HỆ KHUYẾN NGHỊ SẢN PHẨM TRÊN SÀN THƯƠNG MẠI ĐIỆN TỬ KẾT HỢP TIÊU ĐỀ, MÔ TẢ VÀ HÌNH ẢNH SẢN PHẨM

Cao Đình Duy Ngọc, Đỗ Phạm Phúc Tính, Đinh Văn Nguyên, Huỳnh Văn Tín
Khoa Khoa Học và Kỹ Thuật Thông Tin, Trường Đại Học Công Nghệ Thông Tin
Đại Học Quốc Gia Thành Phố Hồ Chí Minh
{20521661,20522020,20520657}gm.uit.edu.vn
tinhv@uit.edu.vn

Abstract

Trong thời kỳ mua sắm trực tuyến phát triển, người dùng thường có xu hướng tìm kiếm nhiều sản phẩm tương tự từ nhiều của hàng khác nhau để so sánh và tìm kiếm những sản phẩm phù hợp với bản thân nhất. Hệ thống khuyến nghị là một phần quan trọng của các nền tảng thương mại điện tử, đóng vai trò quan trọng trong việc cung cấp trải nghiệm mua sắm cá nhân hóa và hiệu quả cho người dùng. Trong bài toán này, chúng tôi giới thiệu một hệ thống khuyến nghị gồm hai thành phần chính: truy xuất (retrieval) và tái xếp hạng (re-rank). Hệ thống khuyến nghị dựa trên ba tiêu chí: tiêu đề sản phẩm, mô tả sản phẩm và hình ảnh sản phẩm. Hệ thống của chúng tôi tận dụng những tiêu chí này để nâng cao độ chính xác của việc khuyến nghị sản phẩm. Sử dụng độ đo precision at K (P@K) làm tiêu chí đánh giá, chúng tôi cố gắng xác định cấu hình trọng số tối ưu cho mỗi tiêu chí. Mục tiêu là tìm ra và đáp ứng nhu cầu của người dùng để khám phá những sản phẩm liên quan trên sàn thương mại điện tử. Ngoài ra, chúng tôi cũng so sánh thêm kết quả đối với việc sử dụng mô hình ngôn ngữ lớn để tóm tắt phần mô tả sản phẩm, cô đọng được những nội dung chính. Kết quả chúng tôi đạt được cho thấy hiệu suất tích cực của phương pháp trong việc cải thiện độ chính xác của hệ thống khuyến nghị.

1 Giới Thiệu

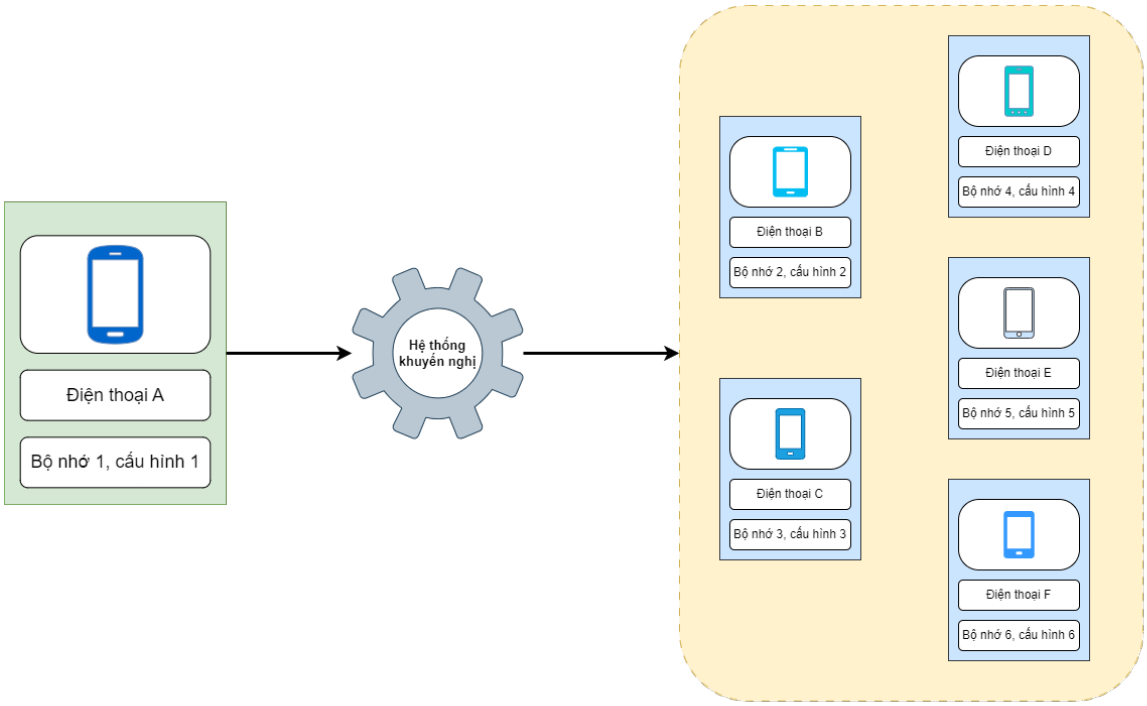
Với sự phát triển nhanh chóng của thương mại điện tử, người tiêu dùng ngày càng được đặt ở trung tâm của trải nghiệm mua sắm trực tuyến, và hệ thống khuyến nghị đóng vai trò quan trọng như một trợ lý thông minh, giúp họ khám phá và chọn lựa những sản phẩm phù hợp nhất.

Hệ thống khuyến nghị có thể được chia thành ba loại chính: Collaborative Filtering (CF), Content-based Filtering (CB) và Hybrid. Hệ thống lọc cộng tác phân tích hành vi của người dùng để tạo ra đề xuất; hệ thống dựa trên nội dung sử dụng chỉ thông tin ý nghĩa từ lĩnh vực của đề xuất, và hệ thống kết hợp đơn giản là sự kết hợp của hai phương pháp này (Berbatova, 2019). Tuy nhiên với quy mô rộng lớn, các sàn thương mại điện tử thường đa dạng về chủng loại lẫn số lượng, điều đó khiến cho khuyến nghị dựa trên phương pháp lọc cộng tác khá khó khăn bởi những đánh giá của người dùng quá thưa thớt so với số lượng sản phẩm. Do đó trong bài toán này chúng tôi sử dụng khuyến nghị dựa trên nội dung (content-based filtering).

Hầu hết các phương pháp khuyến nghị dựa trên nội dung nhận văn bản làm đầu vào duy nhất (ví dụ, tiêu đề sản phẩm, mô tả sản phẩm, cốt truyện của một bộ phim, tóm tắt của một cuốn sách), tuy nhiên, chúng tôi kết hợp thêm hình ảnh sản phẩm để tăng độ chính xác cho hệ khuyến nghị. Điều này được lý giải rằng khi tìm kiếm sản phẩm tương đồng, người dùng thường nhìn vào hình ảnh sản phẩm đầu tiên, thứ có thể trực quan sản phẩm lập tức với hình dạng, kích thước và màu sắc, điều mà tiêu đề sản phẩm hay mô tả không làm tốt như hình ảnh. Và khi kết hợp lại, những thuộc tính của sản phẩm được lấy ra một cách đầy đủ hơn để đánh giá, đảm bảo được những tính chất mà người dùng quan tâm như loại sản phẩm, thuộc tính, màu sắc, hình dáng, từ đó nâng cao hiệu suất cho hệ khuyến nghị. Hình 1 thể hiện nội dung đầu vào - đầu ra của bài toán mà chúng tôi xây dựng.

Chúng tôi cũng xây dựng bộ dữ liệu cho hệ khuyến nghị được thu thập từ sàn thương mại điện tử Lazada ba gồm 3141 sản phẩm với 5 danh mục dựa trên trang chủ của Lazada gồm: Thiết bị điện tử, Phụ kiện điện tử, Thời trang và phụ kiện nam, Thời trang và phụ kiện nữ, Thời trang và phụ kiện trẻ em. Với mỗi sản phẩm được thu thập các trường dữ liệu: tiêu đề sản phẩm, mô tả sản phẩm, hình ảnh, giá, lượt đánh giá, đường dẫn của sản phẩm, loại sản phẩm. Trong đó, đảm bảo 3 thuộc tính không được bỏ trống là: tiêu đề sản phẩm, mô tả sản phẩm và hình ảnh.

Chúng tôi thử nghiệm nhiều phương pháp khuyến nghị khác nhau trên văn bản (tiêu đề và mô tả sản phẩm) và hình ảnh sản phẩm, sau đó kết



Hình 1: Đầu vào - đầu ra của đề tài khuyến nghị mà chúng tôi xây dựng.

hợp lại với bộ trọng số phù hợp để tìm ra phương pháp tốt nhất. Trên hình ảnh sản phẩm, chúng tôi sử dụng 2 mô hình ResNet50 và ConvNeXT-small để embedding. Trên tiêu đề và mô tả sản phẩm, chúng tôi sử dụng mBERT và TF-IDF để tìm ra những đoạn mô tả có độ tương đồng cao nhất. Ngoài ra, để giảm bớt độ dài hoặc loại bỏ những phần không cần thiết, chúng tôi sử dụng mô hình ngôn ngữ lớn để thực hiện tóm tắt lại văn bản trên mô tả. Cuối cùng, kết quả thu được cao nhất là 52,79% trên mô hình ConvNeXT-small kết hợp với TF-IDF với bộ trọng số tiêu đề, mô tả được rút gọn, hình ảnh lần lượt là 0,75:0,2:0,25.

Trong nghiên cứu này, chúng tôi có một số đóng góp chính bao gồm: bộ dữ liệu cho hệ khuyến nghị sử dụng phương pháp content-based; xây dựng hệ khuyến nghị dựa trên nhiều thuộc tính gồm tiêu đề, mô tả và hình ảnh với bộ trọng số phù hợp; thực nghiệm và đánh giá hệ khuyến nghị dựa trên phương pháp contextual embedding và image embedding; sử dụng mô hình ngôn ngữ lớn để tạo ra những tóm tắt có ích trên mô tả để cải thiện hiệu suất.

Những phần trong báo cáo bao gồm: Phần 1 - Giới thiệu; Phần 2 - Các công trình liên quan; Phần 3 - Bộ dữ liệu. Phần 4 - Thiết kế thực nghiệm. Phần 5 - Kết quả, thảo luận. Phần 6 là Kết luận.

2 Các Công Trình Liên Quan

Trong lĩnh vực khuyến nghị dựa trên nội dung và hình ảnh, đã có nhiều nghiên cứu quan trọng đã đưa ra những đóng góp đáng kể. Các nghiên cứu này tập trung vào việc sử dụng thông tin văn bản, hình ảnh và đánh giá của người dùng để cải thiện chất lượng của hệ thống khuyến nghị và mang lại trải nghiệm mua sắm trực tuyến tốt hơn. Một số công trình có thể kể đến như: (Lops et al., 2011), (Iaquinta et al., 2008), (Elsayed et al., 2022), (McAuley et al., 2015), (Burke, 2002)

2.1 Dựa Trên Văn Bản

Có nhiều nghiên cứu khuyến nghị dựa trên nội dung văn bản, tập trung chủ yếu vào biểu diễn sparse hoặc dense như Word2Vec và Doc2Vec. Ahmed Elsafty và cộng sự (Elsafty et al., 2018) đã tiến hành các thử nghiệm sparse và dense vector trên tiêu đề công việc (job title), mô tả (job description) và kết hợp theo trọng số (job title-description). Nhóm tác giả đã chứng minh được việc sử dụng Doc2Vec mang lại hiệu quả cao hơn vượt trội so với Word2Vec và TF-IDF. Đặc biệt với dữ liệu kết hợp giữa tiêu đề và mô tả công việc theo trọng số.

Nghiên cứu của Takashi Yoneya và cộng sự (Yoneya and Mamitsuka, 2007) đã đề xuất một hệ thống khuyến nghị các bài báo hàng ngày theo

sở thích của người dùng được gọi là PURE: A PUbmed artical REcommendation, tác giả sử dụng mô hình xác suất (probabilistic model) với dữ liệu đầu vào là bài báo yêu thích của người dùng. Ngoài việc đề xuất các bài báo tương đồng có cụm từ chính (keywords), điểm đáng chú ý là mô hình có thể đề xuất các báo tương đồng không có keywords. Hệ thống đạt kết quả tốt với số lượng bài báo với số lượng yêu cầu đầu vào rất nhỏ (20 bài báo).

2.2 Dựa Trên Hình Ảnh

So với khuyến nghị dựa trên nội dung văn bản, khuyến nghị dựa trên hình ảnh có nhiều thách thức hơn. Một trong những vấn đề khá lớn gặp phải khi khuyến nghị dựa trên hình ảnh là hình ảnh có thể chứa không đầy đủ thông tin của sản phẩm, chủ yếu về hình thức và màu sắc, ví dụ khách hàng ưu tiên về giá cả hay kích thước của một sản phẩm hơn khi đi mua hàng. Li Yu và cộng sự (Yu et al., 2018) đã đề xuất phương pháp giải quyết vấn đề đó bằng việc sử dụng hình ảnh gốc (hình ảnh mà người dùng tìm kiếm) kết hợp với lịch sử hình ảnh người dùng đã chọn (click) khi tìm kiếm hình ảnh gốc để gợi ý hình ảnh. Hơn nữa, việc tính toán độ tương quan (similarity) của từng hình ảnh đối với nhau tốn rất nhiều thời, nên tác giả đã sử dụng phương pháp tính toán bằng cấu trúc cây (R-tree, SS-tree, B-tree) và sử dụng phương pháp mang tên "Query Samples" để tối ưu thời gian cũng như giảm dung lượng lưu trữ trong quá trình tính toán. Kết quả thực nghiệm mang lại tiềm năng khá cao khi đạt 92,4% đối với hình ảnh của váy và 88,1% đối với hình ảnh giày boot trên phương pháp đánh trọng số trên thuộc tính.

Trong bài báo (Elsayed et al., 2022), nhóm tác giả đã trình bày phương pháp và giới thiệu một phương pháp khuyến nghị trên hình ảnh đơn giản nhưng hiệu quả trên hai bộ dữ liệu Amazon fashion và Amazon men. Mô hình ImgRecEtE được nhóm tác giả đề xuất được lấy cảm hứng từ mô hình GraphRec, sau đó dữ liệu người dùng được mã hoá thành vector và kết nối với những tính năng khác của sản phẩm để thành một vector đầu vào duy nhất. Kết quả thu được cho thấy rằng hiệu suất được cải thiện 2,5% so với tập dữ liệu Amazon fashion và cải thiện 4,8% đối với tập dữ liệu Amazon men mà DVBPR báo cáo.

2.3 Kết Hợp Nhiều Loại Dữ Liệu Đầu Vào

Không chỉ khuyến nghị dựa trên những phương pháp riêng lẻ, trong nghiên cứu của Rafieian và cộng sự (Rafieian and Costa-jussà, 2020), nhóm tác giả đã trình bày phương pháp kết hợp cả hai yếu tố khuyến nghị dựa trên nội dung sản phẩm và khuyến nghị dựa trên những đánh giá của khách hàng, họ sử dụng khuyến nghị dựa trên nội dung để tìm ra những sản phẩm liên quan đến sản phẩm gốc và sử dụng khuyến nghị dựa trên đánh giá để tìm được những danh sách sản phẩm liên quan với nhau để tạo thành một tập hợp các sản phẩm hoàn chỉnh. Với khuyến nghị dựa trên nội dung, họ trình bày một phương pháp mới dựa trên câu hỏi khảo sát, sau đó tách ra thành một tập vector về thông tin sản phẩm, từ đó tạo nên một bảng các tính năng của các sản phẩm. Sau đó sử dụng kNN để đào tạo không giám sát trên tập dữ liệu. Đối với phương pháp đánh giá dựa trên lọc cộng tác, nhóm tác giả đã sử dụng lịch sử click và và phản hồi sản phẩm thay vì phương pháp truyền thống bởi vì quyền riêng tư của lịch sử người dùng.

Nhìn chung, các nghiên cứu trong lĩnh vực này đã làm sáng tỏ tiềm năng của việc tích hợp thông tin từ cả nội dung và hình ảnh để xây dựng hệ thống khuyến nghị mạnh mẽ và đa chiều trên sàn thương mại điện tử. Tuy nhiên, với sự tìm hiểu của chúng tôi, chưa có hệ thống nào kết hợp hình ảnh sản phẩm và văn bản mô tả sản phẩm để đưa ra khuyến nghị cho người dùng. Do đó trong báo cáo này, chúng tôi đề xuất một hệ thống khuyến nghị kết hợp cả hai thuộc tính hình ảnh và văn bản để có thể đưa ra hệ thống khuyến nghị gần gũi với người dùng hơn.

3 Dữ Liệu

3.1 Tổng Quan Về Bộ Dữ Liệu

Bộ dữ liệu chúng tôi xây dựng cần một số lượng lớn với sự đa dạng về nhiều danh mục. Do đó, chúng tôi chọn sàn thương mại điện tử Lazada để thu thập dữ liệu. Qua quá trình thu thập, chúng tôi có được bộ dữ liệu bao gồm 3141 sản phẩm với 5 danh mục: Thiết bị điện tử, Phụ kiện điện tử, Thời trang và phụ kiện nam, Thời trang và phụ kiện nữ, Thời trang và phụ kiện trẻ em với số lượng lần lượt là 632, 522, 199, 1098, 690 sản phẩm. Nhìn chung, các danh mục và sản phẩm được thu thập tập trung vào 2 lĩnh vực chính là điện tử và thời trang; nguyên nhân là phần mô tả của thiết bị và phụ kiện điện tử có nhiều điểm tương đồng (ví dụ điện thoại và ốp lưng của mã điện thoại đó có nhiều sự tương đồng về văn bản) nên danh mục này được chọn để kiểm tra tính nhập nhằng của thuật toán khuyến nghị; lĩnh vực thời trang là lĩnh vực cần sự trực quan cao, nên hình ảnh ở lĩnh vực này đa dạng, phong phú làm

cho việc đánh khuyến nghị dựa trên hình ảnh được trực quan .

Để có một nguồn dữ liệu đa dạng cho hệ khuyến nghị và các công việc tương lai, chúng tôi thu thập các trường dữ liệu sau cho mỗi sản phẩm: Tiêu đề (title), Mô tả sản phẩm (description), Hình ảnh (image), Giá (price), Điểm đánh giá trung bình (rating), Số lượt đánh giá (num_rating), Đường dẫn đến sản phẩm (link_product), Danh mục đầy đủ (category), ID của sản phẩm được đánh số sau khi thu thập (ID).

3.2 Thu Thập và Tiền Xử Lý Dữ Liệu

Để tăng hiệu suất thu thập dữ liệu, chúng tôi tiến hành thu thập dữ liệu từ Lazada bằng ngôn ngữ lập trình Python với sự hỗ trợ của thư viện beautifulsoup4 và selenium.

Đề tài hiện tại chúng tôi sử dụng tiêu đề, mô tả và hình ảnh sản phẩm để xây dựng hệ thống khuyến nghị, nên chúng tôi sẽ loại bỏ những sản phẩm thiếu một trong ba cột này. Trong trường dữ liệu hình ảnh (image), chúng tôi xóa các đoạn phim (video) bởi vì trong thực nghiệm chúng tôi chỉ xử lý trên hình ảnh, sau đó loại bỏ các hình ảnh trùng nhau. Trong trường dữ liệu tiêu đề (title) và mô tả (description), chúng tôi tiến hành loại bỏ các ký tự đặc biệt vì chúng thường không đóng góp nhiều vào ý nghĩa nội dung và có thể gây nhiễu loạn trong quá trình xử lý. Toàn bộ quá trình tiền xử lý dữ liệu về hình ảnh và văn bản đều thực hiện với sự hỗ trợ của ngôn ngữ lập trình Python. Cuối cùng, chúng tôi có bộ dữ liệu hoàn chỉnh để tiến hành các bước tiếp theo.

3.3 Xây Dựng Tập Dữ Liệu Đánh Giá

Chúng tôi lấy ngẫu nhiên 111 sản phẩm (3,5% trong mỗi danh mục của tập dữ liệu) với số lượng là 22, 18, 8, 39, 24 tương ứng với Thiết bị điện tử, Phụ kiện điện tử, Thời trang và phụ kiện nam, Thời trang và phụ kiện nữ, Thời trang và phụ kiện trẻ em và được gán nhãn bởi 3 người gán nhãn (annotators). Người gán nhãn sẽ xem xét toàn bộ sản phẩm trong tập dữ liệu và gán nhãn dựa theo 3 tiêu chí: tiêu đề, mô tả và hình ảnh sản phẩm; mức độ phù hợp trên 3 tiêu chí sẽ do người gán nhãn quyết định theo sự phù hợp với bản thân người gán nhãn, sau đó sắp xếp 5 sản phẩm phù hợp nhất với sản phẩm gốc. Do đó, tập dữ liệu đánh giá của chúng tôi có 2 trường dữ liệu: Mã của sản phẩm gốc (ID), tập hợp 5 ID của 5 sản phẩm liên quan (related_ID).

4 Thí Nghiệm

Quá trình thực nghiệm của chúng tôi được tiến hành qua 4 bước: tóm tắt đặc trưng mô tả bằng Gemini, mục đích để so sánh kết quả giữa việc tóm tắt đặc trưng mô tả có kết quả tốt hơn mô tả nguyên bản hay không; nhúng dữ liệu tiêu đề, mô tả, hình ảnh thành vector; sau đó tính toán độ tương đồng cosine giữa các vector vừa được tính toán và cuối cùng là sắp xếp chúng và đánh giá trên độ đo P@5.

4.1 Tóm tắt đặc trưng mô tả

Gemini[1] là mô hình AI mới nhất được Google giới thiệu vào tháng 12 năm 2023. Mô hình có khả năng xử lý nhiều loại dữ liệu đầu vào như văn bản, hình ảnh, âm thanh, video và mã (ngôn ngữ lập trình). Gemini có 3 phiên bản là Ultra, Pro và Nano. Trong báo cáo này, chúng tôi sử dụng Gemini phiên bản Pro để tóm tắt (summary) đặc trưng mô tả nhằm tạo ra một tóm tắt ngắn gọn nhưng vẫn giữ được thông tin quan trọng từ văn bản gốc. Điều này đặc biệt hữu ích khi làm việc với mô tả sản phẩm, vì nó giúp giảm bớt thông tin không cần thiết và tập trung vào những chi tiết quan trọng nhất.

Trong thực nghiệm, chúng tôi sử dụng đoạn lệnh (prompt) sau để yêu cầu Gemini tóm tắt văn bản trong phần mô tả:

```
prompt=f'''Create a summary of
    the following text with the
    same text language:"{text}"'''
```

Trong đó, text là đoạn mô tả. Bảng 1 là một ví dụ sử dụng mô hình Gemini để tóm tắt phần mô tả sản phẩm. Có thể thấy được mô hình đã giữ lại được những thông tin quan trọng, cô động nhất của phần mô tả.

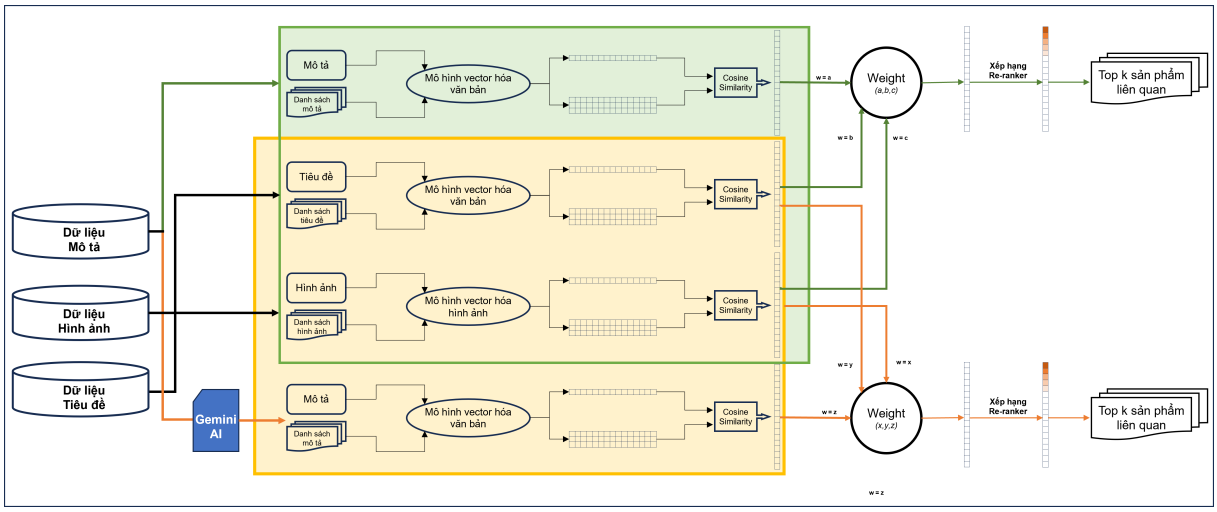
4.2 Nhúng Dữ Liệu

Trong thực nghiệm, chúng tôi sử dụng TF-IDF và Multilingual BERT để nhúng đầu vào dạng văn bản thành vector, đối với dữ liệu hình ảnh, chúng tôi sử dụng ResNet50 và ConvNeXT-small; sau đó dùng vector để tính toán độ tương đồng cosine ở bước tiếp theo. Dưới đây là mô tả về 4 phương pháp mà chúng tôi sử dụng.

4.2.1 TF-IDF

TF-IDF, viết tắt của Term Frequency-Inverse Document Frequency, là một phép đo thống kê được sử dụng rộng rãi trong xử lý ngôn ngữ tự nhiên để đánh giá mức độ quan trọng của một từ

[1] https://deepmind.google/technologies/gemini/



Hình 2: Quy trình thực nghiệm. Quy trình này gồm 4 bước: Bước 1, tóm tắt đặc trưng mô tả bằng Gemini; bước 2, Nhúng dữ liệu thành vector; bước 3, tính toán độ tương đồng cosine; bước 4, sắp xếp độ tương đồng và đánh giá.

trong một văn bản, tương đối so với một tập hợp các văn bản.

TF-IDF được tính bằng cách nhân hai chỉ số: Tần số xuất hiện của một từ trong một văn bản (Term Frequency - TF) và tần số nghịch đảo của từ đó trên tất cả các văn bản (Inverse Document Frequency - IDF). Công thức tính TF-IDF cho một từ t trong một văn bản d được biểu diễn như sau:

$$\text{tf-idf}(t, d) = \text{tf}(t, d) \times \text{idf}(t)$$

Trong đó, (tf(t, d)) là tần số của từ t trong văn bản d, và (idf(t)) được tính bằng cách lấy logarit tự nhiên của tổng số văn bản chia cho số văn bản chứa từ t.

4.2.2 Multilingual BERT

Multilingual BERT (mBERT) là một biến thể của mô hình BERT, được huấn luyện trên 102 ngôn ngữ có số lượng bài viết Wikipedia lớn nhất. Mô hình này không phân biệt chữ hoa và chữ thường. Mô hình mBERT sử dụng kiến trúc Transformer, một kiến trúc mạng nơ-ron sâu được giới thiệu trong bài báo "Attention is All You Need". Kiến trúc Transformer sử dụng cơ chế “attention” để xác định mối quan hệ giữa các từ trong câu, cho phép mô hình học được biểu diễn hai chiều của câu (Vaswani et al., 2023). Mô hình mBERT được huấn luyện với hai mục tiêu: Masked Language Modeling (MLM) và Next Sentence Prediction (NSP). Trong MLM, mô hình sẽ ngẫu nhiên che đi (mask) khoảng 15% số từ trong câu đầu vào. Sau đó, mô hình sẽ cố gắng dự đoán các từ đã bị che dựa trên ngữ cảnh của các từ còn lại trong câu. Trong NSP, mô hình sẽ nhận đầu vào là hai câu đã bị che. Hai câu này có thể là hai câu liên tiếp trong văn bản gốc, hoặc không. Mục tiêu của mô hình là dự đoán xem hai câu này có phải là hai câu liên tiếp trong văn bản gốc hay không. Trong báo cáo này, chúng tôi sử dụng mô hình bert-base-multilingual-uncased[2].

4.2.3 ResNet50

Convolutional Neural Network (CNN) là mô hình chuyên dùng xử lý ảnh, tuy nhiên, khi tăng độ sâu của mạng, CNN thường gặp phải vấn đề "exploding gradient", làm giảm hiệu suất của mô hình. Điều này dẫn đến việc tăng số lượng lớp không nhất thiết mang lại kết quả tốt hơn. Để khắc phục vấn đề này, ResNet (He et al., 2016), với việc sử dụng kết nối tắt (shortcut connections), đã mang lại một bước tiến đáng kể so với CNN. Công thức tính toán trong ResNet50 được biểu diễn như sau: H(x) := F(x) + x

Trong đó, (F(x)) là ánh xạ cần học và (x) là đầu vào. Mỗi lớp học một ánh xạ dư (F(x)) thay vì học ánh xạ gốc (H(x)).

ResNet50, một biến thể của mô hình ResNet, là một mạng CNN với 50 lớp, bao gồm 48 lớp tích chập (convolutional layer), một lớp MaxPool và một lớp Average Pool. Mô hình này đã được huấn luyện trên hơn một triệu hình ảnh từ cơ sở dữ liệu ImageNet. Kiến trúc của ResNet50 có thể được mô tả như Hình 3.

4.2.4 ConvNeXT-small

ConvNeXT-small, một biến thể của mô hình ConvNeXT, là một mạng nơ-ron tích chập sâu được giới thiệu bởi Zhuang Liu và cộng sự (Liu et al., 2022). Mô hình này đã được huấn luyện trên

[2]https://huggingface.co/bert-base-multilingual-uncased

Mô tả gốc
Sản phẩm được nhập trực tiếp từ nước ngoài. Hoá đơn bán hàng không được hỗ trợ trong trường hợp này. Lưu ý: 1-liên kết này chứa nhiều mẫu mã khác nhau, vui lòng chọn cẩn thận trước khi đặt hàng 2-tất cả các Model đều tương thích với ổ cứng/SSD 2.5 ", các Model có bộ đổi nguồn phù hợp với ổ cứng/SSD 2.5/3.5" 3-Bộ chuyển đổi đi kèm là Phích cắm EU, nếu bạn cần thông số kỹ thuật phích cắm khác, vui lòng liên hệ với dịch vụ khách hàng Sản phẩm Đặc Điểm Kỹ Thuật: Ổn định và nhanh chóng Cổng cấp nguồn Type-C 5V dành riêng, ổ đĩa cứng 6TB dễ dàng, truyền dữ liệu ổn định. 1G chuyển trong 3 giây USB3.0 đến SATA3.0 với giao thức UASP, tốc độ lên tới 5Gbps. Tốc độ thử nghiệm có thể đạt tới 480 mét/giây Tương thích rộng rãi Không có ổ đĩa, cắm và chạy, Windows, Linux, Mac OS, tương thích với Android Đặc tính sản phẩm ORICO Bộ Chuyển Đổi Ổ Đĩa HDD Cáp USB 3.0 Sang SATA Bộ Chuyển Đổi SATA Bộ Chuyển Đổi SATA Cho Ổ Đĩa Cứng Gắn Ngoài HDD/SSD 2.5” Thương hiệu Orico SKU 2017030214_VNAMZ-9390492631 Loại Bảo Hành Bảo hành bởi Nhà bán hàng quốc tế Thời gian bảo hành 18 tháng
Mô tả sau khi được tóm tắt
- Sản phẩm là bộ chuyển đổi ổ đĩa cứng HDD/SSD 2.5" từ USB 3.0 sang SATA. - Sản phẩm được nhập trực tiếp từ nước ngoài và không hỗ trợ hóa đơn bán hàng. - Có nhiều mẫu mã khác nhau, cần chọn cẩn thận trước khi đặt hàng. - Tương thích với ổ cứng/SSD 2.5", có bộ đổi nguồn phù hợp với ổ cứng/SSD 2.5/3.5". - Bộ chuyển đổi đi kèm là phích cắm EU, nếu cần thông số kỹ thuật phích cắm khác, vui lòng liên hệ với dịch vụ khách hàng. - Ổn định và nhanh chóng, cổng cấp nguồn Type-C 5V dành riêng, ổ đĩa cứng 6TB dễ dàng, truyền dữ liệu ổn định. - Tốc độ truyền dữ liệu lên tới 5Gbps, tốc độ thử nghiệm có thể đạt tới 480 mét/giây. - Tương thích rộng rãi, không cần ổ đĩa, cắm và chạy, Windows, Linux, Mac OS, tương thích với Android. - Bảo hành bởi Nhà bán hàng quốc tế trong thời gian 18 tháng.

Bảng 1: Một ví dụ đoạn mô tả sản phẩm trước và sau khi tóm tắt bằng Gemini.

hơn một triệu hình ảnh từ cơ sở dữ liệu ImageNet. ConvNeXT-small sử dụng kiến trúc “bottleneck” cho block, điều đó làm giảm số lượng tham số và phép nhân ma trận. Điều này cho phép việc huấn luyện từng lớp nhanh hơn. Tương tự với ResNet50, ConvNeXT-small cũng sử dụng shortcut connections trong mô hình. Hình 4 mô tả kiến trúc của mô hình ConvNeXT.

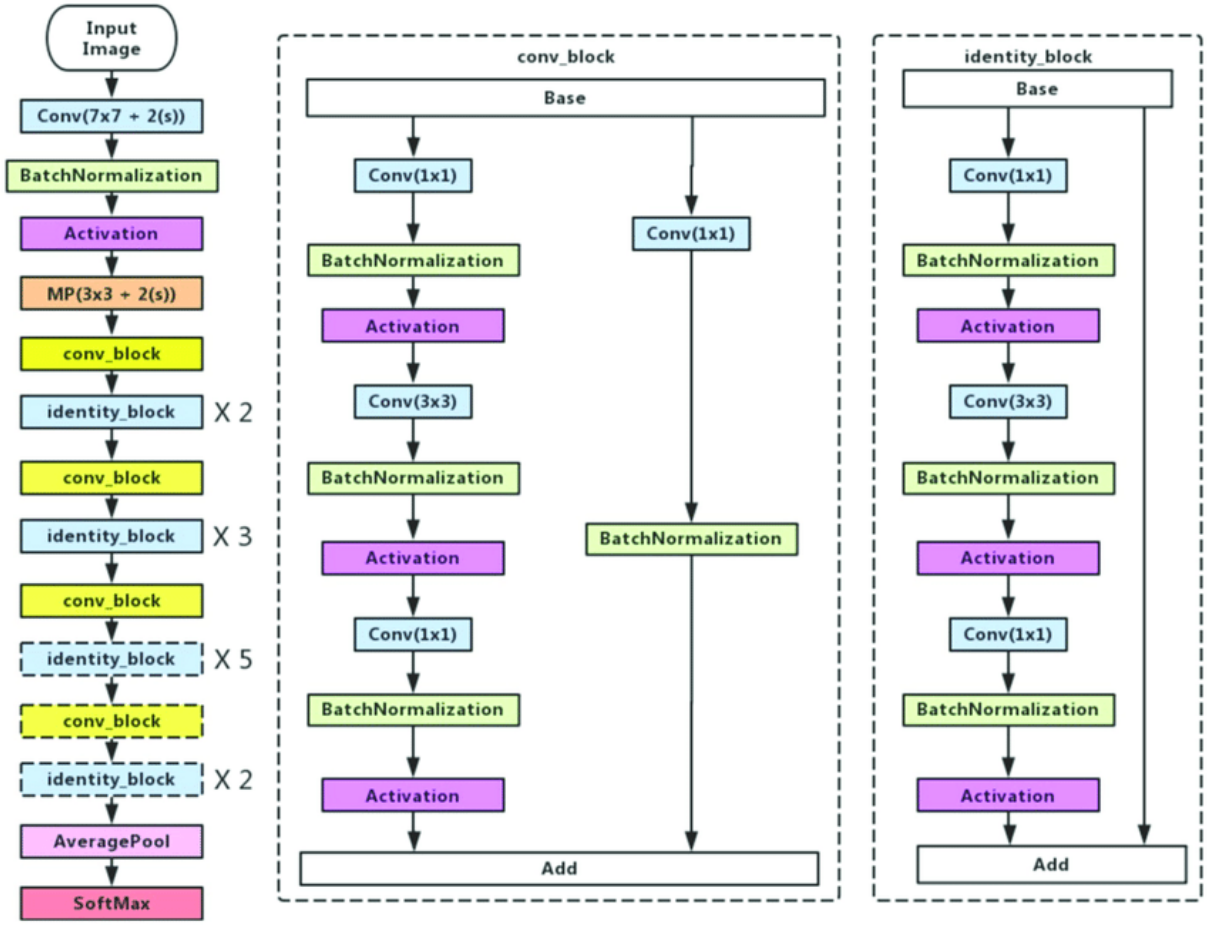
4.3 Tính Toán Độ Tương Đồng Cosine

Trong báo cáo này, chúng tôi sử dụng độ tương đồng Cosine để đánh giá mức độ tương đồng giữa hai vector đa chiều. Các vector này được biểu diễn từ dữ liệu văn bản hoặc hình ảnh thông qua quá trình chuyển đổi thích hợp.

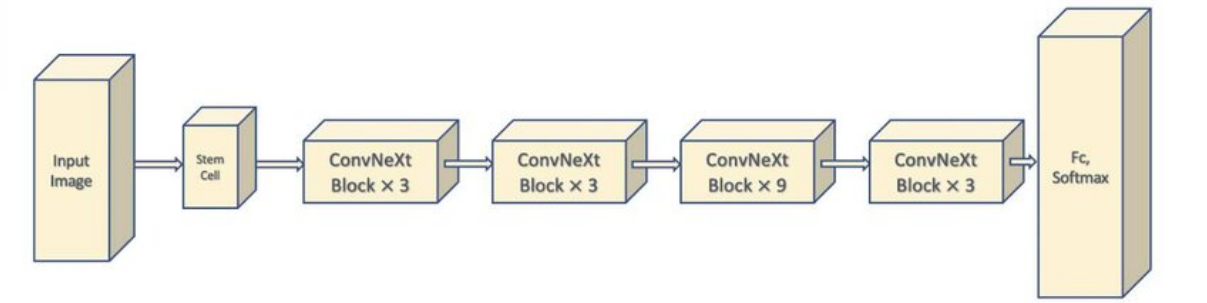
Độ tương đồng Cosine là một phép đo trong không gian vector, dựa trên góc giữa hai vector. Chỉ số này được tính toán bằng cách lấy cosin của góc giữa hai vector, lấy tích vô hướng của hai vector chia cho tích của độ dài của chúng. Công thức tính độ tương đồng Cosine giữa hai vector A và B trong không gian n chiều được biểu diễn như sau:

$$\text{Cosine similarity}(A, B) = \frac{\vec{A} \cdot \vec{B}}{\left\|\vec{A}\right\| \left\|\vec{B}\right\|} \quad (1)$$

Kết quả tương tự nằm trong khoảng từ -1 (hoàn



Hình 3: Kiến trúc mô hình ResNet50.



Hình 4: Kiến trúc mô hình ConvNeXT.

toàn ngược lại) đến 1 (hoàn toàn giống nhau), với 0 chỉ ra sự vuông góc hoặc không tương quan.

4.4 Độ đo đánh giá

Trong ngữ cảnh của hệ thống gợi ý, chúng ta thường quan tâm đến việc gợi ý N mục hàng đầu cho người dùng. Do đó, việc tính toán các chỉ số precision và recall trong N mục hàng đầu thay vì tất cả các mục hàng có ý nghĩa hơn. Trong báo cáo này, chúng tôi sử dụng độ đo P@K. Trong đó, K là số sản phẩm được cho là liên quan với sản phẩm kiểm tra (trong báo cáo này, chúng tôi sử dụng K=5).

$$\text{Precision@}k = \frac{\text{Số sản phẩm đúng}}{K} \quad (2)$$

Đặc trưng	Mô hình nhúng	P@5 (%)
Tiêu đề	TF-IDF	49.91
	mBERT	31.72
Mô tả	TF-IDF	36.58
	mBERT	21.80
Hình ảnh	ResNet50	0.18
	ConvNeXT-small	17.66

Bảng 2: Kết quả cao nhất của các đặc trưng khi sử dụng độc lập và chưa tóm tắt mô tả.

Đặc trưng	Mô hình nhúng	P@5 (%)
Mô tả	TF-IDF	29.36
	mBERT	16.76

Bảng 3: Kết quả của đặc trưng mô tả sau khi thực hiện tóm tắt.

5 Phân tích kết quả và thảo luận

5.1 Phân tích kết quả

5.1.1 Kết quả của từng đặc trưng

Qua Bảng 2 có thể thấy được dự đoán chỉ sử dụng tiêu đề có kết quả cao nhất (49,91% đối với TF-IDF), tiếp theo là mô tả (36,58% đối với TF-IDF) và thấp nhất là hình ảnh (17,66% đối với ConvNeXT). Qua đó chứng minh được những thông tin có được từ văn bản, và đặc biệt là tiêu đề vẫn chiếm một phần rất quan trọng trong việc khuyến nghị ra sản phẩm tương đồng. Tiêu đề thường chứa thông tin ngắn gọn và súc tích về nội dung của sản phẩm. Ví dụ, tiêu đề của một sản phẩm điện thoại có thể là "iPhone 14 Pro Max: 6GB RAM, 256GB ROM". Tiêu đề này cung cấp cho người dùng thông tin quan trọng về các tính năng chính của sản phẩm. Trong khi đó, mô tả chứa đầy đủ thông tin về sản phẩm hơn, nhưng lại có nhiều thông tin gây nhiễu, không liên quan như một đoạn giới thiệu, quảng cáo, hashtag. Hình ảnh có thể chứa một số đặc trưng nhất định như hình dạng, màu sắc, nhưng lại thiếu nhiều thông tin khác quan trọng về sản phẩm. Hơn nữa, hình ảnh cũng có nhiều thông tin gây nhiễu khác như nền ảnh, chữ viết quảng cáo, từ đó cũng gây giảm hiệu suất của phương pháp này.

Đối với các đặc trưng là văn bản, TF-IDF cho kết quả tốt hơn ở cả trên tiêu đề và mô tả. Mô hình mBERT cho kết quả thấp hơn, có thể giải thích điều này bởi chúng tôi không đào tạo lại mô hình mà sử dụng trực tiếp để trích xuất đặc trưng, vì vậy có thể dữ liệu có sẵn của mBERT không phù hợp với dữ liệu đặc trưng của bài toán này.

Ở hai mô hình nhúng cho hình ảnh, chúng ta có thể thấy mô hình ResNet50 có hiệu suất rất thấp, gần như là bằng 0, trong khi đó ConvNeXT cho hiệu khả quan hơn nhiều lần.

Sau khi tóm tắt mô tả dựa trên mô hình ngôn ngữ lớn, kết quả thu được lại kém hơn so với chưa tóm tắt (36,58% của chưa tóm tắt so với 29,36% so với đã tóm tắt trên phương pháp TF-IDF và 21,80% so với 16,76% trên phương pháp mBERT). Điều này có thể được lý giải rằng do mô hình ngôn ngữ lớn sinh ra tóm tắt chứa nhiều từ không có trong mô tả gốc, bị các lỗi về mặt ngữ nghĩa, từ vựng. Vì vậy, các mô tả được tóm tắt của các sản phẩm liên quan với nhau không còn có sự tương đồng cao như ban đầu. Bảng 4 là một ví dụ cho tóm tắt bị đổi ngôn ngữ so với mô tả gốc (một số từ tiếng Việt sang tiếng Anh).

5.1.2 Không áp dụng trọng số

Dựa vào Bảng 5 và Bảng 6, có thể thấy rằng, kết quả cao nhất khi chưa tóm tắt mô tả đạt 44,50% với phương pháp ConvNeXT-small và TF-IDF, trong khí đó kết quả đạt 45,23% với cùng phương pháp ConvNeXT-small - TF-IDF khi đã đã tóm tắt mô tả. Bên cạnh đó, ở cả 2 thí nghiệm, đã tóm tắt và chưa tóm tắt mô tả, sử dụng mô hình ConvNeXT-small để nhúng ảnh thành vector luôn cho kết quả cao hơn so với mô hình ResNet50, dùng TF-IDF để chuyển văn bản thành vector luôn cho kết quả cao hơn với sử dụng mBERT. Phương pháp sử dụng ResNet50 để nhúng ảnh và mBERT cho kết quả thấp nhất, thấp hơn đáng kể so với các phương pháp còn lại.

Đối với thí nghiệm sử dụng mô hình ConvNeXT-small, việc tóm tắt mô tả kết hợp với 2 đặc trưng còn lại làm tăng hiệu suất (tăng 0,73% đối với dùng TF-IDF, 1,09% đối với dùng mBERT). Trong khi đó, thí nghiệm sử dụng ResNet50 lại làm giảm hiệu suất tổng thể (giảm 2,71% đối với dùng TF-IDF, 0,24% đối với dùng mBERT).

5.1.3 Áp dụng trọng số

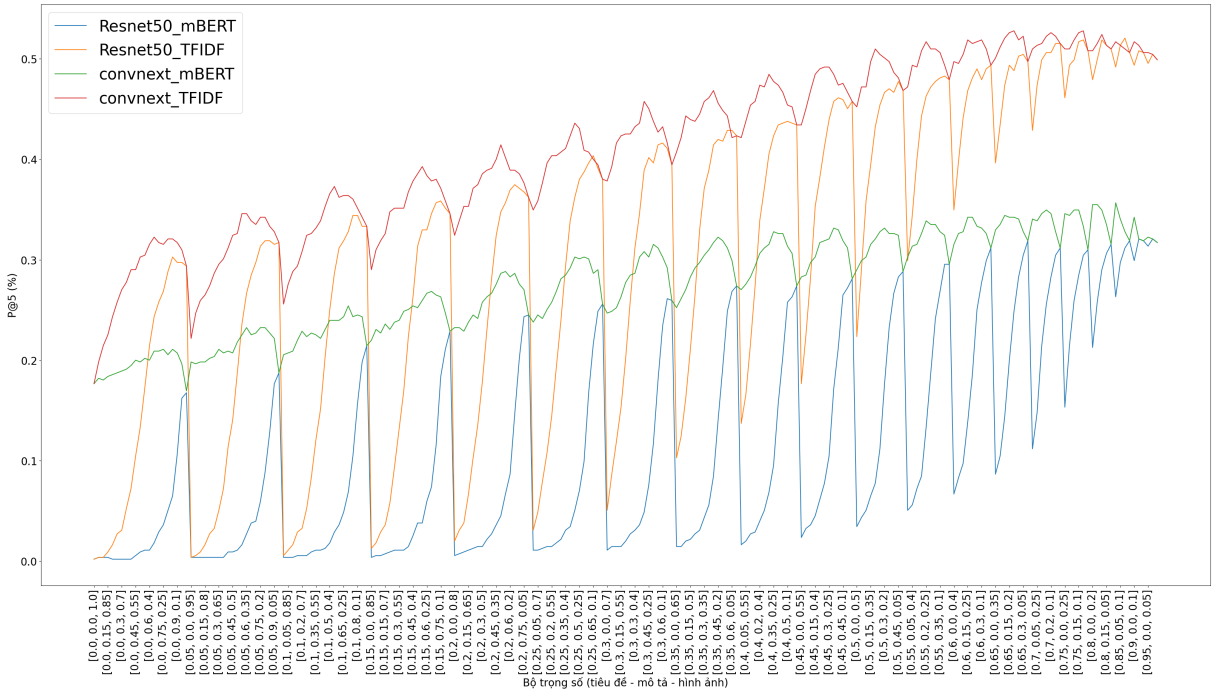
Dựa vào Hình 5 và Hình 6, kết quả cao nhất đối với kết hợp mô tả chưa tóm tắt đạt 52,07% tại bộ trọng số [tiêu đề:mô tả:hình ảnh] là [0,8:0,15:0,05] với ConvNeXT-small và TF-IDF, trong khi đó kết quả cao nhất khi kết hợp với mô tả đã tóm tắt là 52,79% với bộ trọng số [tiêu đề:mô tả:hình ảnh] là [0,75:0,2:0,05] trong kết hợp ConvNeXT-small và TF-IDF. Ở cả 2 thí nghiệm có tóm tắt và không tóm tắt mô tả, kết quả tăng dần khi tăng trọng số

Mô tả gốc Ram XSTAR rgb 8g bus 2666 Hãng sản xuất: XSTAR Chủng loại: RGB Gaming Dung lượng: 8GB Kiểu Ram: DDR4 Bus Ram hỗ trợ: 2666 Mhz Độ trễ: 16 CL(IDD) Tản nhiệt: Yes Điện áp: 1.2V Cam kết: Shop cam kết chỉ bán hàng chính hãng và tuyệt đối không bán các loại hàng fake (hàng loại 1, hàng công ty, hàng nhập khẩu, hàng ngoài, hàng sửa, hàng dựng... như các bên khác vẫn để nhằm qua mắt người dùng) những loại này thường không thể kiểm tra thông tin thương hiệu, serial number và part number trong tab SPD của phần mềm CPU-Z hoặc không kiểm tra được từ trang chủ của hãng. Hỗ trợ đổi lại cho khách trong trường hợp khách mua nhầm hàng và không phù hợp cho máy. Shop sẽ gửi đổi lại sản phẩm khác cho khách, khách hàng chỉ cần gói sản phẩm đó lại và đưa cho nhân viên giao hàng khi giao lại cái khác cho khách và shipper sẽ hoàn về cho shop. Đặc tính sản phẩm Ram PC DDR4 Xstar 8GB bus 2666 tản nhiệt đỏ bảo hành chính hãng 36 tháng Thương hiệu MAXSTAR SKU 2052476660_VNAMZ-9581307717 Thế Hệ Bộ Nhớ 34 rgb lighting Không Hệ Thống Bộ Nhớ 8GB Loại Bảo Hành Bằng Tem bảo hành
Mô tả sau khi được tóm tắt - Product Name: Ram XSTAR rgb 8g bus 2666 - Manufacturer: XSTAR - Category: RGB Gaming - Capacity: 8GB - Type: DDR4 - Supported Bus Speed: 2666 Mhz - Latency: 16 CL(IDD) - Heatsink: Yes - Voltage: 1.2V - Commitment: The shop only sells genuine products and absolutely does not sell fake goods. - Support: The shop supports exchanges in case the customer buys the wrong product and it is not suitable for the machine. - Product Features: - Brand: MAXSTAR - SKU: 2052476660_VNAMZ-9581307717 - Memory Generation: 34 rgb lighting - Memory System: 8GB - Warranty Type: Warranty Seal

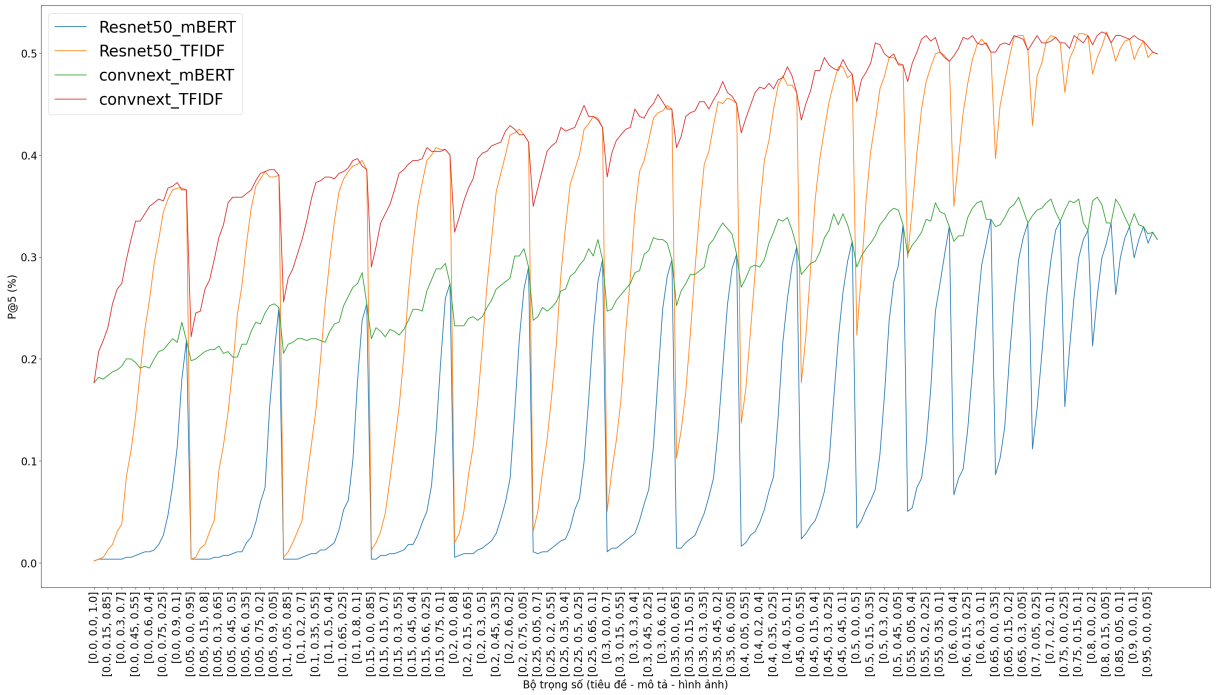
Bảng 4: Lỗi sai ngôn ngữ khi sử dụng Gemini để tóm tắt văn bản.

của tiêu đề. Tuy nhiên, khi chỉ áp dụng tiêu đề lại không có kết quả cao nhất, điều này chứng minh có sự ảnh hưởng của 2 đặc trưng còn lại. Ở mỗi trọng số của tiêu đề, đỉnh nằm ở giữa và lệch về phía bên phải. Do đó có thể kết luận, mô tả có ảnh hưởng lớn hơn hình ảnh.

Sự chênh lệch giữa phương pháp sử dụng mô hình ConvNeXT-small so với mô hình ResNet50 tăng dần khi tóm tắt và không tóm tắt mô tả. Dựa vào Bảng 3 và Bảng 2, có thể kết luận rằng nguyên nhân của điều sự chênh lệch này là do tóm tắt mô tả dẫn đến hiệu suất của đặc trưng mô tả giảm (giảm 7,22% khi dùng TF-IDF và 5,04% khi dùng mBERT), vì vậy, đặc trưng hình ảnh tăng sự ảnh hưởng đến kết quả tổng thể. Hơn nữa, hiệu suất của mô hình ConvNeXT-small cao hơn ResNet50 khi khuyến nghị chỉ dựa trên hình ảnh (cao hơn 17,4%).



Hình 5: Kết quả của thí nghiệm tóm tắt mô tả và có trọng số.



Hình 6: Kết quả của thí nghiệm chưa tóm tắt mô tả có trọng số.

5.1.4 Thảo luận

Vì sao sử dụng ResNet50 và mBERT lại cho kết quả thấp hơn đáng kể so với các phương pháp còn lại?

Thông qua thực nghiệm, chúng tôi nhận thấy hiệu suất thấp hơn đáng kể do kết hợp các nguyên nhân sau:

- Đầu tiên, hiệu suất khi dùng ResNet50 nhúng ảnh thấp hơn ConvNeXT-small đáng kể (thấp hơn 17,4%, Bảng 2 và Bảng 3).

- Thứ hai, chúng tôi chỉ sử dụng mô hình đào tạo trước mBERT để nhúng văn bản thành vector mà không huấn luyện lại trên miền dữ liệu này cho bài toán tương tự ngữ nghĩa của

Mô hình ảnh		Mô hình văn bản		P@5 (%)
ResNet50	ConvNeXT-small	TF-IDF	mBERT	
	✓	✓		45,23
✓		✓		36,93
	✓		✓	30,45
✓			✓	4,80

Bảng 5: Kết quả của hệ thống đối với thí nghiệm có tóm tắt mô tả.

Mô hình ảnh		Mô hình văn bản		P@5 (%)
ResNet50	ConvNeXT-small	TF-IDF	mBERT	
	✓	✓		44,50
✓		✓		39,64
	✓		✓	29,36
✓			✓	5,04

Bảng 6: Kết quả của hệ thống đối với thí nghiệm không tóm tắt mô tả.

văn bản (semantic textual similarity).

- Thứ ba, các mô tả và tiêu đề không phải là một đoạn văn, có ngữ cảnh cụ thể, do vậy, hiệu suất khi sử dụng các vector được mô hình mã hóa không tốt.

Vì sao cùng sử dụng mBERT nhưng ConvNeXT-small lại cho kết quả cao hơn đáng kể so với ResNet50?

Hình 5 và 6 cho thấy kết quả của hệ thống không chỉ dựa vào mô tả, mà có sự ảnh hưởng của hình ảnh. Trong khi đó, hiệu suất chỉ khi sử dụng hình ảnh để khuyến nghị đối với mô hình ConvNeXT-small cao hơn đáng kể so với ResNet50 (cao hơn 17.4%). Do vậy, hiệu suất tổng thể của ConvNeXT-small cao hơn đáng kể so với ResNet50.

6 Kết luận

Trong báo cáo này, chúng tôi đã xây dựng hệ thống khuyến nghị sản phẩm dựa trên vector mã hóa 3 đặc trưng: tiêu đề, mô tả và hình ảnh. Các mô hình nhúng ảnh gồm ConvNeXT-small và ResNet50, các phương pháp nhúng văn bản gồm: TF-IDF và mBERT. Hiệu suất cao nhất của hệ thống đạt 52,79% khi áp dụng trọng số cho 3 đặc trưng và đã tóm tắt mô tả với phương pháp ConvNeXT-small kết hợp TF-IDF. Thông qua các thực nghiệm, chúng tôi đưa ra một số nhận định sau: (1) Phương pháp khuyến nghị sản phẩm khi kết hợp tiêu đề, mô tả và hình ảnh cho hiệu suất tốt hơn so với từng đặc trưng độc lập. (2) Tiêu đề có ảnh hưởng lớn nhất, tiếp theo là mô tả và hình ảnh ít ảnh hưởng nhất khi khuyến nghị sản phẩm sử dụng phương pháp kết hợp 3 đặc trưng. (3) Áp dụng trọng số cho các đặc trưng mang lại hiệu suất tốt hơn khi không áp dụng trọng số. (4) Hiệu suất khi sử dụng ConvNeXT-small để nhúng ảnh tốt hơn đáng kê so với ResNet50. (5) Phương pháp TF-IDF đạt hiệu suất tốt hơn so với sử dụng mBERT để mã hóa mà không huấn luyện lại cho bài toán tương tự ngữ nghĩa văn bản. (6) Hiệu suất khuyến nghị khi chỉ sử dụng mô tả không tóm tắt tốt hơn so với sử dụng Gemini-Pro để tóm tắt.

Trong tương lai, Chúng tôi sẽ cải thiện hiệu suất thông quá các thực nghiệm sau: (1) Sử dụng kỹ thuật Named Entity Recognition (NER) để lấy đặc trưng, giảm nhiễu trong mô tả. (2) Huấn luyện các mô hình ngôn với bài tới semantic textual similarity để nâng chất lượng nhúng văn bản. (3) Sử dụng thêm các kỹ thuật xử lý ảnh. (4) Thử nghiệm mô hình tiên tiến hơn để cải thiện hiệu suất đối với ảnh.

References

Melania Berbatova. 2019. Overview on nlp techniques for content-based recommender systems for books. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 55–61.

Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370.

Ahmed Elsafty, Martin Riedl, and Chris Biemann. 2018. Document-based recommender system for job postings using dense representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 216–224.

Shereen Elsayed, Lukas Brinkmeyer, and Lars Schmidt-Thieme. 2022. End-to-end image-based fashion recommendation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Leo Iaquinta, Marco De Gemmis, Pasquale Lops, Giovanni Semeraro, Michele Filannino, and Piero Molino. 2008. Introducing serendipity in a content-based recommender system. In *2008 eighth international conference on hybrid intelligent systems*, pages 168–173. IEEE.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, pages 73–105.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Bardia Rafieian and Marta R Costa-jussà. 2020. E-commerce content and collaborative-based recommendation using k-nearest neighbors and enriched weighted vectors. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 1–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Takashi Yoneya and Hiroshi Mamitsuka. 2007. Pure: a pubmed article recommendation system based on content-based filtering. *Genome informatics*, 18:267–276.

Li Yu, Fangjian Han, Shaobing Huang, and Yiwen Luo. 2018. A content-based goods image recommendation system. *Multimedia Tools and Applications*, 77:4155–4169.