BOOK RECOMMENDATION SYSTEM WITH SENTIMENT INSIGHT

Bùi Bảo Trân, Phạm Hồng Trà, Nguyễn Lê Thanh Minh, Châu Nguyễn Tri Vũ, Huỳnh Văn Tín
Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{22521503, 22521495,22520875, 22521687 }gm.uit.edu.vn
tinhv@uit.edu.vn

Abstract

Trong thời đại chuyển đổi số, sự gia tăng mạnh mẽ của nội dung trực tuyến và các đánh giá do người dùng tạo ra mang đến cơ hội lớn trong việc hỗ trợ bạn đọc và nhà nghiên cứu tiếp cận thông tin, nhưng đồng thời cũng đặt ra nhiều thách thức. Đặc biệt, các đánh giá sách là nguồn dữ liệu quý giá, cung cấp gợi ý và phản hồi hữu ích để người dùng đưa ra quyết định lựa chọn sách. Tuy nhiên, số lượng đánh giá khổng lồ và đa dạng khiến việc xác định thông tin phù hợp, nắm bắt cảm xúc chung, hoặc tìm kiếm sách trở nên phức tạp. Để giải quyết vấn đề này, đề tài hướng đến xây dựng một hệ thống gợi ý sách thông minh, tích hợp kỹ thuật phân tích cảm xúc từ đánh giá của người dùng. Hệ thống không chỉ cung cấp danh sách sách phù hợp mà còn phân loại các đánh giá theo chiều hướng tích cực, tiêu cực hoặc trung tính, nhằm tối ưu hóa trải nghiệm tìm kiếm và hỗ trợ người đọc đưa ra quyết định chính xác.

1 Giới thiệu

Trong thời đại kỹ thuật số, việc lựa chọn sách phù hợp với sở thích cá nhân không còn là một nhiệm vụ đơn giản khi mà lượng sách xuất bản ngày càng gia tăng, cùng với đó là sự phong phú trong phản hồi và đánh giá từ cộng đồng độc giả. Điều này đặt ra thách thức lớn trong việc khai thác thông tin từ đánh giá để xây dựng một hệ thống gợi ý chính xác và hiệu quả. Để giải quyết vấn đề trên, hệ thống *"Khuyến nghị sách dựa trên thông điệp trong các đánh giá từ người dùng"* được thiết kế, kết hợp giữa các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) tiên tiến và học máy, nhằm mang đến trải nghiệm cá nhân hóa cho độc giả.

Một trong những yếu tố cốt lõi của hệ thống là phân tích cảm xúc, công cụ mạnh mẽ giúp tự động phân loại các đánh giá sách thành những nhóm cảm xúc chính như tích cực, tiêu cực hoặc trung tính. Không chỉ dừng lại ở việc cung cấp cái nhìn tổng quan về sự hài lòng hay không hài lòng của độc giả, phân tích cảm xúc còn đóng vai trò quan trọng trong việc tinh chỉnh hệ thống gợi ý sách. Bằng cách nắm bắt tâm trạng và sở thích của người dùng, hệ thống có thể đưa ra các gợi ý sách mang tính cá nhân hóa cao, giúp trải nghiệm đọc trở nên thú vị và hiệu quả hơn. Hơn nữa, chức năng tìm kiếm thông minh được tích hợp trong hệ thống, với khả năng tối ưu hóa theo ngữ cảnh và nhu cầu cụ thể của người dùng, cho phép họ nhanh chóng tìm thấy các tựa sách phù hợp mà không cần tốn thời gian lọc qua những đánh giá không liên quan.

Hệ thống khuyến nghị sách này không chỉ hữu ích cho người đọc cá nhân mà còn là nguồn thông tin quý giá cho các nhà xuất bản và tác giả. Dựa trên phân tích xu hướng từ đánh giá người dùng, họ có thể hiểu rõ hơn về thị hiếu của độc giả, từ đó cải thiện chiến lược marketing và phát triển nội dung sách. Việc kết hợp giữa phân tích cảm xúc và tìm kiếm thông minh đã tạo nên một giải pháp toàn diện, đảm bảo tính mở rộng, chính xác, và hiệu quả, đáp ứng nhu cầu của nhiều nhóm đối tượng trong hệ sinh thái sách số.

Tiến bộ trong công nghệ phân tích cảm xúc Trong lĩnh vực phân tích cảm xúc, các thuật toán truyền thống như Naive Bayes, Support Vector Machine (SVM) từng đóng vai trò quan trọng trong giai đoạn đầu phát triển. Tuy nhiên, sự ra đời của các công nghệ mới, đặc biệt là mạng nơ-ron nhân tạo, đã mở ra cơ hội cải tiến mạnh mẽ về hiệu suất và độ chính xác. Các mô hình hiện đại như mạng nơ-ron tích chập (CNN) kết hợp với mạng nơ-ron hồi quy dài-ngắn hạn (LSTM) đã được chứng minh là hiệu quả vượt trội. Theo nghiên cứu của Ayutthaya và Pasupa (2018), việc tích hợp các đặc trưng như Word2Vec, POS tagging, và SenticNet trong mô hình CNN-LSTM đã cải thiện đáng kể hiệu suất phân tích cảm xúc. Cụ thể, mô hình kết hợp này đạt F1-score lên tới 0.9, cao hơn rõ rệt so với các mô hình riêng lẻ hoặc thiếu tính năng, vốn chỉ đạt khoảng 0.86. Đồ thị ROC cũng cho thấy hiệu quả vượt trội của mô hình với chỉ số AUC đạt

0.9588, minh chứng cho khả năng phân biệt cảm xúc đáng tin cậy.

Nhằm nâng cao hơn nữa hiệu quả và trải nghiệm người dùng, đề tài này đề xuất tích hợp mô hình ngôn ngữ lớn (Large Language Model - LLM) cùng với các module tìm kiếm và phân loại cảm xúc. Các công nghệ tiên tiến như BM25, Sentence Embedding được sử dụng trong việc tối ưu hóa truy vấn, trong khi các mô hình phân tích cảm xúc mạnh mẽ như ViSoBERT và PhoBERT đảm bảo khả năng phân loại cảm xúc chính xác. Kết hợp cả hai yếu tố này, hệ thống không chỉ đưa ra gợi ý sách phù hợp nhất với nhu cầu cá nhân mà còn cung cấp cái nhìn toàn diện hơn qua phân tích chi tiết các đánh giá liên quan đến từng cuốn sách. Điều này không chỉ giúp người dùng dễ dàng tìm thấy cuốn sách yêu thích mà còn hỗ trợ họ trong việc đưa ra quyết định đọc dựa trên dữ liệu thực tế và đáng tin cậy.

2 Các công trình liên quan

2.1 UIT-VSFC

Bộ dữ liệu UIT-VSFC (Vietnamese Students' Feedback Corpus) (Nguyen et al., 2018) là một trong những tài nguyên quan trọng phục vụ cho các nghiên cứu về *phân tích cảm xúc* và *xử lý ngôn ngữ tự nhiên (NLP)* trong tiếng Việt. Bộ dữ liệu này được phát triển bởi *Trường Đại học Công nghệ Thông tin, ĐHQG-HCM (UIT)* và tập trung vào phản hồi từ sinh viên, với mục tiêu *phân loại cảm xúc* và *phân tích quan điểm (opinion mining)*.

Đặc điểm của bộ dữ liệu UIT-VSFC:

- Bộ dữ liệu bao gồm các phản hồi và đánh giá của sinh viên được gán nhãn với 3 loại cảm xúc chính:
  - **Tích cực** (Positive)
  - **Tiêu cực** (Negative)
  - **Trung tính** (Neutral)
- *UIT-VSFC* là một trong những bộ dữ liệu tiếng Việt đầu tiên có gán nhãn cảm xúc một cách đầy đủ và chi tiết, hỗ trợ cho các mô hình học máy và học sâu trong việc phân loại cảm xúc.

Ứng dụng và kết quả nghiên cứu:

- Trong nghiên cứu của (Nguyen et al., 2018), bộ dữ liệu UIT-VSFC đã được sử dụng để huấn luyện và đánh giá các mô hình học máy như *Naive Bayes*, *MaxEnt classification* đạt kết quả khả quan khi áp dụng các kỹ thuật tiền xử lý văn bản và trích xuất đặc trưng.
- Gần đây, các nghiên cứu đã thử nghiệm các mô hình học sâu như *CNN, LSTM*, và *BERT* trên bộ dữ liệu *UIT-VSFC*. Kết quả cho thấy mô hình BERT và các phương pháp kết hợp Word Embedding như *Word2Vec* đã cải thiện đáng kể độ chính xác so với các phương pháp truyền thống.

2.2 Sentiment Analysis on Book Reviews Using Convolutional Neural Network (CNN) Long Short-Term Memory (LSTM) Hybrid

Nghiên cứu "**Sentiment Analysis on Book Reviews Using Convolutional Neural Network (CNN) Long Short-Term Memory (LSTM) Hybrid**" (Alharbi and de Doncker, 2019) tập trung vào việc xây dựng mô hình kết hợp *CNN* và *LSTM* để phân tích cảm xúc trong các đánh giá sách. **Mục tiêu chính của nghiên cứu:**

- Xây dựng mô hình kết hợp *CNN-LSTM*: Mô hình này được thiết kế để phân loại cảm xúc trong các đánh giá sách, nhằm hiểu rõ hơn về quan điểm của người đọc.
- Sử dụng các đặc trưng ngôn ngữ: Nghiên cứu áp dụng các đặc trưng như *Word2Vec*, *POS (Part-of-Speech)* và *SenticNet* để cải thiện độ chính xác của mô hình.

Phương pháp tiếp cận:

- Kết hợp *CNN* và *LSTM*: *CNN* được sử dụng để trích xuất các đặc trưng cục bộ từ văn bản, trong khi *LSTM* giúp nắm bắt các mối quan hệ ngữ cảnh dài hạn, tạo nên một mô hình mạnh mẽ cho phân tích cảm xúc.
- Tích hợp các đặc trưng ngôn ngữ: Việc kết hợp *Word2Vec*, *POS* và *SenticNet* giúp mô hình hiểu sâu hơn về ngữ nghĩa và cấu trúc của văn bản, từ đó nâng cao hiệu quả phân loại cảm xúc.

Kết quả đạt được:

- Hiệu suất cao: Mô hình kết hợp *CNN-LSTM* với các đặc trưng ngôn ngữ đạt *F1-score* lên đến *0.9*, vượt trội so với các mô hình chỉ sử dụng *CNN* hoặc *LSTM* đơn lẻ.
- Đánh giá bằng *ROC*: *Đường cong ROC* của mô hình cho thấy giá trị *AUC* đạt *0.9588*, chứng tỏ khả năng phân loại xuất sắc.

Đóng góp của nghiên cứu:

- Cải tiến phương pháp phân tích cảm xúc: Nghiên cứu chứng minh rằng việc kết hợp các đặc trưng ngôn ngữ và mô hình *CNN-LSTM* có thể nâng cao đáng kể hiệu quả phân tích cảm xúc trong các đánh giá sách.
- Ứng dụng thực tiễn: Mô hình này có thể được áp dụng để hiểu rõ hơn về phản hồi của người đọc, hỗ trợ các nhà xuất bản và tác giả trong việc cải thiện chất lượng nội dung.

2.3 ViSoBERT

ViSoBERT (Nguyen et al., 2023) được phát triển dựa trên kiến trúc BERT (Bidirectional Encoder Representations from Transformers) - một mô hình học sâu mang tính cách mạng trong lĩnh vực NLP. Tuy nhiên, ViSoBERT được thiết kế và huấn luyện dành riêng cho tiếng Việt với các đặc điểm như:

- Tối ưu hóa cho văn bản tiếng Việt từ mạng xã hội: Dữ liệu huấn luyện của ViSoBERT được lấy từ các nền tảng mạng xã hội, nơi văn bản thường không có cấu trúc, chứa nhiều lỗi chính tả, từ viết tắt, tiếng lóng và các biểu tượng cảm xúc.
- Hiểu ngữ cảnh sâu sắc: ViSoBERT có khả năng hiểu ngữ cảnh hai chiều, giúp nó nắm bắt tốt các nghĩa ngầm hoặc sự mơ hồ trong văn bản tiếng Việt.
- Hiệu suất vượt trội: So với các mô hình ngôn ngữ trước đó như PhoBERT hoặc các phương pháp truyền thống, ViSoBERT cho thấy độ chính xác cao hơn trong nhiều tác vụ NLP.

3 Tổng quan bài toán

Trong bối cảnh thị trường sách ngày càng mở rộng với sự gia tăng không ngừng của các tựa sách và đánh giá từ cộng đồng đọc giả, việc xây dựng một hệ thống hỗ trợ người dùng tìm kiếm và lựa chọn sách phù hợp trở thành một thách thức đáng kể. Đề tài này tập trung giải quyết bài toán lớn bao gồm hai nhiệm vụ chính: **gợi ý sách dựa trên truy vấn của người dùng** và **phân loại cảm xúc từ các đánh giá liên quan đến sách**. Bài toán tổng quát được mô tả như sau:

- **Input:** Một câu truy vấn của người dùng, biểu đạt ý muốn tìm kiếm thông tin về sách hoặc những nội dung liên quan.
- **Output:** Danh sách các tựa sách phù hợp với truy vấn kèm theo các đánh giá đã được phân loại cảm xúc, thể hiện phản hồi của độc giả đối với mỗi tựa sách.

Để giải quyết bài toán trên một cách hiệu quả, nhóm đã phân tích và chia bài toán thành hai bài toán nhỏ, độc lập nhưng có sự liên kết chặt chẽ, bao gồm: **truy vấn thông tin dựa trên các đánh giá sách** và **phân loại cảm xúc của các đánh giá sách**

3.1 Truy vấn thông tin dựa trên các đánh giá sách

Truy vấn thông tin từ các đánh giá sách là nhiệm vụ giúp người dùng tìm kiếm và nhận được các gợi ý sách phù hợp dựa trên nội dung đánh giá. Bài toán này không chỉ nhằm tìm ra những tựa sách liên quan đến yêu cầu của người dùng mà còn trích xuất, tổng hợp thông tin hữu ích từ cộng đồng độc giả để đưa ra các đề xuất chất lượng cao.

Bài toán được định nghĩa như sau:

- **Input:** Một câu truy vấn của người dùng, có thể là một câu hỏi hoặc một yêu cầu cụ thể liên quan đến sách.
- **Output:** Danh sách những tựa sách có liên quan đến truy vấn, được lựa chọn và sắp xếp dựa trên nội dung và ngữ cảnh trong các đánh giá sách.

Mục tiêu chính của bài toán truy vấn thông tin là xây dựng một hệ thống không chỉ cung cấp kết quả chính xác mà còn đảm bảo độ bao phủ cao, tức là cung cấp đủ thông tin để người dùng dễ dàng tìm thấy cuốn sách phù hợp nhất với nhu cầu của họ. Đây là một thành phần quan trọng giúp hoàn thiện hệ thống khuyến nghị, tăng cường sự tin cậy và hiệu quả của nền tảng.

3.2 Phân loại cảm xúc của các đánh giá sách

Sau khi nhận được gợi ý là những cuốn sách phù hợp với truy vấn của người dùng. Hệ thống sẽ cung cấp các đánh giá khách quan từ độc giả về cuốn sách đó. Bài toán còn lại ở đây là - phân loại cảm xúc từ các đánh giá sách, một bài toán quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), nhằm xác định trạng thái cảm xúc mà người dùng thể hiện trong những nhận xét về một tựa sách. Kết quả phân tích cảm xúc không chỉ cung cấp cái nhìn rõ nét về phản ứng của độc giả mà còn là cơ sở để

cá nhân hóa gợi ý sách, từ đó cải thiện trải nghiệm đọc sách.

Bài toán được định nghĩa như sau:

- **Input:** Một câu hoặc đoạn văn bản là đánh giá (review) của người dùng về một cuốn sách.
- **Output:** Một nhãn cảm xúc thuộc một trong ba nhóm chính: **Tích cực (Positive)**, **Tiêu cực (Negative)**, hoặc **Trung tính (Neutral)**, biểu thị ý nghĩa cảm xúc của đánh giá đó.

Việc phân loại cảm xúc đóng vai trò quan trọng trong hệ thống khuyến nghị, giúp phản ánh mức độ hài lòng hoặc không hài lòng của độc giả đối với mỗi tựa sách. Kết quả này không chỉ giúp người dùng khác dễ dàng hơn trong việc ra quyết định mà còn cung cấp thông tin hữu ích cho các nhà xuất bản và tác giả trong việc cải thiện chất lượng sản phẩm.

Như vậy, sự kết hợp giữa hai bài toán nhỏ này sẽ tạo nên một hệ thống toàn diện, vừa hỗ trợ người dùng tìm kiếm thông tin, vừa cung cấp phân tích cảm xúc chi tiết, giúp họ đưa ra quyết định tốt hơn khi tìm kiếm cuốn sách phù hợp với nhu cầu của mình.

4 Dữ liệu

Nhóm xây dựng hai bộ dữ liệu để giải quyết bài toán, bao gồm: bộ dữ liệu phân loại cảm xúc đánh giá sách và bộ dữ liệu Q&A. Sau đây là thông tin chi tiết về hai bộ dữ liệu này.

4.1 Xây dựng bộ dữ liệu

4.1.1 Bộ dữ liệu phân loại cảm xúc của các đánh giá sách

Quy trình xây dựng bộ dữ liệu được nhóm tham khảo bài báo Emotion Recognition for Vietnamese Social Media Text (Ho et al., 2020) như hình minh họa (Hình 1) với 3 giai đoạn:

- Giai đoạn 1 - Thu thập dữ liệu: Nhóm thu thập dữ liệu bao gồm các bình luận review về sách trên website Goodreads - website dành cho người đọc sách với loạt sách đa dạng thể loại, cho phép người dùng viết bài cảm nhận, đánh giá sách và nhiều ứng dụng khác. Nhóm thực hiện crawl các bình luận Tiếng Việt với 8 thể loại sách khác nhau.
- Giai đoạn 2 - Tiền xử lý: vì tính chất là ngôn ngữ mạng xã hội, nên cần loại bỏ các biểu tượng cảm xúc, dấu cách dư thừa, giải nghĩa các từ viết tắt,..
- Giai đoạn 3 - Gán Nhãn Dữ Liệu: Quá trình gán nhãn gồm hai bước chính:
  - Bước 1: Xây dựng bộ hướng dẫn gán nhãn và đào tạo bốn người gán nhãn. Nhóm tiến hành quá trình gán nhãn thử nghiệm và chỉnh sửa bộ hướng dẫn nhiều lần (lần lượt trên bộ dữ liệu gồm 100, 200 va 200 câu đánh giá sách) cho đến khi đạt độ đồng thuận trên 80%.
  - Bước 2: Chia 17,440 câu bình luận đều cho bốn người gán nhãn, sau đó kiểm tra chéo để đảm bảo độ chính xác đạt trên 80%.

4.1.2 Bộ dữ liệu hỏi đáp Q&A

Để xây dựng bộ bộ dữ liệu hỏi đáp, nhóm đã chọn ra ngẫu nhiên 50 cuốn sách và nhờ các mô hình ngôn ngữ lớn tạo ra 371 bộ câu hỏi và câu trả lời gợi ý dựa trên các bài review và nội dung của những cuốn sách này.

Bộ dữ liệu bao gồm các cặp prompt (đầu vào thể hiện cảm xúc, tình huống hoặc mong muốn của người dùng) và nội dung sách gợi ý (lời khuyên về các loại sách phù hợp với nhu cầu tâm lý và cảm xúc được thể hiện trong prompt). Bộ dữ liệu giúp xác định nhu cầu tìm kiếm sách của người dùng để đưa ra những cuốn sách phù hợp.

Bộ dữ liệu gồm cột:

- Prompt: Một chuỗi văn bản ngắn, thường được viết dưới dạng câu miêu tả cảm xúc, suy nghĩ hoặc nhu cầu của người dùng.
- Nội dung sách gợi ý: Một đoạn văn bản miêu tả chi tiết về loại sách phù hợp với cảm xúc hoặc tình huống được đề cập trong prompt được LLM tạo ra. Đoạn gợi ý này thường:
  - Tập trung vào các chủ đề giúp người dùng giải quyết vấn đề cảm xúc hoặc tâm lý của họ.
  - Đề xuất thể loại sách cụ thể (như sách tự lực, sách về phát triển bản thân, hoặc sách văn học có ý nghĩa sâu sắc).

Ví dụ:

Prompt: *'Tôi đang cảm thấy lạc lõng và cô đơn.'*

Nội dung sách gợi ý: *'Dựa trên cảm xúc của bạn, tôi gợi ý một cuốn sách có nội dung tập trung vào việc khám phá bản thân và xây dựng kết nối với chính mình cũng như với người khác. Cuốn sách nên giúp bạn hiểu rõ hơn về cảm xúc, cách vượt qua sự cô đơn, và cách tìm kiếm ý nghĩa trong những*

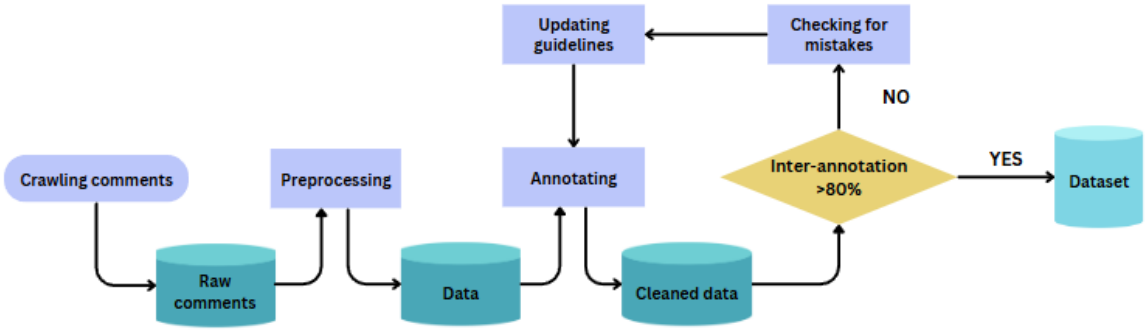

Figure 1: Quy trình xây dựng bộ dữ liệu phân loại cảm xúc của các đánh giá sách

mối quan hệ cũng như trong cuộc sống. Ngoài ra, nội dung nên khuyến khích bạn thực hành lòng biết ơn, trân trọng những gì mình có và mở lòng với những cơ hội mới để gắn kết với thế giới xung quanh.'

Dựa trên nội dung sách được gợi ý bởi mô hình ngữ lớn, hệ thống sẽ truy vấn vào cơ sở dữ liệu để tìm ra những cuốn sách liên quan.

4.2 Bộ quy tắc và hướng dẫn gán nhãn dữ liệu

Sau đây là bộ quy tắc và hướng dẫn gán nhãn cho tập dữ liệu loại cảm xúc review sách với ba nhãn bao gồm: tích cực (Positive), tiêu cực (Negative) và trung tính (Neutral).

- Nhãn tích cực: Thể hiện sự yêu thích rõ ràng đối với cuốn sách. Sử dụng các từ ngữ khen ngợi hoặc biểu đạt cảm xúc tích cực. Đề cập đến các yếu tố cụ thể khiến họ hài lòng, ví dụ: cốt truyện, nhân vật, phong cách viết, hoặc cảm xúc nhận được. Khuyến khích người khác đọc cuốn sách,..

  Ví dụ: *"Xúc động và nể thầy lắm. Hoc sinh nên đọc, người lớn cũng nên đọc để hiểu hơn và tạo động lực cho con cháu yêu quý sự học. Mình mới đọc tập Tôi đi học mà đã hình dung phần nào tính cách của thầy rồi. Một người luôn nỗ lực, kiên trì trong mọi việc dù là nhỏ nhất."*

- Nhãn tiêu cực: Thể hiện rõ sự không hài lòng hoặc chê bai cuốn sách. Sử dụng các từ ngữ phê bình hoặc biểu đạt cảm xúc tiêu cực. Đề cập đến các yếu tố cụ thể gây khó chịu, ví dụ: cốt truyện dở, nhân vật nhàm chán, lỗi viết, hoặc không đáp ứng kỳ vọng. Khuyên người khác không nên đọc cuốn sách,..

  Ví dụ: *"Mình chỉ đọc đến trang 45 là không muốn đọc tiếp, cách kể rất nhảm nhí, không cuốn hút và không ấn tượng."*

- Nhãn trung tính: Không thể hiện rõ ràng sự thích hoặc không thích cuốn sách. Đưa ra nhận xét chung chung, khách quan, không mang tính cảm xúc mạnh. Có thể nêu ưu và nhược điểm một cách cân bằng mà không nghiêng về bên nào. Chỉ đơn thuần mô tả nội dung hoặc các yếu tố của cuốn sách mà không bày tỏ cảm xúc mạnh mẽ,..

  Ví dụ: *"Vẫn chưa cảm thấy nó thực sự hay như người ta vẫn đánh giá mặc dù tác phẩm thành công trong việc miêu tả những nhân vật rất con người với những mâu thuẫn và đấu tranh nội tâm sâu sắc. Ngoài ra việc tác giả miêu tả về những gì diễn ra trong tầng lớp quý tộc Nga và những chuyển biến cả về tư tưởng và vật chất trong xã hội Nga báo hiệu những chuyển biến mang tính cách mạng trong xã hội Nga sau đó."*

Ngoài ra, chúng tôi gán nhãn cũng áp dụng một vài lưu ý ngoài lề như: ưu tiên đánh giá, cảm xúc trong các phần có những từ mang tính tổng kết như "tóm lại", "nhìn chung là",...

4.3 Thống kê về bộ dữ liệu phân loại cảm xúc review sách

Bộ dữ liệu có tổng cộng 17,440 câu review sách. Với 8 thể loại sách khác nhau và 663 cuốn sách với số lượng sách của từng thể loại được miêu tả trực quan qua hình (figure) 2:
Phần trăm các nhãn của các câu review sách được thể hiện qua hình (figure) 3:

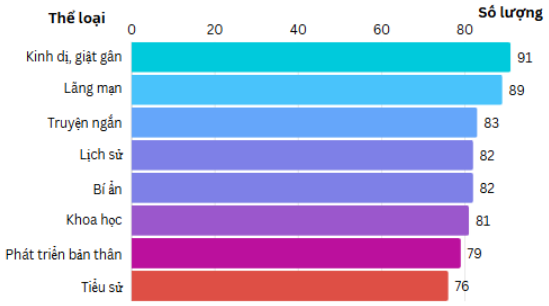

Figure 2: Thể loại sách và số lượng tương ứng

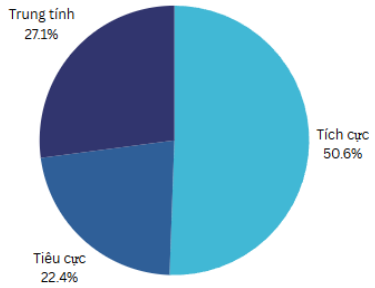

Figure 3: Biểu đồ thể hiện số lượng từng nhãn trong bộ dữ liệu

Tập dữ liệu trên được chia ra tập test và tập train với tỉ lệ 80/20.
Ngoài ra, chúng tôi còn ghi chép các từ viết tắt, các teencode. Ví dụ như:'cx' - cũng, 'nhma' - nhưng mà, 'vde' - vấn đề, 'xd' - xây dựng, 'khs' - không hiểu sao,...
Trong quá trình thu thập dữ liệu và gán nhãn, chúng tôi nhận thấy đây bộ dữ liệu này có nhiều thách thức do độ phức tạp cao. Cụ thể:

- Nhiều bình luận có độ dài lớn, mang tính phân tích nội dung sách, dẫn đến khó xác định cảm xúc chính.
- Ngôn ngữ không đồng nhất: Sử dụng từ ngữ đa dạng, dễ gây hiểu lầm khi bình luận chứa cả ý khen ngợi lẫn chỉ trích.
- Chứa nhiều dạng câu tu từ, chơi chữ, ca dao, tục ngữ.
- Sai lệch trong đánh giá: Điểm số sao không luôn phản ánh cảm xúc. Ví dụ, một số người coi 3 sao là tích cực trong khi người khác cho rằng đó là mức trung bình.
- Yếu tố cá nhân: Bình luận tích cực về sách nhưng kèm theo nhận xét cá nhân tiêu cực ("Sách hay nhưng mình không thích thể loại này").

Vì các lý do trên mà quá trình gán nhãn của nhóm gặp một số khó khăn nhất định trong việc lên bảng hướng dẫn, thống nhất nhãn trong quá trình gán vì tính chất review sách dài và khó nắm bắt, cần nhiều thời gian để đánh giá,...

5 Thí nghiệm

Ở phần này, nhóm sẽ thực hiện hai module riêng biệt để phù hợp với hai yêu cầu khuyến nghị sách dựa trên đánh giá và phân loại cảm xúc của đánh giá sách để phù hợp với bài toán "Khuyến nghị sách dựa trên thông điệp trong các đánh giá từ người dùng".

Module đầu tiên là về truy xuất dữ liệu (Information Retrieval) và module thứ hai là phân tích cảm xúc. Do không có kinh phí và thời gian để huấn luyện lại một mô hình sinh văn bản gợi ý nội dung sách nên nhóm sẽ sử dụng API có sẵn của do Google AI Studio cung cấp là các model hỏi đáp như Gemini 1.5 Pro, Gemini 1.5 Flash, Gemini 1.0 Pro.

5.1 Phương pháp thực nghiệm

5.1.1 Module truy vấn

LLM

Đầu tiên từ truy vấn của người dùng, nhóm sẽ xây dựng prompt và tìm đến các prompt tương tự trong bộ dữ liệu Q&A để lấy ra các nội dung gợi ý tương ứng. Sau đó, các nội dung gợi ý sẽ được sử dụng để truy vấn đến cơ sở dữ liệu để tìm ra các tựa sách phù hợp với truy vấn đầu vào nhất.

Về phương pháp truy vấn, nhóm đã thực nghiệm hai phương pháp khác nhau để thử nghiệm tính hiệu quả của từng phương pháp, tổng quan về các phương pháp như sau:

- BM25
- Sentence Embedding

BM25

Okapi BM25 là hàm xếp hạng các kết quả tìm được theo mức độ phù hợp với một truy vấn nhất định được dùng trong truy vấn và truy xuất thông tin. Hàm được phát minh vào những năm 1970-1980 và xếp hạng dựa trên mô hình truy vấn xác suất và đánh giá trên mức độ liên quan giữa văn bản với truy vấn. BM25 (best match 25) được mở rộng từ mô hình Okapi, là phương pháp xếp hạng tựa như tf-idf, được sử dụng rộng rãi trong tìm kiếm với hai

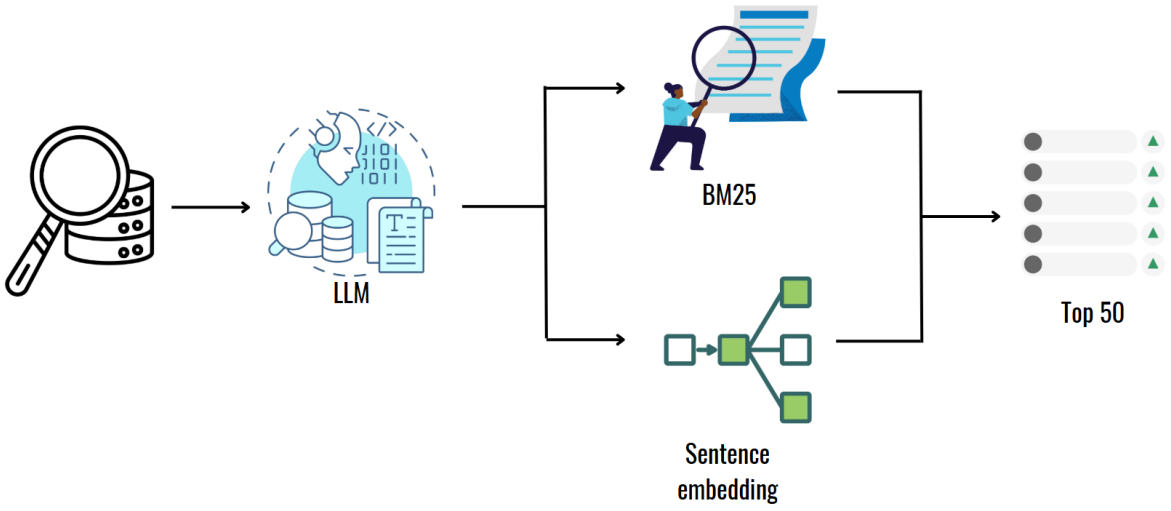

Figure 4: Các bước thực hiện module truy vấn

thông số điều chỉnh chính là k1 (độ nhạy của thuật toán với tần suất từ) và b (mức độ ảnh hưởng của độ dài tài liệu) (Robertson and Zaragoza, 2009).

Công thức cơ bản của BM25 được thể hiện như sau:

$$\text{score}(Q, D) = \tag{1}$$

$$\sum_{t \in Q} IDF(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{2}$$

Trong đó:

- Q: Truy vấn chứa các từ t,
- D: Văn bản được đánh giá,
- $f(t, D)$: Tần suất xuất hiện của từ t trong văn bản D,
- k_1: Tham số điều chỉnh độ nhạy với tần suất từ,
- b: Tham số điều chỉnh độ dài của văn bản,
- $|D|$: Độ dài của văn bản D,
- avgdl: Độ dài trung bình của các văn bản trong tập dữ liệu,
- $IDF(t)$: Độ đo nghịch đảo tần suất của từ t (Inverse Document Frequency).

$$IDF(t) = \log\left(\frac{N - n(t) + 0.5}{n(t) + 0.5}\right) + 1$$

(Sparck Jones, 1972)

  - N: Tổng số tài liệu trong tập dữ liệu.
  - $n(t)$: Số tài liệu chứa từ t.

BM25 được biết đến với hiệu suất cao trong các bài toán xếp hạng tài liệu, đặc biệt khi áp dụng trong các hệ thống truy vấn văn bản như Elasticsearch và Lucene.

Sentence embedding

Sentence embedding trong ngữ cảnh về xử lý ngôn ngữ tự nhiên (NLP – Natural Language Processing) là việc biểu diễn câu dưới dạng vector số hóa trong không gian liên tục, sao cho các câu có ý nghĩa hoặc ngữ cảnh tương tự được biểu diễn gần nhau. Đây là kỹ thuật quan trọng trong các bài toán xử lý ngôn ngữ tự nhiên vì nó giúp máy tính "hiểu" được mối quan hệ giữ các từ, thường được dùng trong tìm kiếm ngữ nghĩa, dịch máy, và xếp hạng văn bản. (Le and Mikolov, 2014) Các phương pháp xây dựng sentence embedding phổ biến bao gồm:

- Bag-of-Words (mô hình túi từ): Tổng hợp vector của các từ trong câu mà không quan tâm đến trật tự, ngữ pháp và ngữ cảnh nhưng vẫn giữ được tính đa dạng.
- Neural Sentence Embeddings: Sử dụng các mô hình học sâu như BERT, RoBERTa hoặc

Sentence-BERT (SBERT) để học biểu diễn vector hóa, trong đó ý nghĩa ngữ cảnh của câu được giữ lại thông qua cơ chế attention.

Sentence embedding thường được chuẩn hóa về kích thước vector (ví dụ: 512 hoặc 768 chiều) để dễ dàng so sánh hoặc sử dụng trong các mô hình downstream.

5.1.2 Module phân loại cảm xúc

Với module phân loại cảm xúc, nhóm thí nghiệm trên các kiến trúc cơ bản như BiGRU, BiLSTM và cả các kiến trúc transformer hiện đại như mô hình PhoBert và ViSoBERT. Với 2 model BiGRU và BiLSTM nhóm sẽ huấn luyện lại từ đầu, với 2 model ViSoBERT và PhoBert nhóm sẽ tiến hành fine-tune lại trên bộ dữ liệu của nhóm để đạt được hiệu suất tốt nhất.

5.2 Độ đo đánh giá

Trong đồ án này, nhóm dùng các độ đo:

5.2.1 Accuracy

Accuracy (độ chính xác) là một thước đo đánh giá hiệu suất phổ biến, đặc biệt trong các bài toán về phân loại. (Mosteller, 1948) Giá trị accuracy biểu thị tỷ lệ dự đoán chính xác trên tổng số dự đoán và được xác định bởi công thức:

$$\text{Accuracy} = \frac{\text{Số lượng dự đoán đúng}}{\text{Tổng số dự đoán}}$$

5.2.2 F-1 score

F1-score là trung bình điều hòa của Precision (độ chính xác) và Recall (độ nhạy). F1 score đặc biệt hữu ích khi bạn cần cân bằng giữa Precision và Recall, đặc biệt trong các bài toán mà dữ liệu không cân bằng. (van Rijsbergen, 1979) Công thức tính F1-score là:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.2.3 Top-k Accuracy

Top-k Accuracy được dùng để đo lường tỷ lệ các truy vấn mà tài liệu đúng xuất hiện trong top k kết quả trả về của hệ thống tìm kiếm. Công thức tính Top-k Accuracy là:

$$\text{Top-k Accuracy} = \frac{1}{N} \sum_{i=1}^{N} I(y_i \in \text{Top-}k(i))$$

Trong đó:

- N là số lượng truy vấn.
- y_i là tài liệu đúng cho truy vấn thứ i.
- I là hàm chỉ thị, với $I(y_i \in \text{Top-}k(i)) = 1$ nếu tài liệu đúng y_i nằm trong top k của truy vấn i, ngược lại $I = 0$.

5.3 Các thông số cài đặt

5.3.1 Module truy vấn

Ở module này nhóm sử dụng thuật toán BM25 của thư viện Okapi, về phần Sentence embedding nhóm sử dụng mô hình 'dangvantuan/vietnamese-embedding' (Dang et al., 2020) - mô hình được huấn luyện riêng cho tiếng Việt. Các mô hình được đều được công khai trên nền tảng hugging face - một thư viện mã nguồn mở cho các lập trình viên, nhà nghiên cứu trên khắp thế giới.

5.3.2 Module phân loại cảm xúc

Trong module này, khi huấn luyện (train) và tinh chỉnh (fine-tune) lại các mô hình nhóm điều chỉnh các tham số: Optimizer (Trình tối ưu hóa), Learning Rate (lr) - Tốc độ học, Weight Decay (Giảm trọng số), Loss Function (Hàm mất mát), Scheduler (Bộ điều chỉnh tốc độ học), Epochs (Số vòng lặp huấn luyện) và Warmup Steps. Cụ thể được thể hiện trong các bảng (table) 1 và 2. Các việc này đều được thực hiện trên GPU P100 miễn phí trên nền tảng Kaggle.

Mô Hình	ViSoBERT / PhoBERT
Optimizer	AdamW
Learning Rate (lr)	2e-5
Weight Decay	0.01
Loss Function	N/A
Scheduler	Linear Warmup
Epochs	5
Warmup Steps	2

Table 1: Thông số huấn luyện của mô hình ViSoBERT / PhoBERT

6 Kết quả thực nghiệm

Để đánh giá kết quả của hệ thống nhóm chia làm 2 phần đánh giá riêng biệt trên 2 tác vụ là phân loại cảm xúc và truy xuất thông tin.

Ở tác vụ phân loại cảm xúc nhóm đã huấn luyện và tinh chỉnh trên 80% bộ dữ liệu và đánh giá hiệu suất trên 20% dữ liệu còn lại. Bảng dưới đây thể hiện hiệu suất của các mô hình trên tác

Mô Hình	BiGRU / BiLSTM
Optimizer	Adam
Learning Rate (lr)	0.01
Weight Decay	N/A
Loss Function	Cross-Entropy
Scheduler	N/A
Epochs	20
Warmup Steps	N/A

Table 2: Thông số huấn luyện của mô hình BiGRU / BiLSTM

vụ phân loại cảm xúc. Các mô hình như BiLSTM và BiGRU thể hiện khá tệ trên bộ dữ liệu của nhóm độ chính xác lần lượt là 56% và 52%. Với 2 mô hình winrax/phobert và SCD-AI/Vietnamese-Sentiment-visobert thì hiệu suất đã cải thiện rõ rệt lần lượt là 71% và 78%. Kết quả chi tiết được thể hiện ở bảng 3.

Với tác vụ truy vấn, nhóm đã thử nghiệm trên các trường hợp như loại bỏ từ dừng (stop words), chỉ sử dụng câu truy vấn hoặc sử dụng cả câu truy vấn và câu trả lời từ mô hình ngôn ngữ lớn. Ở BM25 nhóm đã sử dụng công cụ pyvi để tách từ cho Tiếng Việt. Bảng 4 là kết quả khi truy vấn bằng các kiến trúc mà nhóm đã nêu ra ở phần trước. Ở BM25 nhóm đã sử dụng công cụ pyvi để tách từ. tabularx

Từ bảng kết quả có thể thấy được việc tách từ ở cả 2 trường hợp loại hoặc giữ từ dừng đều cho kết quả cao hơn. Bên cạnh đó việc loại bỏ từ dừng sẽ giúp thông tin được cô đọng lại nên cho kết quả khả quan hơn. Ngoài ra việc sử dụng Embedding cho kết quả tệ hơn hẳn so với các phương pháp sử dụng BM25.

7 Kết luận

Mặc dù bài toán phân loại cảm xúc đã được nghiên cứu từ lâu, các mô hình hiện tại vẫn chưa thể bao quát hết sự đa dạng của dữ liệu trong xã hội. Đặc biệt, việc xử lý cảm xúc trong các đánh giá sách đặt ra một thách thức riêng biệt. Khác với các dạng văn bản ngắn gọn, đánh giá sách thường sử dụng ngôn ngữ phong phú, giàu hình ảnh, ẩn dụ, và các biện pháp tu từ khác. Bối cảnh của văn bản cũng phức tạp hơn, kéo dài hơn, đòi hỏi mô hình phải có khả năng hiểu được ngữ cảnh rộng hơn để đưa ra phân loại chính xác. Điều này gây khó khăn cho việc khái quát hóa và áp dụng các mô hình hiện có.

Nhóm chúng tôi tập trung vào bài toán phân loại cảm xúc trong ngữ cảnh đánh giá sách, nhận thấy rõ những hạn chế của các phương pháp tiếp cận truyền thống. Sự phức tạp của ngôn ngữ văn chương, sự đa dạng trong cách diễn đạt cảm xúc, và độ dài của văn bản đánh giá tạo ra một bài toán đầy thách thức. Ví dụ, một câu văn có thể chứa đựng nhiều tầng ý nghĩa, vừa khen ngợi vừa phê bình, hoặc sử dụng ẩn dụ để thể hiện cảm xúc một cách gián tiếp. Điều này đòi hỏi mô hình không chỉ đơn thuần phân tích từ ngữ mà còn phải hiểu được ý nghĩa sâu xa và ngữ cảnh của toàn bộ văn bản.

Về vấn đề truy xuất thông tin, nhóm nhận thấy tiềm năng to lớn trong việc cải thiện độ chính xác bằng cách tinh chỉnh (fine-tune) các mô hình Sentence Embedding. Trong tương lai, nhóm sẽ tập trung vào việc nghiên cứu và áp dụng các kỹ thuật fine-tuning tiên tiến nhất cho các mô hình Sentence Embedding, như BERT, RoBERTa, hay Sentence-BERT. Mục tiêu là tạo ra một hệ thống có khả năng biểu diễn ngữ nghĩa của câu một cách chính xác và hiệu quả hơn, từ đó nâng cao độ chính xác của việc truy vấn và phân loại cảm xúc. Việc tinh chỉnh này sẽ được thực hiện trên tập dữ liệu đánh giá sách chuyên biệt, giúp mô hình thích ứng tốt hơn với đặc thù của loại văn bản này.

Ngoài ra, nhóm cũng sẽ xem xét kết hợp các phương pháp khác, chẳng hạn như:

- Sử dụng các mô hình ngôn ngữ lớn (LLM): Khám phá khả năng của các LLM như GPT-3, PaLM trong việc hiểu và phân loại cảm xúc trong văn bản dài.
- Áp dụng các kỹ thuật attention: Tăng cường khả năng của mô hình trong việc tập trung vào các phần quan trọng của văn bản, đặc biệt là các cụm từ hoặc câu thể hiện cảm xúc.
- Xây dựng tập dữ liệu chuyên biệt: Tiếp tục mở rộng và làm giàu tập dữ liệu đánh giá sách, bao gồm cả các đánh giá bằng tiếng Việt và các ngôn ngữ khác, để mô hình được huấn luyện trên một tập dữ liệu đa dạng và phong phú hơn.

Bằng cách kết hợp các phương pháp này, nhóm hy vọng sẽ tạo ra một hệ thống phân loại cảm xúc mạnh mẽ và chính xác hơn, đặc biệt là trong bối cảnh phức tạp của đánh giá sách.

Lời cảm ơn

Chúng tối xin cảm ơn sự hướng dẫn nhiệt tình của thầy Huỳnh Văn Tín trong môn học và đồ án này.

Model	Accuracy	F1	Precision	Recall
BiGru	0.56	0.55	0.56	0.56
BiLSTM	0.52	0.52	0.52	0.51
winrax/phobert	0.7177	0.7119	0.7106	0.7177
SCD-AI/Vietnamese-Sentiment-visobert	**0.7842**	**0.7817**	**0.7804**	**0.7842**

Table 3: Model performance comparison.

Method	Top-5 Acc	Top-10 Acc	Top-20 Acc	Top-50 Acc
BM25-query	0.41	0.5	0.54	0.61
BM25-query+answer	0.38	0.45	0.54	0.63
BM25-query-stopwords	0.48	0.46	0.62	0.71
BM25-query+answer-stopwords	**0.48**	**0.57**	**0.63**	**0.74**
Embeddings	0.29	0.37	0.45	0.56

Table 4: Accuracy for different methods

References

Hamad Alharbi and Erwin de Doncker. 2019. Sentiment analysis on book reviews using convolutional neural network (cnn) and long short-term memory (lstm) hybrid model. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(9):269–278.

Van Tuan Dang et al. 2020. Vietnamese embedding models for text representation and natural language understanding. *GitHub Repository*.

Vong Ho, Duong Nguyen, Danh Nguyen, Linh Pham, Duc-Vu Nguyen, Kiet Nguyen, and Ngan Nguyen. 2020. *Emotion Recognition for Vietnamese Social Media Text*, pages 319–333.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1188–1196. PMLR.

Frederick Mosteller. 1948. Questions and answers: How do you measure "goodness" in classification problems? *The American Statistician*, 2(5):14–16.

Kiet Van Nguyen, Vu Duc Nguyen, Phu Xuan-Vinh Nguyen, Tham Thi-Hong Truong, and Ngan Luu-Thuy Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24. IEEE.

Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. 2023. ViSoBERT: A pre-trained language model for Vietnamese social media text processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5191–5207, Singapore. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. In *Journal of Documentation*, volume 28, pages 11–21.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann.