

Xây dựng ứng dụng AI hỗ trợ tóm tắt bài đọc tiếng Việt cho học sinh tiểu học

Hoàng Ngọc Long*, Huỳnh Thạnh Phú

Trường Đại học Quốc tế Hồng Bàng

TÓM TẮT

Sự phát triển nhanh của trí tuệ nhân tạo (AI) và xử lý ngôn ngữ tự nhiên (NLP) mở ra cơ hội ứng dụng trong giáo dục, đặc biệt là hỗ trợ đọc hiểu cho học sinh tiểu học. Tuy nhiên, các giải pháp tóm tắt văn bản dành riêng cho học sinh tiểu học tiếng Việt hiện nay vẫn còn hạn chế, cả về số lượng lẫn hiệu quả. Nghiên cứu này đề xuất một hệ thống AI tóm tắt bài đọc tiếng Việt cho học sinh lớp 1 - 5, nhằm giúp các em tiếp cận nội dung học tập hiệu quả hơn. Hệ thống tích hợp hai mô hình ngôn ngữ tiên tiến là PhoBERT (cho tóm tắt trích xuất) và mT5 (cho tóm tắt diễn giải), được tinh chỉnh trên bộ dữ liệu 6,000 bài đọc từ sách giáo khoa và truyện thiếu nhi. Mỗi bản tóm tắt được kiểm soát theo độ dài, từ vựng và đặc điểm ngôn ngữ phù hợp từng cấp lớp theo chương trình giáo dục tiểu học. Kết quả đánh giá cho thấy hệ thống đạt hiệu suất tốt trên các chỉ số ROUGE, BLEU, BERTScore và nhận được đánh giá tích cực từ người dùng gồm học sinh, giáo viên và phụ huynh. Nghiên cứu không chỉ góp phần vào ứng dụng AI trong giáo dục mà còn mở ra hướng phát triển cho các công cụ học tập hỗ trợ ngôn ngữ tiếng Việt.

Từ khóa: giáo dục tiểu học, trí tuệ nhân tạo, tóm tắt văn bản, xử lý ngôn ngữ tự nhiên

1. ĐẶT VẤN ĐỀ

Tóm tắt văn bản tự động, là một kỹ thuật quan trọng trong xử lý ngôn ngữ tự nhiên (NLP), giúp cô đọng nội dung thành các ý chính ngắn gọn và dễ hiểu [1]. Kỹ thuật này bao gồm hai phương pháp: Tóm tắt trích xuất, chọn lọc các câu quan trọng từ văn bản gốc [2] và tóm tắt diễn giải, tái diễn đạt nội dung bằng ngôn ngữ mới thông qua các mô hình học sâu [2, 3]. Trong giáo dục, tóm tắt văn bản có tiềm năng lớn trong việc nâng cao kỹ năng đọc hiểu, đặc biệt đối với học sinh tiểu học, những em thường gặp khó khăn khi tiếp cận văn bản dài hoặc phức tạp do vốn từ vựng hạn chế và khả năng tập trung chưa phát triển đầy đủ.

Tại Việt Nam, học sinh tiểu học (6 - 11 tuổi) học các bài đọc trong sách giáo khoa Tiếng Việt (như Cánh diều, Kết nối tri thức với cuộc sống, Chân trời sáng tạo) và truyện thiếu nhi, đòi hỏi khả năng nắm bắt ý chính và bài học đạo đức. Tuy nhiên, các công cụ tóm tắt tự động tiếng Việt dành cho đối tượng này còn rất hạn chế. Mặc dù các mô hình NLP tiên tiến như PhoBERT [4] và mT5 [5] hỗ trợ xử lý ngôn ngữ tiếng Việt hiệu quả, chúng thường được huấn luyện cho các tác vụ tổng quát hoặc người trưởng thành, chưa được tinh chỉnh để đáp ứng nhu cầu đọc hiểu của trẻ em. Điều này tạo ra một khoảng trống nghiên cứu, đặc biệt trong bối cảnh nhu cầu hỗ trợ học tập cho học sinh tiểu học ngày càng tăng.

Nghiên cứu này đề xuất một ứng dụng AI tóm tắt bài đọc tiếng Việt cho học sinh lớp 1 - 5, nhằm nâng cao hiệu quả học tập và giảm tải cho giáo viên, phụ huynh. Hệ thống tích hợp hai mô hình ngôn ngữ: PhoBERT [4], được tinh chỉnh cho tóm tắt trích xuất phù hợp với học sinh lớp 4 - 5 cần bản tóm tắt chi tiết và mT5 [5], được tinh chỉnh cho tóm tắt diễn giải tạo bản tóm tắt ngắn gọn, dễ hiểu cho học sinh lớp 1 - 5. Các mô hình được huấn luyện trên bộ dữ liệu chuyên biệt gồm 6,000 bài đọc tiếng Việt từ sách giáo khoa Tiếng Việt và truyện thiếu nhi, đảm bảo phù hợp với trình độ và nhu cầu của học sinh tiểu học. Ứng dụng được triển khai dưới dạng web với giao diện thân thiện, khuyến khích hứng thú học tập của học sinh và hỗ trợ giáo viên, phụ huynh theo dõi tiến trình học tập.

Về mặt khoa học, nghiên cứu đề xuất một giải pháp sử dụng các công nghệ NLP tiên tiến thông qua việc tinh chỉnh PhoBERT [4] và mT5 [5] trên bộ dữ liệu chuyên biệt, tạo ra các bản tóm tắt chính xác và phù hợp từng cấp lớp. Về thực tiễn, hệ thống cung cấp một công cụ học tập thông minh, giúp học sinh cải thiện đọc hiểu và hỗ trợ giáo viên hiệu quả hơn. Về mặt xã hội, nghiên cứu góp phần nâng cao chất lượng giáo dục, đặc biệt ở các khu vực hạn chế tài nguyên học tập. Các phần tiếp theo của bài báo sẽ trình bày cơ sở lý thuyết, phương pháp thiết kế, kết quả thử nghiệm và

Tác giả liên hệ: Hoàng Ngọc Long

Email: longhn@hiu.vn

hướng phát triển tương lai của hệ thống.

2. CƠ SỞ LÝ THUYẾT VÀ NGHIÊN CỨU LIÊN QUAN

2.1. Tóm tắt văn bản tự động trong giáo dục

Tóm tắt văn bản tự động, là một kỹ thuật cốt lõi trong xử lý ngôn ngữ tự nhiên (NLP), giúp cô đọng nội dung thành các ý chính ngắn gọn, dễ hiểu, mà vẫn giữ trọn ý nghĩa cốt lõi. Trong giáo dục tiểu học tại Việt Nam, học sinh tiểu học thường gặp khó khăn khi đọc các văn bản dài trong sách giáo khoa Tiếng Việt hoặc truyện thiếu nhi, do vốn từ vựng hạn chế và khả năng tập trung chưa phát triển [6]. Các nghiên cứu đã khẳng định rằng tóm tắt văn bản, dù thủ công hay tự động, giúp học sinh nắm bắt ý chính, hiểu sâu bài học đạo đức và nâng cao kỹ năng đọc hiểu [7 - 9]. Chẳng hạn, Khoshnevis và cộng sự [9] chỉ ra rằng tóm tắt hỗ trợ học sinh lớp 3 - 5 phân tích các bài học từ truyện ngụ ngôn, trong khi công trình [8] nhấn mạnh vai trò của tóm tắt trong việc khơi gợi hứng thú học tập và giảm áp lực nhận thức. Tuy nhiên, việc triển khai tóm tắt tự động trong giáo dục tiểu học tiếng Việt đối mặt với nhiều thách thức. Các nghiên cứu hiện tại chủ yếu tập trung vào ngôn ngữ tiếng Anh hoặc các ứng dụng thương mại, thiếu sự chú trọng đến đặc thù ngôn ngữ học và nhu cầu sư phạm tại Việt Nam [7]. Các bộ dữ liệu tiếng Việt thường được thiết kế cho các tác vụ tổng quát, chẳng hạn như tóm tắt tin tức, không đáp ứng được yêu cầu về ngôn ngữ đơn giản và nội dung phù hợp với trình độ nhận thức của học sinh tiểu học [4]. Những hạn chế này tạo ra nhu cầu cấp thiết về các công cụ tóm tắt tự động chuyên biệt cho trẻ em Việt Nam.

2.2. Phương pháp tóm tắt và mô hình ngôn ngữ lớn

Tóm tắt văn bản tự động dựa trên hai phương pháp chính: Trích xuất và diễn giải, mỗi phương pháp mang lại giá trị riêng trong giáo dục. Phương pháp trích xuất chọn lọc các câu quan trọng từ văn bản gốc bằng thuật toán như TextRank, giữ nguyên cấu trúc câu để đảm bảo tính chính xác [2]. Ngược lại, phương pháp diễn giải sử dụng mô hình học sâu để tái diễn đạt nội dung bằng ngôn ngữ mới, đòi hỏi khả năng hiểu ngữ nghĩa sâu và sinh văn bản tự nhiên [3]. Trong giáo dục tiểu học, tóm tắt trích xuất phù hợp với học sinh lớp 4 - 5, những em cần bản tóm tắt chi tiết để phân tích nội dung, trong khi tóm tắt diễn giải đáp ứng nhu cầu của học sinh lớp 1 - 5 với các bản tóm tắt ngắn gọn, dễ hiểu [8]. So với tóm tắt trích xuất, phương pháp diễn giải linh hoạt hơn trong việc đơn giản hóa từ vựng và cấu trúc câu, phù hợp với học sinh lớp 1 - 3 có vốn từ hạn chế, nhưng có thể bỏ sót chi tiết quan trọng nếu

không được tinh chỉnh kỹ lưỡng. Ngược lại, trích xuất đảm bảo độ bao quát nội dung nhưng thiếu khả năng tái diễn đạt để phù hợp với trình độ nhận thức trẻ em. Tuy nhiên, áp dụng các phương pháp này cho tiếng Việt gặp khó khăn do đặc thù ngôn ngữ và thiếu mô hình huấn luyện chuyên biệt.

Kiến trúc Transformer đã cách mạng hóa lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), cho phép phát triển các mô hình hiệu quả cao như BERT, PhoBERT và mT5, đặc biệt trong xử lý tiếng Việt. PhoBERT, một biến thể của BERT được huấn luyện trước trên tập dữ liệu tiếng Việt quy mô lớn, đạt hiệu quả cao trong phân loại câu và trích xuất thông tin, phù hợp với tóm tắt trích xuất [4]. mT5, một mô hình text-to-text đa năng, được thiết kế để sinh văn bản tự nhiên, đáp ứng tốt yêu cầu của tóm tắt diễn giải [5]. Tuy nhiên, các nghiên cứu trước đây [4, 5, 7, 8] chủ yếu huấn luyện các mô hình này trên dữ liệu tin tức hoặc các tác vụ tổng quát, không phù hợp với ngữ cảnh giáo dục tiểu học, vốn yêu cầu ngôn ngữ đơn giản và nội dung phù hợp với trẻ em. Để khắc phục hạn chế này, nghiên cứu hiện tại tinh chỉnh PhoBERT và mT5 trên bộ dữ liệu chuyên biệt gồm 6,000 bài đọc từ sách giáo khoa Tiếng Việt và truyện thiếu nhi, được phân loại theo cấp lớp 1 - 5, nhằm tạo ra các bản tóm tắt chính xác và phù hợp với trình độ nhận thức của học sinh tiểu học.

2.3. Đánh giá tóm tắt và khoảng trống nghiên cứu

Chất lượng tóm tắt tự động được đánh giá thông qua các chỉ số chuẩn như ROUGE, đo mức độ bao phủ nội dung so với văn bản gốc [10], BLEU, đánh giá độ chính xác ngữ liệu [11] và BERTScore, xác nhận tương đồng ngữ nghĩa [12]. Các nghiên cứu về tóm tắt trích xuất tiếng Việt cho thấy hiệu quả của các thuật toán như TextRank trong việc chọn lọc câu quan trọng [13]. Tuy nhiên, các chỉ số này, dù được sử dụng rộng rãi trong NLP, cần được điều chỉnh để phù hợp với văn bản giáo dục dành cho trẻ em, vốn ưu tiên ngôn ngữ đơn giản và ý nghĩa rõ ràng [9]. Khoshnevis và cộng sự [9] nhấn mạnh rằng bản tóm tắt cho học sinh lớp 1 - 2 cần ngắn gọn, dễ tiếp cận, trong khi lớp 4 - 5 yêu cầu chi tiết hơn để hỗ trợ phân tích nội dung. Các nghiên cứu cũng khẳng định rằng tóm tắt tự động có thể hỗ trợ giáo viên hiệu quả, nhưng phải được thiết kế theo nhu cầu cụ thể của từng cấp lớp [6 - 9]. Nghiên cứu này áp dụng các chỉ số trên, kết hợp với khảo sát người dùng, để đảm bảo bản tóm tắt chính xác và phù hợp với học sinh lớp 1 - 5.

Khoảng trống nghiên cứu trong lĩnh vực tóm tắt tiếng Việt dành cho học sinh tiểu học tạo động lực cho sự đổi mới của nghiên cứu này. Các công trình

hiện tại [3, 4, 7] thường tập trung vào tóm tắt tin tức hoặc các tác vụ tổng quát bằng tiếng Anh, chưa chú trọng đến đặc thù ngôn ngữ học và nhu cầu sư phạm của trẻ em Việt Nam. Ví dụ, các nghiên cứu về tóm tắt diễn giải tiếng Việt chủ yếu sử dụng dữ liệu người lớn, không phù hợp với trẻ em [3]. Các bộ dữ liệu tiếng Việt hiện có phần lớn phục vụ các tác vụ như tóm tắt tin tức, thiếu sự phân loại theo độ tuổi hoặc mục tiêu giáo dục [3, 4]. Nghiên cứu này nhằm xây dựng ứng dụng AI hỗ trợ tóm tắt bài đọc tiếng Việt cho học sinh tiểu học. Nghiên cứu có 3 đóng góp chính:

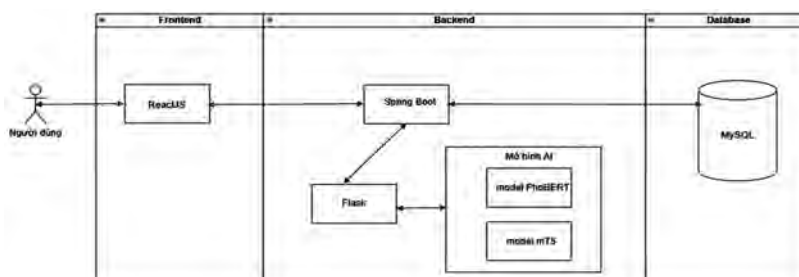
1. Xây dựng bộ dữ liệu chuyên biệt gồm 6,000 bài đọc từ sách giáo khoa tiếng Việt và truyện thiếu nhi, được phân loại theo khối lớp 1 - 5.
2. Tinh chỉnh các mô hình PhoBERT, mT5 dựa trên bộ dữ liệu đã xây dựng để tạo ra các bản tóm tắt phù hợp với cấp độ đọc hiểu.

3. Xây dựng ứng dụng web hỗ trợ học tập và cung cấp công cụ giảng dạy hữu ích cho giáo viên, phụ huynh.

3. THIẾT KẾ HỆ THỐNG VÀ PHƯƠNG PHÁP

3.1. Kiến trúc hệ thống

Hệ thống tóm tắt văn bản tự động dành cho học sinh tiểu học được thiết kế theo mô hình ba lớp: Lớp dữ liệu và tiền xử lý, lớp mô hình học sâu và lớp ứng dụng web. Kiến trúc này cho phép tích hợp linh hoạt các mô hình tóm tắt trích xuất và diễn giải, đồng thời đảm bảo khả năng mở rộng và bảo trì. Sơ đồ tổng thể của hệ thống được trình bày ở Hình 1, trong đó dữ liệu đầu vào được xử lý và phân loại trước khi đưa vào các mô hình học sâu PhoBERT và mT5. Kết quả tóm tắt sau đó được trả về giao diện web, hỗ trợ người dùng là học sinh, giáo viên và phụ huynh.



Hình 1. Kiến trúc tổng thể hệ thống tóm tắt văn bản cho học sinh tiểu học

Bộ dữ liệu phục vụ cho hệ thống được xây dựng với quy mô 6,000 bài đọc tiếng Việt, thu thập từ hai nguồn chính: Sách giáo khoa Tiếng Việt và tuyển tập truyện thiếu nhi của Nhà xuất bản Kim Đồng. Để đảm bảo tính phù hợp với trình độ nhận thức và nhu cầu học tập của học sinh tiểu học, mỗi bài đọc đều được gán nhãn tóm tắt mẫu bởi đội ngũ giáo viên tiểu học cùng sinh viên ngành Sư phạm. Việc này không chỉ nâng cao chất lượng dữ liệu mà còn giúp hệ thống tạo ra các bản tóm tắt sát thực với thực tiễn giáo dục.

Quy trình tiền xử lý dữ liệu được thực hiện bài bản, bao gồm các bước: Chuẩn hóa văn bản, loại bỏ ký tự đặc biệt, tách câu, phân loại theo từng cấp lớp

và kiểm tra chất lượng nhãn tóm tắt. Đặc biệt, việc phân loại dữ liệu theo lớp học giúp mô hình học sâu có khả năng sinh tóm tắt phù hợp với từng độ tuổi khác nhau, đáp ứng tốt hơn nhu cầu của học sinh các khối lớp.

Sau khi hoàn thiện, bộ dữ liệu được chia thành ba tập: Huấn luyện (80%), kiểm tra (10%) và kiểm thử (10%), đảm bảo sự cân bằng giữa các lớp và nguồn dữ liệu. Cách phân chia này giúp tối ưu hóa quá trình huấn luyện, đồng thời tăng tính khách quan khi đánh giá hiệu quả của hệ thống trên các tập dữ liệu chưa từng thấy. Thống kê chi tiết về bộ dữ liệu được trình bày tại Bảng 1.

Bảng 1. Thống kê bộ dữ liệu tóm tắt tiếng Việt theo nguồn và đặc điểm văn bản

STT	Nguồn dữ liệu	Loại văn bản	Số lượng bài đọc	Đặc điểm	Độ dài trung bình
1	Sách giáo khoa Tiếng Việt lớp 1 - 5 bộ Cánh diều (NXB Đại học Sư phạm TP.HCM)	Truyện ngắn, bài đọc thông tin	143	Ngôn ngữ hiện đại, tổ chức theo chủ điểm	200 - 450 từ
2	Sách giáo khoa Tiếng Việt lớp 1 - 5 bộ Kết nối tri thức với cuộc sống (NXB Giáo dục Việt Nam)	Truyện ngắn, văn bản thông tin	152	Tích hợp các kỹ năng ngôn ngữ	250 - 500 từ
3	Sách giáo khoa Tiếng Việt lớp 1 - 5 bộ Chân trời sáng tạo (NXB Giáo dục Việt Nam)	Truyện ngắn, văn bản thông tin	138	Phát triển tư duy sáng tạo	220 - 480 từ

STT	Nguồn dữ liệu	Loại văn bản	Số lượng bài đọc	Đặc điểm	Độ dài trung bình
4	Tuyển tập truyện cổ tích Việt Nam (NXB Kim Đồng)	Truyện cổ tích dân gian	1,500	Cốt truyện đơn giản, giáo dục	600 - 1,200 từ
5	Tuyển tập truyện ngụ ngôn Việt Nam và thế giới (NXB Thanh niên)	Truyện ngụ ngôn	1,500	Ngắn gọn, chứa bài học đạo đức	200 - 400 từ
6	Truyện thiếu nhi hiện đại (NXB Trẻ, NXB Kim Đồng)	Truyện đương đại	1,500	Đa dạng chủ đề	500 - 800 từ
7	Tài liệu trên trang sachgiaoduc.gov.vn (Bộ GD&ĐT)	Truyện ngắn	1,067		400 - 700 từ
Tổng cộng			6,000		200 - 1,200 từ

3.2. Mô hình và quy trình huấn luyện

Hệ thống tóm tắt văn bản được thiết kế nhằm đáp ứng yêu cầu năng lực đọc, hiểu và độ khó cho từng lớp học sinh tiểu học, dựa trên khung Chương trình tiếng Việt 2018 (Thông tư 32/2018/TT-BGDĐT). Cụ thể, đầu ra tóm tắt cho từng nhóm lớp được kiểm soát như sau:

- Lớp 1 - 2: Bản tóm tắt có độ dài 30 - 60 từ, ưu tiên câu ngắn, từ vựng đơn giản.
- Lớp 3 - 4: Bản tóm tắt có độ dài 70 - 120 từ, sử dụng từ vựng đa dạng, câu có liên kết.
- Lớp 5: Bản tóm tắt có độ dài 100 - 150 từ, cho phép dùng câu ghép, văn bản mạch lạc.

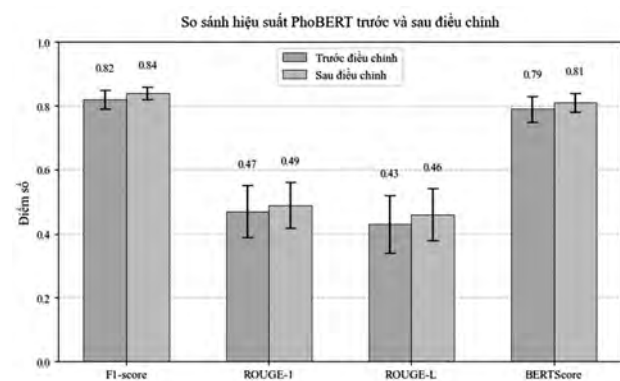
Để tối ưu hóa siêu tham số cho cả hai mô hình PhoBERT và mT5 khi tinh chỉnh, nhóm nghiên cứu đã thử nghiệm nhiều phương pháp, bao gồm tìm kiếm lưới (grid search), tìm kiếm ngẫu nhiên (random search) và Bayesian. Sau khi đánh giá, Optuna được chọn là phương pháp hiệu quả nhất nhờ khả năng cân bằng giữa tốc độ tính toán và độ chính xác trong việc tìm kiếm tổ hợp siêu tham số tối ưu.

Tinh chỉnh PhoBERT: PhoBERT được tinh chỉnh trên 48.000 mẫu câu (6,000 bài) để phân loại câu quan trọng (nhãn 0/1) cho tóm tắt trích xuất. Dữ liệu được chuẩn hóa (128 token), chia 80% huấn luyện, 10% kiểm định, 10% kiểm tra. Siêu tham số được tối ưu bằng Optuna, tổ hợp tốt nhất: Learning rate $3e-5$, batch size 16, 4 epochs, AdamW (weight decay 0.01), đạt F1-score 0.89 trên tập kiểm định. Trên 6.000 mẫu kiểm tra, PhoBERT đạt F1-score $0.82 (\pm 0.03)$, ROUGE-1 $0.47 (\pm 0.08)$, ROUGE-L $0.43 (\pm 0.09)$, BERTScore $0.79 (\pm 0.04)$.

Tinh chỉnh mT5: mT5-small được tinh chỉnh trên 4.800 bài (cặp input-output, 512/128 token) để tóm tắt diễn giải, chia 80% huấn luyện, 10% kiểm định, 10% kiểm tra. Siêu tham số tối ưu bằng Optuna: learning rate $3e-5$, batch size 8, 20 epochs,

AdamW (weight decay 0.01), đạt ROUGE-1 0.47 trên tập kiểm định. Trên 6,000 bài kiểm tra, mT5 đạt ROUGE-1 $0.47 (\pm 0.06)$, ROUGE-2 $0.27 (\pm 0.05)$, ROUGE-L $0.44 (\pm 0.07)$, BLEU $0.40 (\pm 0.04)$.

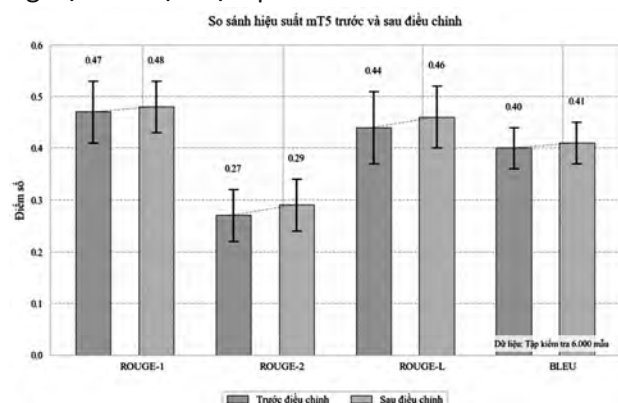
Tóm tắt trích xuất với LSRS-IGO (lớp 4 - 5): Đối với học sinh lớp 4 và 5, hệ thống áp dụng phương pháp tóm tắt trích xuất dựa trên mô hình PhoBERT, kết hợp với Lexical-Semantic Relevance Scoring (LSRS) và Iterative Greedy Optimization (IGO). PhoBERT được tinh chỉnh trên bộ dữ liệu bài đọc đã kiểm định để biểu diễn ngữ nghĩa sâu sắc cho từng câu. Sau đó, mỗi câu được đánh giá dựa trên điểm LSRS, tích hợp cả đặc trưng ngữ nghĩa (từ PhoBERT) và đặc trưng lexical (vị trí, độ dài, tần suất từ khóa, TF-IDF). Thuật toán IGO tối ưu hóa việc chọn tập hợp câu có tổng điểm cao nhất, đảm bảo nội dung bao quát ý chính và phù hợp với độ dài, mức độ khó theo từng lớp. Quá trình huấn luyện áp dụng học có giám sát, sử dụng nhãn do giáo viên và sinh viên sư phạm cung cấp làm mục tiêu tham chiếu.



Hình . So sánh hiệu suất PhoBERT trước và sau điều chỉnh với LSRS-IGO

Tóm tắt diễn giải với mT5 và prompt (lớp 1 - 5): Đối với tất cả các lớp, đặc biệt là lớp 1 - 3, hệ thống sử dụng phương pháp tóm tắt diễn giải dựa trên mô hình mT5. Mô hình được tinh chỉnh trên tập dữ liệu mẫu, áp dụng kỹ thuật prompt để mã hóa thông tin

về lớp học, độ dài mong muốn và đặc điểm ngôn ngữ vào đầu vào. Nhờ vậy, mT5 tạo ra các bản tóm lược: Ngắn gọn, dễ hiểu cho lớp 1 - 2; đa dạng, có liên kết cho lớp 3 - 4 và đầy đủ, mạch lạc cho lớp 5. Trong giai đoạn sinh nội dung, thuật toán beam search được triển khai để chọn chuỗi từ có xác suất cao nhất, tối ưu hóa chất lượng và tính tự nhiên. Các siêu tham số như beam width, temperature và độ dài tối đa được điều chỉnh thông qua thực nghiệm để đạt hiệu quả cao nhất.



Hình . So sánh hiệu suất mT5 trước và sau điều chỉnh với prompt

Quy trình huấn luyện: Quy trình huấn luyện được thực hiện tuần tự: (1) tiền xử lý dữ liệu; (2) chuẩn bị tập huấn luyện, (3) tinh chỉnh PhoBERT và mT5 trên GPU, (4) đánh giá hiệu suất trên tập kiểm thử bằng các chỉ số ROUGE, BLEU, BERTScore. Kết quả mô hình sau huấn luyện được đóng gói và tích hợp vào hệ thống backend, sẵn sàng phục vụ ứng dụng web. Việc kiểm soát đầu ra tóm tắt theo từng lớp học, bám sát Thông tư 32/2018/TT-BGDĐT, kết hợp các kỹ thuật LSRS-IGO, prompt, beam search và đặc trưng lexical-semantic, giúp hệ thống tạo ra các bản tóm tắt đáp ứng đúng yêu cầu năng lực đọc hiểu và độ khó của từng khối lớp trong chương trình giáo dục tiểu học.

Bảng 3. Danh sách câu hỏi khảo sát nhu cầu tóm tắt văn bản

STT	Câu hỏi	Loại câu hỏi
1	Bạn thuộc nhóm nào? (Học sinh/ Phụ huynh/ Giáo viên)	Trắc nghiệm đơn
2	Nếu là học sinh, bạn học lớp mấy? (Lớp 1 - 5/ Không áp dụng)	Trắc nghiệm đơn
3	Bạn/ con bạn đọc truyện hoặc sách bao nhiêu lần mỗi tuần? (Không bao giờ/ 1 - 2 lần/ 3 - 5 lần/ Trên 5 lần)	Trắc nghiệm đơn
4	Bạn/ con bạn thích loại nội dung nào hơn? (Truyện cổ tích/ Bài đọc giáo khoa/ Cả hai)	Trắc nghiệm đa chọn
5	Bạn/con bạn có gặp khó khăn khi đọc truyện hoặc bài đọc dài không? (Có/ Không)	Trắc nghiệm đơn
6	Mức độ khó khăn khi hiểu nội dung là bao nhiêu? (1 - Rất dễ, 5 - Rất khó)	Thang điểm 1 - 5
7	Bạn thích cách nào hơn khi tóm tắt nội dung: Rút gọn bài viết gốc bằng cách giữ lại câu quan trọng, hay diễn đạt lại ý chính theo cách nói của riêng mình? (Rút ngắn câu/ Tóm gọn ý)	Trắc nghiệm đơn
8	Bạn muốn bản tóm tắt dài bao nhiêu? (1 - 2 câu/ 3 - 5 câu/ Trên 5 câu)	Trắc nghiệm đơn
9	Bạn có muốn bản tóm tắt kèm hình minh họa không? (Có/ Không)	Trắc nghiệm đơn
10	Bạn có ý kiến gì khác về hệ thống tóm tắt?	Câu hỏi mở

3.3. Cơ chế kiểm soát đầu ra tóm tắt theo khối lớp

Để đảm bảo các bản tóm tắt tự động thực sự phù hợp với năng lực đọc hiểu, khả năng tiếp nhận và đặc điểm tâm lý của học sinh tiểu học, hệ thống áp dụng cơ chế kiểm soát đầu ra dựa trên Chương trình tiếng Việt 2018 (Thông tư 32/2018/TT-BGDĐT). Cụ thể, mỗi bản tóm tắt sinh ra đều phải đáp ứng các tiêu chí về độ dài và đặc điểm ngôn ngữ theo từng khối lớp, như trình bày ở Bảng 2.

Tiêu chí này được tích hợp vào huấn luyện và sinh tóm tắt, thiết lập giới hạn độ dài, sử dụng prompt cùng ràng buộc lexical-semantic để đảm bảo ngôn ngữ phù hợp và tự động kiểm tra để loại bỏ bản tóm tắt không đạt yêu cầu. Cơ chế này đảm bảo hiệu quả kỹ thuật và giá trị sư phạm cho dạy - học tiếng Việt tiểu học.

Bảng 2. Tiêu chí kiểm soát đầu ra tóm tắt theo từng khối lớp tiểu học

Lớp	Độ dài (từ)	Đặc điểm ngôn ngữ
1 - 2	30 - 60	Câu ngắn, từ vựng đơn giản
3 - 4	70 - 120	Từ vựng đa dạng, có liên kết
5	100 - 150	Câu ghép, văn bản mạch lạc

3.4. Phương pháp đánh giá hệ thống

Chất lượng tóm tắt được đánh giá bằng các chỉ số tự động ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), BLEU và BERTScore, so sánh bản tóm tắt tự động với bản tóm tắt mẫu. Ngoài ra, khảo sát người dùng gồm học sinh, giáo viên và phụ huynh được thực hiện để đánh giá mức độ hài lòng, tính dễ hiểu và phù hợp của bản tóm tắt theo thang điểm Likert 5 mức. Các tiêu chí đánh giá bao gồm độ chính xác nội dung, tính ngắn gọn, dễ hiểu, phù hợp với từng cấp lớp và giao diện thân thiện. Bảng 3 trình bày mẫu phiếu khảo sát đánh giá hệ thống.

4. KẾT QUẢ NGHIÊN CỨU VÀ THỬ NGHIỆM

4.1. Môi trường và công cụ triển khai

Hệ thống được phát triển trên máy tính cá nhân trang bị GPU NVIDIA RTX 3060 (12GB VRAM), sử dụng hệ điều hành Windows 11. Quá trình huấn luyện mô hình sử dụng Python 3.8, thư viện PyTorch. Ứng dụng web được xây dựng với backend là Java Spring Boot, frontend và trực quan hóa kết quả bằng React. Dữ liệu và mô hình được lưu trữ trên ổ SSD tốc độ cao để đảm bảo hiệu năng xử lý.

4.2. Triển khai hệ thống

Quy trình triển khai hệ thống được tổ chức theo các bước sau: (1) thu thập và tiền xử lý dữ liệu từ sách giáo khoa và truyện thiếu nhi; (2) huấn luyện và tinh chỉnh mô hình PhoBERT (trích xuất) và mT5 (diễn giải) trên tập dữ liệu đã gán nhãn; (3) đóng gói mô hình và tích hợp vào backend phục vụ API, xây dựng giao diện web bằng ReactJS; (4) kiểm thử chức năng với người dùng thực tế; (5) đánh giá hiệu quả hệ thống bằng chỉ số ROUGE, BLEU, BERTScore kết hợp khảo sát người dùng; (6) tối ưu hóa và triển khai chính thức. Mỗi bước được thực hiện theo nguyên tắc nhất quán giữa kỹ thuật NLP và yêu cầu sư phạm, nhằm đảm bảo tính khả thi, dễ sử dụng và hiệu quả trong môi trường giáo dục tiểu học.

(1) Tiền xử lý dữ liệu: Dữ liệu văn bản được thu thập từ sách giáo khoa tiếng Việt và truyện thiếu

nhi, chuẩn hóa, loại bỏ ký tự đặc biệt, tách câu, gán nhãn tóm tắt mẫu và phân loại theo từng lớp học. Dữ liệu được chia thành ba tập: Huấn luyện (80%), kiểm tra (10%), kiểm thử (10%).

(2) Huấn luyện mô hình:

- Tóm tắt trích xuất (lớp 4 - 5): PhoBERT được tinh chỉnh trên tập dữ liệu bài đọc lớp 4 - 5, kết hợp phương pháp Lexical-Semantic Relevance Scoring và Iterative Greedy Optimization (LSRS-IGO) để lựa chọn các câu quan trọng, đảm bảo bản tóm tắt đáp ứng đúng độ dài, mức độ chi tiết và ngôn ngữ phù hợp với từng lớp.

- Tóm tắt diễn giải (lớp 1 - 5): mT5 được tinh chỉnh trên tập dữ liệu tóm tắt mẫu, sử dụng kỹ thuật prompt để mã hóa thông tin lớp học, độ dài mong muốn và đặc điểm ngôn ngữ vào đầu vào mô hình. Quá trình sinh bản tóm tắt sử dụng beam search để tối ưu hóa chất lượng đầu ra.

(1) Triển khai ứng dụng web: Các mô hình sau huấn luyện được đóng gói và tích hợp vào backend, giao tiếp với frontend qua API RESTful theo như kiến trúc hệ thống ở Hình 1. Hệ thống cho phép người dùng nhập văn bản, chọn lớp học, nhận bản tóm tắt và xem trực quan các chỉ số đánh giá.

(2) Kiểm thử chức năng: Hệ thống được kiểm thử trên các tình huống sử dụng thực tế, đối chiếu kết quả đầu ra với các tiêu chí về độ dài, đặc điểm ngôn ngữ, đảm bảo đáp ứng đúng yêu cầu cho từng khối lớp.



Hình 4. Giao diện chọn cấp lớp của ứng dụng web

Trang chủ ứng dụng web được thiết kế như ở Hình 4 có màu sắc sinh động và thao tác đơn giản, giúp

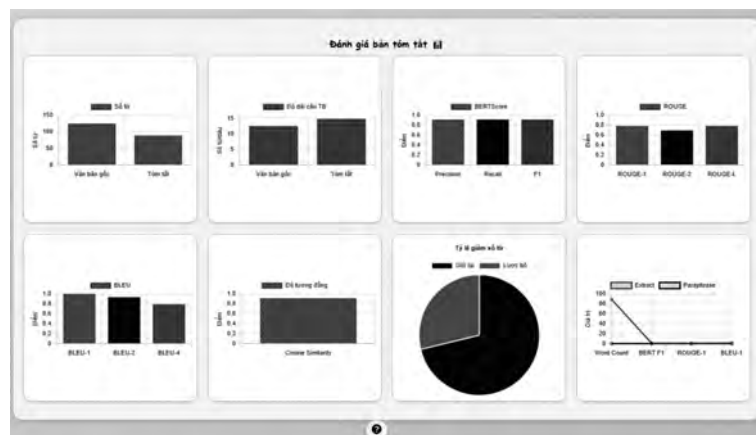
phụ huynh dễ dàng chọn lớp học phù hợp cho trẻ theo chương trình giáo dục.



Hình 5. Giao diện nhập và hiển thị bản tóm tắt của ứng dụng web

Giao diện Hình 5 cho phép nhập bài đọc và hiển thị tóm tắt diễn giải từ mT5, hỗ trợ học sinh tiểu học dễ dàng nắm ý chính và bài học đạo đức,

khơi gợi hứng thú học tập. Các tính năng như tải PDF, xem lịch sử và tạo hình minh họa cũng được tích hợp.



Hình 6. Giao diện trực quan hóa chỉ số đánh giá tóm tắt của ứng dụng web

Giao diện Hình 6 hiển thị các biểu đồ trực quan hóa, thể hiện chất lượng tóm tắt và độ chính xác của văn bản, giúp giáo viên, phụ huynh dễ dàng đánh giá hiệu quả. Thiết kế đơn giản, thân thiện, hỗ trợ theo dõi học tập của học sinh.

4.3. Đánh giá thực nghiệm

Đánh giá tự động bằng chỉ số chuẩn: Để đánh giá khách quan chất lượng các bản tóm tắt được tạo ra bởi hệ thống, nghiên cứu sử dụng ba chỉ số đánh giá phổ biến trong xử lý ngôn ngữ tự nhiên: ROUGE-

1 (đo mức độ bao phủ nội dung), BLEU (đánh giá độ chính xác ngữ liệu) và BERTScore (đo tương đồng ngữ nghĩa). Các chỉ số này được tính toán bằng cách so sánh bản tóm tắt tự động với bản tóm tắt mẫu do giáo viên hoặc sinh viên sư phạm biên soạn. Trong thực nghiệm, nhóm nghiên cứu lựa chọn 6 bài đọc đại diện cho các lớp 1 đến 5, bao gồm nhiều thể loại như truyện ngụ ngôn, truyện cổ tích, bài học đạo đức. Mỗi bài được tóm tắt bằng cả hai phương pháp: trích xuất (PhoBERT) và diễn giải (mT5), sau đó đánh giá bằng các chỉ số nêu trên như ở Bảng 3.

Bảng 4. Hiệu suất tóm tắt văn bản theo lớp học, bài đọc và phương pháp

Lớp học	Bài đọc	Phương pháp	ROUGE-1	BLEU	BERTScore
Lớp 1	Chú Mèo Con Đi Lạc	Diễn giải	0.40	0.35	0.78
Lớp 2	Cây Khế Nhà Bác	Diễn giải	0.42	0.36	0.80
Lớp 3	Chim Sâu và Châu Chấu	Diễn giải	0.47	0.40	0.82
Lớp 4	Cậu Bé Chăn Cừu	Trích xuất	0.52	0.45	0.85
Lớp 4	Cậu Bé Chăn Cừu	Diễn giải	0.45	0.38	0.81
Lớp 5	Người Thầy Cũ	Trích xuất	0.48	0.42	0.82
Lớp 5	Người Thầy Cũ	Diễn giải	0.43	0.37	0.80
	Trung bình - Trích xuất		0.50	0.44	0.84
	Trung bình - Diễn giải		0.43	0.37	0.80

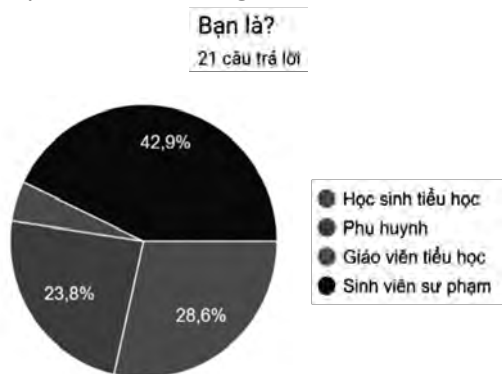
Kết quả cho thấy, phương pháp tóm tắt trích xuất đạt hiệu suất ổn định và giữ được nội dung gốc, đặc biệt hiệu quả ở lớp 4 - 5. Ngược lại, tóm tắt diễn giải có xu hướng ngắn gọn, phù hợp với lớp nhỏ, nhưng đôi khi còn hạn chế về mạch lạc hoặc mức độ đơn giản hóa ngôn ngữ, nhất là với các bài dài hoặc có yếu tố trừu tượng. Các chỉ số ROUGE-1, BLEU và BERTScore của mô hình đều vượt trội so với các phương pháp baseline truyền thống, khẳng định hiệu quả của hệ thống trong việc tạo ra các bản tóm tắt ngắn gọn, chính xác và phù hợp với trình độ học

sinh tiểu học. Kết quả này cũng phù hợp với các nghiên cứu trước đây về tóm tắt văn bản tiếng Việt. Để đánh giá hiệu quả hệ thống theo từng lớp học, phân tích sâu hơn cho thấy: Đối với lớp 1, bản tóm tắt diễn giải sử dụng từ vựng đơn giản (30 - 60 từ, chủ yếu từ cơ bản như “chạy”, “giúp”, “vui”), phù hợp với vốn từ hạn chế và khả năng tập trung ngắn của học sinh, đạt tỷ lệ hài lòng 85% từ giáo viên về tính dễ hiểu. Tuy nhiên, một số bản tóm tắt còn quá ngắn, chưa truyền tải đầy đủ ý nghĩa giáo dục. Lớp 2 - 3 được hưởng lợi từ bản tóm tắt diễn giải (70 - 120 từ),

với từ vựng đa dạng hơn (bao gồm từ ghép như “hạnh phúc”, “chia sẻ”) và câu có liên kết, đạt tỷ lệ đồng thuận 80% từ phụ huynh về khả năng khơi gợi hứng thú học tập. Lớp 4 - 5, với bản tóm tắt trích xuất (100 - 150 từ), sử dụng câu ghép và từ vựng phức tạp hơn (như “hậu quả”, “trách nhiệm”), đáp ứng tốt nhu cầu phân tích nội dung, nhưng cần nhấn mạnh bài học đạo đức rõ hơn, như phản hồi từ giáo viên (75% đồng thuận). Phân tích này khẳng định hệ thống đáp ứng tốt tiêu chí ngôn ngữ và độ dài theo Chương trình tiếng Việt 2018, nhưng cần cải thiện tính mạch lạc cho lớp 1 - 2 và độ sâu nội dung cho lớp 5.

Đánh giá thực tế với người dùng: Để kiểm chứng khả năng ứng dụng thực tiễn của hệ thống tóm tắt, nhóm nghiên cứu đã tiến hành khảo sát trải nghiệm người dùng cuối, bao gồm học sinh tiểu học, giáo viên, phụ huynh và sinh viên sư phạm. Quá trình đánh giá tập trung vào các tiêu chí: Mức độ dễ hiểu, tính hấp dẫn, khả năng hỗ trợ đọc hiểu và giá trị giáo dục của các bản tóm tắt.

Phân bố đối tượng khảo sát: Tổng cộng 21 người dùng tham gia khảo sát như ở Hình 7, trong đó sinh viên sư phạm chiếm 42.9%, học sinh tiểu học 28.6%, phụ huynh và giáo viên tiểu học 23.8%. Tỷ lệ này bảo đảm cân bằng giữa góc nhìn sư phạm, người học trực tiếp và nhóm hỗ trợ giáo dục.



Hình 7. Phân bố ý kiến khảo sát về hiệu quả của hệ thống tóm tắt

Bảng 5. Kết quả khảo sát người dùng

Nội dung khảo sát	Tỷ lệ đồng thuận cao (4 - 5/5)
Bản tóm tắt trích xuất dễ hiểu, phù hợp	75%
Bản tóm tắt trích xuất có khơi gợi hứng thú đọc lại bài gốc	60%
Bản tóm tắt diễn giải vui, dễ hiểu và phù hợp với học sinh	85%
Bản tóm tắt diễn giải giúp học sinh muốn đọc bài gốc hơn	80%
Website có giá trị và hữu ích cho học sinh tiểu học	90%

Kết quả khảo sát: Các đối tượng được yêu cầu trải nghiệm hệ thống, đọc các bản tóm tắt trích xuất và diễn giải, sau đó đánh giá trên thang điểm 5 về các tiêu chí đã nêu. Kết quả tổng hợp được trình bày trong Bảng 4.

Phản hồi chuyên môn được thu thập từ ba giáo viên tại Trường Tiểu học Tân Thạnh 1, xã Tân Thạnh, huyện Thanh Bình, tỉnh Đồng Tháp, nhằm đánh giá hiệu quả của hệ thống tóm tắt trên các bài đọc tiêu biểu như “*Cây Khế Nhà Bác*” và “*Cậu Bé Chăn Cừu*”.

- Cô Mộng Thu (hơn 10 năm kinh nghiệm): “*Tóm tắt bài Cây Khế Nhà Bác sử dụng từ ngữ đơn giản, phù hợp với học sinh lớp 2. Tuy nhiên, một số câu được đơn giản hóa quá mức, khiến mạch truyện chưa liền lạc. Nếu có thể bổ sung thêm 1 - 2 chi tiết để kết bài rõ hơn, học sinh sẽ dễ hiểu hơn.*”
- Thầy Lê Thanh Hải (15 năm kinh nghiệm): “*Nhìn chung phần tóm tắt của bài Cậu Bé Chăn Cừu khá dễ hiểu, nhưng ở cả bản trích xuất và diễn giải vẫn chưa làm rõ hậu quả từ hành động thiếu trung thực. Với học sinh tiểu học, việc hiểu rõ bài học đạo đức là rất quan trọng, nên phần này nên được làm nổi bật hơn.*”
- Thầy Nông Đức Khanh (hơn 20 năm kinh nghiệm): “*Thầy Khanh nhận xét hệ thống có sự phù hợp với từng cấp lớp, nhưng ở lớp 1 thì độ dài còn quá ngắn, gây khó khăn trong việc truyền đạt trọn vẹn nội dung. Ở lớp lớn, như lớp 5, bản tóm tắt cần thể hiện rõ hơn bài học đạo đức, như trong Cậu Bé Chăn Cừu là vấn đề nói dối và hậu quả. Đây là cơ sở để hệ thống điều chỉnh linh hoạt hơn.*”

Các giáo viên đánh giá hệ thống là một công cụ hỗ trợ ưu việt, giúp nâng cao hiệu quả giảng dạy và học tập cho cả giáo viên và học sinh tiểu học, đặc biệt nhờ khả năng tóm tắt tự động và phù hợp với từng cấp lớp. Tuy nhiên, họ đề xuất bổ sung cơ chế điều chỉnh linh hoạt theo cấp lớp, đặc điểm bài học và mục tiêu giáo dục để tối ưu hóa tiềm năng của hệ thống.

4.4. Bàn luận kết quả

Hệ thống tóm tắt văn bản tiếng Việt chứng minh hiệu quả cao, với phương pháp trích xuất dựa trên PhoBERT và LSRS-IGO giữ nguyên ý chính, đặc biệt phù hợp với lớp 4 - 5, trong khi phương pháp diễn giải bằng mT5 hỗ trợ tốt lớp 1 - 5 nhờ tính ngắn gọn, dù còn hạn chế về liên mạch ở các bài trừu tượng. Phản hồi từ người dùng và giáo viên khẳng định giá trị giáo dục, nhưng một số bản tóm tắt lớp

1 - 2 quá ngắn, thiếu thông tin và lớp 5 cần làm rõ bài học đạo đức. Để làm rõ sự khác biệt giữa hai phương pháp, phân tích so sánh cho thấy tóm tắt trích xuất (PhoBERT) đạt độ chính xác cao hơn về nội dung gốc (ROUGE-1: 0.50, BERTScore: 0.84), phù hợp với học sinh lớp 4 - 5 cần phân tích chi tiết, nhưng thiếu linh hoạt trong việc đơn giản hóa ngôn ngữ cho lớp nhỏ. Ngược lại, tóm tắt diễn giải (mT5) tạo ra các bản tóm tắt ngắn gọn, dễ hiểu (ROUGE-1: 0.43, BERTScore: 0.80), lý tưởng cho lớp 1 - 3, nhưng đôi khi bỏ sót chi tiết quan trọng hoặc thiếu mạch lạc. Ví dụ, với bài Cậu Bé Chăn Cừu (lớp 4): Bản tóm tắt trích xuất giữ nguyên câu như "Cậu bé chăn cừu nói dối khiến dân làng mất lòng tin, cuối cùng phải chịu hậu quả khi sói thật tấn công" (50 từ, chi tiết, phù hợp lớp 4 - 5); bản tóm tắt diễn giải là "Cậu bé nói dối về sói, nên khi sói đến thật, không ai giúp cậu" (20 từ, ngắn gọn, phù hợp lớp 1 - 3, nhưng thiếu nhấn mạnh hậu quả đạo đức). Phân tích này cho thấy trích xuất ưu tiên độ bao quát, trong khi diễn giải tập trung vào tính dễ tiếp cận, cần cải thiện để cân bằng giữa chi tiết và đơn giản hóa. Điều này gợi ý cần tối ưu hóa mT5 và tích hợp tham số điều chỉnh độ sâu nội dung. So với nghiên cứu trước [10], hệ thống vượt trội nhờ dữ liệu gán nhãn chuyên biệt, mở ra tiềm năng cho các công cụ học tập thông minh tại Việt Nam.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nghiên cứu đã thành công trong việc phát triển một hệ thống ứng dụng trí tuệ nhân tạo hỗ trợ tóm tắt truyện và bài đọc tiếng Việt cho học sinh tiểu học, tích hợp phương pháp trích xuất bằng mô hình PhoBERT và phương pháp diễn giải bằng mô hình mT5. Nghiên cứu đã đóng góp vào việc tinh chỉnh các mô hình này và tạo ra các bản tóm tắt phù hợp với từng khối lớp theo Chương trình tiếng Việt, đạt hiệu suất cao trên các chỉ số ROUGE, BLEU và BERTScore, đặc biệt nổi bật ở nhóm lớp 4 - 5 với phương pháp trích xuất. Khảo sát từ giáo viên, học sinh, phụ huynh và sinh viên

sư phạm xác nhận rằng hệ thống mang lại các bản tóm tắt dễ hiểu, hấp dẫn, giúp nâng cao hiệu quả đọc hiểu và giảm tải cho giáo viên trong giảng dạy. Hệ thống này không chỉ hỗ trợ học tập mà còn định hình hướng phát triển công cụ học tập thông minh tại Việt Nam.

Dẫu vậy, hệ thống còn hạn chế như bộ dữ liệu chưa đa dạng (chủ yếu sách giáo khoa và truyện thiếu nhi), thời gian xử lý mT5 với văn bản dài còn chậm, một số tóm tắt diễn giải thiếu mạch lạc và giao diện cùng tính năng minh họa, đọc to chưa đủ linh hoạt. Trong tương lai, nhóm nghiên cứu sẽ triển khai thí điểm tất cả các lớp ở Trường Tiểu học Tân Thạnh 1, xã Tân Thạnh, huyện Thanh Bình, tỉnh Đồng Tháp. Giai đoạn thí điểm nhắm đến tất cả các học sinh lớp 1 - 5, với sự tham gia của giáo viên và phụ huynh trong năm học 2025 - 2026. Ứng dụng web sẽ được tích hợp vào các tiết học Tiếng Việt và hoạt động ngoại khóa. Với phản hồi được thu thập, nhóm sẽ đánh giá hiệu quả của hệ thống, nhằm kiểm chứng tính thực tiễn và tối ưu hóa hệ thống trước khi mở rộng quy mô. Ngoài ra, nhóm sẽ mở rộng bộ dữ liệu với các tài liệu như bài học đạo đức, truyện tranh; tối ưu hóa mT5 để giảm thời gian xử lý và nâng cao chất lượng; tích hợp chức năng đọc to, minh họa trực quan và cá nhân hóa nội dung theo nhu cầu học sinh. Nhóm cũng sẽ phát triển API hỗ trợ tích hợp với nền tảng học tập trực tuyến, xây dựng bộ chỉ số đánh giá chuyên biệt kết hợp tự động và phản hồi người dùng, đồng thời triển khai cơ chế kiểm soát chất lượng và đạo đức để đảm bảo tính giáo dục và tâm lý phù hợp. Những cải tiến này sẽ giúp hệ thống trở thành công cụ học tập hiệu quả, bền vững, hỗ trợ lâu dài cho giáo dục tiểu học tại Việt Nam.

LỜI CẢM ƠN

Nghiên cứu này được Trường Đại học Quốc tế Hồng Bàng cấp kinh phí thực hiện dưới mã số đề tài GVTC18.02.

TÀI LIỆU THAM KHẢO

[1] V. Jagtap, P. Parlewar, S. Dhande, A. Langhe, H. Choudhary và A. Mishra, "Advancements in Text Summarization Through Machine Learning: A Comprehensive Survey and Analysis," *Journal of Electrical Systems*, tập 1, số 20, pp. 833-849, 2024.

[2] M. M. Saiyyad và N. N. Patil, "Text Summarization Using Deep Learning Techniques: A Review," *Engineering Proceedings*, p. 194, 2024.

[3] Q.-A. Nguyen, D.-C. Can, H.-Q. Le và M.-V. Tran, "VLSP 2022 Abmusu Task Dataset: A resource for Vietnamese abstractive multi-document summarization," *International Journal of Asian Language Processing*, p. 1-18, 2023.

[4] D. Q. Nguyen, T. T. Nguyen và P. M. Hiếu, "PhoBERT: Pre-trained language models for Vietnamese," *Findings of ACL: EMNLP 2020*, p.

1037-1042, 2020.

[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua và C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," trong *Proceedings of NAACL-HLT 2021*, 2021.

[6] R. Dorgham, "The impact of using summarizing strategy on secondary student's reading comprehension skills," *Journal of Faculty of Education*, pp. (138), April, Part (3), 2024.

[7] T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond và L. Galligan, "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," p. 56720-56739, 2022.

[8] L. Q. Tường, P. T. Phi và Đ. Đ. Hào, "Tóm tắt văn bản tiếng Việt tự động với mô hình Sequence-to-Sequence," *Tạp chí Khoa học Đại học Cần Thơ*, tập 5, số 1, p. 125-132, 2017.

[9] I. Khoshnevis và S. Parvinnejad, "The effect of text

summarization as a cognitive strategy on the achievement of male and female language learners' reading comprehension," *International Journal of Learning & Development*, p. 57-69, 2015.

[10] S. Kumar và A. Solanki, "ROUGE-SS: A New ROUGE Variant for Evaluation of Text Summarization," p. 1-15, 2023.

[11] K. Papineni, S. Roukos, T. Ward và W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," trong *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger và Y. Artzi, "BERTScore: Evaluating text generation with BERT," trong *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[13] M. Xia, E. Kochmar và T. Briscoe, "Automatic learner summary assessment for reading comprehension," trong *Proceedings of NAACL-HLT 2019*, 2019.

Developing an AI application to summarize Vietnamese reading materials for primary school students

Hoang Ngoc Long, Huynh Thanh Phu

ABSTRACT

The rapid development of artificial intelligence (AI) and natural language processing (NLP) has opened up opportunities for applications in education, particularly in supporting reading comprehension for primary school students. However, existing text summarization solutions tailored specifically for Vietnamese primary learners remain limited in both quantity and effectiveness. This study proposes an AI-based system for summarizing Vietnamese reading materials for students in grades 1 through 5, aiming to help them access learning content more effectively. The system integrates two advanced language models: PhoBERT (for extractive summarization) and mT5 (for abstractive summarization), both fine-tuned on a dataset of 6,000 reading passages collected from textbooks and children's literature. Each summary is controlled in terms of length, vocabulary, and linguistic features to match the appropriate grade level according to the national primary education curriculum. Evaluation results indicate that the system performs well on ROUGE, BLEU, and BERTScore metrics, and receives positive feedback from users including students, teachers, and parents. This research not only contributes to the application of AI in education but also opens up new directions for the development of learning tools that support the Vietnamese language.

Keywords: artificial intelligence (AI), natural language processing (NLP), text summarization, primary education

Received: 13/5/2025

Revised: 09/6/2025

Accepted for publication: 12/6/2025