại điện tử. Đề tài gồm hai phần chính là thu thập, phân tích dữ liệu, đánh giá tình hình cũng như xu hướng mua hàng hiện nay và từ đó thiết kế hệ thống gợi ý sản phẩm phù hợp.

ABSTRACT

With the rapid growth of e-commerce, the demand for online fashion shopping, particularly in women's fashion, has significantly increased. However, the huge amount of information available has led to information overload, making it confusing for customers to make purchasing decisions. The aim of this project is to solve the challenge that customers are facing in making decisions on e-commerce platforms. The project consists of two main parts: collecting and analyzing data to evaluate current shopping trends and designing a recommendation system to assist customers in selecting suitable products.

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

TỪ VIẾT TẮT	CỤM TỪ ĐẦY ĐỦ	Ý NGHĨA
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên.
CSV	Comma-separated Values	Tên một loại file có định dạng cụ thể cho phép lưu dữ liệu ở dạng bảng có cấu trúc.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	Tên một loại thuật toán gom cụm.
DFD	Data Flow Diagram	Sơ đồ luồng dữ liệu.
ERD	Entity Relationship Diagram	Biểu đồ thực thể - quan hệ, một loại biểu đồ cấu trúc sử dụng trong thiết kế cơ sở dữ liệu.
RDM	Relational Data Model	Mô hình dữ liệu quan hệ.

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

MỞ ĐẦU

Thời trang nữ luôn là lĩnh vực được quan tâm hàng đầu, đặc biệt trong bối cảnh thương mại điện tử phát triển mạnh mẽ. Với sự gia tăng không ngừng của các sản phẩm đa dạng mẫu mã và giá cả trên các nền tảng trực tuyến, người tiêu dùng gặp phải với nhiều khó khăn trong việc tìm kiếm sản phẩm phù hợp. Hơn nữa, sự thay đổi liên tục của các xu hướng thời trang tạo ra nhiều sự lựa chọn nhưng cũng đồng thời gây ra khó khăn trong lựa chọn. Vì vậy, một hệ thống đề xuất sản phẩm là giải pháp thiết yếu, không chỉ giúp người tiêu dùng có trải nghiệm mua sắm tốt hơn mà còn hỗ trợ các nhà bán lẻ tăng trưởng doanh thu và hiệu quả kinh doanh.

Thương mại điện tử đã và đang phát triển, thay đổi cách người tiêu dùng tiếp cận và mua sắm. Khả năng tiếp cận đa dạng sản phẩm từ nhiều nơi cùng với việc cung cấp đầy đủ thông tin về sản phẩm đã thu hút một lượng lớn người mua hàng trực tuyến. Tuy nhiên, sự đa dạng này cũng dẫn đến hiện tượng nhiễu và quá tải thông tin, khiến người tiêu dùng cảm thấy bối rối và mất thời gian trong quá trình lựa chọn sản phẩm. Điều này tạo ra nhu cầu cho các giải pháp hỗ trợ quyết định mua hàng hiệu quả.

Hệ thống đề xuất sản phẩm đóng vai trò quan trọng trong việc giải quyết vấn đề này. Bằng cách sử dụng các thuật toán phân tích dữ liệu và học máy, hệ thống có thể hiểu rõ hơn về xu hướng, thói quen mua sắm và các yếu tố khác của từng người tiêu dùng. Từ đó, hệ thống có khả năng đề xuất những sản phẩm phù hợp nhất, giúp tối ưu hóa trải nghiệm mua sắm của khách hàng.

Ngoài việc nâng cao trải nghiệm người dùng, hệ thống đề xuất còn mang lại nhiều lợi ích cho các nhà bán lẻ. Việc đề xuất sản phẩm phù hợp không chỉ giúp tăng doanh số bán hàng mà còn gia tăng lòng trung thành của khách hàng. Khi khách hàng cảm thấy đáp ứng đúng nhu cầu, họ sẽ có xu hướng quay lại và tiếp tục mua sắm trên nền tảng đó. Hơn nữa, hệ thống đề xuất hiệu quả giúp các nhà bán lẻ tối ưu hóa các công đoạn quản lý kho hàng, giảm thiểu tình trạng tồn kho và tăng cường hiệu quả kinh doanh.

Chương 1. TỔNG QUAN VỀ ĐỀ TÀI

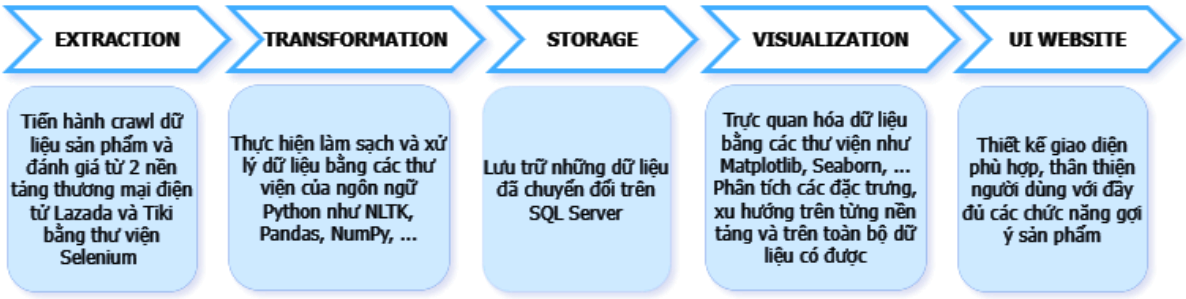
1.1. Mục tiêu của đề tài

Thời trang, đặc biệt là thời trang nữ, luôn là một lĩnh vực thu hút sự quan tâm lớn của người tiêu dùng. Nhu cầu mua sắm các sản phẩm thời trang trên các nền tảng thương mại điện tử ngày càng nhiều, việc tìm kiếm sản phẩm phù hợp cũng trở nên khó khăn bởi các mẫu mã với đa dạng mức giá đang phát triển không ngừng. Hơn nữa, sự biến đổi không ngừng các xu hướng thời trang mang lại cho khách hàng nhiều sự lựa chọn hơn, nhưng cũng tạo ra những khó khăn trong việc tìm kiếm sản phẩm thỏa mãn nhu cầu cá nhân. Trong bối cảnh đó, một hệ thống đề xuất sản phẩm mang tính cá nhân hóa sẽ không chỉ giúp nâng cao trải nghiệm mua sắm của khách hàng mà còn giúp các cửa hàng bán lẻ trên các nền tảng thương mại điện tử tăng doanh thu và hiệu quả kinh doanh.

1.2. Phạm vi đề tài

Đề tài tập trung vào việc phát triển một hệ thống hỗ trợ người dùng trong tìm kiếm và lựa chọn sản phẩm phù hợp. Đề tài bao gồm các bước thu thập và xử lý dữ liệu từ các nền tảng thương mại điện tử phổ biến ở Việt Nam có bán các mặt hàng thời trang nữ đồng thời phân tích hành vi mua sắm và đặc điểm các sản phẩm có ở những nền tảng này mà cụ thể ở đây là Lazada và Tiki. Các nguồn dữ liệu sẽ được chọn lọc và kiểm tra kỹ lưỡng để đảm bảo độ tin cậy và đa dạng. Dựa trên những gì phân tích được, hệ thống sẽ phát triển các mô hình máy học và thuật toán gợi ý thích hợp. Cuối cùng là xây dựng hệ thống gợi ý đề xuất sản phẩm từ hành vi mua hàng của người dùng.

1.3. Các bước thực hiện đề tài



Hình 1.1 Các bước thực hiện đề tài

1.4. Các công nghệ và lý do chọn công nghệ

1.4.1. Thư viện Selenium

Hình 1.2 Logo Selenium

Selenium là một thư viện mã nguồn mở, là một công cụ tự động hóa tương tác với trình duyệt web với hai chức năng chính là tương tác với trình duyệt và kiểm thử ứng dụng web. Selenium hỗ trợ nhiều ngôn ngữ như Python, C#, Java, ...

Trong đề tài này, ta sử dụng trình duyệt Chrome và ngôn ngữ Python để tương tác và thu thập dữ liệu từ các nền tảng thương mại điện tử. Với Python, ta có thể cài đặt Selenium bằng pip `pip install selenium`. Để thu thập dữ liệu từ trình duyệt, ta cần tải xuống driver tương ứng với trình duyệt (ở đề tài này là ChromeDriver) và cài đặt thư viện webdriver-manager (có thể cài đặt bằng pip `pip install webdriver-manager`).

Lợi thế của Selenium so với các thư viện với chức năng tương tự là Selenium hỗ trợ đa ngôn ngữ và nhiều trình duyệt; là công cụ mã nguồn mở và miễn phí phù hợp với nhiều đối tượng mà ở đây cụ thể là sinh viên đồng thời cho phép cộng đồng cải tiến liên tục; vì có một cộng đồng người dùng lớn, việc tìm hiểu cách sử dụng thông qua các hướng dẫn, mã code ví dụ và những hỗ trợ từ cộng đồng sẽ dễ dàng hơn [1].

1.4.2. Thư viện NLTK

NLTK (Natural Language Toolkit) là một thư viện mã nguồn mở, là công cụ phổ biến trong lĩnh vực NLP bằng ngôn ngữ Python. Là một nền tảng làm việc với dữ liệu là ngôn ngữ con người, NLTK có các chức năng chính: xử lý văn bản, phân tích câu trúc từ vựng và ngữ pháp, phân tích cảm xúc, … Ta có thể cài đặt thư viện NLTK thông qua pip `pip install nltk`, tải các tài nguyên và dữ liệu cần thiết có sẵn của thư viện bằng câu lệnh `nltk.download()`. Vì NLTK chủ yếu hỗ trợ tiếng Anh nên trong quá trình xử lý tiếng Việt cần cài đặt thêm thư viện Unicodedata để đảm bảo các thao tác NLP được nhất quán và chính xác. Có thể cài đặt thư viện Unicodedata bằng pip `pip install unicodedata` [2].

Điểm mạnh của NLTK là lượng tài nguyên phong phú và đầy đủ các công cụ phục vụ các nhiệm vụ NLP từ cơ bản đến nâng cao; khả năng tích hợp với những thư viện khác như Pandas trong xử lý dữ liệu, Matplotlib trong trực quan hóa dữ liệu, …; đồng thời vì là thư viện mã nguồn mở, được sử dụng và hỗ trợ bởi số lượng lớn cộng đồng người dùng nên giúp dễ dàng tham khảo các giáo trình hướng dẫn, các mã code có sẵn miễn phí và các diễn đàn lớn từ cộng đồng.

1.4.3. Thư viện Numpy

Hình 1.3 Logo NumPy

NumPy là một thư viện mã nguồn mở phổ biến, được sử dụng chủ yếu cho tính toán khoa học và xử lý các mảng (arrays) trong Python. NumPy được xem là nền tảng của nhiều thư viện khác như Pandas, Matplotlib và SciPy. Để cài đặt NumPy. Có thể sử dụng pip với lệnpip install numpy.

Ưu điểm của NumPy là khả năng tính toán nhanh chóng nhờ việc sử dụng mảng đa chiều (ndarray), giúp tiết kiệm thời gian và bộ nhớ so với việc sử dụng danh sách thông thường trong Python. NumPy hỗ trợ nhiều phép toán số học và ma trận, đồng thời tích hợp tốt với các thư viện khoa học khác. Cộng đồng lớn, nhiều tài liệu hướng dẫn và mã ví dụ cũng giúp người dùng dễ dàng làm quen và sử dụng. Có thể cài đặt NumPy bằng pip `pip install numpy` [3].

1.4.4. Thư viện Pandas

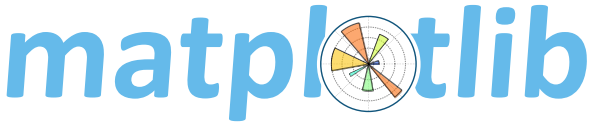
Hình 1.4 Logo Pandas

Pandas là một thư viện mã nguồn mở chuyên dùng để xử lý và phân tích dữ liệu. Với khả năng nổi bật là làm việc với dữ liệu dạng bảng, Pandas là một công cụ xử lý dữ

liệu tốt với đầy đủ các tính năng lọc, tổng hợp, xử lý thiếu, nhiễu, lọc, … dữ liệu. Có thể cài đặt Pandas bằng pip `pip install pandas`.

Ưu điểm của Pandas là khả năng linh hoạt và tích hợp tốt với các thư viện khác giúp hỗ trợ tốt hơn trong thao tác với dữ liệu. Thư viện này cũng là mã nguồn mở và được hỗ trợ bởi cộng đồng lớn, cung cấp nhiều tài liệu hướng dẫn và ví dụ mã, giúp dễ dàng tìm kiếm và giải quyết các vấn đề liên quan đến dữ liệu [4].

1.4.5. Thư viện Matplotlib



Hình 1.5 Logo Matplotlib

Matplotlib là một thư viện mã nguồn mở được sử dụng để tạo các biểu đồ và hình ảnh hóa, trực quan hóa dữ liệu. Matplotlib phổ biến nhờ việc hỗ trợ nhiều loại biểu đồ, giúp người dùng dễ dàng tạo ra các loại biểu đồ một cách trực quan và chuyên nghiệp. Có thể cài đặt Matplotlib bằng pip `pip install matplotlib`.

Hỗ trợ nhiều kiểu dáng, màu sắc và khả năng tùy chỉnh các chú thích, tiêu đề, trục, … Matplotlib giúp người dùng dễ dàng cá nhân hóa cũng như điều chỉnh các biểu đồ theo nhu cầu cụ thể. Hơn nữa khả năng tích hợp tốt với các thư viện khác như Numpy, Pandas giúp quá trình xử lý và hình ảnh hóa dữ liệu trở nên dễ dàng hơn [5].

1.4.6. Thư viện Seaborn

Hình 1.6 Logo Seaborn

Seaborn là một thư viện mã nguồn mở được xây dựng dựa trên Matplotlib, chuyên dùng để tạo các biểu đồ và trực quan hóa dữ liệu. Seaborn phổ biến nhờ việc cung cấp các biểu đồ thống kê nâng cao và trực quan hơn, giúp người dùng dễ dàng tạo ra các biểu đồ đẹp mắt và giàu thông tin. Có thể cài đặt Seaborn bằng pip `pip install seaborn`.

Seaborn hỗ trợ nhiều loại biểu đồ trực quan như biểu đồ phân tán, biểu đồ tương quan, biểu đồ hộp (box plot), và biểu đồ mật độ, với khả năng tùy chỉnh cao về màu sắc, kích thước và chú thích. Thư viện này giúp người dùng dễ dàng tạo ra các biểu đồ phức tạp chỉ với vài dòng mã. Seaborn tích hợp tốt với Pandas và Numpy, giúp việc xử lý và trực quan hóa dữ liệu từ các tập dữ liệu trở nên mượt mà và hiệu quả hơn [6].

1.4.7. Thư viện Scikit-learn

Hình 1.7 Logo Scikit-learn

Scikit-learn là một thư viện mã nguồn mở nổi bật trong lĩnh vực học máy (machine learning) và phân tích dữ liệu. Được xây dựng dựa trên các công cụ khoa học tính toán của Python như NumPy, SciPy và Matplotlib, scikit-learn cung cấp một bộ công cụ phong phú cho các bài toán học máy, từ hồi quy và phân loại đến phân cụm và giảm chiều. Thư viện này được yêu thích nhờ vào giao diện đơn giản, dễ sử dụng và tích hợp tốt với các thư viện khác trong hệ sinh thái Python. Có thể cài đặt scikit-learn bằng pip với lệnh `pip install scikit-learn`.

Scikit-learn hỗ trợ nhiều thuật toán học máy, bao gồm hồi quy tuyến tính, cây quyết định, hồi quy logistic, và các phương pháp ensemble như Random Forest và Gradient Boosting. Thư viện này cũng cung cấp các công cụ để đánh giá mô hình, lựa chọn đặc trưng và tối ưu hóa siêu tham số, giúp người dùng dễ dàng xây dựng, kiểm tra và triển khai các mô hình học máy hiệu quả. Scikit-learn tích hợp tốt với Pandas và NumPy, làm cho việc tiền xử lý dữ liệu và phân tích trở nên nhanh chóng và thuận tiện [7].

1.4.8. Thư viện Gensim

Hình 1.8 Logo Gensim

Gensim là một thư viện mã nguồn mở chuyên dùng để xử lý ngôn ngữ tự nhiên (NLP). Thư viện này nổi bật với khả năng xây dựng các mô hình vector hóa văn bản

như Word2Vec, Doc2Vec, và các mô hình phân phối chủ đề như Latent Dirichlet Allocation (LDA). Có thể cài đặt Gensim qua pip `pip install gensim`.

Ưu điểm của Gensim là khả năng xử lý văn bản lớn một cách hiệu quả, cùng với sự tích hợp dễ dàng với các công cụ học máy khác. Gensim hỗ trợ nhiều thuật toán hiện đại cho NLP và có thể dễ dàng mở rộng cho các tập dữ liệu lớn. Đồng thời cũng có cộng đồng hỗ trợ mạnh mẽ và tài liệu phong phú, giúp người dùng nhanh chóng làm quen và triển khai các mô hình NLP phức tạp [8].

1.4.9. Thư viện Streamlit

Hình 1.9 Logo Streamlit

Streamlit là một thư viện mã nguồn mở hỗ trợ việc xây dựng và triển khai các ứng dụng web tương tác một cách nhanh chóng, đặc biệt trong lĩnh vực phân tích dữ liệu và học máy. Thư viện này cho phép các nhà phát triển biến các tập lệnh Python đơn giản thành giao diện web đầy đủ tính năng mà không cần kiến thức sâu về lập trình web. Streamlit có thể được cài đặt qua pip với lệnh `pip install streamlit`.

Ưu điểm của Streamlit nằm ở việc cho phép xây dựng các giao diện người dùng đơn giản và hỗ trợ việc hiển thị các biểu đồ, bảng dữ liệu, và các mô hình học máy trực tiếp trên trình duyệt. Thư viện này tương thích tốt với nhiều công cụ phân tích và học máy khác như Matplotlib, Plotly, và Scikit-learn. Cộng đồng mạnh mẽ và tài liệu phong phú giúp Streamlit trở thành một công cụ hữu ích cho việc trình bày và tương tác với các mô hình và kết quả phân tích dữ liệu [9].

1.5. Thông tin về nguồn dữ liệu

Dữ liệu phân tích được thu thập bằng thư viện Selenium từ hai trang thương mại điện tử lớn ở Việt Nam là Lazada và Tiki.

1.5.1. Lazada

Hình 1.10 Logo Lazada

Lazada là một trong những công ty điều hành thương mại điện tử lớn nhất Đông Nam Á. Thành lập từ năm 2012 bởi Maximilian Bittner, Lazada cho phép các nhà bán lẻ và bên thứ ba bán sản phẩm trên nền tảng của mình. Lazada hoạt động tại nhiều quốc gia trong khu vực, trong đó có Việt Nam [10].

Thị trường thời trang trên Lazada đang phát triển mạnh mẽ, trở thành một trong những danh mục sản phẩm phổ biến nhất trên nền tảng này. Với nhiều nhà bán lẻ cung cấp đủ loại sản phẩm từ quần áo, giày dép, phụ kiện cho cả nam, nữ và trẻ em, Lazada thu hút đông đảo người mua sắm với đa dạng lựa chọn [11].

1.5.2. Tiki

Hình 1.11 Logo Tiki

Tiki là một trong những nền tảng thương mại điện tử hàng đầu tại Việt Nam, được thành lập vào năm 2010 bởi Trần Ngọc Thái Sơn. Ban đầu, Tiki tập trung vào việc bán sách trực tuyến, nhưng sau đó đã mở rộng sang nhiều ngành hàng khác. Tiki không chỉ phục vụ thị trường Việt Nam mà còn không ngừng nâng cao chất lượng dịch vụ để cạnh tranh với các đối thủ quốc tế [12].

Thị trường thời trang trên Tiki cũng đang ngày càng thu hút sự chú ý của người tiêu dùng. Tiki cung cấp đa dạng các sản phẩm thời trang từ quần áo, giày dép đến phụ kiện cho mọi lứa tuổi. Với các chương trình khuyến mãi, giao hàng nhanh và dịch vụ chăm sóc khách hàng tốt, Tiki dần trở thành một địa chỉ đáng tin cậy cho người mua sắm thời trang trực tuyến tại Việt Nam.

1.6. Mô tả dữ liệu

Dữ liệu thu thập được qua quá trình trích xuất dữ liệu tự động từ trang web bằng thư viện Selenium sẽ được lưu trữ dưới dạng file text (*.txt) để backup dữ liệu và file CSV (*.csv) để dễ dàng chuyển đổi thành dạng DataFrame trong quá trình xử lý và phân tích, với các file có tên chứa 'cleaned_' là các file đã được tiền xử lý và những file *.csv còn lại là file dữ liệu gốc khi crawl từ web. Tất cả file dữ liệu được lưu trữ trên Github. Bao gồm:

- Laz_product_urls.txt
- lazada_products.csv
- lazada_feedbacks.csv
- cleaned_lazada_products.csv
- cleaned_lazada_feedbacks.csv
- Tiki_product_urls.txt
- tiki_products.csv
- tiki_feedbacks.csv
- cleaned_tiki_products.csv
- cleaned_tiki_feedbacks.csv

1.6.1. Các sản phẩm của Lazada

Bộ dữ liệu gồm 120 sản phẩm với từ khóa 'Quần áo nữ'. Gồm 5 trường:

- Url: đường dẫn của từng sản phẩm. (string)
- Name: tên sản phẩm. (string)
- Price: đơn giá mỗi sản phẩm được rao bán trên sàn Lazada tại thời điểm thu thập dữ liệu. (float)
- Sold: Số lượng sản phẩm đã bán tại thời điểm thu thập dữ liệu. (int)
- Category: Loại sản phẩm được phân loại bằng phương pháp n-grams dựa trên trường 'Name' và quy ước được quy định sẵn được lưu trong biến '*categories_keywords*' bằng kiểu dữ liệu 'dict'. (string)

```
for c, k in categories_keywords.items():
    print('{0}: {1}'.format(c, k))
```

```
đồ bộ: ['đồ bộ', 'set', 'bộ', 'quần áo', 'sét']
đồ lót: ['đồ lót', 'underwear', 'quần lót', 'áo lót', 'áo ngực', 'bra', 'panty', 'boxer', 'quần chíp']
váy/quần: ['váy', 'quần', 'skirt', 'pants', 'jeans']
áo: ['áo', 't-shirt', 'shirt', 'croptop', 'yếm', 'khoác', 'áo khoác', 'vest', 'blazer']
đầm: ['đầm', 'dress', 'váy ngủ']
nón: ['nón', 'mũ', 'hat', 'helmet']
vớ: ['vớ', 'tất', 'socks']
giày/dép: ['giày', 'dép', 'bata', 'cao gót', 'shoes', 'slipper', 'guốc', 'boots', 'xăng đan', 'sandals', 'sneakers']
trang sức: ['trang sức', 'nhẫn', 'vòng', 'dây chuyền', 'earrings', 'necklace']
```

Hình 1.12 Các giá trị trong ‘categories_keywords’

1.6.2. Các đánh giá sản phẩm của Lazada

Bộ dữ liệu gồm 9786 đánh giá của 120 sản phẩm ở bộ dữ liệu trên. Dữ liệu đã được xử lý xóa trùng và xóa rỗng. Gồm 2 trường:

- Product_Url: đường dẫn của sản phẩm. (string)
- Content: đánh giá của khách hàng đối với sản phẩm. (string)

1.6.3. Các sản phẩm của Tiki

Tương tự như ở Lazada, bộ dữ liệu gồm 120 sản phẩm với từ khóa ‘Quần áo nữ’. Gồm 5 trường:

- Url: đường dẫn của từng sản phẩm. (string)
- Name: tên sản phẩm. (string)
- Price: đơn giá mỗi sản phẩm được rao bán trên sàn Lazada tại thời điểm thu thập dữ liệu. (float)
- Sold: Số lượng sản phẩm đã bán tại thời điểm thu thập dữ liệu. (int)
- Category: Loại sản phẩm được phân loại bằng phương pháp n-grams dựa trên trường ‘Name’ và quy ước được quy định sẵn được lưu trong biến ‘*categories_keywords*’ bằng kiểu dữ liệu ‘dict’ như Hình 1.12. (string)

1.6.4. Các đánh giá sản phẩm của Tiki

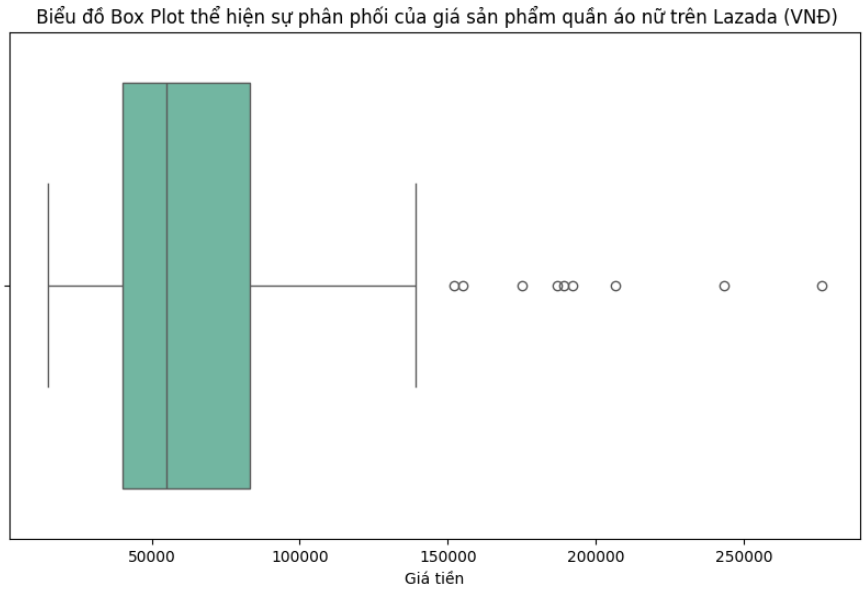
Bộ dữ liệu gồm 399 đánh giá của 120 sản phẩm ở bộ dữ liệu trên. Dữ liệu đã được xử lý xóa trùng và xóa rỗng. Gồm 2 trường:

- Product_Url: đường dẫn của sản phẩm. (string)
- Content: đánh giá của khách hàng đối với sản phẩm. (string)

1.7. Trực quan hóa dữ liệu

1.7.1. Các sản phẩm của Lazada

Hình 1.13 Biểu đồ cột các thông số thống kê giá tiền sản phẩm trên Lazada



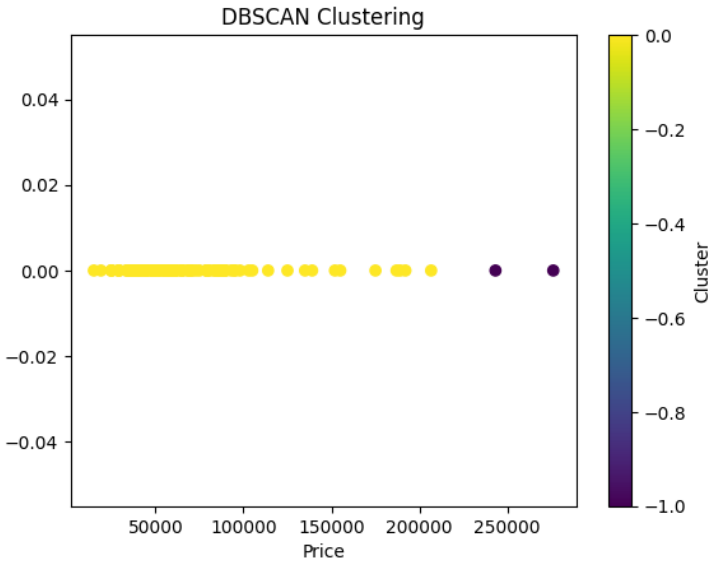
Hình 1.14 Biểu đồ box plot biểu thị sự phân phối giá tiền sản phẩm trên Lazada

Giá tiền cao nhất cho sản phẩm quần áo nữ là 276.051 VNĐ và thấp nhất là 15.000 VNĐ. Điều này cho thấy sự chênh lệch giá của các sản phẩm khá lớn.

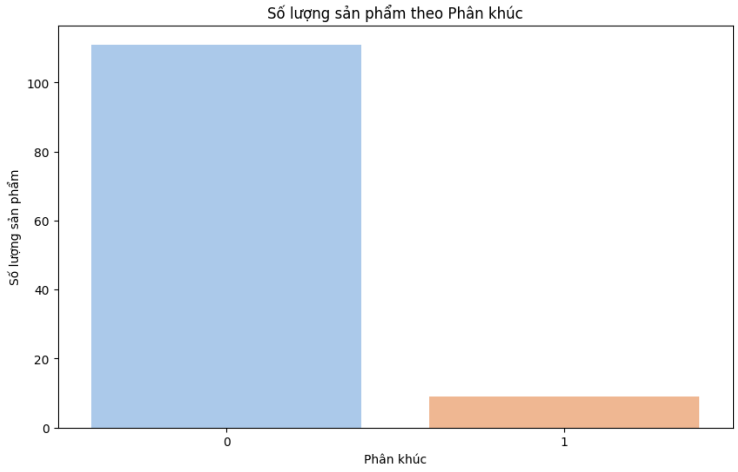
Giá trung bình của sản phẩm là khoảng 68.981 VNĐ, trong khi độ lệch chuẩn là 44.870 VNĐ. Tỷ lệ giữa độ lệch chuẩn và giá trị trung bình đạt khoảng 0.65, cho thấy mức độ phân tán của dữ liệu là tương đối cao. Điều này chỉ ra rằng dữ liệu có sự biến

động đáng kể. Hơn nữa, như đã trình bày trong Hình 1.14, trường dữ liệu ‘Price’ có khá nhiều giá trị ngoại lai.

Sự xuất hiện nhiều giá trị ngoại lai trong giá tiền của các sản phẩm quần áo nữ cho thấy sự đa dạng của sản phẩm được bán trong nền tảng này. Điều này có thể được lý giải bởi sự khác biệt về thương hiệu, chất liệu, kiểu dáng, và phân khúc khách hàng mà mỗi sản phẩm hướng đến. Sự đa dạng này tạo điều kiện cho người tiêu dùng có nhiều lựa chọn, từ các sản phẩm bình dân đến hàng cao cấp, phục vụ nhu cầu và sở thích đa dạng của thị trường.



Hình 1.15 Biểu đồ phân cụm bằng DBSCAN theo giá tiền sản phẩm trên Lazada

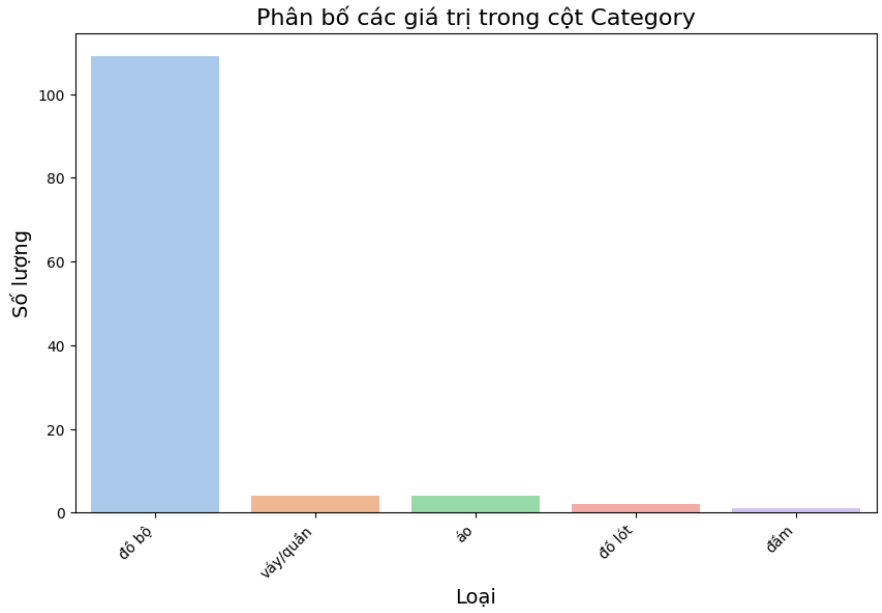


Hình 1.16 Biểu đồ thể hiện số lượng sản phẩm trong mỗi phân khúc theo giá tiền trên Lazada

Do dữ liệu giá tiền chứa nhiều giá trị ngoại lai và mục tiêu là phân tích và phân khúc khách hàng dựa trên giá tiền (với số lượng cụm chưa được xác định trước), thuật toán DBSCAN là sự lựa chọn phù hợp để thực hiện phân cụm.

Sau khi phân cụm bằng DBSCAN, ta có 2 cụm: cụm 1 (màu vàng) với giá tiền nằm dưới 207.000 VNĐ chiếm đa số và cụm 2 (màu tím) với những giá trị còn lại (Hình 1.15).

Tuy nhiên, dữ liệu tập trung chủ yếu ở mức giá dưới 150.000 VNĐ nên nếu chỉ xét theo giá tiền ta có 2 phân khúc khách hàng chính trên nền tảng Lazada là phân khúc giá thấp (dưới 150.000 VNĐ) và phân khúc giá cao (trên 150.000 VNĐ) theo Hình 1.16.



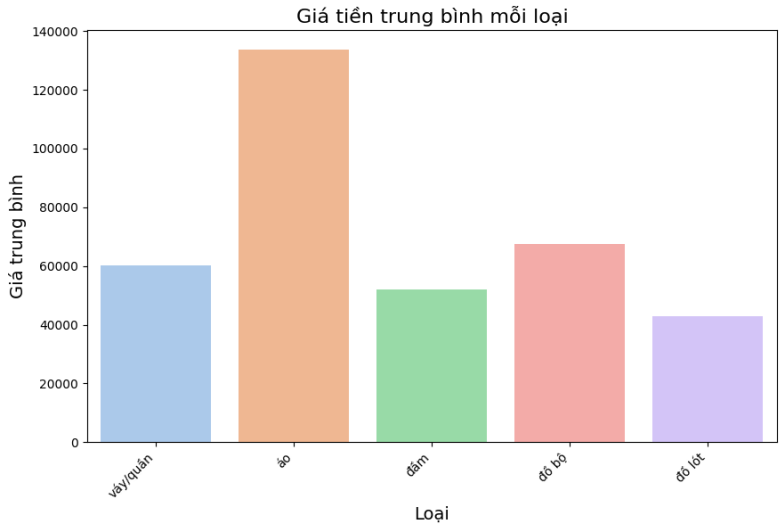
Hình 1.17 Số lượng sản phẩm ở mỗi loại sản phẩm quần áo nữ trên Lazada

Giá trị trong cột ‘Category’ được phân loại từ các giá trị trong cột ‘Name’ bằng phương pháp n-grams với chuẩn được quy định sẵn (Hình 1.12). Do việc xử lý NLP Tiếng Việt chưa được thực hiện một cách toàn diện, các giá trị không thể phân loại bằng phương pháp n-grams đã được xử lý thủ công.

Sau khi phân loại, trường ‘Category’ trong bộ dữ liệu các sản phẩm của Lazada có 5 loại: đồ bộ (109 sản phẩm), váy/quần (4 sản phẩm), áo (4 sản phẩm), đồ lót (2 sản phẩm) và đầm (1 sản phẩm).

Từ Hình 1.17, có thể thấy loại ‘đồ bộ’ chiếm đa số và chiếm số lượng quá lớn so với các loại quần áo nữ còn lại. Cho thấy xu hướng đăng bán sản phẩm trên nền tảng

Lazada tập trung vào các mặt hàng đồ bộ như đồ ngủ và các set đồ đã phối sẵn nhằm thuận tiện hơn cho khách hàng và tăng độ đa dạng trong chọn lựa sản phẩm.



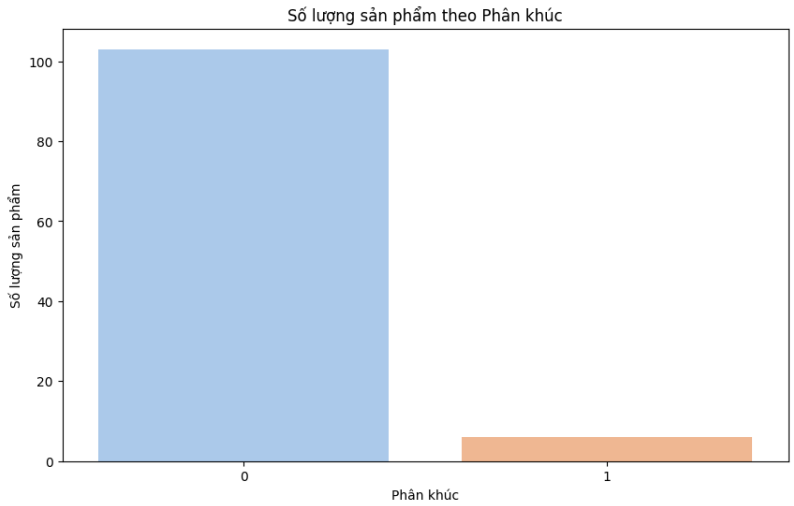
Hình 1.18 Giá tiền trung bình mỗi loại sản phẩm quần áo nữ trên Lazada

	Name	Price	Sold	Category
26	Đồ Ngủ Gợi Cảm Vải Satiin Đẹp BT FASHION (Yếm ...	35400.0	561	áo
39	ZANZEA Women Vintage Casual Sleeveless Top Ela...	243200.0	0	áo
40	Áo khoác nam Hoodie Chống Nắng WHENEVER Nỉ Bôn...	48999.0	537	áo
98	ZANZEA Korean Style Women's 2pcs Suits New Fas...	206646.0	31	áo

Hình 1.19 Các sản phẩm loại áo trên Lazada

Đối với áo, vì số lượng sản phẩm trên tổng số sản phẩm là quá thấp (4/120 sản phẩm) và giá tiền các sản phẩm trải đều ở cả hai phân khúc giá cao và giá thấp nên giá tiền trung bình của sản phẩm không phản ánh được ý nghĩa thực tế của sản phẩm áo trên nền tảng này.

Tương tự với các loại sản phẩm váy/quần, đầm và đồ lót, vì số lượng sản phẩm trên tổng số là quá ít nên ở đây ta chỉ tập trung phân tích loại sản phẩm đồ bộ.

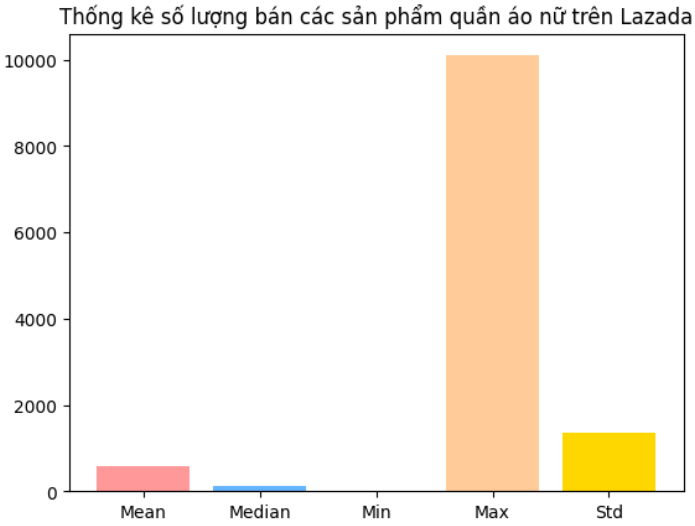


Hình 1.20 Biểu đồ thể hiện số lượng sản phẩm đồ bộ trong mỗi phân khúc theo giá tiền trên Lazada

	Name	Price	Sold	Category
23	Recool Short Sleeve Top and Wide Leg Pants Kni...	276051.0	9	đồ bộ
56	MỘT NÚT THÊU KATE, Trang phục truyền thống, Qu...	152000.0	1400	đồ bộ
63	Đồ Bộ Nữ Thiết Kế Mới 2024, Set Đồ Nữ, Áo Sát ...	175000.0	53	đồ bộ
107	Bộ Áo Cộc Cổ Sơ Mi, Quần Dài, Thiết Kế Lịch Sự...	187000.0	18	đồ bộ
110	Set Bộ Sát Nách Nữ Thiết Kế, Cổ Áo Sơ Mi Quần ...	192000.0	9	đồ bộ
114	Quần Áo Nữ Sang Chảnh, Áo Sát Nách Quần suông,...	189000.0	12	đồ bộ

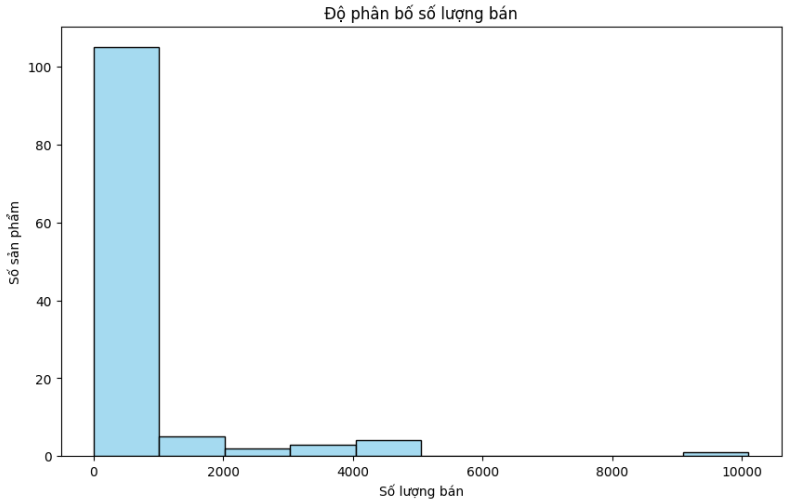
Hình 1.21 Các sản phẩm đồ bộ có phân khúc giá cao trên Lazada

Trong 9 sản phẩm thuộc phân khúc giá cao có 6 sản phẩm là loại đồ bộ (Hình 1.21). Trong đó, sản phẩm có giá tiền cao nhất thuộc phân loại đồ bộ và sản phẩm đồ bộ rẻ nhất có giá tiền là 19.000 VNĐ (chênh lệch không quá nhiều so với mức giá thấp nhất là 15.000 VNĐ). Hơn nữa, sự chênh lệch giữa số lượng sản phẩm loại đồ bộ phân khúc giá thấp và sản phẩm phân khúc giá cao (Hình 1.20) cũng tương tự với số liệu trên toàn bộ sản phẩm (Hình 1.16). Có thể nói sản phẩm loại này phản ánh sự phân bố của cả 120 sản phẩm trên Lazada.



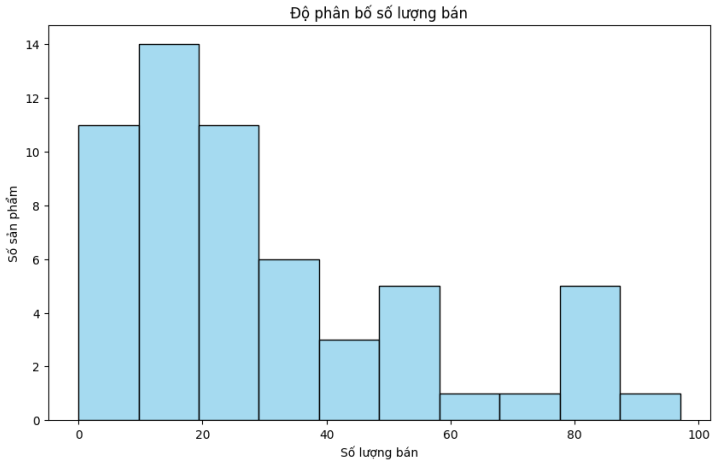
Hình 1.22 Biểu đồ cột các thông số thống kê số lượng bán sản phẩm trên Lazada

Giá trị nhỏ nhất là 0 và lớn nhất 10100 cho thấy số lượng bán giữa các sản phẩm quần áo nữ có độ chênh lệch khá cao.



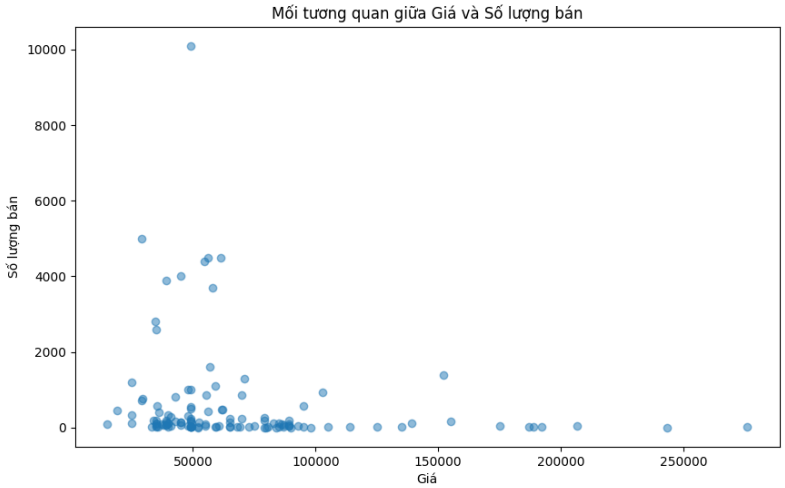
Hình 1.23 Biểu đồ cột biểu thị độ phân bố số lượng bán sản phẩm trên Lazada

Nhìn chung, giá tiền ảnh hưởng khá nhiều đến doanh thu bán hàng. Phần lớn sản phẩm có số lượng bán rất thấp, tập trung ở mức gần bằng 0. Điều này có nghĩa là hầu hết các sản phẩm trong tập dữ liệu không bán được nhiều. Có 6 sản phẩm chưa được bán ra (số lượng bán là 0) ở thời điểm dữ liệu được thu thập. Phân phối lệch phải rất rõ ràng, số lượng sản phẩm có số lượng bán cao (lên tới 2000, 4000 hoặc 10000) là rất nhỏ. Các sản phẩm có số lượng bán lớn chiếm một tỷ lệ rất nhỏ so với tổng số.



Hình 1.24 Biểu đồ cột biểu thị độ phân bố số lượng bán từ 0 đến dưới 100 của các sản phẩm quần áo nữ trên Lazada

Ở Hình 1.24, ta xem xét chi tiết hơn về độ phân bố số lượng bán của các sản phẩm trong khoảng từ 0 đến dưới 100. Biểu đồ tiếp tục cho thấy sự phân bố không đồng đều, với phần lớn sản phẩm có số lượng bán từ 0 đến 40. Có thể thấy rằng nhóm sản phẩm có số lượng bán từ 10 đến dưới 30 chiếm tỷ lệ lớn nhất, với đỉnh ở khoảng 20 - 30. Sau đó, số lượng sản phẩm bán giảm dần khi số lượng bán tăng. Một số sản phẩm có số lượng bán cao hơn 60 vẫn xuất hiện, mặc dù không nhiều. Các sản phẩm bán từ 60 đến 100 tuy ít nhưng có thể là những mặt hàng nổi bật, do đó cần tập trung khai thác thêm tiềm năng từ những sản phẩm này.

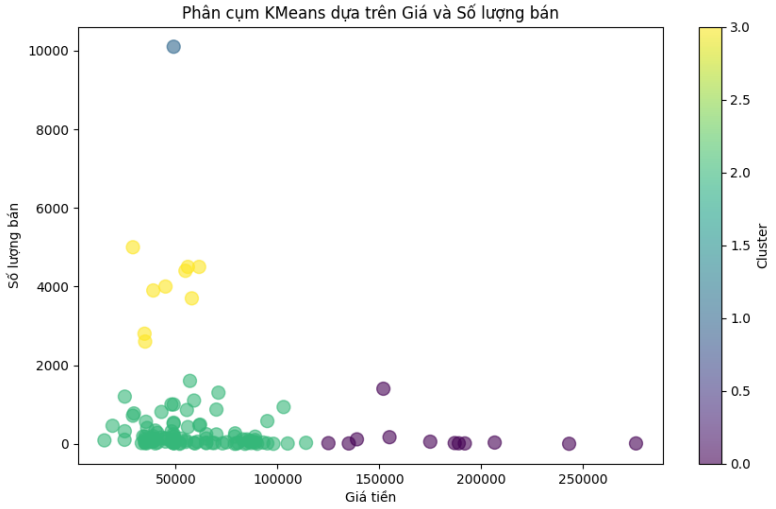


Hình 1.25 Biểu đồ tương quan giữa giá tiền và số lượng bán sản phẩm trên Lazada

Nhìn chung những sản phẩm có số lượng bán cao tập trung ở phân khúc giá thấp và đặc biệt tập trung ở những sản phẩm có mức giá bán 50.000 VNĐ. Khi giá sản phẩm tăng lên từ 50.000 VNĐ đến khoảng 150.000 VNĐ, số lượng bán bắt đầu giảm rõ rệt.

Rất ít sản phẩm trong khoảng giá này có số lượng bán đáng kể. Ý nghĩa của kết quả này cho thấy khách hàng có xu hướng mua nhiều hơn các sản phẩm giá thấp, nhất là những sản phẩm quần áo nữ có mức giá dưới 150.000 VNĐ.

Dù có xu hướng giảm số lượng bán khi giá tăng, nhưng không có một mối tương quan tuyến tính rõ ràng giữa giá tiền và số lượng bán. Một số sản phẩm có giá thấp vẫn có số lượng bán rất ít, trong khi một số sản phẩm giá thấp lại bán cực kỳ nhiều. Điều này chỉ ra rằng giá cả không phải là yếu tố duy nhất quyết định số lượng bán. Nhiều yếu tố khác có thể ảnh hưởng đến quyết định mua hàng của khách hàng.



Hình 1.26 Biểu đồ phân cụm Kmeans dựa trên giá tiền và số lượng bán sản phẩm quần áo nữ trên Lazada

Sau khi phân cụm bằng thuật toán K-means với số cụm là 4, ta có được 4 cụm tương ứng với 4 màu như Hình 1.26. Biểu đồ cho thấy dữ liệu đã được chia thành 4 cụm:

- Cụm 1 (xanh lá): gồm các sản phẩm có mức giá thấp và số lượng bán < 2000 sản phẩm. Đây là cụm có số lượng sản phẩm tập trung dày đặc nhất.
- Cụm 2 (tím): gồm các sản phẩm có mức giá trung bình – cao (xấp xỉ 120.000 VNĐ trở lên). Tất cả sản phẩm ở cụm này đều có số lượng bán ở mức dưới 2000 sản phẩm.
- Cụm 3 (vàng): gồm các sản phẩm có mức giá thấp (xấp xỉ dưới 60.000 VNĐ) và có số lượng bán từ 2000 đến 6000 sản phẩm.
- Cụm 4 (xanh dương): gồm chỉ 1 sản phẩm với mức giá xấp xỉ 50.000 VNĐ và số lượng bán 10100 – mức bán cao nhất trên 120 sản phẩm trong bộ dữ liệu.

Tương ứng với 4 cụm trên là 3 đối tượng sản phẩm với các chiến lược kinh doanh phù hợp. Trong đó, vì cụm 4 chỉ có 1 phần tử và đặc điểm về mức giá giống nhau nên đối tượng 3 sẽ gồm cụm 3 và 4.

Tóm lại, qua những thông tin được trực quan hóa trên của bộ dữ liệu thông tin sản phẩm quần áo nữ trên nền tảng thương mại điện tử Lazada, ta rút ra được những đặc điểm như sau:

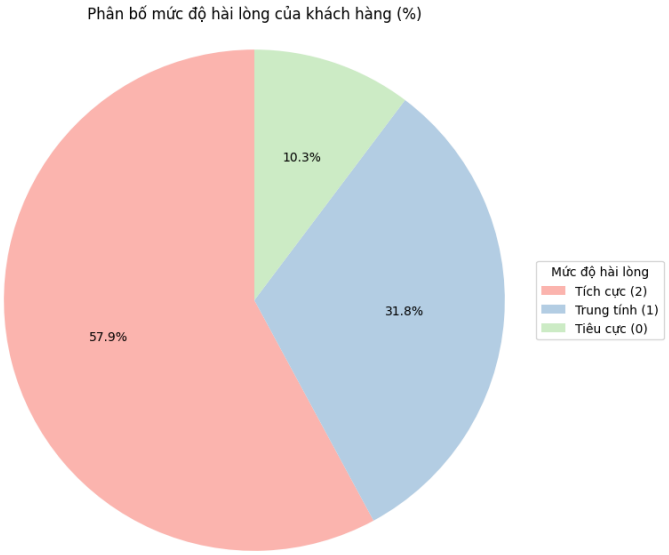
- Ở thời điểm thu thập dữ liệu, trên thị trường thương mại điện tử Lazada đồ bộ là mặt hàng được đăng bán nhiều nhất và chiếm đa số.
- Có 3 đối tượng sản phẩm chính với đối tượng 1 chiếm ưu thế về mặt số lượng.
- Đối với đối tượng 1 là những sản phẩm có mức giá thấp và số lượng bán < 2000, vì số lượng bán còn thấp, việc tối ưu hóa trải nghiệm mua sắm, tăng cường đánh giá tích cực, và cải thiện hình ảnh sản phẩm có thể giúp tăng doanh số. Đối tượng khách hàng mua những sản phẩm này có xu hướng mua sắm sản phẩm giá rẻ, vì thế cần tập trung hơn về các chiến lược quảng cáo và các chương trình khuyến mãi.
- Đối tượng sản phẩm 2 là đối tượng có mức giá trung bình – cao với số lượng bán tương đối thấp. Vì là sản phẩm giá cao hơn, khách hàng thường yêu cầu cao hơn về chất lượng sản phẩm. Thêm vào đó là yếu tố chăm sóc khách hàng và các chính sách bảo hành sản phẩm của các cửa hàng bán lẻ đăng bán trên Lazada cũng như các quy trình đổi trả, bảo hành của hệ thống.
- Cuối cùng, đối tượng 3 gồm các sản phẩm ở mức giá thấp và số lượng bán khá cao (từ 2000 sản phẩm trở lên). Trong đó, có sản phẩm có số lượng bán cao vượt trội (10100 sản phẩm). Với các sản phẩm thuộc đối tượng này, cần tập trung vào việc duy trì đà tăng trưởng và mở rộng đối tượng khách hàng. Các sản phẩm này đã có lượng bán tốt, vì vậy cần tiếp tục quảng bá và tận dụng lợi thế về giá.

1.7.2. Các đánh giá sản phẩm của Lazada

Ở bộ dữ liệu này ta tập trung phân tích những đánh giá của khách hàng với sản phẩm.

Sử kỹ thuật Word2Vec của thư viện Gensim nhằm xác định những từ khóa tích cực và tiêu cực trong các phản hồi của khách hàng. Bên cạnh đó, kỹ thuật n-grams của

thư viện NLTK được áp dụng để tính điểm hài lòng của khách hàng với sản phẩm dựa trên những từ khóa đã xác định. Trong đó, điểm ban đầu mỗi sản phẩm được thiết lập ở mức 0 biểu thị trạng thái trung tính, điểm âm cho thấy cảm xúc tiêu cực của khách hàng chiếm ưu thế và ngược lại điểm là số dương khi cảm xúc tích cực của khách hàng trong trải nghiệm sản phẩm nhiều hơn.



Hình 1.27 Biểu đồ phân bố mức độ hài lòng của khách hàng với các sản phẩm trên Lazada

Để dễ dàng phân tích và đánh giá, từ điểm hài lòng đã tính trên, ta tiến hành rời rạc hóa dữ liệu về dạng [0, 1, 2] với 3 trạng thái hài lòng của khách hàng tương ứng là tiêu cực, trung tính và hài lòng.

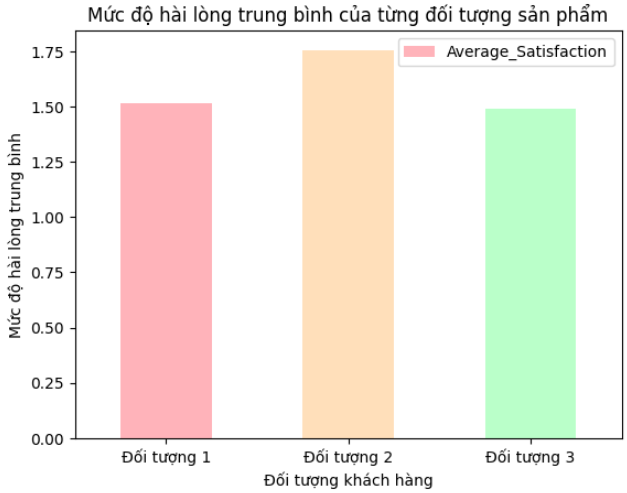
Hình 1.28 Wordcloud các từ khóa đánh giá tích cực các sản phẩm trên Lazada

Hình 1.29 Wordcloud các từ khóa đánh giá tiêu cực các sản phẩm trên Lazada

Kết quả Wordcloud cho các từ khóa đánh giá tích cực nổi bật với các từ như "thoải mái", "hoàn hảo", "chất lượng", … cho thấy sự hài lòng cao của người tiêu dùng với sản phẩm và dịch vụ. Ngược lại, Wordcloud cho các từ khóa tiêu cực chứa các từ như "xấu", "chất liệu", "chất liệu", … phản ánh những vấn đề mà khách hàng gặp phải.

Tỷ lệ của từng mức độ hài lòng được biểu diễn ở Hình 1.27. Trong đó, có 5239 (59%) đánh giá tích cực, 2946 (33.2%) đánh giá trung tính và 688 (7.8%) đánh giá tiêu cực.

Nhìn chung, lượng đánh giá tích cực chiếm đa số, cho thấy các sản phẩm đáp ứng được nhu cầu phần lớn khách hàng. Tuy nhiên, vẫn còn một phần nhỏ những đánh giá tiêu cực về chất lượng sản phẩm đến từ nhiều yếu tố như chất lượng một số sản phẩm còn chưa tốt, mẫu mã không phù hợp thị hiếu khách hàng, giá tiền còn chưa phù hợp với chất lượng, … mà chiếm đa số trong những đánh giá tiêu cực là những phàn nàn về chất lượng sản phẩm. Cho thấy rằng ngoài giá cả, chất lượng sản phẩm là yếu tố được khách hàng quan tâm hơn cả.



Hình 1.30 Biểu đồ cột trung bình mức độ hài lòng của khách hàng với các đối tượng sản phẩm trên Lazada

Mức độ hài lòng trung bình trên toàn bộ dữ liệu đánh giá sản phẩm của Lazada là 1.51, nghĩa là ở mức trung tính – tích cực. Trong đó, đối tượng sản phẩm 1 có mức độ hài lòng trung bình là 1.52, nghĩa là phản hồi về sản phẩm này thiên về trung tính nhưng vẫn có phần tích cực. Đối tượng 2 có mức hài lòng trung bình cao hơn, 1.76, gần mức tích cực, cho thấy sản phẩm được đánh giá tốt hơn. Và cuối cùng, đối tượng 3 có mức hài lòng trung bình 1.49, nghĩa là khách hàng cảm thấy trung bình về sản phẩm.

Nhìn tổng quát thì mức hài lòng trung bình các sản phẩm ở mức trung bình, có thiên hướng tích cực và sự chênh lệch giữa các đối tượng sản phẩm là không quá lớn.

Hình 1.31 Biểu đồ tròn phân bố mức độ hài lòng ở mỗi đối tượng sản phẩm trên Lazada

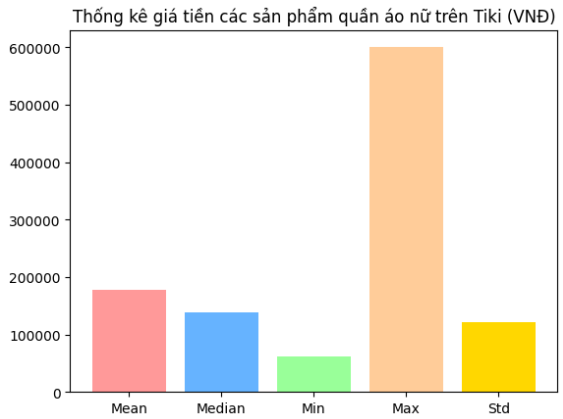
Đối tượng sản phẩm 2 tuy có giá tiền cao và lượng bán thấp nhưng phản hồi từ khách hàng tương đối tốt. Điều này có thể chỉ ra rằng sản phẩm có chất lượng tốt, đáp ứng hoặc vượt quá kỳ vọng của khách hàng, khiến họ sẵn sàng bỏ ra số tiền lớn và cảm

thấy hài lòng sau khi mua. Khách hàng có thể ít nhưng chất lượng phản hồi của họ cao, cho thấy sản phẩm có thể phù hợp với một nhóm nhỏ khách hàng nhưng đáp ứng rất tốt nhu cầu của nhóm này.

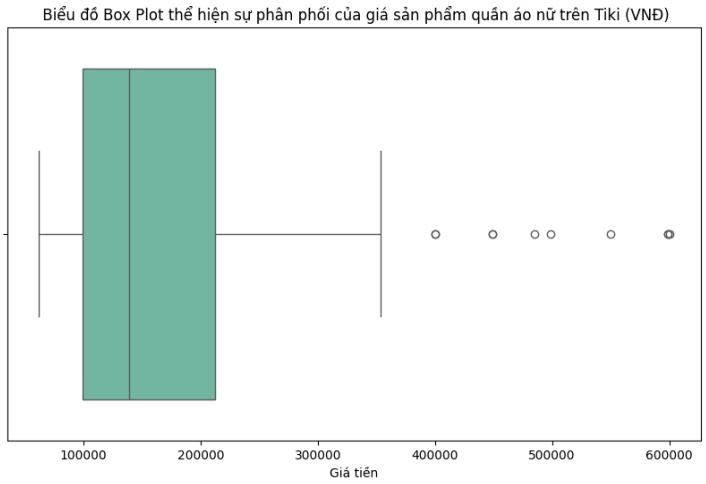
Đối với đối tượng sản phẩm 3 gồm những sản phẩm có giá thấp và số lượng bán cao, mặc dù giá sản phẩm thấp, nhưng độ hài lòng trung bình chỉ ở mức 1.49, gần với mức trung tính. Điều này chứng tỏ giá rẻ là yếu tố chính giúp sản phẩm loại này có số lượng bán cao, khách hàng mua sản phẩm ở đối tượng 3 ưu tiên giá tiền hơn trải nghiệm mua sắm và chất lượng sản phẩm. Mức độ hài lòng 1.49, nghĩa là trải nghiệm của khách hàng về sản phẩm không được như kỳ vọng. Dù số lượng bán cao nhưng nếu không cải thiện các yếu tố khác thì số lượng bán sẽ có nguy cơ giảm.

Cuối cùng, sản phẩm đối tượng 1 với mức giá rẻ và số lượng bán thấp có mức hài lòng trung bình 1.52 cho thấy khách hàng có trải nghiệm ở mức trung bình, không hoàn toàn tích cực nhưng cũng không quá tiêu cực. Sản phẩm có thể đáp ứng một số nhu cầu cơ bản của người dùng nhưng chưa thực sự nổi bật hoặc tạo ấn tượng tốt. Mức độ hài lòng trung bình cho thấy có tiềm năng để cải thiện sản phẩm. Nếu cải thiện được chất lượng hoặc nâng cấp, cải thiện thêm các yếu tố chăm sóc khách hàng, mẫu mã, … sản phẩm có thể thu hút nhiều người mua hơn và tăng số lượng bán.

1.7.3. Các sản phẩm của Tiki



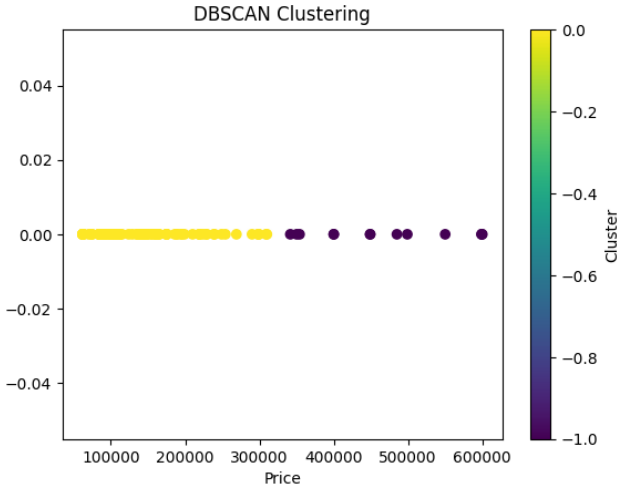
Hình 1.32 Biểu đồ cột các thông số thống kê giá tiền sản phẩm trên Tiki



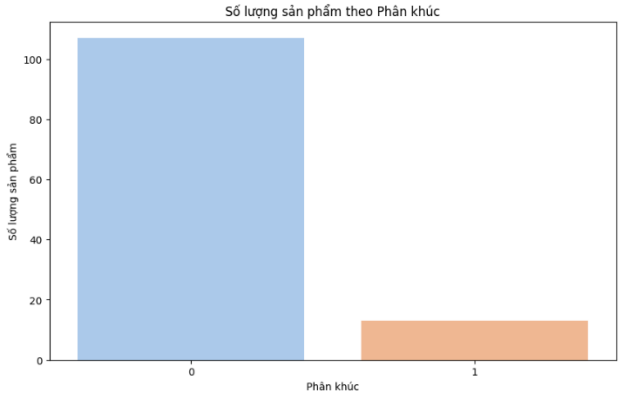
Hình 1.33 Biểu đồ box plot biểu thị sự phân phối giá tiền sản phẩm trên Tiki

Giá tiền cao nhất cho sản phẩm quần áo nữ là 600.000 VNĐ và thấp nhất là 62.000 VNĐ cho thấy sự chênh lệch giá của các sản phẩm khá lớn.

Trung bình giá tiền của sản phẩm là 177.410 VNĐ, một mức giá tương đối phù hợp với nhiều phân khúc người tiêu dùng. Tuy nhiên, độ lệch chuẩn 121.830 VNĐ cho thấy mức độ phân tán giá khá cao. Tỷ lệ giữa độ lệch chuẩn và giá trị trung bình là 0.69, cho thấy sự biến động giá khá mạnh. Điều này có thể xuất phát từ sự khác biệt về chất lượng hoặc thương hiệu của các sản phẩm quần áo nữ được đăng bán trên nền tảng. Ngoài ra, các giá trị ngoại lai như trong Hình 1.33 tuy nhiều nhưng không đáng kể.

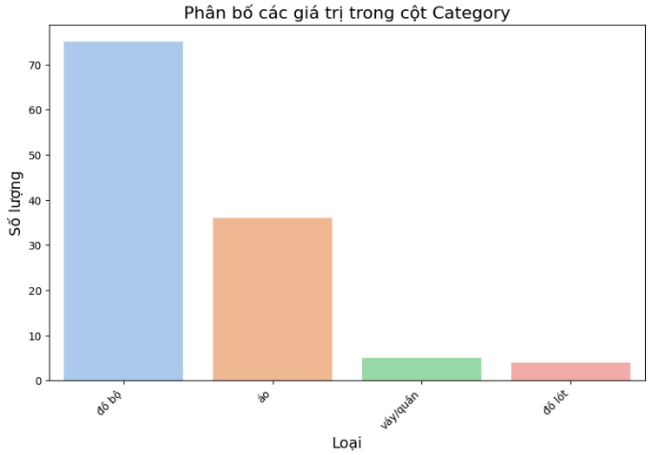


Hình 1.34 Biểu đồ phân cụm bằng DBSCAN theo giá tiền sản phẩm trên Tiki



Hình 1.35 Biểu đồ thể hiện số lượng sản phẩm trong mỗi phân khúc theo giá tiền trên Tiki

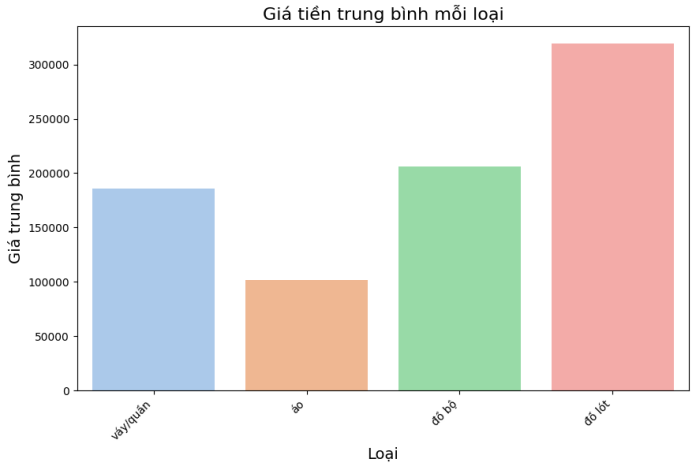
Sau khi phân cụm bằng DBSCAN, ta có 2 cụm: cụm 1 (màu vàng) với giá tiền ít hơn 310.000 VNĐ chiếm tỷ lệ cao hơn hẳn cụm 2 (màu tím) với những giá trị còn lại. ỞHình 1.36 Hình 1.35, phân khúc 0 đại diện cho cụm 1 và phân khúc 1 đại diện cho cụm 2.



Hình 1.36 Số lượng sản phẩm ở mỗi loại sản phẩm quần áo nữ trên Tiki

Sau khi phân loại, trường ‘Category’ trong bộ dữ liệu các sản phẩm của Lazada có 4 loại: đồ bộ (75 sản phẩm), áo (36 sản phẩm), váy/quần (5 sản phẩm) và đồ lót (4 sản phẩm).

Từ Hình 1.36, có thể thấy loại ‘đồ bộ’ chiếm đa số so với các loại quần áo nữ còn lại. Bên cạnh đó, sản phẩm loại áo cũng chiếm không ít, nhiều hơn hẳn 2 loại còn lại là váy/quần và đồ lót. Cho thấy xu hướng đăng bán sản phẩm trên nền tảng Tiki tập trung nhiều hơn vào các mặt hàng áo và đồ bộ.



Hình 1.37 Giá tiền trung bình mỗi loại sản phẩm quần áo nữ trên Tiki

	Name	Price	Category
24	Combo 10 Quần Lót Nữ Lưng Cao Modal Phối Ren M...	449000.0	đồ lót
39	Combo 5 quần lót nữ Modal Miley Lingerie BCS0704	189000.0	đồ lót
84	Combo 10 Quần Lót Nữ Lưng Cao Modal Phối Ren M...	449000.0	đồ lót
99	Combo 5 quần lót nữ Modal Miley Lingerie BCS0704	189000.0	đồ lót

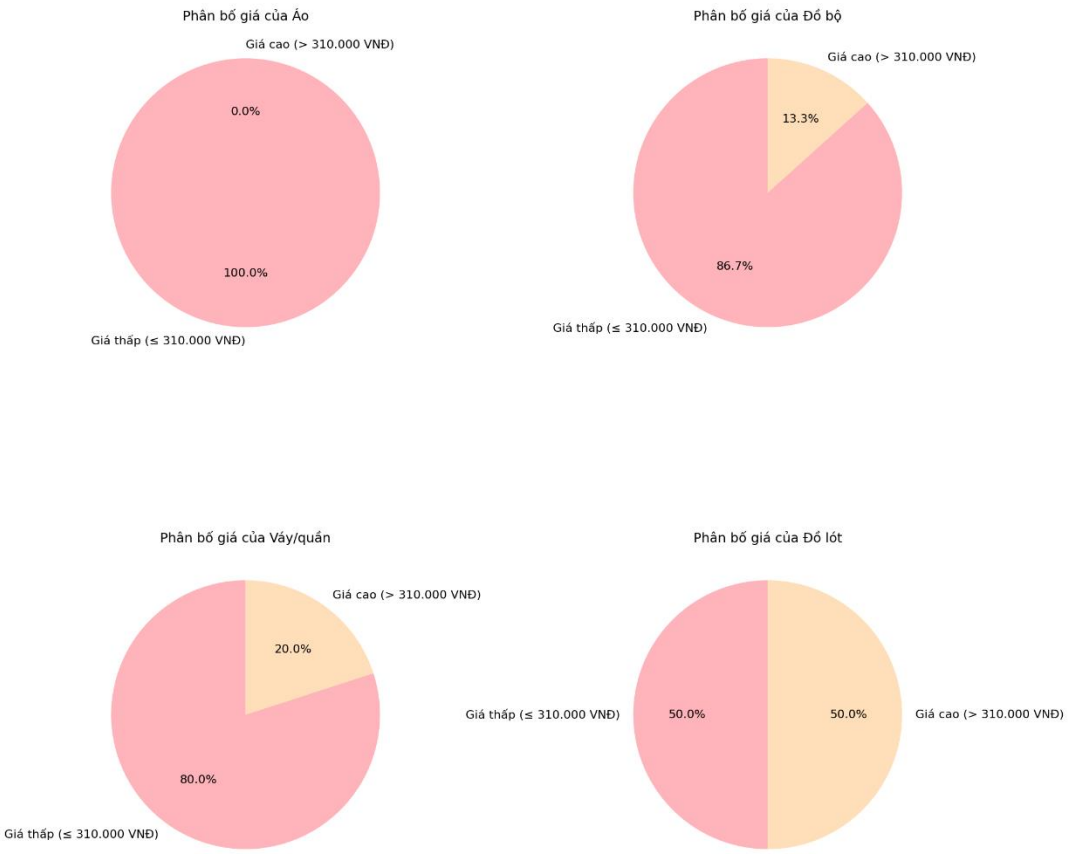
Hình 1.38 Các sản phẩm loại đồ lót trên Tiki

Qua Hình 1.37 ta có một số đánh giá khách quan về giá tiền của từng loại sản phẩm trên nền tảng Tiki. Mặc dù chỉ có 4 sản phẩm, giá trung bình của đồ lót lên đến 319.000 VNĐ, cao nhất trong các loại trang phục. Bên cạnh đó, giá tiền của các sản phẩm loại đồ lót cũng nằm trong mức trung bình – cao (trên mức giá trung bình). Điều này cho thấy các sản phẩm đồ lót có thể thuộc phân khúc cao cấp, nhắm vào đối tượng khách hàng tìm kiếm sản phẩm chất lượng cao.

Với số lượng 75 sản phẩm, chiếm tỷ trọng lớn nhất, đồ bộ có giá trung bình 205.782 VNĐ. Đây là một mức giá trung bình khá cao so với áo và váy/quần cho thấy đồ bộ được tập trung đăng bán nhiều để đáp ứng nhu cầu tiện lợi và thời trang của khách hàng.

Chỉ có 5 sản phẩm nhưng giá trung bình là 185.490 VNĐ, sự hạn chế về số lượng cho thấy váy/quần chưa phải là danh mục tập trung của Tiki, mặc dù mức giá trung bình vẫn thuộc phân khúc trung cấp.

Với 36 sản phẩm, áo có giá trung bình 101.444 VNĐ, thấp nhất trong các loại trang phục. Sự đa dạng về số lượng cho thấy áo là một mặt hàng phổ biến, nhưng giá thành thấp có thể phản ánh nhiều sản phẩm ở phân khúc giá thấp, phục vụ cho nhu cầu mua sắm với giá cả phải chăng.



Hình 1.39 Biểu đồ tròn phân bố phân khúc giá của mỗi loại sản phẩm trên Tiki

Tất cả các sản phẩm áo (100%) đều thuộc phân khúc giá thấp (≤ 310.000 VNĐ). Điều này cho thấy áo là loại sản phẩm có xu hướng nằm ở phân khúc giá thấp, phù hợp với người tiêu dùng có nhu cầu tìm kiếm các sản phẩm giá rẻ

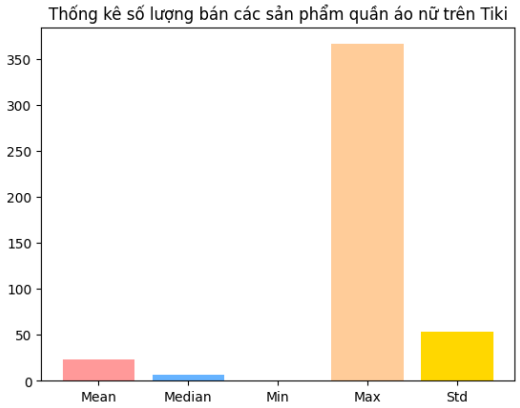
Với 86.7% sản phẩm thuộc phân khúc giá thấp, đồ bộ chủ yếu được bán ở mức giá phải chăng tuy nhiên vẫn có 13.3% nằm trong phân khúc giá cao. Điều này cho thấy sự đa dạng về giá của loại sản phẩm này.

Phân bố giá của đồ lót khá cân bằng, với 50% sản phẩm thuộc phân khúc giá thấp và 50% thuộc phân khúc giá cao. Như đã phân tích ở trên, sản phẩm loại đồ lót nhắm vào đối tượng khách hàng ưu tiên chất lượng cao hơn mức giá.

	Name	Price	Category
19	COMBO ĐỒ LÍNH 3 IN 1 ÁO QUẦN KHOÁC CAO CẤP	341550.0	váy/quần
34	Jumpsuit (Áo Liền Quần) Lụa Cổ Tròn Tay Ngắn S...	199000.0	váy/quần
54	Áo Jean Liền Quần Thời Trang	195000.0	váy/quần
91	Áo Sơ Mi Nữ Form Dài Che Quần Cá Tính SM013 Ma...	128900.0	váy/quần
119	áo thun nữ,áo phông nữ cộc tay phối màu in ...	63000.0	váy/quần

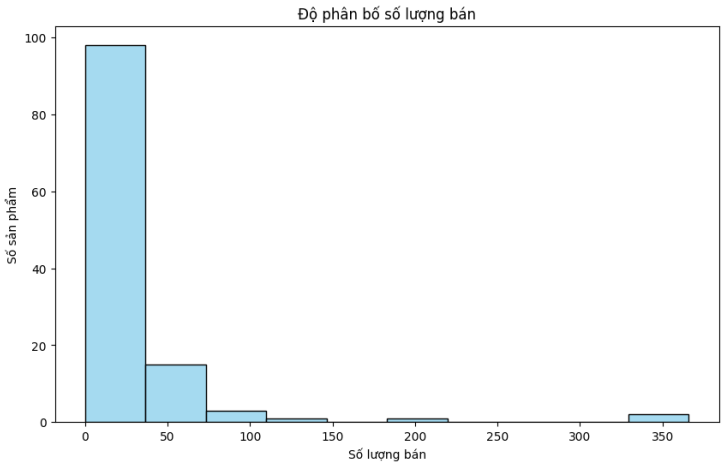
Hình 1.40 Các sản phẩm loại váy/quần trên Tiki

Tương tự như đồ bộ, khoảng 80% sản phẩm váy/quần nằm trong phân khúc giá thấp, trong khi 20% thuộc phân khúc giá cao. Tỷ lệ này cho thấy váy/quần cũng được phân phối chủ yếu ở mức giá phải chăng với một số sản phẩm có giá thành cao hơn. Tuy nhiên trong thực tế chỉ có 1 sản phẩm thuộc phân khúc giá cao (như trong Hình 1.40). Vì số lượng sản phẩm quá ít (5/120 sản phẩm) nên biểu đồ tròn ở trên không phản ánh được ý nghĩa thực tế của loại sản phẩm này.



Hình 1.41 Biểu đồ cột các thông số thống kê số lượng bán sản phẩm trên Tiki

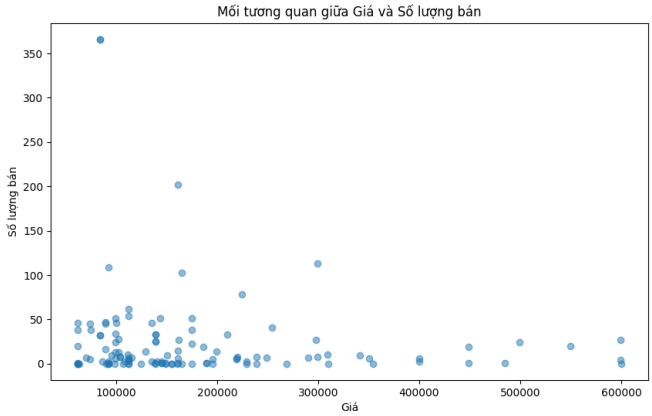
Số lượng bán các sản phẩm trên Tiki tại thời điểm thu thập dữ liệu có sự chênh lệch đáng kể với mức thấp nhất là 0 và cao nhất là 366 sản phẩm.



Hình 1.42 Biểu đồ cột biểu thị độ phân bố số lượng bán sản phẩm trên Tiki

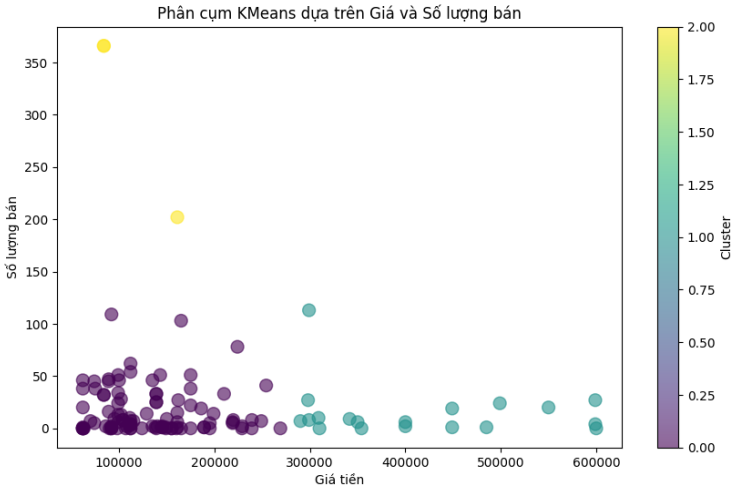
Phần lớn các sản phẩm có số lượng bán rất thấp, chủ yếu dưới 50 đơn vị. Số lượng sản phẩm giảm dần khi số lượng bán tăng, cho thấy sự phân phối không đều. Rất ít sản phẩm có số lượng bán trên 150. Chỉ có một vài sản phẩm bán hơn 300 đơn vị, thể hiện qua các cột rất nhỏ ở phía bên phải của biểu đồ. Nhìn tổng quát, phần lớn sản phẩm

trong tập dữ liệu không bán chạy, trong khi chỉ có một số ít sản phẩm đạt được số lượng bán lớn hơn.



Hình 1.43 Biểu đồ tương quan giữa giá tiền và số lượng bán sản phẩm trên Tiki

Nhìn chung những sản phẩm có số lượng bán cao tập trung ở phân khúc giá thấp và đặc biệt tập trung ở những sản phẩm có mức giá bán xấp xỉ 100.000 VNĐ. Khi giá sản phẩm tăng lên, số lượng bán bắt đầu giảm rõ rệt. Rất ít sản phẩm có số lượng bán đáng kể, chỉ có 3 sản phẩm có số lượng bán trên 150. Kết quả này cho thấy mặt hàng quần áo nữ trên nền tảng Tiki chưa thu hút được phần nhiều khách hàng, khách hàng ở đây có xu hướng mua nhiều hơn các sản phẩm giá thấp, nhất là những sản phẩm quần áo nữ có mức giá xấp xỉ 100.000 VNĐ.



Hình 1.44 Biểu đồ phân cụm Kmeans dựa trên giá tiền và số lượng bán sản phẩm quần áo nữ trên Tiki

Sau khi xác định số cụm là 3 bằng phương pháp Elbow, ta tiến hành phân cụm bằng thuật toán K-means, ta có được 3 cụm tương ứng với 3 màu như Hình 1.44. Đặc điểm từng cụm như sau:

- Cụm 1 (tím): gồm các sản phẩm có mức giá thấp và số lượng bán < 150 sản phẩm. Đây là cụm có số lượng sản phẩm tập trung dày đặc nhất.
- Cụm 2 (xanh): gồm các sản phẩm có mức giá ở phân khúc giá cao (xấp xỉ 300.000 VNĐ trở lên). Tất cả sản phẩm ở cụm này đều có số lượng bán ở mức dưới 150 sản phẩm.
- Cụm 3 (vàng): là cụm có số lượng sản phẩm ít nhất (3 sản phẩm) gồm các sản phẩm có mức giá thấp (xấp xỉ 100.000 VNĐ) và có số lượng bán từ 150 sản phẩm trở lên.

Tương ứng với 3 cụm trên là 3 đối tượng sản phẩm với các chiến lược kinh doanh phù hợp.

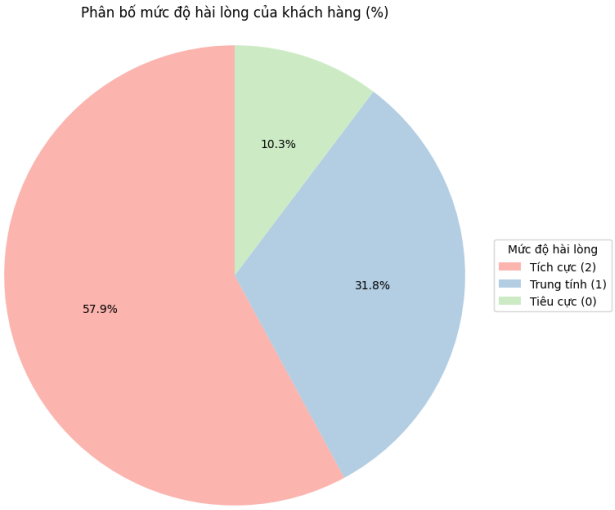
Tóm lại, qua những thông tin được trực quan hóa trên của bộ dữ liệu thông tin sản phẩm quần áo nữ trên nền tảng thương mại điện tử Tiki, ta rút ra được những đặc điểm như sau:

- Ở thời điểm thu thập dữ liệu, trên thị trường thương mại điện tử Tiki, mặt hàng được đăng bán nhiều nhất và chiếm đa số là áo và đồ bộ mà chiếm ưu thế hơn là loại sản phẩm đồ bộ.
- Có 3 đối tượng sản phẩm với đối tượng 1 chiếm ưu thế về mặt số lượng (98 sản phẩm).
- Đối với đối tượng 1 là những sản phẩm có mức giá thấp và số lượng bán < 150, vì số lượng bán còn thấp, việc tối ưu hóa trải nghiệm mua sắm, tăng cường đánh giá tích cực, và cải thiện hình ảnh sản phẩm có thể giúp tăng doanh số. Đối tượng khách hàng mua những sản phẩm này có xu hướng mua sắm sản phẩm giá rẻ, vì thế cần tập trung hơn về các chiến lược quảng cáo và các chương trình khuyến mãi.
- Đối tượng sản phẩm 2 là đối tượng có mức giá cao với số lượng bán tương đối thấp. Khách hàng mua những sản phẩm ở đối tượng này có xu hướng yêu cầu cao hơn về chất lượng tương ứng với giá tiền.
- Cuối cùng, đối tượng 3 gồm các sản phẩm ở mức giá thấp và số lượng bán khá từ 150 sản phẩm trở lên. Vì chỉ có 3 sản phẩm và số lượng bán ra chênh lệch đáng kể với các sản phẩm còn lại. Điều này có thể chỉ ra rằng mặc dù giá cả thấp, nhưng sản phẩm có thể đáp ứng nhu cầu thị trường hoặc có giá trị cao hơn so với

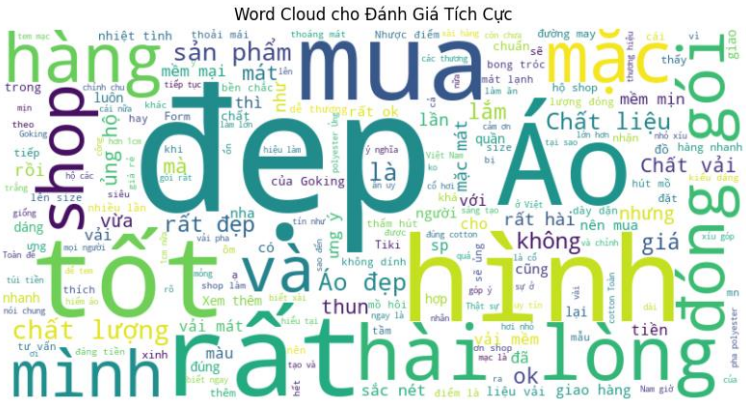
mức giá. Với các sản phẩm thuộc đối tượng này, việc tập trung vào mở rộng đối tượng khách hàng là rất cần thiết, vì thế cần tiếp tục quảng bá và tận dụng các lợi thế về giá.

1.7.4. Các đánh giá sản phẩm của Tiki

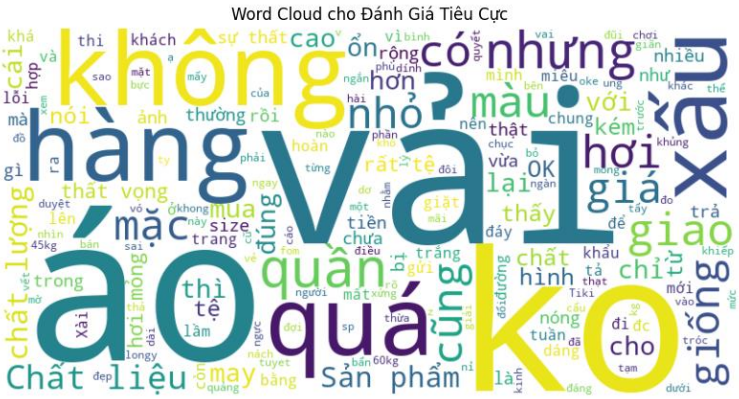
Ở bộ dữ liệu này ta tập trung phân tích những đánh giá của khách hàng với 120 sản phẩm quần áo nữ trên nền tảng Tiki.



Hình 1.45 Biểu đồ phân bố mức độ hài lòng của khách hàng với các sản phẩm trên Tiki



Hình 1.46 Wordcloud các từ khóa đánh giá tích cực các sản phẩm trên Tiki



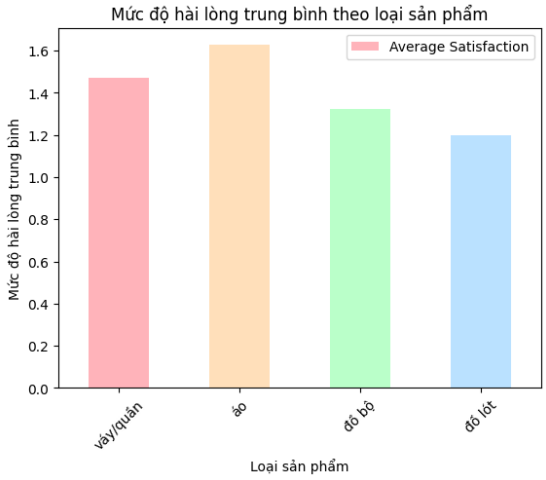
Hình 1.47 Wordcloud các từ khóa đánh giá tiêu cực các sản phẩm trên Tiki

Tương tự các thao tác đánh giá mức độ hài lòng ở bộ dữ liệu trên Lazada, ta có cái nhìn khách quan về mức độ hài lòng của khách hàng trên nền tảng Tiki như sau.

Kết quả Wordcloud ở Hình 1.46 cho các từ khóa đánh giá tích cực nổi bật với các từ như "đẹp", "hài lòng", "đóng gói", … thể hiện sự hài lòng cao của người tiêu dùng với mẫu mã sản phẩm và dịch vụ. Ngược lại, trong Hình 1.47 Wordcloud cho các từ khóa tiêu cực chứa các từ như "vải", "xấu", "chất lượng", … phản ánh những vấn đề mà khách hàng gặp phải mà đa phần là những phàn nàn về chất lượng.

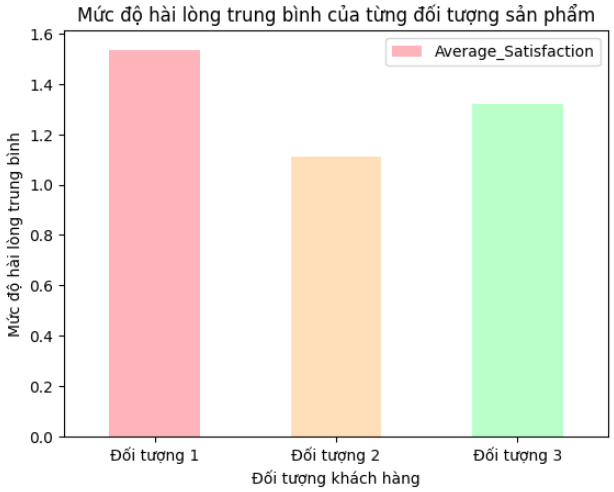
Tỷ lệ của từng mức độ hài lòng được biểu diễn ở Hình 1.45. Trong đó, có 231 (57.9%) đánh giá tích cực, 127 (31.8%) đánh giá trung tính và 41 (10.3%) đánh giá tiêu cực.

Lượng đánh giá tích cực chiếm đa số có nghĩa là phần lớn các sản phẩm đáng ứng tốt nhu cầu mua sắm của khách hàng ở mặt hàng quần áo nữ trên nền tảng này. Bênh cạnh đó vẫn còn một phần nhỏ những đánh giá tiêu cực về chất lượng sản phẩm như chất vải chưa đủ tốt, kiểu dáng chưa đáp ứng được thị hiếu khách hàng, … Kết quả này cho thấy ngoài cân nhắc giá cả, khách hàng còn quan tâm hơn vào chất lượng sản phẩm và trải nghiệm mua hàng.



Hình 1.48 Biểu đồ cột trung bình mức độ hài lòng của khách hàng với các loại sản phẩm trên Tiki

Mức độ hài lòng trung bình của sản phẩm loại áo là 1.63, nghĩa là ở rất gần mức tích cực cho thấy các sản phẩm loại này khá được khách hàng ưa thích. Có thể lý do nằm ở mẫu mã hợp thời và dịch vụ làm hài lòng khách hàng như đã phân tích ở trên. Cao thứ hai là sản phẩm loại váy/quần với mức độ hài lòng trung bình là 1.47 cũng là mức độ tương đối cao (trung tính thiến hướng cực tích). Và với hai loại còn lại là đồ bộ và đồ lót, mức hài lòng là xấp xỉ nhau (lần lượt là 1.32 và 1.2) và chỉ cao hơn mức trung tính một chút. Điều này thể hiện rằng có khách hàng hài lòng và ưa thích các sản phẩm loại này nhưng cũng có khách hàng không vừa ý về yếu tố nào đó của sản phẩm ở mức không đáng kể.



Hình 1.49 Biểu đồ cột trung bình mức độ hài lòng của khách hàng với các đối tượng sản phẩm trên Tiki

Xét trên toàn bộ dữ liệu đánh giá sản phẩm quần áo nữ trên nền tảng Tiki, mức độ hài lòng trung bình là 1.48 nghĩa là đang ở mức trung tính có thiên hướng tích cực. Mức độ hài lòng trung bình của khách hàng ở mỗi đối tượng sản phẩm 1, 2, 3 lần lượt là 1.54, 1.11 và 1.32.

Nhìn chung, mức độ hài lòng của khách hàng là trung bình. Có nghĩa là chất lượng sản phẩm và trải nghiệm mua sắm của khách hàng không đủ tốt nhưng cũng không quá tệ. Tuy nhiên, đối tượng sản phẩm 2 và 3 có mức độ hài lòng còn chưa cao, chỉ ở gần mức trung tính. Cần xem xét và khắc phục những vấn đề của 2 đối tượng sản phẩm này. Trái lại mức độ hài lòng trung bình của sản phẩm đối tượng 1 nằm ở mức trung tính – tích cực, cho thấy nhiều sản phẩm được đánh giá khá tốt.

Hình 1.50 Biểu đồ tròn phân bố mức độ hài lòng ở mỗi đối tượng sản phẩm trên Tiki

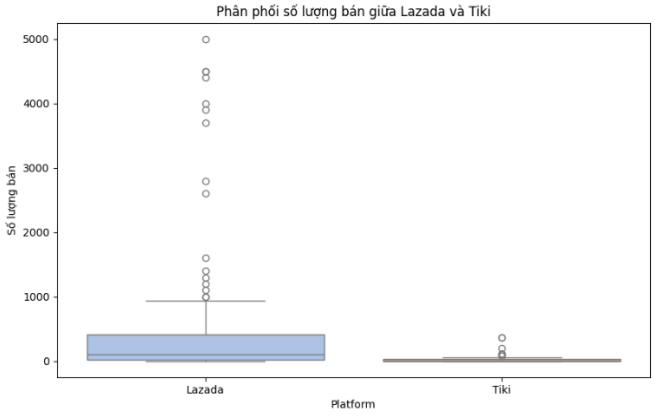
Đối với đối tượng 1, đây là đối tượng sản phẩm có nhiều đánh giá tích cực nhất, cho thấy sản phẩm này khá thành công trong việc làm hài lòng khách hàng. Số lượng đánh giá tích cực (195) chiếm phần lớn, phản ánh sản phẩm này có chất lượng tốt hoặc dịch vụ chăm sóc khách hàng hiệu quả. Đây cũng là những sản phẩm có giá tiền khá thấp và số lượng bán dưới 150 sản phẩm. Có thể thấy tuy số lượng bán không cao nhưng những phản hồi từ khách hàng khá khả quan. Đối tượng này có tiềm năng khá lớn để mở rộng đối tượng khách hàng và tăng doanh thu.

Đối tượng 2 là những sản phẩm ở mức giá khá cao và số lượng bán ra không nhiều đang có số lượng đánh giá khá thấp, nhưng sự phân bổ giữa các mức độ hài lòng tương đối cân bằng. Điều này có thể cho thấy sản phẩm chưa nổi bật trên thị trường, nhưng cũng không gặp quá nhiều phàn nàn từ khách hàng.

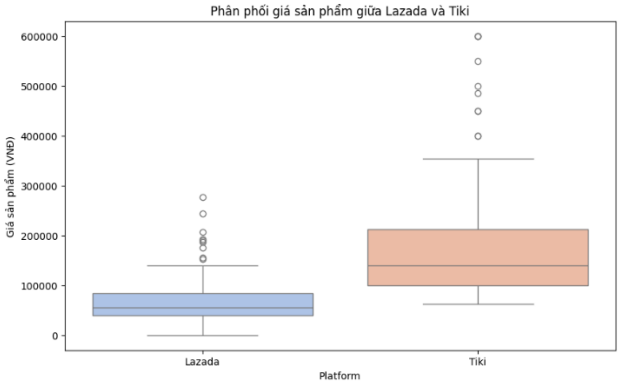
Cuối cùng, đối tượng sản phẩm 3 có số lượng đánh giá trung tính cao hơn một chút so với đánh giá tích cực. Điều này cho thấy khách hàng có xu hướng cảm thấy hài lòng ở mức độ trung bình, có thể sản phẩm vẫn còn tiềm năng để cải thiện nhằm gia tăng sự hài lòng.

1.7.5. Trực quan hóa toàn dữ liệu

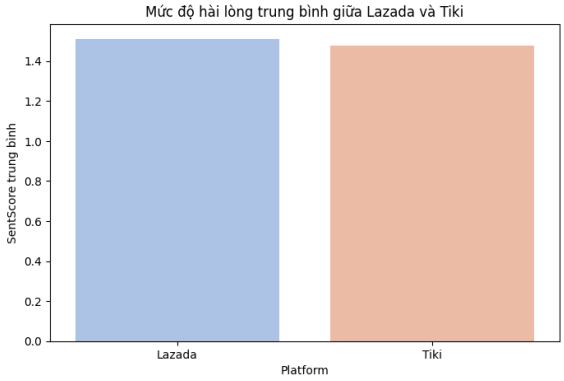
Từ những thông tin được hình ảnh hóa và phân tích ở trên, ta có cái nhìn chung về các sản phẩm và đánh giá của sản phẩm quần áo nữ trên nền tảng thương mại điện tử như sau:



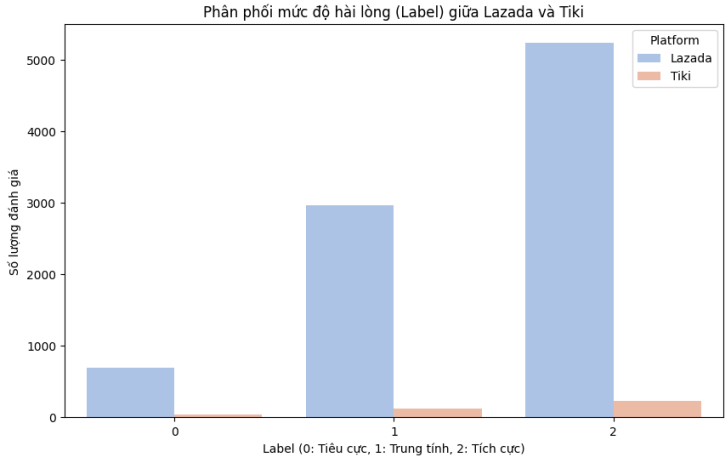
Hình 1.51 Biểu đồ so sánh phân phối số lượng bán giữa Lazada và Tiki



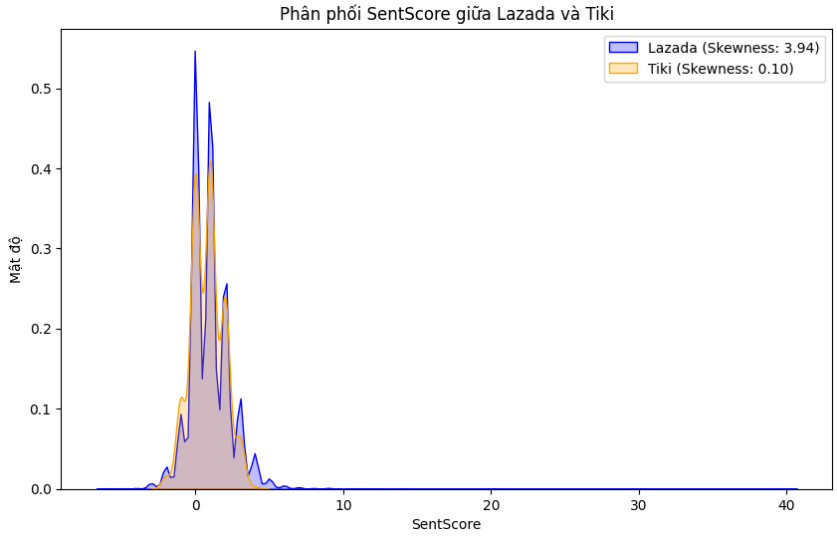
Hình 1.52 Biểu đồ so sánh phân phối giá sản phẩm giữa Lazada và Tiki



Hình 1.53 Biểu đồ so sánh mức độ hài lòng trung bình giữa Lazada và Tiki



Hình 1.54 Biểu đồ so sánh phân phối mức độ hài lòng giữa Lazada và Tiki



Hình 1.55 Biểu đồ so sánh phân phối SentScore giữa Lazada và Tiki

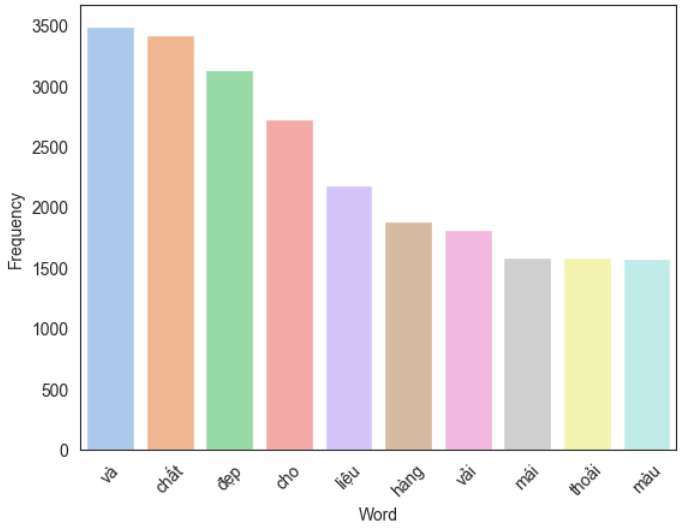
Số lượng bán cao nhất trên nền tảng Tiki tương đối thấp so với Lazada, giá trị ngoại lai cũng ít hơn. Hình 1.51 biểu diễn rõ ràng sự chênh lệch về số lượng bán giữa hai nền tảng. Ý nghĩa của kết quả này cho thấy các sản phẩm được đăng bán trên Lazada

tiếp cận được với số lượng khách hàng đông hơn và đáp ứng nhu cầu nhiều khách hàng hơn.

Trái lại, mức giá sản phẩm ở Tiki nhìn chung cao hơn Lazada. Độ chênh lệch về mức giá thấp nhất và cao nhất của hai nền tảng này cũng lớn. Có thể thấy sản phẩm quần áo nữ trên Tiki có khoảng giá rộng hơn, đa dạng về giá hơn (Hình 1.52).

Nhìn Hình 1.53, ta dễ dàng nhận thấy mức độ hài lòng trung bình của khách hàng là xấp xỉ nhau trên cả 2 nền tảng. Múc độ hài lòng trung bình là 1.5, nghĩa là ở mức trung tính thiên hướng tích cực. Đây là một tín hiệu tốt cho việc tiếp cận gần hơn nữa đến nhu cầu từng khách hàng. Dù mức độ hài lòng là ngang nhau nhưng chênh lệch số lượng bán còn quá lớn. Cho thấy độ nhận diện của nền tảng Tiki còn yếu và có lẽ vẫn có những vấn đề đến từ giá sản phẩm.

Hình 1.56 Wordcloud các từ khóa đánh giá sản phẩm



Hình 1.57 Biểu đồ thống kê 10 từ khóa có tần suất xuất hiện cao nhất trong đánh giá sản phẩm

Ta có số lần xuất hiện của 10 từ khóa phổ biến nhất trong đánh giá sản phẩm ở cả 2 nền tảng như sau:

Từ khóa	**Số lần xuất hiện**
‘và’	3492
‘chất’	3426
‘đẹp’	3132
‘cho’	2730
‘liệu’	2182
‘hàng’	1882
‘vải’	1814
‘mái’	1591
‘thoải’	1583
‘màu’	1576

Bảng 1.1 Bảng thống kê 10 từ khóa có tần suất xuất hiện cao nhất trong đánh giá sản phẩm

Hình 1.57 đã thể hiện khá rõ ràng những đặc trưng trong nhận xét khách hàng với mặt hàng quần áo nữ. Trong đó, có 3 đặc điểm khách hàng quan tấm nhất là chất liệu, thiết kế và cảm giác khi sử dụng.

Các từ nối ‘và’, ‘cho’ tuy xuất hiện nhiều nhưng vì chỉ là những từ nối thông dụng nên không mang ý nghĩa đặc biệt trong phân tích này.

Các từ khóa ‘chất’, ‘liệu’, ‘vải’ có tần suất xuất hiện khá cao (lần lượt là 3426, 2182, 1814 như trong Bảng 1.1**Error! Reference source not found.**) cho thấy sự quan tâm của khách hàng về chất liệu sản phẩm, nhất là sản phẩm quần áo nữ là khá cao.

‘đẹp’ và ‘màu’ là những từ khóa liên quan đến thiết kế của sản phẩm. Đối với phái nữ, đây là yếu tố khá được cân nhắc. So với việc mua sắm ở các cửa hàng trực tiếp, khách hàng lựa chọn mua sắm trực tuyến trên các nền tảng thương mại điện tử đa phần là vì sự đa dạng trong kiểu dáng và thiết kế của sản phẩm. Hình 1.56 cũng cho thấy thời trang, thiết kế, màu sắc đều chiếm sự quan tâm đáng kể của khách hàng, mà ở đây đối tượng sử dụng sản phẩm là nữ giới.

Tương tự, ‘mái’ (1591) và ‘thoải’ (1583) đều ám chỉ cảm giác khi sử dụng sản phẩm của khách hàng. Tuy ít hơn hai yếu tố trên nhưng đây cũng là yếu tố chiếm nhiều sự quan tâm của khách hàng.

Chương 2. XÂY DỰNG HỆ THỐNG GỢI Ý SẢN PHẨM

2.1. Tổng quan giải pháp xây dựng hệ thống

Qua những phân tích sản phẩm và xu hướng mua hàng, ta có cái nhìn tổng quát về thị trường quần áo nữ trên các sàn mua sắm online. Trong phạm vi đề tài này, hệ thống gợi ý sẽ được phát triển thông qua việc xây dựng mô hình dựa trên thuật toán gợi ý theo nội dung (Content-based Filtering). Mục tiêu là đào tạo mô hình sao cho đưa ra gợi ý sản phẩm phù hợp dựa trên các yếu tố từ khóa khách hàng đưa ra, sản phẩm khách hàng đã mua và xu hướng thời trang hiện tại. Và để tối ưu hóa hiệu suất hệ thống, những phản hồi của người dùng trong thực tế sẽ được cập nhật và điều chỉnh liên tục.

2.2. Xác định các tác nhân

Hệ thống gợi ý sản phẩm gồm có hai tác nhân chính là khách hàng tìm kiếm và mua hàng trên các sàn thương mại điện tử (Customers) và các cửa hàng bán lẻ đang đăng bán sản phẩm trên đó (Shops). Trong đó, đề tài xây dựng các chức năng tập trung vào hỗ trợ khách hàng trong ra gợi ý sản phẩm trong cả quá trình tìm kiếm và mua sắm.

2.3. Xác định các kho dữ liệu

[1] Sản phẩm (Products): thông tin các sản phẩm được đăng bán trên sàn thương mại điện tử, ở đề tài này là mặt hàng quần áo nữ. Gồm tên, giá bán, số lượng đã bán, loại sản phẩm và số lượng hàng tồn.

[2] Đánh giá (Feedbacks): những phản hồi của khách hàng về sản phẩm đã mua. Bao gồm nội dung đánh giá, điểm hài lòng và mức độ hài lòng được tính dựa trên nội dung.

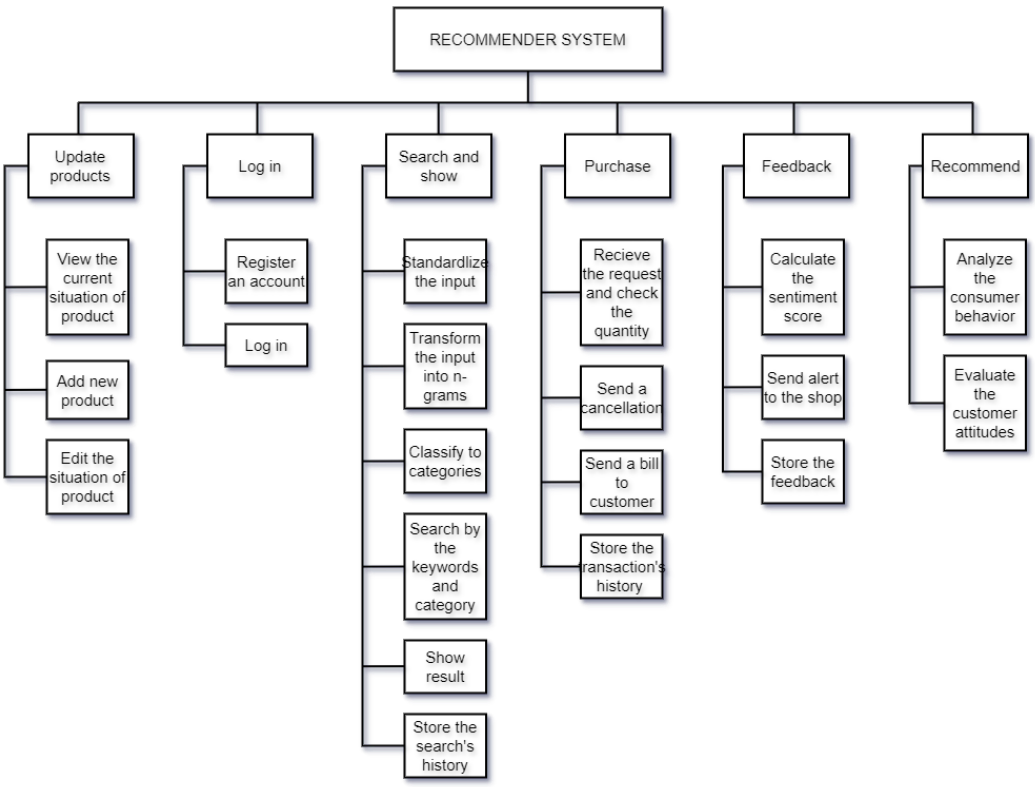
[3] Lịch sử tìm kiếm (History Of Searches): nội dung tìm kiếm sản phẩm của khách hàng.

[4] Lịch sử giao dịch (History Of Transactions): thông tin về giao dịch mua hàng của khách hàng.

[5] Khách hàng (Customers): thông tin khách hàng đã đăng ký tài khoản.

2.4. Tiến trình hệ thống

2.4.1. Phân rã chức năng



Hình 2.1 Sơ đồ phân rã chức năng

2.4.2. Mô tả chi tiết các chức năng lá

(1.1) Xem tình trạng sản phẩm (View the current situation of product): hệ thống sẽ truy cập vào kho dữ liệu, cho phép các cửa hàng theo dõi tình trạng sản phẩm.

(1.2) Thêm sản phẩm (Add new product): hệ thống nhận yêu cầu thêm sản phẩm từ cửa hàng, tiến thành thêm sản phẩm mới và thông tin sản phẩm vào kho lưu trữ dữ liệu.

(1.3) Thay đổi tình trạng sản phẩm (Edit the situation of product): hệ thống khi nhận yêu cầu chỉnh sửa từ của hàng sẽ chỉnh sửa thông tin, tình trạng sản phẩm trên kho dữ liệu.

(2.1) Đăng ký tài khoản (Register an account): hệ thống tiếp nhận thông tin của khách hàng và tạo tài khoản.

(2.2) Đăng nhập (Log in): với khách hàng đã đăng ký tài khoản, hệ thống tiếp nhận tên đăng nhập và mật khẩu để đăng nhập.

(3.1) Chuẩn hóa (Standardlize the input): làm sạch và chuẩn hóa dữ liệu tìm kiếm của khách hàng.

(3.2) Chuyển đổi thành dạng n-grams (Transform the input into n-grams): dùng các công cụ NLP để biến đổi dữ liệu đã chuẩn hóa thành dạng n-grams.

(3.3) Phân loại sản phẩm (Classify to categories): phân loại sản phẩm theo các từ khóa tương tự như Hình 1.12.

(3.4) Tìm kiếm bằng từ khóa và loại sản phẩm (Search by the keywords and category):

(3.5) Hiển thị (Show result): hiển thị ra gia diện cho người dùng.

(3.6) Lưu trữ lịch sử tìm kiếm (Store the search's history): hệ thống cập nhật dữ liệu khách hàng vừa tìm kiếm vào kho lưu trữ.

(4.1) Tiếp nhận đơn hàng và kiểm tra tình trạng sản phẩm (Recieve the request and check the quantity): kiểm tra tình trạng sản phẩm trong kho.

(4.2) Từ chối đơn hàng (Send a cancellation): trong trường hợp tình trạng sản phẩm không đáp ứng được đơn hàng của khách, hệ thống sẽ gửi thông báo.

(4.3) Gửi hóa đơn mua hàng (Send a bill to customer): trường hợp tình trạng sản phẩm đáp ứng đủ đơn hàng của khách, sau khi nhận tiền thanh toán hệ thống sẽ trả về khách hàng hóa đơn.

(4.4) Lưu trữ lịch sử mua hàng (Store the transaction's history): hệ thống cập nhật thông tin đơn hàng vào kho lưu trữ.

(5.1) Tính điểm hài lòng (Calculate the sentiment score): khi hệ thống nhận đánh giá sản phẩm từ khách hàng sẽ tính điểm hài lòng dựa trên nội dung đánh giá,

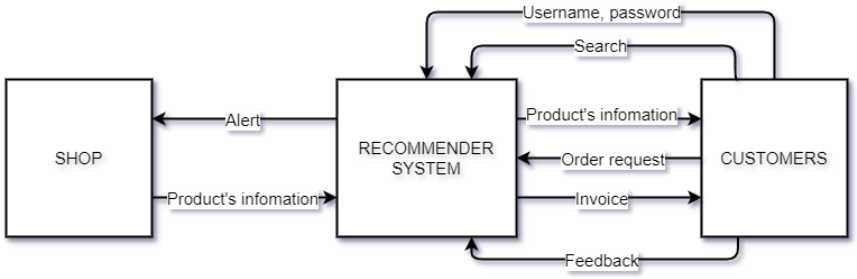
(5.2) Gửi thông báo cho cửa hàng (Send alert to the shop): các phản hồi từ khách hàng sẽ được gửi về cửa hàng để kịp thời hỗ trợ giải quyết nếu có tình huống bất trắc và để cửa hàng dễ dàng theo dõi mức độ hài lòng về các sản phẩm mình đăng bán.

(5.3) Lưu trữ các đánh giá (Store the feedback): hệ thống cập nhật dữ liệu về phản hồi của khách hàng về sản phẩm vào kho lưu trữ.

(6.1) Phân tích hành vi khách hàng (Analyze the consumer behavior): xây dựng các mô hình đánh giá xu hướng mua hàng của khách hàng theo thời gian thực.

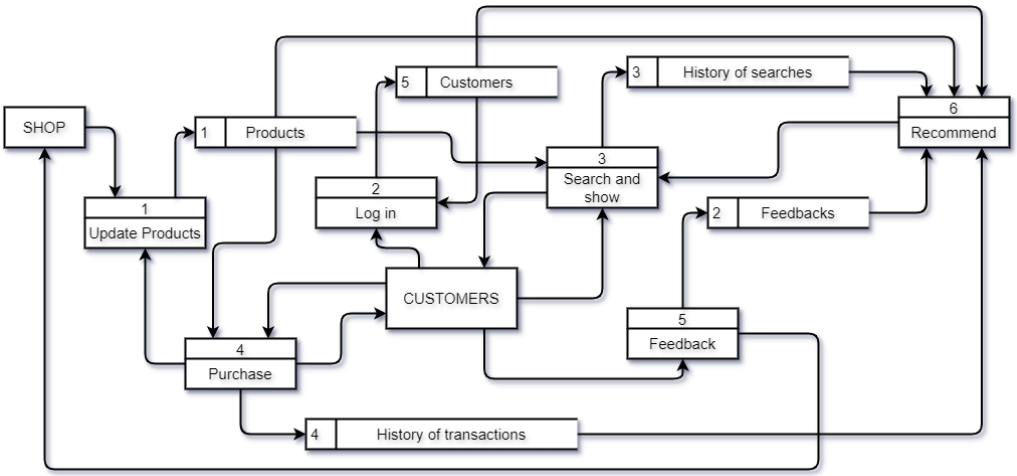
(6.2) Đánh giá thị hiếu khách hàng (Evaluate the customer attitudes): sử dụng các công cụ NLP phân tích cảm xúc, mức độ hài lòng của khách hàng với từng loại sản phẩm và đánh giá tỷ lệ mua lại của sản phẩm để đưa ra gợi ý phù hợp nhất.

2.4.3. Biểu đồ mức ngữ cảnh



Hình 2.2 Biểu đồ mức ngữ cảnh

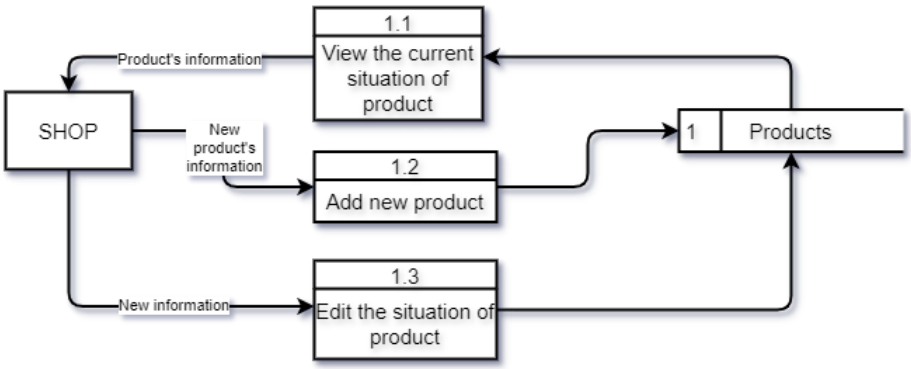
2.4.4. Sơ đồ DFD mức đỉnh



Hình 2.3 Sơ đồ DFD mức đỉnh

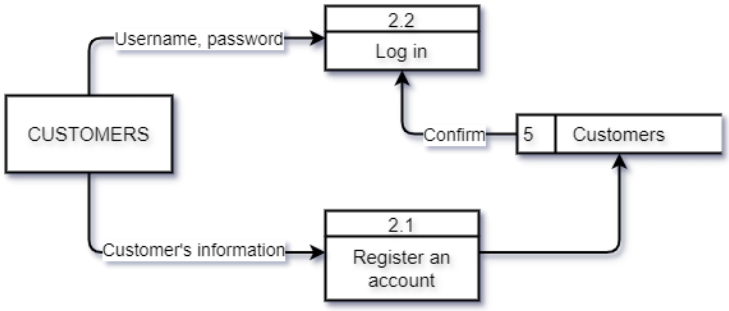
2.4.5. Sơ đồ DFD mức 1

2.4.5.1. Tiến trình 1:



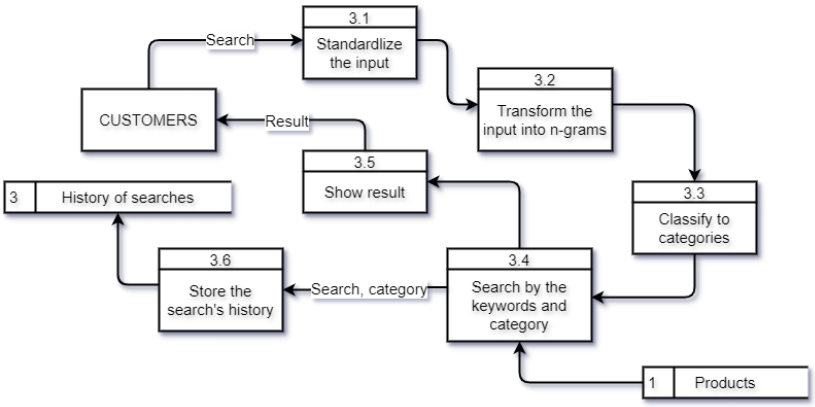
Hình 2.4 Sơ đồ DFD mức 1 tiến trình 1

2.4.5.2. Tiến trình 2:



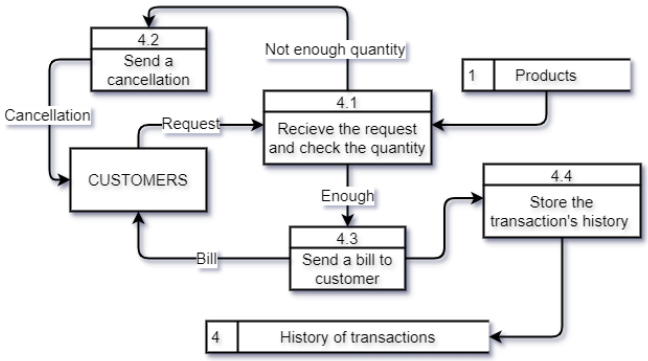
Hình 2.5 Sơ đồ DFD mức 1 tiến trình 2

2.4.5.3. Tiến trình 3:



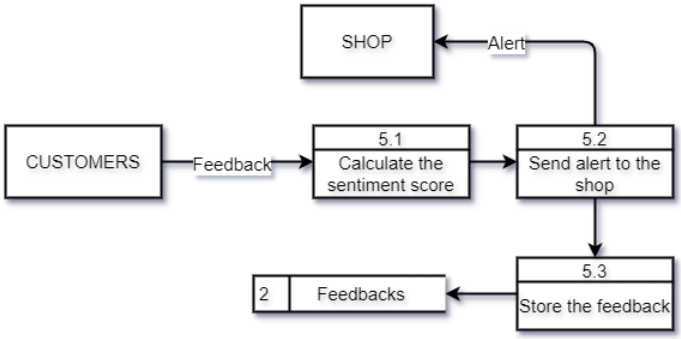
Hình 2.6 Sơ đồ DFD mức 1 tiến trình 3

2.4.5.4. Tiến trình 4:



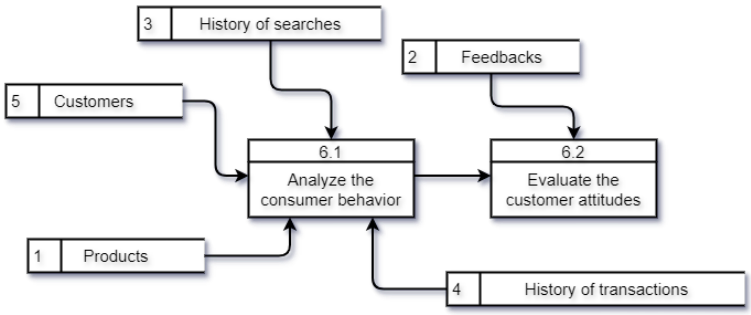
Hình 2.7 Sơ đồ DFD mức 1 tiến trình 4

2.4.5.5. Tiến trình 5:



Hình 2.8 Sơ đồ DFD mức 1 tiến trình 5

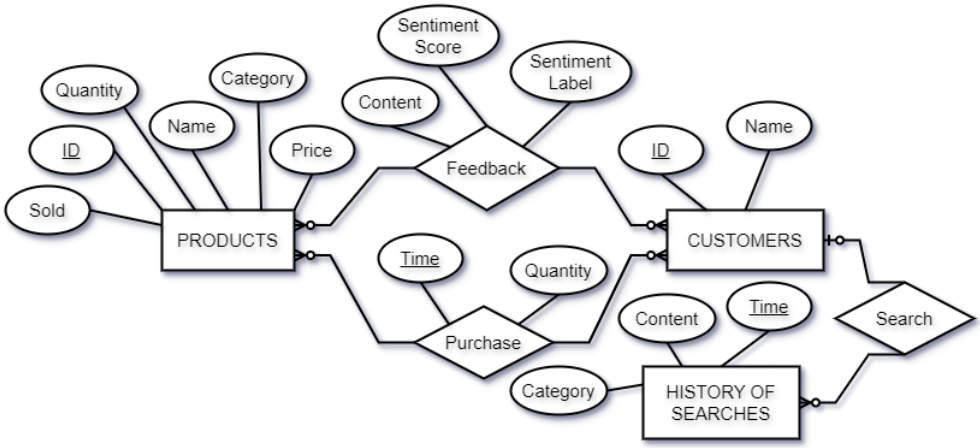
2.4.5.6. Tiến trình 6:



Hình 2.9 Sơ đồ DFD mức 1 tiến trình 6

2.5. Thiết kế cơ sở dữ liệu

2.5.1. Mô hình ERD



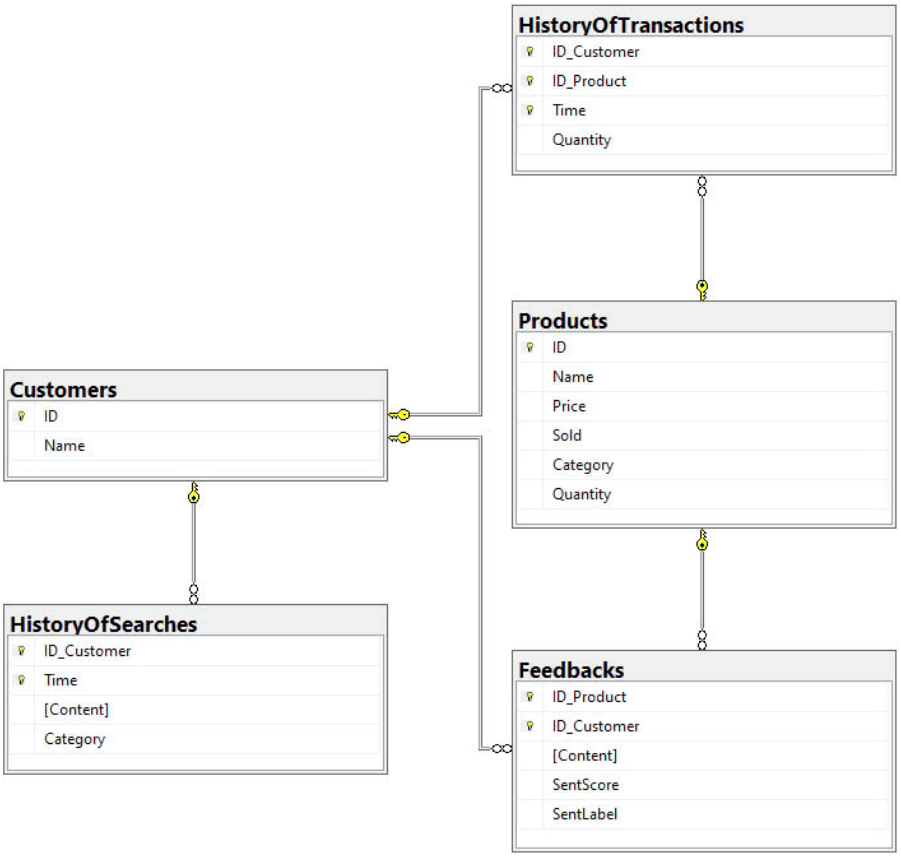
Hình 2.10 Mô hình ERD

2.5.2. Thiết kế cơ sở dữ liệu vật lý

- Products (ID, Name, Price, Sold, Category, Quantity)
- Feedbacks (#ID_Product, Content, SentScore, SentLabel, #ID_Customer)
- HistoryOfTransactions (#ID_Customer, #ID_Product, Time, Quantity)

- HistoryOfSearches (#ID_Customer, Time, Content, Category)
- Customers (ID, Name)

2.5.3. Mô hình RDM



Hình 2.11 Mô hình RDM cơ sở dữ liệu RCM_System

Chương 3. TỔNG KẾT VÀ ĐÁNH GIÁ

3.1. Ưu điểm

Quá trình thu thập và làm sạch dữ liệu tương đối thuận lợi, ít trường hợp dữ liệu trùng lắp và trống, lượng dữ liệu bị hao hụt thấp. Trong quá trình phân tích, tuy các số liệu ở hai nền tảng có sự chênh lệch nhưng vẫn có nhiều nét tương đồng phản ánh xu hướng chung hỗ trợ đánh giá và xác định tình hình chung trong việc mua sắm trực tuyến ở thị trường Việt Nam.

3.2. Hạn chế

Bên cạnh đó, vẫn còn những trở ngại khiến đề tài chưa thể đạt được trạng thái tốt nhất. Đầu tiên là sự đa dạng về dữ liệu đầu vào bị hạn chế. Thị trường thương mại điện tử mà sản phẩm cụ thể quần áo nữ ở Việt Nam chưa thật sự đa dạng, chỉ tập trung ở vài

nền tảng quen thuộc. Trong đó, ngoài hai nền tảng đã thu thập được dữ liệu như trên, những nền tảng còn lại đều có những hạn chế về kinh phí và thời gian thực hiện đồ án. Hơn nữa, quá trình phân tích dữ liệu vẫn còn những công đoạn thủ công do sự phức tạp trong xử lý ngôn ngữ tự nhiên, mà ở đây là tiếng Việt. Việc chưa xử lý được hết những trường hợp liên quan đến tính đặc thù của ngôn ngữ tiếng Việt gây ra những sai số trong kết quả đánh giá của phân tích trên.

3.3. Nhận xét tổng quát

Nhìn chung, đề tài đã đạt được những kết quả nhất định trong việc phân tích hành vi mua sắm trực tuyến của khách hàng Việt Nam đối với sản phẩm quần áo nữ. Tuy nhiên, vẫn còn một số hạn chế cần được khắc phục để nâng cao chất lượng hệ thống. Để hệ thống hoàn thiện hơn, có các hướng có thể phát triển như sử dụng công cụ phù hợp hơn để xử lý ngôn ngữ tiếng Việt giúp nâng cao tính chính xác của phân tích; giảm thiểu sự thủ công, tăng tính tự động hóa trong xử lý và phân tích dữ liệu bằng cách lựa chọn các mô hình huấn luyện hoặc các công cụ hiện đại giúp tăng hiệu quả. Việc đầu tư vào các công cụ và kỹ thuật phân tích dữ liệu sẽ giúp nâng cao chất lượng và độ tin cậy của hệ thống.

TÀI LIỆU THAM KHẢO

[1] Selenium, “About,” [Trực tuyến]. Available: https://www.selenium.dev/about/. [Đã truy cập 31 8 2024].

[2] NLTK Team, “NLTK Documentation,” [Trực tuyến]. Available: https://www.nltk.org/. [Đã truy cập 31 8 2024].

[3] K. J. M. S. J. V. D. W. R. G. P. V. D. C. C. R. Harris, “Array programming with NumPy,” *Nature*, tập 585, số 7825, p. 357–362, 2020.

[4] H. Stepanek, “Thinking in Pandas,” trong *About pandas*, Portland, OR, USA, Apress, 2020, pp. 1-8.

[5] S. Tosi, Matplotlib for Python Developers, UK: Packt Publishing Ltd, 2009.

[6] M. Waskom, “seaborn: statistical data visualization,” 2021. [Trực tuyến]. Available: https://seaborn.pydata.org/. [Đã truy cập 1 9 2024].

[7] G. V. A. G. V. M. B. T. O. G. F. Pedregosa, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, tập 12, p. 2825–2830, 2011.

[8] R. Řehůřek, “Gensim,” 2024. [Trực tuyến]. Available: https://radimrehurek.com/gensim/intro.html#what-is-gensim. [Đã truy cập 3 9 2024].

[9] Streamlit, “Streamlit,” Snowflake Inc, [Trực tuyến]. Available: https://streamlit.io/. [Đã truy cập 3 9 2024].

[10] Wikipedia, “Wikipedia,” [Trực tuyến]. Available: https://en.wikipedia.org/wiki/Lazada. [Đã truy cập 4 9 2024].

[11] P. V. Le-Hoang, “Factors affecting online purchase intention: the case of e-commerce on Lazada,” *Independent Journal of Management & Production*, tập 11, số 3, p. 1018–1033, 2020.

[12] Wikipedia, “Wikipedia,” [Trực tuyến]. Available: https://vi.wikipedia.org/wiki/Tiki. [Đã truy cập 4 9 2024].