UIT Trường Đại học Công nghệ Thông tin | Khoa Khoa học và Kỹ thuật Thông tin

DS307 - NHÓM 10

Social data analysis

DS107 - Social Data Analysis

BOOK RECOMMENDATION SYSTEM WITH SENTIMENT INSIGHT

Nhóm 10:

Bùi Bảo Trân - 22521503

Phạm Hồng Trà - 22521495

Châu Nguyễn Tri Vũ - 22521687

Nguyễn Lê Thanh Minh - 22520875

GVHD: Thầy Huỳnh Văn Tín

NỘI DUNG

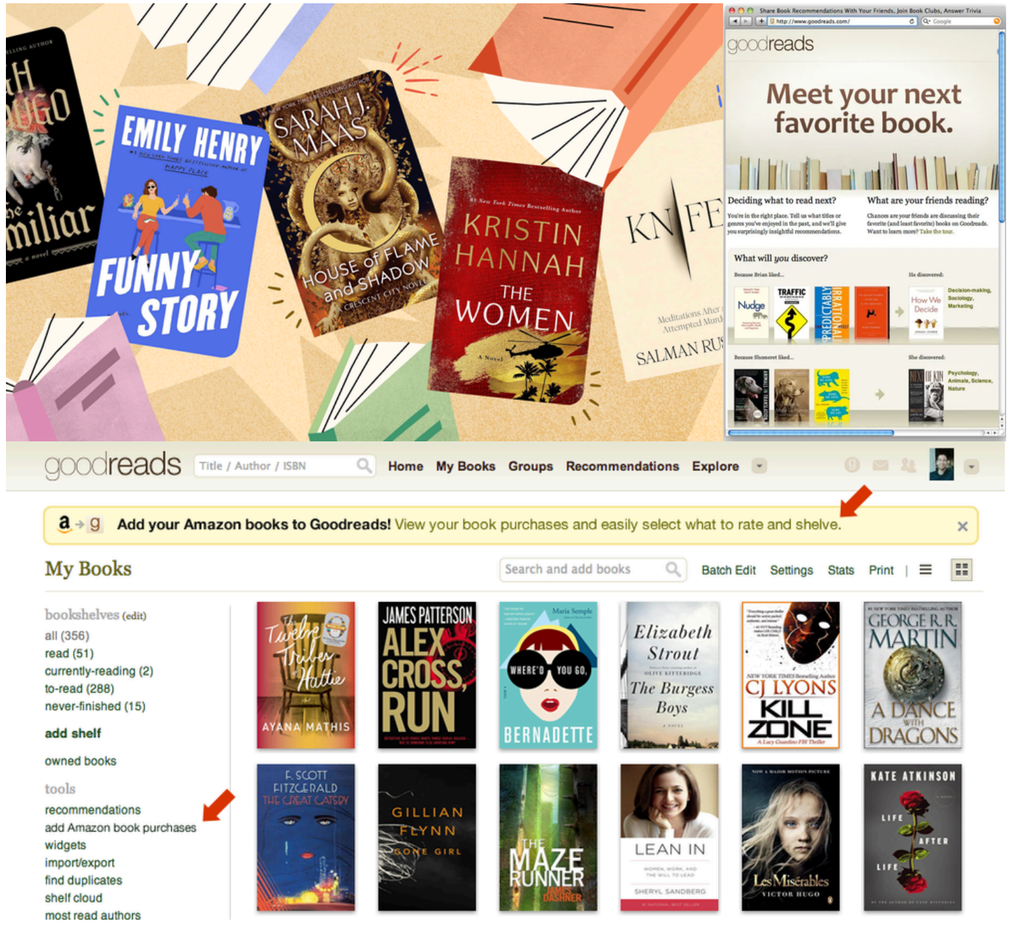
1. GIỚI THIỆU

Lượng thông tin đồ sộ trên Internet + phương pháp tra cứu thông tin tiên tiến

↳ dễ dàng hơn trong tìm thông tin, đánh giá về sách

NHƯNG

Hiện đại = "hại điện" ?



↳ Đề tài ***"Khuyến nghị sách dựa trên thông điệp trong các đánh giá từ người dùng" :*** áp dụng NLP và các model như BiGRU, BiLSTM, Bert, vietnamese-embedding,..

2. CÁC CÔNG TRÌNH LIÊN QUAN

UIT-VSFC	Sentiment Analysis on Book Reviews Using Convolutional Neural Network (CNN) Long Short-Term Memory (LSTM) Hybrid	ViSoBERT
• Bộ dữ liệu Tiếng Việt, bao gồm phản hồi từ sinh viên • Mục tiêu: phân loại cảm xúc + quan điểm • 3 nhãn: Tích cực, tiêu cực, trung tính	• Xây dựng mô hình kết hợp CNN và LSTM • Mục tiêu: phân tích cảm xúc trong các đánh giá sách • Dữ liệu thu thập từ Goodreads với 3 thể loại sách khác nhau	• Mô hình học sâu, được phát triển dựa trên kiến trúc BERT • Được thiết kế dành riêng cho Tiếng Việt: -- Tối ưu hóa văn bản từ MXH -- Hiểu ngữ cảnh -- Hiệu suất cao

3.Tổng quan bài toán

Gồm 2 bài toán nhỏ:

Phân loại cảm xúc

Mục tiêu: xác định cảm xúc ẩn chứa trong nội dung đánh giá của người dùng về một cuốn sách.

- **Input:** Một câu hoặc đoạn văn bản là review về cuốn sách.
- **Output:** Một trong 3 nhãn cảm xúc (Positive, Negative, Neutral).

Truy vấn thông tin

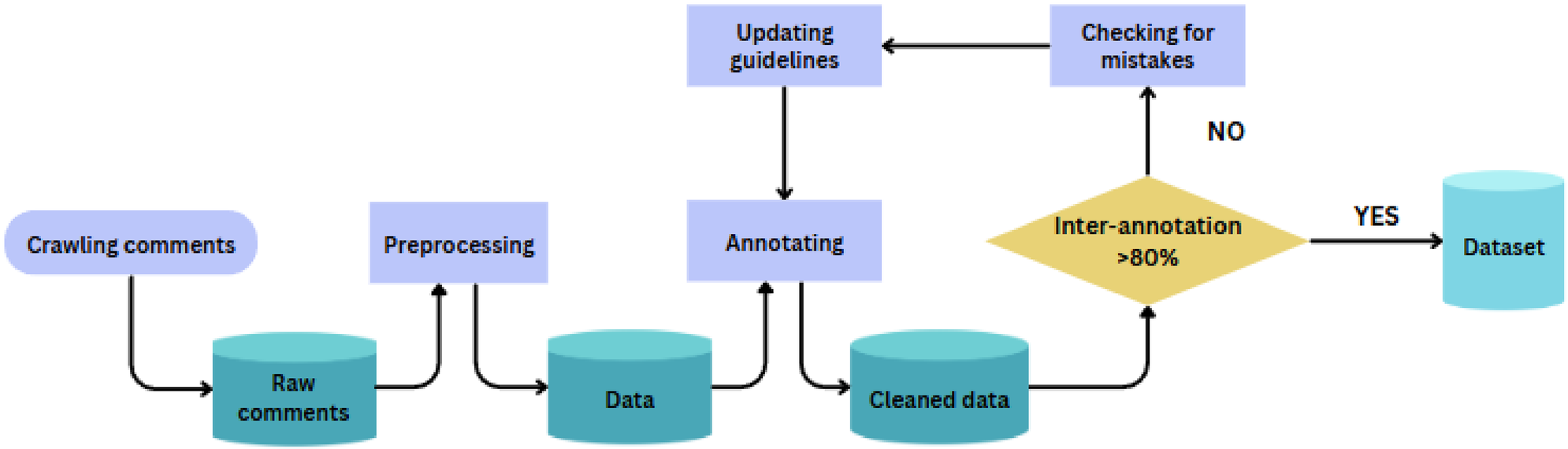
Mục tiêu: đề xuất những cuốn sách có quan đến một câu hỏi hoặc yêu cầu cụ thể của người dùng, dựa trên nội dung các review (đánh giá) sách

- **Input:** Câu truy vấn của người dùng.
- **Output:** Những tựa sách liên quan đến truy vấn.

4. DỮ LIỆU

Bao gồm 2 bộ dữ liệu tự xây dựng:

1. Bộ dữ liệu phân loại cảm xúc review sách



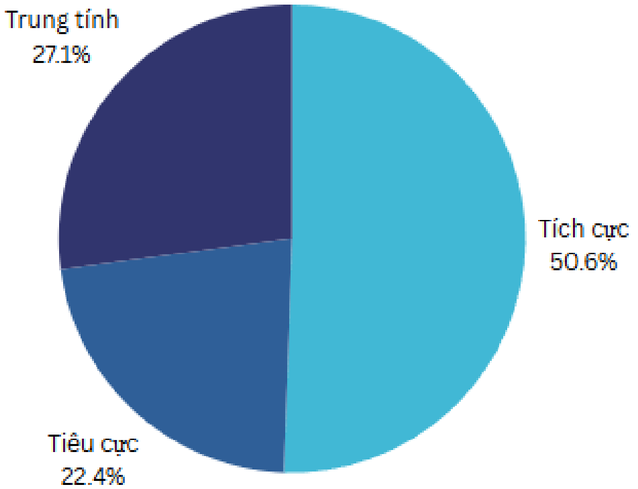
Quy trình xây dựng bộ dữ liệu phân loại cảm xúc review sách

4. DỮ LIỆU

1. Bộ dữ liệu phân loại cảm xúc review sách

- Bao gồm **3 nhãn**: tích cực, tiêu cực và trung tính
- Xây dựng **guidelines gán nhãn** cho **17,440 câu review sách** → Độ đồng thuận: **0.82**

- Tổng cộng **663 cuốn** sách với **8 thể loại** chính
- **Nhận xét:**



-- Độ dài lớn
-- Ngôn ngữ đa dạng trong cách sử dụng từ và diễn đạt
-- Chứa các câu tu từ, chơi chữ
-- Khác biệt trong thang điểm đánh giá
-- Yếu tố cá nhân

- **Khó khăn:** 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,...

4. DỮ LIỆU

1. Bộ dữ liệu phân loại cảm xúc review sách

Ví dụ:

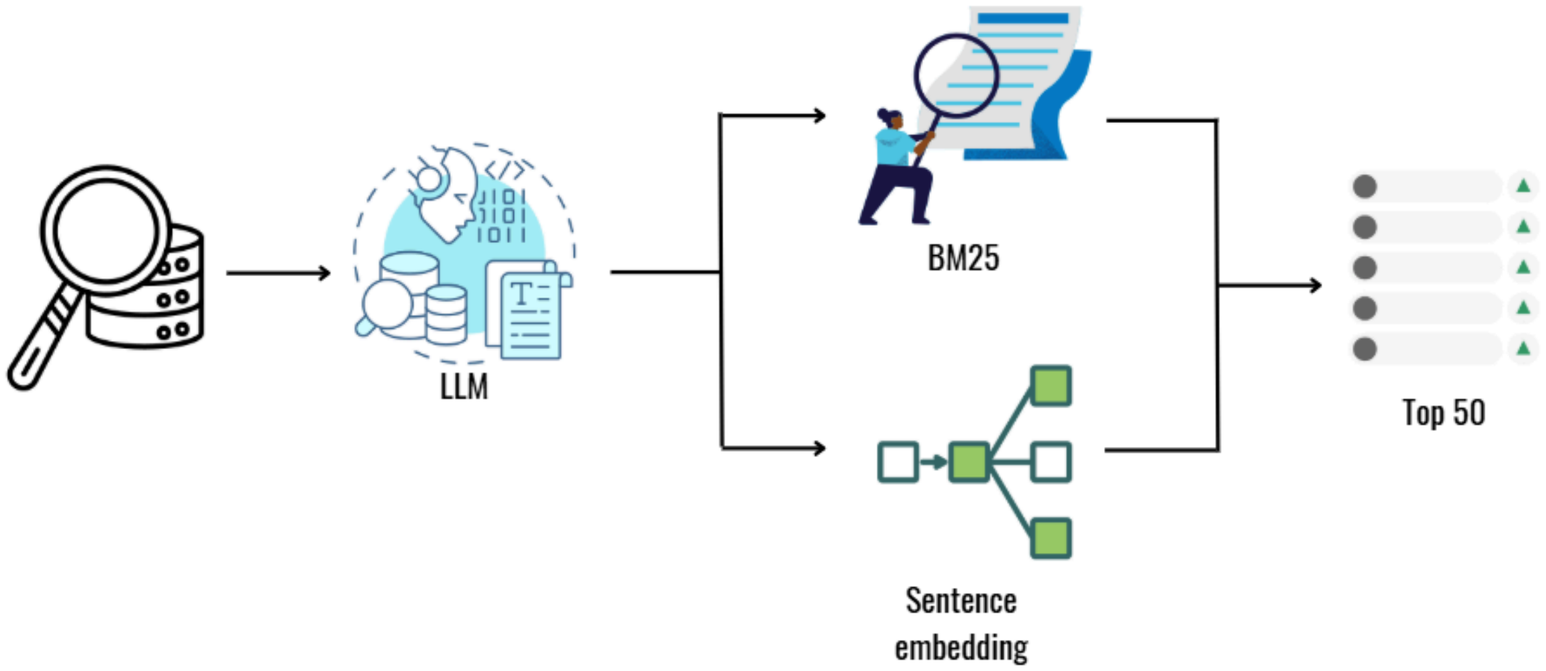
- Tích cực: *"Xúc động và nể thầy lắm. Hoc sinh nên đọc, người lớn cũng nên đọc để hiểu hơn và tạo động lực cho con cháu yêu quý sự học. Mình mới đọc tập Tôi đi học mà đã hình dung phần nào tính cách của thầy rồi. Một người luôn nỗ lực, kiên trì trong mọi việc dù là nhỏ nhất."*
- *Tiêu cực: "Mình chỉ đọc đến trang 45 là không muốn đọc tiếp, cách kể rất nhảm nhí, không cuốn hút và không ấn tượng."*
- Trung tính: *"Thấy các bạn khen nên tò mò đọc thử. Cũng có một số đoạn đọc có giá trị, nhưng phần lớn khá là nhàm."*

2. Bộ dữ liệu Q&A

- Bao gồm các câu hỏi người dùng đặt ra và có liên quan đến nội dung sách, giúp xác định nhu cầu tìm kiếm sách.
- Gồm **200 bộ câu hỏi và câu trả lời**

5. PHƯƠNG PHÁP THÍ NGHIỆM

Truy vấn thông tin



6. KẾT QUẢ THÍ NGHIỆM

Phân loại cảm xúc

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
5CD-AI/Vietnamese-Sentiment-visobert	**0.7842**	**0.7817**	**0.7804**	**0.7842**

Truy vấn thông tin

Method	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy
BM25	0.3	0.35	0.4
BM25+Embedding	0.32	0.40	0.45
BM25+Embedding+Cross reranker	**0.4**	**045**	**0.55**

7. KẾT LUẬN

Phân loại cảm xúc

- Các model như GRU, LSTM thuần hoạt động không tốt trên ngữ cảnh dài.
- Model được fine-tune trên ViSoBert cho ra kết quả tốt nhất nhưng độ chính xác vẫn chưa cao lắm.

Truy vấn thông tin

- Sử dụng tổ hợp BM25-Sentence Embedding - Cross encoder cho ra kết quả tốt nhất.

Bộ dữ liệu: Có nhiều từ ngữ khá phức tạp, biện pháp tu từ nhiều, độ dài của các câu là khá lớn -> gây khó khăn cho mô hình.

THANK YOU

FOR LISTENING