VINEWREC: BỘ DỮ LIỆU TIN TỨC XÃ HỘI TIẾNG VIỆT CHẤT LƯỢNG CAO DÀNH CHO HỆ THỐNG GỢI Ý DỰA TRÊN NỘI DUNG VĂN BẢN

Nguyễn Ngọc Yến Nhi[1,2], **Nguyễn Đức Anh**[1,2],
Nguyễn Thị Huyền Trang[1,2], **Huỳnh Văn Tín**[1,2]

[1] *Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam*

[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

Hệ thống đề xuất tin tức cá nhân hóa đóng vai trò quan trọng trong việc giúp người dùng tiếp cận thông tin phù hợp. Tuy nhiên, các hệ thống này vẫn còn hạn chế với tiếng Việt do thiếu dữ liệu chất lượng cao. Nghiên cứu này giới thiệu ViNewRec, một bộ dữ liệu tích hợp nội dung tin tức và tương tác người dùng như bình luận, lượt đọc, chia sẻ. ViNewRec hỗ trợ phát triển các hệ thống gợi ý cá nhân hóa và xử lý ngôn ngữ tự nhiên (NLP). Hai nhiệm vụ chính được thực hiện gồm gợi ý dựa trên nội dung và gợi ý dựa trên cộng tác. Mô hình Meta-LLaMA đạt Precision@1 cao nhất (0.3290), vượt Baseline PhoBERT 18.4%. Trong gợi ý dựa trên cộng tác, Text-based NeuCF cho thấy hiệu suất vượt trội với Recall và Precision cao hơn NeuCF cơ bản tại top@5 và top@10. Kết quả nhấn mạnh tiềm năng của ViNewRec trong việc cải thiện hệ thống gợi ý tin tức tiếng Việt và mở ra hướng nghiên cứu mới trong xử lý ngôn ngữ tự nhiên và hệ thống đề xuất.

Từ khóa: ViNewRec, hệ thống gợi ý, lọc nội dung, lọc cộng tác, NLP, dữ liệu tiếng Việt.

1 Giới thiệu

Báo điện tử đã trở thành một phần không thể thiếu trong đời sống hiện đại, cung cấp thông tin nhanh chóng, chính xác và đa dạng cho độc giả. Tại Việt Nam, các nền tảng như VnExpress[1] không chỉ đóng vai trò là nguồn thông tin chính thống mà còn là nơi tập trung lượng lớn người dùng thường xuyên. Tuy nhiên, với sự phát triển nhanh chóng của nội dung số, người dùng phải đối mặt với tình trạng quá tải thông tin, dẫn đến khó khăn trong việc tìm kiếm nội dung phù hợp với nhu cầu cá nhân. Để giải quyết vấn đề này, các hệ thống đề xuất tin tức đã được phát triển nhằm cá nhân hóa nội dung và cải thiện trải nghiệm người dùng.

Mặc dù đã có nhiều nghiên cứu về hệ thống gợi ý tin tức trên thế giới, phần lớn tập trung vào dữ liệu tiếng Anh và ít chú ý đến ngôn ngữ tiếng Việt. Điều này dẫn đến một số thách thức trong việc phát triển và triển khai hệ thống đề xuất tin tức tại Việt Nam. Thứ nhất, dữ liệu tiếng Việt chất lượng cao còn rất hạn chế, đặc biệt là các bộ dữ liệu tích hợp cả nội dung tin tức và thông tin tương tác người dùng. Thứ hai, tiếng Việt có cấu trúc ngữ pháp phức tạp, khó khăn trong phân tách từ, và sự linh hoạt trong ngữ nghĩa, khiến việc xử lý ngôn ngữ tự nhiên (NLP) trở nên phức tạp hơn. Ngoài ra, bài toán cold-start—khi bài viết mới xuất hiện mà chưa có dữ liệu hành vi người dùng—là một thách thức lớn trong việc duy trì hiệu suất và tính cá nhân hóa của hệ thống.

Trước những hạn chế này, nghiên cứu giới thiệu **ViNewRec**, một bộ dữ liệu tiếng Việt phong phú kết hợp nội dung tin tức và tương tác người dùng, được thu thập từ nền tảng VnExpress. Bộ dữ liệu bao gồm các bài viết chất lượng cao thuộc các lĩnh vực như kinh tế, giáo dục, khoa học, công nghệ và văn hóa. ViNewRec không chỉ cung cấp dữ liệu phong phú cho các nghiên cứu học máy và xử lý ngôn ngữ tự nhiên, mà còn hỗ trợ phát triển các hệ thống gợi ý tin tức cá nhân hóa. Bên cạnh đó, nghiên cứu còn tập trung khám phá các giải pháp giảm thiểu thiên lệch (bias) trong hệ thống gợi ý, nhằm đảm bảo tính công bằng và hiệu quả trong việc phục vụ người dùng.

Nghiên cứu này có bốn đóng góp chính, gồm:

1. Đầu tiên, nghiên cứu giới thiệu **ViNewRec**, bộ dữ liệu tiếng Việt đầu tiên tích hợp nội dung tin tức và thông tin tương tác người dùng, cung cấp nền tảng mạnh mẽ cho các nghiên cứu hệ thống gợi ý và NLP.

2. Thứ hai, kết quả thực nghiệm cho thấy ViNewRec vượt trội hơn các baseline trong việc cải thiện độ chính xác và khả năng cá nhân hóa của các hệ thống gợi ý tin tức.

3. Thứ ba, các nghiên cứu chuyên sâu trên bộ dữ

[1] https://vnexpress.net/

liệu, nhằm khai phá thêm thông tin trong lĩnh vực tin tức và bài viết của các chuyên gia.

4. Cuối cùng, nghiên cứu mở ra tiềm năng phát triển các hệ thống gợi ý đa phương tiện và đa ngôn ngữ, phù hợp với xu hướng quốc tế.

Phần còn lại của bài báo được tổ chức như sau. Phần 2 trình bày các nghiên cứu liên quan về hệ thống gợi ý và xử lý ngôn ngữ tự nhiên. Phần 3 mô tả chi tiết bộ dữ liệu **ViNewRec**, bao gồm quá trình thu thập, chuẩn hóa và làm sạch dữ liệu. Phần 4 giới thiệu các phương pháp gợi ý, bao gồm chiến lược dựa trên nội dung và lọc cộng tác. Phần 5 trình bày kết quả thực nghiệm và so sánh hiệu suất của các mô hình gợi ý. Phần 6 thảo luận ý nghĩa của nghiên cứu, các hạn chế hiện tại và đề xuất cải tiến. Cuối cùng, Phần 7 kết luận và định hướng nghiên cứu tương lai.

2 Các công trình liên quan

Hệ thống gợi ý văn bản đóng vai trò ngày càng quan trọng trong việc cải thiện trải nghiệm người dùng bằng cách cung cấp nội dung phù hợp với sở thích cá nhân. Sự phát triển của xử lý ngôn ngữ tự nhiên (NLP), các phương pháp học sâu, và đặc biệt là các mô hình ngôn ngữ lớn (LLMs) đã thúc đẩy đáng kể hiệu quả và khả năng cá nhân hóa của các hệ thống này. Trong phần này, chúng tôi phân tích các công trình liên quan theo bốn khía cạnh: dữ liệu, phương pháp, ngôn ngữ lớn, và hệ thống gợi ý.

2.1 Công trình liên quan đến dữ liệu

Dữ liệu đóng vai trò trung tâm trong việc xây dựng và đánh giá hệ thống gợi ý tin tức. Các tập dữ liệu chất lượng cao không chỉ cung cấp thông tin phong phú mà còn tạo điều kiện kiểm thử các mô hình mới. Dưới đây là các tập dữ liệu nổi bật, được thể hiện qua Bảng 1:

- **Engage Corpus:** Tập trung vào dữ liệu xã hội từ Reddit với hàng trăm nghìn bài đăng và bình luận. Dữ liệu này hỗ trợ nghiên cứu về tương tác xã hội và hệ thống gợi ý dựa trên văn bản (Cheng et al., 2022).

- **MIND (Microsoft News Dataset):** Một tập dữ liệu lớn chứa hơn 1 triệu người dùng và 160.000 bài viết tiếng Anh từ Microsoft News. MIND tích hợp nội dung văn bản như tiêu đề, tóm tắt, và sử dụng các mô hình NLP tiên tiến như BERT (Wu et al., 2020).

- **Adressa Dataset:** Tập dữ liệu tin tức tiếng Na Uy được thiết kế cho hệ thống gợi ý ngữ cảnh. Nó bao gồm các hành vi người dùng như lượt xem và thời gian đọc (Gulla et al., 2017).

- **ViNewRec:** Một tập dữ liệu tiếng Việt kết hợp nội dung tin tức và tương tác người dùng. Đây là tập dữ liệu đầu tiên tại Việt Nam hỗ trợ nghiên cứu NLP và hệ thống gợi ý trên ngữ cảnh tiếng Việt.

2.2 Công trình liên quan đến phương pháp

Các phương pháp xây dựng hệ thống gợi ý tin tức có thể chia thành ba nhóm chính: dựa trên mức độ tương đồng, học máy, và học sâu.

2.2.1 Phương pháp dựa trên mức độ tương đồng

Các phương pháp này tập trung vào việc đo lường mức độ tương đồng giữa hồ sơ người dùng và các bài viết tin tức dựa trên nội dung văn bản. Các nghiên cứu sử dụng các chỉ số tương đồng như khoảng cách Euclidean, Cosine similarity và Jaccard similarity để so sánh hồ sơ người dùng với các bài viết.

2.2.2 Phương pháp dựa trên học máy

Các phương pháp này sử dụng các kỹ thuật học máy để phân loại và gợi ý các bài viết dựa trên nội dung văn bản.

- **Hồi quy**: Kỹ thuật học máy này được sử dụng để phân loại và gợi ý các bài báo khoa học dựa trên nội dung của tóm tắt.

- **SULM (Sentiment Utility Logistic Model)**: Mô hình này không chỉ gợi ý bài viết mà còn xác định các khía cạnh liên quan nhất của bài viết đối với người dùng, dựa trên phân tích cảm xúc từ các đánh giá.

2.2.3 Công trình liên quan đến học sâu (Deep Learning)

Học sâu (Deep Learning) đã trở thành một trong những phương pháp cốt lõi trong việc xây dựng hệ thống gợi ý tin tức nhờ khả năng học các biểu diễn ngữ nghĩa phức tạp từ dữ liệu lớn. Các kỹ thuật học sâu chính bao gồm:

- **Nhúng từ (Word Embedding):** Các phương pháp như Word2Vec, fastText, và GloVe chuyển đổi từ hoặc cụm từ thành các vector số thực, nắm bắt mối quan hệ ngữ nghĩa. Mặc dù là kỹ thuật truyền thống, nhúng từ vẫn được sử dụng rộng rãi trong các hệ thống gợi ý.

Bảng 1: So sánh các tập dữ liệu hệ thống gợi ý văn bản

Tên tập dữ liệu	Ngôn ngữ	Quy mô	Mục tiêu sử dụng	Đặc điểm nổi bật
Engage Corpus	Tiếng Anh	Hàng trăm nghìn bài đăng	Dự đoán tương tác cộng đồng	Sử dụng lọc cộng tác
MIND (Microsoft News Dataset)	Tiếng Anh	Hơn 1 triệu người dùng, 160k bài viết	Gợi ý tin tức	Tích hợp các kỹ thuật SOTA
Adressa Dataset	Tiếng Na Uy	Hàng nghìn bài viết	Gợi ý tin tức	Tập trung ngôn ngữ NaUy
Globo Dataset	Tiếng Bồ Đào Nha	Hàng chục nghìn bài viết	Gợi ý tin tức	Thiếu thông tin từ nội dung
ViNewRec	Tiếng Việt	Hàng trăm nghìn bài viết	Gợi ý tin tức tiếng Việt	Cung cấp dữ liệu phong phú và hỗ trợ nghiên cứu NLP cho tiếng Việt

- **Mạng nơ-ron tích chập (Convolutional Neural Networks - CNN):** Được sử dụng để trích xuất đặc trưng cục bộ từ văn bản, như tiêu đề hoặc đoạn văn, mang lại hiệu quả cao trong các bài toán phân loại và gợi ý.
- **Mạng nơ-ron tái phát (Recurrent Neural Networks - RNN):** Các biến thể như LSTM và GRU hỗ trợ học biểu diễn chuỗi, đặc biệt hiệu quả trong việc dự đoán nội dung tiếp theo dựa trên lịch sử duyệt tin tức của người dùng.
- **Học tăng cường sâu (Deep Reinforcement Learning - DRL):** Kỹ thuật này giải quyết các bài toán thay đổi động trong dữ liệu tin tức, đồng thời tối ưu hóa gợi ý dựa trên phản hồi của người dùng.
- **Multi-Layer Perceptron - MLP:** Kết hợp các đặc trưng từ nội dung và tương tác người dùng, cho phép xếp hạng và cá nhân hóa các gợi ý một cách hiệu quả.
- **Transformer:** Một kỹ thuật tiên tiến dựa trên cơ chế tự chú ý (self-attention), giúp nắm bắt mối quan hệ ngữ nghĩa dài hạn giữa các từ hoặc câu. Transformer là nền tảng của nhiều mô hình ngôn ngữ lớn hiện nay.

Những kỹ thuật học sâu này đã được áp dụng hiệu quả trong hệ thống gợi ý tin tức trên các ngôn ngữ khác như tiếng Anh và tiếng Na Uy. Trong ngữ cảnh tiếng Việt, việc kết hợp các kỹ thuật học sâu với tập dữ liệu phong phú như ViNewRec có tiềm năng lớn trong việc cải thiện khả năng cá nhân hóa và tối ưu hóa gợi ý tin tức, đặc biệt khi kết hợp với các phương pháp hiện đại như Transformer.

2.3 Công trình liên quan đến hệ thống gợi ý

Hệ thống gợi ý tin tức đã giải quyết hiệu quả các bài toán như cá nhân hóa và xử lý dữ liệu lớn. Một số hệ thống tiêu biểu:

- **NRMS (Neural News Recommendation with Self-attention):** Sử dụng cơ chế self-attention để học các mối quan hệ giữa tiêu đề, nội dung bài viết và sở thích người dùng. Mô hình này đã đạt hiệu suất cao trong các bài toán gợi ý tin tức (Wu et al., 2019).
- **LSTUR (Long- and Short-term User Representation):** Kết hợp sở thích dài hạn và ngắn hạn của người dùng để tạo ra biểu diễn toàn diện hơn, giúp cải thiện khả năng cá nhân hóa (An et al., 2019).

2.4 Công trình liên quan đến ngôn ngữ lớn (LLMs)

Các mô hình ngôn ngữ lớn (LLMs) đã mang lại những tiến bộ vượt bậc trong xử lý ngôn ngữ tự nhiên (NLP) và hệ thống gợi ý tin tức. LLMs khai thác khả năng học các biểu diễn ngữ nghĩa từ khối lượng lớn dữ liệu, giúp tối ưu hóa các tác vụ liên quan đến gợi ý, tóm tắt, và dự đoán nội dung. Những kỹ thuật chính được sử dụng trong LLMs bao gồm:

- **Học biểu diễn ngữ nghĩa sâu (Deep Semantic Representation):** Sử dụng các tầng mã hóa để nắm bắt ngữ nghĩa phức tạp trong văn bản, giúp cải thiện chất lượng biểu diễn dữ liệu.
- **Tiền huấn luyện và tinh chỉnh (Pre-training and Fine-tuning):** Kỹ thuật tiền huấn luyện trên khối lượng lớn dữ liệu và tinh chỉnh trên tập dữ liệu cụ thể giúp LLMs thích nghi với các ngữ cảnh khác nhau, đặc biệt hiệu quả với dữ liệu tiếng Việt.
- **Chuyển giao tri thức đa ngôn ngữ (Cross-lingual Transfer Learning):** Tận dụng khả năng tổng quát hóa của LLMs để áp dụng trên

nhiều ngôn ngữ, bao gồm cả các ngôn ngữ ít tài nguyên như tiếng Việt.

- **Học tăng cường với phản hồi người dùng (Reinforcement Learning with Human Feedback - RLHF):** Kỹ thuật này giúp LLMs điều chỉnh gợi ý dựa trên phản hồi thực tế, tăng cường khả năng cá nhân hóa và độ chính xác.

Những kỹ thuật trên không chỉ cải thiện hiệu suất gợi ý mà còn giúp hệ thống thích nghi linh hoạt hơn với sự thay đổi của dữ liệu và hành vi người dùng. Trong bối cảnh tiếng Việt, việc ứng dụng LLMs kết hợp với tập dữ liệu chuẩn hóa như ViNewRec sẽ mở ra cơ hội mới để tối ưu hóa hệ thống gợi ý tin tức.

3 Bộ dữ liệu

3.1 Tổng quan về bộ dữ liệu

Để tạo điều kiện thuận lợi cho việc nghiên cứu đề xuất tin tức, chúng tôi đã xây dựng Bộ dữ liệu tin tức ViNewRec, một trong những trang tin tức lớn và uy tín tại Việt Nam. Bộ dữ liệu này được xây dựng bằng cách thu thập các bài báo được xuất bản trên nền tảng báo điện tử VnExpress. Nó chứa 463 bài báo và 11.954 dữ liệu bình luận của người dùng. Các bài báo trong tập dữ liệu này bằng tiếng Việt và người dùng chủ yếu là người Việt Nam. Chúng tôi đã lấy mẫu ngẫu nhiên 463 bài báo và 11.954 dữ liệu bình luận từ ngày 28 tháng 12 năm 2021 đến ngày 07 tháng 11 năm 2024. Bộ dữ liệu bao gồm các bài viết từ nhiều lĩnh vực như kinh tế, khoa học, giáo dục, công nghệ, văn hóa và xã hội. Các bài viết được lựa chọn dựa trên các tiêu chí cụ thể nhằm đảm bảo tính đa dạng, độ dài phù hợp và chất lượng nội dung cao. Cụ thể:

- Độ dài phù hợp: Các bài viết có độ dài trung bình từ 3286 đến 12306 từ, đảm bảo đủ nội dung để thực hiện các phân tích và thuật toán xử lý.
- Chất lượng nội dung: Chỉ thu thập các bài viết từ các chuyên mục chính với thông tin rõ ràng, ngữ pháp chính xác và do các chuyên gia hoặc cây viết nổi bật của VnExpress thực hiện.
- Đa dạng lĩnh vực: Tập dữ liệu bao gồm các lĩnh vực kinh tế, giáo dục, khoa học, công nghệ, đời sống văn hóa và chính sách, phản ánh bức tranh toàn diện về thông tin hiện nay.

3.2 Thu thập và xử lý dữ liệu

Chúng tôi tiến hành thu thập dữ liệu từ các chuyên gia trên nền tảng VNEpress. Dữ liệu sau khi được thu thập sẽ trải qua quy trình kiểm tra và tiền xử lý nghiêm ngặt nhằm đảm bảo tính chính xác và nhất quán. Các bước cụ thể bao gồm: loại bỏ thẻ HTML và các ký tự đặc biệt, chuẩn hóa dấu tiếng Việt để đồng nhất về mặt ngôn ngữ, phân đoạn câu và tách từ bằng các công cụ như VnCoreNLP, và loại bỏ các phần thông tin không liên quan.

Mỗi bài viết và đoạn văn đều được chú thích cẩn thận với các trường thông tin quan trọng như đường dẫn URL, phân loại theo danh mục từ trang web gốc, tiêu đề, tóm tắt, nội dung bài viết, ngày đăng và các chỉ số tương tác (bao gồm lượt đọc, chia sẻ và bình luận).

Sau khi hoàn tất thu thập, bước tiếp theo là làm sạch và chuẩn hóa văn bản để đảm bảo dữ liệu đầu vào có chất lượng tốt cho các thuật toán xử lý tiếp theo. Quy trình này diễn ra qua ba giai đoạn chính. Thứ nhất, chúng tôi loại bỏ các ký tự không cần thiết như ký tự đặc biệt, thẻ HTML và khoảng trắng thừa, giúp văn bản trở nên gọn gàng và dễ xử lý hơn. Tiếp theo, quá trình chuẩn hóa văn bản bao gồm chuyển toàn bộ nội dung về dạng chữ thường, sửa các lỗi chính tả đơn giản và xử lý các vấn đề liên quan đến encoding. Cuối cùng, văn bản được tách từ để xử lý ngữ nghĩa cơ bản bằng các công cụ chuyên biệt như VnCoreNLP, tạo tiền đề vững chắc cho các bước phân tích sâu hơn trong giai đoạn tiếp theo.

Bên cạnh việc làm sạch và chuẩn hóa văn bản, chúng tôi tiếp tục kiểm tra lại giai đoạn nhãn dữ liệu. Các bài viết được gắn nhãn theo các danh mục lĩnh vực chính như Kinh tế, Khoa học, Giáo dục, Công nghệ và Văn hóa. Các nhãn này được thực hiện dựa trên thông tin từ thẻ tags hoặc chuyên mục gốc của từng bài viết, đảm bảo tính chính xác và hiệu quả trong quá trình phân loại.

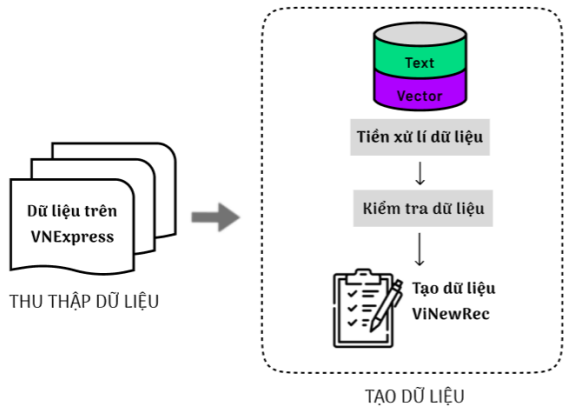
Kết quả của quá trình được thể hiện trong Hình 1 là một bộ dữ liệu đầy đủ và chi tiết, bao gồm các trường thông tin quan trọng như tiêu đề, tóm tắt, thân bài, ngày đăng và lượt tương tác. Bộ dữ liệu này được thiết kế để đáp ứng nhu cầu phân tích thống kê, trực quan hóa và xây dựng hệ thống đề xuất tin tức. Các thuộc tính cụ thể được mô tả chi tiết trong bảng 1.

3.3 Phân tích dữ liệu

Thống kê mô tả của bộ dữ liệu ViNewRec được trình bày trong bảng 2. Bộ dữ liệu này chứa 11.954

Tên thuộc tính	Mô tả
description	Một câu giới thiệu ngắn về bài báo
article_id	Mã định danh duy nhất cho từng bài báo
author_name	Tên của tác giả viết bài báo
author_url	Liên kết đến trang cá nhân hoặc thông tin của tác giả
author_description	Mô tả nghề nghiệp hoặc thông tin chi tiết về tác giả
publish_date	Ngày đăng tải bài báo
url	Liên kết dẫn đến bài báo trên trang web
category	Phân loại của bài báo theo chủ đề
tags	Các từ khóa hoặc tag liên quan đến nội dung bài báo
content	Nội dung chính của bài báo
avata_coment_href	Đường dẫn đến ảnh đại diện của người đọc để lại bình luận
user_comment	Nội dung bình luận của người đọc
nickname	Tên hoặc biệt danh của người dùng để lại bình luận
user_reacted	Số lượng tương tác của bình luận từ người đọc khác
time_com	Thời gian người đọc để lại bình luận trên bài báo
Title	Tiêu đề của bài báo

Bảng 2: Bảng mô tả các thuộc tính của bộ dữ liệu ViNewRec.



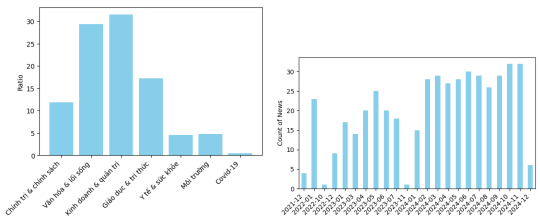
Hình 1: Qui trình thu thập dữ liệu ViNewRec.

Số lượng bài báo	463
Số lượng người đọc	11.954
Số lượng tác giả	151
Tổng tương tác	427.706
Số lượng nghề nghiệp của tác giả	111
Số lượng tag	831
Số lượng danh mục bài báo	7
Chiều dài trung bình tiêu đề	5.03 từ
Chiều dài trung bình mô tả bài báo	35.9 từ
Chiều dài trung bình nội dung	1178.86 từ
Chiều dài trung bình bình luận	44.58 từ

Bảng 3: Bảng thống kê mô tả của bộ dữ liệu ViNewRec.

người dùng và 463 bài viết tin tức dưới sự đóng góp của 151 tác giả. Chúng ta có thể thấy rằng tiêu đề tin tức thường rất ngắn và độ dài trung bình chỉ khoảng 5.03 từ.

Hình 1 cho thấy số lượng các bài viết và phân bố của các bài viết theo danh mục. Biểu đồ cho thấy sự phân bổ không đồng đều của các bài viết giữa các danh mục. Một số danh mục như Kinh doanh & quản trị, Văn hóa & lối sống chiếm tỷ lệ cao, trong khi các danh mục khác như Môi trường và Y tế & Sức khỏe có tỷ lệ thấp hơn. Điều này phản ánh sự tập trung vào các chủ đề được độc giả quan tâm nhiều hơn. Số lượng bài viết được đăng tải có xu hướng đồng đều nhau giữa các tháng. Điều này cho thấy tần suất xuất bản bài viết được duy trì ổn định theo thời gian, phản ánh chiến lược phân phối nội dung đều đặn của nền tảng để đảm bảo thông tin luôn được cập nhật liên tục cho độc giả.



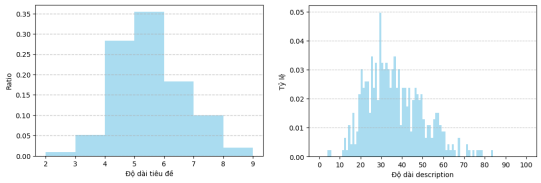
(a) Category (b) Month

Hình 2: Biểu đồ phân bố bài báo theo danh mục và thời gian.

Qua hình 2 ta thấy phần lớn tiêu đề bài báo có độ dài rơi vào khoảng 5-6 ký tự, cho thấy xu hướng tạo tiêu đề ngắn gọn, súc tích để thu hút sự chú ý của độc giả. Rất ít tiêu đề có độ dài trên 7 ký tự.

Phân tích độ dài của phần mô tả cho thấy phần lớn các mô tả bài báo có độ dài từ 20-50 ký tự. Điều này cho thấy xu hướng viết mô tả ngắn gọn, tóm lược ý chính của bài báo nhằm thu hút độc giả và cung cấp thông tin nhanh chóng. Một số mô tả bài báo có độ dài vượt trội (trên 150 ký tự) thường xuất hiện ở các bài viết chuyên sâu hoặc mang tính phân tích, yêu cầu mô tả chi tiết hơn để làm rõ nội dung. Việc tối ưu độ dài của description giúp cải thiện trải nghiệm người dùng, đồng thời ảnh hưởng đến khả năng hiển thị của bài viết trên các nền tảng tìm kiếm và mạng xã hội.

Nội dung bài báo chủ yếu tập trung trong khoảng 1000-1200 từ, đáp ứng nhu cầu thông tin không quá dài nhưng đầy đủ. Một số bài viết có nội dung dài hơn, thường thuộc các chuyên mục chuyên sâu hoặc phân tích. Bình luận của người dùng có xu hướng ngắn gọn, chủ yếu nằm trong khoảng 10-40 ký tự. Điều này phản ánh hành vi tương tác nhanh, tập trung vào ý kiến ngắn thay vì thảo luận chi tiết.



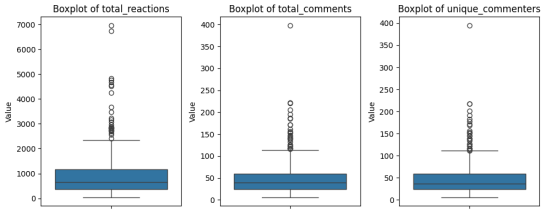
(a) Độ dài tiêu đề (b) Độ dài description

(c) Độ dài content (d) Độ dài user_comment

Hình 3: Các số liệu thống kê về độ dài của Tiêu đề, Mô tả bài báo, Nội dung bài báo, Nội dung bình luận của người đọc.

Ở hình 3 là ba biểu đồ boxplot thể hiện sự phân bố của ba chỉ số: tổng số phản ứng, tổng số bình luận, và số người bình luận. Ta thấy sự phân tán lớn trong các tương tác của người dùng. Phần lớn các giá trị của cả ba chỉ số đều tập trung ở mức thấp (gần gốc tọa độ), nhưng có sự xuất hiện của nhiều outliers (giá trị ngoại lệ), cho thấy có một số nội dung thu hút sự tương tác vượt trội so với mặt bằng chung. Ở biểu đồ boxplot tổng số tương tác trong mỗi bài viết, phần lớn giá trị nằm dưới 1000, nhưng một số trường hợp đạt mức trên 4000 - 7000. Điều này thể hiện sự khác biệt lớn trong mức độ thu hút phản ứng. Hai biểu đồ còn lại cũng có xu hướng tương tự, với phần lớn dữ liệu nằm ở mức thấp và các outliers nổi bật, đặc biệt ở các giá trị trên 100 - 400.

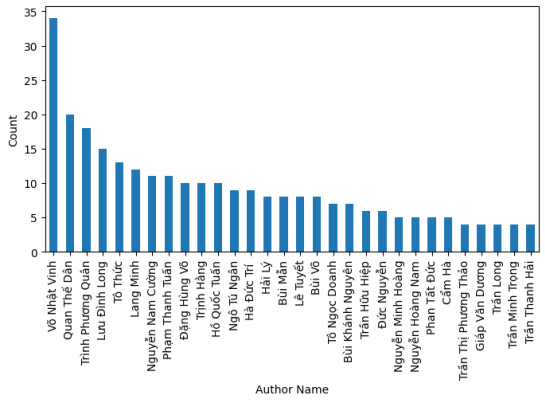
Dữ liệu cho thấy sự phân hóa lớn trong mức độ tương tác. Trong khi hầu hết các nội dung chỉ nhận được tương tác trung bình hoặc thấp, một số ít nội dung có khả năng tạo ra sự chú ý và tương tác rất cao, thể hiện qua các giá trị ngoại lệ trên cả ba biểu đồ.



Hình 4: Biểu đồ boxplot của tổng số reaction, tổng số bình luận, số lượng người thảo luận ở mỗi bài báo.

Biểu đồ ở hình 4 cho thấy 20% tác giả hàng đầu đóng góp số lượng bài viết lớn nhất trong tập dữ liệu. Một vài số ít tác giả có số lượng bài viết vượt trội so với các tác giả khác, cho thấy vai trò quan trọng của họ trong việc cung cấp nội dung cho nền tảng. Điều này có thể phản ánh mức độ hoạt động tích cực của những tác giả này hoặc họ phụ trách các chuyên mục có tần suất xuất bản cao.

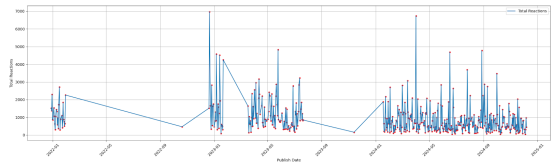
Tuy nhiên, sự tập trung số lượng bài viết vào một nhóm nhỏ tác giả có thể dẫn đến bias về góc nhìn và phong cách viết, làm giảm tính đa dạng trong nội dung. Vì vậy, cần xem xét thêm các yếu tố như chất lượng nội dung và phản hồi từ độc giả để đánh giá đầy đủ hơn.



Hình 5: Biểu đồ thể hiện sự phân bố số lượng bài báo của nhóm 20% tác giả có số bài viết nhiều nhất.

Sự biến động trong tổng số tương tác theo thời gian của mỗi bài báo là rất lớn. Có thể nhận thấy qua hình 5, các đợt tăng mạnh tương ứng với các thời điểm đăng tải bài viết có chủ đề nóng hoặc sự

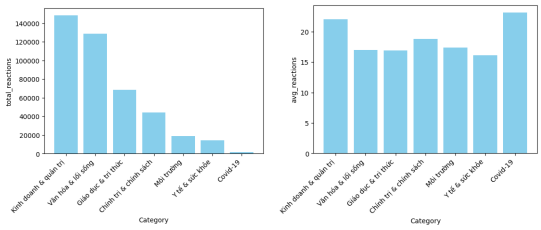
kiện nổi bật.



Hình 6: Biểu đồ thể hiện xu hướng tương tác của bài báo theo thời gian.

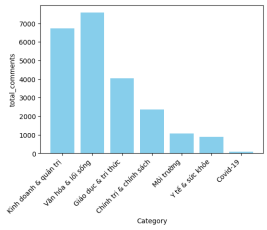
Hình 6a và 6b cho t thấy tổng số reactions cao nhất thuộc danh mục Kinh doanh & quản trị và Văn hóa & lối sống, trong khi Covid-19 có tổng tương tác thấp. Đáng chú ý, mặc dù số lượng bài viết ít, danh mục Covid-19 lại có trung bình reactions cao nhất, cho thấy các bài viết về chủ đề này nhận được sự quan tâm đặc biệt từ độc giả. Điều này thể hiện sự mất cân đối rõ rệt trong phân bổ nội dung, khi các chủ đề có tương tác cao lại không được xuất bản nhiều, dẫn đến bias trong hệ thống.

Hình 6c và 6d cho thấy tổng số comments cao nhất ở danh mục Văn hóa & lối sống và Giáo dục & tri thức, phản ánh sự sôi nổi trong thảo luận của độc giả. Trung bình comments cao nhất thuộc về Văn hóa & lối sống, cho thấy độc giả thường thảo luận nhiều hơn trong các bài viết thuộc danh mục này.

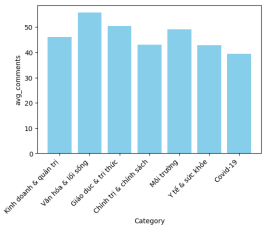


(a) Độ dài tiêu đề

(b) Độ dài description



(c) Độ dài content



(d) Độ dài user_comment

Hình 7: Phân tích tương tác và bình luận theo danh mục bài báo (sắp xếp theo độ phổ biến giảm dần).

4 Phương pháp thực nghiệm

Quy trình xây dựng hệ thống gợi ý được chia thành hai nhiệm vụ chính, mỗi nhiệm vụ là một tác vụ độc lập: Gợi ý dựa trên nội dung (Content-based Filtering) và Gợi ý dựa trên cộng tác (Collaborative Filtering). Mỗi nhiệm vụ đảm nhận một vai trò cụ thể trong việc cải thiện độ chính xác và tính cá nhân hóa của gợi ý.

4.1 Nhiệm vụ 1: Chiến lược dựa trên nội dung (Content-Based Recommendation)

4.1.1 Định nghĩa bài toán

Hệ thống gợi ý dựa trên nội dung (content-based recommendation system) được thiết kế để gợi ý các danh mục bài báo phù hợp với sở thích của người dùng. Bài toán đặt ra là tìm kiếm và xếp hạng các danh mục tiềm năng dựa trên lịch sử tương tác và phản hồi của người dùng.

Giả sử $\mathcal{U}$ là tập hợp người dùng, $\mathcal{A}$ là tập hợp bài báo, và $\mathcal{C}$ là tập hợp danh mục. Mỗi người dùng $u \in \mathcal{U}$ tương tác với một tập con bài báo $\mathcal{A}_u \subset \mathcal{A}$ và có thể để lại bình luận $\mathcal{K}_u$ trên một số bài báo. Mỗi bài báo $a \in \mathcal{A}$ được đặc trưng bởi mô tả văn bản d_a (bao gồm tiêu đề, tóm tắt hoặc nội dung) và được gắn với danh mục $c_a \in \mathcal{C}$. Tập danh mục $\mathcal{C}$ chứa tất cả các danh mục mà bài báo có thể thuộc về.

Cho trước một người dùng $u \in \mathcal{U}$ với lịch sử tương tác $\mathcal{A}_u \subseteq \mathcal{A}$ và tập bình luận $\mathcal{K}_u$, cùng một tập bài báo mục tiêu $\mathcal{A}_t \subseteq \mathcal{A}$, cần được đánh giá mức độ phù hợp với người dùng u. Mục tiêu là đề xuất danh sách xếp hạng các danh mục $\mathcal{C}_u \subseteq \mathcal{C}$ mà người dùng u có khả năng quan tâm nhất. Danh mục được sắp xếp theo thứ tự giảm dần của mức độ liên quan.

Để đạt được mục tiêu này, cần học một hàm xếp hạng $f : \mathcal{U} \times \mathcal{A}_t \to \mathbb{R}$, trong đó $f(u, a_t)$ biểu diễn mức độ liên quan giữa người dùng u và bài báo $a_t \in \mathcal{A}_t$. Điểm liên quan của mỗi danh mục $c \in \mathcal{C}$ được tính dựa trên điểm số của các bài báo thuộc danh mục đó.

Hồ sơ người dùng $\mathbf{u}$ được xây dựng dựa trên lịch sử tương tác và bình luận của người dùng. Cụ thể, hồ sơ người dùng được biểu diễn như sau:

$$\mathbf{u} = \frac{\alpha}{|\mathcal{A}_u|} \sum_{a \in \mathcal{A}_u} \mathbf{v}_a + \frac{\beta}{|\mathcal{K}_u|} \sum_{k \in \mathcal{K}_u} \mathbf{v}_k, \qquad (1)$$

trong đó $\mathbf{v}_a$ là vector embedding của bài báo $a \in \mathcal{A}_u$, $\mathbf{v}_k$ là vector embedding của bình luận $k \in \mathcal{K}_u$, và α, β là các trọng số điều chỉnh mức độ ảnh hưởng của bài báo và bình luận.

Độ tương đồng giữa hồ sơ người dùng $\mathbf{u}$ và bài báo mục tiêu $a_t \in \mathcal{A}_t$ được tính bằng cosine similarity:

$$\text{sim}(\mathbf{u}, \mathbf{v}_{a_t}) = \frac{\mathbf{u} \cdot \mathbf{v}_{a_t}}{\|\mathbf{u}\| \|\mathbf{v}_{a_t}\|}, \qquad (2)$$

trong đó $\mathbf{v}_{a_t}$ là vector embedding của bài báo mục tiêu a_t.

Điểm liên quan của danh mục $c \in \mathcal{C}$ được tính dựa trên trung bình điểm tương đồng của các bài báo thuộc danh mục đó:

$$\text{score}(c) = \frac{1}{|\mathcal{A}_t^c|} \sum_{a \in \mathcal{A}_t^c} \text{sim}(\mathbf{u}, \mathbf{v}_a), \quad (3)$$

trong đó $\mathcal{A}_t^c \subseteq \mathcal{A}_t$ là tập các bài báo thuộc danh mục c.

Cuối cùng, các danh mục $c \in \mathcal{C}$ được xếp hạng theo thứ tự giảm dần của score(c), và danh mục có điểm số cao nhất sẽ được gợi ý đầu tiên.

4.1.2 Chiến lược Sử dụng Mô hình Ngôn ngữ Tiền Huấn Luyện

Nghiên cứu này sử dụng các mô hình ngôn ngữ tiền huấn luyện (Pretrained Language Models - PLM) để biểu diễn bài báo và phản hồi người dùng dưới dạng vector embedding ngữ nghĩa. Các vector này được sử dụng để tính toán độ tương đồng giữa hồ sơ người dùng và các bài báo mục tiêu, từ đó xếp hạng và gợi ý danh mục bài báo phù hợp.

Lý do sử dụng mô hình ngôn ngữ tiền huấn luyện Mô hình ngôn ngữ tiền huấn luyện được lựa chọn do khả năng biểu diễn ngữ nghĩa sâu sắc và tính linh hoạt trong xử lý ngôn ngữ tự nhiên:

- **PhoBERT:** Đây là mô hình chuyên biệt cho tiếng Việt, được huấn luyện trên tập dữ liệu tiếng Việt lớn. PhoBERT có khả năng nắm bắt tốt các đặc trưng ngữ pháp và ngữ nghĩa đặc thù của tiếng Việt. (Nguyen and Nguyen, 2020)
- **XLM-R:** Đây là mô hình đa ngôn ngữ mạnh mẽ, hỗ trợ hơn 100 ngôn ngữ, bao gồm tiếng Việt. XLM-R được sử dụng để đánh giá khả năng tổng quát hóa so với mô hình chuyên biệt tiếng Việt. (Conneau et al., 2019)

Phương pháp được triển khai để đánh giá hiệu quả của mô hình ngôn ngữ qua 1, và cụ thể gồm các bước sau:

(1) Biểu diễn vector embedding. Mỗi bài báo $a \in \mathcal{A}$ và phản hồi $k \in \mathcal{K}_u$ được biểu diễn dưới dạng vector embedding ngữ nghĩa thông qua mô hình PLM:

$$\mathbf{v}_a = T(d_a), \quad \forall a \in \mathcal{A}, \quad \mathbf{v}_k = T(k), \quad \forall k \in \mathcal{K}_u,$$

trong đó T là PhoBERT hoặc XLM-R.

(2) Xây dựng hồ sơ người dùng. Hồ sơ người dùng $\mathbf{u}$ được tính toán bằng cách tổng hợp vector embedding của các bài báo đã đọc $\mathcal{A}_u$ và phản hồi $\mathcal{K}_u$:

$$\mathbf{u}_{\text{no-weighted}} = \frac{1}{|\mathcal{A}_u|} \sum_{a \in \mathcal{A}_u} \mathbf{v}_a + \frac{1}{|\mathcal{K}_u|} \sum_{k \in \mathcal{K}_u} \mathbf{v}_k. \quad (4)$$

$$\mathbf{u}_{\text{weighted}} = \frac{\alpha}{|\mathcal{A}_u|} \sum_{a \in \mathcal{A}_u} \mathbf{v}_a + \frac{\beta}{|\mathcal{K}_u|} \sum_{k \in \mathcal{K}_u} \mathbf{v}_k, \quad (5)$$

trong đó α, β là trọng số điều chỉnh mức độ quan trọng giữa bài báo và phản hồi.

(3) Tính toán độ tương đồng. Sử dụng cosine similarity để đo lường mức độ tương đồng giữa hồ sơ người dùng $\mathbf{u}$ và vector embedding của bài báo mục tiêu $\mathbf{v}_{a_t}$:

$$\text{sim}(\mathbf{u}, \mathbf{v}_{a_t}) = \frac{\mathbf{u} \cdot \mathbf{v}_{a_t}}{\|\mathbf{u}\| \|\mathbf{v}_{a_t}\|}. \quad (6)$$

(4) Tính điểm danh mục. Điểm liên quan của từng danh mục $c \in \mathcal{C}$ được tính bằng trung bình điểm tương đồng của các bài báo trong danh mục đó:

$$\text{score}(c) = \frac{1}{|\mathcal{A}_t^c|} \sum_{a \in \mathcal{A}_t^c} \text{sim}(\mathbf{u}, \mathbf{v}_a), \quad (7)$$

trong đó $\mathcal{A}_t^c \subseteq \mathcal{A}_t$ là tập các bài báo thuộc danh mục c.

(5) Xếp hạng danh mục. Các danh mục $c \in \mathcal{C}$ được xếp hạng theo thứ tự giảm dần của score(c).

4.1.3 Chiến lược Sử dụng Prompting Mô hình Ngôn ngữ Lớn (LLM)

Phương pháp này sử dụng các mô hình ngôn ngữ lớn (Large Language Models - LLM) như LLaMA (Touvron et al., 2023) để thực hiện dự đoán danh mục bài báo dựa trên lịch sử người dùng thông qua chiến lược prompting. Không giống các phương pháp sử dụng vector embedding truyền thống, chiến lược này tận dụng khả năng hiểu ngữ cảnh sâu và linh hoạt của LLM để tạo ra các gợi ý mang tính ngữ nghĩa cao.

Phương pháp được triển khai để đánh giá hiệu quả của mô hình ngôn ngữ qua 2, và cụ thể gồm các bước sau:

Algorithm 1 Chiến lược Sử dụng Mô hình Ngôn ngữ Tiền Huấn Luyện

1. **Đầu vào:** Tập bài báo $\mathcal{A}$ với mô tả d_a, bình luận người dùng $\mathcal{K}_u$, tập bài báo mục tiêu $\mathcal{A}_t$, trọng số α, β.

2. **Sinh vector nhúng:** Sinh vector nhúng $\mathbf{v}_a = T(d_a)$ cho mỗi bài báo $a \in \mathcal{A}$ và $\mathbf{v}_k = T(k)$ cho mỗi bình luận $k \in \mathcal{K}_u$, sử dụng mô hình tiền huấn luyện T (PhoBERT hoặc XLM-R).

3. **Tính hồ sơ người dùng:** Tính hồ sơ người dùng $\mathbf{u}$ bằng cách tổng hợp các vector nhúng từ bài báo đã đọc và các bình luận, có điều chỉnh trọng số:

$$\mathbf{u} = \frac{\alpha}{|\mathcal{A}_u|} \sum_{a \in \mathcal{A}_u} \mathbf{v}_a + \frac{\beta}{|\mathcal{K}_u|} \sum_{k \in \mathcal{K}_u},$$

   trong đó $\mathcal{A}_u$ là tập bài báo đã đọc của người dùng và $\mathcal{K}_u$ là tập bình luận của người dùng.

4. **Tính độ tương đồng:** Với mỗi bài báo mục tiêu $a_t \in \mathcal{A}_t$, tính độ tương đồng cosine giữa hồ sơ người dùng $\mathbf{u}$ và vector nhúng của bài báo:

$$\text{sim}(\mathbf{u}, \mathbf{v}_{a_t}) = \frac{\mathbf{u} \cdot \mathbf{v}_{a_t}}{\|\mathbf{u}\| \|\mathbf{v}_{a_t}\|}.$$

5. **Tổng hợp điểm danh mục:** Tính điểm liên quan của mỗi danh mục $c \in \mathcal{C}$ dựa trên trung bình độ tương đồng của các bài báo thuộc danh mục đó:

$$\text{score}(c) = \frac{1}{|\mathcal{A}_t^c|} \sum_{a \in \mathcal{A}_t^c} \text{sim}(\mathbf{u}, \mathbf{v}_a),$$

   trong đó $\mathcal{A}_t^c \subseteq \mathcal{A}_t$ là tập các bài báo thuộc danh mục c.

6. **Xếp hạng danh mục:** Sắp xếp các danh mục $c \in \mathcal{C}$ theo thứ tự giảm dần của $\text{score}(c)$.

7. **Đầu ra:** Danh sách các danh mục $c \in \mathcal{C}$ được xếp hạng.

(1) Xây dựng lịch sử người dùng. Lịch sử người dùng $\mathcal{H}_u$ bao gồm các bài báo d_a đã đọc và danh mục tương ứng c_a. Chuỗi tóm tắt được tạo thành:

$$\text{Prompt} = \text{“Người dùng đã đọc: ”} + \sum_{a \in \mathcal{H}_u} \text{“}d_a \text{ (Danh mục: } c_a\text{)”}. \quad (8)$$

(2) Bổ sung yêu cầu dự đoán danh mục. Sau khi tóm tắt lịch sử, thêm một yêu cầu để hướng dẫn mô hình thực hiện dự đoán danh mục: “Dựa trên lịch sử, hãy dự đoán danh mục tiếp theo.”

(3) Nhập prompt vào mô hình ngôn ngữ lớn. Chuỗi prompt được đưa vào mô hình ngôn ngữ lớn M.

(4) Phân tích đầu ra của mô hình. Kết quả từ M sẽ là danh sách các danh mục $\mathcal{C}_p$. Các danh mục này được trích xuất và xử lý để đảm bảo phù hợp với tập danh mục mục tiêu $\mathcal{C}_t$.

(5) Đầu ra: Danh sách các danh mục $\mathcal{C}_p$ được sắp xếp theo mức độ liên quan, với danh mục có độ ưu tiên cao nhất xếp đầu.

Algorithm 2 Chiến lược sử dụng Prompting Mô hình Ngôn ngữ Lớn

1. **Đầu vào:** Lịch sử người dùng $\mathcal{H}_u$: Bao gồm các bài báo d_a đã đọc và danh mục tương ứng c_a. Tập danh mục mục tiêu $\mathcal{C}_t$.

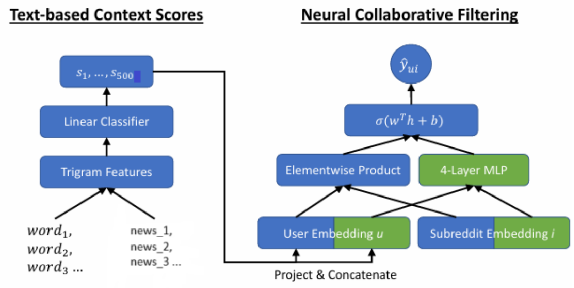
2. **Tạo prompt tóm tắt lịch sử người dùng:** Kết hợp lịch sử bài báo đã đọc và danh mục tương ứng thành một chuỗi văn bản.

3. **Nhập prompt vào mô hình ngôn ngữ lớn:** Cung cấp chuỗi prompt đã tạo vào mô hình ngôn ngữ lớn M.

4. **Phân tích đầu ra của mô hình:** Trích xuất danh mục dự đoán $\mathcal{C}_p$ từ kết quả trả về của mô hình M.

5. **Xếp hạng danh mục dự đoán:** Đánh giá và xếp hạng các danh mục $\mathcal{C}_p$.

6. **Đầu ra:** Danh sách danh mục $\mathcal{C}_p$ được xếp hạng theo mức độ liên quan.



Hình 8: Mô hình cơ bản neural collaborative filtering (NCF) (phía phải) và bộ phân loại tuyến tính dựa trên văn bản (phía trái). Mô hình NCF học các embedding đặc trưng cho từng người dùng và từng bài báo, sau đó sử dụng chúng để tính toán điểm số cho mỗi cặp người dùng - bài báo. Mô hình NCF được huấn luyện dựa trên các cặp tương tác tích cực và tiêu cực giữa người dùng và subreddit.

4.2 Nhiệm vụ 2: Chiến lược lọc cộng tác (Collaborative Filtering)

Neural Collaborative Filtering (NCF) Neural Collaborative Filtering (NCF) là một cách tiếp cận hiện đại, được xây dựng trên nền tảng Collaborative Filtering truyền thống. Chúng tôi sử dụng mô hình NCF (He et al., 2017) làm hệ thống baseline để giải quyết bài toán gợi ý bài báo. Mô hình này được thiết kế nhằm khắc phục các hạn chế của Collaborative Filtering dựa trên phân rã ma trận (Matrix Factorization), bằng cách thay thế phép nhân ma trận tuyến tính bằng một mạng nơ-ron phi tuyến.

Cơ sở lý thuyết của Collaborative Filtering Collaborative Filtering (CF) dựa trên giả định rằng sở thích của một người dùng đối với một bài báo có thể được suy ra từ sở thích của những người dùng khác đã tương tác với các bài báo tương tự.

Đầu vào của CF là một ma trận tương tác người dùng - bài báo, thường được biểu diễn dưới dạng:

$$\mathbf{R} \in \mathbb{R}^{M \times N} \tag{9}$$

trong đó:

- M: Số lượng người dùng.
- N: Số lượng bài báo.
- $r_{u,i}$: Giá trị tương tác của người dùng u với bài báo i, có thể là nhị phân (đã tương tác hoặc chưa tương tác) hoặc giá trị thực (mức độ tương tác, ví dụ: số lượt xem, lượt thích).

Mô hình phân rã ma trận (Matrix Factorization) tìm cách biểu diễn ma trận $\mathbf{R}$ dưới dạng tích của hai ma trận nhỏ hơn:

$$\mathbf{R} \approx \mathbf{P} \cdot \mathbf{Q}^\top \tag{10}$$

trong đó:

- $\mathbf{P} \in \mathbb{R}^{M \times K}$: Ma trận embedding người dùng.
- $\mathbf{Q} \in \mathbb{R}^{N \times K}$: Ma trận embedding bài báo.
- K: Số chiều của không gian embedding.

Dự đoán điểm tương tác $\hat{r}_{u,i}$ giữa người dùng u và bài báo i được thực hiện qua:

$$\hat{r}_{u,i} = \mathbf{p}_u \cdot \mathbf{q}_i^\top \tag{11}$$

4.2.1 Neural Collaborative Filtering

NCF mở rộng CF bằng cách sử dụng mạng nơ-ron phi tuyến để học mối quan hệ giữa người dùng và bài báo. Thay vì sử dụng tích vô hướng $\mathbf{p}_u \cdot \mathbf{q}_i^\top$, NCF kết hợp embedding của người dùng và bài báo thông qua một mạng nơ-ron nhiều tầng (MLP).

Embedding người dùng $\mathbf{p}_u \in \mathbb{R}^K$ và embedding bài báo $\mathbf{q}_i \in \mathbb{R}^K$ được kết hợp bằng cách nối vector:

$$\mathbf{z}_{u,i} = [\mathbf{p}_u; \mathbf{q}_i] \tag{12}$$

trong đó $[\mathbf{p}_u; \mathbf{q}_i]$ là phép nối các vector $\mathbf{p}_u$ và $\mathbf{q}_i$.

Mạng MLP sau đó được sử dụng để dự đoán điểm tương tác:

$$\hat{r}_{u,i} = \sigma(\mathbf{h}_L) \tag{13}$$

trong đó:

- $\mathbf{h}_L$: Đầu ra của lớp cuối cùng trong MLP.
- σ: Hàm kích hoạt để chuẩn hóa đầu ra trong khoảng $[0, 1]$.

Mạng MLP bao gồm nhiều lớp ẩn:

$$\begin{aligned} \mathbf{h}_1 &= f(\mathbf{W}_1 \mathbf{z}_{u,i} + \mathbf{b}_1), \\ \mathbf{h}_l &= f(\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l), \quad l = 2, \ldots, L \end{aligned} \tag{14}$$

trong đó:

- $\mathbf{W}_l$, $\mathbf{b}_l$: Trọng số và độ lệch của lớp l.
- f: Hàm kích hoạt phi tuyến (ví dụ: ReLU).

Huấn luyện NCF Mô hình được huấn luyện bằng cách tối ưu hóa hàm mất mát nhị phân cross-entropy:

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \Big[r_{u,i} \log(\hat{r}_{u,i}) + (1 - r_{u,i}) \log(1 - \hat{r}_{u,i}) \Big] \quad (15)$$

trong đó:

- $\mathcal{D}$: Tập dữ liệu huấn luyện.
- $r_{u,i}$: Nhãn tương tác thực tế (1 nếu tương tác, 0 nếu không tương tác).
- $\hat{r}_{u,i}$: Dự đoán của mô hình.

4.2.2 Tích hợp dữ liệu văn bản vào NCF

Việc kết hợp thông tin văn bản có thể giúp cải thiện hiệu suất gợi ý, đặc biệt khi dữ liệu tương tác còn hạn chế.

Phương pháp tiếp cận gồm hai giai đoạn:

- **Trích xuất embedding từ nội dung bài báo**: Sử dụng một mô hình ngôn ngữ lớn như BERT hoặc Sentence Transformers để tạo embedding nội dung của bài báo. Các embedding này cung cấp thông tin ngữ nghĩa phong phú nội dung của bài báo.
- **Kết hợp embedding văn bản vào NCF**: Các embedding nội dung của bài báo được sử dụng để bổ sung vào embedding người dùng trong mô hình NCF. Ví dụ, vector kết hợp có thể được biểu diễn dưới dạng:

$$\mathbf{q}_i^{\text{final}} = [\mathbf{q}_i^{\text{NCF}}; \mathbf{q}_i^{\text{text}}] \quad (16)$$

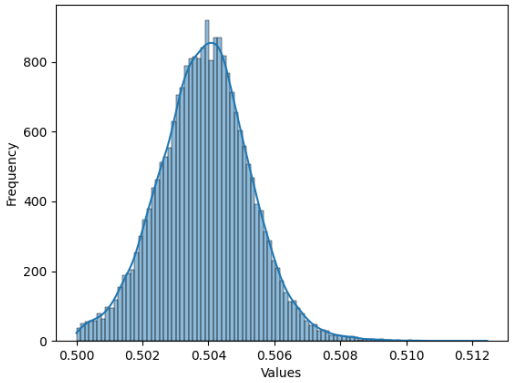
  trong đó $\mathbf{q}_i^{\text{text}}$ là embedding từ văn bản bài báo, và $\mathbf{q}_i^{\text{NCF}}$ là embedding học được từ NCF.

NCF là một mô hình mạnh mẽ cho bài toán gợi ý bài báo, nhờ khả năng kết hợp phi tuyến giữa người dùng và bài báo. Việc tích hợp embedding văn bản vào mô hình NCF không chỉ tăng cường khả năng cá nhân hóa mà còn khai thác hiệu quả nội dung bài báo, giúp nâng cao chất lượng gợi ý trong các tình huống dữ liệu hạn chế.

5 Thực nghiệm

5.1 Cài đặt thực nghiệm

Nhiệm vụ 1: Chiến lược dựa trên nội dung Trong nghiên cứu này, dữ liệu được chia theo thời gian để tạo tập train-dev-test, đảm bảo tính thực tế và phản ánh đúng ngữ cảnh của bài toán gợi ý tin tức. Dữ liệu được sắp xếp theo ngày và chia thành ba tập không trùng lặp với tỷ lệ 70%-15%-15%, trong đó tập train bao gồm 70% ngày đầu tiên, tập dev 15% ngày kế tiếp, và tập test 15% ngày cuối cùng. Điều này đảm bảo rằng dữ liệu trong các tập dev và test không bị rò rỉ từ tập train, giúp đánh giá hiệu suất mô hình trên các bài viết mới và người dùng chưa từng thấy. Các thực nghiệm được thực hiện với các cấu hình trọng số α, β khác nhau để tìm kiếm cài đặt tối ưu.



Hình 9: Phân phối mức độ tương quan

5.1.1 Nhiệm vụ 2: Chiến lược lọc cộng tác

Trong nghiên cứu này, chúng tôi sẽ tập trung vào việc dự đoán xem liệu người dùng có đọc bài báo không. Bộ dữ liệu được chia làm 3 phần gồm các tập train, dev, test với tỉ lệ lần lượt là 60% - 20% - 20%. Tiêu chí để chia tập dữ liệu là dựa trên số lượng tương tác của mỗi người dùng, trong bối cảnh của nghiên cứu này là số lượng bình luận của mỗi người ở các bài báo khác nhau với 1 lượt tương tác là 1 bình luận.

Để đánh giá mức độ tương quan của bình luận người dùng đối với bài báo, nhóm đã sử dụng mô hình XLM-RoBERTa, kết quả thu được nằm trong khoảng 0.5 đến 0.508 (Hình 8). Do đó, các mẫu dữ liệu sẽ được gán nhãn 1 (có quan tâm). Trong việc huấn luyện các mô hình phân loại nhị phân, sự có mặt của lớp negative là không thể thiếu, do đó nhóm đã sử dụng phương pháp tạo mẫu negative cho tập train. Các mẫu negative được tạo theo 2 hướng, nếu số lượng tương tác của người dùng là 1 thì tỉ lệ sẽ là 1 positive:2 negative, với các lượt tương tác lớn hơn 1, tỉ lệ sẽ là 1:1. Các mẫu negative sẽ được tạo ra bằng cách lấy ngẫu nhiên các bài báo

Bình luận
Bài viết này không liên quan đến sở thích của tôi.
Tôi không quan tâm đến chủ đề này.
Chủ đề này không thực sự hấp dẫn tôi.
Tôi không hiểu ý nghĩa của bài viết này.
Nội dung này có vẻ không có sức thuyết phục.
Tôi thấy bài viết này không rõ ràng.
Thể loại này chẳng thu hút tôi.
Tôi không thấy bài viết này có giá trị.
Bài viết không có độ mới mẻ để giữ tôi lại.

Bảng 4: Các bình luận tiêu cực

mà người dùng chưa đọc kết hợp ngẫu nhiên với các bình luận tiêu cực ở trong Bảng 4.

Nhóm sẽ chọn những người có số lượng bài báo đã đọc lớn hơn hoặc bằng 4. Sau đó, nhóm sẽ chọn thêm các bài báo ngẫu nhiên mà người dùng chưa đọc, sao cho mỗi người sẽ có tổng cộng 100 bài báo để tiến hành dự đoán. Việc chọn các bài báo ngẫu nhiên thay vì một tập bài báo cố định cho mỗi người tham gia giúp giảm thiểu sự thiên lệch có thể xảy ra trong quá trình đánh giá mô hình. Một tập bài báo cố định có thể dẫn đến các đánh giá không chính xác nếu các bài báo trong tập đó có nội dung tương tự hoặc không đủ đa dạng. Bằng cách sử dụng tập bài báo ngẫu nhiên, mô hình sẽ được kiểm tra trong một bối cảnh phong phú hơn, phản ánh nhiều chủ đề và thể loại khác nhau. Điều này giúp đánh giá khả năng tổn

5.2 Độ đo đánh giá

Để đánh giá hiệu suất của hệ thống gợi ý, chúng tôi sử dụng hai độ đo cơ bản và phổ biến là **Precision** và **Recall**. Hai độ đo này được tính toán trên danh sách các mục được gợi ý cho mỗi người dùng, nhằm đo lường khả năng gợi ý chính xác và toàn diện của mô hình.

Precision @k: Precision là tỷ lệ giữa số lượng mục gợi ý đúng (relevant items) so với tổng số mục được gợi ý trong danh sách top-k. Công thức tính:

$$\text{Precision@k} = \frac{\text{Số lượng mục đúng trong top-}k}{k}$$

Precision phản ánh mức độ chính xác của các gợi ý do mô hình cung cấp. Giá trị Precision cao cho thấy mô hình tập trung vào các mục gợi ý phù hợp nhất với sở thích người dùng.

Recall @k: Recall là tỷ lệ giữa số lượng mục gợi ý đúng trong top-k so với tổng số mục đúng trong toàn bộ tập kiểm tra. Công thức tính:

$$\text{Recall@k} = \frac{\text{Số lượng mục đúng trong top-}k}{\text{Tổng số mục đúng trong tập kiểm tra}}$$

Recall đo lường khả năng của mô hình trong việc bao phủ tất cả các mục liên quan đến người dùng. Giá trị Recall cao cho thấy mô hình không bỏ sót các mục quan trọng trong danh sách gợi ý.

Cả hai độ đo đều được tính trên tập kiểm tra với các giá trị $k = \{1, 5, 10\}$. Precision thường được sử dụng để đánh giá chất lượng gợi ý khi danh sách các mục được giới hạn, trong khi Recall phù hợp để đánh giá khả năng bao phủ của mô hình. Để đạt được hiệu quả tốt nhất, hệ thống gợi ý cần cân bằng giữa Precision và Recall.

5.3 Kết quả thực nghiệm

5.3.1 Nhiệm vụ 1: Chiến lược dựa trên nội dung

Mô hình	Tỉ lệ trọng số	P@1	Chênh lệch
PhoBERT	Mặc định	0.2778	-
PhoBERT	0.5 - 0.5	0.1075	-0.1703
XLM-Roberta	0.5 - 0.5	0.0466	-0.2312
Meta-LLaMA	0.5 - 0.5	**0.3290**	**0.0512**

Bảng 5: Kết quả Precision@1 (**P@1**) của các mô hình, chênh lệch so với Baseline (PhoBERT), và cấu hình tỉ lệ trọng số α, β. Giá trị tốt nhất được in đậm.

Kết quả thực nghiệm được trình bày trong Bảng 5 và Bảng 6 đã cung cấp thông tin cụ thể về hiệu suất của các mô hình trên bài toán gợi ý. Điểm nổi bật đầu tiên là sự vượt trội củ LLaMA với Precision@1 (**P@1**) đạt **0.3290**, vượt qua Baseline (PhoBERT, mặc định) với chênh lệch **+0.0512**. Điều này cho thấy rằng Meta-LLaMA có khả năng khai thác ngữ cảnh vượt trội, tập trung vào các gợi ý chính xác nhất. Đây là bằng chứng mạnh mẽ cho thấy tiềm năng ứng dụng của các mô hình ngôn ngữ lớn trong việc tối ưu hóa độ chính xác gợi ý trong các bài toán phức tạp.

Mô hình	Tỉ lệ trọng số	P@5	P@10	R@5	R@10
PhoBERT	Mặc định	**0.1900**	**0.1250**	**0.9501**	**1.0000**
PhoBERT	0.5 - 0.5	0.1427	0.1250	0.7136	1.0000
XLM-Roberta	0.5 - 0.5	0.1399	0.1250	0.6996	1.0000

Bảng 6: Kết quả Precision và Recall tại các giá trị top-k (P@5, P@10, R@5, R@10) và cấu hình tỉ lệ trọng số α, β của các mô hình. Giá trị cao nhất trong mỗi cột được in đậm.

Ngược lại, các mô hình PhoBERT và XLM-Roberta có sự tinh chỉnh trọng số đều cho thấy sự

suy giảm hiệu suất đáng kể, lần lượt giảm **-0.1703** và **-0.2312**, nhấn mạnh tầm quan trọng của việc tinh chỉnh trọng số hợp lý để đảm bảo hiệu suất tối ưu cho các mô hình khác nhau.

Ở các độ đo Precision và Recall tại các giá trị top-k (P@5, P@10, R@5, R@10), mô hình PhoBERT (mặc định) tiếp tục chứng minh sự ổn định vượt trội. Với P@5 đạt **0.1900** và Recall@5 đạt **0.9501**, PhoBERT duy trì hiệu suất cao hơn đáng kể so với các cấu hình trọng số. Điều này làm nổi bật vai trò của việc sử dụng mô hình đã được huấn luyện trên dữ liệu tiếng Việt trong các bài toán yêu cầu cả độ chính xác và độ bao phủ. Tuy nhiên, sự chênh lệch rõ ràng tại Recall@5 giữa PhoBERT (mặc định) và các mô hình còn lại nhấn mạnh rằng khả năng bao phủ ở top-k nhỏ hơn vẫn cần được tối ưu hóa, đặc biệt với các bài toán yêu cầu độ chính xác cao.

5.3.2 Nhiệm vụ 2: Chiến lược lọc cộng tác

Top-k	**Text-based NeuCF**		**NeuCF**	
	Precision	**Recall**	**Precision**	**Recall**
Top@1	**1.000**	0.192	0.991	0.189
Top@5	0.902	**0.836**	0.830	0.758
Top@10	0.567	**0.951**	0.545	0.918

Bảng 7: So sánh Precision và Recall giữa mô hình Text-based NeuCF và NeuCF ở các giá trị Top-k. Các giá trị tốt nhất được in đậm.

Bảng 7 trình bày kết quả so sánh giữa hai mô hình *Text-based NeuCF* và *NeuCF cơ bản* trên các chỉ số *Precision* và *Recall* tại các mức top-k (top@1, top@5, top@10). Mô hình *Text-based NeuCF* cho thấy sự cải thiện rõ rệt so với *NeuCF cơ bản* ở cả hai chỉ số.

Cụ thể, tại mức top@1, *Text-based NeuCF* đạt *Precision* tối đa (1.000 so với 0.991 của NeuCF), trong khi *Recall* của hai mô hình là tương đương (0.192 so với 0.189). Ở mức top@5, *Text-based NeuCF* thể hiện ưu thế vượt trội với *Recall* cao hơn 7.8% (0.836 so với 0.758) và *Precision* cao hơn 7.2% (0.902 so với 0.830). Tương tự, tại mức top@10, mô hình *Text-based NeuCF* tiếp tục duy trì hiệu suất vượt trội với *Recall* đạt 0.952 (cao hơn 3.6% so với NeuCF) và *Precision* đạt 0.567 (cao hơn 3.9% so với NeuCF).

Kết quả này cho thấy rằng việc tích hợp thông tin văn bản giúp mô hình Text-based NeuCF không chỉ cải thiện khả năng xếp hạng các mục liên quan mà còn đảm bảo hiệu suất tốt hơn ở cả các mức top thấp (top@1) và cao hơn (top@5, top@10). Hiệu suất này làm nổi bật tiềm năng của Text-based NeuCF trong việc cung cấp các gợi ý chính xác và toàn diện hơn so với NeuCF cơ bản.

Hiệu suất tại các mức top-k Dựa trên kết quả trong Bảng 7, mô hình *Text-based NeuCF* vượt trội hơn so với *NeuCF cơ bản* ở hầu hết các tiêu chí đánh giá. Sự cải thiện rõ ràng nhất được ghi nhận tại mức top@5, với *Recall* của Text-based NeuCF đạt 0.836, cao hơn 7.8% so với NeuCF (0.758), và *Precision* đạt 0.902, cao hơn 7.2% so với NeuCF (0.830). Kết quả này nhấn mạnh khả năng xếp hạng tốt hơn của mô hình Text-based NeuCF trong danh sách đề xuất giữa (*mid-range recommendations*).

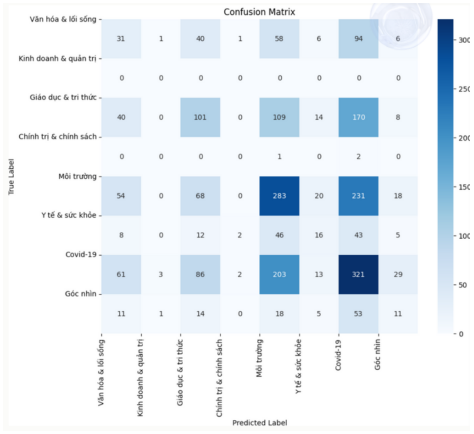
Ở mức top@10, Text-based NeuCF tiếp tục thể hiện hiệu suất cao hơn với *Recall* đạt 0.951 (cao hơn 3.6% so với NeuCF, 0.918) và *Precision* đạt 0.567 (cao hơn 3.9% so với NeuCF, 0.545). Điều này cho thấy Text-based NeuCF có khả năng bao quát tốt hơn trong việc thu thập các mục liên quan khi danh sách đề xuất mở rộng.

Xu hướng Precision và Recall Xu hướng phân phối của các chỉ số *Precision* và *Recall* phản ánh sự đánh đổi điển hình trong các hệ thống gợi ý.

- *Precision* đạt giá trị rất cao ở top@1 (gần 1.0 cho cả hai mô hình) nhưng giảm dần khi số lượng mục được đề xuất tăng lên, chỉ còn khoảng 0.54–0.57 ở top@10.
- Ngược lại, *Recall* bắt đầu từ giá trị thấp ở top@1 ($\sim$0.19) nhưng tăng mạnh ở các mức cao hơn, lên tới $\sim$0.95 tại top@10.

Sự đánh đổi này cho thấy rằng, mặc dù cả hai mô hình không luôn dự đoán chính xác mục quan trọng nhất, chúng vẫn có khả năng bao quát nhiều mục liên quan hơn khi danh sách mở rộng. Text-based NeuCF cho thấy khả năng duy trì cân bằng tốt hơn giữa *Precision* và *Recall*, đặc biệt tại mức top@5.

Hiệu suất vượt trội của mô hình Text-based NeuCF có thể được lý giải bởi khả năng khai thác các đặc tính văn bản, giúp mô hình hiểu sâu hơn mối quan hệ ngữ nghĩa giữa các bài báo. Thay vì chỉ dựa trên dữ liệu lịch sử người dùng, việc tích hợp thông tin văn bản cho phép mô hình xây dựng kết nối ngữ nghĩa giữa các mục, từ đó cung cấp các gợi ý chính xác hơn và đa dạng hơn. Điều này đặc biệt quan trọng ở các mức top@5 và top@10, nơi người dùng thường kỳ vọng vào sự cân bằng giữa tính liên quan và tính đa dạng trong danh sách gợi ý.



Hình 10: Ma trận nhầm lẫn của mô hình LLaMa-3.2-1B.

6 Phân tích & thảo luận

6.1 Nhiệm vụ 1: Chiến lược dựa trên nội dung

Dựa trên ma trận nhầm lẫn ở Hình 10, mô hình hoạt động tốt ở một số nhãn như **"Covid-19"** và **"Môi trường"**, với giá trị đúng lần lượt là 321 và 283. Điều này cho thấy khả năng nhận diện tương đối chính xác các bài viết thuộc những chủ đề này. Tuy nhiên, vẫn tồn tại sự nhầm lẫn đáng kể, đặc biệt giữa hai nhãn này. Cụ thể, có 203 bài viết từ "Covid-19" bị nhầm lẫn thành "Môi trường" và 231 bài viết từ "Môi trường" bị nhầm lẫn thành "Covid-19." Nguyên nhân có thể xuất phát từ sự tương đồng về nội dung hoặc ngữ cảnh giữa các bài viết thuộc hai chủ đề này.

Ngoài ra, mô hình gặp khó khăn trong việc nhận diện các nhãn như **"Kinh doanh & quản trị"** và **"Chính trị & chính sách"**, với giá trị đúng lần lượt là 0 và 1. Đây là lỗi nghiêm trọng, có thể xuất phát từ việc thiếu dữ liệu hoặc phân phối dữ liệu không đồng đều giữa các nhãn. Một số nhãn khác, chẳng hạn **"Văn hóa & lối sống"** và **"Giáo dục & tri thức"**, cũng có sự nhầm lẫn đáng kể (40 bài viết từ "Văn hóa & lối sống" bị nhầm thành "Giáo dục & tri thức"), cho thấy mô hình chưa phân biệt tốt giữa các chủ đề có nội dung chồng chéo.

Để cải thiện, cần tập trung cân bằng dữ liệu huấn luyện, đặc biệt tăng cường dữ liệu cho các nhãn có hiệu suất thấp như "Kinh doanh & quản trị" và "Chính trị & chính sách." Ngoài ra, việc giảm nhầm lẫn giữa các nhãn như "Covid-19" và "Môi trường" có thể được thực hiện thông qua các kỹ thuật tinh chỉnh mô hình, chẳng hạn sử dụng các mô hình ngôn ngữ lớn hoặc áp dụng các phương pháp phân loại đa nhiệm. Tổng quan, mặc dù mô hình đạt hiệu suất tốt trên một số nhãn chính, vẫn cần có các cải tiến để giảm nhầm lẫn và nâng cao độ chính xác tổng thể.

6.2 Nhiệm vụ 2: Chiến lược lọc cộng tác

Qua việc phân tích kết quả đánh giá và bảng lỗi thể loại Bảng 8, ta phát hiện một số hạn chế đáng chú ý của mô hình. Đáng kể nhất là vấn đề thiếu đa dạng trong gợi ý thể loại, thể hiện rõ qua trường hợp người dùng 2714, khi mô hình chỉ đề xuất bài viết thuộc thể loại *Văn hóa & lối sống* mặc dù người dùng còn quan tâm đến *Y tế & sức khỏe* và *Giáo dục & tri thức*. Tương tự, với người dùng 55, ở top@10 mô hình bỏ sót thể loại *Kinh doanh & quản trị* trong danh sách gợi ý.

Một vấn đề khác được phản ánh qua chỉ số *recall* thấp (0.192) ở top@1, cho thấy mô hình gặp khó khăn trong việc nắm bắt toàn bộ sở thích của người dùng khi chỉ đưa ra một gợi ý. Hiện tượng này đặc biệt rõ với những người dùng có sở thích đa dạng như trường hợp người dùng 2714. Ngoài ra, còn xuất hiện hiện tượng "filter bubble", khi mô hình có xu hướng tập trung vào một thể loại đơn lẻ mà người dùng đã tương tác nhiều (như trường hợp người dùng 49 với thể loại *Giáo dục & tri thức*), có thể hạn chế khả năng khám phá nội dung mới của người dùng.

Đáng chú ý, mô hình hoạt động tốt với người dùng có sở thích tập trung (người dùng 49) nhưng kém hiệu quả với người dùng có sở thích đa dạng (người dùng 2714, 55). Những hạn chế này gợi ý các hướng cải thiện tiềm năng như: tăng cường đa dạng hóa gợi ý, cải thiện khả năng nắm bắt sở thích đa chiều của người dùng, và phát triển cơ chế cân bằng giữa khai thác (*exploitation*) và khám phá (*exploration*) trong quá trình gợi ý.

7 Thách thức & Hướng phát triển

7.1 Đóng góp

Nghiên cứu này đã giới thiệu **ViNewRec**, một tập dữ liệu tiếng Việt chất lượng cao và đa dạng, được thiết kế đặc biệt để hỗ trợ phát triển các hệ thống đề xuất tin tức và nghiên cứu xử lý ngôn ngữ tự nhiên (NLP). Khác biệt so với các bộ dữ liệu truyền thống, ViNewRec không chỉ bao gồm nội dung văn bản mà còn tích hợp các thông tin quan trọng như bình luận, lượt đọc, và chia sẻ của người dùng. Tập dữ liệu này không chỉ mang lại cơ hội cho việc xây dựng các hệ thống gợi ý hiện đại mà còn giải quyết tình trạng thiếu hụt dữ liệu chuẩn hóa cho tiếng Việt—một ngôn ngữ phức tạp và giàu ngữ nghĩa.

Bằng cách cung cấp dữ liệu phong phú và có cấu trúc, ViNewRec tạo nền tảng cho các nghiên cứu

Bảng 8: Phân tích lỗi thể loại của mô hình Text-based NeuCF

Top-k	User ID	Ground Truth Categories	Generated Categories
1	2714	Văn hóa & lối sống	Văn hóa & lối sống
	49	Giáo dục & tri thức	Giáo dục & tri thức
	55	Chính trị & chính sách	Chính trị & chính sách
5	2714	Văn hóa & lối sống, Y tế & sức khỏe	Văn hóa & lối sống
	49	Giáo dục & tri thức	Giáo dục & tri thức
	55	Chính trị & chính sách, Văn hóa & lối sống	Chính trị & chính sách, Văn hóa & lối sống
10	2714	Giáo dục & tri thức, Văn hóa & lối sống, Y tế & sức khỏe	Văn hóa & lối sống
	49	Giáo dục & tri thức	Giáo dục & tri thức
	55	Chính trị & chính sách, Kinh doanh & quản trị, Văn hóa & lối sống	Chính trị & chính sách, Văn hóa & lối sống

nâng cao trong phân tích sentiment, phân loại tin tức, và tóm tắt văn bản tự động. Hơn nữa, nó mở ra khả năng ứng dụng vào các hệ thống gợi ý tin tức mang tính cá nhân hóa cao, cải thiện trải nghiệm người dùng và đáp ứng nhu cầu thông tin trong thời đại số.

7.2 Thách thức

Trong quá trình phát triển ViNewRec và hệ thống đề xuất tin tức, chúng tôi đã gặp phải nhiều thách thức, đặc biệt trong xử lý ngôn ngữ tiếng Việt và thu thập dữ liệu hành vi người dùng.

Thách thức trong xử lý ngôn ngữ tiếng Việt Tiếng Việt, với cấu trúc ngữ pháp phức tạp và sự linh hoạt trong ngữ nghĩa, đòi hỏi các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến. Một thách thức lớn là từ đồng âm khác nghĩa, khi ngữ cảnh đóng vai trò quan trọng trong việc giải nghĩa chính xác. Ví dụ, từ "bàn" có thể mang nghĩa "đồ vật" hoặc "thảo luận" tùy thuộc vào ngữ cảnh. Ngoài ra, quá trình phân tách từ (tokenization) trong tiếng Việt phức tạp hơn so với tiếng Anh do sự liên kết giữa các từ không được thể hiện bằng khoảng trắng. Các công cụ hiện có như VnCoreNLP hoặc Underthesea đã hỗ trợ đáng kể, nhưng hiệu quả vẫn phụ thuộc vào chất lượng dữ liệu đầu vào.

Thách thức trong thu thập và xử lý dữ liệu hành vi Mặc dù dữ liệu về lượt đọc, chia sẻ, và bình luận tương đối dễ thu thập, nhưng các thông tin chi tiết như thời gian đọc bài viết hay mức độ cuộn trang thường bị bỏ sót. Hơn nữa, hành vi người dùng đôi khi không phản ánh đúng sở thích thực sự, ví dụ như các bình luận tiêu cực hoặc tranh cãi làm tăng tương tác nhưng không đại diện cho nhu cầu thông tin. Điều này đòi hỏi các phương pháp phân tích ngữ nghĩa và mô hình hóa hành vi người dùng chính xác hơn.

7.3 Hướng phát triển

Dựa trên những kết quả đạt được, nghiên cứu này sẽ được mở rộng với các hướng phát triển sau:

Mở rộng tập dữ liệu Trong tương lai, ViNewRec sẽ được mở rộng để bao gồm thêm dữ liệu hình ảnh và video đi kèm với nội dung bài viết, hỗ trợ phát triển các hệ thống gợi ý tin tức đa phương tiện. Việc này không chỉ đáp ứng nhu cầu đa dạng của người dùng mà còn tạo điều kiện cho các nghiên cứu về học sâu (Deep Learning) trong các tác vụ như nhận diện nội dung đa phương tiện.

Tích hợp hành vi người dùng sâu hơn Chúng tôi sẽ tiếp tục nghiên cứu và tích hợp thêm các yếu tố hành vi như thời gian đọc, mức độ tương tác (reaction, bookmark) và tần suất truy cập. Điều này không chỉ cải thiện khả năng cá nhân hóa mà còn giúp hệ thống dự đoán tốt hơn các xu hướng thông tin, mang lại gợi ý phù hợp và chính xác hơn.

8 Kết luận

Nghiên cứu này đã giới thiệu **ViNewRec**, một bộ dữ liệu tiếng Việt chất lượng cao tích hợp nội dung tin

tức và thông tin tương tác người dùng, mở ra tiềm năng phát triển các hệ thống gợi ý cá nhân hóa và ứng dụng NLP. Hai nhiệm vụ chính được thực hiện là gợi ý dựa trên nội dung (*Content-based Filtering*) và gợi ý dựa trên cộng tác (*Collaborative Filtering*). Trong đó, mô hình **LLaMA** đạt *Precision@1* là 0.3290, cao hơn Baseline PhoBERT 18.4%, khẳng định khả năng khai thác ngữ cảnh hiệu quả của các mô hình ngôn ngữ lớn. Đối với gợi ý dựa trên cộng tác, **Text-based NeuCF** vượt trội với các chỉ số *Precision* và *Recall* tại *top@5* và *top@10*, nhờ tích hợp thông tin văn bản và dữ liệu tương tác.

ViNewRec không chỉ giải quyết vấn đề thiếu hụt dữ liệu tiếng Việt mà còn mở ra tiềm năng mở rộng sang dữ liệu hình ảnh, video, và áp dụng ở các ngôn ngữ khác, hướng tới xây dựng các hệ thống gợi ý đa phương tiện và đa ngôn ngữ trong tương lai.

8.1 Vấn đề đạo đức

Trong quá trình xây dựng và triển khai **ViNewRec**, nhóm nghiên cứu đã tuân thủ nghiêm ngặt các nguyên tắc đạo đức trong xử lý dữ liệu. Dữ liệu được thu thập từ các nguồn công khai, minh bạch và không bao gồm thông tin nhạy cảm hoặc dữ liệu cá nhân có thể nhận diện người dùng. Chúng tôi đảm bảo rằng việc sử dụng **ViNewRec** sẽ tuân theo các quy định bảo vệ dữ liệu cá nhân và chỉ phục vụ mục đích nghiên cứu.

Việc phát triển các hệ thống gợi ý cũng cần thận trọng để tránh các tác động tiêu cực như khuếch đại thiên lệch hoặc bóp méo thông tin. Nhóm nghiên cứu khuyến nghị các nhà phát triển nên áp dụng các biện pháp giảm thiểu thiên lệch và đảm bảo tính minh bạch trong các thuật toán gợi ý.

8.2 Lời cảm ơn

Nhóm tác giả xin gửi lời cảm ơn đến thầy Huỳnh Văn Tín đã hỗ trợ trong quá trình thực hiện nghiên cứu này. Chúng tôi cũng cảm ơn cộng đồng nghiên cứu NLP đã cung cấp các công cụ mã nguồn mở như **VnCoreNLP** và các tài nguyên hỗ trợ xử lý ngôn ngữ tiếng Việt. Ngoài ra, chúng tôi xin cảm ơn các đồng nghiệp và chuyên gia trong lĩnh vực đã đóng góp để cải thiện chất lượng nghiên cứu và tập dữ liệu **ViNewRec**.

References

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, Florence, Italy. Association for Computational Linguistics.

Daniel Cheng, Kyle Yan, Phillip Keung, and Noah A. Smith. 2022. The engage corpus: A social media dataset for text-based recommender systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1885–1889, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The adressa dataset for news recommendation. In *Proceedings of the International Conference on Web Intelligence*, WI '17, page 1042–1048, New York, NY, USA. Association for Computing Machinery.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. 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 6389–6394, Hong Kong, China. Association for Computational Linguistics.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.