RESEARCH ARTICLE

GOM CỤM BÀI BÁO THEO CHỦ ĐỀ

Nguyễn Lê Minh Hòa[1], Trần Văn Lăng[2]

[1]Trường Đại học Lạc Hồng
[2]Trường Đại học Ngoại ngữ -Tin học TP.HCM
hoanlmbd@gmail.com, langtv@huflit.edu.vn

TÓM TẮT— Trong hội nghị khoa học, ban tổ chức chương trình phải dành thời gian để chọn lọc, sắp xếp chủ đề vào từng phiên họp phù hợp. Trước sự gia tăng đáng kể về số lượng bài báo được gửi về trong một hội nghị khoa học, việc sử dụng công cụ hỗ trợ để tối ưu hóa thời gian trở thành một nhu cầu cấp thiết. Bài báo tập trung vào việc nghiên cứu và áp dụng mô hình học sâu PhoBERT nhằm gom cụm các bài báo khoa học, với thử nghiệm được tiến hành trên bộ dữ liệu của hội nghị khoa học FAIR. Kết quả của thử nghiệm cho thấy rằng PhoBERT đã đạt được hiệu suất cao, với điểm F1-score đạt tỷ lệ cao trên bộ dữ liệu bài báo của hội nghị FAIR.

Từ khóa— PhoBERT, phân loại văn bản, tiền xử lý, Softmax

I. GIỚI THIỆU

Khi gửi bài tham dự hội nghị khoa học, các nhà nghiên cứu thường phải viết tóm tắt công trình (abstract) của mình cùng với một số từ khoá (keyword) và có thể đề xuất chủ đề liên quan trong số những chủ đề (topic) được liệt kê. Tóm tắt thể hiện vấn đề gì mà công trình giải quyết, việc giải quyết bằng cách gì, kết quả đạt được là gì, kết quả đó ra sao. Còn từ khoá là những cụm từ mang tính đặc trưng của bài báo, qua đó giúp người quản lý biết được công trình này thuộc chủ đề hẹp là gì; thông thường cũng chỉ có tối đa là 5 cụm từ.

Tại một hội nghị khoa học quy mô lớn, các chủ đề thường được phân loại và bố trí vào các phiên làm việc riêng (session); từ đó sắp xếp để công trình chỉ trình bày trong một phiên làm việc cụ thể. Khi số lượng bài báo gửi về tham dự vô cùng lớn, vì vậy người quản trị với tư cách là ban chương trình rất vất vả trong việc làm sao chọn chủ đề đúng cho phiên làm việc. Bên cạnh đó, việc cung cấp từ khoá của tác giả bài báo cũng chưa đáp ứng mong đợi, nên cần phải xem xét thêm nội dung tóm tắt mới sắp xếp phù hợp vào từng phiên làm việc. Nhưng việc này đòi hỏi ban chương trình phải có chuyên môn sâu.

Từ đó, với sự phát triển của chủ đề xử lý ngôn ngữ tự nhiên, đặc biệt là tiếng Việt; cùng với các công nghệ tương đối phong phú giúp cho máy tính có thể xử lý một số công việc thay cho con người nếu được huấn luyện trước đó. Nên vấn đề đặt ra ở đây là xây dựng ứng dụng giúp cho ban tổ chức hội nghị khoa học quy mô lớn thu xếp việc phân loại bài báo theo chủ đề, mà cụ thể là theo phiên làm việc.

Đầu vào của bài toán gom cụm bài báo khoa học theo chủ đề là phần tóm tắt và từ khóa, đầu ra là kết quả dự đoán đoạn văn này thuộc phiên làm việc nào. Trong bài báo này, bài báo chọn phân tích đầu ra theo 6 loại chủ đề phổ biến trong ngành công nghệ thông tin gồm: Xử lý ngôn ngữ tự nhiên (NLP), thị giác máy tính (CV), trí tuệ nhân tạo (AI), mạng máy tính (NS), hệ thống thông tin (IS), khoa học dữ liệu (DS). Dữ liệu trong bài báo này được thu thập từ Hội nghị khoa học Quốc gia về Nghiên cứu cơ bản và Ứng dụng Công nghệ thông tin (FAIR) trong những năm gần đây.

Phần báo cáo được trình bày theo cấu trúc sau: Phần II và phần III lần lượt trình bày các công trình liên quan cũng như các kiến thức nền của bài báo. Phần xử lý dữ liệu được trình bày trong phần IV. Về phần thực nghiệm và kết luận được trình bày theo thứ tự ở phần V và VI.

II. CÔNG TRÌNH LIÊN QUAN

Công trình của Susie Xi Rao và cộng sự [1] đã đặt ra 4 thách thức: đánh giá dữ liệu danh sách từ khóa tham chiếu, đánh giá các từ khóa được trích xuất, số lượng bài báo khoa học, trích xuất từ khóa theo miền cụ thể. Bài báo dựa vào tập dữ liệu gồm 46985 bản ghi trên trang Web of Science Dataset (WOS) với cấu trúc như Bảng 1.

Bảng 1. *Dữ liệu mẫu tiêu chuẩn của WOS*

Y1	Y2	Y	Domain	area	keywords	Abstract
5	50	122	Medical	Sports Injuries	Elastic therapeutic tape; Material properties; Tension test	The aim of this study was to analyze stabilometry in athletes...
5	48	120	Medical	Senior Health	Sports injury; Athletes; Postural stability	This study examined the influence of range of motion of the ankle joints on elderly people's balance ability...

Công trình này có các ưu điểm là:

- So sánh giữa thuật toán TextRank và các phương pháp trích xuất cụm từ không giám sát.
- Cung cấp cơ sở nền tảng các cách trích xuất từ khóa.
- Xây dựng các phương pháp để mở rộng sang trích xuất từ khóa cho các chủ đề như pháp lý, tin tức, ...

Tuy nhiên đây chỉ là tiền đề tìm giải các giải pháp vượt qua những thách thức được nêu ra trong công trình này.

Công trình của tác giả Thaer Sahmoud, Dr. Mohammad Mikki [2] đã phát triển công cụ để phát hiện thư rác, đã được đánh giá về hiệu suất dựa trên tập dữ liệu của Enron, SpamAssasin, Ling-spam, tập dữ liệu SMS rác v.1 (Bảng 2), và đã thu được kết quả F1-Score lần lượt là 98.62%, 97.83%, 99.13%, và 99.28%. Các con số trên là kết quả của việc sử dụng mô hình được tiền huấn luyện của BERT, cho phép phát hiện thư rác theo nhiều ngữ cảnh khác nhau trong nội dung, từ đó phân loại tốt hơn với nhiều thể loại thư.

Bảng 2. *Thống kê kho ngữ liệu*

Corpus	Emails/SMSs	Spam	Ham
SMS Spam Collection v.1	5572	747	4825
SpamAssasin	5796	3900	1896
Ling-Spam	2893	481	2412
Enron	33716	16852	16493

Công trình này có ưu điểm là phát hiện được thư rác với tỉ lệ khá cao (trong khoảng từ 98 đến 99%) nhờ vào sử dụng mô hình tiền huấn luyện BERT cho phép hệ thống phát hiện thư rác dựa trên nội dung và ngữ cảnh của từng câu trong thư. Bên cạnh đó công trình còn một hạn chế là chưa đưa ra sự so sánh giữa các mô hình như BERT, PHOBERT và SVM

Tuy nhiên, để áp dụng cho tiếng Việt, còn nhiều vấn đề cần phải giải quyết như xử lý từ ngữ, văn phong sao cho phù hợp với ngữ cảnh của từng câu trong phần tóm tắt.

Với tiếng Việt, có luận văn "Phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT" của tác giả Nguyễn Diệu Linh [3]. Luận văn này xây dựng hệ thống câu hỏi và trả lời tự động liên quan đến ba chủ đề chính: xử lý ngôn ngữ tự nhiên, truy xuất thông tin và khai thác thông tin. Phân loại câu hỏi là giai đoạn đầu tiên trong kiến trúc chung của hệ thống hỏi đáp, có nhiệm vụ tìm kiếm thông tin cần thiết làm đầu vào cho quá trình xử lý các giai đoạn sau (trích xuất tài liệu, trích xuất câu, v.v.). Hệ thống hỗ trợ hỏi và trả lời một vấn đề pháp lý cần phải tra cứu, tìm kiếm nhiều văn bản pháp luật có liên quan. Qua đó giúp rút ngắn thời gian, cần phân loại câu hỏi pháp luật theo các chủ đề pháp luật. Luận văn thực hiện phân loại câu hỏi bằng cách tiếp cận học máy có giám sát, cụ thể là sử dụng một số mô hình SVM truyền thống và mô hình BERT.

Công trình này có các ưu điểm là:

- Là tiền đề đánh giá thực hiện các nghiên cứu tiếp theo với bộ dữ liệu được xây dựng trên nền tảng này.
- Đưa ra kết quả thực nghiệm so sánh 3 mô hình SVM, PHOBERT và BERT đa ngôn ngữ thông qua 3 chỉ số đánh giá chất lượng là độ chuẩn xác, độ bao phủ và thang đo F1.

Kết quả thực nghiệm tốt nhất đạt được khi sử dụng mô hình BERT là 89,47% (thang đo F1), tiếp theo là SVM 87.39% và cuối cùng là PhoBERT 86.65%.

Tuy nhiên, công trình vẫn còn một số hạn chế như sau:

- Với một số nhãn như “Cán bộ, công chức, viên chức”, kết quả dự đoán bằng 0%, nhãn “Tổ chức cơ quan, chính quyền”, ..., kết quả dự đoán không chính xác khi sử dụng mô hình BERT. Đối với mô hình SVM và BERT có kết quả khá thấp (khoảng 36%).
- Tập dữ liệu huấn luyện, xác thực và kiểm tra còn khá ít và phân bố không đồng đều dẫn đến kết quả dự đoán thấp.

Trong việc phân loại bài báo khoa học có công trình "Giải pháp phân loại bài báo khoa học bằng kĩ thuật máy học" của tác giả Trần Thanh Điện, Thái Nhựt Thanh và Nguyễn Thái Nghe [4]. Trong công trình này, tác giả đề xuất các giải pháp để rút trích thông tin và phân loại bài báo dựa vào 3 kỹ thuật: kỹ thuật máy học vectơ hỗ trợ (SVM), Navïe Bayes, kNN. Mặc dù SVM đạt tỷ lệ phân lớp cao nhất trong 3 kỹ thuật (khoảng 91%), nhưng không phải tất cả chủ đề đều có sự phân lớp tốt, đặc biệt là khi một bài báo liên quan đến nhiều chủ đề khác nhau, dẫn đến sự chồng lấn và gây ra khó khăn trong quá trình phân loại bài báo. Đối với các chủ đề có tính đặc thù, riêng biệt, tỉ lệ chính xác và độ bao phủ đạt được khá cao.

Công trình này có các ưu điểm là:

- Đánh giá kết quả dựa trên 3 kỹ thuật phân loại SVM, Navïe Bayes và kNN.

- Kết quả thu được cho thấy kỹ thuật SVM đã đạt được tỉ lệ chính xác trung bình hơn 91%, Navïe Bayes đạt 80.9% và kNN chỉ đạt 76.5%.

Bên cạnh đó tập dữ liệu còn tương đối nhỏ, do vậy để đạt độ chính xác hơn mô hình cần được thực nghiệm trên tập dữ liệu lớn hơn. Ngoài ra bài báo cũng chưa đưa ra hướng giải quyết cho trường hợp chồng lấn nêu trên.

III. PHƯƠNG PHÁP

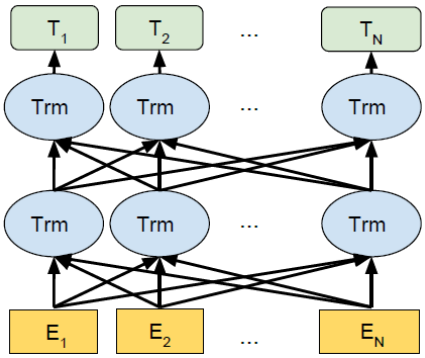
Năm 2018 là một năm đột phá trong NLP. Các mô hình như ELMo của Allen AI, Open-GPT của OpenAI và BERT của Google cho phép các nhà nghiên cứu vượt qua nhiều điểm chuẩn với mức tinh chỉnh tối thiểu cho từng tác vụ cụ thể và cung cấp cho cộng đồng NLP các mô hình được đào tạo trước dễ dàng (với ít dữ liệu hơn và thời gian tính toán ít hơn) được tinh chỉnh và thực hiện để tạo ra kết quả tốt.

A. BERT

BERT (Mô hình mã hóa hai chiều từ các khối Transformer), được tác giả J. Devlin và cộng sự công bố vào cuối năm 2018 [5]. BERT là một phương pháp huấn luyện trước các biểu diễn ngôn ngữ, mà chúng dựa trên mô hình mạng mô phỏng theo hệ thống nơ-ron thần kinh con người. Nói một cách đơn giản, mô hình này giúp phân biệt rõ hơn ngữ cảnh của các từ xuất hiện trong một câu hay đoạn văn. Ví dụ như mô hình này được sử dụng để tinh chỉnh sao cho phù hợp với nhiệm vụ cụ thể như phân loại, trả lời câu hỏi, ... với dữ liệu được xử lý theo nhu cầu từ cá nhân cho đến tổ chức.

Điểm mạnh của BERT nằm ở khả năng huấn luyện các mô hình ngôn ngữ dựa trên toàn bộ tổ hợp các từ trong một câu hay có thể hiểu là huấn luyện hai chiều, thay cho cách huấn luyện truyền thống từ trái sang phải. BERT cho phép học ngữ cảnh từ các từ xung quanh nó, thay vì chỉ quan tâm từ trước và ngay sau nó.

Hình 1 mô tả kiến trúc mô hình BERT, trong đó E_n, Trm và T_n là các lớp biểu diễn trong quá trình xử lý ngôn ngữ tự nhiên. E_n (Embeddings Layer) là lớp biểu diễn các từ trong văn bản thành các vectơ có chiều thấp hơn để có thể xử lý được bởi mô hình. Trong lớp Trm (Transformer Encoder) chứa các khối mã hóa Transformer giúp biến đổi dữ liệu đầu vào. Bên trong chúng chứa nhiều lớp tự chú ý (Self-attention) và lớp Feedforward Neural Network. Lớp cuối cùng là T_n được sử dụng để giải mã đầu ra, tức là chuyển đổi các vectơ biểu diễn ngôn ngữ đã được mã hóa trước đó thành chuỗi các từ hoặc câu.



Hình 1. *Kiến trúc mô hình BERT*

Có 2 vấn đề thách thức dẫn đến mô hình BERT huấn luyện cho tiếng Việt không đạt hiệu quả cao:

Thứ nhất, kho ngữ liệu của Wikipedia tiếng Việt chỉ dùng dữ liệu để huấn luyện mô hình đơn ngôn ngữ. Kho dữ liệu này không đại diện cho việc sử dụng một ngôn ngữ chung, và dữ liệu tiếng Việt trên Wikipedia có kích thước khá nhỏ, khoảng 1GB, không đủ lớn để cải thiện mô hình ngôn ngữ tiền huấn luyện.

Thứ hai, các mô hình ngôn ngữ $BERT_{base}$ được công bố không nhận biết được từ ngữ tiếng Việt, ví dụ như từ ghép, từ đơn. Sự mơ hồ này đến từ việc khoảng trắng còn được dùng để ngăn cách các âm tiết cấu thành từ khi viết trong tiếng Việt. Ví dụ, một văn bản có 6 âm tiết "Tôi là một nhà nghiên cứu" tạo thành 4 từ "$Tôi_I$ $là_{am}$ $một_a$ $nhà\_nghiên\_cứu_{researcher}$". Đoạn văn bản này không được tiền xử lý phân đoạn cho tiếng Việt, điều này dẫn đến cụm từ "nhà_nghiên_cứu" được tách ra thành từng từ không đúng với ý nghĩa cần được hiểu. Vì vậy, dữ liệu ở cấp độ âm tiết (syllable-level) sẽ cho ra kết quả không được tốt so với mô hình ngôn ngữ huấn luyện trước trên dữ liệu ở cấp độ từ (word-level).

B. PHOBERT

Mô hình RoBERTa [6] là một cải tiến quan trọng với BERT là nền tảng. Mô hình này vẫn giữ nguyên các tham số và cấu trúc *Transformers*, chỉ thay đổi các siêu tham số và loại bỏ đi khối dự đoán câu tiếp theo (NSP). Trong dự

đoán câu tiếp theo, mô hình được đào tạo để dự đoán câu tiếp theo có cùng hàm ý ý nghĩa và mối quan hệ với câu trước đó. Trong bài báo về RoBERTa, các tác giả đã thử nghiệm với việc loại bỏ/thêm độ mất mát (loss) NSP và kết luận rằng việc loại bỏ độ mất mát này là phù hợp, vì cải thiện được hiệu năng của các tác vụ tiếp theo. Cùng với đó, việc huấn luyện các chuỗi câu dài hơn và kích thước của batch lớn hơn, điều này đã giúp nhận biết các từ ngữ bị che theo nhiều trường hợp khác nhau, cũng như tăng độ chính xác ở các tác vụ về sau. Cuối cùng, sự linh hoạt trong việc che từ ở mỗi epoch khác nhau đã tạo cho mô hình học được nhiều trường hợp từ hơn. Cụ thể với kiến trúc BERT, việc tạo lớp che từ ngẫu nhiên chỉ được thực hiện trong quá trình tiền xử lý dữ liệu, còn gọi là lớp che tĩnh. Tuy nhiên với kiến trúc RoBERTa, việc che từ được thực hiện ngẫu nhiên qua mỗi epoch, từ đó tạo thành lớp che động không bị trùng lắp.

PhoBERT là mô hình huấn luyện trước dành riêng cho tiếng Việt, do tác giả Nguyen và cộng sự trình bày và thực hiện ở viện nghiên cứu VinAI Việt Nam vào tháng 03/2020. Mô hình này dựa trên kiến trúc và cách tiếp cận giống RoBERTa được giới thiệu bởi Facebook trong năm 2019 và áp dụng rộng rãi trong các bài bài toán NLP như tóm tắt văn bản, trả lời câu hỏi, ...

Để giải quyết vấn đề thứ nhất liên quan tới BERT, tác giả đã thực hiện huấn luyện trên tập dữ liệu khoảng 20GB, bao gồm 1GB dữ liệu trên Wikipedia Việt Nam và 19GB dữ liệu còn lại lấy từ ngữ liệu tin tức tiếng Việt. Với vấn đề thứ 2, tác giả đã sử dụng RDRSegmenter của VnCoreNLP để tách từ và phân đoạn từ trước khi thực hiện mã hóa fastBPE (khác với BPE trên RoBERTa). Bảng 3 mô tả thông tin cơ bản hình thành nên kiến trúc BERT, RoBERTa, và PhoBERT.

Bảng 3. *So sánh các mô hình kiến trúc trong NLP*

So sánh	BERT	RoBERTa	PhoBERT
Tham số	Base: 110M Large: 340M	Base: 110M Large: 340M	Base: 110M Large: 340M
Số lớp / Hidden Dimensions / Self-Attention heads	Base: 12/768/12 Large: 24/1024/16	Base: 12/768/12 Large: 24/1024/16	Base: 12/768/12 Large: 24/1024/16
Dữ liệu	16GB	160GB	20G
Mô hình ngôn ngữ	Đa ngôn ngữ	Đa ngôn ngữ	Đơn ngữ - Tiếng Việt
Huấn luyện	Mô hình ngôn ngữ bị che (MLM) Dự đoán câu tiếp theo (NSP)	MLM	MLM

C. KIẾN TRÚC HỆ THỐNG

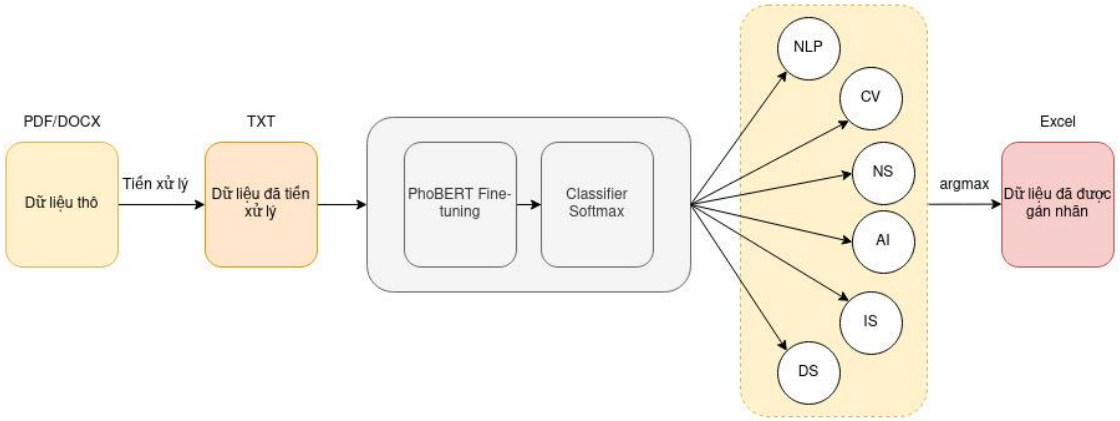
Bởi vì số lượng đề tài khoa học mỗi năm càng tăng, việc dùng máy móc thay cho con người cũng càng được chú trọng và phổ biến trong xã hội ngày nay, bài toán gom cụm bài báo khoa học rất hữu ích cho các ban tổ chức hội nghị khoa học. Cùng với đó, PhoBERT là một kiến trúc NLP rất nổi trội dùng cho tiếng Việt, thu hút khá nhiều bài báo khoa học, bài báo học thuật nhắc đến và ứng dụng vào thực tế.

Bài báo sử dụng mô hình $\text{PhoBERT}_{\text{base}}$ để thực nghiệm việc gom cụm bài báo khoa học trên tập dữ liệu của hội nghị FAIR.

Kiến trúc hệ thống phân loại bài báo được mô hình hóa như hình 2. Sau khi các tác giả nộp bài báo dạng docx/pdf lên hệ thống để ban chương trình xét duyệt, hệ thống sẽ tự động thực hiện tiền xử lý tệp này thành tệp có dạng txt. Dữ liệu sau đó sẽ được đưa vào mô hình phân lớp PhoBERT với các lớp đại diện cho nhóm chủ đề đã được chuẩn bị trước. Cuối cùng, một tập tin kết quả (một tập tin Excel) chứa các thông tin như tên bài báo, tên tập tin và kết quả dự đoán được ban chương trình của hội nghị sử dụng để sắp xếp vào từng phiên làm việc phù hợp.

Đầu vào: một đoạn văn chứa các từ khóa và đoạn tóm tắt trong bài báo.

Đầu ra: bài báo là một trong 6 chủ đề đã chuẩn bị trước.



Hình 2. *Mô hình phân loại bài báo khoa học*

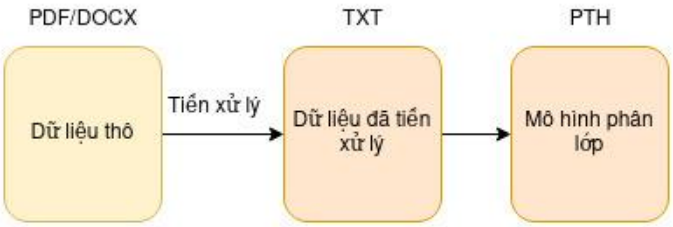
Mặc dù các bài báo đã được viết theo cấu trúc chung, nhưng vẫn còn một số vấn đề về định dạng không hoàn toàn giống nhau. Vì vậy, các bài báo khi đưa vào hệ thống cần xử lý để rút trích phần từ khóa và tóm tắt trước khi được đưa vào mô hình. Các vấn đề trong quá trình tiền xử lý sẽ được mô tả chi tiết ở phần *TIỀN XỬ LÝ DỮ LIỆU*.

D. GIAI ĐOẠN PHÂN LOẠI

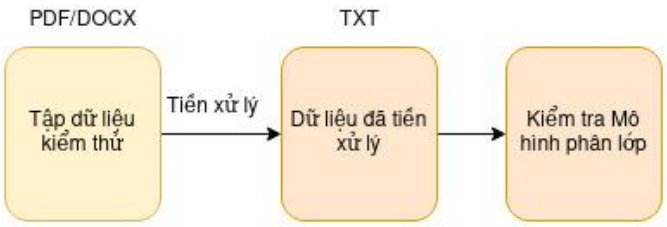
Quá trình thực hiện bài báo này gồm hai giai đoạn chính:

- **Giai đoạn huấn luyện**: Dựa vào tập dữ liệu đã tiền xử lý, tiến hành cho máy học để tạo ra mô hình phân lớp.
- **Giai đoạn kiểm thử**: Dựa vào mô hình phân lớp trên, tiến hành phân loại bài báo dựa vào tập dữ liệu xác thực.

Cả 2 giai đoạn đều được thực hiện cụ thể với mô tả lần lượt như hình 3 và hình 4.



Hình 3. *Giai đoạn huấn luyện cho ra mô hình phân lớp*



Hình 4. *Giai đoạn kiểm thử kiểm tra mô hình phân lớp*

Giai đoạn đầu tiên được mô tả ở hình 3, dữ liệu sẽ được thu thập, tiền xử lý và chia thành các tập dữ liệu huấn luyện, tập dữ liệu xác thực và kiểm tra. Sau đó, mô hình sẽ được xây dựng và cài đặt các tham số. Quá trình huấn luyện để tạo ra mô hình phân lớp sẽ được thực hiện nhiều lần, kết hợp với việc thay đổi những thông số như tốc độ học, số lượng epoch, kích thước batch, ... cùng với việc đánh giá tập dữ liệu xác thực để có thể đạt kết quả tốt nhất. Giai đoạn kiểm thử kiểm tra trong hình 4 là giai đoạn quan trọng tiếp theo trong quá trình xây dựng mô hình phân lớp. Tại đây, mô hình được đánh giá bằng cách sử dụng tập dữ liệu độc lập để đảm bảo tính hiệu quả và độ chính xác. Quá trình này thường dùng các chỉ số đánh giá hiệu suất như độ chính xác, độ phủ, độ đo F1.

1. TIỀN XỬ LÝ DỮ LIỆU

Tập dữ liệu được tiền xử lý bằng cách chuyển định dạng từ docx/pdf sang txt, sau đó lọc tách những cụm từ có chứa chữ tóm tắt và từ khóa trong toàn bộ nội dung văn bản thành 2 tập tin riêng biệt. Để mô hình được huấn luyện tốt, dữ liệu cần được xử lý để giải quyết các vấn đề sau:

- Lọc và sửa các từ đồng nghĩa nhưng khác định dạng ở phần tóm tắt (ví dụ như: "Tóm Tắt", "tóm tắt", "TÓM TẮT", "Abstract" ...) thành một cụm từ *"Tóm tắt"* đứng đầu câu, tương tự với phần từ khóa cũng cần được chuyển thành cụm từ cố định là *"Từ khóa"*.
- Phần tóm tắt và từ khóa có chứa những ngôn ngữ là tiếng Anh, bài báo đã sử dụng công cụ tác từ AutoTokenizer và AutoModelForSeq2SeqLM để phiên dịch toàn bộ ngôn ngữ sang tiếng Việt.
- Tách từ trong tiếng Việt khác với tiếng Anh vì có những từ, cụm từ kết hợp với nhau tạo ra ý nghĩa khác. Ví dụ như từ vựng "học sinh" được tạo thành từ "học" và "sinh" đều có ý nghĩa riêng. Vì vậy để giải quyết vấn đề này, bài báo đề xuất sử dụng thư viện VnCoreNLP để tách từ trong dữ liệu ở cả 2 tập tin.

Dữ liệu sau khi được xử lý có dạng như sau: "tên chủ đề" + "từ khóa" + "tóm tắt".

Ví dụ: câu "__label__NS Internet vạn_vật (IoT), An_toàn thông_tin, Mạng cảm_biến không dây (WSN), DTLS, Overhearing, An_ninh mạng. Internet vạn_vật (IoT) đang phát_triển cả về số_lượng và chất_lượng, ...", trong đó:

- *__label__NS* là nhãn thuộc chủ đề hệ thống mạng
- *Internet vạn_vật (IoT), An_toàn thông_tin, Mạng cảm_biến không dây (WSN), DTLS, Overhearing, An_ninh mạng* là các từ khóa rút trích từ bài báo
- *Internet vạn_vật (IoT) đang phát_triển cả về số_lượng và chất_lượng, ...* là nội dung phần tóm tắt đã được dịch sang tiếng Việt trong bài báo.

Trong tập dữ liệu, ngoài những bài báo đã được gán nhãn sẵn được tổng hợp từ Ban chương trình của hội nghị FAIR, một số thành viên khác cũng đã hỗ trợ gán nhãn khoảng 45% lượng dữ liệu còn lại. Vì số lượng dữ liệu nhỏ, bài báo được thực hiện gán nhãn bằng tay dựa vào kinh nghiệm và kiến thức từ trước đến nay.

2. MÔ HÌNH PHÂN LỚP

Hình 5 mô phỏng quá trình thực hiện việc phân loại một câu vào chủ đề AI được mô tả chi tiết như sau:

Bước 1: Đầu vào là câu "IoT được ứng dụng rộng rãi trong đời sống xã hội", sử dụng RDRSegmenter của VnCoreNLP để tách từ trong câu.

Bước 2: Thêm token [CLS] để bắt đầu câu đầu tiên và kết thúc mỗi câu là token [SEP] để nhận biết từng câu trong đoạn văn. Sử dụng BPE để mã hóa câu dưới dạng từ phụ (subword) thành các chỉ mục (index) dạng số dùng cho mô hình huấn luyện.

Bước 3: Đưa tokenizer vào mô hình huấn luyện trước của PhoBERT, dùng hàm Softmax để tính xác suất cho từng chủ đề.

Bước 4: Lấy giá trị lớn nhất trong vectơ ở bước 3 và trả kết quả như hình 5.

Hàm Softmax tính toán xác suất xảy ra sự kiện, nghĩa là hàm này tính toán xác suất của một lớp xuất hiện trong tổng các lớp đã cho. Xác suất này được sử dụng để xác định lớp mục tiêu của đầu vào.

Cụ thể, hàm Softmax chuyển đổi bất kỳ vectơ k chiều có giá trị thực bất kỳ thành vectơ k chiều có giá trị thực với tổng bằng một. Giá trị đầu vào có thể dương, âm, 0 hoặc lớn hơn 1, nhưng hàm Softmax luôn chuyển đổi nó thành một giá trị trong phạm vi *(0:1)*.

- Nếu bất kỳ giá trị đầu vào nào rất nhỏ hoặc âm, hàm Softmax sẽ chuyển đổi chúng thành các giá trị có xác suất nhỏ.
- Nếu đầu vào lớn, nó chuyển thành xác suất lớn.

Công thức hàm Softmax:

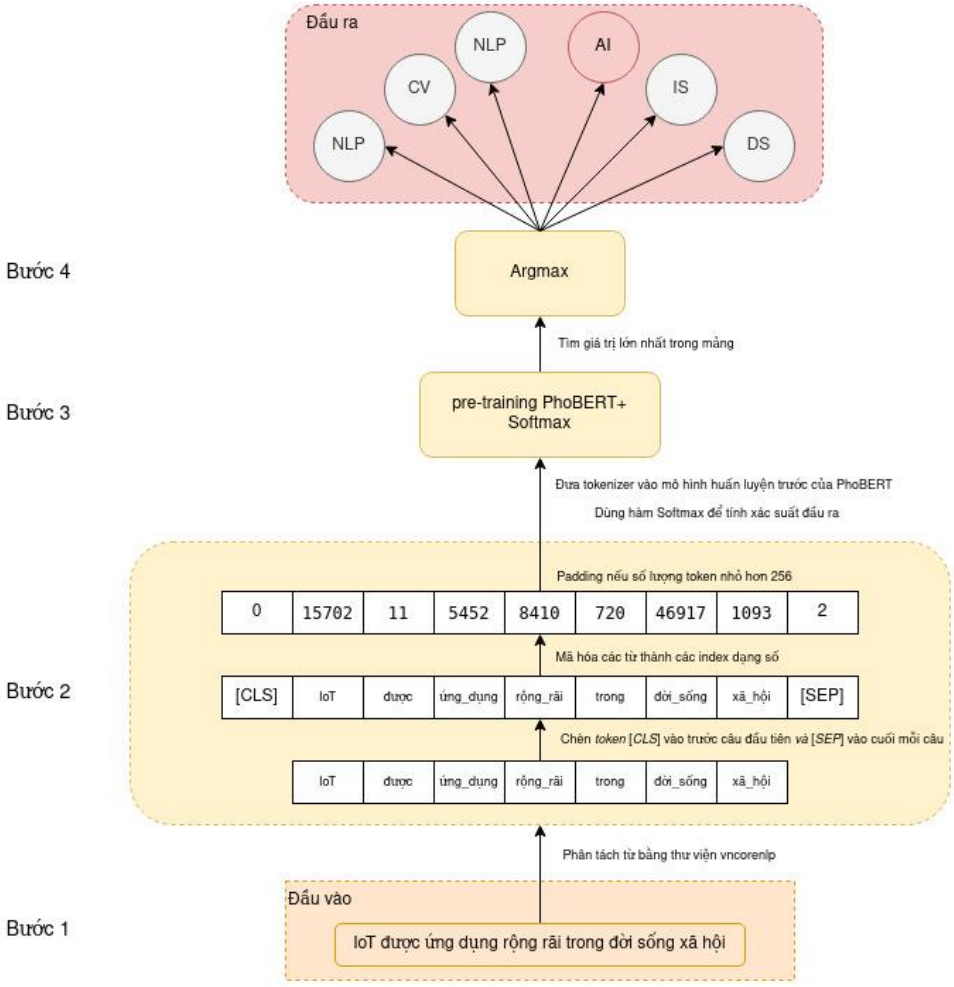
$$softmax(Z_i) = \frac{exp^{Z_i}}{\sum_{j=1}^{C} exp^{Z_i}}$$

Trong đó:

Với i nằm trong khoảng từ 1 đến C (C là số lượng lớp cần được phân loại của mô hình)

- Z_i: gồm các giá trị phần tử của vectơ đầu vào
- $\exp^{Z_i}$ là kết quả trong phạm vi từ 0 đến 1

Trong bài báo này, hàm Softmax được đặt ở lớp cuối của mạng để đánh giá xác suất phân loại các chủ đề của đầu vào.



Hình 5. *Mô phỏng phân loại một câu thuộc chủ đề AI*

3. PHƯƠNG PHÁP ĐÁNH GIÁ

Trong học máy, có nhiều cách đánh giá mô hình phân loại được sử dụng phổ biến. Độ chính xác (Precision) và Độ bao phủ (Recall) được thực hiện theo công thức (1) và (2):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (2)$$

True Positive (TP): số lượng điểm của lớp positive được phân loại đúng là positive.

True Negative (TN): số lượng điểm của lớp negative được phân loại đúng là negative.

False Positive (FP): số lượng điểm của lớp negative bị phân loại nhầm thành positive.

False Negative (FN): số lượng điểm của lớp positive bị phân loại nhầm thành negative.

Tuy nhiên, cả hai phương pháp đánh giá trên không hoàn toàn đánh giá được cả mô hình. Với độ chuẩn xác, mô hình chỉ đưa ra dự đoán một điểm mà nó chắc chắn nhất, khi đó Độ chuẩn xác bằng 1, ta không thể nói đây là mô hình tốt. Nếu chỉ dùng Độ bao phủ, mô hình dự đoán tất cả các điểm đều là tích cực (Positive), khi đó, Độ bao phủ bằng 1, điều này cũng không thể hiện rằng mô hình này tốt. Khi đó Điểm F1 (F1-Score) được sử dụng để trung bình điều hòa (Harmonic Mean) cho 2 phương pháp. F1-Score được tính theo công thức như sau:

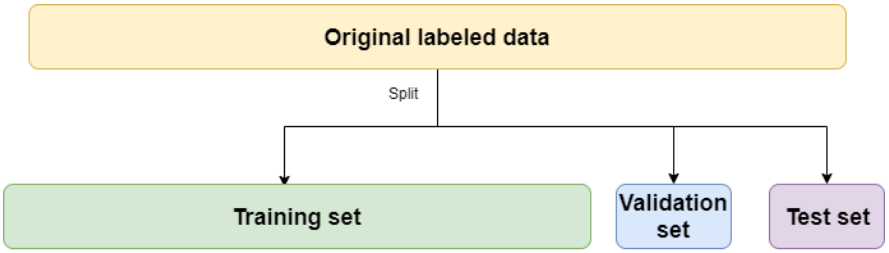
$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2\frac{Precision \times Recall}{Precision + Recall} \qquad (3)$$

Trong phạm vi bài báo đang thực hiện là bài toán phân loại nhiều lớp; và với mỗi lớp, dữ liệu thuộc lớp đó sẽ có nhãn là tích cực, tất cả các dữ liệu còn lại có nhãn là tiêu cực. Phép đánh giá được sử dụng thêm trong bài báo này là trung bình trọng số tối đa (weight macro-average), thường được dùng trong trường hợp mất cân bằng giữa các nhãn, và được tính dựa trên hai tham số độ chính xác trung bình (macro-average precision) và độ bao phủ trung bình (macro-average recall). Kết quả của chúng là trung bình cộng của kết quả độ chính xác và độ bao phủ cho từng nhãn được tính ở công thức (1), (2), lần lượt thể hiện qua công thức tổng quát (4) và (5) như bên dưới, trong đó TP_c, FP_c, FN_c lần lượt là TP, FP, FN của lớp C:

$$\text{macro average precision} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C}(TP_c + FP_c)} \quad (4)$$

$$\text{macro average recall} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C}(TP_c + FN_c)} \quad (5)$$

Trong quá trình xây dựng mô hình phân loại, việc chia tập dữ liệu là một bước quan trọng để đảm bảo tính chính xác và khả năng tổng quát hóa của mô hình. Để đảm bảo tính khách quan trong quá trình đánh giá mô hình, tập dữ liệu được chia ngẫu nhiên thành các tập nhỏ bao gồm tập dữ liệu huấn luyện (Training set), tập dữ liệu xác thực (Validation set) với tỉ lệ là 9:1, còn lại là tập dữ liệu kiểm tra (Test set) như hình 6.



Hình 6. *Chia dữ liệu để huấn luyện, xác thực và kiểm tra*

E. CÀI ĐẶT VÀ KẾT QUẢ THỰC NGHIỆM

Quá trình huấn luyện và kiểm thử được cài đặt trên môi trường Google Colab với GPU A100-SXM4-40GB, sử dụng ngôn ngữ lập trình Python, thư viện Huggingface, framework Pytorch và một số thư viện khác.

Dữ liệu mẫu sử dụng trong bài báo này là tập dữ liệu gồm 286 bài báo được tổng hợp trong hội nghị FAIR từ năm 2020 đến 2021. Bài báo cũng đề xuất sử dụng phương pháp gia tăng dữ liệu (Data Augmentation) nhằm tăng độ chính xác trong quá trình học. Tổng số bài báo trong tập dữ liệu tăng lên từ 286 bài báo thành 855 bài báo được trình bày như bảng 4.

Bảng 4 mô tả tập dữ liệu các bài báo được chia ra 3 phần gồm tập dữ liệu kiểm tra, tập dữ liệu huấn luyện và tập dữ liệu xác thực. Trong đó, tập dữ liệu kiểm tra gồm 18 bài báo, mỗi chủ đề chọn ra 3 bài báo. Đối với tập dữ liệu huấn luyện và tập dữ liệu xác thực, tổng số bài báo còn lại là 837, được chia ngẫu nhiên theo tỉ lệ 90:10, tương ứng với 751 và 86 bài báo. Số bài báo trong các chủ đề không đồng đều, thấp nhất là chủ đề về khoa học dữ liệu chỉ có 88 bài báo, cao nhất là chủ đề về hệ thống mạng có 183 bài báo.

Bảng 4. *Các chủ đề được tổng hợp trong hội nghị FAIR*

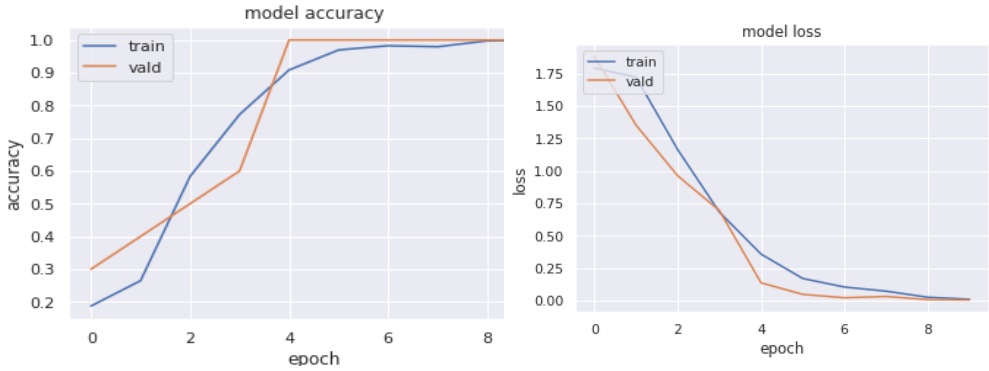
Số thứ tự	Chủ đề	Tập dữ liệu kiểm tra	Tập dữ liệu xác thực	Tập dữ liệu huấn luyện	Tổng số
1	NLP (Xử lý ngôn ngữ tự nhiên)	3	15	134	152
2	CV (Thị giác máy tính)	3	16	136	155
3	AI (Trí tuệ nhân tạo)	3	16	140	159
4	NS (Hệ thống mạng)	3	18	162	183
5	IS (Hệ thống thông tin)	3	12	103	118
6	DS (Khoa học dữ liệu)	3	9	76	88
	Tổng cộng	18	86	751	855

Bài báo đã sử dụng thuật toán tối ưu AdamW cho mô hình PhoBERT với các siêu tham số cấu hình trong quá trình fine-tuning PhoBERT được mô tả ở bảng 5. Đây là các tham số được tác giả bài báo BERT đề xuất sử dụng trong quá trình huấn luyện.

Bảng 5. *Bộ siêu tham số của mô hình PhoBERT*

Learning Rate	Batch size	Max length	Epoch
2e-5	32	256	10

Hình 7 trình bày kết quả đánh giá trên tập dữ liệu huấn luyện (train) và tập dữ liệu xác thực (vald) đánh giá hiệu suất trong việc phân loại chủ đề của các bài báo. Mô hình đạt độ chính xác tối đa là 100% tại epoch thứ 4 trên tập dữ liệu xác thực và tại epoch thứ 8 với tập dữ liệu huấn luyện. Đối với mô hình mất mát, độ mất mát trong quá trình huấn luyện giảm dần và tiệm cận với giá trị 0 từ epoch thứ 6.



Hình 7. *Độ chuẩn xác và độ lỗi của mô hình tốt nhất*

Mô hình được đánh giá dựa vào chỉ số độ chuẩn xác (Precision), độ bao phủ (Recall) và độ đo F1. Thực hiện đánh giá mô hình theo từng chủ đề bằng thư viện phân lớp Sklearn, sau đó tiến hành thống kê kết quả. Như bảng 6 thể hiện kết quả thực nghiệm, cho thấy mặc dù số lượng tập dữ liệu huấn luyện và tập dữ liệu xác thực rất ít, nhưng tất các chủ đề của mô hình PhoBERT đều được phân loại rất tốt và cho ra kết quả F1-Score xấp xỉ 100%. Đối với mô hình BERT, F1-Score cho cả mô hình chỉ đạt được 94%, trong đó cao nhất là chủ đề NLP, AI và DS đạt 100%, thấp nhất là chủ đề IS đạt 80%.

Bảng 6. *Kết quả phân loại tập dữ liệu xác thực trên mô hình PhoBERT*

Chủ đề	BERT			PhoBERT			Support
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
NLP	1.00	1.00	1.00	1.00	1.00	1.00	3
CV	0.75	1.00	0.86	1.00	1.00	1.00	3
AI	1.00	1.00	1.00	1.00	1.00	1.00	3
NS	1.00	1.00	1.00	1.00	1.00	1.00	3
IS	1.00	0.67	0.8	0.99	1.00	1.00	3
DS	1.00	1.00	1.00	1.00	0.99	1.00	3
Tỷ lệ chính xác			0.94			1.00	18
Weight macro-average	0.96	0.94	0.94	1.00	1.00	1.00	18

Đối với chủ đề CV và IS trên hai mô hình BERT và PhoBERT, ta thấy rằng chủ đề CV trên mô hình BERT có kết quả F1-Score thấp hơn 14% so với mô hình PhoBERT. Tương tự, chủ đề IS cũng đạt điểm thấp hơn 20%. Cả hai chủ đề này chiếm lần lượt 13,80% và 18,596% trên bộ dữ liệu.

IV. KẾT LUẬN

Bài báo đã đề xuất mô hình PhoBERT giải quyết tốt dữ liệu trên, việc rút trích thông tin và tự động phân loại bài báo khoa học khi các tác giả gửi đăng trên các tạp chí là hoàn toàn khả thi. Ngoài ra, bài báo cũng thử nghiệm trên mô hình BERT để đánh giá và so sánh kết quả thực nghiệm của hai mô hình. Việc huấn luyện mô hình tinh chỉnh PhoBERT đã đạt kết quả cao khi cho đánh giá với tập dữ liệu xác thực tại epoch thứ 6 với F1-Score đạt 100%, độ chuẩn xác – bao phủ và độ chính xác cũng đã đạt đỉnh 100%, và tập dữ liệu kiểm tra chứa 3 bài báo cho mỗi chủ đề cũng đã thành công với kết quả tương tự.

Tuy nhiên việc thực nghiệm bài báo nên được tiếp tục thực hiện ở bộ dữ liệu kiểm tra lớn không thông qua quá trình tăng cường dữ liệu để đảm bảo duy trì các chỉ số đánh giá chất lượng. Các kết quả cho thấy sự hiệu quả vượt trội của kiến trúc PhoBERT đối với các bài toán phân loại dữ liệu tiếng Việt. Các kết quả nghiên cứu trong đề tài này cho thấy các mô hình máy học có thể dễ dàng áp dụng vào các bài toán thực tế trong hội nghị khoa học FAIR.

V. TÀI LIỆU THAM KHẢO

[1] Susie Xi Rao, et al., Keyword Extraction in Scientific Documents, SwissText 2022, arXiv:2207.01888, https://doi.org/10.48550/arXiv.2207.01888, 2022.
[2] Thaer Sahmoud, Dr. Mohammad Mikki, Spam Detection Using BERT, Jun 07, 2022, arXiv:2206.02443, https://arxiv.org/abs/2206.02443, 2022.
[3] Nguyễn Diệu Linh, PGS. TS. Ngô Xuân Bách, Phân loại câu hỏi pháp quy tiếng Việt sử dụng mô hình BERT, Học viện Công nghệ Bưu chính Viễn thông, http://dlib.ptit.edu.vn/handle/HVCNBCVT/3186, 2022.
[4] Trần Thanh Điện, Thái Nhựt Thanh và Nguyễn Thái Nghe, Giải pháp phân loại bài báo khoa học bằng kĩ thuật máy học, Tạp chí Khoa học Trường Đại học Cần Thơ. 55(4A): 29-37, http://sj.ctu.edu.vn/ql/docgia/ tacgia-256/baibao-64782.html, 2019.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 24, arXiv:1810.04805, https://arxiv.org/abs/1810.04805, 2019.
[6] Yinhan L., Myle O., Naman G., Jingfei D., Mandar J., Danqi C., Omer L., Mike L., Luke Z., Veselin S., RoBERTa: A Robustly Optimized BERT Pretraining Approach, Jun 26, arXiv:1907.11692, https://arxiv.org/abs/1907.11692, 2019.

SCIENTIFIC ARTICLE CLASSIFICATION BY TOPIC

Nguyen Le Minh Hoa, Tran Van Lang

ABSTRACT— In the scientific conference, the program organizers must spend time selecting and arranging topics for each proper session. Given a substantial number of submissions, it is necessary to use a tool that can streamline the process is imperative for organizers of scientific conferences to efficiently classify topics for each proper session. Within the scope of this article, I focus on: Building a scientific article clustering application based on PhoBERT deep learning model to evaluate the dataset built as above. The test results show that PhoBERT gives results on the FAIR conference paper dataset with an F1-score of 100%.

Nguyễn Lê Minh Hòa, tốt nghiệp kỹ sư kỹ thuật máy tính tại trường đại học Công Nghệ Thông Tin - ĐHQG HCM, thạc sĩ công nghệ thông tin tại trường Đại học Lạc Hồng. Hiện đang là chuyên viên IT tại công ty TNHH VBTech Việt Nam.

Trần Văn Lăng, tốt nghiệp đại học ngành Toán học tại Trường Đại học Tổng hợp TPHCM năm 1982 (nay là Trường ĐHKH Tự nhiên–ĐHQG-HCM); tiến sĩ Toán–Lý năm 1996 cũng tại Trường này. Hiện nay là phó giáo sư tin học tại Trường Đại học HUFLIT