|  |  |
| --- | --- |
|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HỒ CHÍ MINH** |

**ĐỒ ÁN TỐT NGHIỆP**

**XÂY DỰNG MÔ HÌNH LỌC ĐÁNH GIÁ NGƯỜI DÙNG NHẰM CẢI THIỆN CHẤT LƯỢNG CỦA HỆ THỐNG ĐỀ XUẤT**

Ngành: **CÔNG NGHỆ THÔNG TIN**

Chuyên ngành: **CÔNG NGHỆ PHẦN MỀM**

Giảng viên hướng dẫn : PGS.TS Võ Đình Bảy

Sinh viên thực hiện : Trương Văn Đạt

MSSV: 2180608812 Lớp: 21DTHD6

TP. Hồ Chí Minh, 2025

|  |  |
| --- | --- |
|  | BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HỒ CHÍ MINH** |

**ĐỒ ÁN TỐT NGHIỆP**

**XÂY DỰNG MÔ HÌNH LỌC ĐÁNH GIÁ NGƯỜI DÙNG NHẰM CẢI THIỆN CHẤT LƯỢNG CỦA HỆ THỐNG ĐỀ XUẤT**

Ngành: **CÔNG NGHỆ THÔNG TIN**

Chuyên ngành: **CÔNG NGHỆ PHẦN MỀM**

Giảng viên hướng dẫn : PGS.TS Võ Đình Bảy

Sinh viên thực hiện : Trương Văn Đạt

MSSV: 2180608812 Lớp: 21DTHD6

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn** | **Giảng viên phản biện** | **Chủ tịch hội đồng** |

TP. Hồ Chí Minh, 2025

**LỜI CAM ĐOAN**

Tôi xin cam đoan đồ án tốt nghiệp với đề tài “Xây dựng mô hình lọc đánh giá người dùng nhằm cải thiện chất lượng của hệ thống đề xuất” là kết quả nghiên cứu và làm việc nghiêm túc của cá nhân tôi dưới sự hướng dẫn của giảng viên hướng dẫn.

Toàn bộ nội dung, số liệu, kết quả trình bày trong báo cáo này đều là do tôi trực tiếp thực hiện hoặc trích dẫn từ các nguồn tài liệu đáng tin cậy, có ghi rõ nguồn gốc. Tôi hoàn toàn chịu trách nhiệm trước Hội đồng chấm đồ án và Nhà trường về tính trung thực và nguyên bản của đồ án.

Tôi xin cam kết đồ án này chưa từng được sử dụng để bảo vệ tại bất kỳ nơi nào khác và không vi phạm quy định về đạo văn của Trường Đại học Công nghệ TP. Hồ Chí Minh (HUTECH). Nếu có bất kỳ sai phạm nào liên quan đến bản quyền, tôi xin hoàn toàn chịu trách nhiệm.

**MỤC LỤC**

[CHƯƠNG 1: TỔNG QUAN 7](#_Toc204202178)

[1.1. Giới thiệu đề tài 7](#_Toc204202179)

[1.2. Tổng quan nghiên cứu liên quan 7](#_Toc204202180)

[1.3. Nhiệm vụ của đồ án 8](#_Toc204202181)

[1.4. Cấu trúc của đồ án 9](#_Toc204202182)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ 10](#_Toc204202183)

[2.1. Tổng quan về hệ thống đề xuất 10](#_Toc204202184)

[2.1.1. Hệ thống lọc cộng tác (Collaborative Filtering) 10](#_Toc204202185)

[2.1.2. Hệ thống lọc theo nội dung (Content-Based Filtering) 10](#_Toc204202186)

[2.1.3. Hệ thống lai ghép (Hybrid Recommendation Systems) 11](#_Toc204202187)

[2.2. Kỹ thuật phân rã ma trận (Matrix Factorization) 11](#_Toc204202188)

[2.2.1. Khái niệm tổng quan 11](#_Toc204202189)

[2.2.2. Mô hình toán học 11](#_Toc204202190)

[2.2.3. Thuật toán huấn luyện 12](#_Toc204202191)

[2.2.4. Ưu điểm của Matrix Factorization 12](#_Toc204202192)

[2.2.5. Hạn chế của Matrix Factorization 13](#_Toc204202193)

[2.2.6. Ứng dụng thực tế 13](#_Toc204202194)

[2.3. Phân tích cảm xúc trong đánh giá người dùng 13](#_Toc204202195)

[2.3.1. Mục tiêu của phân tích cảm xúc 13](#_Toc204202196)

[2.3.2. Các mức độ phân tích cảm xúc 13](#_Toc204202197)

[2.3.3. Các kỹ thuật phân tích cảm xúc 14](#_Toc204202198)

[2.3.4. Áp dụng trong hệ thống đề xuất 14](#_Toc204202199)

[2.4. Các mô hình phân tích cảm xúc từ văn bản 15](#_Toc204202200)

[2.4.1. Tổng quan bài toán phân tích cảm xúc 15](#_Toc204202201)

[2.4.2. Mô hình phân loại truyền thống 16](#_Toc204202202)

[2.4.3. Mô hình học sâu trong phân tích cảm xúc: PhoBERT 18](#_Toc204202203)

[2.4.4. Ứng dụng trong đề tài 18](#_Toc204202204)

[2.5. Thư viện và công nghệ sử dụng 18](#_Toc204202205)

[2.5.1. Các thư viện xử lý và phân tích dữ liệu 18](#_Toc204202206)

[2.5.2. Thư viện học máy và phân tích cảm xúc 19](#_Toc204202207)

[2.5.3. Các công nghệ triển khai hệ thống đề xuất 19](#_Toc204202208)

[2.5.4. Quy trình tích hợp hệ thống 20](#_Toc204202209)

[CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG 21](#_Toc204202210)

[3.1. Thu thập dữ liệu từ nền tảng thương mại điện tử Tiki.vn 21](#_Toc204202211)

[**3.1.1.** Mục tiêu thu thập dữ liệu 21](#_Toc204202212)

[**3.1.2.** Danh mục dữ liệu mục tiêu 21](#_Toc204202213)

[**3.1.3.** Phương pháp thu thập dữ liệu 21](#_Toc204202214)

[**3.1.4.** Tổng quan kết quả thu thập 22](#_Toc204202215)

[**3.1.5.** Mục tiêu tiền xử lý 22](#_Toc204202216)

[**3.1.6.** Các bước xử lý văn bản tiếng Việt 23](#_Toc204202217)

[**3.1.7.** Triển khai luật cụm từ (phrase rules) 23](#_Toc204202218)

[**3.1.8.** Lọc người dùng có ít đánh giá 24](#_Toc204202219)

[3.2. Thiết kế pipeline xử lý 24](#_Toc204202220)

[**3.2.1.** Gán nhãn cảm xúc cho dữ liệu 24](#_Toc204202221)

[**3.2.2.** Biểu diễn đặc trưng văn bản 25](#_Toc204202222)

[**3.2.3.** Huấn luyện mô hình phân tích cảm xúc 25](#_Toc204202223)

[**3.2.4.** Phát hiện đánh giá không hợp lệ 26](#_Toc204202224)

[**3.2.5.** Tổng quan pipeline xử lý 26](#_Toc204202225)

[3.3. Mô tả kiến trúc hệ thống: sơ đồ tổng quát các khối chức năng 27](#_Toc204202226)

[**3.3.1.** Tổng quan hệ thống 27](#_Toc204202227)

[**3.3.2.** Luồng dữ liệu và tương tác 28](#_Toc204202228)

[**3.3.3.** Khả năng mở rộng và tích hợp 29](#_Toc204202229)

[3.4. Các yếu tố thiết kế quan trọng 29](#_Toc204202230)

[**3.4.1.** Dữ liệu đầu vào và đầu ra 29](#_Toc204202231)

[**3.4.2.** Mô-đun huấn luyện và dự đoán 29](#_Toc204202232)

[**3.4.3.** Cơ chế lưu trữ và quản lý kết quả 30](#_Toc204202233)

[**3.4.4.** Tính mở rộng và bảo trì 30](#_Toc204202234)

[CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 32](#_Toc204202235)

[4.1. Thực nghiệm mô hình phân tích cảm xúc 32](#_Toc204202236)

[**4.1.1.** Tập dữ liệu và tiền xử lý 32](#_Toc204202237)

[**4.1.2.** Huấn luyện mô hình và đánh giá 34](#_Toc204202238)

[**4.1.3.** Phân tích kết quả 35](#_Toc204202239)

[**4.1.4.** Minh hoạ kết quả dự đoán 36](#_Toc204202240)

[**4.1.5.** Kết luận phần thực nghiệm cảm xúc 37](#_Toc204202241)

[4.2. Thực nghiệm hệ thống đề xuất 37](#_Toc204202242)

[**4.2.1.** Thiết lập thực nghiệm 37](#_Toc204202243)

[**4.2.2.** Các kịch bản thực nghiệm 37](#_Toc204202244)

[**4.2.3.** Kết quả thực nghiệm 38](#_Toc204202245)

[**4.2.4.** Phân tích và nhận xét. 38](#_Toc204202246)

[**4.2.5.** Kết luận phần thực nghiệm đề xuất 38](#_Toc204202247)

[4.3. Giao diện website demo 38](#_Toc204202248)

[**4.3.1.** Trang chủ và tìm kiếm sản phẩm 38](#_Toc204202249)

[**4.3.2.** Trang danh mục sản phẩm 39](#_Toc204202250)

[**4.3.3.** Trang chi tiết sản phẩm và đánh giá người dùng 40](#_Toc204202251)

[**4.3.4.** Trang gợi ý sản phẩm cá nhân hóa 40](#_Toc204202252)

[**4.3.5.** Trang quản trị và thống kê hệ thống 41](#_Toc204202253)

[**4.3.6.** Phân tích đánh giá và lọc nội dung bất thường 42](#_Toc204202254)

[**4.3.7.** Tổng kết phần giao diện 43](#_Toc204202255)

[CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 44](#_Toc204202256)

[5.1. Kết luận 44](#_Toc204202257)

[5.2. Hướng phát triển 44](#_Toc204202258)

[TÀI LIỆU THAM KHẢO 46](#_Toc204202259)

**Danh sách hình vẽ**

* 1. Sơ đồ pipeline tích hợp mô hình lọc đánh giá và hệ thống đề xuất 17
  2. Pipeline xử lý văn bản và phát hiện đánh giá không hợp lệ 24
  3. Sơ đồ kiến trúc hệ thống lọc đánh giá người dùng và gợi ý sản phẩm 26
  4. Giao diện trang chủ và tìm kiếm sách 39
  5. Giao diện trang danh mục sản phẩm với chức năng tìm kiếm và lọc theo thể loại 40
  6. Trang chi tiết sản phẩm và các đánh giá liên quan 40
  7. Giao diện gợi ý sách theo cá nhân hóa người dùng 41
  8. Giao diện trang quản trị hệ thống với biểu đồ và bảng thống kê tương tác 42
  9. Giao diện hiển thị kết quả phân tích cảm xúc và cảnh báo đánh giá không hợp lệ 42

**Danh sách bảng**

* 1. Phân bố nhãn cảm xúc trong tập dữ liệu huấn luyện 34
  2. Tổng hợp độ chính xác và F1-score của từng mô hình 35
  3. Một số mẫu đầu vào và nhãn dự đoán cảm xúc 37

# CHƯƠNG 1: TỔNG QUAN

## Giới thiệu đề tài

Trong kỷ nguyên số hiện nay, các hệ thống đề xuất (recommender systems) đóng vai trò quan trọng trong việc cung cấp các nội dung phù hợp cho người dùng trên các nền tảng thương mại điện tử, dịch vụ truyền thông, mạng xã hội và nhiều lĩnh vực khác. Việc cá nhân hóa trải nghiệm người dùng không chỉ nâng cao hiệu quả kinh doanh mà còn giúp tiết kiệm thời gian tìm kiếm và tăng mức độ hài lòng của khách hàng.

Một trong những yếu tố quan trọng ảnh hưởng đến chất lượng của hệ thống đề xuất chính là dữ liệu đánh giá từ người dùng. Tuy nhiên, trên thực tế, các đánh giá này không phải lúc nào cũng phản ánh đúng cảm nhận thật sự, do tồn tại các đánh giá thiếu trung thực, cảm xúc tiêu cực vô lý hoặc thậm chí là giả mạo. Việc xử lý các đánh giá không hợp lệ này trở thành một vấn đề cấp thiết nhằm cải thiện chất lượng dự đoán của hệ thống đề xuất.

Đề tài “Xây dựng mô hình lọc đánh giá người dùng nhằm cải thiện chất lượng của hệ thống đề xuất” được thực hiện với mục tiêu nhận diện các đánh giá không phản ánh đúng cảm xúc thực sự, từ đó loại bỏ hoặc điều chỉnh để tăng độ chính xác và tin cậy của mô hình đề xuất.

## Tổng quan nghiên cứu liên quan

Các hệ thống đề xuất hiện nay thường dựa trên các phương pháp phổ biến như lọc cộng tác (Collaborative Filtering), lọc theo nội dung (Content-Based Filtering) và mô hình lai (Hybrid Model). Trong đó, phương pháp lọc cộng tác sử dụng đánh giá của người dùng để dự đoán mức độ yêu thích đối với các sản phẩm chưa từng tương tác. Một trong các kỹ thuật tiêu biểu là phân rã ma trận (Matrix Factorization), giúp giảm chiều dữ liệu và tìm ra mối quan hệ tiềm ẩn giữa người dùng và sản phẩm [1].

Nghiên cứu về chất lượng dữ liệu trong hệ thống đề xuất chỉ ra rằng sự hiện diện của đánh giá sai lệch ảnh hưởng lớn đến kết quả dự đoán. Một số công trình đề xuất sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) để phân tích nội dung đánh giá nhằm phát hiện dấu hiệu bất thường [2].

Đặc biệt, việc sử dụng mô hình học máy như SVM (Support Vector Machine) cho bài toán phân tích cảm xúc từ văn bản đánh giá đã được chứng minh mang lại hiệu quả cao [3]. Trong đó, việc chuẩn hóa văn bản, loại bỏ từ dừng, xử lý ngôn ngữ không chính thức là các bước quan trọng trong pipeline xử lý dữ liệu văn bản [4].

Ngoài ra, nhiều nghiên cứu gần đây đã tập trung vào việc kết hợp mô hình lọc đánh giá với hệ thống đề xuất để nâng cao chất lượng đầu ra. Ví dụ, tác giả [5] đề xuất phương pháp phân tích cảm xúc kết hợp lọc đánh giá không phù hợp nhằm cải thiện độ chính xác của mô hình đề xuất.

## Nhiệm vụ của đồ án

**Lý do chọn đề tài**

Trong quá trình phát triển hệ thống đề xuất dựa trên dữ liệu thực tế từ các nền tảng thương mại điện tử như Tiki.vn, tôi nhận thấy có nhiều đánh giá chứa nội dung không phản ánh đúng chất lượng sản phẩm hoặc mang tính thiên lệch. Việc sử dụng nguyên trạng dữ liệu này trong mô hình đề xuất dẫn đến kết quả dự đoán thiếu chính xác. Do đó, đề tài được chọn với mong muốn xây dựng một pipeline xử lý có khả năng lọc ra các đánh giá không hợp lệ nhằm cải thiện hiệu năng của hệ thống.

**Mục tiêu nghiên cứu**

* + - Xây dựng pipeline xử lý và phân tích cảm xúc từ các đánh giá văn bản của người dùng.
    - Kết hợp mô hình phân tích cảm xúc và điểm đánh giá để phát hiện các đánh giá bất thường.
    - Đề xuất hai phương pháp xử lý: (1) loại bỏ đánh giá không hợp lệ; (2) điều chỉnh điểm đánh giá dựa trên cảm xúc.
    - Đánh giá tác động của việc lọc đánh giá đến hiệu năng của mô hình đề xuất sử dụng Matrix Factorization.

**Ý nghĩa khoa học và thực tiễn**

* + - Về khoa học: Góp phần mở rộng nghiên cứu trong lĩnh vực phân tích cảm xúc và cải thiện chất lượng hệ thống đề xuất.
    - Về thực tiễn: Ứng dụng mô hình vào hệ thống thương mại điện tử giúp nâng cao độ tin cậy của đề xuất, tăng trải nghiệm người dùng và hỗ trợ doanh nghiệp ra quyết định chiến lược hiệu quả hơn.

**Đối tượng và phạm vi nghiên cứu**

* + - **Đối tượng**: Các đánh giá văn bản từ người dùng trên nền tảng thương mại điện tử.
    - **Phạm vi**: Dữ liệu thu thập từ website Tiki.vn, gồm thông tin sản phẩm sách và các đánh giá kèm điểm số. Mô hình đề xuất sử dụng phương pháp Matrix Factorization.

## Cấu trúc của đồ án

Đồ án được chia thành 5 chương chính như sau:

* + - **Chương 1: Tổng quan** – Trình bày lý do chọn đề tài, tổng quan các nghiên cứu liên quan, xác định nhiệm vụ, mục tiêu và cấu trúc đồ án.
    - **Chương 2: Cơ sở lý thuyết và công nghệ** – Trình bày các kiến thức nền tảng như hệ thống đề xuất, phân tích cảm xúc, các mô hình phân tích cảm xúc, Matrix Factorization và các thư viện liên quan.
    - **Chương 3: Phân tích và thiết kế hệ thống** – Mô tả quá trình thu thập dữ liệu, xử lý ngôn ngữ tự nhiên, xây dựng pipeline phát hiện đánh giá không hợp lệ và tích hợp vào hệ thống đề xuất.
    - **Chương 4: Thực nghiệm và đánh giá kết quả** – Trình bày các kịch bản đánh giá mô hình: giữ nguyên dữ liệu gốc, loại bỏ đánh giá, từ đó phân tích tác động đến hệ thống đề xuất.
    - **Chương 5: Kết luận và hướng phát triển** – Tổng kết những đóng góp, nêu ra giới hạn và đề xuất hướng phát triển trong tương lai như tích hợp mô hình lai, sử dụng dữ liệu lớn hoặc deep learning.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT VÀ CÔNG NGHỆ

## Tổng quan về hệ thống đề xuất

Hệ thống đề xuất (Recommender Systems) là một trong những ứng dụng phổ biến của trí tuệ nhân tạo nhằm cá nhân hóa trải nghiệm người dùng. Chúng được sử dụng rộng rãi trong các nền tảng thương mại điện tử, dịch vụ truyền phát nội dung (như Netflix, Spotify), mạng xã hội và nhiều lĩnh vực khác nhằm cung cấp các đề xuất phù hợp với sở thích và hành vi của từng người dùng.

Ba nhóm chính trong các hệ thống đề xuất bao gồm: **lọc cộng tác (Collaborative Filtering)**, **lọc theo nội dung (Content-Based Filtering)** và **lai ghép (Hybrid)**.

### Hệ thống lọc cộng tác (Collaborative Filtering)

Lọc cộng tác là phương pháp đề xuất dựa trên hành vi hoặc đánh giá của người dùng khác có sở thích tương tự. Ý tưởng cốt lõi là: nếu người A và người B cùng thích các sản phẩm giống nhau, thì những sản phẩm mà A thích nhưng B chưa xem có khả năng cao sẽ phù hợp với B.

Có hai dạng chính của Collaborative Filtering:

* **Dựa trên người dùng (User-Based CF)**: Xác định các người dùng có thị hiếu tương đồng, sau đó sử dụng đánh giá của họ để đưa ra đề xuất.
* **Dựa trên sản phẩm (Item-Based CF)**: Xác định các sản phẩm tương tự dựa trên đánh giá của người dùng và đề xuất sản phẩm tương tự với những gì người dùng đã thích.

Một số kỹ thuật điển hình sử dụng trong CF bao gồm:

* Tính toán độ tương đồng (similarity) giữa người dùng hoặc sản phẩm bằng các độ đo như Cosine, Pearson, Jaccard.
* Xây dựng ma trận người dùng – sản phẩm (user–item matrix) và áp dụng các kỹ thuật giảm chiều như Matrix Factorization.

### Hệ thống lọc theo nội dung (Content-Based Filtering)

Content-Based Filtering đề xuất sản phẩm dựa trên đặc điểm của sản phẩm và hồ sơ người dùng. Hệ thống học từ các thuộc tính của các sản phẩm mà người dùng đã yêu thích trong quá khứ và từ đó dự đoán sản phẩm mới có thể phù hợp.

Các đặc điểm kỹ thuật:

* Mỗi sản phẩm được biểu diễn bằng một vector đặc trưng (feature vector) như thể loại, mô tả, thương hiệu, v.v.
* Hồ sơ người dùng (user profile) được xây dựng từ các sản phẩm mà họ đánh giá cao.
* Đề xuất được đưa ra bằng cách tính độ tương đồng giữa sản phẩm mới và hồ sơ người dùng.

Ưu điểm của phương pháp này là khả năng đề xuất cho người dùng mới (cold-start user) nếu có thông tin về sản phẩm họ đã tương tác. Tuy nhiên, phương pháp này thường bị giới hạn bởi không gian đặc trưng và dễ rơi vào tình trạng "bong bóng lọc" (filter bubble) — người dùng chỉ nhận được những đề xuất quá giống với lịch sử tương tác.

### Hệ thống lai ghép (Hybrid Recommendation Systems)

Hệ thống lai ghép kết hợp cả Collaborative Filtering và Content-Based Filtering để tận dụng ưu điểm và khắc phục hạn chế của từng phương pháp.

Có nhiều chiến lược lai ghép:

* **Kết hợp tuần tự (Sequential Hybrid)**: Áp dụng một phương pháp đầu tiên (ví dụ Content-Based) để lọc trước, sau đó áp dụng CF để xếp hạng.
* **Kết hợp song song (Parallel Hybrid)**: Áp dụng cả hai phương pháp đồng thời và trộn kết quả theo trọng số.
* **Kết hợp tích hợp (Feature Augmentation)**: Kết hợp đặc trưng từ một phương pháp làm đầu vào cho phương pháp còn lại.

Hybrid Recommendation Systems có khả năng mở rộng tốt hơn, nâng cao độ chính xác và giảm thiểu hiệu ứng cold-start. Đây là xu hướng được sử dụng phổ biến trong các hệ thống đề xuất hiện đại.

## Kỹ thuật phân rã ma trận (Matrix Factorization)

### Khái niệm tổng quan

Matrix Factorization (MF) là một trong những kỹ thuật quan trọng và phổ biến nhất trong hệ thống đề xuất, đặc biệt trong các phương pháp Collaborative Filtering. Ý tưởng chính là phân rã ma trận tương tác người dùng – sản phẩm (user–item matrix) thành hai ma trận mật độ thấp đại diện cho đặc trưng tiềm ẩn của người dùng và sản phẩm.

MF đã được sử dụng thành công trong nhiều hệ thống thực tế, đáng chú ý nhất là chiến thắng của nhóm BellKor tại cuộc thi Netflix Prize 2009.

### Mô hình toán học

Giả sử ta có một ma trận đánh giá , trong đó biểu thị điểm số mà người dùng đánh giá sản phẩm . Trong thực tế, ma trận này rất thưa (sparse), tức là có nhiều giá trị bị thiếu (unknown ratings).

Mục tiêu của MF là tìm hai ma trận ẩn:

* : ma trận đặc trưng của người dùng (user-feature matrix)
* : ma trận đặc trưng của sản phẩm (item-feature matrix)

sao cho:

(2.1)

trong đó:

* là vector đặc trưng của người dùng
* là vector đặc trưng của sản phẩm
* là số lượng đặc trưng tiềm ẩn (latent factors)

Để tìm các ma trận và , ta giải bài toán tối ưu hàm mất mát:

(2.2)

trong đó:

* là tập hợp các cặp người dùng – sản phẩm có đánh giá thực tế.
* là hệ số điều chuẩn (regularization parameter) nhằm tránh hiện tượng overfitting.

### Thuật toán huấn luyện

Một trong các phương pháp phổ biến để tối ưu bài toán trên là **Stochastic Gradient Descent (SGD)**. Quy trình cập nhật tham số như sau:

Lỗi dự đoán: *eui* = *Rui* − *R*ˆ*ui* (2.3)

*Pu* ← *Pu* + *η*(*euiQi* − *λPu*) (2.4)

*Qi* ← *Qi* + *η*(*euiPu* − *λQi*) (2.5) Trong đó *η* là tốc độ học (learning rate).

### Ưu điểm của Matrix Factorization

* + - * **Tính khái quát cao:** MF có khả năng khai thác thông tin tiềm ẩn trong mối quan hệ người dùng – sản phẩm, không chỉ đơn thuần dựa trên điểm số.
      * **Hiệu suất tốt với dữ liệu thưa:** Dễ dàng mở rộng cho các tập dữ liệu lớn và thưa như MovieLens, Netflix.
      * **Có thể kết hợp thêm thông tin phụ:** : MF có thể được mở rộng thành các mô hình phức tạp hơn như SVD++, Factorization Machines để tích hợp thông tin hồ sơ người dùng, ngữ cảnh, thời gian.

### Hạn chế của Matrix Factorization

* + - * **Không xử lý tốt với người dùng hoặc sản phẩm mới (Cold-start problem):** Cần có đủ đánh giá để học đặc trưng ẩn.
      * **Không tận dụng đặc trưng nội dung:** Không dùng thông tin mô tả sản phẩm hoặc hồ sơ người dùng nếu không mở rộng mô hình.
      * **Yêu cầu xử lý dữ liệu đầu vào kỹ lưỡng:** Cần chuẩn hóa ma trận đầu vào, xử lý mất mát và nhiễu (ví dụ: đánh giá không hợp lệ).

### Ứng dụng thực tế

Matrix Factorization thường được sử dụng như nền tảng cho hệ thống gợi ý sản phẩm trong các nền tảng thương mại điện tử, dịch vụ streaming, mạng xã hội. Trong khuôn khổ đề tài này, MF được áp dụng để xây dựng mô hình đề xuất dựa trên dữ liệu đánh giá đã được xử lý cảm xúc, nhằm cải thiện độ tin cậy của mô hình đề xuất tổng thể.

## Phân tích cảm xúc trong đánh giá người dùng

### Mục tiêu của phân tích cảm xúc

Phân tích cảm xúc (Sentiment Analysis), hay còn gọi là phân loại quan điểm (opinion mining), là một kỹ thuật trong xử lý ngôn ngữ tự nhiên (NLP) nhằm xác định thái độ, quan điểm hoặc cảm xúc của người dùng đối với một chủ thể nào đó (sản phẩm, dịch vụ, nội dung...).

Trong ngữ cảnh hệ thống đề xuất, phân tích cảm xúc giúp đánh giá mức độ tích cực hay tiêu cực của các bình luận người dùng. Điều này đặc biệt hữu ích khi điểm số đánh giá (rating) không phản ánh đúng nội dung bình luận — một tình huống phổ biến dẫn đến đánh giá sai lệch, ảnh hưởng đến hiệu quả gợi ý.

### Các mức độ phân tích cảm xúc

Phân tích cảm xúc có thể được thực hiện ở nhiều mức độ:

* + - * **Mức văn bản (Document-level):** Xác định cảm xúc tổng thể của toàn bộ bình luận.
      * **Mức câu (Sentence-level):** Phân tích cảm xúc trong từng câu đơn lẻ.
      * **Mức khía cạnh (Aspect-level):** Xác định cảm xúc liên quan đến các thành phần cụ thể của sản phẩm (ví dụ: pin, màn hình, giá cả...).

Trong đề tài này, tập trung chính vào **phân tích cảm xúc ở mức văn bản**, phù hợp với độ dài và cấu trúc ngữ nghĩa của các bình luận đánh giá sản phẩm.

### Các kỹ thuật phân tích cảm xúc

Có nhiều phương pháp phân tích cảm xúc, được chia thành ba nhóm chính:

*Phương pháp dựa trên từ điển (Lexicon-based)*

Sử dụng các từ điển cảm xúc chứa danh sách các từ tích cực và tiêu cực. Tổng điểm cảm xúc được tính dựa trên số lượng và trọng số của các từ xuất hiện trong văn bản.

**Ưu điểm:** Đơn giản, dễ triển khai.

**Hạn chế:** Phụ thuộc mạnh vào chất lượng từ điển, kém hiệu quả với từ lóng, teencode, ngôn ngữ địa phương.

*Phương pháp học máy truyền thống (Machine Learning)*

Xem phân tích cảm xúc là một bài toán phân loại văn bản. Các bước bao gồm:

1. Tiền xử lý dữ liệu: loại bỏ stopword, emoji, từ đặc biệt, chuẩn hóa teencode.
2. Trích xuất đặc trưng: TF-IDF, Bag-of-Words, n-gram.
3. Áp dụng mô hình phân loại: Na¨ıve Bayes, Logistic Regression, SVM...

**Ưu điểm:** Dễ huấn luyện, kiểm soát được mô hình.

**Hạn chế:** Hiệu quả phụ thuộc vào chất lượng đặc trưng.

*Phương pháp học sâu (Deep Learning)*

Sử dụng các mô hình như RNN, LSTM, hoặc Transformer (BERT, PhoBERT) để học biểu diễn ngữ nghĩa ngữ cảnh. Đây là hướng hiện đại và có độ chính xác cao.

**Ưu điểm:** Tự động học đặc trưng ngữ nghĩa phức tạp.

**Hạn chế:** Yêu cầu tài nguyên tính toán và dữ liệu lớn.

### Áp dụng trong hệ thống đề xuất

Trong đề tài này, phân tích cảm xúc được sử dụng nhằm:

* + - * Phát hiện các đánh giá không nhất quán giữa văn bản và điểm số (ví dụ: bình luận tiêu cực nhưng cho 5 sao).
      * Lọc bỏ hoặc điều chỉnh các đánh giá sai lệch nhằm cải thiện độ tin cậy của dữ liệu đầu vào cho mô hình đề xuất.
      * Làm sạch dữ liệu (data cleaning) trước khi huấn luyện mô hình MF.

Quy trình áp dụng gồm các bước:

1. Thu thập dữ liệu đánh giá và bình luận từ Tiki.vn.
2. Áp dụng mô hình phân tích cảm xúc đã huấn luyện (SVM) để xác định nhãn cảm xúc.
3. Đối chiếu với điểm số gốc để phát hiện các đánh giá bất thường.
4. Tùy chọn: loại bỏ hoặc điều chỉnh điểm số trong ma trận đánh giá.

Việc tích hợp phân tích cảm xúc vào pipeline hệ thống gợi ý giúp gia tăng tính chính xác và phản ánh đúng trải nghiệm thực tế của người dùng, thay vì chỉ dựa vào điểm số.

## Các mô hình phân tích cảm xúc từ văn bản

### Tổng quan bài toán phân tích cảm xúc

Phân tích cảm xúc (Sentiment Classification) là một nhánh quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), nhằm mục tiêu xác định thái độ, quan điểm hoặc cảm xúc của người viết thông qua văn bản. Trong bối cảnh các nền tảng số ngày càng phát triển, việc người dùng để lại đánh giá, bình luận, hoặc phản hồi bằng văn bản trở nên phổ biến. Do đó, khả năng tự động phân tích và hiểu được cảm xúc ẩn sau các đoạn văn bản là một yêu cầu thiết yếu đối với nhiều hệ thống, đặc biệt là hệ thống đề xuất.

Bài toán phân tích cảm xúc thường được mô hình hóa như một bài toán phân loại văn bản, trong đó mỗi văn bản đầu vào được gán một nhãn thuộc một tập hợp hữu hạn các trạng thái cảm xúc. Các trạng thái này có thể ở mức nhị phân (ví dụ: tích cực và tiêu cực), đa lớp (tích cực, tiêu cực, trung lập), hoặc mức độ liên tục (ví dụ: thang điểm cảm xúc từ 1 đến 5). Phân tích cảm xúc không chỉ giúp đánh giá thái độ người dùng đối với sản phẩm hoặc dịch vụ mà còn hỗ trợ phát hiện hành vi bất thường, quản lý thương hiệu, và cải thiện trải nghiệm người dùng.

Khác với phân loại văn bản truyền thống, phân tích cảm xúc gặp nhiều thách thức đặc thù như:

* + **Ngôn ngữ phi cấu trúc và cảm tính**: Người dùng thường sử dụng từ lóng, từ viết tắt, biểu tượng cảm xúc hoặc cách nói gián tiếp, khiến việc diễn giải nội dung trở nên phức tạp.
  + **Nội dung ngắn và thiếu ngữ cảnh**: Các bình luận thường ngắn gọn, rời rạc, làm giảm khả năng trích xuất đặc trưng hiệu quả.
  + **Sự không nhất quán giữa nội dung và điểm số**: Một số đánh giá có nội dung tích cực nhưng lại kèm theo điểm số thấp (hoặc ngược lại), gây khó khăn trong việc gán nhãn và huấn luyện mô hình.

Để giải quyết bài toán này, các kỹ thuật học máy và học sâu thường được kết hợp với các bước tiền xử lý ngôn ngữ chuyên biệt như: tách từ, chuẩn hóa cú pháp, loại bỏ từ dừng và biểu diễn văn bản bằng các mô hình vector. Sau khi biểu diễn, văn bản được đưa vào các mô hình phân loại như SVM, Logistic Regression, hoặc các kiến trúc học sâu như BERT để xác định nhãn cảm xúc tương ứng.

Trong phạm vi đề tài này, phân tích cảm xúc giữ vai trò là bước tiền xử lý trung gian, giúp phát hiện các đánh giá không phản ánh đúng trải nghiệm thực tế. Nhờ đó, hệ thống có thể loại bỏ hoặc điều chỉnh các đánh giá sai lệch trước khi sử dụng chúng trong mô hình đề xuất, từ đó nâng cao độ chính xác và độ tin cậy của hệ thống tổng thể.

### Mô hình phân loại truyền thống

Trước khi các mô hình học sâu (deep learning) trở nên phổ biến, phần lớn các hệ thống phân tích cảm xúc dựa trên các mô hình học máy truyền thống kết hợp với kỹ thuật trích xuất đặc trưng từ văn bản như Bag-of-Words, TF-IDF hoặc n-gram. Các mô hình này vẫn giữ vai trò quan trọng trong nhiều ứng dụng thực tế nhờ tính đơn giản, dễ triển khai và hiệu quả trên tập dữ liệu vừa và nhỏ. Dưới đây là các mô hình truyền thống được sử dụng trong đề

**(1) Support Vector Machine (SVM)**

* **Nguyên lý hoạt động**: Tìm siêu phẳng phân tách tối ưu giữa các lớp, tối đa hóa khoảng cách (margin) giữa các điểm dữ liệu gần nhất.
* **Phù hợp với**: Dữ liệu thưa (TF-IDF), không gian chiều cao.
* **Ưu điểm**:
  + - Hiệu suất ổn định với văn bản.
    - Tổng quát hóa tốt, ít overfitting.
* **Hạn chế**:
  + - Cần chọn kernel và tham số phù hợp (C, gamma).
    - Thời gian huấn luyện chậm nếu dữ liệu lớn.

**(2) Logistic Regression**

* **Nguyên lý hoạt động**: Ước lượng xác suất đầu ra bằng hàm sigmoid, phân loại dựa vào ngưỡng (thường là 0.5).
* **Phù hợp với**: Bài toán phân loại nhị phân, văn bản đơn giản.
* **Ưu điểm**:
  + - Dễ triển khai và lý giải.
    - Huấn luyện nhanh, ít yêu cầu tài nguyên.
* **Hạn chế**:
  + - Kém hiệu quả với dữ liệu không tuyến tính.
    - Phụ thuộc nhiều vào kỹ thuật biểu diễn đặc trưng.

**(3) Naive Bayes**

* **Nguyên lý hoạt động**: Áp dụng định lý Bayes cùng với giả định độc lập giữa các đặc trưng (từ).
* **Phù hợp với**: Dữ liệu rời rạc như từ xuất hiện, văn bản ngắn.
* **Ưu điểm**:
  + - Nhanh, đơn giản, phù hợp làm baseline.
    - Khả năng xử lý tốt khi dữ liệu lớn và phân bố lệch.
* **Hạn chế**:
  + - Giả định độc lập giữa từ là không thực tế.
    - Kém hiệu quả với văn bản dài và ngữ nghĩa phức tạp.

**(4) K-Nearest Neighbors (KNN)**

* **Nguyên lý hoạt động**: Phân loại một điểm mới dựa trên đa số phiếu của *k* điểm gần nhất trong không gian đặc trưng.
* **Phù hợp với**: Dữ liệu có cấu trúc rõ ràng, không quá nhiều chiều.
* **Ưu điểm**:
  + - Không cần huấn luyện, dễ hiểu và trực quan.
* **Hạn chế**:
  + - Hiệu suất thấp với dữ liệu lớn và nhiều chiều (như TF-IDF).
    - Tốn bộ nhớ, thời gian dự đoán chậm.

**(5) Random Forest**

* **Nguyên lý hoạt động**: Tổ hợp nhiều cây quyết định huấn luyện trên các tập con dữ liệu và đặc trưng khác nhau.
* **Phù hợp với**: Dữ liệu phức tạp, yêu cầu ổn định và khả năng khái quát hóa cao.
* **Ưu điểm**:
  + - Hiệu quả cao, giảm overfitting.
    - Có thể xử lý cả dữ liệu định tính và định lượng.
* **Hạn chế**:
  + - Tốn thời gian huấn luyện và tài nguyên hơn các mô hình đơn.
    - Khó giải thích hơn logistic hoặc Naïve Bayes.

### Mô hình học sâu trong phân tích cảm xúc: PhoBERT

Sự phát triển mạnh mẽ của các mô hình học sâu, đặc biệt là các kiến trúc dựa trên Transformer như BERT, đã mở ra hướng tiếp cận mới trong phân tích cảm xúc với độ chính xác vượt trội. Trong ngữ cảnh tiếng Việt, **PhoBERT** là một trong những mô hình pre-trained đầu tiên và hiệu quả nhất, được huấn luyện trên tập dữ liệu tiếng Việt lớn theo kiến trúc BERT của Google.

PhoBERT là mô hình **language model pre-trained** theo phương pháp masked language modeling (MLM), được phát triển bởi nhóm nghiên cứu tại VinAI Research. PhoBERT sử dụng tập dữ liệu tiếng Việt lớn (Vietnamese Wikipedia, news, etc.) và tokenizer theo kiểu **Byte Pair Encoding (BPE)** để xử lý tiếng Việt dưới dạng subword.

**Các đặc điểm chính:**

* Kiến trúc tương đương BERT-base (12 layers, 768 hidden size, 12 attention heads).
* Được huấn luyện từ đầu bằng tiếng Việt nên hiểu rõ ngữ cảnh, ngữ pháp và cách dùng từ tiếng Việt hơn các mô hình đa ngôn ngữ như mBERT.
* Có khả năng biểu diễn sâu sắc các mối liên hệ ngữ nghĩa và cú pháp trong văn bản.

### Ứng dụng trong đề tài

Trong đề tài, mô hình ML được sử dụng như bộ phân tích cảm xúc chính, áp dụng sau bước tiền xử lý tiếng Việt. Mô hình giúp xác định các bình luận tích cực hoặc tiêu cực một cách chính xác, từ đó hỗ trợ quá trình lọc và điều chỉnh đánh giá người dùng nhằm nâng cao chất lượng dữ liệu đầu vào cho hệ thống đề xuất.

## Thư viện và công nghệ sử dụng

Trong quá trình xây dựng mô hình lọc đánh giá người dùng nhằm cải thiện chất lượng hệ thống đề xuất, đề tài sử dụng nhiều thư viện và công nghệ từ học máy, xử lý ngôn ngữ tự nhiên cho đến các công cụ triển khai ứng dụng web. Mục tiêu là đảm bảo hiệu quả trong xử lý dữ liệu, huấn luyện mô hình và tích hợp hệ thống khuyến nghị vào ứng dụng thực tế.

### Các thư viện xử lý và phân tích dữ liệu

* + - * **Pandas:** thư viện mạnh mẽ dùng để thao tác dữ liệu dạng bảng (DataFrame). Được sử dụng trong các bước xử lý dữ liệu đầu vào, phân tích thống kê và chuẩn bị tập huấn luyện.
      * **NumPy:** hỗ trợ tính toán ma trận, vector và các thao tác số học hiệu năng cao.
      * **Matplotlib, Seaborn:** phục vụ việc trực quan hóa dữ liệu và kết quả mô hình (biểu đồ tần suất đánh giá, độ chính xác, phân phối điểm số...).

### Thư viện học máy và phân tích cảm xúc

* **scikit-learn:** thư viện học máy mã nguồn mở, được sử dụng cho:
  + Tiền xử lý đặc trưng văn bản (TF-IDF, chuẩn hóa).
  + Huấn luyện mô hình phân tích cảm xúc.
  + Đánh giá mô hình (cross-validation, accuracy, confusion matrix...).
* **Surprise:** thư viện chuyên biệt cho hệ thống đề xuất, đặc biệt là các mô hình dựa trên ma trận. Được sử dụng để xây dựng mô hình *Matrix Factorization* với thuật toán **SVD** hoặc **SVD++**.
* **Underthesea, PyVi, VnCoreNLP:** các công cụ xử lý ngôn ngữ tự nhiên tiếng Việt, phục vụ:
  + Tách từ tiếng Việt (word segmentation).
  + Chuẩn hóa teencode và ngôn ngữ không chuẩn.
  + Gán nhãn từ loại, phân tích câu.
* **emoji, unidecode, re:** thư viện hỗ trợ làm sạch dữ liệu văn bản như loại bỏ biểu tượng cảm xúc, ký tự đặc biệt, chuyển đổi dấu.

### Các công nghệ triển khai hệ thống đề xuất

Sau khi xây dựng mô hình lọc đánh giá và mô hình đề xuất, hệ thống được triển khai thành một ứng dụng web cho phép người dùng tương tác. Các công nghệ sử dụng gồm:

*Backend: NodeJS*

* + - * Được sử dụng để xây dựng API trung gian giữa frontend và cơ sở dữ liệu.
      * Triển khai các hàm gọi mô hình dự đoán từ server.
      * Xử lý logic xác thực người dùng và truy vấn sản phẩm.

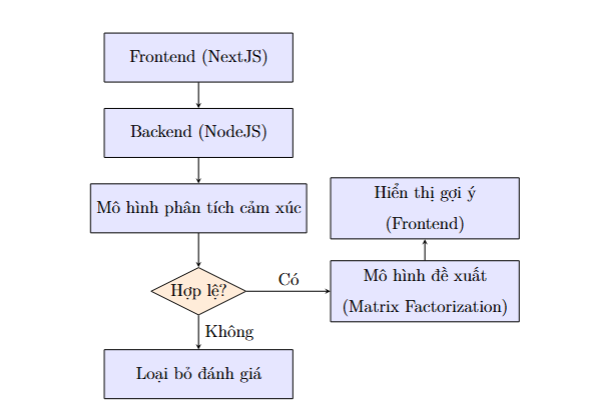
*Frontend: NextJS*

* + - * Framework ReactJS được tối ưu hóa cho hiệu suất và SEO.
      * Giao diện hiển thị danh sách sản phẩm được đề xuất, điểm đánh giá, bình luận.
      * Giao tiếp với API backend thông qua RESTful hoặc GraphQL.

*Cơ sở dữ liệu: MongoDB*

* + - * Cơ sở dữ liệu NoSQL dạng document, phù hợp với dữ liệu phi cấu trúc như đánh giá, hồ sơ người dùng.
      * Lưu trữ các sản phẩm, người dùng, đánh giá đã được gán nhãn cảm xúc và kết quả đề xuất.
      * Tích hợp tốt với NodeJS qua thư viện Mongoose.

### Quy trình tích hợp hệ thống



Hình 2.1: Sơ đồ pipeline tích hợp mô hình lọc đánh giá và hệ thống

đề xuất

1. Người dùng truy cập hệ thống qua giao diện NextJS.
2. Dữ liệu đánh giá được gửi về backend NodeJS.
3. Backend xử lý dữ liệu, gọi mô hình SVM để xác định cảm xúc, kiểm tra độ tin cậy.
4. Nếu đánh giá hợp lệ, điểm được đưa vào hệ thống đề xuất MF.
5. Kết quả gợi ý sản phẩm trả về frontend và hiển thị.

Hệ thống đảm bảo tính mở rộng và có thể nâng cấp để tích hợp các mô hình hiện đại hơn như Deep Learning hoặc mô hình học liên tục (online learning).

# CHƯƠNG 3: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

## Thu thập dữ liệu từ nền tảng thương mại điện tử Tiki.vn

### Mục tiêu thu thập dữ liệu

Trong bối cảnh xây dựng hệ thống lọc đánh giá người dùng nhằm cải thiện chất lượng của hệ thống đề xuất, việc thu thập một tập dữ liệu thực tế, đa dạng và phản ánh đúng hành vi đánh giá của người dùng là điều kiện tiên quyết. Nền tảng thương mại điện tử Tiki.vn được lựa chọn bởi vì đây là một trong những website thương mại điện tử hàng đầu tại Việt Nam, sở hữu số lượng lớn sản phẩm và bình luận đánh giá từ phía người dùng, đặc biệt trong lĩnh vực sách – nơi cảm xúc đóng vai trò quan trọng trong hành vi tiêu dùng.

### Danh mục dữ liệu mục tiêu

Tập dữ liệu cần thu thập bao gồm hai thành phần chính:

* + - * **Thông tin sản phẩm**: bao gồm tiêu đề, mô tả, tác giả (đối với sách), danh mục, giá bán, và điểm đánh giá trung bình.
      * **Đánh giá người dùng (user reviews)**: bao gồm nội dung bình luận, điểm số (rating), thời gian đánh giá, và các phản hồi (nếu có).

Tập trung vào nhóm sản phẩm sách, sáu danh mục được chọn gồm: sách văn học, sách kinh tế, sách kỹ năng sống, nuôi dạy con, sách kiến thức tổng hợp và sách lịch sử địa lý.

### Phương pháp thu thập dữ liệu

Việc thu thập dữ liệu được thực hiện thông qua cách tiếp cận **web crawling** sử dụng các API công khai của Tiki. Để đảm bảo khả năng tái lập và tính hệ thống, tôi đã xây dựng một tập hợp script Python tự động hoá toàn bộ quá trình thu thập sản phẩm và đánh giá tương ứng. Quá trình được chia thành các bước như sau:

*Bước 1: Lấy danh sách sản phẩm theo danh mục*

Dữ liệu được truy xuất từ endpoint:

https://tiki.vn/api/personalish/v1/blocks/listings

API này trả về danh sách các sản phẩm thuộc danh mục được chỉ định thông qua cate- gory\_id. Với mỗi danh mục, hệ thống sẽ lặp qua nhiều trang kết quả (tối đa 3 trang đầu tiên để đảm bảo tính thực thi và giảm tải).

*Bước 2: Truy xuất chi tiết sản phẩm*

Với mỗi sản phẩm thu được, hệ thống tiếp tục truy xuất thông tin chi tiết qua API:

https://tiki.vn/api/v2/products/{product\_id}

Dữ liệu chi tiết được lưu trữ dưới định dạng JSON với cấu trúc cây, bao gồm các trường như tiêu đề, mô tả, giá bán, số lượng đánh giá, số điểm trung bình, thuộc tính chi tiết.

*Bước 3: Truy xuất đánh giá người dùng*

Sau khi thu được mã sản phẩm (product\_id), đánh giá của người dùng được truy xuất qua endpoint:

https://tiki.vn/api/v2/reviews?product\_id={id}&page={n}

Quá trình này được thực hiện theo trang, cho đến khi không còn trang dữ liệu mới hoặc đạt đến trang cuối cùng được chỉ định trong phản hồi API. Các đánh giá được trích xuất bao gồm nội dung văn bản, điểm số, thời gian đánh giá, và các tương tác đi kèm (nếu có).

*Bước 4: Lưu trữ và tổ chức dữ liệu*

Toàn bộ dữ liệu được lưu dưới cấu trúc cây thư mục rõ ràng:

* data/raw/products/{category}: chứa các tệp JSON sản phẩm theo danh mục.
* data/raw/reviews/{category}: chứa các tệp JSON đánh giá theo danh mục sản phẩm.

Mỗi sản phẩm và tệp đánh giá tương ứng được lưu tách biệt, đảm bảo dễ dàng thao tác và tiền xử lý về sau.

### Tổng quan kết quả thu thập

Sau khi hoàn tất tiến trình thu thập, hệ thống đã xây dựng được một tập dữ liệu thực tế có quy mô lớn bao gồm:

* + - * **Tổng số đánh giá:** 296.836 đánh giá người dùng.
      * **Số lượng người dùng duy nhất:** 181.741 người dùng.
      * **Số lượng sản phẩm sách duy nhất:** 881 sản phẩm.

Tập dữ liệu có độ bao phủ cao, phản ánh hành vi và quan điểm tiêu dùng từ nhiều phân khúc người dùng khác nhau. Đây là cơ sở vững chắc để triển khai các bước phân tích cảm xúc, phát hiện đánh giá không hợp lệ và đánh giá hiệu quả mô hình đề xuất trong các phần tiếp theo của đồ án.Quy trình tiền xử lý dữ liệu

### Mục tiêu tiền xử lý

Dữ liệu văn bản thu thập từ các đánh giá người dùng trên nền tảng Tiki.vn thường mang tính tự do, không chuẩn hóa, chứa nhiều ký hiệu, từ viết tắt (teencode), từ ngữ địa phương hoặc cảm tính. Để đảm bảo chất lượng đầu vào cho mô hình học máy, đặc biệt là mô hình phân tích cảm xúc và phát hiện đánh giá không hợp lệ, việc xây dựng một pipeline tiền xử lý ngôn ngữ tự nhiên tiếng Việt chuyên biệt là yêu cầu thiết yếu.

### Các bước xử lý văn bản tiếng Việt

Pipeline tiền xử lý văn bản được thiết kế theo các bước sau:

* **Làm sạch dữ liệu:** Loại bỏ các ký tự không cần thiết như emoji, biểu tượng cảm xúc, liên kết URL, đoạn văn bị lỗi (ví dụ: \_x000D\_), và chuẩn hóa định dạng văn bản về chữ thường.
* **Chuẩn hóa từ viết tắt và teencode:** Áp dụng từ điển ánh xạ gồm các từ teencode phổ biến sang tiếng Việt chuẩn. Ví dụ: “hok” “không”, “tks” “cảm ơn”.
* **Chuẩn hóa cụm từ phổ biến (phrase rules):** Các cụm từ thường xuất hiện trong đánh giá được nối lại thành đơn vị từ ghép bằng dấu gạch dưới. Ví dụ: “giao hàng” “giao\_hàng”, “chất lượng” “chất\_lượng”. Luật này đặc biệt quan trọng nhằm bảo toàn ngữ nghĩa trong phân tích cảm xúc và cải thiện độ chính xác mô hình. Bộ luật được trích xuất từ dữ liệu thực tế và bao gồm hàng trăm cụm từ quan trọng như: “bao bọc”, “bìa sách”, “dịch vụ”, “giá cả”, “mùa dịch”, “giao nhầm”, “bản dịch”, “móp méo”, “sách đẹp”, …
* **Chuẩn hóa ký tự lặp:** Loại bỏ hoặc giảm số lượng ký tự lặp không cần thiết nhằm xử lý hiện tượng nhấn mạnh cảm xúc (ví dụ: “hayyyy” “hay”).
* **Tách từ:** Sử dụng công cụ tách từ tiếng Việt ViTokenizer để phân đoạn văn bản thành các từ đơn vị, hỗ trợ bước biểu diễn đặc trưng.
* **Xử lý phủ định và cụm ngữ cảnh tiêu cực:** Nhằm bảo toàn ngữ nghĩa cảm xúc, hệ thống thực hiện nối cụm từ có chứa tiền tố phủ định như “không”, “chưa”, “đừng” với động từ đi sau, ví dụ: “không tốt” “không\_tốt”, “không có bọc” “không\_có\_bọc”. Đồng thời xử lý các chỉ dấu tiêu cực như “bị”, “thiếu”, “nên”, “được” để xác định rõ thái độ đánh giá.
* **Xoá từ dừng (stopwords):** Áp dụng danh sách từ dừng tiếng Việt được biên soạn thủ công, bao gồm các từ không mang ý nghĩa phân biệt trong cảm xúc như “là”, “một”, “có”, “những”, “và”, …nhằm giảm nhiễu.
* **Xử lý cụ thể cho một số cụm thường gặp:** Một số cụm từ có nhiều biến thể như “giao hàng”, “bao bọc”, “bọc sách”, “giao nhanh” được xử lý riêng để tăng tính nhất quán trong biểu diễn văn bản.

### Triển khai luật cụm từ (phrase rules)

Bộ luật cụm từ được xây dựng từ tập dữ liệu thực tế bằng cách trích xuất các n-gram có tần suất cao và ngữ nghĩa rõ ràng. Các cụm từ này có tác dụng quan trọng trong việc giữ nguyên khối ngữ nghĩa, ví dụ:

* + - * **Giao nhận và đóng gói:** *giao hàng*, *đóng gói*, *gói hàng*, *vận chuyển*, *giao nhầm*, *giao chậm*.
      * **Chất lượng sản phẩm:** *chất lượng*, *sách mới*, *sách rách*, *sách đẹp*, *sách bẩn*.
      * **Cảm nhận của người dùng:** *hài lòng*, *thất vọng*, *xuất sắc*, *phí tiền*, *tuyệt vời*, *đáng kể*.
      * **Thuật ngữ chuyên ngành:** *sách kỹ năng*, *sách giáo khoa*, *truyện tranh*, *bản dịch*, *kiến thức*.
      * **Cụm từ nhấn mạnh cảm xúc:** *100 điểm*, *siêu nhanh*, *tuyệt vời ông mặt trời*, *toẹt zời*, *must have*.

Việc chuẩn hóa các cụm từ này không chỉ cải thiện tính đồng nhất trong biểu diễn dữ liệu mà còn hỗ trợ mô hình phát hiện các đánh giá giả mạo vốn thường lặp lại cụm từ mẫu, thiếu sự tự nhiên.

### Lọc người dùng có ít đánh giá

Một bước quan trọng trong quá trình tiền xử lý là loại bỏ các người dùng có số lượng đánh giá quá ít (ít hơn 3 đánh giá), nhằm đảm bảo rằng các đặc trưng liên quan đến hành vi người dùng có đủ thông tin để học. Việc lọc này nhằm:

* + - * Giảm thiểu độ thưa của ma trận người dùng–sản phẩm.
      * Loại bỏ các tài khoản khả nghi hoặc đánh giá thử nghiệm (noise).
      * Cải thiện hiệu quả và độ tin cậy của các thuật toán học máy.

Tổng thể, sau khi áp dụng các bước tiền xử lý, dữ liệu văn bản được đưa về định dạng sạch, thống nhất, dễ học và sẵn sàng cho bước tiếp theo: gán nhãn cảm xúc và phân loại.

## Thiết kế pipeline xử lý

Sau khi hoàn thiện bước thu thập và tiền xử lý dữ liệu, hệ thống tiến hành xây dựng pipeline xử lý nhằm phục vụ cho ba nhiệm vụ chính: **(1)** gán nhãn cảm xúc cho đánh giá, **(2)** huấn luyện mô hình phân loại văn bản dựa trên cảm xúc, và **(3)** phát hiện các đánh giá không hợp lệ dựa trên sự sai lệch giữa nội dung cảm xúc và điểm số.

### Gán nhãn cảm xúc cho dữ liệu

Để xây dựng tập huấn luyện cho bài toán phân tích cảm xúc, hệ thống dựa vào điểm đánh giá (rating) do người dùng cung cấp. Mỗi đánh giá được gán một nhãn cảm xúc như sau:

* + - * **Tích cực (Positive):** nếu rating ≥ 4.
      * **Tiêu cực (Negative):** nếu rating ≤ 2.
      * **Trung lập (Neutral):** nếu rating = 3 (các đánh giá này được loại bỏ trong giai đoạn huấn luyện để tăng tính rõ ràng cho mô hình phân loại nhị phân).

Tập dữ liệu huấn luyện cuối cùng bao gồm các cặp (văn bản đánh giá, nhãn cảm xúc) với hai lớp: tích cực và tiêu cực.

### Biểu diễn đặc trưng văn bản

Dữ liệu sau khi tiền xử lý được biểu diễn thành dạng số sử dụng mô hình **TF-IDF (Term Frequency – Inverse Document Frequency)**. Kỹ thuật này giúp đánh giá mức độ quan trọng của một từ trong một văn bản, tương quan với toàn bộ tập dữ liệu.  
Một số cấu hình được sử dụng cho bộ biểu diễn TF-IDF:

* + - * **max\_features = 3000:** giới hạn số lượng đặc trưng tối đa.
      * **min\_df = 5:** loại bỏ các từ xuất hiện ít hơn 5 lần trong toàn bộ tập.
      * **max\_df = 0.8:** loại bỏ các từ xuất hiện trong hơn 80% tài liệu (có thể là stopword).
      * **sublinear\_tf = True:** áp dụng phép biến đổi logarit để làm mượt tần suất.

Kết quả là mỗi văn bản được ánh xạ thành một vector không gian đặc trưng thưa (sparse vector), phản ánh mức độ nổi bật của từng cụm từ trong ngữ cảnh cảm xúc.

### Huấn luyện mô hình phân tích cảm xúc

Sau khi hoàn tất giai đoạn tiền xử lý và biểu diễn đặc trưng văn bản bằng mô hình TF-IDF, hệ thống tiến hành huấn luyện và kiểm thử trên nhiều mô hình phân tích cảm xúc khác nhau. Mục tiêu không phải là triển khai duy nhất một mô hình cụ thể, mà là đánh giá hiệu năng của từng thuật toán học máy và học sâu, từ đó lựa chọn mô hình phù hợp nhất để tích hợp vào hệ thống.

Cụ thể, các mô hình học máy truyền thống như **Logistic Regression**, **Naïve Bayes**, **Random Forest**, **K-Nearest Neighbors** và **Support Vector Machine** được huấn luyện trên cùng một tập dữ liệu chuẩn hóa. Đồng thời, mô hình học sâu **PhoBERT** – một mô hình ngôn ngữ tiền huấn luyện mạnh mẽ dành riêng cho tiếng Việt – cũng được fine-tune để so sánh.

Tất cả các mô hình được huấn luyện và đánh giá dưới cùng một quy trình thống nhất, sử dụng tập huấn luyện và tập kiểm thử độc lập, với các chỉ số chính như độ chính xác (accuracy), F1-score và confusion matrix. Quá trình đánh giá này cho phép xác định mô hình nào có khả năng phân tích cảm xúc người dùng tốt nhất trong điều kiện dữ liệu thực tế, phục vụ mục tiêu nâng cao chất lượng lọc đánh giá trong hệ thống đề xuất.

Chi tiết về kết quả huấn luyện, độ chính xác của từng mô hình và phân tích so sánh sẽ được trình bày cụ thể trong Chương 4.

### Phát hiện đánh giá không hợp lệ

Sau khi xây dựng được mô hình phân tích cảm xúc từ nội dung đánh giá, hệ thống tiến hành so sánh kết quả dự đoán với điểm số (rating) gốc do người dùng cung cấp. Dựa trên giả định rằng nội dung và điểm số phải nhất quán về mặt cảm xúc, một đánh giá được xem là **không hợp lệ** nếu thỏa mãn điều kiện:

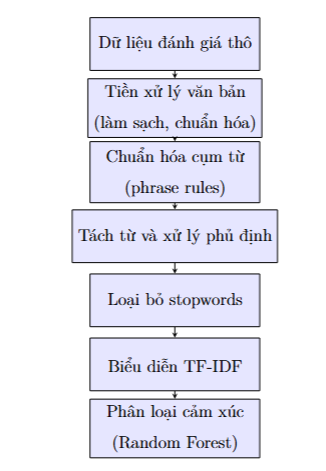
* + - * **Nội dung mang cảm xúc tích cực nhưng rating thấp (1 hoặc 2).**
      * **Nội dung mang cảm xúc tiêu cực nhưng rating cao (4 hoặc 5).**

Các đánh giá không hợp lệ có thể đến từ:

* + - * Lỗi hệ thống hoặc thao tác sai của người dùng khi chọn rating.
      * Hành vi bất thường như spam, đánh giá giả mạo, hoặc cảm xúc thay đổi trong nội dung.
      * Người dùng sử dụng ngôn ngữ mỉa mai, gián tiếp (được xử lý hạn chế trong bước phân tích ngữ nghĩa).

Việc phát hiện các đánh giá không hợp lệ đóng vai trò quan trọng trong việc **lọc nhiễu đầu vào** cho hệ thống đề xuất, từ đó cải thiện độ chính xác và tin cậy của các gợi ý sản phẩm.

### Tổng quan pipeline xử lý



Hình 3.1: Pipeline xử lý văn bản và phát hiện đánh giá không hợp lệ

Pipeline xử lý này là xương sống cho toàn bộ hệ thống, kết nối trực tiếp giữa dữ liệu thô và các mô-đun phân tích nâng cao được trình bày trong các chương sau.

## Mô tả kiến trúc hệ thống: sơ đồ tổng quát các khối chức năng

### Tổng quan hệ thống

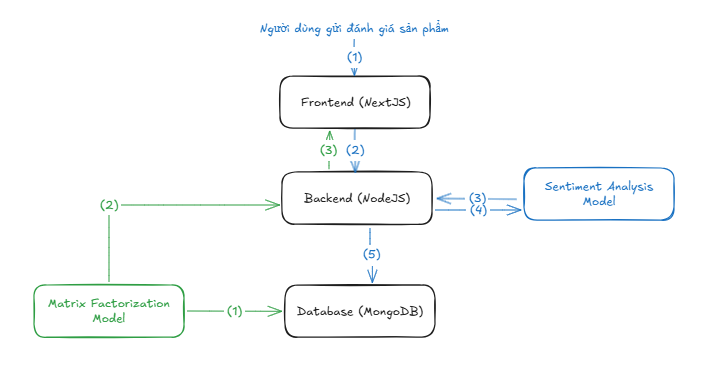
Hệ thống được triển khai nhằm thực hiện nhiệm vụ phân tích đánh giá người dùng, phát hiện các đánh giá không hợp lệ, và cải thiện chất lượng gợi ý sản phẩm. Kiến trúc tổng thể gồm nhiều thành phần hoạt động độc lập nhưng liên kết chặt chẽ thông qua API. Các khối chức năng chính bao gồm:

* + - * **Frontend (Next.js):** là giao diện người dùng nơi người dùng nhập đánh giá, tương tác với hệ thống và nhận kết quả gợi ý. Giao diện gửi dữ liệu đến backend và hiển thị phản hồi từ hệ thống.
      * **Backend (Node.js):** xử lý các yêu cầu từ frontend, định tuyến dữ liệu, gọi các mô-đun xử lý cảm xúc và mô hình gợi ý. Backend đóng vai trò trung tâm điều phối.
      * **Sentiment Analysis Model:** mô hình học máy được huấn luyện để phân tích cảm xúc từ văn bản đánh giá. Mô hình này nhận dữ liệu từ backend, dự đoán nhãn cảm xúc, và trả kết quả lại.
      * **Database (MongoDB):** cơ sở dữ liệu lưu trữ toàn bộ đánh giá, thông tin sản phẩm, kết quả phân tích cảm xúc và kết quả lọc.
      * **Matrix Factorization Model:** mô hình gợi ý sản phẩm, sử dụng dữ liệu đã lọc để xây dựng ma trận người dùng–sản phẩm. Mô hình này đầu ra là danh sách sản phẩm được cá nhân hóa cho từng người dùng.

### Luồng dữ liệu và tương tác

Hình 3.2 mô tả trực quan luồng dữ liệu giữa các thành phần trong hệ thống.   
Cụ thể:

1. Người dùng nhập đánh giá sản phẩm thông qua frontend.
2. Dữ liệu được chuyển đến backend để xử lý.
3. Backend lưu dữ liệu xuống cơ sở dữ liệu MongoDB.
4. Đồng thời, đánh giá được gửi đến mô hình phân tích cảm xúc để xác định nhãn cảm xúc.
5. Kết quả phân tích cảm xúc được trả về backend và lưu cùng dữ liệu gốc.
6. Backend sử dụng dữ liệu đã được dán nhãn cảm xúc để phát hiện các đánh giá không hợp lệ.
7. Những đánh giá hợp lệ sẽ được đưa vào mô hình gợi ý (Matrix Factorization) để huấn luyện hoặc cập nhật dự đoán.
8. Cuối cùng, kết quả gợi ý được gửi về frontend để hiển thị cho người dùng.



Hình 3.2: Sơ đồ kiến trúc hệ thống lọc đánh giá người dùng và gợi ý sản phẩm

### Khả năng mở rộng và tích hợp

Kiến trúc hiện tại được thiết kế theo nguyên tắc module hoá, cho phép dễ dàng mở rộng trong tương lai:

* + - * Thay thế mô hình sử dụng ML bằng các mô hình học sâu như BERT hoặc LSTM để nâng cao độ chính xác.
      * Tích hợp các công nghệ mới như Redis cache hoặc GraphQL để tối ưu hiệu năng truy vấn.
      * Kết nối với các nguồn dữ liệu khác (ví dụ: đánh giá từ mạng xã hội hoặc diễn đàn) để tăng độ bao phủ dữ liệu.

Hệ thống mang tính linh hoạt cao và có thể áp dụng vào nhiều lĩnh vực thương mại điện tử khác nhau với cấu trúc tương tự.

## Các yếu tố thiết kế quan trọng

Quá trình thiết kế và triển khai hệ thống không chỉ dừng lại ở các thành phần chức năng, mà còn đòi hỏi sự chú trọng đến các yếu tố kỹ thuật cốt lõi ảnh hưởng trực tiếp đến độ tin cậy, khả năng mở rộng và hiệu quả xử lý. Ba yếu tố chính bao gồm: dữ liệu đầu vào/đầu ra, mô-đun huấn luyện và cơ chế lưu trữ kết quả.

### Dữ liệu đầu vào và đầu ra

* + - * **Dữ liệu đầu vào:** Hệ thống tiếp nhận hai loại dữ liệu đầu vào chính:
        + Dữ liệu đánh giá từ người dùng (bình luận văn bản và điểm số).
        + Dữ liệu sản phẩm từ sàn thương mại điện tử (tên sách, danh mục, mô tả).

Dữ liệu được xử lý tiền xử lý trước khi đưa vào pipeline học máy, nhằm loại bỏ nhiễu, chuẩn hóa ngữ nghĩa và đảm bảo tính đồng nhất.

* + - * **Dữ liệu đầu ra:** Tùy theo giai đoạn xử lý, hệ thống tạo ra ba loại kết quả:
        + Nhãn cảm xúc (tích cực hoặc tiêu cực) gán cho mỗi đánh giá.
        + Cảnh báo đánh giá không hợp lệ nếu có sự mâu thuẫn giữa nội dung và điểm số.
        + Danh sách gợi ý sản phẩm cá nhân hóa cho từng người dùng.

### Mô-đun huấn luyện và dự đoán

Hệ thống bao gồm hai mô-đun học máy độc lập, mỗi mô-đun đảm nhận vai trò cụ thể trong pipeline xử lý:

* + - * **Mô hình phân tích cảm xúc:** sử dụng dữ liệu đánh giá văn bản đã được gán nhãn (dựa vào rating) để huấn luyện mô hình phân loại nhị phân. Mô hình Random Forest được lựa chọn nhờ khả năng phân biệt rõ ràng trong không gian đặc trưng thưa (TF-IDF) và tốc độ huấn luyện nhanh.
      * **Mô hình đề xuất (Matrix Factorization):** sử dụng tập người dùng – sản phẩm với điểm số đã được lọc hoặc điều chỉnh để xây dựng ma trận dự đoán ẩn. Việc huấn luyện mô hình được thực hiện theo chiến lược offline, với khả năng cập nhật định kỳ hoặc theo chu kỳ học online trong tương lai.

Các mô hình được đóng gói thành module độc lập (bằng Python hoặc thư viện Scikit-learn), có thể dễ dàng tích hợp vào backend thông qua API hoặc script xử lý nền.

### Cơ chế lưu trữ và quản lý kết quả

Hệ thống sử dụng cơ sở dữ liệu **MongoDB** để lưu trữ tất cả thông tin liên quan, bao gồm:

* + - * Dữ liệu thô: thông tin sản phẩm, đánh giá gốc của người dùng.
      * Dữ liệu đã xử lý: văn bản đã chuẩn hóa, kết quả phân tích cảm xúc, nhãn phân loại.
      * Kết quả mô hình: ma trận người dùng–sản phẩm, danh sách sản phẩm gợi ý, chỉ số đánh giá mô hình (MAE, RMSE).

Cơ sở dữ liệu được thiết kế theo hướng phi quan hệ (NoSQL), linh hoạt trong việc mở rộng schema, cho phép lưu trữ cả văn bản, mảng số liệu và các đối tượng kết quả trung gian.

Các bản ghi có thể được truy vấn trực tiếp từ backend để phục vụ hiển thị trên giao diện người dùng hoặc làm đầu vào cho các mô hình tiếp theo. Ngoài ra, hệ thống có khả năng xuất kết quả ra định dạng CSV hoặc JSON để phục vụ trực quan hóa và phân tích bên ngoài.

### Tính mở rộng và bảo trì

Toàn bộ hệ thống được tổ chức theo kiến trúc module, mỗi thành phần có thể dễ dàng bảo trì, cập nhật hoặc thay thế độc lập. Một số định hướng nâng cấp trong tương lai gồm:

* + - * Tích hợp mô hình học sâu (Deep Learning) cho phân tích cảm xúc nâng cao.
      * Áp dụng học liên tục (online learning) cho hệ thống đề xuất theo thời gian thực.
      * Tự động gán nhãn cảm xúc bằng mô hình pre-trained thay vì dựa vào rating.
      * Lưu trữ dữ liệu trên hệ thống phân tán như MongoDB Cluster hoặc ElasticSearch.

Nhờ tính linh hoạt này, hệ thống có khả năng thích ứng với khối lượng dữ liệu lớn hơn cũng như các bài toán tương tự trong các miền ứng dụng khác.

# CHƯƠNG 4: THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Chương này trình bày quá trình thực nghiệm đánh giá hiệu quả của hệ thống đề xuất và mô hình phân tích cảm xúc. Các kết quả được so sánh và phân tích nhằm rút ra vai trò của bước lọc đánh giá trong việc nâng cao chất lượng đề xuất sản phẩm.

## Thực nghiệm mô hình phân tích cảm xúc

### Tập dữ liệu và tiền xử lý

Để huấn luyện mô hình phân tích cảm xúc, tôi xây dựng một tập dữ liệu gồm **4.024** đánh giá văn bản tiếng Việt từ người dùng trên nền tảng thương mại điện tử Tiki.vn. Mỗi đánh giá bao gồm hai thông tin chính: nội dung bình luận và điểm số (rating) từ 1 đến 5. Tuy nhiên, thay vì sử dụng trực tiếp điểm số làm nhãn cảm xúc, tôi áp dụng phương pháp gán nhãn bán tự động kết hợp giữa mô hình ngôn ngữ lớn (ChatGPT) và kiểm tra thủ công nhằm đảm bảo độ tin cậy cho tập huấn luyện.

*Quy trình gán nhãn cảm xúc*

* + - 1. **Gán nhãn sơ bộ bằng ChatGPT:** Từng đánh giá được đưa vào mô hình ChatGPT để phân tích cảm xúc thành hai nhãn: pos (tích cực) hoặc neg (tiêu cực). Việc gán nhãn dựa trên nội dung ngữ nghĩa của đánh giá, không phụ thuộc vào điểm số gốc. Ví dụ:

“Hãy đọc đoạn đánh giá sau và phân tích cảm xúc người viết là tích cực (pos) hoặc tiêu cực (neg). Trả lời duy nhất bằng pos hoặc neg.”

* + - 1. **Hiệu chỉnh thủ công:** Sau khi ChatGPT gán nhãn cho toàn bộ tập dữ liệu, tôi tiến hành rà soát thủ công lại toàn bộ các mẫu đã được dán nhãn. Những trường hợp có cảm xúc mơ hồ, giọng điệu mỉa mai, hoặc mâu thuẫn với rating được điều chỉnh lại để đảm bảo tính nhất quán và chất lượng dữ liệu đầu vào.

Việc kết hợp gán nhãn tự động và kiểm tra thủ công giúp tiết kiệm đáng kể thời gian nhưng vẫn duy trì được độ chính xác của bộ dữ liệu.

*Phân bố nhãn cảm xúc cuối cùng*

Sau quá trình hiệu chỉnh, dữ liệu được phân thành hai lớp cảm xúc như sau:

Bảng 4.1: Phân bố nhãn cảm xúc trong tập dữ liệu huấn luyện

|  |  |
| --- | --- |
| **Nhãn cảm xúc** | **Số lượng đánh giá** |
| Tích cực (pos) | 1.911 |
| Tiêu cực (neg) | 2.113 |
| Tổng cộng | 4.024 |

*Tiền xử lý văn bản*

Trước khi đưa vào mô hình học máy, dữ liệu văn bản được chuẩn hóa thông qua pipeline tiền xử lý bao gồm:

1. **Làm sạch văn bản:** Loại bỏ emoji, ký tự đặc biệt, liên kết URL và định dạng lỗi (ví dụ:

\_x000D\_).

1. **Chuẩn hóa ngôn ngữ:** Thay thế các từ viết tắt (teencode), từ tiếng Anh phổ biến bằng tương đương tiếng Việt.
2. **Tách từ và xử lý cấu trúc ngữ nghĩa:** Áp dụng công cụ VnTokenizer để tách từ, chuẩn hóa phủ định và cụm từ cố định.
3. **Xoá từ dừng (stopwords):** Loại bỏ các từ không mang ngữ nghĩa phân loại như “là”, “có”, “một”, “và”, . . .
4. **Trích xuất đặc trưng:** Sử dụng mô hình TF-IDF Vectorizer để biểu diễn văn bản thành dạng vector số phục vụ cho mô hình học máy.

Nhờ quy trình tiền xử lý kỹ lưỡng, tập dữ liệu đầu vào đạt chất lượng cao, phục vụ hiệu quả cho việc huấn luyện mô hình phân tích cảm xúc được trình bày trong phần tiếp theo.

*Các bước tiền xử lý*

Vì dữ liệu đánh giá mang tính tự nhiên, đa dạng ngôn ngữ và biểu cảm, việc tiền xử lý là bước thiết yếu nhằm chuẩn hóa và làm sạch đầu vào. Các bước thực hiện cụ thể như sau:

1. **Làm sạch văn bản:** loại bỏ các ký tự đặc biệt, biểu tượng cảm xúc (emoji), đoạn văn lỗi định dạng như “\_x000D\_”, đường dẫn URL và chuyển toàn bộ văn bản sang chữ thường.
2. **Chuẩn hóa teencode và từ viết tắt:** sử dụng từ điển ánh xạ để thay thế các từ như “tks”

→ “cảm ơn”, “hok” → “không”.

1. **Ghép cụm từ quan trọng (phrase rules):** nối các cụm từ như “giao hàng” → “giao\_hàng” để giữ nguyên ngữ nghĩa trong quá trình tách từ.
2. **Xử lý phủ định và cụm tiêu cực:** nối cụm từ phủ định như “không tốt” → “không\_tốt” hoặc “bị móp” → “bị\_móp” nhằm bảo tồn cảm xúc.
3. **Tách từ tiếng Việt:** sử dụng thư viện pyvi.ViTokenizer để tách các câu thành danh sách từ đơn vị.
4. **Xoá từ dừng:** áp dụng danh sách stopword tiếng Việt để loại bỏ các từ không có tính phân biệt về mặt cảm xúc như “là”, “nhưng”, “đã”.
5. **Biểu diễn đặc trưng:** sử dụng **TF-IDF Vectorizer** với cấu hình:
   * max\_features = 3000
   * min\_df = 5, max\_df = 0.8
   * sublinear\_tf = True
   * N-gram: (1, 2) (unigram + bigram)

Kết quả, mỗi đánh giá văn bản được biểu diễn thành một vector thưa (sparse vector) phản ánh mức độ nổi bật của các cụm từ quan trọng trong ngữ cảnh cảm xúc.

### Huấn luyện mô hình và đánh giá

Sau khi hoàn tất quá trình tiền xử lý và gán nhãn cảm xúc cho tập dữ liệu đánh giá văn bản, hệ thống tiến hành huấn luyện các mô hình phân loại cảm xúc với mục tiêu tìm ra mô hình có hiệu suất tốt nhất để ứng dụng vào pipeline lọc đánh giá không hợp lệ.

Các mô hình được lựa chọn bao gồm cả thuật toán học máy truyền thống (**Logistic Regression**, **Naïve Bayes**, **Support Vector Machine**, **Random Forest**, **K-Nearest Neighbors**) và mô hình học sâu hiện đại (**PhoBERT**) nhằm so sánh toàn diện về hiệu năng.

**Quy trình huấn luyện:**

* **Tập dữ liệu**: được chia theo tỷ lệ 80% dùng để huấn luyện và 20% để kiểm thử. Hai lớp cảm xúc (tích cực và tiêu cực) được phân bố cân bằng nhằm đảm bảo tính công bằng trong đánh giá.
* **Biểu diễn đặc trưng**:
  + Đối với các mô hình học máy truyền thống: sử dụng biểu diễn văn bản bằng TF-IDF với cấu hình max\_features = 3000, n-gram = (1,2), min\_df = 5, max\_df = 0.8.
  + Đối với PhoBERT: văn bản được token hóa bằng tokenizer của mô hình, sau đó fine-tune toàn bộ mạng Transformer với một lớp phân loại ở đầu ra.
* **Huấn luyện**:
  + Mỗi mô hình được huấn luyện với tập dữ liệu giống nhau để đảm bảo tính so sánh.
  + Các mô hình được triển khai bằng thư viện scikit-learn (đối với ML) và transformers (đối với PhoBERT).
  + Với PhoBERT, quá trình fine-tune sử dụng optimizer AdamW, learning rate thấp (2e-5) và chạy trên GPU để tối ưu hiệu quả.

Bảng 4.2 Tổng hợp độ chính xác và F1-score của từng mô hình

| **Mô hình** | **Accuracy** | **F1-score (neg)** | **F1-score (pos)** | **F1-score trung bình** |
| --- | --- | --- | --- | --- |
| **Random Forest** | **0.9085** | 0.9132 | 0.9032 | **0.9082** |
| **SVM** | 0.8960 | 0.9023 | 0.8889 | 0.8956 |
| **Logistic Regression** | 0.8943 | 0.9031 | 0.8838 | 0.8935 |
| **Naïve Bayes** | 0.8869 | 0.8942 | 0.8784 | 0.8863 |
| **K-Nearest Neighbors** | 0.6972 | 0.6293 | 0.7440 | 0.6867 |

### Phân tích kết quả

Kết quả thực nghiệm trên năm mô hình học máy cho bài toán phân loại cảm xúc đánh giá văn bản đã phản ánh rõ hiệu quả của từng phương pháp trong việc nhận diện cảm xúc tích cực và tiêu cực. Các chỉ số đánh giá như độ chính xác (accuracy), F1-score, và sự cân bằng giữa hai lớp cảm xúc được sử dụng để làm cơ sở phân tích, nhằm lựa chọn mô hình tối ưu cho hệ thống.

* Random Forest – Mô hình tốt nhất
  + Với độ chính xác đạt 90.85%, Random Forest là mô hình có hiệu suất cao nhất trong số các mô hình được kiểm thử. Mô hình đạt F1-score cao và gần bằng nhau cho cả hai lớp cảm xúc (neg: 0.9132, pos: 0.9032), cho thấy khả năng phân loại đồng đều và không thiên lệch. Đây là lợi thế quan trọng khi áp dụng vào hệ thống thực tế, nơi cả hai loại cảm xúc đều cần được phát hiện chính xác để đảm bảo chất lượng dữ liệu đầu vào.
  + Sự vượt trội của Random Forest có thể lý giải bởi khả năng học tổ hợp nhiều cây quyết định từ các đặc trưng ngẫu nhiên, từ đó giảm thiểu overfitting và cải thiện khả năng tổng quát hóa trên dữ liệu phức tạp.
* SVM và Logistic Regression – Ổn định và hiệu quả
  + Hai mô hình tuyến tính này cho kết quả gần tương đương, với độ chính xác khoảng 89.4% – 89.6% và F1-score dao động từ 0.88 – 0.90. Cụ thể, SVM có xu hướng phân loại lớp tiêu cực tốt hơn (recall neg = 0.9130), trong khi Logistic Regression ưu tiên phát hiện lớp tích cực (precision pos = 0.9235). Điều này phản ánh sự khác biệt nhỏ trong chiến lược tối ưu hóa giữa hai mô hình.
  + Dù không đạt độ chính xác cao nhất, SVM và Logistic Regression vẫn là các lựa chọn hợp lý trong các tình huống đòi hỏi hiệu suất tốt với chi phí tính toán thấp và dễ triển khai.
* Naïve Bayes – Hiệu quả đơn giản
  + Naïve Bayes, dù dựa trên giả định độc lập giữa các từ, vẫn cho kết quả đáng ghi nhận với độ chính xác 88.69%. Mô hình này có ưu thế rõ rệt về tốc độ huấn luyện và mức tiêu tốn tài nguyên thấp, phù hợp cho các hệ thống yêu cầu thời gian phản hồi nhanh hoặc xử lý trên thiết bị có cấu hình hạn chế. Tuy nhiên, so với các mô hình còn lại, độ chính xác và F1-score thấp hơn một chút khiến nó không phải là lựa chọn ưu tiên trong bối cảnh yêu cầu độ chính xác cao.
* K-Nearest Neighbors – Không phù hợp
  + KNN cho thấy hiệu suất phân loại kém với độ chính xác chỉ 69.72% và sự mất cân bằng rõ rệt giữa hai lớp (recall lớp neg chỉ đạt 48.89%, trong khi lớp pos lại đạt 92.81%). Điều này cho thấy mô hình gặp khó khăn khi xử lý không gian đặc trưng có chiều cao như dữ liệu văn bản TF-IDF, đồng thời dễ bị ảnh hưởng bởi nhiễu và phân bố không đều của các điểm dữ liệu trong không gian.

Dựa trên phân tích này, KNN bị loại khỏi các mô hình đề xuất trong hệ thống.

### Minh hoạ kết quả dự đoán

Một số ví dụ thực tế từ tập kiểm tra được minh họa trong Bảng 4.3:

Bảng 4.3: Một số mẫu đầu vào và nhãn dự đoán cảm xúc

|  |  |  |
| --- | --- | --- |
| **Rating** | **Bình luận** | **Label (dự đoán)** |
| 5 | Một quyển sách hay | pos |
| 5 | Sách đẹp, hài lòng | pos |
| 5 | Ok sách đẹp | pos |
| 2 | Giao hàng tệ, sách rách | neg |
| 1 | Mất thời gian, không đáng đọc | neg |

### Kết luận phần thực nghiệm cảm xúc

Từ kết quả thực nghiệm, có thể kết luận rằng **Random Forest** là mô hình phù hợp nhất để triển khai trong hệ thống phân tích cảm xúc của đề tài. Bên cạnh đó, **SVM** và **Logistic Regression** cũng là các lựa chọn thay thế đáng tin cậy nếu cần tối ưu chi phí huấn luyện hoặc đơn giản hóa triển khai. Trong khi đó, **Naïve Bayes** phù hợp làm baseline nhanh, còn **KNN** không đáp ứng yêu cầu về hiệu quả phân loại và không được khuyến nghị sử dụng trong bài toán này.

## Thực nghiệm hệ thống đề xuất

### Thiết lập thực nghiệm

Trong phần này, tôi tiến hành đánh giá hiệu quả của hệ thống đề xuất khi áp dụng các phương pháp xử lý đánh giá người dùng. Cụ thể, mô hình đề xuất sử dụng là **Matrix Factorization (MF)** được tối ưu bằng **thuật toán Gradient Descent**, được hiện thực trong class MFRecommender tự xây dựng.

* **Ngôn ngữ**: Python 3.10
* **Thư viện**: pandas, numpy, matplotlib
* **Tập dữ liệu**: 73.271 đánh giá thực từ người dùng trên Tiki.vn
* **Tập huấn luyện / kiểm thử**: chia theo tỉ lệ 80% / 20%
* **Số yếu tố tiềm ẩn (latent factors)**: 20
* **Số epoch huấn luyện**: 20

### Các kịch bản thực nghiệm

Ba kịch bản được xây dựng để kiểm tra hiệu quả của bước xử lý đánh giá:

* **K1 – Dữ liệu gốc**: giữ nguyên toàn bộ đánh giá ban đầu, không loại bỏ hay điều chỉnh.
* **K2 – Lọc đánh giá không hợp lệ**: loại bỏ các đánh giá có mâu thuẫn giữa cảm xúc và điểm số..

### Kết quả thực nghiệm

Sau khi huấn luyện mô hình MF trên từng kịch bản, tôi đánh giá hiệu suất dựa trên chỉ số **RMSE (Root Mean Squared Error)**.

Bảng 4.5: So sánh RMSE giữa các kịch bản

|  |  |
| --- | --- |
| **Kịch bản** | **RMSE (↓)** |
| K1 – Dữ liệu gốc | 0.550866 |
| K2 – Lọc đánh giá không hợp lệ | 0.545250 |

### Phân tích và nhận xét.

* **Kịch bản K2**, trong đó các đánh giá không hợp lệ được loại bỏ khỏi tập dữ liệu huấn luyện, cho kết quả RMSE thấp hơn so với kịch bản sử dụng toàn bộ dữ liệu (K1). Điều này cho thấy việc **lọc bỏ các đánh giá không phản ánh đúng cảm xúc thực tế của người dùng** giúp mô hình đề xuất học được thông tin chính xác và nhất quán hơn từ dữ liệu đầu vào.
* **Chênh lệch RMSE giữa hai kịch bản tuy không lớn**, nhưng mang ý nghĩa quan trọng: chỉ cần một lượng nhỏ nhiễu trong tập dữ liệu (các đánh giá lệch cảm xúc) cũng có thể ảnh hưởng đến độ chính xác của mô hình đề xuất.
* Việc lọc đánh giá không hợp lệ không chỉ giúp giảm sai số dự đoán, mà còn **nâng cao độ tin cậy** của hệ thống trong việc cá nhân hóa gợi ý cho người dùng.

### Kết luận phần thực nghiệm đề xuất

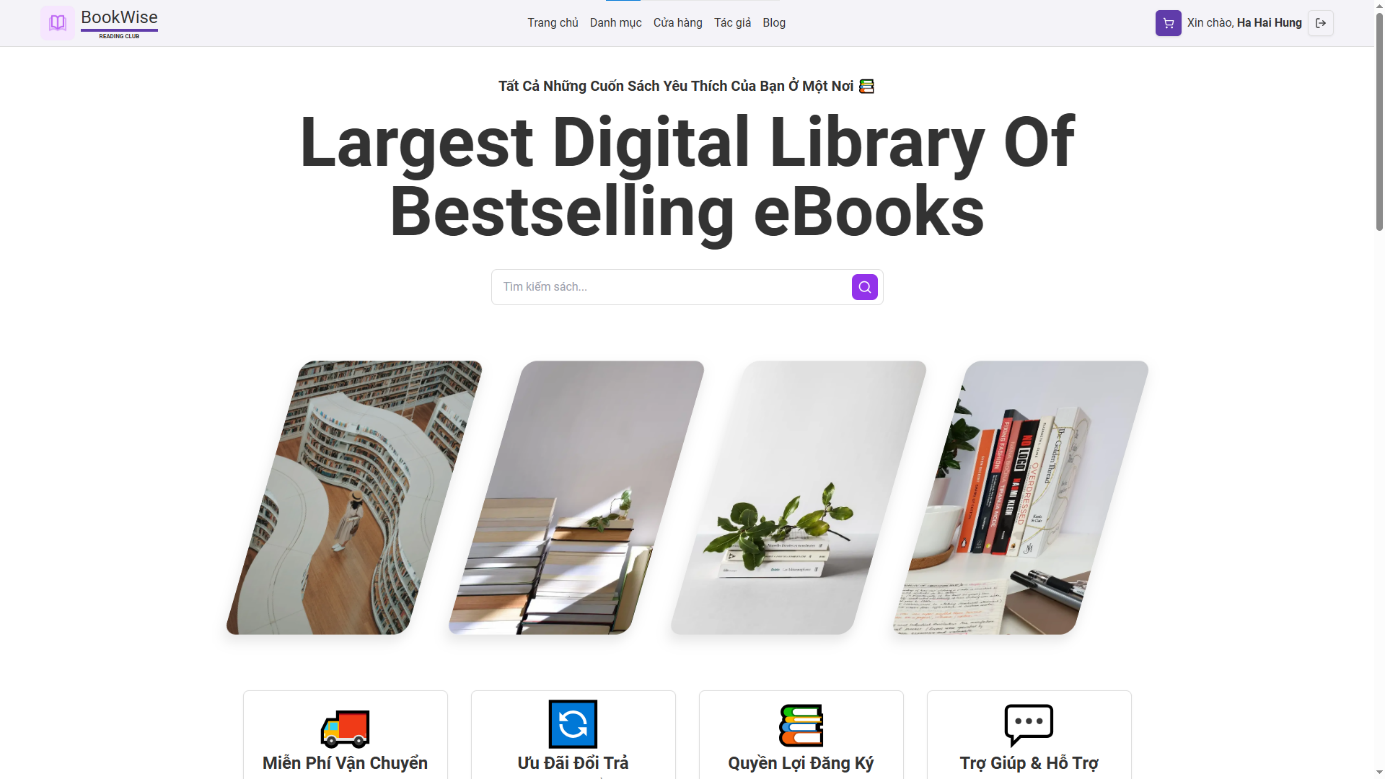
Kết quả cho thấy, việc kết hợp giữa **phân tích cảm xúc vào hệ thống đề xuất** là một hướng đi hợp lý và khả thi. Các phương pháp làm sạch và kiểm soát chất lượng dữ liệu sẽ tiếp tục được xem xét mở rộng trong các hướng phát triển tương lai của hệ thống.

## Giao diện website demo

Để minh họa tính ứng dụng thực tế của hệ thống lọc đánh giá người dùng và mô hình đề xuất sản phẩm, tôi đã xây dựng một **giao diện web demo** nhằm thử nghiệm luồng hoạt động từ phía người dùng đến hệ thống backend. Hệ thống sử dụng các công nghệ như Next.js cho frontend, Node.js cho backend và MongoDB cho lưu trữ dữ liệu.

### Trang chủ và tìm kiếm sản phẩm

Giao diện chính cung cấp chức năng tìm kiếm sản phẩm sách, hiển thị thông tin cơ bản như tên, ảnh bìa và điểm đánh giá trung bình. Người dùng có thể chọn sản phẩm để xem chi tiết hoặc gửi đánh giá mới.



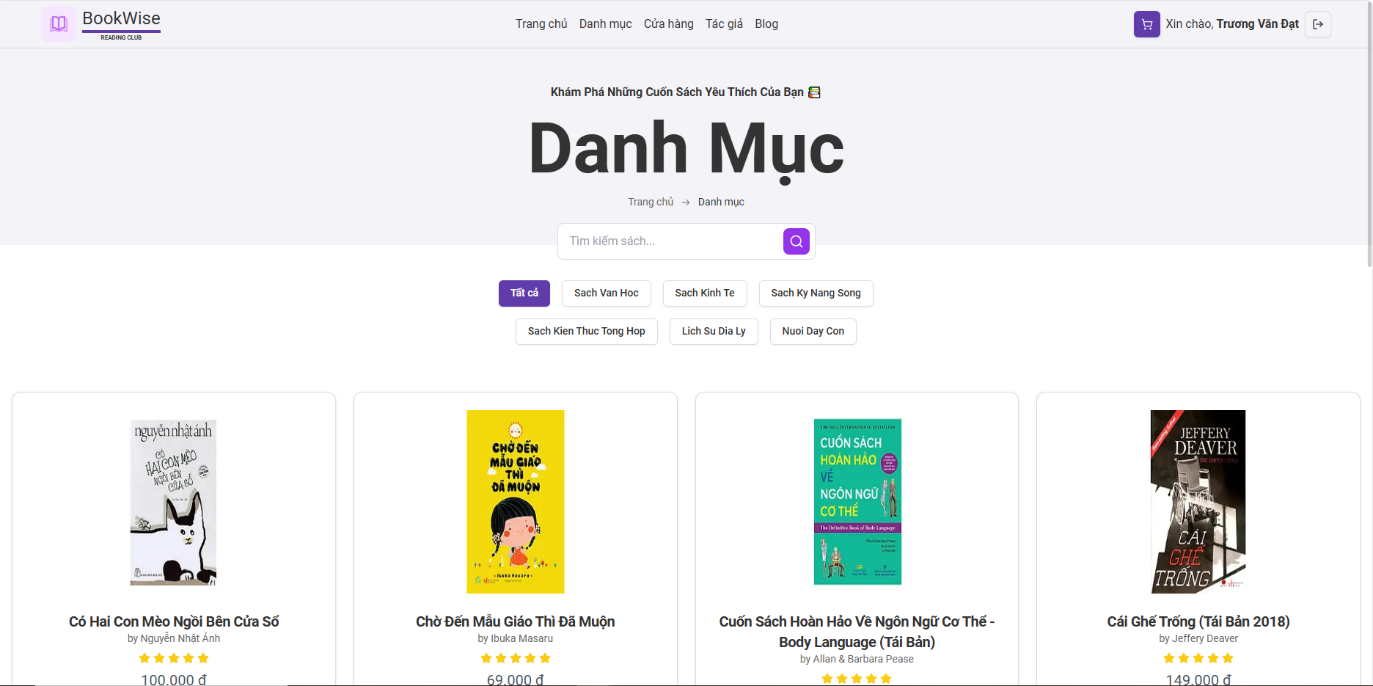
Hình 4.1: Giao diện trang chủ và tìm kiếm sách

### Trang danh mục sản phẩm

Trang danh mục cho phép người dùng duyệt qua các nhóm sách được phân loại theo chủ đề như: Văn học, Kinh tế, Kỹ năng sống, Nuôi dạy con, Lịch sử – Địa lý, Kiến thức tổng hợp, v.v. Giao diện trực quan hỗ trợ tìm kiếm và lọc nhanh theo loại sách.

* **Thanh tìm kiếm:** cho phép nhập từ khoá để lọc danh sách theo tiêu đề sách.
* **Bộ lọc theo danh mục:** mỗi nút lọc đại diện cho một thể loại sách, cho phép hiển thị các sản phẩm thuộc loại tương ứng.
* **Danh sách sản phẩm:** hiển thị hình ảnh bìa sách, tiêu đề, tác giả, giá và điểm đánh giá trung bình dưới dạng sao.

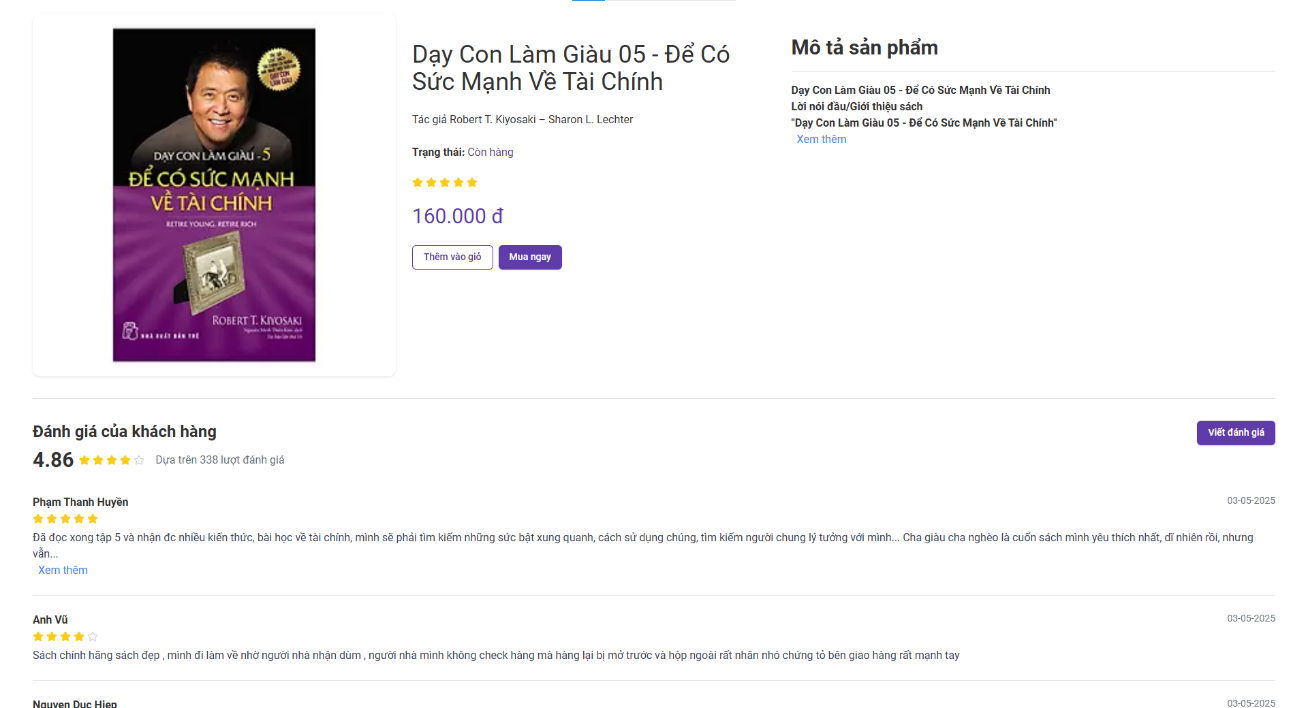
Giao diện này đóng vai trò là điểm khởi đầu chính để người dùng tiếp cận kho sách, từ đó có thể chọn sản phẩm yêu thích và thực hiện đánh giá.



Hình 4.2: Giao diện trang danh mục sản phẩm với chức năng tìm kiếm và lọc theo thể loại

### Trang chi tiết sản phẩm và đánh giá người dùng

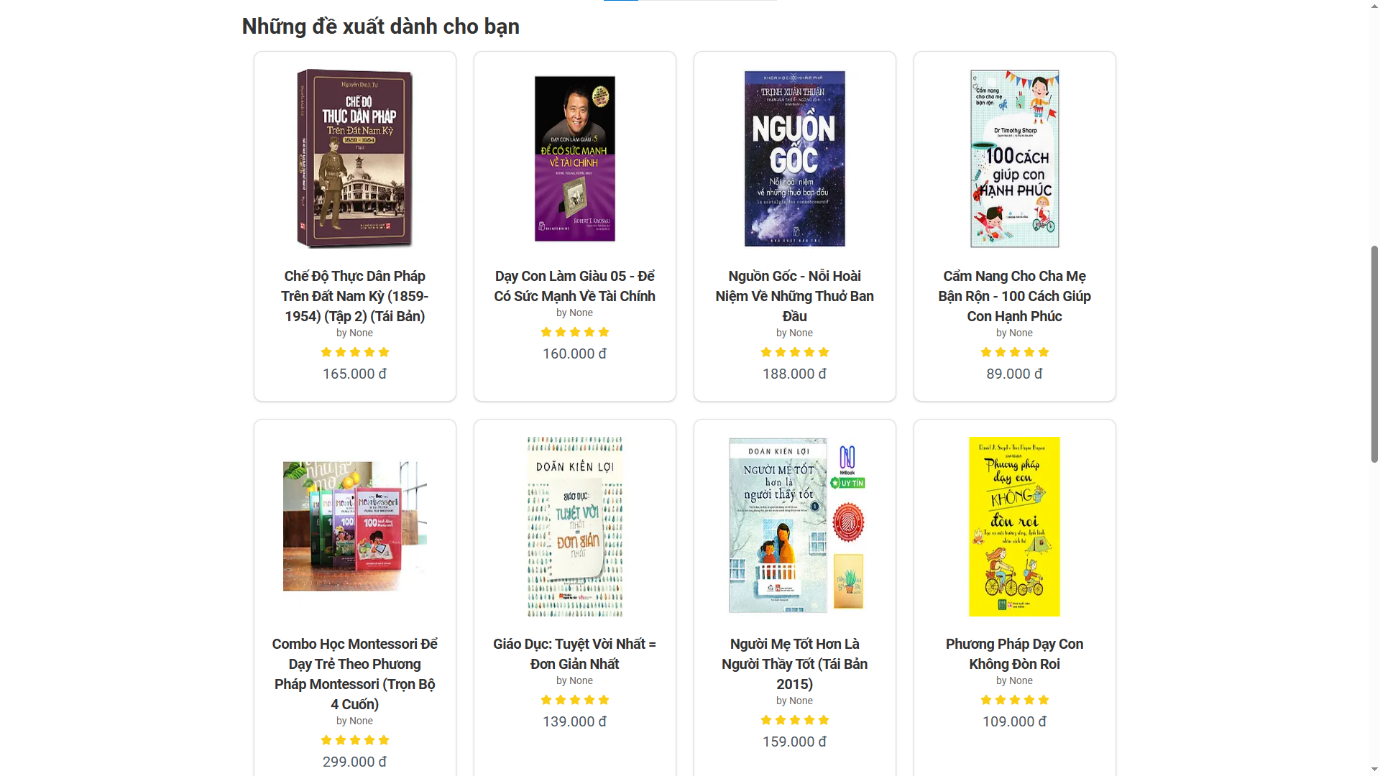
Người dùng có thể truy cập vào chi tiết một sản phẩm cụ thể để xem toàn bộ các đánh giá từ người dùng khác. Đồng thời, hệ thống cung cấp biểu mẫu để người dùng nhập bình luận mới và chấm điểm sản phẩm.



Hình 4.3: Trang chi tiết sản phẩm và các đánh giá liên quan

### Trang gợi ý sản phẩm cá nhân hóa

Cuối cùng, người dùng có thể truy cập mục “Sản phẩm dành cho bạn” để nhận các gợi ý từ hệ thống dựa trên hành vi đánh giá cá nhân. Mô hình Matrix Factorization sử dụng các đánh giá đã lọc để tạo ra các gợi ý chính xác hơn.

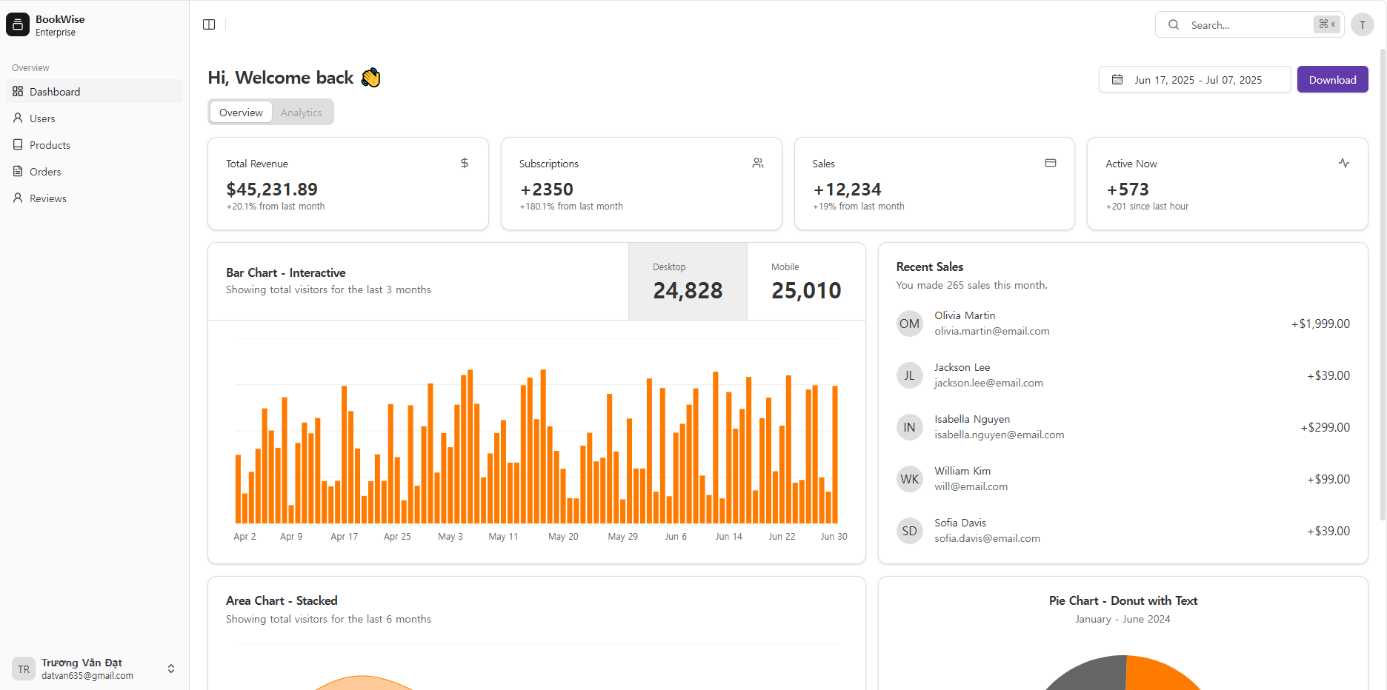


Hình 4.4: Giao diện gợi ý sách theo cá nhân hóa người dùng

### Trang quản trị và thống kê hệ thống

Ngoài các chức năng dành cho người dùng cuối, hệ thống còn cung cấp một giao diện quản trị (admin dashboard) cho phép người quản trị theo dõi và kiểm soát dữ liệu hệ thống, bao gồm thông tin đánh giá, hoạt động người dùng và hiệu quả gợi ý. Giao diện này được xây dựng bằng framework Next.js, tích hợp biểu đồ và bảng thống kê tương tác.

* **Tổng quan hoạt động:** thống kê số lượng doanh thu, lượt đánh giá, người dùng đang hoạt động và tần suất truy cập gần đây.
* **Biểu đồ tương tác:** trực quan hoá dữ liệu bằng biểu đồ cột (bar chart), biểu đồ vùng (area chart), biểu đồ tròn (pie chart), phục vụ theo dõi xu hướng sử dụng.
* **Lịch sử đánh giá:** hiển thị danh sách các đánh giá mới nhất cùng địa chỉ email người đánh giá.
* **Tính năng lọc và xuất dữ liệu:** người quản trị có thể chọn khoảng thời gian để lọc thống kê và xuất báo cáo.

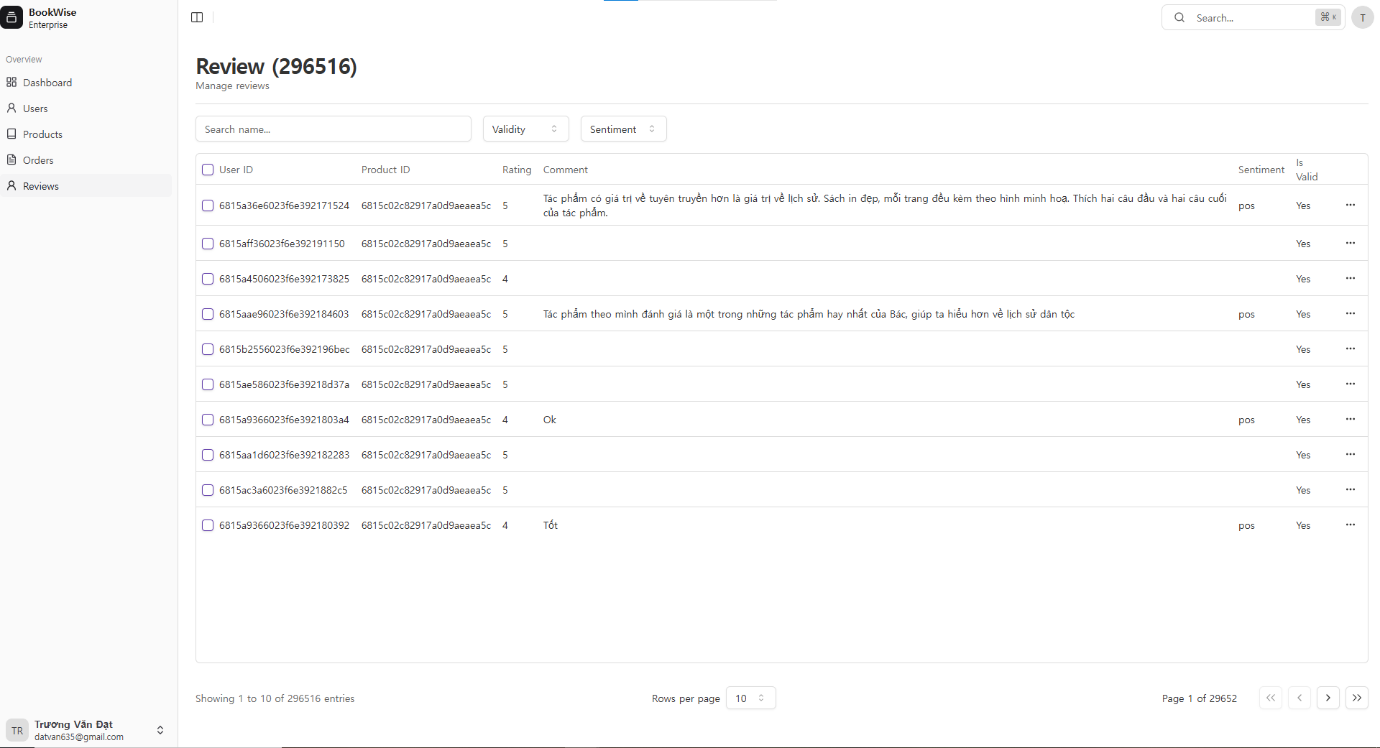


Hình 4.5: Giao diện trang quản trị hệ thống với biểu đồ và bảng thống kê

Trang quản trị đóng vai trò hỗ trợ người vận hành kiểm tra dữ liệu đầu vào, giám sát các lượt tương tác, và theo dõi hiệu quả của hệ thống lọc đánh giá và đề xuất. Trong tương lai, trang này có thể được mở rộng thêm để quản lý người dùng, phân tích mức độ tin cậy của đánh giá, hoặc trực tiếp hiệu chỉnh dữ liệu bất thường.

### Phân tích đánh giá và lọc nội dung bất thường

Sau khi người dùng gửi đánh giá, hệ thống thực hiện phân tích cảm xúc tự động. Nếu phát hiện mâu thuẫn giữa nội dung và điểm số, đánh giá sẽ được đánh dấu là “không hợp lệ” và có thể bị loại bỏ hoặc điều chỉnh lại điểm trong hệ thống gợi ý.



Hình 4.6: Giao diện hiển thị kết quả phân tích cảm xúc và cảnh báo đánh giá không hợp lệ

### Tổng kết phần giao diện

Giao diện web được xây dựng để minh họa trực quan quá trình vận hành của hệ thống, từ việc tiếp nhận dữ liệu người dùng, xử lý cảm xúc cho đến gợi ý sản phẩm. Hệ thống có thể dễ dàng mở rộng để tích hợp vào các nền tảng thương mại điện tử thực tế.

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết luận

Trong khuôn khổ đồ án này, tôi đã xây dựng và triển khai một hệ thống lọc đánh giá người dùng nhằm nâng cao chất lượng của hệ thống đề xuất sản phẩm. Cụ thể, đề tài tập trung vào việc phát hiện và xử lý các đánh giá không phản ánh đúng cảm xúc thực tế của người dùng – một yếu tố thường bị bỏ qua trong các hệ thống gợi ý truyền thống.

Hệ thống được thiết kế theo hướng modular, bao gồm các bước: thu thập dữ liệu từ nền tảng thương mại điện tử Tiki.vn, tiền xử lý ngôn ngữ tự nhiên tiếng Việt, gán nhãn và phân loại cảm xúc bằng nhiều mô hình học máy khác nhau, phát hiện đánh giá không hợp lệ, và cuối cùng là huấn luyện mô hình đề xuất trên tập dữ liệu đã được lọc.

Một số kết quả tiêu biểu đạt được:

* + - **Mô hình Random Forest** đạt độ chính xác cao nhất (**90.85%**) trong bài toán phân loại cảm xúc, vượt qua các mô hình khác như SVM, Logistic Regression và Naïve Bayes.
    - Việc **loại bỏ các đánh giá không hợp lệ** dựa trên mâu thuẫn giữa nội dung cảm xúc và điểm số đã giúp **cải thiện RMSE** của mô hình đề xuất từ **0.5509** xuống **0.5453**.
    - Giao diện web demo minh họa được toàn bộ quy trình xử lý, từ người dùng nhập đánh giá cho đến việc hệ thống gợi ý sản phẩm phù hợp sau khi đã lọc dữ liệu.

Từ đó, có thể kết luận rằng việc kết hợp phân tích cảm xúc với hệ thống đề xuất không chỉ khả thi về mặt kỹ thuật mà còn mang lại hiệu quả rõ rệt về mặt chất lượng đầu ra. Đây là minh chứng cho tính ứng dụng thực tiễn của các kỹ thuật xử lý ngôn ngữ tự nhiên và học máy trong việc cải thiện trải nghiệm người dùng trên nền tảng số.

## Hướng phát triển

Mặc dù đã đạt được một số kết quả tích cực, đề tài vẫn còn nhiều hướng mở rộng và cải tiến trong tương lai:

* + - **Ứng dụng mô hình học sâu trong phân tích cảm xúc**: Thay vì chỉ sử dụng SVM, việc áp dụng các kiến trúc mạng nơ-ron như BiLSTM, BERT hoặc PhoBERT có thể nâng cao độ chính xác phân tích cảm xúc, đặc biệt trong ngôn ngữ tiếng Việt giàu ngữ cảnh.
    - **Kết hợp thông tin nội dung sản phẩm**: Việc khai thác thêm các đặc trưng nội dung của sản phẩm như thể loại, mô tả, từ khóa sẽ giúp cải thiện chất lượng đề xuất, đặc biệt đối với người dùng mới (cold-start).
    - **Tích hợp các phương pháp embedding hoặc attention-based**: Việc sử dụng các kỹ thuật embedding hiện đại (như word2vec, fastText) hoặc mô hình dựa trên attention (VD: Transformer, SASRec) có thể giúp xây dựng hệ thống đề xuất dựa trên ngữ nghĩa sâu hơn, phù hợp với xu hướng hiện tại trong lĩnh vực Recommender Systems.
    - **Triển khai và đánh giá trên dữ liệu lớn hơn**: Để kiểm định khả năng tổng quát, hệ thống cần được thử nghiệm trên các tập dữ liệu quy mô lớn hơn và thuộc nhiều miền sản phẩm khác nhau, ví dụ như điện tử, thời trang, thực phẩm,...

Tóm lại, đề tài đã đưa ra một phương pháp tiếp cận khả thi để lọc đánh giá người dùng và chứng minh được tác động tích cực của nó trong hệ thống đề xuất. Với các cải tiến kể trên, hệ thống có tiềm năng phát triển thành một giải pháp mạnh mẽ và linh hoạt trong thực tiễn.

# TÀI LIỆU THAM KHẢO

1. Y. Koren và R. Bell, “Advances in collaborative filtering,” trong Recommender Systems Handbook, ấn bản lần 3, F. Ricci, L. Rokach và B. Shapira (biên tập), Springer, 2022, trang 91–142.
2. J. Zhang, Y. Guo và Y. Chen, “Trust-aware collaborative filtering for fake review detection,” IEEE Transactions on Knowledge and Data Engineering, tập 32, số 12, trang 2481–2494, 2020.
3. Z. Yang, K. Zhang và M. Sun, “Sentiment classification using deep learning approaches: A survey,” IEEE Transactions on Affective Computing, tập 10, số 3, trang 328–350, 2019.
4. B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, ấn bản lần 2, Cambridge University Press, 2020.
5. Z. Fan, X. Liu và Q. Yang, “Enhancing recommender systems with opinion-aware review filtering,” trong Proceedings of the Web Conference (WWW), 2021, trang 1129–1140.