Cải tiến Hệ Thống Khuyến nghị Dựa Trên Nội Dung với Các Phương Pháp Embedding Hiện Đại

Nguyễn Lê Vy, Trần Thị Mỹ Duyên, Nguyễn Thị Mai Trinh
Trường Đại học Công Nghệ Thông Tin, Đại học Quốc gia TP Hồ Chí Minh, Việt Nam
{21522811, 21522017, 21522718}@gm.uit.edu.vn

Abstract

Trong bối cảnh dữ liệu số ngày càng bùng nổ, hệ khuyến nghị đã trở thành một công cụ đắc lực giúp người dùng tìm kiếm nội dung phù hợp giữa vô vàn thông tin. Tuy nhiên, "bài toán khởi đầu lạnh" *(cold start)* – khi mà dữ liệu hạn chế khiến việc gợi ý cho người dùng mới hoặc sản phẩm mới trở nên khó khăn – vẫn là một thách thức lớn cần giải quyết. Nghiên cứu này triển khai các phương pháp embedding từ cơ bản như TF-IDF, đến các các mô hình hiện đại dựa trên transformer như Sentence-T5 (ST5), Sentence-BERT (SBERT), Sentence-GPT (SGPT) trên tập dữ liệu Goodreads của Google. Chúng tôi sử dụng UMAP để giảm chiều dữ liệu cùng hai phương pháp phân cụm HDBSCAN và K-means (k=10), tạo thành 8 tổ hợp khác nhau để đánh giá hiệu quả trong nhiệm vụ khuyến nghị. Kết quả thực nghiệm cho thấy các mô hình transformer vượt trội hơn hẳn so với mô hình embedding cơ bản, với SGPT kết hợp UMAP + K-means(k=10) trên Dataset3 đạt Recall@50 = 0.4768, các tổ hợp khác của các mô hình transformer cũng cho kết quả tốt hơn rõ rệt so với kết quả khuyến nghị dựa trên TF-IDF. Những phát hiện này khẳng định tiềm năng của các phương pháp sentence-embedding hiện đại cùng các phương pháp giảm chiều, phân cụm trong việc nâng cao hiệu quả gợi ý.

1 GIỚI THIỆU

Trong thời đại số hóa, người dùng không chỉ mong đợi việc truy cập thông tin mà còn mong muốn thông tin tiếp nhận thật sự có ý nghĩa, phù hợp với sở thích và nhu cầu cá nhân. Từ khuyến nghị những bài hát yêu thích, những bộ phim thú vị cho đến những cuốn sách hay mỹ phẩm dưỡng da phù hợp,... các hệ khuyến nghị đã trở thành người bạn đồng hành không thể thiếu, mang lại sự tiện lợi và trải nghiệm cá nhân hóa vượt trội cho người sử dụng.

Tuy nhiên, không phải lúc nào các hệ thống khuyến nghị cũng vận hành trơn tru. Một trong những thách thức lớn nhất là "bài toán khởi đầu lạnh" *(cold start)*, khi mà dữ liệu người dùng hoặc sản phẩm mới quá ít ỏi để hệ thống đưa ra các gợi ý chính xác. Điều này khó khăn giống như việc đoán sở thích của một người hoàn toàn xa lạ. Để giải quyết vấn đề này, các hệ thống khuyến nghị dựa trên nội dung (content-based hay CB) được xem là giải pháp tiềm năng, nhờ khả năng phân tích đặc điểm của sản phẩm và so khớp chúng với sở thích người dùng, từ đó đưa ra các gợi ý phù hợp với nhu cầu người dùng.

Nghiên cứu này đặt mục tiêu khám phá và tối ưu hóa hiệu suất của các hệ thống khuyến nghị dựa trên nội dung. Thay vì chỉ dừng lại ở phương pháp embedding cơ bản như TFIDF, nhóm thử nghiệm cả các mô hình embedding tiên tiến dựa trên transformer, bao gồm Sentence-T5 (ST5), Sentence-BERT (SBERT) và Sentence-GPT (SGPT). Những phương pháp này không chỉ giúp chuyển đổi văn bản thành các vector mà còn giúp hệ thống tăng khả năng nắm bắt mối liên hệ giữa sản phẩm và người dùng, đặc biệt là trong bối cảnh dữ liệu hạn chế.

Để đảm bảo tính toàn diện, nghiên cứu áp dụng-phương pháp giảm chiều dữ liệu UMAP cùng hai kỹ thuật phân cụm: HDBSCAN và K-Means (k=10), thực nghiệm trên 3 bộ dữ liệu khác nhau (Dataset1, Dataset2, Dataset3 - được xây dựng từ tập dữ liệu nổi tiếng Goodreads của Google) tạo ra hai mươi tư tổ hợp khác nhau để thử nghiệm và đánh giá trên tác vụ khuyến nghị sách. Thực nghiệm được thực hiện với mục tiêu so sánh hiệu quả các phương pháp và xác định cách tiếp cận tối ưu nhất trong việc cải thiện hệ thống khuyến nghị.

Hình (1) minh họa quy trình xây dựng hệ thống khuyến nghị dựa trên nội dung của chúng tôi. Khi người dùng truy cập vào hệ thống, danh sách các cuốn sách mà họ đã đọc sẽ được hệ thống lưu trữ và xử lý thông qua mô-đun embedding, nhằm tạo ra embedding người dùng, đại diện cho sở thích và đặc trưng riêng của họ. Embedding này sau đó được so sánh với các embedding của các cuốn sách

khác trong cơ sở dữ liệu để tính toán mức độ tương đồng. Dựa trên kết quả đánh giá, hệ thống đề xuất top K cuốn sách có mức độ tương đồng cao nhất, phù hợp nhất với sở thích của người dùng. Những điểm nổi bật và đóng góp chính trong hệ thống của chúng tôi có thể được tóm tắt như sau:

- Thử nghiệm đa dạng các phương pháp embedding, từ truyền thống đến hiện đại, cùng phương pháp giảm chiều dữ liệu và gom cụm trên nhiệm vụ: khuyến nghị dựa trên nội dung.
- Cung cấp phân tích sâu sắc, đánh giá hiệu quả và hạn chế của từng phương pháp, đồng thời mở ra hướng nghiên cứu trong tương lai.

Bài báo cáo này được chia thành 6 phần chính. Phần 1 giới thiệu tổng quan về đề tài. Phần 2 tổng hợp các công trình nghiên cứu liên quan, cung cấp nền tảng lý thuyết về hệ khuyến nghị và các phương pháp embedding. Bộ dữ liệu nhóm sử dụng - Goodreads và quy trình xử lý dữ liệu được mô tả chi tiết trong phần 3. Phần 4 trình bày chi tiết quy trình, kết quả thực nghiệm và phân tích. Sản phẩm minh họa cho bài toán được trình bày trong phần 5. Cuối cùng, tại phần 6, chúng tôi đưa ra kết luận, chỉ ra hạn chế, và đề xuất các hướng phát triển trong tương lai.

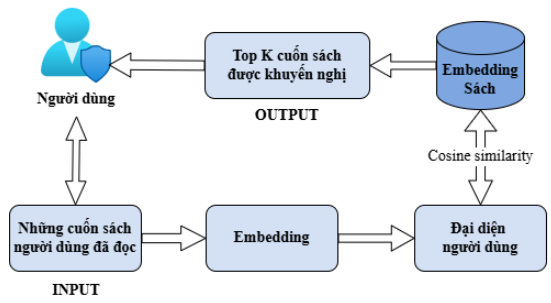

Figure 1: Quy trình hệ khuyến nghị dựa trên nội dung

Với cách tiếp cận này, nghiên cứu không chỉ dừng lại ở việc xây dựng và cải thiện các hệ thống khuyến nghị mà còn đưa ra những góc nhìn mới mẻ và đầy hứa hẹn trong lĩnh vực này.

2 CÔNG TRÌNH LIÊN QUAN

Trong kỷ nguyên hiện đại, việc dựa vào đánh giá của người dùng khác đã trở thành một xu hướng phổ biến để khám phá sản phẩm và nội dung yêu thích. Từ việc mua sắm trên các sàn thương mại điện tử cho đến chọn lựa phim ảnh hay sách vở, chúng ta thường xem xét các "review" như một dạng khuyến nghị quan trọng. Những đánh giá này không chỉ giúp tiết kiệm thời gian mà còn mang lại trải nghiệm cá nhân hóa, giúp người dùng đưa ra quyết định sáng suốt hơn (Resnick and Varian, 1997). Đây cũng chính là lý do vì sao các hệ thống gợi ý (Recommender Systems - RS) ngày càng khẳng định vai trò thiết yếu trong đời sống số.

Các nghiên cứu đã minh chứng rõ ràng hiệu quả của RS trong nhiều lĩnh vực. Ví dụ, hệ thống gợi ý top-K được xây dựng trên YouTube (Chen et al., 2019) đã cải thiện đáng kể khả năng cá nhân hóa nội dung cho người dùng, trong khi nghiên cứu của (Xu and Hu, 2023) ứng dụng các mô hình ngôn ngữ tiên tiến để nâng cao chất lượng gợi ý sản phẩm trong thương mại điện tử. Những thành tựu này nhấn mạnh vai trò không thể thiếu của RS trong việc tạo ra trải nghiệm tối ưu cho người dùng.

Trong số các phương pháp RS, Collaborative Filtering (CF) nổi bật như một kỹ thuật phổ biến nhất (Yang et al., 2016). CF hoạt động dựa trên sự tương đồng giữa sở thích và mối quan tâm của người dùng với những người dùng khác có chung quan điểm trong quá khứ. Tuy nhiên, CF phải đối mặt với một bài toán lớn: vấn đề khởi động lạnh (*cold-start problem*), tức là thiếu thông tin từ người dùng và sản phẩm (Hasan and Khatwal, 2022). Nhiều nghiên cứu đã nỗ lực khắc phục vấn đề này, chẳng hạn (Nguyen et al., 2014) sử dụng xếp hạng của người dùng làm thông tin ngữ cảnh và đề xuất thuật toán LinUCB. Một giải pháp khác là khuyến nghị xuyên miền (cross-domain recommendation) của (Bi et al., 2020), nơi thông tin được chuyển từ một miền nguồn sang miền đích để giảm thiểu tình trạng khan hiếm dữ liệu. Cách tiếp cận này kết hợp cơ chế xuyên miền và mạng thông tin không đồng nhất để mang lại các khuyến nghị cá nhân hóa cho người dùng mới trong lĩnh vực bảo hiểm.

Phần lớn các nghiên cứu trước đây tập trung vào kỹ thuật nhúng từ (word embedding), mã hóa siêu dữ liệu sản phẩm bằng cách lấy trung bình các từ riêng lẻ để tạo thành nhúng câu (sentence embedding). Một số nghiên cứu nổi bật như (Mediani et al., 2023) áp dụng nhúng từ để nâng cao chất lượng gợi ý nội dung giáo dục, hay (Birunda and Devi, 2021) khám phá cách sử dụng nhúng từ theo ngữ cảnh. Tuy nhiên, nhúng câu (sentence embedding) đang dần chiếm ưu thế nhờ khả năng nắm bắt toàn bộ ý nghĩa của văn bản, giúp bảo toàn ngữ cảnh tốt hơn. Ví dụ, (Wang and Kuo, 2020) đã giới thiệu phương pháp nhúng hai câu với cả mô hình không tham số và có tham số. Nhiều nghiên cứu còn tích hợp BERT để tạo nhúng câu, như công

trình của (Cygan, 2021) và (Juarto and Girsang, 2021). Đặc biệt, (Mendes de Melo et al., 2022) đã phát triển hệ thống khuyến nghị hỗ trợ COVID-19, sử dụng nhúng câu để đảm bảo chất lượng gợi ý trong các cuộc trò chuyện qua chat.

Mục tiêu báo cáo của chúng tôi là khai thác sức mạnh của nhúng câu (sentence embedding) kết hợp với các mô hình học sâu như BERT, T5, GPT để nâng cao chất lượng gợi ý trong các hệ thống khuyến nghị. Bằng cách tận dụng khả năng nắm bắt ngữ cảnh toàn diện của nhúng câu, nhóm không chỉ mong muốn cải thiện độ chính xác của các gợi ý mà còn đưa ra giải pháp hiệu quả cho bài toán khởi động lạnh (*cold-start*), giúp hệ thống đáp ứng tốt hơn với người dùng và sản phẩm mới.

3 BỘ DỮ LIỆU

Nghiên cứu của chúng tôi tập trung vào việc xây dựng hệ thống Khuyến nghị Sách dựa trên nội dung. Chúng tôi thực nghiệm bài toán trên bộ dữ liệu Goodreads của Google, tập trung vào danh mục "truyện tranh và đồ họa" (comic-graphic), từ đó đánh giá và cải thiện hiệu suất của hệ thống khuyến nghị. Tập dữ liệu này cung cấp một lượng thông tin phong phú về sách, người dùng và đánh giá, mang lại một góc nhìn sâu sắc và thực tiễn cho nghiên cứu của chúng tôi.

Để nâng cao hiệu quả của hệ thống khuyến nghị, chúng tôi đã sử dụng các kỹ thuật tiền xử lý văn bản như chuyển đổi chữ cái thành chữ thường, tách từ (tokenization), loại bỏ stopword, loại bỏ các khoảng trắng dư thừa và và chuẩn hóa từ gốc (stemming/ lemmatization).

Sau khi tiền xử lý, nhóm thu được hai bộ dữ liệu: MetaBooks và UsersHistory. Trong đó:

MetaBooks: bao gồm siêu dữ liệu của sách như tiêu đề, tác giả và mô tả.

UsersHistory: chứa các tương tác giữa người dùng và sách, bao gồm user_id, book_id và đánh giá người dùng cho sản phẩm sách từ 1 đến 5.

Sau đó, nhóm kết hợp hai bảng dữ liệu MetaBooks và UsersHistory, tạo thành bộ dữ liệu mới gồm 5 thuộc tính: user-id, book-id, rating, text-feature, và index. Trong đó, giá trị text-feature được tạo theo 3 công thức khác nhau, tương ứng 3 bộ dữ liệu cho bài toán:

- **Dataset1**: text-feature là giá trị thuộc tính description trong bảng MetaBooks.
- **Dataset2**: text-feature là giá trị kết hợp của hai thuộc tính description và title trong bảng MetaBooks.
- **Dataset3**: text-feature là giá trị kết hợp của ba thuộc tính description, title, và author trong bảng MetaBooks.

Thống kê của 3 bộ dữ liệu: Dataset1 (description), Dataset2 (Description+Title), Dataset3 (Description+Title+Author) được trình bày trong bảng 1. Các tập dữ liệu này đóng vai trò quan trọng trong việc đánh giá và cải thiện hệ thống gợi ý, đặc biệt trong các bài toán gợi ý dựa trên nội dung.

Tập trung phân tích bộ dữ liệu Dataset3, ta nhận thấy phân phối độ dài của cột text_features (xem hình 2) có độ lệch phải, với một đuôi dài mở rộng về phía các tài liệu dài hơn, độ dài của cột text_features thường tập trung trong khoảng 80-140 từ. Cột text_features có phân phối Zipfian (Newman, 2005), (xem hình 3) về tần suất từ, có nghĩa là có một vài từ xuất hiện rất thường xuyên và nhiều từ khác xuất hiện chỉ một vài lần. Nhóm cũng xuất ra 50 từ xuất hiện nhiều nhất trong thuộc tính text_features (xem hình 4).

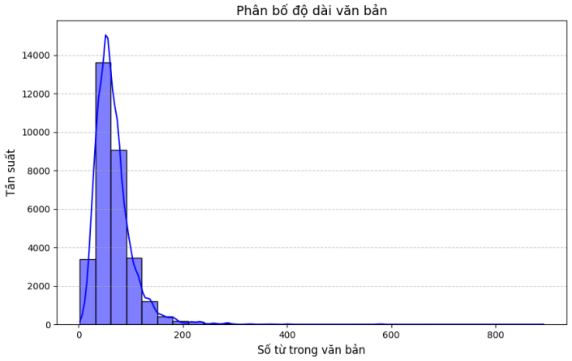

Figure 2: Phân phối độ dài tài liệu-Dataset3

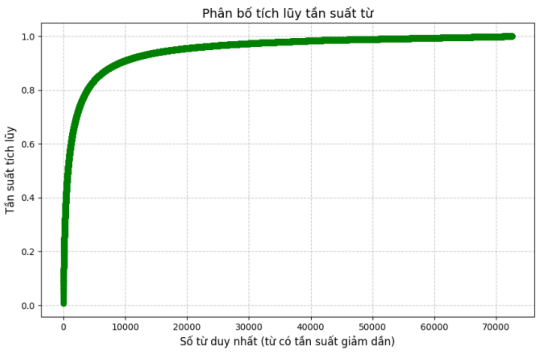

Figure 3: Phân phối tích lũy tần suất từ-Dataset3

Thông tin	Dataset1	Dataset2	Dataset3
Tổng số tài liệu	74,125	74,125	74,125
Kích thước từ vựng	289,425	299,379	302,159
Số từ tối đa trong tài liệu	888	890	892
Số từ tối thiểu trong tài liệu	1	1	3
Số lượng từ trung bình mỗi tài liệu	64.48	66.17	68.2

Table 1: Thống kê bộ dữ liệu: Dataset1, Dataset2, Dataset3

Figure 4: 50 từ xuất hiện nhiều nhất-Dataset3

4 THỰC NGHIỆM VÀ KẾT QUẢ

4.1 Các phương pháp

Nghiên cứu này so sánh phương pháp embedding cổ điển - TFIDF với các embedding kiến trúc hiện đại dựa trên transformer. Các mô hình embedding hiện đại này thường mang lại hiệu suất cao hơn nhiều trong các tác vụ NLP như: phân loại văn bản, nhận dạng thực thể, và trả lời câu hỏi,... Bài nghiên cứu cũng khám phá ưu điểm phương pháp embedding cổ điển- TFIDF, đặc biệt là trong các ứng dụng yêu cầu tính toán nhanh và dễ dàng triển khai. Chi tiết về các phương pháp được sử dụng trong báo cáo được cung cấp dưới đây.

TF-IDF (Term Frequency - Inverse Document Frequency) là một kỹ thuật phổ biến trong truy hồi thông tin, được sử dụng để gán trọng số cho các từ dựa trên mức độ quan trọng của chúng trong tài liệu. TF-IDF kết hợp hai yếu tố: tần suất xuất hiện của từ trong tài liệu (TF) và sự khan hiếm của từ trong toàn bộ tập dữ liệu (IDF). Phương pháp này giúp xác định từ khóa đặc trưng cho tài liệu bằng cách gán trọng số cao hơn cho các từ xuất hiện thường xuyên trong một tài liệu cụ thể nhưng lại hiếm gặp trên toàn bộ bộ sưu tập. Điều này làm nổi bật những từ mang tính đặc trưng thay vì các từ thông dụng. Theo một nghiên cứu (Beel et al., 2015), có tới 83% các hệ thống gợi ý dựa trên văn bản trong thư viện kỹ thuật số ứng dụng TF-IDF. Điều này minh chứng rõ ràng cho hiệu quả vượt trội của TF-IDF trong việc tổ chức và tìm kiếm thông tin, đặc biệt trong các hệ thống xử lý văn bản lớn.

Sentence-T5 (ST5): (Ni et al., 2022) là phương pháp tạo nhúng câu tiên tiến dựa trên mô hình Text-to-Text Transfer Transformer (T5), một mô hình mạnh mẽ đã được huấn luyện trước cho các tác vụ xử lý ngôn ngữ tự nhiên. ST5 mở rộng mô hình T5 (Raffel et al., 2019) để tạo ra các nhúng câu, mang đến một phương pháp đồng nhất và hiệu quả để mã hóa câu thành các đại diện vector dày đặc.

Sentence-BERT (SBERT): (Reimers and Gurevych, 2019) là viết tắt của Sentence Embeddings using Siamese BERT-Networks, là một phương pháp tạo các embedding có ý nghĩa cho câu sử dụng mô hình ngôn ngữ mạnh mẽ BERT. Khác với BERT (Devlin et al., 2019), vốn tập trung vào các từ đơn lẻ, SBERT nhắm đến việc nắm bắt ý nghĩa ngữ nghĩa tổng thể của một câu trong một vector dày đặc. Phương pháp này giúp cải thiện đáng kể hiệu suất trong các tác vụ yêu cầu hiểu câu, chẳng hạn như đo lường độ tương đồng câu hoặc phân loại câu.

GPT Sentence Embeddings (SGPT): (Muennighoff, 2022) Mô hình GPT (Brown et al., 2020) là một công cụ xử lý ngôn ngữ mạnh mẽ, được huấn luyện trên một lượng dữ liệu văn bản khổng lồ để tạo ra văn bản tự nhiên và nắm bắt chính xác các sắc thái ngữ nghĩa cũng như cú pháp. SGPT, như một kỹ thuật nhúng câu, tận dụng các ưu điểm của GPT để mã hóa câu thành các vector dày đặc, phản ánh chính xác ý nghĩa ngữ nghĩa. Với khả năng tạo ra các văn bản tự nhiên, hiểu ngữ cảnh và biểu diễn tương đồng câu hiệu quả, SGPT mang lại cải tiến đáng kể trong các ứng dụng như tìm kiếm thông tin, phân tích cảm xúc và gợi ý nội dung.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP): (McInnes et al., 2020) là một kỹ thuật giảm chiều được sử dụng cho việc trực quan hóa tương tự như t-SNE, nhưng cũng có thể được sử dụng cho giảm chiều phi tuyến tính tổng quát. UMAP có thể duy trì cấu trúc dữ liệu trong không gian có số chiều nhỏ một cách hiệu quả, giúp trực quan hóa các dữ

liệu lớn và phức tạp, đồng thời giảm thiểu mất mát thông tin.

HDBSCAN: là một thuật toán phân cụm được phát triển bởi (Campello et al., 2013). Nó mở rộng DBSCAN (Ester et al., 1996) bằng cách chuyển nó thành thuật toán phân cụm phân cấp, sau đó sử dụng một kỹ thuật để trích xuất một phân cụm phẳng dựa trên độ ổn định của các cụm. HDBSCAN giúp giải quyết vấn đề phân cụm dữ liệu không đồng nhất và có khả năng xử lý dữ liệu có độ nhiễu cao.

K-Means (Jin and Han, 2010): là một thuật toán phân cụm phổ biến được sử dụng trong phân tích dữ liệu. Nó nhóm các điểm dữ liệu tương tự vào 'k' cụm dựa trên các đặc trưng của chúng, nhằm giảm thiểu phương sai trong các cụm. Thuật toán này có thể ứng dụng rộng rãi trong phân tích dữ liệu lớn, phân loại và trích xuất thông tin. Thuật toán Kmeans sẽ phân bổ các điểm dữ liệu vào trung tâm cụm gần nhất và cập nhật lại trung tâm cho đến khi hội tụ. Kết quả là các cụm có sự đồng nhất cao và có thể được sử dụng cho các bài toán phân loại, dự báo và tìm kiếm.

4.2 Quy trình thực nghiệm

Hệ thống khuyến nghị dựa trên nội dung (Content-based hay CB) sử dụng thông tin về đặc điểm của sản phẩm hoặc dịch vụ để gợi ý các mục tương tự cho người dùng. Bằng cách phân tích các thuộc tính như mô tả, thể loại hoặc từ khóa, hệ thống tạo ra các vector đại diện cho từng sản phẩm và tính toán sự tương đồng giữa chúng để đưa ra các đề xuất phù hợp với sở thích người dùng.

Ưu điểm của phương pháp này là tính cá nhân hóa cao và khả năng giải thích rõ ràng về lý do các sản phẩm được đề xuất. Tuy nhiên, nhược điểm rất lớn của Hệ thống Khuyến nghị dựa trên nội dung (CB) là khó đề xuất các sản phẩm mới hoặc đa dạng nếu người dùng chỉ tương tác với một nhóm sản phẩm nhất định.

Chúng tôi xây dựng một hệ thống Khuyến nghị dựa trên nội dung với quy trình như mô tả trong hình 5. Trong giai đoạn nhúng (embedding), đầu tiên, các đặc trưng văn bản từ các cuốn sách được nhúng vào các vector dày đặc bằng những phương pháp embedding: TF-IDF, ST5, SBERT, SGPT (đã được thảo luận trong phần 4.1). Tiếp theo, chúng tôi áp dụng hai tổ hợp UMAP + HDBSCAN và UMAP + KMeans (k=10) để giảm chiều và phân cụm các cuốn sách vào các nhóm tương ứng. Cuối cùng, chúng tôi lưu trữ các embedding của người dùng cùng với các cuốn sách và nhãn cụm của chúng vào bộ dữ liệu, phục vụ cho các giai đoạn đánh giá sau này.

Để đánh giá hệ thống trên tác vụ khuyến nghị, chúng tôi tái sử dụng các embedding sách đã được lưu trữ từ giai đoạn embedding. Sau đó, bộ dữ liệu được chia thành các tập huấn luyện và kiểm tra theo người dùng theo tỉ lệ 70:30. Đối với mỗi người dùng, chúng tôi xem xét các cuốn sách mà họ đã mua. Cụ thể, 70% các cuốn sách đầu tiên mà người dùng mua sẽ được dùng để áp dụng các phương pháp embedding. Sau đó, chúng tôi thực hiện pooling trung bình để tạo ra một đại diện người dùng (embedding người dùng). Tiếp theo, độ tương đồng cosine giữa đại diện người dùng và các embedding của tất cả các cuốn sách trong cơ sở dữ liệu được tính toán và sắp xếp theo thứ tự giảm dần. Cuối cùng, chúng tôi so sánh k cuốn sách hàng đầu (với K có thể là 5, 10 hoặc 50) với 30% các cuốn sách trong tập kiểm tra, dựa trên các nhóm phân loại trong giai đoạn nhúng. Cuối cùng, hiệu suất hệ khuyến nghị được đánh giá thông qua chỉ số recall@K (Cụ thể, trong bài này, chúng tôi sử dụng các chỉ số recall@5, recall@10, recall@50). Quy trình đánh giá tác vụ Khuyến nghị được mô tả chi tiết trong hình 6.

Ngoài việc so sánh hiệu suất của các phương pháp embedding, thực nghiệm của nhóm còn được thiết kế để chạy trên ba bộ dữ liệu thông tin sách khác nhau, phương pháp giảm chiều UMAP, kết hợp với hai phương pháp gom cụm là HDBSCAN và Kmeans. Việc thử nghiệm hệ khuyến nghị trên ba bộ thông tin sách khác nhau nhằm mục đích tìm hiểu và phân tích ảnh hưởng của dữ liệu đầu vào đối với hiệu suất của hệ thống khuyến nghị. Đối với tác vụ phân cụm, ngoài HDBSCAN, chúng tôi còn sử dụng K-Means cho phân cụm vì HDBSCAN không thể xác định số lượng chủ đề cụ thể. Mỗi phương pháp phân cụm đều có những ưu điểm và nhược điểm riêng, do đó, nhóm thực nghiệm nhằm tìm ra phương pháp gom cụm phù hợp nhất với bài toán và bộ dữ liệu đang sử dụng.

Tổng quan, các thực nghiệm của nhóm được thiết kế để đưa ra phân tích sâu sắc về ảnh hưởng của dữ liệu đầu vào, so sánh hiệu suất của hai phương pháp gom cụm là HDBSCAN và KMeans (k=10), đồng thời đánh giá hiệu suất của các phương pháp embedding trong tác vụ khuyến nghị sách.

4.3 Cài đặt thực nghiệm

Chúng tôi không quy định độ dài chuỗi tối đa cho các phương pháp TFIDF do các phương pháp này

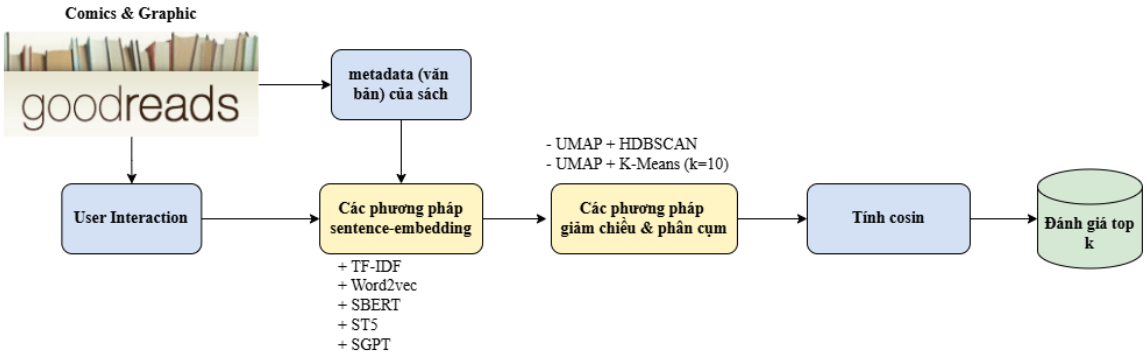

Figure 5: Quy trình thực nghiệm tổng quan

tạo embedding bằng cách lấy trung bình các từ trong câu. Ngược lại, các phương pháp khác nhúng toàn bộ câu. Đáng chú ý, độ chiều của TFIDF lớn hơn đáng kể so với các phương pháp khác. Tuy nhiên, dựa trên số lượng từ vựng của bộ dữ liệu và nguyên lý hoạt động của TFIDF, chúng tôi chọn giá trị 8000 để biểu diễn hiệu quả.

Cấu hình các phương pháp Embedding và Cấu hình các phương pháp giảm chiều phân cụm được trình bày chi tiết trong Bảng 2, 3.

Phương pháp	**Độ dài chuỗi tối đa**	**Kích thước embedding**	**Cấu hình**
TFIDF	Không giới hạn	8000	min_df=4, max_df=200, max_features=8000
ST5	512	768	pretrained: sentence-t5-base
SBERT	512	768	pretrained: bert-base-uncased, bert-base-multilingual-uncased
SGPT	512	768	pretrained: SGPT-125M-weighted-mean-nli-bitfit

Table 2: Cấu hình các phương pháp Embedding

4.4 Độ đo đánh giá

Để đánh giá hiệu suất các phương pháp embedding trên nhiệm vụ gợi ý, nhóm sử dụng độ đo *Recall*. Chỉ số Recall được tính dựa trên số lượng sách đã được người dùng tiêu thụ trong danh sách gợi ý so với tổng số sách mà người dùng đã tiêu thụ. Chỉ số này được gọi là Recall@k, trong đó k biểu thị kích thước danh sách gợi ý.

	Cấu hình
UMAP	n_neighbors=15, n_components=5, metric=cosine
HDBSCAN	min_cluster_size=10, metric=euclidean, cluster_method=eom
K-Means	n_clusters = [5,7,9], k = 10

Table 3: Cấu hình các phương pháp giảm chiều và phân cụm

$$\text{Recall@k} = \frac{S_{\text{relevant-in-top-K}}}{S_{\text{total-relevant}}}$$

Trong đó:

- $S_{\text{relevant-in-top-K}}$ là số lượng sách liên quan trong top K gợi ý.
- $S_{\text{total-relevant}}$ là tổng số sách liên quan trong toàn bộ bộ dữ liệu.

Chúng tôi định nghĩa sách liên quan là những sách có đánh giá từ 3 trở lên trong bộ dữ liệu (rating >= 3).

Trong bài này, nhóm sử dụng 3 độ đo là recall@5, recall@10 và recall@50 để đánh giá hiệu suất của hệ khuyến nghị.

4.5 Kết quả thực nghiệm và Phân tích

Chúng tôi đã thực hiện các thí nghiệm nhằm đánh giá hiệu suất hệ khuyến nghị với các phương pháp embedding khác nhau (TF-IDF, ST5, SBERT, SGPT) trên thông tin sách, sử dụng các độ đo như Recall@5, Recall@10, và Recall@50 để đánh giá hiệu suất. Đầu vào hệ khuyến nghị là 3 bộ dữ liệu mà nhóm xây dựng: Dataset1 (Description), Dataset2 (Description + Title), và Dataset3 (Description + Title + Author), trình bày chi tiết tại

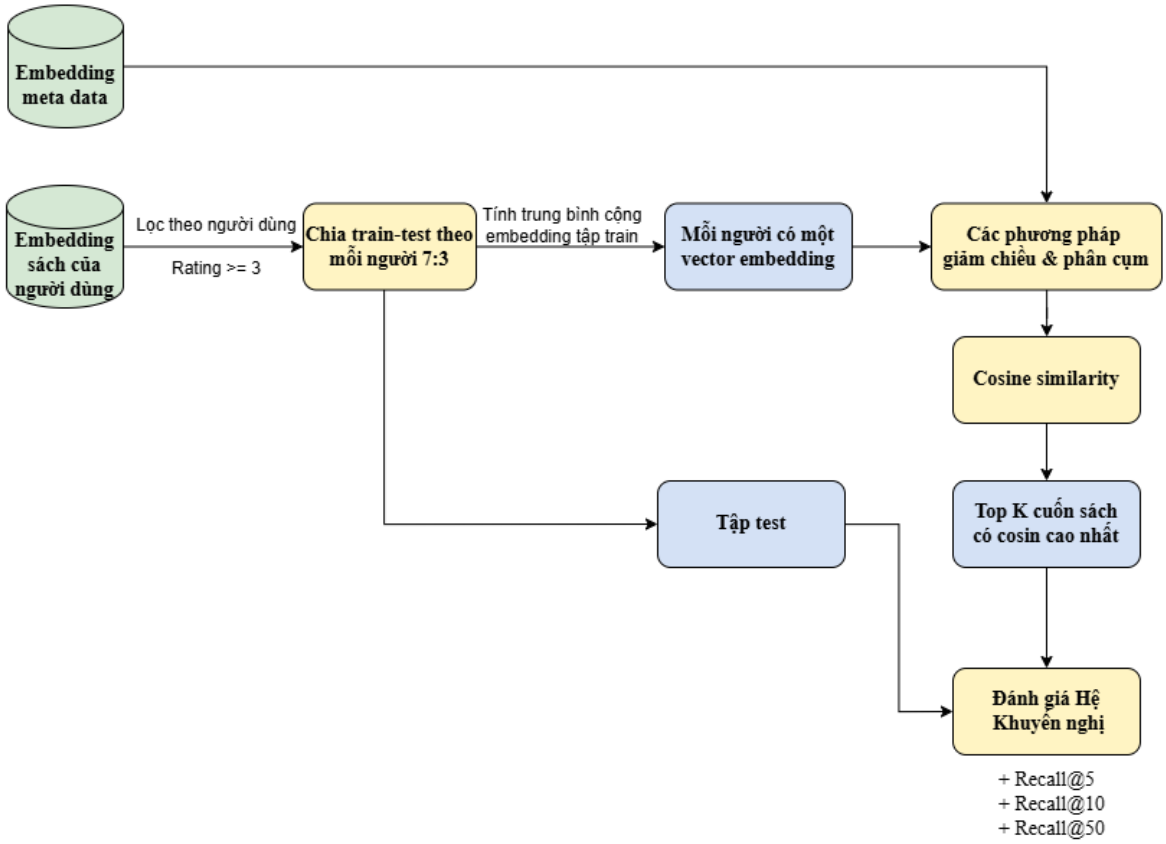

Figure 6: Quy trình Đánh giá Hệ khuyến nghị chi tiết

mục 3. Kết quả thực nghiệm được trình bày chi tiết trong các bảng (4), (5), và (6).

4.5.1 Phân tích ảnh hưởng của dữ liệu đầu vào

Thực nghiệm của nhóm được thiết kế để kiểm tra mức độ ảnh hưởng của dữ liệu đầu vào lên hiệu suất hệ thống khuyến nghị, cụ thể nhóm cho thực nghiệm trên cả 3 bộ dữ liệu: Dataset1, Dataset2, Dataset3. Quan sát từ các kết quả cho thấy, khi sử dụng TF-IDF, hiệu suất có xu hướng giảm khi dữ liệu ngày càng đa dạng. cụ thể, phương pháp *TF-IDF + UMAP + HDBSCAN* đạt Recall@5 = 0.1353 với Dataset1 -'Description', giảm xuống 0.1081 với Dataset2 - 'Title + Description', và chỉ còn 0.0846 với Dataset3 - 'Title + Description + Author'. Điều này được cho là do TF-IDF không xử lý tốt được mối liên kết ngữ nghĩa phức tạp, đặc biệt khi dữ liệu chứa nhiều yếu tố bổ sung như Title và Author, làm phân tán đặc trưng trong không gian vector. Kết quả TF-IDF giảm rõ rệt với Dataset3 có thể do nó không xử lý được mối quan hệ giữa Author và 2 thuộc tính 'Description' và 'Title' của cuốn sách.

Trong khi đó, các phương pháp embedding hiện đại là ST5, SBERT và SGPT đạt kết quả ổn định hơn. Hiệu suất của hai phương pháp này có xu hướng ổn định giữa hai bộ dữ liệu Dataset1 và Dataset3, giảm nhẹ khi thực nghiệm trên Dataset2. Ví dụ, hiệu suất của *SGPT + UMAP + KMeans (k=10)* trên ba dạng dữ liệu lần lượt là Recall@50 = 0.4757, 0.4373, và 0.4768. Điều này chứng minh rằng các mô hình embedding hiện đại có khả năng học và giữ được mối quan hệ ngữ nghĩa trong các ngữ cảnh đa dạng. Đặc biệt, SGPT và SBERT có khả năng mã hóa toàn diện ngữ cảnh văn bản, giúp mô hình duy trì chất lượng khuyến nghị khi dữ liệu đầu vào bao gồm nhiều nguồn thông tin hơn, ngay cả khi bổ sung thêm yếu tố tác giả (Author).

4.5.2 Phân tích ảnh hưởng của phương pháp phân cụm

Việc tích hợp các phương pháp embedding hiện đại với các kỹ thuật giảm chiều (UMAP) và phân cụm (HDBSCAN, KMeans) không chỉ cho phép chúng tôi đánh giá toàn diện hiệu suất, mà còn hỗ trợ phân tích sự khác biệt giữa hai phương pháp phân cụm

Embedding	Topic	Recall@5	Recall@10	Recall@50
TF-IDF	UMAP + HDBSCAN	0.1353	0.1624	0.2197
ST5		0.2278	0.2305	0.2409
SBERT		0.2357	0.2430	0.2549
SGPT		0.2474	0.2569	0.2802
TF-IDF	UMAP + K-means (k=10)	0.2756	0.3251	0.3889
ST5		0.2296	0.2703	0.3905
SBERT		0.2569	0.3224	0.4428
SGPT		0.2682	0.3380	0.4757

Table 4: Kết quả thực nghiệm trên Dataset1: Description

Embedding	Topic	Recall@5	Recall@10	Recall@50
TF-IDF	UMAP + HDBSCAN	0.1081	0.1299	0.1719
ST5		0.1910	0.2056	0.2236
SBERT		0.2277	0.2324	0.2468
SGPT		0.2394	0.2478	0.2698
TF-IDF	UMAP + K-means (k=10)	0.2575	0.3173	0.3883
ST5		0.2197	0.2694	0.3914
SBERT		0.2595	0.3199	0.4346
SGPT		0.2967	0.3519	0.4373

Table 5: Kết quả thực nghiệm trên Dataset2: Title + Description

Embedding	Topic	Recall@5	Recall@10	Recall@50
TF-IDF	UMAP + HDBSCAN	0.0846	0.1101	0.1549
ST5		0.2070	0.2171	0.2321
SBERT		0.2358	0.2387	0.2529
SGPT		0.2381	0.2494	0.2682
TF-IDF	UMAP + K-means (k=10)	0.2477	0.3057	0.3798
ST5		0.2315	0.2804	0.4058
SBERT		0.2586	0.3276	0.4531
SGPT		0.2853	0.3633	0.4768

Table 6: Kết quả thực nghiệm trên Dataset3: Title + Description + Author

chính được sử dụng trong bài. Kết quả chỉ ra rằng KMeans (k=10) luôn có hiệu suất tốt hơn so với HDBSCAN. Ví dụ, trên dạng dữ liệu 3 (Description + Title + Author), phương pháp *TF-IDF + UMAP + HDBSCAN* chỉ đạt Recall@5 = 0.0846, Recall@10 = 0.1101, và Recall@50 = 0.1549. Ngược lại, *TF-IDF + UMAP + KMeans (k=10)* cải thiện đáng kể với Recall@5 = 0.2477, Recall@10 = 0.3057, và Recall@50 = 0.3798, cao hơn từ 2 đến 3 lần độ đo khi kết hợp với HDBSCAN.

4.5.3 Đánh giá hiệu suất các phương pháp Embedding

Chúng tôi đã đánh giá hiệu suất của một số phương pháp sentence-embedding trong tác vụ gợi ý sản phẩm sách và so sánh chúng với phương pháp embedding cổ điển - TFIDF, phương pháp này đóng vai trò là cơ sở so sánh. Kết quả cho thấy các phương pháp sentence-embedding hiện đại như SBERT và SGPT vượt trội về khả năng gợi ý khi so sánh với các phương pháp embedding truyền thống như TF-IDF. Chẳng hạn,SGPT đạt Recall@5 = 0.2853, Recall@10 = 0.3633, và Recall@50 = 0.4768, cao hơn đáng kể so với các phương pháp truyền thống. Ngoài ra, mô hình SBERT (pre-trained bert-multilingual-uncased) cũng đạt kết quả tốt, với điểm số recall@k tương đương hoặc tốt hơn so với các phương pháp embedding cơ sở.

4.5.4 Tổng kết

Nhìn chung, nghiên cứu của chúng tôi cho thấy việc sử dụng các phương pháp sentence-embedding tiên

tiến có thể nâng cao đáng kể hiệu suất hệ thống khuyến nghị, đặc biệt là trong bối cảnh dữ liệu phức tạp và đa dạng. Đồng thời, kết quả này mở ra hướng nghiên cứu mới trong việc tối ưu hóa thời gian thực thi mà vẫn duy trì hiệu suất gợi ý cao.

Bảng 7 trình bày chi tiết ưu nhược điểm của 2 phương pháp: HDBSCAN và K-means trong việc ứng dụng cho bài toán khuyến nghị sách, từ đó lựa chọn phương pháp phù hợp tùy thuộc vào đặc điểm của dữ liệu và yêu cầu bài toán.

Nhóm thực hiện so sánh ưu nhược điểm của các phương pháp embedding sử dụng trong bài, trình bày chi tiết trong bảng 8. Mỗi phương pháp có ưu, nhược riêng cho bài toán khuyến nghị, vì vậy ta cần lựa chọn phương pháp phù hợp tùy thuộc vào đặc điểm của dữ liệu và yêu cầu bài toán.

5 DEMO

Trong phần này, chúng tôi xây dựng một trang web đơn giản về hệ thống khuyến nghị dựa trên nội dung với embedding ST5 trên nền tảng streamlit. Điều này nhằm thể hiện rõ hơn các tính năng tương tác, chức năng cũng như tính ứng dụng của mô hình gợi ý của chúng tôi trong giao diện người dùng thân thiện của streamlit. Cụ thể người dùng có thể nhập vào nội dung (từ khóa, tên cuốn sách hoặc nội dung cuốn sách), điều chỉnh số lượng cuốn sách mong muốn (điều chỉnh giá trị k), sau đó hệ thống sẽ thực hiện tính toán theo quy trình sau đó trả về kết quả top k, theo định dạng danh sách, thông tin các cuốn sách trả về gồm: Tên sách, Hình bìa sách, Mô tả, Độ tương đồng (Cosine) với nội dung người dùng nhập. Giao diện web được thiết kế như hình 7 Đây chỉ là một thiết kế đơn giản, chưa thể hiện hết được tiềm năng ứng dụng của hệ khuyến nghị sách. Hệ khuyến nghị sách có thể ứng dụng trong việc khuyến nghị trên các trang thương mại điện tử,...

6 KẾT LUẬN

Chúng tôi đã thành công xây dựng hệ thống gợi ý sách dựa trên bộ dữ liệu Goodreads. Hệ thống gợi ý là 8 tổ hợp từ 4 phương pháp embedding (TFIDF, ST5, SBERT và SGPT), phương pháp giảm chiều dữ liệu UMAP và 2 phương pháp gom cụm (HDBSCAN, Kmean (k=10)). Kết quả bài toán cho thấy các phương pháp sentence-embedding hiện đại có thể là công cụ hữu ích trong việc nâng cao độ chính xác của hệ thống khuyến nghị, đồng thời mở ra một tương lai đầy hứa hẹn thay thế cho các phương pháp truyền thống như TFIDF. Việc tận dụng các mô hình đã được huấn luyện trước cho phép tạo ra cho ra các gợi ý chất lượng cao một cách nhanh chóng. Tổng thể, kết quả thực nghiệm của chúng tôi làm nổi bật tiềm năng của các phương pháp sentence-embedding trong tác vụ: khuyến nghị. Kết quả này cung cấp cơ sở vững chắc cho các nghiên cứu tiếp theo trong lĩnh vực này.

Trong nhiệm vụ gợi ý, chúng tôi sử dụng hai cấu hình UMAP+HDBSCAN và UMAP+Kmean (k=10) để so sánh cả các phương pháp gom cụm và các phương pháp embedding khác nhau. Mỗi phương pháp lại tạo ra các embedding với kích thước chiều khác nhau. Vì vậy, trong các nghiên cứu tương lai, chúng tôi sẽ tiếp tục khám phá ảnh hưởng của việc giảm chiều và các thuật toán phân cụm đối với các phương pháp embedding này. Ngoài ra, nhóm dự định triển khai hệ thống khuyến nghị sách hiện tại và thực hiện đánh giá trực tuyến, chẳng hạn như thử nghiệm A/B, để kiểm tra hiệu quả thực tế của nó.

6.1 Hạn chế của bài toán

Mặc dù nghiên cứu của chúng tôi đã chỉ ra hiệu quả rõ rệt của các phương pháp embedding hiện đại cùng các kỹ thuật gom cụm trong việc nâng cao chất lượng hệ thống khuyến nghị, nhưng vẫn tồn tại một số hạn chế cần được xem xét. Một trong những vấn đề nổi bật là tác động không lường trước của yếu tố ngẫu nhiên trong các thuật toán phân cụm, chẳng hạn như việc phương pháp HDBSCAN không thể kiểm soát được số lượng chủ đề, điều này có thể dẫn đến các kết quả sai lệch. Sự biến thiên vốn có trong những thuật toán này có thể gây ra những sai sót, làm giảm độ tin cậy và tính ổn định của kết quả phân cụm. Để khắc phục vấn đề này, các nghiên cứu trong tương lai có thể tập trung vào việc phát triển các chiến lược giảm thiểu ảnh hưởng của yếu tố ngẫu nhiên và nâng cao tính ổn định của các phương pháp đã được đề xuất.

References

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2015. Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, pages 1–34.

Ye Bi, Liqiang Song, Mengqiu Yao, Zhenyu Wu, Jianming Wang, and Jing Xiao. 2020. A heterogeneous information network based cross domain insurance recommendation system for cold start users. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2211–2220, New York, NY, USA. Association for Computing Machinery.

Phương pháp	Ưu điểm	Nhược điểm
HDBSCAN	- Không yêu cầu số lượng cụm cố định. - Có thể xử lý các dữ liệu có hình dạng không đồng đều. - Phân cụm tự động xác định các điểm nhiễu (outliers).	- Có thể gặp khó khăn trong việc giải thích số lượng cụm được tìm thấy. - Phụ thuộc vào các tham số như min_samples và min_cluster_size, việc lựa chọn các tham số này có thể ảnh hưởng lớn đến kết quả. - Thời gian tính toán có thể lâu hơn khi so với K-means đối với các tập dữ liệu lớn.
K-means (k=10)	- Đơn giản và dễ triển khai. - Hiệu quả trong việc phân cụm dữ liệu có hình dạng tròn đều. - Thời gian tính toán nhanh đối với tập dữ liệu nhỏ đến trung bình.	- Yêu cầu người dùng xác định trước số lượng cụm (k). - Không hoạt động tốt với dữ liệu không đồng nhất. - Dễ bị ảnh hưởng bởi nhiễu và các điểm ngoại lệ (outliers). - Không thể tự động nhận diện và xử lý các điểm nhiễu như HDBSCAN.

Table 7: So sánh ưu nhược điểm của HDBSCAN và K-means (k=10)

Selva Birunda and R.Kanniga Devi. 2021. *A Review on Word Embedding Techniques for Text Classification*, pages 267–281.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. 2019. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 456–464, New York, NY, USA. Association for Computing Machinery.

Natalie K. Cygan. 2021. Sentence-bert for interpretable topic modeling in web browsing data.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Sayed Nasir Hasan and Ravi Khatwal. 2022. Cold start problem in recommendation system: A solution model based on clustering and association rule

Phương pháp	Ưu điểm	Nhược điểm	Ứng dụng
TF-IDF	- Đơn giản, dễ hiểu và dễ triển khai. - Không yêu cầu huấn luyện mô hình phức tạp. - Hiệu quả với dữ liệu nhỏ và các từ có tần suất thấp.	- Không nắm bắt được nghĩa ngữ cảnh của từ. - Không hiệu quả với các câu dài hoặc phức tạp. - Không thể xử lý các nghĩa đa nghĩa. - Dễ bị ảnh hưởng bởi từ ngữ không quan trọng.	- Phân loại văn bản. - Tìm kiếm tài liệu. - Phân tích và trích xuất đặc trưng trong văn bản.
ST5	- Sử dụng mô hình T5 mạnh mẽ đã được huấn luyện trước. - Cung cấp nhúng câu chất lượng cao với khả năng hiểu ngữ cảnh tốt. - Hiệu quả cho nhiều tác vụ NLP.	- Yêu cầu tài nguyên tính toán lớn. - Cần thời gian huấn luyện dài và mô hình phức tạp.	- Tổng hợp văn bản. - Phân tích cảm xúc. - Dịch máy.
SBERT	- Tạo sentence-embedding chất lượng cao. - Hiệu quả cao trong các tác vụ như tìm kiếm và phân loại văn bản. - Hỗ trợ tính toán embedding theo cặp câu.	- Cần huấn luyện trước và không thể tối ưu cho từng tác vụ cụ thể nếu không có dữ liệu huấn luyện. - Cần nhiều tài ngueyen tính toán để huấn luyện - Tốc độ huấn luyện chậm.	- Tìm kiếm câu tương tự. - Phân loại văn bản. - Hệ thống khuyến nghị
SGPT	- Kết hợp giữa GPT và Sentence-embedding. - Mô hình mạnh mẽ - Có thể tạo văn bản mượt mà và chính xác.	- Cần tài nguyên tính toán rất lớn. - Chi phí huấn luyện cao. - Đôi khi không ổn định khi xử lý dữ liệu không chuẩn.	- Tạo văn bản tự động. - Sinh câu - Tóm tắt văn bản.

Table 8: So sánh các phương pháp embedding

techniques. In *2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, pages 1–8.

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

Budi Juarto and Abba Girsang. 2021. Neural collaborative with sentence bert for news recommender system. *JOIV : International Journal on Informatics Visualization*, 5:448.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Chahrazed Mediani, Saad Harous, and Mahieddine Djoudi. 2023. Content-based recommender system using word embeddings for pedagogical resources. In *2023 5th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–8.

Saulo Mendes de Melo, André Lima Férrer de Almeida, Lívia Almada Cruz, and Ticiana Linhares Coelho da Silva. 2022. A chat recommender system for covid-19 support based in textual sentence embeddings. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, page 248–252, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embed-

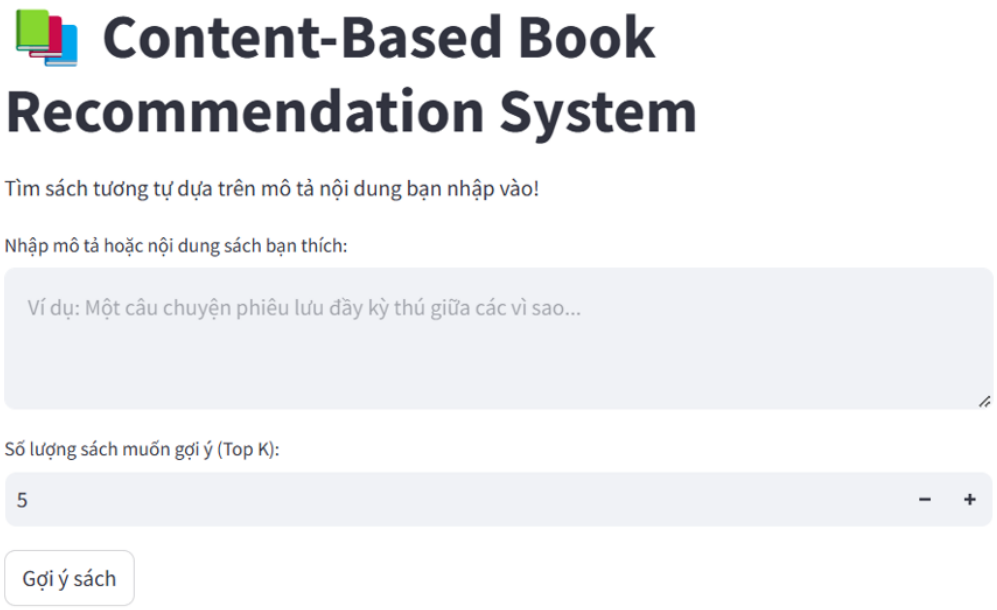

Figure 7: Giao diện web demo

dings for semantic search.

M. E. J. Newman. 2005. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351.

Hai Nguyen, Jérémie Mary, and Philippe Preux. 2014. Cold-start problems in recommendation systems via contextual-bandit algorithms.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Paul Resnick and Hal R. Varian. 1997. Recommender systems. *Commun. ACM*, 40(3):56–58.

Bin Wang and C.-C. Jay Kuo. 2020. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28:2146–2157.

Nuofan Xu and Chenhui Hu. 2023. Enhancing e-commerce recommendation using pre-trained language model and fine-tuning.

Chong Yang, Xiaohui Yu, Yang Liu, Yanping Nie, and Yuanhong Wang. 2016. Collaborative filtering with weighted opinion aspects. *Neurocomputing*, 210.