PHÂN CỤM TIN TỨC BẰNG HỌC MÁY

Dương Trần Đức
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Trong thời kỳ bùng nổ thông tin trực tuyến hiện nay, hàng chục nghìn bài báo, tin tức được đăng tải mỗi ngày. Ngoài các tờ báo điện tử, các nguồn tin khác như mạng xã hội cũng cung cấp các nguồn tin tức trực tuyến nhanh chóng và đa dạng, đáp ứng nhu cầu thông tin của người dùng. Tuy nhiên, các bài báo, tin tức được tạo ra với số lượng lớn cũng mang lại vấn đề quá tải thông tin, trong đó có nhiều tin tức trùng lặp hoặc được đưa về cùng một vụ việc. Để có thể nhanh chóng nắm bắt được các thông tin về các tin tức nổi bật, việc tổ chức, phân cụm lại các tin tức theo vụ việc từ các nguồn thông tin khác nhau là một thao tác quan trọng đem lại trải nghiệm tốt hơn cho người đọc. Bài báo này trình bày một phương pháp phân cụm tin tức sử dụng thuật toán phân cụm DBSCAN dựa trên các nhúng từ trích xuất từ mô hình PhoBERT. Các thực nghiệm được thực hiện trên tập dữ liệu hơn 1.000 bản tin thu thập từ các báo điện tử của Việt Nam qua hệ thống thu thập tin tức tự động đã được gán nhãn cụm. Kết quả thực nghiệm được đánh giá theo các độ đo precision, recall, F-measure với độ chính xác lần lượt là 92.3%, 78.5%, và 85.1%.

Từ khóa: phân cụm tin tức, học máy, xử lý ngôn ngữ tự nhiên.

I. MỞ ĐẦU

Tin tức trực tuyến là một kênh thông tin phổ biến và dễ sử dụng nhất hiện nay trong lĩnh vực truyền thông. Các kênh tin tức trực tuyến như báo điện tử, mạng xã hội, trang cá nhân v.v. cung cấp các thông tin đa dạng, nhanh chóng cho người đọc. Tuy nhiên, mặt trái của nó là sự tràn ngập thông tin. Điều này có thể dẫn đến việc người dùng bị mất nhiều thời gian khi tiếp cận lần lượt các nguồn tin và khó có khả năng kết hợp thông tin từ quá nhiều nguồn tin khác nhau.

Để hỗ trợ và cung cấp các trải nghiệm tốt cho người dùng trong vấn đề tiếp cận thông tin, việc phân cụm các tin tức theo các vụ việc giúp cho người dùng có thể tiếp cận thông tin một cách có hệ thống và tiết kiệm thời gian trong việc nắm bắt, tổng hợp, kết nối các thông tin từ nhiều nguồn khác nhau [1]. Ngoài ra, việc phân cụm tin tức cũng giúp phát hiện ra các xu hướng truyền thông, giúp người dùng nắm bắt được các vụ việc đang là vấn đề quan tâm của xã hội [2, 3].

Bài báo này trình bày phương pháp phân cụm tin tức từ các báo điện tử tiếng Việt sử dụng thuật toán phân cụm DBSCAN [4]. DBSCAN là một thuật toán phân cụm không yêu cầu cấu hình trước số lượng cụm, do vậy thuật toán này phù hợp với các bài toán như phân cụm bản tin (vốn không biết trước có bao nhiêu vụ việc trong tập bản tin cần phân cụm). Để thực hiện phân cụm văn bản nói chung và phân cụm bản tin nói riêng, cần chuyển đổi các văn bản hay bản tin này thành các nhúng từ để làm đầu vào cho thuật toán phân cụm. Đã có nhiều kỹ thuật tạo nhúng từ cho văn bản hoặc bản tin, tuy nhiên các mô hình dựa trên BERT [5] được chứng minh tính vượt trội so với các kỹ thuật được dùng trước đây như Word2Vec [6], GloVe [7] trong việc tạo các nhúng từ giàu ngữ cảnh. Trong nghiên cứu này, chúng tôi sử dụng mô hình PhoBERT [8], một mô hình dựa trên BERT được huấn luyện đặc thù cho tiếng Việt, để tạo các nhúng từ. Tập dữ liệu được sử dụng trong nghiên cứu được thu thập từ các báo điện tử phổ biến ở Việt Nam như VnExpress, Vietnamnet, Dân trí v.v. Tổng cộng 1.057 bản tin trong đa dạng các lĩnh vực như Kinh tế, Xây dựng, Xã hội, Giáo dục, Quản lý đô thị (mỗi chủ đề hơn 200 bản tin), được chia thành 70 cụm, được sử dụng để huấn luyện và kiểm thử mô hình phân cụm.

Bài báo có cấu trúc như sau. Phần II trình bày về các nghiên cứu liên quan trong lĩnh vực phân cụm văn bản và phân cụm tin tức. Phần III mô tả phương pháp. Phần IV trình bày về các kết quả và thảo luận. Cuối cùng, các kết luận sẽ được trình bày trong phần V của bài báo.

II. TỔNG QUAN

Phân cụm bản tin là một vấn đề được nghiên cứu từ nhiều năm trước. Các nghiên cứu phân cụm bản tin chủ yếu hướng tới các mục tiêu trợ giúp người đọc trong việc nắm bắt thông tin hoặc phát hiện ra các sự kiện hoặc vụ việc được báo chí đưa tin [1, 2, 9, 10, 11, 12]. Mục đích của việc phân cụm là chọn ra các bản tin chứa thông tin về một vụ việc trong một tập các bản tin lớn từ các phương tiện truyền thông. Mỗi nhóm (cụm) bản tìn chứa các tin tức về cùng một vụ việc hoặc sự kiện. Liu et al. [2] thực hiện nghiên cứu phát hiện các sự kiện tin tức, trong đó các tác giả đề xuất phương pháp sử dụng tích chập đồ thị và phân cụm trên tập dữ liệu Ebola cho kết quả có độ chính xác F-measure 82%. Tarekegn et al. [9] đề xuất phương pháp sử dụng mô hình ngôn ngữ lớn để cải tiến phân cụm bản tin, thông qua tác vụ trích xuất từ khoa và sinh nhúng từ. Các tác giả cũng sử dụng thuật toán DBSCAN để phân cụm các bản tin từ tập dữ liệu GDELT với độ chính xác theo độ đo Silhouette là 0.32.

Một số phương pháp về khai phá văn bản và xử lý ngôn ngữ tự nhiên đã được nghiên cứu trong phân cụm bản tin và áp dụng trên một số loại kênh truyền thông như báo điện tử [9, 11] hay mạng xã hội [2, 10, 12]. Carta et al. [11] đề xuất phương pháp phân cụm theo phân cấp cho các tin tức tài chính với kết quả theo độ đo Silhouette là 0.45. Hettiarachchi et al. [10] đề xuất phương pháp có tên gọi Embed2Detect để phát hiện các sự kiện trên mạng xã hội nhờ phương pháp kết hợp giữa các nhúng từ và phân cụm kết tụ theo phân cấp. Các tác giả đã tiến hành các thực nghiệm trên các kỹ thuật nhứng từ khác nhau và cho thấy BERT có kết quả tốt nhất và đạt độ chính xác cao hơn so

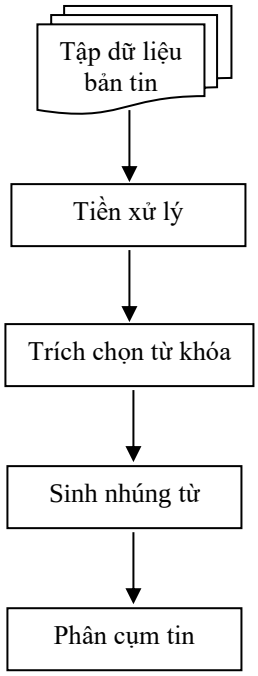
Tác giả liên hệ: Dương Trần Đức,
Email: ducdt@ptit.edu.vn
Đến tòa soạn: 8/2024, chỉnh sửa: 9/2024, chấp nhận đăng: 10/2024.

với các nghiên cứu trước. Do tính chất động của tin tức, hầu hết các phương pháp đã được nghiên cứu là phân cụm không giám sát với số lượng cụm không cố định [2, 9, 10, 11, 12].

Trong những năm gần đây, các mô hình ngôn ngữ được phát triển dựa trên Transformer như BERT, GPT đã chứng minh được sự nổi trội trong các tác vụ xử lý ngôn ngữ quan trọng như phân loại văn bản, sinh văn bản, hệ thống hỏi đáp v.v. Tuy nhiên, ứng dụng của các mô hình này với tác vụ phân cụm là tương đối hạn chế. Trong bài báo này, chúng tôi nghiên cứu sử dụng mô hình dựa trên PhoBERT cho một số bước trong quá trình phân cụm như các bước trích chọn từ khóa hay tạo nhúng từ cho các bản tin cần phân cụm.

III. PHƯƠNG PHÁP

Phần này sẽ trình bày chi tiết về phương pháp thực hiện, bao gồm các bước thực hiện trong quá trình thu thập và phân cụm tin tức, như được minh họa trong hình số 1. Quá trình bắt đầu với hoạt động thu thập dữ liệu từ các báo điện tử tiếng Việt và sau đó là quá trình tiền xử lý với các tác vụ làm sạch và chuẩn hóa dữ liệu. Bước tiếp theo các bản tin được tiến hành trích chọn từ khóa và sinh nhúng từ sử dụng PhoBERT. Cuối cùng, thuật toán phân cụm DBSCAN [4] được sử dụng để phân chia tập bản tin ra thành các cụm thể hiện các vụ việc khác nhau được đưa tin.



Hình 1. Quá trình thực hiện phân cụm tin tức

A. Thu thập và tiền xử lý dữ liệu

Đối tượng nghiên cứu của bài báo là các tin tức trực tuyến, do vậy nguồn thu thập thông tin chính là các báo điện tử (các tin tức từ mạng xã hội chưa được sử dụng do tính chính thống của các nguồn tin này chưa cao). Tổng nguồn dữ liệu được thu thập bao gồm hơn 20 báo điện tử phổ biến như VNExpress, Vietnamnet, Dân trí, Tiền phong, Thanh niên v.v. thông qua công cụ thu thập thông tin tự động do nhóm nghiên cứu tự xây dựng.

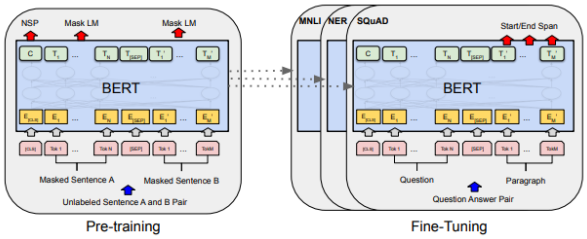
Tiền xử lý là một bước quan trọng trong quá trình xử lý văn bản, đặc biệt là với các loại văn bản tự động thu thập từ mạng Internet, vốn có thể chứa nhiều các ký tự và định dạng không mong muốn. Một số bước tiền xử lý ban đầu có thể được thực hiện như chuẩn hóa câu, lọc bỏ các ký tự lạ, từ viết tắt, dấu câu, các liên kết (links). Bước cuối cùng trong quá trình tiền xử lý là hoạt động phân tách từ (word segmentation) nhằm tạo đầu vào cho các bước xử lý tiếp theo.

B. Trích chọn từ khóa và sinh nhúng từ

Trích chọn từ khóa là một tác vụ quan trọng trong xử lý ngôn ngữ tự nhiên. Các từ khóa được trích chọn từ văn bản không những cung cấp các thông tin chính và quan trọng của tài liệu mà còn là tiền đề cho nhiều tác vụ tiếp theo như tóm tắt hay phân cụm. Đối với việc phân cụm bản tin, việc trích chọn từ khóa trước khi phân loại là một thao tác quan trọng, do các bản tin thường chứa những thông tin phụ và sử dụng tất cả các thông tin này để phân cụm sẽ làm hạn chế kết quả của thuật toán phân cụm. Các kỹ thuật dùng để trích chọn từ khóa và sinh nhúng từ đều được thực hiện sử dụng các mô hình dựa trên BERT. Các phần tiếp theo sẽ trình bày về BERT và các mô hình liên quan.

1) BERT

BERT (Bidirectional Encoder Representations from Transformer) là một mô hình biểu diễn từ được phát triển sử dụng kỹ thuật Transformer bằng cách tạo các lớp mã hóa (transformer encoder) và xếp chồng chúng với nhau để tạo thành một kiến trúc mới [5]. Tương tự transformer, BERT có thể được học chuyển giao (transfer learning) và có thể được huấn luyện với các dữ liệu không cần gán nhãn. Hai kiểu huấn luyện BERT có thể được thực hiện đồng thời, đó là mặt nạ mô hình ngôn ngữ (Mask Language Model) và dự đoán câu kế tiếp (Next Sentence Prediction). BERT có thể được huấn luyện trước bằng một lượng lớn dữ liệu văn bản không gán nhãn để tạo ra một mô hình có tri thức tổng quát về các mối quan hệ giữa các từ và các câu. Sau đó, mô hình có thể được tinh chỉnh thêm (fine-tuned) bằng cách cho học chuyển giao trên các tập dữ liệu đặc thù, thường là các tập dữ liệu nhỏ hơn và có gán nhãn cho các tác vụ cụ thể. Quá huấn luyện trước và tinh chỉnh mô hình của BERT được mô tả trong hình số 2.



Hình 2. Quá trình huấn luyện và tinh chỉnh BERT [5]

Như đã nói ở trên, BERT sử dụng các transformers, một kiến trúc có khả năng học các mối quan hệ giữa các từ sử dụng một cơ chế dựa trên sự tập trung (attention). Transformer bao gồm một bộ mã hóa (encoder) có nhiệm vụ đọc các văn bản đầu vào. Nó cũng có một bộ giải mã (decoder) có nhiệm vụ dự đoán dựa trên tác vụ cần thực hiện. Khác với các mô hình theo cấu trúc một chiều, vốn đọc đầu vào theo thứ tự tuần tự, transformer có khả năng đọc và xử lý tất cả các từ đầu vào cùng lúc và làm cho nó trở thành một mô hình có thể học ngữ cảnh của các tất cả các từ xung quanh theo cả hai chiều. Với kiến trúc đặc biệt đó và nhờ sử dụng một khối lượng khổng lồ dữ liệu huấn luyện, BERT đã cho kết quả tốt nhất trên 11 tác vụ phổ biến trong xử lý ngôn ngữ tự nhiên [5].

BERT có nhiều phiên bản được huấn luyện trước (pre-trained) cho các trường hợp sử dụng khác nhau. Hai phiên bản được sử dụng phổ biến nhất là BERT-base và BERT-large.

- BERT-base: Gồm 12 lớp mã hóa + 768 nút ẩn (hidden units) + 12 nút tập trung (attention heads). Tổng cộng 110 triệu tham số.
- BERT-large: Gồm 24 lớp mã hóa + 1024 nút ẩn (hidden units) + 12 nút tập trung (attention heads). Tổng cộng 340 triệu tham số.

2) PhoBERT

BERT là một mô hình đa ngôn ngữ, đã được huấn luyện và sử dụng cho nhiều ngôn ngữ khác nhau. Tuy nhiên, hầu hết các ngôn ngữ ngoài tiếng Anh đều được các nhóm nghiên cứu phát triển mô hình đặc thù cho ngôn ngữ đó dựa trên mô hình BERT ban đầu.

RoBERTa là một tiếp cận kế thừa kiến trúc và thuật toán của mô hình BERT nhưng mạnh và tối ưu hơn. Dự án này của Facebook hỗ trợ việc huấn luyện lại các mô hình BERT trên những bộ dữ liệu mới cho các ngôn ngữ khác ngoài một số ngôn ngữ phổ biến. Hiện đã có rất nhiều các mô hình huấn luyện trước cho những ngôn ngữ khác nhau được huấn luyện trên RoBERTa, trong đó có tiếng Việt với mô hình PhoBERT.

PhoBERT [8] là một mô hình đơn ngôn ngữ cho tiếng Việt. PhoBERT dựa trên kiến trúc RoBERTa [13] và cho thấy hiệu suất tốt hơn hẳn so với các phương pháp dựa trên mô hình BERT đa ngôn ngữ khi làm việc với văn bản tiếng Việt. Trong nghiên cứu này, chúng tôi sử dụng PhoBERT để sinh các nhúng từ cho các bản tin và sử dụng làm đầu vào cho thuật toán phân cụm.

3) KeyBERT

Trong nghiên cứu này, chúng tôi sử dụng kỹ thuật KeyBERT [14] để thực hiện trích chọn từ khóa của bản tin. KeyBERT là một công cụ trích xuất từ khóa dựa trên BERT. Kỹ thuật này sử dụng các nhúng từ được tạo từ BERT và độ tương tự cosine cơ bản để xác định các cụm từ con trong một tài liệu nào đó có độ tương tự cao nhất với chính tài liệu đó. Cách thức hoạt động của KeyBERT như sau:

- Trích xuất các nhúng từ của văn bản bằng BERT để có được biểu diễn cấp tài liệu.
- Sử dụng các nhúng từ để trích xuất các từ/cụm từ N-gram.
- Sử dụng độ tương tự cosine để xác định các từ/cụm từ nào có độ tương tự cao nhất với tài liệu. Các cụm từ tương ứng nhất được chọn là những cụm từ mô tả tốt nhất toàn bộ tài liệu.

Do phiên bản gốc của KeyBERT sử dụng mô hình BERT huấn luyện trước cho tiếng Anh, chúng tôi đã thay đổi mô hình trong KeyBERT phiên bản gốc về sử dụng mô hình PhoBERT để có thể sinh các nhúng từ tiếng Việt chất lượng hơn, từ đó có thể trích xuất các từ khóa trong văn bản tiếng Việt tốt hơn.

C. Thuật toán phân cụm

Để thực hiện nhóm các bản tin lại theo các sự kiện hoặc vụ việc. Các thuật toán phân cụm sẽ được áp dụng trên tập bản tin đã được mã hóa thành các véc tơ nhúng từ. Các thuật toán phân cụm phổ biến hiện nay như K-Means, DBSCAN sử dụng các cách thức khác nhau để tiến hành so sánh và phát hiện ra các thành phần có đặc điểm giống nhau và phân chia chúng vào các cụm. K-Means được biết là thuật toán phân cụm phổ biến nhất hiện nay, hoạt động dựa trên kỹ thuật phân cụm dựa trên phân vùng (partition-based clustering) [15]. DBSCAN là một thuật toán phổ biến khác thuộc loại phân cấp (hierachical clustering), hoạt động dựa trên kỹ thuật tích tụ [16]. Kỹ thuật này ban đầu giải thiết mỗi điểm dữ liệu là một cụm và các cụm dần dần được trộn vào nhau dựa trên sự tương đồng (được xác định dựa trên số đo khoảng cách giữa các điểm) cho đến khi đạt được số cụm mong muốn hoặc các cụm có khoảng cách xa ở một ngưỡng nhất định [17]. Mặc dù K-Means là một thuật toán phân cụm đơn giản, hiệu quả nhưng nó phụ thuộc vào cấu hình ban đầu như số lượng cụm phải được định trước. Trong khi đó, DBSCAN không yêu cầu cấu hình trước số lượng cụm, và đặc điểm này phù hợp với tác vụ phân cụm bản tin do chúng ta không biết trước có bao nhiêu sự kiện hay vụ việc được đưa trong số bản tin cần phân cụm. Với các phân tích trên, DBSCAN được lựa chọn làm thuật toán sử dụng trong nghiên cứu này.

D. Đánh giá kết quả phân cụm

Đánh giá kết quả phân cụm là một vấn đề khá phức tạp do tính chất không giám sát của thuật toán phân cụm. Để có thể đánh giá độ chính xác của phương pháp phân cụm một cách tốt nhất, chúng tôi đã tiến hành gán nhãn cụm cho tập bản thi thu thập được. Từ các bản thu thập được, các cụm được chọn bằng phương pháp thủ công. Các bản tin nói về cùng một vụ việc sẽ được gom thành cụm. Các cụm có số lượng tin lớn hơn 5 được chọn vào tập dữ liệu thực nghiệm.

Để đánh giá độ chính xác phân cụm, ba độ đo được sử dụng là precision, recall, và F-measure và sử dụng cách tính tương tự như đề xuất trong [2]. Cụ thể, giả sử có l cụm tin tức được dự đoán là $\mathcal{K}=\{K_1, K_2, \ldots, K_l\}$. Tiếp theo cần xem xét 5 bài viết hàng đầu trong cụm tin tức K_i (vì vậy số lượng bài viết trong mỗi cụm cần phải lớn hơn 5), trong đó mỗi K_i là một tập hợp các tin tức được xếp hạng theo mức độ liên quan với vụ việc chính, và thứ hạng càng cao thì mức độ liên quan đến vụ việc chính càng cao. Giả sử có N vụ việc chính thực tế trong D, được ký hiệu là $G=\{G_1, G_2, \ldots, G_N\}$. Nếu hơn một nửa số bài viết trong top 5 của K_j thuộc về G_i, thì K_j và G_i được coi là khớp với nhau, được ghi nhận là $Match(G_i, K_j) = 1$, và $Match(G_i, K_j) = 0$ nếu không có sự khớp. Các công thức để tính độ chính xác (precision) và độ bao phủ (recall) được định nghĩa dưới đây [2].

$$check(G_i) = \{1 \; if \; \exists G_i | Match(G_i, K_j) = 1, 0 \; else\} \quad (1)$$

$$precision = \frac{\sum_{G_i} check(K_i)}{1} \quad (2)$$

$$precision = \frac{\sum_{G_i} check(K_i)}{N} \quad (3)$$

Cuối cùng, F-measure là trung bình của 2 độ đo precision và recall với công thức như sau:

$$F - measure = \frac{2(precision * recall)}{precision + recall} \quad (4)$$

IV. THỰC NGHIỆM VÀ KẾT QUẢ

A. Dữ liệu và môi trường thực nghiệm

Trong nghiên cứu này, chúng tôi sử dụng tập dữ liệu được thu thập từ các trang báo điện tử tại Việt Nam. Nội dung của các bài báo đa dạng về các chủ đề y tế, chính trị, xã hội, môi trường v.v. Từ các bài báo thu thập được, chúng tôi tiến hành bước gán nhãn thủ công để tạo các cụm tin với nhãn thực. Cụ thể, với hơn 4.000 bản tin thu thập được, chúng tôi tiến hành phân cụm thủ công và thu được 70 cụm với số tin/cụm khoảng từ 5-30 tin. Các bản tin không tạo thành các cụm có số lượng lớn hơn 5 được loại bỏ. Tổng cộng có hơn 1.057 tin được sử dụng trong các cụm được tạo.

B. Kết quả thực nghiệm

Thuật toán phân cụm DBSCAN có ưu điểm là không yêu cầu định trước số lượng cụm trước khi thực hiện. Tuy nhiên, thuật toán này khá nhạy cảm với việc lựa chọn tham số. Có hai tham số chính cần lựa chọn và tối ưu khi tiến hành phân cụm theo thuật toán DBSCAN như mô tả dưới đây:

- eps: Khoảng cách lớn nhất giữa hai điểm dữ liệu để được xem là nằm trong cùng một cụm.
- min_samples: số lượng tối thiểu các điểm để có thể xem là một cụm.

Do tập dữ liệu kiểm thử được xây dựng có tối thiểu năm bản tin (điểm dữ liệu) cho mỗi cụm, trong thực nghiệm này chúng tôi thiết lập min_samples với giá trị 5. Đối với tham số eps, các kết quả thực nghiệm cho thấy eps đạt tối ưu ở giá trị 1.1.

Bảng 1 cho kết quả của thuật toán phân cụm DBSCAN với các bộ tham số tối ưu ở các độ đo precision, recall, và F-measure, với các nhúng từ được tạo ra sử dụng PhoBERT.

Bảng 1. *Kết quả phân cụm theo DBSCAN và PhoBERT*

Tham số		**Prec. (%)**	**Rec. (%)**	**F1 (%)**
eps	*min_sps*			
0.8	5	92.3	78.5	85.1

Các kết quả trên cho thấy hiệu quả của phương pháp đề xuất, trên cơ sở so sánh với các nghiên cứu trước đây về phân cụm tin (có kết quả base-line khoảng hơn 80%). Kết quả tốt ở cả các độ đo precision, recall, và F-measure cho thấy phương pháp có độ chính xác cả ở vấn đề phân hoạch bản tin vào cụm và số lượng cụm được tạo ra.

Để đánh giá hiệu quả của mô hình sinh nhúng từ dựa trên PhoBERT so với các cách tiếp cận khác như mô hình nhúng từ truyền thống hoặc mô hình LLM (Large Language Model), chúng tôi cũng tiến hành thực nghiệm với các mô hình GloVe và Vistra 7B [18]. Kết quả ở bảng 2 cho thấy PhoBERT có kết quả tốt hơn so với việc sinh nhúng từ sử dụng GloVe hay mô hình ngôn ngữ lớn Vistra 7B. Các kết quả cho thấy phân cụm từ nhúng từ được tạo theo PhoBERT có kết quả vượt trội mô hình truyền thống GloVe. So với mô hình ngôn ngữ lớn Vistra 7B, PhoBERT cũng có kết quả tốt hơn, mặc dù không quá vượt trội nhưng còn có thêm ưu điểm về tốc độ và khối lượng xử lý so với mô hình ngôn ngữ lớn như Vistra 7B.

Bảng 2. *Kết quả phân cụm dựa trên các mô hình sinh nhúng từ khác nhau*

Mô hình sinh nhúng từ	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
GloVe	88.3	75.2	82.7
Vistra 7B	91.7	77.8	84.7
PhoBERT	**92.3**	**78.5**	**85.1**

V. KẾT LUẬN

Bài báo thực hiện nghiên cứu phân cụm tin tức theo các vụ việc hoặc sự kiện truyền thông sử dụng PhoBERT để trích chọn từ khóa và sinh nhúng từ, đồng thời sử dụng DBSCAN để phân cụm bản tin. Các thực nghiệm được thực hiện trên tập dữ liệu thu thập từ mạng Internet và gán nhãn thủ công.

Các kết quả thực nghiệm cho thấy việc phân cụm các tin tức là bài toán có tiềm năng ứng dụng thực tế và mang lại nhiều hiệu quả trong công tác quản lý thông tin mạng cũng như mang lại trải nghiệm tốt hơn cho người dùng.

Các hướng phát triển tiếp theo bao gồm việc thực hiện bổ sung các tác vụ tiền xử lý như tóm tắt trước khi phân cụm hoặc ứng dụng nhiều hơn các mô hình ngôn ngữ lớn trong các bước của hoạt động phân cụm.

VI. LỜI CẢM ƠN

Nghiên cứu được thực hiện dưới sự hỗ trợ của Viện Inres.AI (www.inres.ai) và đã được ứng dụng thử nghiệm trên trang tổng hợp tin tại địa chỉ Web www.8am.vn.

TÀI LIỆU THAM KHẢO

[1] J. Allan, R. Papka, and V. Lavrenko, "Online New EventDetection and Tracking," in SIGIR 1998 - Proceedings ofthe 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.doi: 10.1145/290941.290954.

[2] Z. Liu, Y. Zhang, Y. Li, and Chaomurilige, "Key NewsEvent Detection and Event Context Using GraphicConvolution, Clustering, and Summarizing Methods," Appl.Sci., 2023, doi: 10.3390/app13095510.

[3] R. Toujani and J. Akaichi, "Event news detection andcitizens community structure for disaster management insocial networks," Online Inf. Rev., 2019, doi: 10.1108/OIR-03-2018-0091.

[4] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchicaldensity based clustering," J. Open Source Softw., 2017, doi:10.21105/joss.00205.

[5] J. Devlin, M.W. Chang, K. Lee and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2019.

[6] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv (Cornell University), Jan. 2013, doi: 10.48550/arxiv.1301.3781.

[7] J. Pennington, R. Socher, and C. D. Manning, "GloVe:Global vectors for word representation," in EMNLP 2014 -2014 Conference on Empirical Methods in NaturalLanguage Processing, Proceedings of the Conference, 2014.doi: 10.3115/v1/d14-1162

[8] N.Q.Dat and N.A.Tuan. PhoBERT: Pre-trained language models for Vietnamese. arXiv preprint arXiv:2003.00744. 2020.

[9] A. N. Tarekegn, "Large Language Model Enhanced

Clustering for News Event Detection," *arXiv.org*, Jun. 15,
[10] 2024. https://arxiv.org/abs/2406.10552
[11] H. Hettiarachchi, M. Adedoyin-Olowe, J. Bhogal, and M. M. Gaber, "Embed2Detect: temporally clustered embeddedwords for event detection in social media," Mach. Learn.,2022, doi: 10.1007/s10994-021-05988-7.
[12] S. Carta, S. Consoli, L. Piras, A. S. Podda, and D. R.Recupero, "Event Detection in Finance Using HierarchicalClustering Algorithms on News and Tweets," PeerJ Comput.Sci., 2021, doi: 10.7717/PEERJ-CS.438.
[13] P. Goyal, P. Kaushik, P. Gupta, D. Vashisth, S. Agarwal,and N. Goyal, "Multilevel Event Detection, StorylineGeneration, and Summarization for Tweet Streams," IEEETrans. Comput. Soc. Syst., 2020, doi:10.1109/TCSS.2019.2954116.
[14] Y. Liu, et al., "Roberta: A robustly optimized bert pretraining approach," arXiv 2019, arXiv preprint arXiv: 1907.11692.
[15] MaartenGr, "GitHub - MaartenGr/KeyBERT: Minimal keyword extraction with BERT," GitHub. https://github.com/MaartenGr/KeyBERT
[16] A. K. Jain, "Data clustering: 50 years beyond K-means,"Pattern Recognit. Lett., 2010, doi:10.1016/j.patrec.2009.09.011.
[17] M. Vichi, C. Cavicchia, and P. J. F. Groenen, "HierarchicalMeans Clustering," J. Classif., 2022, doi: 10.1007/s00357-022-09419-7.
[18] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in Lecture Notes in Computer Science, 2013. doi: 10.1007/978-3-642-37456-2_14
[19] Vistral-7b-chat – Towards a state-of-the-art large language model for Vietnamese

VIETNAMESE NEWS ARTICLE CLUSTERING USING MACHINE LEARNING

Abstract: In the current era of online information explosion, tens of thousands of articles and news reports are published daily. In addition to electronic newspapers, other sources like social media also provide rapid and diverse online news, meeting users' information needs. However, the large volume of news articles produced also leads to information overload, with many duplicate news articles on the same story or event. To quickly grasp information on prominent news, organizing and clustering news according to story or events from different sources is an essential task that enhances the reading experience. This paper presents a method for news clustering using the DBSCAN clustering algorithm based on word embeddings extracted from the PhoBERT model. Experiments were conducted on a dataset of over 1,000 news articles collected from Vietnamese electronic newspapers through an automatic news collection system. The experimental results were evaluated using precision, recall, and F-measure metrics, with accuracies of 92.3%, 78.5%, and 85.1%, respectively.

Keywords: news articles clustering, phobert, dbscan.

Dương Trần Đức Tốt nghiệp Thạc sỹ chuyên ngành Hệ thống thông tin tại Đại học Tổng hợp Leeds, Vương Quốc Anh năm 2004, và Tiến sỹ chuyên ngành Kỹ thuật máy tính tại Học viện Công nghệ Bưu chính Viễn thông năm 2018. Hiện đang công tác tại Khoa Công nghệ Thông tin, Học viện Công nghệ Bưu chính Viễn thông.