KHUYẾN NGHỊ BÀI VIẾT CHO DIỄN ĐÀN TRỰC TUYẾN SỬ DỤNG HỌC SÂU

Nguyễn Đỗ Hải[*], Nguyễn Thị Yến[+], Ngô Xuân Bách[‡], Từ Minh Phương[‡]
[*]Học viện An ninh nhân dân
[+]Học viện Ngân hàng
[‡]Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Diễn đàn là một kênh thu hút được sự tương tác của một số lượng lớn người dùng hàng ngày trên Internet. Các diễn đàn ở Việt Nam hiện nay thường gợi ý cho người dùng đọc những bài viết mới được đăng; người dùng sẽ không tương tác với những bài viết này vì chúng không chứa những nội dung mà người dùng quan tâm. Các hệ thống khuyến nghị sẽ dự đoán và giới thiệu những bài viết mà người dùng có thể quan tâm và bình luận, qua đó giải quyết phần nào vấn đề này. Trong bài báo này, chúng tôi đề xuất một mô hình khuyến nghị các bài viết cho người dùng diễn đàn dựa trên lịch sử bình luận của người dùng trước đó. Phương pháp đề xuất gồm 3 phần chính bao gồm: Phần thứ nhất sử dụng mạng tích chập CNN với cơ chế Attention cho phép biểu diễn nội dung các bài viết trên diễn đàn; phần thứ hai sử dụng cơ chế Attention để biểu diễn sở thích của người dùng thông qua lịch sử bình luận và phần cuối cùng là so sánh sở thích của người dùng với nội dung bài viết để tìm ra bài viết người dùng quan tâm. Thử nghiệm trên dữ liệu thực cho thấy Phương pháp đề xuất có khả năng khuyến nghị các bài báo tốt hơn rất nhiều so với các mô hình khuyến nghị truyền thống như phương pháp khuyến nghị dựa trên nội dung hoặc phương pháp khuyến nghị dựa trên lọc cộng tác.

Từ khóa: Cơ chế Attention, CNN, mạng tích chập, Diễn đàn, Hệ khuyến nghị.

I. GIỚI THIỆU

Ngày nay người dùng Internet có thể sử dụng nhiều kênh khác nhau để thu thập thông tin và chia sẻ quan điểm như blog, wiki, các trang mạng xã hội và các diễn đàn Internet truyền thống khác. Trong các diễn đàn, người dùng vừa có thể đăng tải các nội dung mà mình muốn chia sẻ, vừa có thể bình luận vào các bài viết của những người dùng khác mà họ thấy quan tâm. Hàng ngày, một diễn đàn có thể có hàng trăm bài viết mới được đăng bởi nhiều người dùng khác nhau. Đối với một người dùng, rất khó để lựa chọn một bài viết mà thực sự quan tâm trong số hàng trăm bài viết mới mỗi ngày. Do đó, việc khuyến nghị bài viết cá nhân hóa là rất quan trọng cho các diễn đàn để giúp người dùng tìm thấy những bài viết họ quan tâm và giảm bớt tình trạng quá tải thông tin.

Đã có rất nhiều nghiên cứu giải quyết bài toán xây dựng hệ khuyến nghị sản phẩm ở nhiều dạng khác nhau như video, bài báo điện tử hay các bài viết trên diễn đàn. Trước đây, các hệ khuyến nghị thường sử dụng hai phương pháp chính là khuyến nghị dựa trên nội dung và khuyến nghị dựa trên lọc cộng tác. Hệ khuyến nghị dựa trên nội dung [1][2][3] thường biểu diễn nội dung của các đối tượng từ dạng văn bản thành dạng vector; sau đó biểu diễn sở thích người dùng dựa trên việc tổng hợp vector của các đối tượng mà người dùng đó đã tương tác và cuối cùng tính toán độ tương đồng giữa vector sở thích của người dùng với vector nội dung của các đối tượng mới để tìm ra các đối tượng mà người dùng quan tâm. Hệ khuyến nghị dựa trên lọc cộng tác [4][5][6] tập trung phân tích ma trận lịch sử tương tác giữa người dùng và sản phẩm để tìm ra sự tương đồng giữa sở thích của một số người dùng với nhau hoặc tìm ra những sản phẩm thường có xu hướng được tương tác cùng nhau. Dựa trên những sự tương đồng này, mô hình lọc cộng tác sẽ đưa ra được những sản phẩm mà người dùng quan tâm.

Hiện nay, một số mô hình học sâu [7][8][9] cũng được sử dụng để xây dựng các hệ khuyến nghị. Các mô hình học sâu được ứng dụng vào nhiều quá trình khác nhau khi xây dựng mô hình khuyến nghị như việc sử dụng các mạng CNN [7] để biểu diễn nội dung bài viết, hay sử dụng mạng RNN [10] để phát hiện ra các chuỗi sản phẩm thường được tương tác cùng nhau. Các mô hình khuyến nghị dựa trên học sâu cho hiệu quả tốt hơn rất nhiều so với các mô hình khuyến nghị truyền thống nhờ khả năng phân tích nhiều lớp thông tin ẩn trong dữ liệu.

Trong hệ khuyến nghị bài viết trên diễn đàn nói chung, thông tin của bài viết thường nằm trong tiêu đề và nội dung bài viết. Mặc dù tiêu đề và nội dung bài viết đều ở dạng ký tự nhưng hai thành phần này có cách biểu diễn thông tin khác nhau. Trong khi phần tiêu đề thường ngắn gọn và biểu diễn thông tin về bài viết qua một số từ khóa đặc trưng thì phần nội dung thường có số lượng từ lớn và thông tin dàn trải ở nhiều đoạn. Do đó với mỗi một thành phần khác nhau của bài viết cần được biểu diễn một cách khác nhau và cần có cơ chế thích hợp để kết hợp các thông tin này lại với nhau. Bên cạnh đó, sở thích của người dùng thường được biểu diễn thông qua các bài viết mà họ đã tương tác (tự đăng tải hoặc có bình luận). Trong thực tế, với mỗi bài viết người dùng lại có mức độ quan tâm khác nhau tùy thuộc vào việc nội dung bài viết có chứa nhiều thông tin mà người dùng thích hay không. Do đó để hệ khuyến nghị đạt hiệu quả tốt cần có cơ chế thích hợp để lựa chọn được những bài viết mà người dùng thực sự quan tâm trong danh sách các bài viết mà người dùng đã tương tác và tổng hợp chúng thành một vector biểu diễn sở thích của người dùng.

Trong bài báo này, chúng tôi đưa ra một hệ khuyến nghị cho người dùng trên diễn đàn sử dụng mô hình học sâu để biểu diễn thông tin bài viết và biểu diễn sở thích của người dùng. Để biểu diễn thông tin tiêu đề và nội dung bài

Tác giả liên hệ: Nguyễn Đỗ Hải,
Email: nguyendohai@gmail.com
Đến tòa soạn: 24/6/2022, chỉnh sửa: 25/8/2022, chấp nhận đăng: 5/9/2022

viết, một lớp CNN sẽ được sử dụng để trích xuất thông tin về ngữ cảnh của từng từ trong câu từ đó phát hiện những từ khóa quan trọng. Sau khi đã biểu diễn tiêu đề và nội dung bài viết thành các vector, phương pháp đề xuất sử dụng cơ chế Attention để tổng hợp những thông tin quan trọng từ cả tiêu đề và nội dung của bài viết thành một vector biểu diễn thông tin chung của cả bài viết. Cơ chế Attention sẽ có nhiệm vụ đánh trọng số các phần thông tin trong tiêu đề và nội dung bài viết để từ đó lựa chọn những phần thông tin quan trọng nhất cho vào vector biểu diễn thông tin bài viết. Để biểu diễn sở thích của người dùng, phương pháp đề xuất sử dụng cơ chế Attention ở mức bài viết để đánh trọng số mức độ quan trọng của từng bài viết mà người dùng đã tương tác và từ đó lựa chọn những thông tin quan trọng đưa vào vector biểu diễn sở thích của người dùng.

Để thử nghiệm hiệu quả của phương pháp đề xuất, chúng tôi xây dựng một bộ dữ liệu thực tế thu được từ diễn đàn VOZ. Bộ dữ liệu bao gồm thông tin về bài viết (tiêu đề và nội dung), thông tin về người dùng và lịch sử tương tác giữa người dùng và bài viết trong vòng một tháng. Phương pháp đề xuất sẽ được so sánh với các mô hình khuyến nghị truyền thống như mô hình khuyến nghị dựa trên nội dung sử dụng TF-IDF hay LDA và mô hình khuyến nghị dựa trên lọc cộng tác sử dụng thống kê. Chúng tôi sẽ đanh giá hiệu quả của các mô hình dựa trên các độ đo hiệu quả phổ biến như AUC, MRR, nDCG@5 và nDCG@10. Kết quả thử nghiệm trên bộ dữ liệu VOZ cho thấy phương pháp đề xuất trong bài báo cho hiệu quả khuyến nghị tốt hơn nhiều so với các mô hình khuyến nghị truyền thống.

Với mục tiêu như vậy, những nội dung tiếp theo của bài báo được tổ chức như sau: Phần II trình bày về một số nghiên cứu có liên quan đến bài toán xây dựng hệ khuyến nghị, Phần III trình bày về mô hình khuyến nghị bài viết cho người dùng diễn đàn mà nhóm tác giả đề xuất, Phần IV mô tả về bộ dữ liệu được sử dụng trong bài báo và quá trình cài đặt thử nghiệm mô hình, Phần V sẽ đưa ra một số đánh giá về kết quả thực nghiệm và cuối cùng là phần VI tổng hợp lại bài báo và đưa ra một số hướng nghiên cứu tiếp theo.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Các hệ khuyến nghị đưa ra gợi ý cho người dùng một danh sách các sản phẩm (sách, báo, video, bài viết,…) mà họ có thể quan tâm. Các hệ khuyến nghị không chỉ giúp người dùng dễ dàng tìm được những sản phẩm mà mình mong muốn mà nó còn giúp cho các nhà sản xuất đưa ra được những chiến lược quảng cáo phù hợp với từng khách hàng. Có hai cách tiếp cận truyền thống để xây dựng các hệ khuyến nghị đó là: phương pháp lọc dựa trên nội dung (CBF) và phương pháp lọc cộng tác (CF). Ngoài ra hiện nay, các mô hình học sâu cũng được ứng dụng vào để xây dựng các hệ khuyến nghị và cho kết quả tốt hơn so với các phương pháp truyền thống.

Các kỹ thuật CBF và CF đã được sử dụng rộng rãi để đưa ra các khuyến nghị cho nhiều loại phương tiện trực tuyến khác nhau. Các đề xuất được cá nhân hóa cho tin tức là một trong những hệ thống đầu tiên của hướng này [10][11]. Resnick và cộng sự [5] trình bày một trong những giải pháp CF đầu tiên để đề xuất netnews, được gọi là mô hình CF dựa trên bộ nhớ. Các công trình gần đây thường sử dụng phương pháp kết hợp. Hệ thống được mô tả trong [12] kết hợp thông tin nội dung của các bài báo đã truy cập với lịch sử xem của người dùng để cung cấp các đề xuất tin tức cho người dùng của Google News. Li và cộng sự [13] sử dụng cả nội dung các tin tức và lịch sử tương tác trong đó: các tin bài được phân nhóm đầu tiên và các chủ đề được trích xuất từ chúng; các chủ đề sau đó được sử dụng với kiểu truy cập của người dùng để đưa ra các đề xuất. Chu và Park [14] cũng đã thêm thông tin nhân khẩu học của người dùng để giải quyết vấn đề cold-start.

Với nguồn dữ liệu khổng lồ thu được từ tương tác của người dùng với hệ thống, các mô hình học sâu cũng đã được sử dụng để xây dựng nên các hệ khuyến nghị cho các hệ thống này. Hầu hết các phương pháp đều dự đoán sở thích của người dùng từ lịch sử tương tác của họ với các sản phẩm trong hệ thống. Hai mô hình NAML [7] và KRED [15] tìm hiểu phương pháp biểu diễn sở thích của người dùng từ các biểu diễn nội dung của tin tức mà người dùng đó đã đọc bằng cách sử dụng mạng Attention [16]. Mạng nơ-ron hồi tiếp (RNN) là một lựa chọn phổ biến để lập mô hình sự phụ thuộc tuần tự giữa các tin tức được tương tác bởi người dùng [10][17][18]. Ví dụ: EBNR [10] xây dựng một phương pháp biểu diễn sở thích của người dùng từ các tin tức được được người dùng đó tương tác sử dụng một mạng GRU. Các nghiên cứu về hệ khuyến nghị dựa trên mô hình học sâu đã khắc phục được nhược điểm của các mô hình khuyến nghị dựa trên CF và CBF trong việc biểu diễn thông tin về ngữ cảnh trong nội dung văn bản.

Các mô hình khuyến nghị dựa trên học sâu nói trên chủ yếu xây dựng cho các hệ thống báo điện tử - hệ thống có đầy đủ thông tin của một bài viết như tiêu đề, nội dung, các nhãn phân loại chủ đề (thể thao, thời sự, giáo dục…) và các hình thức tương tác khác nhau của người dùng (bình luận, thể hiện cảm xúc, đăng bài…). Tuy nhiên, các bài viết trên diễn đàn thường không có chủ đề rõ ràng, nên việc thu thập đầy đủ các thông tin bài viết như vậy rất khó khăn. Ngoài ra việc thu thập tương tác của người dùng trên diễn đàn cũng không dễ dàng, thường chúng ta chỉ có thể thu được thông tin về việc người dùng đã đăng tải bài viết nào hoặc đã bình luận vào bài viết nào. Để giải quyết vấn đề đó, phương pháp đề xuất trong bài báo này sử dụng các mô hình khuyến nghị học sâu đã có nhưng loại bỏ đi những thành phần biểu diễn thông tin không xuất hiện trong bài viết của diễn đàn. Chúng tôi cũng áp dụng mô hình xây dựng được vào một bộ dữ liệu thực tế của một diễn đàn ở Việt Nam.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Mô tả bài toán khuyến nghị cho diễn đàn

Với một diễn đàn, chúng ta có một tập người dùng $U = \{u_1, u_2, \dots, u_{|U|}\}$ trong đó $|U|$ là số lượng người dùng của diễn đàn. Bên cạnh đó, diễn đàn cung cấp cho người dùng một danh sách các bài viết $D = \{D_1, D_2, \dots, D_{|D|}\}$ trong đó $|D|$ là số lượng bài viết. Mỗi bài viết trong diễn đàn bao gồm hai thành phần chính là tiêu đề và nội dung của bài viết, chúng ta ký hiệu $W^t = [w_1^t, w_2^t, \dots, w_n^t]$ là vector tiêu đề bài viết với w_i^t là từ thứ i trong tiêu đề bài viết và n là tổng số từ trong tiêu đề bài viết; $W^b = [w_1^b, w_2^b, \dots, w_m^b]$ là vector nội dung bài viết với w_j^b là từ thứ j trong nội dung bài viết và m là tổng số từ trong nội dung bài viết.

Trong phạm vi một diễn đàn, một người dùng có thể tương tác với một bài viết theo nhiều cách khác nhau như đăng tải bài viết mới, đọc một bài viết, bình luận một bài viết hay trả lời lại một bình luận của người dùng khác. Tuy

nhiên chỉ có người quản trị của các diễn đàn mới có thể theo dõi được toàn bộ thông tin đó. Trong bài báo này, chúng tôi chỉ sử dụng những tương tác có thể thu thập được một cách công khai bao gồm người dùng đăng một bài viết mới và người dùng bình luận vào một bài viết để xây dựng bộ dữ liệu xây dựng mô hình khuyến nghị. Như vậy, với mỗi người dùng $u \in U$ chúng tôi xác định một tập bài viết $D^u = \{D_1^u, D_2^u, \dots, D_{|D^u|}^u\}$ mà người dùng đó đã tương tác, trong đó $|D^u|$ là số lượng bài viết mà người dùng đó đã tương tác và D_k^u là bài viết mà người dùng u đã đăng tải hoặc bình luận. D^u còn được gọi là lịch sử tương tác của người dùng u. Các bài viết trong tập D^u sẽ được sắp xếp theo thứ tự thời gian tăng dần mà người dùng u tương tác với bài viết; ví dụ bài viết D_1^u sẽ được người dùng bình luận trước bài viết D_2^u.

Nhiệm vụ của hệ khuyến nghị bài viết cho diễn đàn là từ lịch sử tương tác D^u của người dùng u, hệ thống phải đưa ra danh sách các bài viết $D_{rec}^u \subset D/D^u$mà người dùng có thể quan tâm (có thể bình luận). Những bài viết đưa ra khuyến nghị cho người dùng phải là những bài viết mà người dùng chưa từng tương tác.

B. Mô hình khuyến nghị bài viết cho diễn đàn trực tuyến sử dụng học sâu

Mô hình khuyến nghị bài viết cho diễn đàn trực tuyến sử dụng học sâu sẽ biểu diễn thông tin bài viết từ dạng ký tự thành một vector ở dạng số, bên cạnh đó mô hình còn phải biểu diễn được sở thích của người dùng dưới dạng một vector thông qua việc tổng hợp thông tin từ danh sách các bài viết mà người dùng đã tương tác. Để giải quyết vấn đề này, mô hình khuyến nghị này sử dụng một mạng học sâu. Phương pháp này dựa trên việc tìm hiểu cách biểu diễn nội dung văn bản trong nghiên cứu [19]; nhưng thay vì sử dụng một lớp LSTM để biểu diễn thông tin của nội dung bài viết thì hệ thống sẽ sử dụng một lớp CNN như trong phương pháp NAML [7]. Việc lựa chọn sử dụng lớp CNN thay vì lớp LSTM vì mạng CNN có khả năng bắt được các n-grams chứa những từ khóa quan trọng của câu với n là độ dài của filter, thông qua đó thì việc biểu diễn thông tin của câu chính xác hơn, thay vì phải lưu thông tin của cả câu khi dùng LSTM.

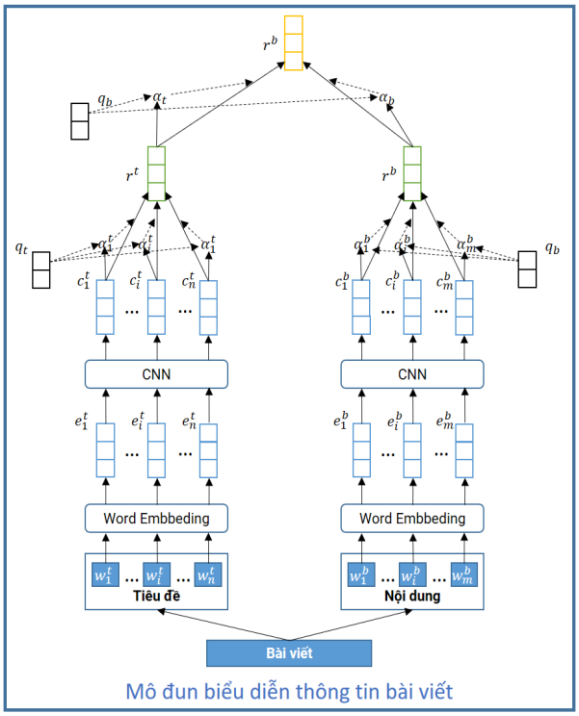
Trong phương pháp đề xuất có ba mô đun chính bao gồm: Mô đun biểu diễn thông tin bài viết có nhiệm vụ tổng hợp thông tin bài viết từ tiêu đề và nội dung của bài viết; mô đun biểu diễn sở thích của người dùng có nhiệm vụ biểu diễn sở thích của người dùng dựa trên lịch sử tương tác của người dùng đó và mô đun tính độ tương đồng có nhiệm vụ so sánh độ tương đồng giữa sở thích của người dùng với nội dung của bài viết để xác định xem người dùng có quan tâm đến bài viết hay không.

1) Mô đun biểu diễn bài viết:

Mô đun này có nhiệm vụ biểu diễn thông tin các bài viết từ hai nguồn thông tin thu được từ mỗi bài viết đó là tiêu đề bài viết và nội dung bài viết. Trong mô đun này có ba thành phần chính đó là bộ mã hóa tiêu đề biểu diễn thông tin trong tiêu đề, bộ mã hóa nội dung biểu diễn thông tin trong nội dung và lớp tổng hợp thông tin để biểu diễn thông tin của cả bài viết thông qua việc kết hợp thông tin tiêu đề và thông tin nội dung của bài viết.

- Bộ mã hóa tiêu đề

Để trích xuất thông tin từ tiêu đề bài viết, một bộ mã hóa tiêu đề 3 lớp được đề xuất. Lớp đầu tiên trong bộ mã hóa tiêu đề là lớp WordEmbedding, lớp này có nhiệm vụ chuyển tiêu đề của bài viết từ định dạng chuỗi ký tự thành dạng chuỗi các vector biểu diễn ý nghĩa của các từ tương ứng. Giả sử chúng ta có đoạn tiêu đề bài viết là $[w_1^t, w_2^t, \dots, w_n^t]$ thì sau khi chúng ta đưa qua lớp Embedding sẽ thu được một chuỗi các vector tương ứng $[e_1^t, e_2^t, \dots, e_n^t]$. Trong đó e_i^t là vector biểu diễn thông tin của từ thứ i trong câu đầu vào.



Hình 1: Mô đun biểu diễn thông tin bài viết

Lớp thứ hai trong bộ mã hóa tiêu đề là một mạng nơ ron tích chập [20], lớp này có nhiệm vụ biểu diễn thông tin ngữ cảnh của các từ xuất hiện trong tiêu đề. Trong một câu, nội dung chính của câu không thể thu được khi phân tích từng từ trong câu mà phải kết hợp các từ lân cận nhau lại với nhau thì mới có thể thu được nội dung chính xác của câu, đây chính là ý nghĩa của việc biểu diễn thông tin ngữ cảnh của câu. Lớp này sẽ nhận đầu vào là chuỗi vector biểu diễn thông tin của từng từ trong câu ở lớp thứ nhất, và đầu ra của lớp này là một chuỗi các vector biểu diễn thông tin ngữ cảnh của từng từ trong câu $[c_1^t, c_2^t, \dots, c_n^t]$. Thông tin ngữ cảnh của từng từ trong câu sẽ được tính bằng công thức sau:

$$c_i^t = ReLU(F_t \times e_{(i-K:i+K)}^t + b_t \qquad (1)$$

Trong đó:

+ $e_{(i-K:i+K)}^t$ là sự kết hợp chuỗi các vector biểu diễn thông tin của các từ lân cận với từ thứ i, bắt đầu từ vị trí $i-K$ đến vị trí $i+K$;

+ $F_t \in R^{N_f \times (2K+1)D}$ và $b_t \in R^{N_f}$ là các tham số của mạng CNN.

Lớp cuối cùng trong bộ mã hóa tiêu đề là lớp tổng hợp thông tin sử dụng cơ chế Attention. Trong một câu, không phải từ nào trong câu cũng chứa lượng thông tin như nhau, có những từ chứa nhiều thông tin quan trọng để thể hiện nội dung của câu nhưng cũng có những từ không thể hiện nội dung gì của câu. Do đó cần phải phát hiện được những từ có chứa nhiều thông tin quan trọng để có thể biểu diễn

thông tin của câu tốt hơn. Để giải quyết vấn đề này chúng tôi sử dụng một mạng Attention ở mức từ [21] để lựa chọn được những từ quan trọng trong tiêu đề bài viết. Mục tiêu của lớp Attention này là xây dựng được bộ trọng số đánh giá độ quan trọng của từng từ trong câu $[\alpha_1^t, \alpha_2^t, \dots, \alpha_n^t]$, bộ trọng số này được tính bằng công thức sau:

$$\alpha_i^t = \frac{exp\ (a_i^t)}{\sum_{j=1}^{n} exp\ (a_j^t)} \quad (2)$$

$$a_i^t = q_t^T tanh\ (V_t \times c_i^t + v_t) \quad (3)$$

Như vậy thông tin từ tiêu đề bài viết cuối cùng sẽ là sự tổng hợp của vector biểu diễn thông tin ngữ cảnh của từng từ trong tiêu đề và trọng số tương ứng của nó, cụ thể vector biểu diễn thông tin tiêu đề bài viết sẽ có dạng $r^t = \sum_{j=1}^{n}(\alpha_i^t \times c_i^t)$.

- Bộ mã hóa nội dung:

Bộ mã hóa nội dung có nhiệm vụ trích xuất thông tin từ nội dung của bài báo. Bộ mã hóa nội dung cũng bao gồm 3 lớp giống với bộ mã hóa tiêu đề. Cụ thể, lớp đầu tiên trong bộ mã hóa nội dung là lớp WordEmbedding, lớp này sẽ biến nội dung bài viết từ dạng chuỗi các từ $[w_1^b, w_2^b, \dots, w_m^b]$ thành chuỗi các vector biểu diễn thông tin của từng từ trong nội dung bài viết $[e_1^b, e_2^b, \dots, e_m^b]$. Sau khi đã biểu diễn nội dung bài viết từ dạng ký tự thành chuỗi các vector chứa thông tin của từng từ, bộ mã hóa nội dung sử dụng một mạng nơ-ron tích chập để tổng hợp thông tin về ngữ cảnh của các từ trong nội dung bài viết. Đầu ra của mạng nơ ron tích chập là chuỗi các vector biểu diễn thông tin ngữ cảnh của các từ trong nội dung bài viết $[c_1^b, c_2^b, \dots, c_m^b]$. Lớp cuối cùng trong bộ mã hóa nội dung là lớp Attention, lớp này có nhiệm vụ tổng hợp thông tin từ các từ có trong nội dung bài viết để lựa chọn ra những từ có chứa nhiều thông tin quan trọng về nội dung của cả bài viết. Đầu ra của lớp Attention là bộ trọng số đánh giá độ quan trọng của từng từ trong nội dung bài viết $[\alpha_1^b, \alpha_2^b, \dots, \alpha_m^b]$. Cuối cùng thông tin từ nội dung bài viết là sự tổng hợp của vector biểu diễn thông tin ngữ cảnh của từng từ trong nội dung bài viết và trọng số tương ứng của nó, cụ thể vector biểu diễn thông tin nội bài viết sẽ có dạng

$$r^b = \sum_{j=1}^{m}(\alpha_i^b \times c_i^b) \quad (4)$$

- Lớp tổng hợp thông tin:

Nhiệm vụ tiếp theo sau khi tổng hợp thông tin từ từng nguồn khác nhau của bài viết đó là kết hợp các thông tin khác nhau lại với nhau để cho ra biểu diễn cuối cùng của một bài viết. Vấn đề đặt ra khi tổng hợp thông tin cho cả bài viết đó là làm sao lựa chọn được những nội dung quan trọng trong thông tin thu được để biểu diễn chính xác thông tin của cả bài viết. Ví dụ có những bài viết thì phần tiêu đề sẽ chứa nhiều thông tin quan trọng hơn so với phần nội dung bài viết; ngược lại sẽ có những bài tiêu đề bài viết quá ngắn không chứa nhiều thông tin quan trọng mà phần nội dung lại chứa rất nhiều thông tin quan trọng. Để giải quyết vấn đề này, nhóm tác giả sử dụng một cơ chế Attention để kết hợp thông tin từ các nguồn khác nhau lại với nhau dựa trên sự quan sát và lựa chọn những phần thông tin quan trọng để giữ lại và bỏ qua những phần không quan trọng trong bài viết.

Với hai vector đầu vào là vector biểu diễn thông tin tiêu đề r^t và vector biểu diễn thông tin nội dung r^b, lớp Attention sẽ đi tính toán bộ trọng số đánh giá mức độ quan trọng của thông tin từ hai nguồn thông tin này lần lượt là α_t và α_b như sau:

$$\alpha_t = \frac{exp\ (a_t)}{exp\ (a_t)\ + exp\ (a_b)} \quad (5)$$

$$\alpha_b = \frac{exp\ (a_b)}{exp\ (a_t)\ + exp\ (a_b)} \quad (6)$$

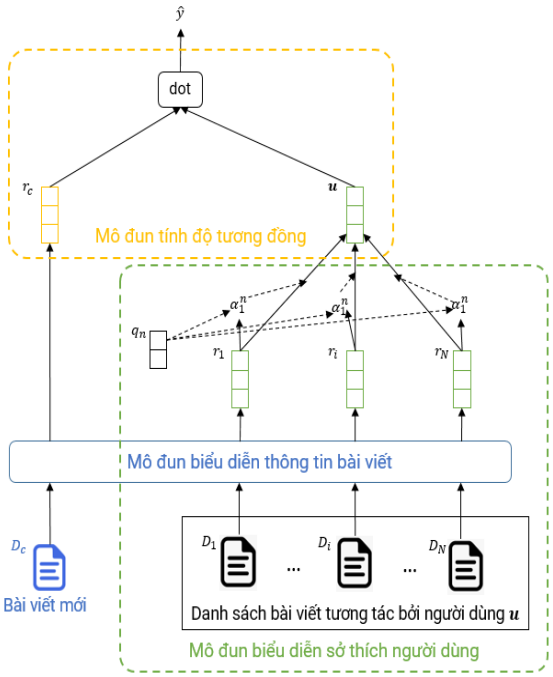
$$a_t = q_v^T tanh\ (U_v \times r^t + u_v) \quad (7)$$

$$a_b = q_v^T tanh\ (U_v \times r^b + u_v) \quad (8)$$

Và cuối cùng, vector biểu diễn thông tin của một bài viết sẽ được tính dựa trên hai vector biểu diễn thông tin tiêu đề và nội dung kết hợp với trọng số của chúng, cụ thể như sau:

$$r = \alpha_t \times r^t + \alpha_b \times r^b \quad (9)$$

Mô đun biểu diễn thông tin bài viết này sẽ được sử dụng để biểu diễn thông tin cho tất cả các bài viết đã được đăng (các bài viết mà những người dùng trong tập người dùng đã tương tác) và những bài viết mới (các bài viết mà người dùng chưa tương tác).



Hình 2: Mô hình khuyến nghị bài viết cho người dùng diễn đàn sử dụng học sâu

2) Mô đun biểu diễn sở thích người dùng:

Sau khi đã biểu diễn được nội dung các bài viết từ dạng văn bản thành các vector chứa thông tin của bài viết thì nhiệm vụ tiếp theo cần làm của một hệ thống khuyến nghị đó là biểu diễn sở thích của người dùng dựa trên những bài viết mà người dùng đã tương tác. Có nhiều cách khác nhau để giải quyết vấn đề này, trong phạm vi đề tài này sẽ đi xây dựng một mô đun biểu diễn sở thích người dùng sử dụng cơ chế Attention để lựa chọn những bài viết chứa nhiều thông tin quan trọng trong tất cả những bài viết mà

người dùng đã tương tác. Cụ thể đầu vào của mô đun biểu diễn sở thích người dùng sẽ là danh sách các bài viết mà người dùng đã tương tác $D = [D_1, D_2, \dots, D_N]$, các bài viết này sau đó sẽ được cho qua mô đun biểu diễn thông tin bài viết ở mục 2.1 để chuyển thành tập các vector thông tin bài viết $[r_1, r_2, \dots, r_N]$. Tập các vector này sẽ được đưa qua lớp Attention để xây dựng nên một bộ trọng số đánh giá mức độ quan trọng của các bài viết mà người dùng đã tương tác $[\alpha_1^n, \alpha_2^n, \dots, \alpha_N^n]$.

$$\alpha_i^n = \frac{exp\ (a_i^n)}{\sum_{j=1}^{N} exp\ (a_j^n)} \quad (10)$$

$$a_i^n = q_n^T tanh\ (W_n \times r_i + b_n) \quad (11)$$

Từ bộ trọng số này, kết hợp với vector thông tin các bài viết mà người dùng, mô đun biểu diễn sở thích người dùng sẽ tổng hợp thành vector sở thích của người dùng u dựa trên công thức: $u = \sum_{i=1}^{N} \alpha_i^n r_i$.

3) Mô đun tính độ tương đồng:

Nhiệm vụ cuối cùng của hệ thống khuyến nghị đó là tìm ra những bài viết có nội dung trùng khớp nhất với sở thích của người dùng. Để giải quyết vấn đề này có thể đơn giản chỉ cần sử dụng độ đo Cosine [22] để tính khoảng cách giữa vector biểu diễn sở thích của người dùng với vector biểu diễn thông tin của những bài viết mới và lựa chọn những bài viết có khoảng cách gần nhất với vector biểu diễn sở thích của người dùng. Một cách khác đó là sử dụng một số mô hình học máy như SVM hay một mạng học sâu để phân loại xem người dùng có bình luận vào bài viết hay không [23].

Tuy nhiên, trong bài báo này, chúng tôi sử dụng phương pháp tính tích vô hướng của hai vector để đánh giá độ tương đồng giữa sở thích của người dùng và nội dung bài viết. Lý do nhóm tác giả lựa chọn phương pháp này vì theo nghiên cứu của Chu-han Wu [7] thì phương pháp này cho đạt hiệu quả cao về cả mặt thời gian lẫn độ chính xác của mô hình xây dựng được. Cụ thể, với một người dùng có vector biểu diễn sở thích là u và một bài viết có vector biểu diễn thông tin là r_c thì giá trị độ tương đồng giữa người dùng và bài viết được tính bằng công thức $\hat{y} = u^T r_c$. Sau khi tính toán được độ tương đồng giữa người dùng và bài viết, hệ thống sẽ lựa chọn ra K bài viết có độ tương đồng cao nhất với vector biểu diễn sở thích của người dùng để đưa ra khuyến nghị cho người dùng.

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Chuẩn bị dữ liệu

Để kiểm tra hiệu quả của phương pháp đề xuất, trong phạm vi bài báo, chúng tôi xây dựng một bộ dữ liệu thực tế thu được từ diễn đàn VOZ[1]. Diễn đàn VOZ là một diễn đàn công nghệ lớn Việt Nam. Số người thường xuyên truy cập diễn đàn VOZ là khá lớn. Bên cạnh đó "Điểm báo" là một chuyên mục khá hot của diễn đàn VOZ. Các user hoạt động khá sôi nổi, comment vào nhiều bài post khác nhau. Các bài báo mạng hay sẽ được người dùng VOZ chọn lọc và đăng tải lại vào diễn đàn này. Đây là các bài báo về nhiều lĩnh vực khác nhau như: kinh tế, văn hóa, giáo dục, thể thao. Các bài báo này có chủ đề rõ ràng, nội dung có chất lượng.

Dữ liệu được thu thập trong vòng một tháng bắt đầu từ ngày 01/01/2022 đến ngày 01/02/2022 trên chuyên mục "Điểm báo" của diễn đàn VOZ. Sau khi thu thập dữ liệu từ VOZ, dữ liệu ở dạng văn bản (bao gồm tiêu đề và nội dung bài viết) sẽ được tiền xử lý để làm sạch dữ liệu; và dữ liệu về tương tác người dùng cũng được xử lý để chỉ giữ lại những người dùng có ít nhất 5 tương tác. Sau khi tiền xử lý, bộ dữ liệu thu được từ diễn đàn VOZ bao gồm 4.103 bài viết, 4.460 người dùng và số lượng tương tác của người dùng vào bài viết là 106.253. Những thông tin này được thể hiện cụ thể ở Bảng 1.

Bảng I. Bảng thống kê bộ dữ liệu voz

Số lượng bài viết	4.103
Số lượng người dùng	4.460
Số lượng tương tác	106.253
Số từ trung bình trong tiêu đề	10
Số từ trung bình trong nội dung	5747

Để phục vụ quá trình thực nghiệm, với mỗi người dùng, các bài viết trong tập lịch sử tương tác sẽ được sắp xếp theo thứ tự thời gian mà người dùng tương tác với bài viết. Trong quá trình huấn luyện mô hình, chúng tôi chia bộ dữ liệu làm 3 phần: Phần thứ nhất bao gồm 80% bài viết người dùng bình luận đầu tiên được sử dụng để biểu diễn sở thích của người dùng; phần thứ hai bao gồm 10% bài viết được người dùng bình luận tiếp theo được sử dụng để tinh chỉnh mô hình và phần cuối cùng là dữ liệu kiểm thử bao gồm 10% bài viết mới nhất mà người dùng đã bình luận.

Để tiết kiệm thời gian cũng như không gian tính toán, thay vì đi tính độ tương đồng của mỗi người dùng với tất cả những bài viết mà người đó chưa tương tác và đưa ra khuyến nghị thì chúng tôi sử dụng kỹ thuật lấy mẫu âm tính (negative sampling) [24][25]. Cụ thể, với bộ dữ liệu tinh chỉnh và bộ dữ liệu kiểm thử, ngoài việc chứa những bài viết $D^+ \in D^u$ là những bài viết mà người dùng đã bình luận được gán nhãn là dương tính, chúng tôi sẽ đi lựa chọn thêm một số bài viết $D^- \in D/D^u$ là những bài viết mà người dùng u chưa bao giờ bình luận để gán nhãn âm tính. Số lượng mẫu âm tính sẽ được lấy theo tỉ lệ K, tức là 1 mẫu D^+ sẽ lấy thêm K mẫu D^-. Sau khi lựa chọn xong các mẫu âm tính sẽ được trộn với các mẫu dương tính để tạo ra bộ dữ liệu tinh chỉnh cũng như bộ dữ liệu đánh giá hiệu quả mô hình huấn luyện.

B. Cài đặt thực nghiệm

Trong phần thực nghiệm, chúng tôi đã cài đặt một số tham số cho các mô hình học sâu như sau: Thứ nhất lớp Word Embedding sẽ lựa chọn số lượng chiều vector đầu ra sẽ là 300 và chúng tôi không sử dụng bất cứ bộ mô hình Word Embedding nào được huấn luyện trước mà sẽ sử dụng chính dữ liệu thực tế để huấn luyện. Thứ hai, với mạng nơ-ron tích chập, số filter được lựa chọn là 300 và kích thước cửa sổ trượt là 3. Thứ 3 với lớp Attention, số lượng chiều trong vector truy vấn được đặt là 200. Cuối cùng, trong quá trình lấy mẫu của kỹ thuật lấy mẫu âm tính thì tham số K được đặt bằng 9. Thứ ba, mô hình sẽ được huấn luyện nhiều lần với số lượng epoch khác nhau (từ 5-10 epoch) và kích thước batch thay đổi (16, 32, 64) để lựa chọn ra bộ tham số mà mô hình cho kết quả tốt nhất.

[1] https://voz.vn/f/diem-bao.33/

Mô hình sẽ sử dụng dữ liệu trong bộ dữ liệu tinh chỉnh để tính hàm mất mát của mô hình được huấn luyện và từ đó điều chỉnh các tham số ẩn trong các lớp mạng nơ ron để cải thiện độ chính xác của mô hình sau mỗi batch. Sau mỗi epoch mô hình sẽ sử dụng bộ dữ liệu kiểm thử để đánh giá hiệu quả của mô hình và lựa chọn ra mô hình tốt nhất.

Để đánh giá hiệu quả của mô hình khuyến nghị, bài báo sử dụng 3 độ đo phổ biến thường được sử dụng để đánh giá hiệu quả của các mô hình xếp loại sản phẩm trong các mô hình khuyến nghị đó là: Area Under Curve (AUC), Mean Reciprocal Rank (MRR) và normalized Discounted Cummulative Gain (nDCG@K) [26]. Cụ thể:

$$AUC = \frac{\left|\{(i,j)|Rank(p_i) < Rank(n_j)\}\right|}{N_p N_n} \quad (12)$$

$$MRR = \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{1}{Rank(p_i)} \quad (13)$$

$$nDCG@K = \frac{\sum_{i=1}^{K}(2^{rel_i}-1)/(1+i)}{\sum_{i=1}^{N_p}1/(1+i)} \quad (14)$$

Trong đó:

- N_p, N_n: lần lượt là số lượng mẫu âm tính và mẫu dương tính tương ứng;

- p_i: là giá trị dự đoán của mẫu dương tính thứ i;

- n_j: là giá trị dự đoán của mẫu âm tính thứ j;

- rel_i: là độ chênh lệch giữa vị trí dự đoán của mẫu dương tính so với vị trí thực tế

V. ĐÁNH GIÁ KẾT QUẢ

A. Mô tả các phương pháp sử dụng để so sánh với phương pháp đề xuất

1) Phương pháp khuyến nghị dựa trên nội dung

Để đánh giá hiệu quả của phương pháp đề xuất, chúng tôi tiến hành so sánh hiệu quả của mô hình này với ba mô hình khuyến nghị đối sánh là mô hình khuyến nghị dựa trên nội dung sử dụng TF- IDF [27], mô hình khuyến nghị dựa trên nội dung sử dụng LDA và mô hình khuyến nghị dựa trên lọc cộng tác sử dụng xác suất thống kê. Cụ thể, với phương pháp khuyến nghị dựa trên nội dung sử dụng TF-IDF, các bài viết sẽ sử dụng các vector TF-IDF để biểu diễn độ quan trọng của các từ xuất hiện trong bài viết dựa theo thống kê, vector sở thích của người dùng sẽ được xây dựng bằng cách lấy trung bình các vector TF-IDF của các bài viết mà người dùng đó đã tương tác. Phương pháp khuyến nghị dựa trên nội dung sử dụng LDA xây dựng vector biểu diễn thông tin của một bài viết dựa trên việc xây dựng một vector phân phối nội dung bài viết theo chủ đề. tương tự như phương pháp khuyến nghị dựa trên nội dung sử dụng TF-IDF, vector sở thích của người dùng sẽ được tính bằng trung bình cộng các vector bài viết mà người dùng đã tương tác. Sau khi xây dựng được vector biểu diễn thông tin và vector biểu diễn sở thích người dùng, hai phương pháp khuyến nghị này sử dụng độ đo Cosine để đánh giá độ tương đồng giữa nội dung bài viết và sở thích của người dùng và cuối cùng đưa ra khuyến nghị.

2) Phương pháp khuyến nghị dựa trên lọc cộng tác

Không giống hai phương pháp trên, phương pháp khuyến nghị dựa trên lọc cộng tác sử dụng xác suất thống kê sẽ đi xây dựng ma trận tương tác giữa từng cặp người dùng với nhau. Cụ thể với một người dùng u, hệ thống sẽ đi tính xác suất $P(u|u_i)$ là xác suất mà người dùng u sẽ bình luận vào một bài viết nếu người dùng u_i đã bình luận, với $u_i \in U/\{u\}$. Sau khi xây dựng được ma trận tương tác người dùng, với mỗi một bài viết mới p_c, phương pháp khuyến nghị dựa trên lọc cộng tác sử dụng xác suất thống kê sẽ tính xác suất $P(u, p_c)$ là xác suất người dùng u sẽ bình luận vào bài viết p_c dựa trên các xác suất $P(u|u_j)$, với u_j là tập những người dùng đã bình luận vào bài viết p_c. Cuối cùng, phương pháp này sẽ sắp xếp các bài viết dựa trên các xác suất vừa tính được. Trong phạm vi bài báo, chúng tôi sẽ tiến hành thử nghiệm 3 phương pháp kết hợp các xác suất $P(u|u_j)$ đó là lấy tổng (SUM), lấy giá trị lớn nhất (MAX) và lấy trung bình (AVE) các xác suất này để ra được xác suất $P(u, p_c)$.*MÔ HÌNH DKN [28]*

Mô hình DKN là một mô hình hệ khuyến nghị dựa trên nội dung sử dụng mạng KCNN [20]. Trong mô hình này, để biểu diễn nội dung các bài viết nhóm tác giả đã sử dụng một mạng CNN nhiều kênh để nắm bắt thông tin về ngữ cảnh của một từ trong câu. Bên cạnh đó, thay vì sử dụng vector one-hot của các từ xuất hiện trong bài viết để làm đầu vào cho mạng CNN, mô hình DKN sử dụng một lớp embedding để biểu diễn một từ thành một vector mô tả quan hệ của từ đó với các từ khác trong từ điển. Để biểu diễn sở thích của người dùng mô hình DKN sử dụng một mạng Attention để tổng hợp thông tin từ các bài viết mà người dùng đã từng tương tác.

3) Mô hình TANR [29]

Mô hình TANR cũng là một mô hình hệ khuyến nghị dựa trên nội dun. Trong mô hình này, Wu và cộng sự cũng sử dụng một lớp KCNN để nắm bắt thông tin về ngữ cảnh của từn từ trong tiêu đề bài viết. Sau đó các vector ngữ cảnh của các từ trong bài viết sẽ được đưa qua một lớp Attention để trích lọc ra những thông tin quan trọng trong nội dung bài viết. Sau đó, mô hình này cũng dùng một lớp Attention để tổng hợp thông tin về sở thích của người dùng.

B. So sánh hiệu quả giữa mô hình đề xuất với các phương pháp so sánh

Dựa trên kết quả thực nghiệm ở bảng II, có thể thấy mô hình khuyến nghị dựa trên học sâu được đề xuất trong bài báo cho hiệu quả tốt nhất trong cả 5 phương pháp. Cụ thể, phương pháp này cho độ chính xác AUC, MRR, nDCG@5 và nDCG@10 lần lượt là 85,66; 46,34; 61,1 và 66,61; kết quả này cao hơn hẳn so với hai phương pháp khuyến nghị truyền thống là phương pháp khuyến nghị dựa trên nội dung sử dụng TF-IDF hoặc LDA và phương pháp khuyến nghị dựa trên lọc cộng tác sử dụng thống kê. Có thể lý giải cho việc này là do mô hình đề xuất đã tính toán đến vấn đề ngữ cảnh của câu khi sử dụng rất hai lớp mạng nơ ron để tính toán đến vấn đề ngữ cảnh của một từ trong câu khi biểu diễn thông tin của nội dung bài viết.

So với mô hình khuyến nghị dựa trên học sâu sử dụng mô hình DKN và TANR thì mô hình đề xuất trong phạm vi bài báo cũng cho hiệu quả tốt hơn. Lý giải cho vấn đề này là do mô hình đề xuất trong phạm vi bài báo đã biểu diễn thông tin bài viết từ hai nguồn thông tin quan trọng là tiêu đề của bài viết và nội dung của bài viết, trong khi đó hai mô hình còn lại chỉ sử dụng thông tin thu được từ tiêu đề của bài viết.

Bảng II. Bảng so sánh hiệu quả của các mô hình khuyến nghị

Tên phương pháp			AUC	MRR	nDCG@5	nDCG@10
Mô hình khuyến nghị dựa trên lọc cộng tác	MAX		46.57	20.71	22.33	32.18
	AVE		43.18	20.26	22.93	32.34
	SUM		40.62	18.29	18.06	27.68
Mô hình khuyến nghị dựa trên nội dung	TF-IDF		61.02	27.14	31.80	40.58
	LDA		55.80	21.18	22.96	32.83
DKN			80.88	43.08	56.33	62.13
TANR			81.88	44.31	58.15	63.79
Mô hình đề xuất	Số lớp filter CNN	Kích thước cửa sổ CNN				
	300	2	85.38	46.07	60.68	66.21
	100	3	85.75	46.94	61.77	67.25
	200	3	85.41	46.21	60.72	66.24
	300	3	85.66	46.34	61.10	66.61
	400	3	85.28	45.55	60.06	65.57
	100	4	85.90	46.72	61.66	67.00
	300	4	85.93	47.25	62.04	67.51

Bảng III. Bảng so sánh hiệu quả của mô hình khi thay đổi bộ dữ liệu

m	C	K	Tên mô hình			AUC	MRR	nDCG@5	nDCG@10
4460	5	9	CF	MAX		46,57	20,71	22,33	32,18
				AVE		43,18	20,76	22,93	32,34
				SUM		40,62	18,29	18,06	27,68
			CBF	TF-IDF	MAX	61,02	27,14	31,8	40,58
				LDA	SUM	55,62	21,77	24,1	33,96
			Phương pháp đề xuất			**85,66**	**46,34**	**61,1**	**66,61**
1002	34	9	CF	MAX		31,64	6,3	6,93	9,37
				AVE		34,54	8,47	11,99	14,25
				SUM		28,04	5,85	6,49	8,53
			CBF	TF-IDF	MAX	60,28	13,11	21,28	25,52
				LDA	MAX	56,67	9,66	12,66	17,44
			Phương pháp đề xuất			**78,87**	**20,57**	**37,62**	**44,98**
1002	34	10	CF	MAX		31,4	5,75	6,34	8,49
				AVE		34,21	8,01	11,52	13,22
				SUM		27,82	5,45	5,56	7,52
			CBF	TF-IDF	MAX	60,3	12,43	19,43	23,89
					SUM	55,81	10,09	14,82	18,84
			Phương pháp đề xuất			**79,87**	**19,77**	**35,12**	**42,99**

Trong đó: - ***m*** *là số lượng người dùng;*

- ***C*** *là số lượng tương tác tối thiểu;*

- ***K*** *LÀ SỐ MẪU NEGATIVE ĐƯỢC CHỌN*

Ngoài ra, khi so sánh giữa hai phương pháp khuyến nghị dựa trên nội dung và phương pháp khuyến nghị dựa trên lọc cộng tác thì phương pháp khuyến nghị dựa trên nội dung cho kết quả tốt hơn khi mô hình khuyến nghị dựa trên nội dung cho kết quả tốt nhất với các độ đo AUC, MRR, nDCG@5, nDCG@10 lần lượt là 61,02; 27,13; 31,8 và 40,58 khi sử dụng vector TF-IDF. Trong khi đó, phương lọc cộng tác cho thấy kém hiệu quả hơn với ba độ đo AUC, MRR, nDCG@5, nDCG@10 lần lượt chỉ là 46,59; 20,75; 22,93 và 32,33. Lý giải cho việc này thì chúng ta có thể quan sát ở bảng \ref{tab:VozStat} và thấy rằng số lượng người dùng trên diễn đàn VOZ là khá lớn lên đến hơn 4.000 người dùng; trong khi đó số lượng tương tác thu được từ những người dùng này không cao chỉ khoảng hơn 100.000 tương tác. Với số lượng tương tác như vậy thì rất khó để tìm được sự tương đồng về mặt sở thích giữa những người dùng với nhau. Mặt khác, thông tin về nội dung các bài viết rất chi tiết và đầy đủ (trung bình mỗi bài viết có 10 từ ở tiêu đề, có hơn 5.000 từ ở nội dung bài viết); từ nguồn dữ liệu này có thể dễ dàng xây dựng được các thông tin biểu diễn nội dung bài viết cũng như là vector sở thích của người dùng.

C. So sánh hiệu quả của phương pháp đề xuất với những bộ dữ liệu kích thước khác nhau

Để đánh giá độ ổn định của phương pháp đề xuất,

chúng tôi cũng đã tiến hành một số thay đổi trong bộ dữ liệu huấn luyện mô hình. Cụ thể, thay vì đưa tất cả số lượng tương tác thu được vào huấn luyện, chúng tôi sẽ lựa chọn ra những người dùng có số lượng tương tác nhất định để đem lịch sử tương tác của họ vào huấn luyện. Chúng tôi lấy thử nghiệm với 2 trường hợp: Trường hợp thứ nhất lựa chọn những người dùng có tối thiểu 5 tương tác để huấn luyện và trường hợp thứ hai lựa chọn những người dùng có tối thiểu 34 tương tác để huấn luyện. Bên cạnh đó, chúng tôi cũng thay đổi số lượng mẫu negative được lựa chọn theo phương pháp negative sampling để xem việc lựa chọn mẫu có ảnh hưởng đến quá trình huấn luyện của mô hình không.

Dựa vào kết quả ở bảng III có thể thấy, thứ nhất, phương pháp đề xuất trong bài báo vẫn cho hiệu quả tốt nhất khi so sánh với các phương pháp khác mặc dù bộ dữ liệu đã bị thay đổi. Kết quả này bổ sung minh chứng cho việc ngữ cảnh của các từ trong câu ảnh hưởng rất lớn đến việc biểu diễn nội dung của bài biết.

Bên cạnh nhận xét trên, số liệu trong bảng III còn cho thấy việc loại bỏ đi những người dùng có số lượng tương tác ít cũng ảnh hưởng rất nhiều đến độ chính xác của phương pháp đề xuất. Có thể thấy mặc dù chỉ số AUC của mô hình sau khi thay đổi bộ dữ liệu có sự thay đổi nhẹ từ 85,66 với bộ dữ liệu có 4.460 người dùng xuống còn 78,87 và 79,87 với hai bộ dữ liệu có 1.002 người dùng. Tuy nhiên các chỉ số còn lại như MRR, nDCG@5 và nDCG@10 chứng kiến sự chênh lệch lớn giữa bộ dữ liệu 4.460 người dùng và hai bộ dữ liệu 1.002 người dùng. Nguyên nhân của vấn đề này có thể là do khi giảm số lượng người dùng thì số lượng tương tác cũng bị giảm xuống dẫn đến chất lượng mô hình huấn luyện không được tốt.

VI. KẾT LUẬN

Trong phạm vi bài báo này, chúng tôi đã tiến hành xây dựng một mô hình khuyến nghị bài viết trên diễn đàn cho người dùng bình luận sử dụng học sâu. Phương pháp đề xuất trong bài báo sử dụng một bộ mã hóa thông tin bài viết để biểu diễn các bài viết trên diễn đàn dựa trên tiêu đề và nội dung bài viết. Trong bộ mã hóa bài viết có hai bộ mã hóa thành phần là bộ mã hóa thông tin tiêu đề và bộ mã hóa thông tin nội dung để trích xuất thông tin đặc trưng của bài viết từ tiêu đề và nội dung bài viết ra. Sau khi trích xuất được thông tin từ tiêu đề và nội dung của bài viết ra, bộ mã hóa bài viết sẽ sử dụng một lớp Attention để tổng hợp các thông tin quan trọng từ hai nguồn thông tin này và kết hợp với nhau vào một vector biểu diễn thông tin của toàn bộ bài viết. Sau khi đã biểu diễn được thông tin các bài viết trên diễn đàn, mô hình khuyến nghị dựa trên học sâu này sẽ tiến hành biểu diễn sở thích của từng người dùng trong diễn đàn bằng cách sử dụng bộ mã hóa sở thích người dùng để tổng hợp thông tin của các bài viết mà người dùng đã tương tác sử dụng một lớp Attention. Cuối cùng để đánh giá độ tương đồng giữa vector biểu diễn thông tin bài viết với vector sở thích của người dùng, mô hình này sử dụng phương pháp lấy tích vô hướng của hai vector này.

Khi áp dụng mô hình này vào bộ dữ liệu của diễn đàn VOZ, kết quả cho thấy phương pháp đề xuất cho hiệu quả tốt hơn nhiều so với hai phương pháp được đem đối sánh là phương pháp khuyến nghị dựa trên nội dung sử dụng TF-IDF và LDA; và phương pháp khuyến nghị dựa trên lọc cộng tác sử dụng xác suất thống kê.

Với những kết quả đạt được như vậy, để mở rộng nghiên cứu cho nội dung bài báo, có thể có một số hướng nghiên cứu tiếp theo như: Thứ nhất, để biểu diễn bài viết một cách chi tiết hơn có thể bổ sung thông tin về chuyên mục của bài báo vào mô hình mã hóa thông tin bài viết; thứ hai, thay vì sử dụng toàn bộ các tương tác của người dùng để biểu diễn sở thích của họ thì chỉ quan tâm đến sở thích ngắn hạn của họ khi mà bộ dữ liệu thu thập được mở rộng lên phạm vi vài tháng hoặc vài năm để tiết kiệm không gian bộ nhớ và năng lực xử lý. Nhóm tác giả sẽ lên kế hoạch để triển khai những hướng nghiên cứu này trong thời gian tới.

TÀI LIỆU THAM KHẢO

[1] M. Kompan and M. Bieliková, "Content-based news recommendation," 2010, pp. 61–72.

[2] T. Luostarinen and O. Kohonen, "Using topic models in content-based news recommender systems," 2013, pp. 239–251.

[3] O. Phelan, K. McCarthy, M. Bennett, and B. Smyth, "Terms of a feather: Content-based news recommendation and discovery using twitter," 2011, pp. 448–459.

[4] F. Garcin, K. Zhou, B. Faltings, and V. Schickel, "Personalized news recommendation based on collaborative filtering," 2012, vol. 1, pp. 437–441.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: An open architecture for collaborative filtering of netnews," 1994, pp. 175–186.

[6] J. Sun, Z. Cheng, S. Zuberi, F. Pérez, and M. Volkovs, "Hgcf: Hyperbolic graph convolution networks for collaborative filtering," 2021, pp. 593–601.

[7] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, and X. Xie, "Neural news recommendation with attentive multi-view learning," ArXiv Prepr. ArXiv190705576, 2019.

[8] Q. Jia, J. Li, Q. Zhang, X. He, and J. Zhu, "RMBERT: News Recommendation via Recurrent Reasoning Memory Network over BERT," 2021, pp. 1773–1777.

[9] K. Park, J. Lee, and J. Choi, "Deep neural networks for news recommendations," 2017, pp. 2255–2258.

[10] S. Okura, Y. Tagami, S. Ono, and A. Tajima, "Embedding-based news recommendation for millions of users," 2017, pp. 1933–1942.

[11] F. Wu et al., "Mind: A large-scale dataset for news recommendation," 2020, pp. 3597–3606.

[12] J. Liu, P. Dolan, and E. R. Pedersen, "Personalized news recommendation based on click behavior," 2010, pp. 31–40.

[13] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, "Scene: a scalable two-stage personalized news recommendation system," 2011, pp. 125–134.

[14] W. Chu and S.-T. Park, "Personalized recommendation on dynamic content using predictive bilinear models," 2009, pp. 691–700.

[15] D. Liu et al., "KRED: Knowledge-aware document representation for news recommendations," 2020, pp. 200–209.

[16] Q. Zhang, Q. Jia, C. Wang, J. Li, Z. Wang, and X. He, "Amm: Attentive multi-field matching for news recommendation," 2021, pp. 1588–1592.

[17] V. Kumar, D. Khattar, S. Gupta, M. Gupta, and V. Varma, "Deep Neural Architecture for News Recommendation.," 2017.

[18] S. Raza and C. Ding, "Deep Dynamic Neural Network to trade-off between Accuracy and Diversity in a News Recommender System," ArXiv Prepr. ArXiv210308458, 2021.

[19] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," 2016, pp. 1480–1489.

[20] Y. Kim, "Convolutional neural networks for sentence classification. CoRR abs/1408.5882," ArXiv Prepr. ArXiv14085882, 2014.

[21] C. Wu, F. Wu, J. Liu, S. He, Y. Huang, and X. Xie, "Neural demographic prediction using search query," 2019, pp. 654–662.

[22] F. Goossen, W. IJntema, F. Frasincar, F. Hogenboom, and U. Kaymak, "News personalization using the CF-IDF semantic recommender," 2011, pp. 1–12.

[23] A. Gershman, T. Wolfe, E. Fink, and J. G. Carbonell, "News personalization using support vector machines," 2011.

[24] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck,

"Learning deep structured semantic models for web search using clickthrough data," 2013, pp. 2333–2338.

[25] S. Zhai, K. Chang, R. Zhang, and Z. M. Zhang, "Deepintent: Learning attentions for online advertising with recurrent neural networks," 2016, pp. 1295–1304.

[26] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized News Recommendation: Methods and Challenges," ACM Trans. Inf. Syst. TOIS, 2022.

[27] N. Do Hai, N. X. Bach, T. Q. An, and T. M. Phuong, "What should I comment: Recommending posts for commenting," 2013, pp. 117–122.

[28] H. Wang, F. Zhang, X. Xie, and M. Guo, "DKN: Deep knowledge-aware network for news recommendation," 2018, pp. 1835–1844.

[29] C. Wu, F. Wu, M. An, Y. Huang, and X. Xie, "Neural news recommendation with topic-aware news representation," 2019, pp. 1154–1159.

POST RECOMMENDATION FOR FORUM USING DEEP LEARNING

Abstract: Forum is an active platform which has an enormous amount of user interactive everyday. In Vietnam, forum's users are suggested the newest posts which are not read or commented because these posts do not contain information attracting users. There are many recommendation systems proposed to suggest posts which users may be interested in based on interaction history of users. In this paper, we propose a post recommendation for forum's users using deep models. Our proposed model has three sub-modules: The first module is an content encoder which combines a CNN layer and an attention layer to present context information and select important words in title and content of posts. In the second module, we use an attention layer to present user preferences. The third module is used to calculate similarity between user preference vector and content vector of candidate post. Experimental results show that our proposed model performs better than tradition recommendation models such as Content-based filtering recommendation system and Collaborative filtering recommendation.

Keywords: *Attention, CNN, Convolution neural network, Forum, Recommendation System*

Nguyễn Đỗ Hải, Nhận học vị Thạc sỹ năm 2016. Hiện công tác tại Học viện An ninh nhân dân. Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, học máy, hệ khuyến nghị.

Email: nguyendohai@gmail.com

Nguyễn Thị Yến, Nhận học vị Thạc sỹ năm 2017. Hiện là giảng viên tại Học Viện Ngân hàng. Lĩnh vực nghiên cứu chính: học máy, học sâu ứng dụng trong lọc thông tin..

Email: yenptit1511@gmail.com

Ngô Xuân Bách, Nhận học vị Tiến sỹ năm 2014, học hàm Phó giáo sư năm 2020. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Xử lý ngôn ngữ tự nhiên, học máy, hệ khuyến nghị

Email: bachnx@ptit.edu.vn

Từ Minh Phương, Nhận học vị Tiến sỹ năm 1995, học hàm Giáo sư năm 2019. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Học máy, xây dựng hệ khuyến nghị, tin sinh.

Email: phuongtm@ptit.edu.vn