**KHẢO SÁT VÀ PHÂN TÍCH ĐÁNH GIÁ TÓM TẮT Ý KIẾN CỦA NGƯỜI VIỆT NAM VỀ MẠNG XÃ HỘI SỬ DỤNG MÔ HÌNH HỌC SÂU**

**MỤC LỤC**

[MỞ ĐẦU 2](#_Toc196486822)

[DANH MỤC HÌNH ẢNH 3](#_Toc196486823)

[DANH MỤC BẢNG BIỂU 3](#_Toc196486824)

[DANH MỤC TỪ VIẾT TẮT 5](#_Toc196486825)

[CHƯƠNG 1: GIỚI THIỆU 6](#_Toc196486826)

[1.1 Bối cảnh thực tế của bài toán nghiên cứu 6](#_Toc196486827)

[Hình 1: Kiến trúc transformer 7](#_Toc196486828)

[CHƯƠNG 2: TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU 9](#_Toc196486829)

[CHƯƠNG 3: ĐỀ XUẤT GIẢI PHÁP 10](#_Toc196486830)

[3.1.BERT 10](#_Toc196486831)

[3.2 PEGASUS 11](#_Toc196486832)

[CHƯƠNG 4: ĐÁNH GIÁ GIẢI PHÁP ĐỀ XUẤT 15](#_Toc196486833)

[KẾT LUẬN VÀ KIẾN NGHỊ 16](#_Toc196486834)

[TÀI LIỆU THAM KHẢO 17](#_Toc196486835)

[MINH CHỨNG BÀI BÁO 17](#_Toc196486836)

# **MỞ ĐẦU**

1-2 trang

# **DANH MỤC HÌNH ẢNH**

# **DANH MỤC BẢNG BIỂU**

# **DANH MỤC TỪ VIẾT TẮT**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Từ viết tắt** | **Thuật ngữ tiếng Anh** | **Ý nghĩa** |
| 1 |  |  |  |
| 2 |  |  |  |

# **CHƯƠNG 1: GIỚI THIỆU VỀ BÀI TOÁN NGHIÊN CỨU**

**1.1 Bối cảnh thực tế của bài toán nghiên cứu**

Trong kỷ nguyên thương mại kỹ thuật số, các bài đánh giá do người dùng tạo trên phương tiện truyền thông xã hội và nền tảng thương mại điện tử đã nổi lên như một nguồn thông tin chi tiết về người tiêu dùng và phản hồi về sản phẩm thiết yếu. Những bài đánh giá này ảnh hưởng đến quyết định mua hàng, nhận thức về thương hiệu và phát triển sản phẩm trong nhiều ngành công nghiệp khác nhau.

Tuy nhiên, khối lượng lớn, tính biến đổi và bản chất thường dư thừa của nội dung như vậy đặt ra một thách thức đáng kể cho cả người dùng và doanh nghiệp đang cố gắng trích xuất thông tin có thể hành động. Tóm tắt văn bản tự động, đặc biệt là thông qua các phương pháp dựa trên học sâu (DL), ngày càng được áp dụng để tạo ra các bản tóm tắt ngắn gọn và nhiều thông tin về các bài đánh giá của người dùng. Mặc dù có những tiến bộ đáng kể trong việc tóm tắt tự động trong các ngôn ngữ giàu tài nguyên như tiếng Anh, tiếng Trung và tiếng Nhật, được hỗ trợ bởi các tập dữ liệu chuẩn lớn và các mô hình ngôn ngữ được đào tạo trước mạnh mẽ, nhưng tiến trình trong tiếng Việt vẫn khiêm tốn. Các bài đánh giá bằng tiếng Việt—đặc biệt là các bài đánh giá liên quan đến điện thoại thông minh trên các nền tảng thương mại điện tử địa phương—thường ngắn gọn, không có cấu trúc cao và được truyền tải bằng các biểu thức không chính thức, từ viết tắt và tiếng lóng chuyên ngành. Những đặc điểm về ngôn ngữ và ngữ cảnh này đặt ra những thách thức đáng kể cản trở việc áp dụng trực tiếp các kỹ thuật tóm tắt hiện có.

**1.2. Mục tiêu**

Khảo sát này điều tra các phương pháp tiếp cận dựa trên học sâu để tóm tắt các bài đánh giá sản phẩm của Việt Nam, tập trung cụ thể vào các bài đánh giá được thu thập từ phương tiện truyền thông xã hội và nền tảng thương mại điện tử. Chúng tôi tập trung vào các phương pháp tạo ra các bản tóm tắt vừa ngắn gọn vừa toàn diện, giảm thiểu sự trùng lặp trong khi vẫn nắm bắt được những hiểu biết cần thiết và đôi khi không thường xuyên. Bài báo nhấn mạnh các phương pháp tiếp cận có thể áp dụng cho các bối cảnh có ít tài nguyên và nhiễu, như minh họa bằng dữ liệu tiếng Việt.

Chúng tôi tiếp tục phân tích tính phù hợp của các kiến ​​trúc hiện đại, chẳng hạn như các mô hình dựa trên kiến trúc transformer, transfer learning và các phương pháp kết hợp, trong việc xử lý các nhu cầu riêng biệt của văn bản do người dùng tiếng Việt tạo ra. Ngoài ra, khảo sát này giới thiệu một nghiên cứu trường hợp thực tế sử dụng tập dữ liệu tự xây dựng về các bài đánh giá điện thoại thông minh, chứng minh những hạn chế trong thế giới thực của các mô hình hiện có và nhu cầu quan trọng đối với các giải pháp được thiết kế riêng hơn.

**1.3. Đóng góp**

Khảo sát này đóng góp vào sự phát triển của lĩnh vực tóm tắt văn bản tự động đối với các ngôn ngữ ít tài nguyên bằng cách cung cấp một cái nhìn tổng quan toàn diện về các phương pháp dựa trên học sâu, bao gồm cả hai hướng tiếp cận: trích xuất và trừu tượng. Bài viết đặc biệt nhấn mạnh vào bối cảnh tiếng Việt, nơi mà các đánh giá sản phẩm do người dùng tạo ra đặt ra những thách thức riêng biệt do cách sử dụng ngôn ngữ không chính thống, cấu trúc cú pháp bất thường và sự thiếu hụt dữ liệu huấn luyện đã được gán nhãn.

Thông qua việc đánh giá có hệ thống các kỹ thuật hiện đại, bao gồm kiến trúc dựa trên Transformer, các phương pháp điều hướng bằng prompt và mô hình lai (hybrid), bài viết xem xét khả năng thích ứng và các giới hạn hiệu suất của chúng khi được áp dụng cho dữ liệu tiếng Việt. Qua đó, bài viết cho thấy một khoảng cách đáng kể giữa năng lực của các hệ thống tóm tắt đa ngôn ngữ hiện tại và những yêu cầu đặc thù trong bối cảnh ít tài nguyên, tập trung vào miền nội dung cụ thể.

Ngoài ra, khảo sát cũng xác định một số hướng nghiên cứu tiềm năng trong tương lai, nhấn mạnh tầm quan trọng của các phương pháp dựa trên dữ liệu, chỉ số đánh giá nhạy theo ngữ cảnh, và các mô hình được thiết kế để thích ứng với các đặc điểm ngôn ngữ riêng của tiếng Việt. Tổng thể, những hiểu biết này nhằm định hướng việc phát triển các hệ thống tóm tắt mạnh mẽ hơn, có khả năng mở rộng và nhận thức ngôn ngữ tốt hơn dành cho tiếng Việt và các ngôn ngữ ít tài nguyên khác.

# **CHƯƠNG 2: TỔNG QUAN TÌNH HÌNH NGHIÊN CỨU**

**2.1. Bối cảnh và thuật ngữ**

Tóm tắt văn bản tự động là quá trình tạo ra một phiên bản ngắn gọn và mạch lạc từ một văn bản dài hơn, đồng thời vẫn giữ được những thông tin cốt lõi. Quá trình này được chia thành hai loại chính: **tóm tắt trích xuất** và **tóm tắt trừu tượng**. Phương pháp trích xuất xác định và chọn ra các cụm từ hoặc câu quan trọng trực tiếp từ văn bản gốc, thường dựa trên các tiêu chí thống kê hoặc được học để đánh giá mức độ quan trọng. Những phương pháp này thường dễ triển khai hơn và ít gặp lỗi sai về mặt nội dung, tuy nhiên chúng có thể tạo ra các bản tóm tắt rời rạc, thiếu mạch lạc. Ngược lại, các phương pháp trừu tượng tạo ra các câu mới có thể diễn giải lại hoặc viết lại nội dung gốc, gần giống cách con người thực hiện việc tóm tắt. Mặc dù cho phép tạo ra các bản tóm tắt mượt mà và súc tích hơn, phương pháp này đòi hỏi sự hiểu biết ngữ nghĩa sâu hơn và đặt ra nhiều thách thức về mô hình hóa, đặc biệt là đối với các ngôn ngữ ít tài nguyên.

Bên cạnh sự phân loại cốt lõi đó, bài toán tóm tắt còn được chia thành **tóm tắt một văn bản** hoặc **tóm tắt đa văn bản**. Tóm tắt một văn bản liên quan đến việc rút gọn thông tin từ một nguồn duy nhất, chẳng hạn như một bài đánh giá hoặc mô tả sản phẩm; trong khi đó, tóm tắt đa văn bản tổng hợp thông tin từ nhiều nguồn liên quan. Trong bối cảnh thương mại điện tử và mạng xã hội, tóm tắt đa văn bản đặc biệt có ý nghĩa, vì các đánh giá sản phẩm thường xuất hiện theo cụm dưới một mục sản phẩm và cần được tổng hợp để phản ánh quan điểm đa dạng từ người dùng.

Sự ra đời của học sâu (deep learning) đã thúc đẩy đáng kể khả năng của các hệ thống tóm tắt. Các mô hình học sâu giai đoạn đầu dựa vào mạng nơ-ron hồi quy (RNN) và các biến thể có cổng như LSTM và GRU để mã hóa chuỗi đầu vào và tạo ra tóm tắt theo trình tự. Những mô hình sequence-to-sequence (seq2seq) này cung cấp một hướng tiếp cận dựa trên dữ liệu, thay thế cho các phương pháp tóm tắt dựa trên luật hoặc thống kê. Tuy nhiên, chúng thường gặp khó khăn trong việc xử lý các phụ thuộc dài và ghi nhớ thông tin, đặc biệt trong môi trường đa ngôn ngữ.

Để khắc phục những hạn chế đó, cơ chế chú ý (attention) đã được giới thiệu, cho phép mô hình tập trung chọn lọc vào các phần khác nhau của đầu vào trong quá trình sinh văn bản. Sáng kiến này đã dẫn đến sự phát triển của các kiến trúc dựa trên Transformer, hoàn toàn thay thế sự hồi tiếp bằng các cơ chế tự chú ý. Các mô hình như BERT, GPT, T5 và BART đã xác lập tiêu chuẩn hiện đại trong cả hai hướng tiếp cận trích xuất và trừu tượng. Những mô hình này tận dụng quá trình tiền huấn luyện trên tập dữ liệu lớn và có thể được tinh chỉnh cho các nhiệm vụ cụ thể hoặc điều chỉnh thông qua chiến lược prompt.

Mặc dù các kiến trúc này đã đạt được kết quả ấn tượng đối với tiếng Anh và các ngôn ngữ có nhiều tài nguyên, việc ứng dụng chúng cho tiếng Việt vẫn còn hạn chế. Sự thiếu hụt dữ liệu tóm tắt được gán nhãn quy mô lớn và đặc điểm ngôn ngữ không chính thức, rời rạc của nội dung do người dùng tạo ra gây khó khăn cho cả việc huấn luyện mô hình và đánh giá hiệu quả. Như sẽ được trình bày ở các phần sau, việc điều chỉnh các mô hình nền tảng này cho tiếng Việt đòi hỏi phải giải quyết những khác biệt ngôn ngữ và ràng buộc về dữ liệu – những yếu tố vốn thường bị bỏ qua trong nghiên cứu chính thống.

**2.2. Những gì liên quan**

**2.2.1. Bức tranh nghiên cứu toàn cầu về Tóm tắt văn bản**

Lĩnh vực tóm tắt văn bản tự động đã chứng kiến sự phát triển nhanh chóng trong thập kỷ qua, phần lớn nhờ vào sự sẵn có của các bộ dữ liệu quy mô lớn và các mô hình tiền huấn luyện mạnh mẽ dành cho các ngôn ngữ giàu tài nguyên. Một số tập dữ liệu công khai đã trở thành các chuẩn đánh giá phổ biến cho việc đánh giá mô hình tóm tắt. Chẳng hạn, tập dữ liệu **CNN/DailyMail** gồm các bài báo dài kèm theo các bản tóm tắt nhiều câu, cho phép huấn luyện theo cả hướng trích xuất và trừu tượng. Tập **XSum** đặt ra thách thức cao hơn, yêu cầu mô hình tạo ra bản tóm tắt ngắn gọn chỉ trong một câu. Các tập dữ liệu khác như **Gigaword**, **Reddit TIFU**, và **Multi-News** góp phần đa dạng hóa mô hình khi bao gồm nội dung không chính thức, do người dùng tạo ra, hoặc thông tin từ nhiều văn bản khác nhau. Các bộ dữ liệu này có điểm chung là: định dạng tốt, cấu trúc nhất quán, và ngôn ngữ sạch – những điều hiếm khi đạt được ở các ngôn ngữ ít tài nguyên như tiếng Việt.

**2.2.2. Các chỉ số đánh giá**

Để đánh giá hiệu quả trên các bộ dữ liệu này, hầu hết các nghiên cứu đều sử dụng các chỉ số đánh giá tự động, đặc biệt là họ chỉ số **ROUGE**, đo mức độ trùng khớp giữa các n-gram của bản tóm tắt do hệ thống tạo ra và bản tóm tắt tham chiếu. **ROUGE-1** và **ROUGE-2** đánh giá độ chính xác và độ hồi tưởng theo unigram và bigram, trong khi **ROUGE-L** đo độ dài chuỗi con chung dài nhất. **BLEU**, vốn được đề xuất cho dịch máy, đôi khi cũng được sử dụng nhưng ít nhạy với các nội dung được diễn đạt lại. Gần đây, các chỉ số dựa trên ngữ nghĩa như **METEOR** và **BERTScore** ngày càng được quan tâm vì chúng đánh giá nội dung sâu sắc hơn mức độ trùng khớp bề mặt. Tuy vậy, tất cả các chỉ số này đều có hạn chế khi áp dụng vào các văn bản mang tính chủ quan, không chính thức, và chứa nhiều quan điểm cá nhân như đánh giá sản phẩm. Ngoài ra, chúng phụ thuộc vào ngôn ngữ và độ tin cậy của chúng có thể suy giảm trong tiếng Việt do các vấn đề như tách từ không nhất quán và sự trùng khớp từ vựng thấp giữa tóm tắt của người và máy.

**2.2.3. Xu hướng phương pháp luận**

Từ góc nhìn phương pháp luận, lĩnh vực này đã thay đổi mạnh mẽ cùng với sự phát triển của học sâu. Các nghiên cứu ban đầu sử dụng mô hình sequence-to-sequence (seq2seq) với cơ chế chú ý, nhưng những phương pháp hiện đại hơn đã tận dụng các mô hình dựa trên Transformer được tiền huấn luyện trên tập dữ liệu khổng lồ. Các mô hình như BART và T5 hiện nay thường được tinh chỉnh cho các tác vụ tóm tắt, đạt kết quả tiên tiến trong nhiều bộ đánh giá chuẩn. Trong bối cảnh đa ngôn ngữ, các mô hình như mBART và mT5 đã được đề xuất, mang lại khả năng chuyển giao xuyên ngôn ngữ, có thể hỗ trợ các ngôn ngữ ít tài nguyên thông qua chia sẻ tham số và học chuyển giao.

**2.2.4. Học dựa trên prompt và học đa tác vụ**

Song song đó, **học dựa trên prompt** nổi lên như một lựa chọn nhẹ nhưng hiệu quả so với việc tinh chỉnh mô hình. Thay vì thay đổi trọng số mô hình, phương pháp này định nghĩa nhiệm vụ dưới dạng các prompt ngôn ngữ tự nhiên đưa vào mô hình tiền huấn luyện như **GPT-3** hoặc **T0**. Chiến lược này giúp giảm sự phụ thuộc vào dữ liệu gán nhãn và đã chứng minh khả năng cạnh tranh trong các bài toán tóm tắt. Bên cạnh đó, **học đa tác vụ (multi-task learning)** cũng thu hút sự chú ý nhờ khả năng chia sẻ biểu diễn giữa các nhiệm vụ liên quan – chẳng hạn như phân loại cảm xúc và tóm tắt – đặc biệt hữu ích trong các lĩnh vực như đánh giá sản phẩm, nơi cảm xúc và nội dung tóm tắt thường gắn liền với nhau.

**2.2.5. Hạn chế và tính chuyển giao**

Dù có nhiều tiến bộ, phần lớn nghiên cứu và bộ dữ liệu vẫn tập trung vào tiếng Anh và một số ngôn ngữ giàu tài nguyên khác. Do đó, khả năng tổng quát hóa của các mô hình này sang tiếng Việt – đặc biệt trong các lĩnh vực có văn bản không chính thức và do người dùng tạo ra – vẫn là một câu hỏi bỏ ngỏ. Phần tiếp theo sẽ đi sâu vào các thách thức này, đưa ra góc nhìn tập trung vào tiếng Việt về những hạn chế của công cụ hiện tại, đồng thời chỉ ra các cơ hội cho nghiên cứu trong tương lai.

**2.3. Bối cảnh tiếng Việt**

**2.3.1. Thách thức ngôn ngữ**

Tiếng Việt có một số đặc điểm ngôn ngữ độc đáo khiến việc tóm tắt văn bản trở nên phức tạp. Khác với nhiều ngôn ngữ Ấn-Âu, tiếng Việt là ngôn ngữ phân tích, nghĩa là dựa nhiều vào thứ tự từ thay vì biến hình từ để biểu thị cấu trúc ngữ pháp. Thêm vào đó, tiếng Việt là ngôn ngữ thanh điệu, nghĩa của một từ có thể thay đổi tùy theo thanh điệu sử dụng. Những đặc điểm này, mặc dù làm phong phú giao tiếp hàng ngày, lại đặt ra thách thức cho các nhiệm vụ xử lý ngôn ngữ tự nhiên (NLP), vì mô hình phải phân tích ngữ cảnh và ý nghĩa mà không thể dựa vào các dấu hiệu cú pháp truyền thống.

Hơn nữa, tiếng Việt còn thường xuyên sử dụng các biểu thức không chính thức, bao gồm tiếng lóng và các từ viết tắt, đặc biệt phổ biến trong nội dung người dùng tạo ra như đánh giá sản phẩm. Nhiều người dùng trên các nền tảng thương mại điện tử dùng các cụm từ đời thường và cách viết sáng tạo (ví dụ “pin trâu” để chỉ “pin lâu”), vốn không có trong dữ liệu huấn luyện chuẩn mực. Thêm vào đó, cấu trúc câu tiếng Việt thường không chuẩn, nhiều câu thiếu chủ ngữ hoặc tân ngữ rõ ràng, kèm theo các phép lược bỏ (ellipsis) khiến những thành phần cú pháp quan trọng bị mất đi. Các yếu tố này cản trở khả năng phân tích hiệu quả của các mô hình tóm tắt thông thường, dẫn đến khó khăn trong việc tạo ra bản tóm tắt mạch lạc và giàu ý nghĩa.

**2.3.2. Hạn chế về bộ dữ liệu**

Sự sẵn có của các bộ dữ liệu chú thích chất lượng cao là yếu tố then chốt để huấn luyện mô hình học sâu cho tóm tắt văn bản. Trong khi các bộ dữ liệu quy mô lớn như CNN/DailyMail hay XSum dễ dàng tìm thấy cho tiếng Anh, các bộ dữ liệu tóm tắt tiếng Việt vẫn rất khan hiếm. Phần lớn dữ liệu hiện có được thu thập thủ công hoặc giới hạn trong một số lĩnh vực nhất định, thường đòi hỏi khâu tiền xử lý phức tạp để xử lý ngôn ngữ không chính thức và đảm bảo tính nhất quán trong việc chú thích.

Một thách thức khác là tính đa dạng chủ đề trong các đánh giá tiếng Việt. Ví dụ, trong đánh giá sản phẩm, người dùng có thể thảo luận về nhiều khía cạnh như hiệu năng, thời lượng pin, chất lượng camera, giá cả… nhưng thường theo cách rất chủ quan. Sự đa dạng mang tính cá nhân này khiến việc chuẩn hóa hoặc tự động chú thích nội dung trở nên khó khăn. Hơn nữa, các đánh giá tiếng Việt thường ngắn và rời rạc, làm phức tạp thêm nhiệm vụ tạo ra bản tóm tắt đủ để phản ánh đầy đủ quan điểm. Do đó, các bộ dữ liệu tiếng Việt hiện tại thường thiếu quy mô, độ nhất quán và chiều sâu để huấn luyện các mô hình tóm tắt vững chắc. Những hạn chế này nhấn mạnh nhu cầu về các bộ dữ liệu mới dành riêng cho tóm tắt văn bản tiếng Việt, với sự chú ý kỹ lưỡng đến các thách thức theo từng miền.

**2.3.3. Các bộ dữ liệu và công cụ tiếng Việt hiện có**

Mặc dù gặp nhiều khó khăn, vẫn có một số nỗ lực đáng chú ý trong việc xây dựng tài nguyên cho các nhiệm vụ NLP tiếng Việt. VLSP (Vietnamese Language and Speech Processing) cung cấp các bộ dữ liệu cho các bài toán như gán nhãn từ loại, phân tích phụ thuộc cú pháp và nhận dạng thực thể có tên, nhưng thiếu bộ dữ liệu tóm tắt quy mô lớn. PhoSum và VNSum là những cố gắng cung cấp dữ liệu tóm tắt tiếng Việt, nhưng các tập này tương đối nhỏ và chưa được ứng dụng rộng rãi do hạn chế về phạm vi chủ đề và sự không nhất quán trong chú thích.

Về mô hình ngôn ngữ, PhoBERT là một trong những mô hình tiếng Việt nổi bật, xây dựng trên kiến trúc BERT và tiền huấn luyện trên một kho nhỏng liệu lớn tiếng Việt. PhoBERT đã thể hiện tiềm năng ở nhiều nhiệm vụ hạ nguồn, bao gồm phân tích cảm xúc và phân loại văn bản. ViT5, biến thể đa ngôn ngữ của T5, cũng đã được tiền huấn luyện và tinh chỉnh cho tiếng Việt, nhưng ứng dụng trong tóm tắt vẫn còn nhiều tiềm năng để khám phá. Các mô hình khác như viGPT2 đã được dùng cho nhiệm vụ sinh văn bản nhưng chưa được áp dụng rộng rãi cho tóm tắt tiếng Việt.

Dù có những tiến bộ này, vẫn tồn tại khoảng cách giữa các mô hình ngôn ngữ chung và yêu cầu cụ thể khi tóm tắt nội dung ngắn, tiếng ồn và mang tính chủ quan cao của các đánh giá sản phẩm tiếng Việt. Thách thức trong việc điều chỉnh những mô hình này để tóm tắt các bình luận ngắn gọn, lộn xộn và mang sắc thái cá nhân là hướng nghiên cứu quan trọng mà khảo sát này sẽ tập trung làm rõ trong các phần tiếp theo.

**CHƯƠNG 3: ĐỀ XUẤT GIẢI PHÁP**

# **CHƯƠNG 4: ĐÁNH GIÁ GIẢI PHÁP ĐỀ XUẤT**

# **KẾT LUẬN VÀ KIẾN NGHỊ**

**TÀI LIỆU THAM KHẢO**