PHÁT HIỆN TÍNH CÁCH TRONG VĂN BẢN BẰNG TRÍ TUỆ NHÂN TẠO: TỪ LÝ THUYẾT ĐẾN ỨNG DỤNG

Nguyễn Thanh Nam[1*], Nguyễn Công Hào[2], Lê Mạnh Thạnh[3], Nguyễn Đình Hoa Cương[4]

[1] Văn phòng, Đại học Huế

[2] Ban Thanh tra và Pháp chế, Đại học Huế

[3] Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế

[4] Khoa Công nghệ – Kinh doanh, Trường Đại học Phú Xuân

*Email: ntnam@hueuni.edu.vn

Ngày nhận bài: 3/12/2024; ngày hoàn thành phản biện: 17/12/2024; ngày duyệt đăng: 20/12/2024

TÓM TẮT

Phát hiện tính cách từ văn bản bằng Trí tuệ nhân tạo (AI) là một lĩnh vực liên ngành quan trọng, kết nối giữa xử lý ngôn ngữ tự nhiên, học máy và tâm lý học. Mục tiêu là xác định hoặc suy luận các đặc điểm tính cách của một cá nhân dựa trên văn bản viết hoặc phát ngôn. Bài báo cung cấp một cách tổng quan về phát hiện tính cách, từ khái niệm cơ bản, đến kỹ thuật phát hiện tính cách từ văn bản, đồng thời trình bày chi tiết kiến trúc mô hình dự đoán, bao gồm tiền xử lý dữ liệu, trích xuất đặc trưng và mô hình phân loại. Tiếp theo bài báo chỉ ra những tiến bộ nghiên cứu gần đây trong dự đoán tính, ứng dụng trong nhiều lĩnh vực, cũng như một số hạn chế của các nghiên cứu đã công bố. Cuối cùng, bài báo đề xuất các hướng nghiên cứu tương lai để cải thiện hiệu quả và tính công bằng của các hệ thống phát hiện tính cách.

Từ khóa: Học máy, Học sâu, Phát hiện đặc điểm tính cách, Xử lý ngôn ngữ tự nhiên.

1. MỞ ĐẦU

Khả năng tự động phát hiện tính cách của một cá nhân từ nội dung văn bản của họ đã trở nên quan trọng đáng kể trong những năm gần đây do các ứng dụng tiềm năng của nó trên nhiều lĩnh vực khác nhau [1-5]. Mặc dù đầy tiềm năng, lĩnh vực này vẫn còn nhiều thách thức, thúc đẩy việc ứng dụng các kỹ thuật AI và xử lý ngôn ngữ tự nhiên (NLP) hiện đại [2, 5, 6].

Nghiên cứu trình bày một cách hệ thống về phát hiện tính cách từ văn bản dựa trên AI, từ các khái niệm cơ bản đến sự tiến triển của các kiến trúc tính toán từ phương

pháp ban đầu đến mô hình học sâu hiện đại. Cấu trúc của các hệ thống dự đoán tính cách, chi tiết các giai đoạn quan trọng như tiền xử lý dữ liệu, kỹ thuật trích xuất đặc trưng và các mô hình phân loại, từ thuật toán học máy truyền thống đến các kiến trúc học sâu phức tạp. Bài báo cũng chỉ ra những tiến bộ gần đây trong phát hiện tính cách qua văn bản ở nhiều ứng dụng khác nhau, xem xét các yếu tố thúc đẩy và hạn chế cần khắc phục. Cuối cùng, bài báo tổng hợp các thách thức hiện tại và đề xuất hướng nghiên cứu tương lai nhằm cải thiện độ chính xác và khả năng ứng dụng của AI trong việc hiểu tính cách từ văn bản.

Cấu trúc bài báo được trình bày trong 4 phần, ngoài phần mở đầu, phần 2 trình bày các mô hình phân loại tính cách, các kỹ thuật phát hiện tính cách từ văn bản và kiến trúc của mô hình dự đoán tính cách. Phần 3 trình bày các ứng dụng thực tế của việc phát hiện tính cách và những thách thức cũng như triển vọng trong nghiên cứu. Phần 4 trình bày kết luận cũng như định hướng nghiên cứu.

2. PHÁT HIỆN TÍNH CÁCH TRONG VĂN BẢN BẰNG AI

2.1. Các mô hình phân loại tính cách

2.1.1. Phân loại đặc điểm tính cách

Phân loại đặc điểm tính cách bao gồm việc xác định và phân loại các đặc điểm tính cách riêng biệt của một cá nhân dựa trên khuôn khổ tâm lý cụ thể [5]. Bản thân tính cách bao gồm khuynh hướng nhận thức, cảm xúc và hành vi của một cá nhân khi tương tác với thế giới, hình thành nên một khuôn mẫu độc đáo và dễ nhận biết, ảnh hưởng đến suy nghĩ và hành động của họ. Mục tiêu của việc nghiên cứu dự đoán tính cách là hiểu và dự đoán các kiểu hành vi và đặc điểm này [2, 3].

Phân loại đặc điểm tính cách trong NLP ngày càng quan trọng nhờ sự gia tăng dữ liệu văn bản từ mạng xã hội và các nền tảng số [7, 8]. Những dữ liệu này phản ánh sâu sắc hành vi, cảm xúc và suy nghĩ cá nhân, cung cấp nguồn thông tin phong phú để suy luận tính cách [1]. Việc hiểu tính cách mang lại giá trị học thuật và ứng dụng rộng rãi trong cá nhân hóa nội dung, tiếp thị, tuyển dụng, y tế và các hệ thống AI khác [2, 3, 5, 6, 7, 9]. Thay vì sử dụng bảng hỏi truyền thống, các phương pháp NLP hiện đại cho phép trích xuất đặc trưng ngôn ngữ và áp dụng học máy để phân loại tính cách hiệu quả và mở rộng trên quy mô lớn.

2.1.2. Các mô hình tính cách áp dụng trong nghiên cứu

Các nhà tâm lý học đã phát triển nhiều mô hình để mô tả và phân loại tính cách con người, trong đó có hai mô hình phổ biến nhất được áp dụng trong nghiên cứu trích xuất tính cách từ văn bản.

Mô hình Big Five (OCEAN) gồm năm chiều chính: Cởi mở (Openness), Tận tâm (Conscientiousness), Hướng ngoại (Extraversion), Tính dễ chịu (Agreeableness) và Rối loạn thần kinh (Neuroticism). Mỗi chiều đại diện cho những đặc điểm hành vi và cảm xúc riêng biệt. Big Five được công nhận là một khung mô hình toàn diện, ổn định qua nhiều nền văn hóa và độ tuổi, và được sử dụng rộng rãi trong nghiên cứu phát hiện tính cách từ văn bản [10], [11].

Mô hình Myers-Briggs Type Indicator (MBTI) phân loại cá nhân thành 16 kiểu tính cách dựa trên bốn cặp phân đôi: Hướng nội/Hướng ngoại (I/E), Cảm giác/Trực giác (S/N), Suy nghĩ/Cảm xúc (T/F), và Đánh giá/Nhận thức (J/P). MBTI nhấn mạnh cách cá nhân xử lý thông tin, ra quyết định trong cuộc sống hàng ngày, từ đó dự đoán hành vi và động cơ [9, 10]. Không giống Big Five, MBTI chia thành các nhóm riêng biệt thay vì liên tục, và thường được ứng dụng trong thực tiễn như tuyển dụng, huấn luyện và phát triển cá nhân.

Ngoài hai mô hình trên, còn có các mô hình khác như Allport, Cattell (16 yếu tố), Eysenck (Bộ ba nhân cách), DISC và Bộ ba đen tối [1], [12]. Việc lựa chọn mô hình tùy thuộc vào mục tiêu nghiên cứu hoặc ứng dụng.

2.2. Một số kỹ thuật phát hiện tính cách từ văn bản:

Các phương pháp tiếp cận ban đầu: Các phương pháp phát hiện tính cách ban đầu chủ yếu dựa vào các phương pháp tiếp cận dựa trên từ vựng, đặc biệt là sử dụng các từ vựng tâm lý như LIWC (Linguistic Inquiry and Word Count) [6, 13]. Các hệ thống này phân loại các từ thành các nhóm có ý nghĩa về mặt tâm lý và dự đoán các đặc điểm dựa trên tần suất từ [6, 14]. LIWC đóng vai trò quan trọng trong việc thiết lập mối liên hệ giữa cách sử dụng ngôn ngữ và các đặc điểm tâm lý. Mặc dù cơ bản, các phương pháp này đã đặt nền tảng cho các mô hình phức tạp hơn.

Phương pháp tiếp cận học máy (ML) truyền thống: Giai đoạn này tập trung vào các đặc điểm ngôn ngữ được thiết kế thủ công như số lượng từ, cấu trúc cú pháp và cực tính tình cảm [7, 8]. Các đặc điểm này được sử dụng với các thuật toán học máy như SVM, Naive Bayes, KNN, Decision Trees, Random Forests và XGBoost để phân loại các đặc điểm tính cách [7, 14, 15]. Các nghiên cứu đã chứng minh rằng các đặc điểm thủ công này có mối tương quan tốt với các chiều hướng tính cách khác nhau [5].

Phương pháp tiếp cận học sâu (DL): Sự trỗi dậy của học sâu đã giới thiệu các kỹ thuật mã hóa văn bản (Word2Vec, GloVe, FastText) có thể nắm bắt được các biểu diễn ngữ nghĩa phong phú hơn [5, 6, 14, 16]. Các mô hình sâu như CNN, RNN, LSTM và Bi-LSTM tự động trích xuất tính năng và cải thiện hiệu suất dự đoán [2, 5, 7, 8, 17]. Các nhà nghiên cứu như Majumder và Sun [5] đã chứng minh rằng việc kết hợp các tính năng ngữ nghĩa và phong cách bằng cách sử dụng các mô hình CNN hoặc Bi-LSTM-CNN lai đã tăng cường độ chính xác. Tuy nhiên, các mô hình dựa trên RNN ban đầu có những

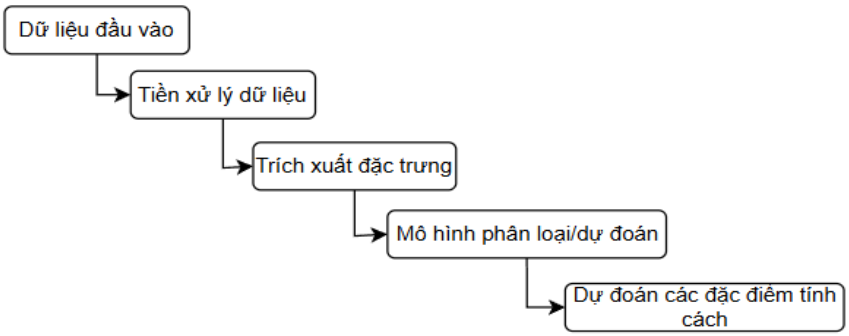
hạn chế, bao gồm đào tạo chậm và khó xử lý tình trạng mất ngữ cảnh và các từ ngoài vốn từ vựng [5, 18]. Những vấn đề này dẫn đến việc tìm kiếm các mô hình nhận thức ngữ cảnh hơn.

Mô hình ngôn ngữ được đào tạo trước (PLM - Pre-trained Language Models): Tiến bộ gần đây được thúc đẩy bởi các mô hình ngôn ngữ được đào tạo trước như BERT, RoBERTa, XLNet và ERNIE [5, 8, 18, 19, 20]. Các mô hình này cung cấp các từ nhúng (word embedding) nhạy cảm với ngữ cảnh và xử lý tốt hơn các sắc thái đa nghĩa và ngôn ngữ. Việc tinh chỉnh PLM trên các tập dữ liệu tính cách đã dẫn đến những cải tiến đáng kể về độ chính xác của dự đoán [6, 18].

Giai đoạn phát triển các mô hình ngôn ngữ lớn (LLMs - Large Language Models) như GPT-4, Claude hay LLaMA đã mở ra một kỷ nguyên mới cho việc phát hiện tính cách từ văn bản. Với khả năng xử lý ngôn ngữ vượt trội, LLMs cho phép phân tích sâu sắc hơn về ngữ cảnh, phong cách và sắc thái tâm lý – những yếu tố quan trọng trong nhận diện đặc điểm tính cách [21]. Không giống như các mô hình truyền thống phụ thuộc mạnh vào dữ liệu gắn nhãn, LLMs có thể hoạt động hiệu quả trong thiết lập zero-shot hoặc few-shot learning, nhờ đó giảm thiểu đáng kể nhu cầu về dữ liệu huấn luyện [13]. Ngoài ra, các LLM còn có khả năng giải thích lý do đằng sau các dự đoán tính cách, góp phần hướng tới các hệ thống AI có khả năng giải thích được (explainable AI). Cuối cùng, LLMs còn hỗ trợ phân tích đa nhiệm và đa phương thức, kết hợp thông tin từ văn bản, hình ảnh và âm thanh để đưa ra nhận định toàn diện hơn về tính cách con người [22-24].

2.3. Kiến trúc của mô hình dự đoán tính cách

Các mô hình dự đoán tính cách thường tuân theo quy trình gồm các bước: thu thập dữ liệu văn bản do người dùng tạo [25], xử lý trước để làm sạch và định dạng dữ liệu, trích xuất đặc trưng ngôn ngữ và ngữ nghĩa, sau đó đưa vào mô hình học máy để dự đoán tính cách theo mô hình định sẵn [1, 2, 5, 15, 17]. Cuối cùng, hiệu suất mô hình được đánh giá bằng các chỉ số để đảm bảo độ chính xác và hiệu quả. Cách tiếp cận này giúp tự động suy luận đặc điểm tính cách từ văn bản liên quan đến người dùng [18].



Hình 1. Khung mô hình dự đoán tính cách

2.3.1. Dữ liệu huấn luyện (Dataset)

Hiệu quả của mô hình phát hiện tính cách phụ thuộc lớn vào chất lượng bộ dữ liệu huấn luyện. Trước hết, dữ liệu cần được gắn nhãn tính cách đáng tin cậy dựa trên các mô hình chuẩn như Big Five hoặc MBTI, với nguồn nhãn đến từ các bài kiểm tra tâm lý học chính thống [7, 11]. Ngoài ra, văn bản trong tập dữ liệu cần mang tính biểu đạt cá nhân cao nhằm phản ánh được cảm xúc, suy nghĩ và hành vi của người viết, vốn là các biểu hiện ngôn ngữ có liên hệ chặt chẽ với đặc điểm tính cách [1, 19].

Bên cạnh đó, tập dữ liệu cần đảm bảo sự đa dạng về nội dung, bối cảnh ngôn ngữ và đặc điểm nhân khẩu học để tăng tính khái quát hóa [3, 16, 25]. Kích thước dữ liệu cũng là yếu tố quan trọng: các mô hình học sâu và mô hình ngôn ngữ lớn thường đòi hỏi tập dữ liệu quy mô lớn hoặc cần sử dụng kỹ thuật Few-shot/Zero-shot (Few-shot learning và Zero-shot learning là hai kỹ thuật học máy giúp mô hình thực hiện nhiệm vụ mà không cần hoặc chỉ cần rất ít dữ liệu huấn luyện có gắn nhãn) [1, 5].

2.3.2. Tiền xử lý dữ liệu (Preprocessing)

Tiền xử lý dữ liệu là bước quan trọng nhằm chuẩn bị văn bản đầu vào cho quá trình trích xuất đặc trưng và huấn luyện mô hình phát hiện tính cách. Các thao tác phổ biến bao gồm loại bỏ stop words (các từ phổ biến như “the”, “is”, “a” không mang nhiều ý nghĩa phân biệt), loại bỏ đường dẫn URL, ký hiệu đặc biệt, chữ số, hashtags (#) và mentions (@ – chỉ đề cập người dùng trong mạng xã hội), chuẩn hoá ký tự và chuyển toàn bộ văn bản về chữ thường để thống nhất định dạng [15, 17].

Văn bản cũng được tách thành các token (đơn vị từ hoặc cụm từ nhỏ nhất) thông qua bước tokenization, sau đó áp dụng lemmatization (đưa từ về dạng gốc có nghĩa) hoặc stemming (cắt gọn từ về gốc) nhằm giảm số lượng biến thể từ vựng không cần thiết. Các thành phần như dấu câu, khoảng trắng dư thừa, biểu tượng cảm xúc và emoji (hình biểu cảm) có thể được loại bỏ hoặc xử lý riêng tuỳ theo mục tiêu phân tích [2]. Mục tiêu chung là làm sạch và chuẩn hoá dữ liệu văn bản, đảm bảo chỉ giữ lại những yếu tố ngôn ngữ mang giá trị nhận diện cho mô hình phát hiện tính cách.

2.3.3. Trích xuất đặc trưng (Feature Extraction)

Trích xuất đặc trưng là bước chuyển văn bản đã tiền xử lý thành định dạng phù hợp với mô hình học máy hoặc học sâu. Các kỹ thuật truyền thống như Bag of Words (BoW) và TF-IDF giúp biểu diễn văn bản dựa trên tần suất và mức độ quan trọng của từ, đặc biệt hữu ích trong dự đoán tính cách như sự thân thiện. Ngoài ra, POS tagging, N-grams, phân tích cảm xúc, các đặc trưng tâm lý-ngôn ngữ (như chiều dài từ, tỷ lệ từ loại) và đặc trưng phong cách viết (như độ dài câu, độ phong phú từ vựng) cũng đóng vai trò quan trọng. Đáng chú ý, công cụ LIWC phân loại từ theo các danh mục tâm lý học để hỗ trợ liên hệ với đặc điểm tính cách [12, 14, 17].

Với các mô hình học sâu, văn bản được biểu diễn thông qua word embeddings - tức các vector liên tục giàu thông tin ngữ nghĩa. Các phương pháp như Word2Vec [17], GloVe [5], và FastText [17] được dùng để ánh xạ từ ngữ vào không gian vector. Ngoài ra, biểu diễn câu đầy đủ thông qua sentence embeddings hoặc các vector ngữ cảnh từ các PLM như BERT, RoBERTa, XLNet, ELECTRA hay ERNIE giúp mô hình hiểu được ngữ nghĩa sâu, các mối quan hệ phân cấp và phụ thuộc ngữ cảnh dài hạn [5, 17].

2.3.4. Mô hình phân loại/dự đoán (Classification/Prediction Models)

Các mô hình phân loại/dự đoán được sử dụng nhằm ánh xạ các đặc trưng trích xuất từ văn bản sang các đặc điểm tính cách dựa trên các mô hình như Big Five hoặc MBTI. Các thuật toán học máy truyền thống như Naïve Bayes, SVM, Logistic Regression và Random Forest đã được sử dụng rộng rãi trong phân loại tính cách từ văn bản [1, 15, 18]. Những mô hình này đòi hỏi đặc trưng được trích xuất thủ công và dữ liệu được gắn nhãn đầy đủ. Tuy nhiên, khả năng học ngữ nghĩa sâu của các mô hình ML vẫn còn hạn chế.

Bên cạnh đó, thuật toán học sâu ngày càng được ứng dụng rộng rãi nhờ khả năng học đặc trưng tự động từ dữ liệu lớn. Các mô hình học sâu như RNN, LSTM, Bi-LSTM và CNN đã chứng minh hiệu quả vượt trội trong phát hiện tính cách nhờ khả năng học đặc trưng tự động và xử lý chuỗi ngôn ngữ. LSTM giúp nắm bắt các phụ thuộc dài hạn trong văn bản, còn CNN khai thác các mẫu ngữ pháp cục bộ [6, 17, 28]. Các mô hình lai như CNN-LSTM thường cho kết quả tốt hơn so với từng mô hình riêng lẻ [9, 17]. Ngoài ra, các mô hình dựa trên transformer như BERT, RoBERTa và XLNet đang dẫn đầu trong nhiều nghiên cứu hiện nay [2, 5, 17].

Đặc biệt, sự xuất hiện của các mô hình ngôn ngữ lớn như GPT, PaLM và Claude đã mở ra một hướng mới trong phát hiện tính cách nhờ khả năng hiểu sâu ngữ cảnh và thực hiện tốt với rất ít hoặc không cần dữ liệu gắn nhãn (zero-shot, few-shot) [5, 29]. Ngoài phân loại, các mô hình này còn giải thích được lý do dự đoán, giúp tăng tính minh bạch. LLMs cũng hỗ trợ tích hợp dữ liệu đa phương thức như văn bản, hình ảnh, âm thanh để cung cấp phân tích tính cách toàn diện hơn [13, 22].

2.3.5. Dự đoán các đặc điểm tính cách

Bước cuối cùng liên quan đến việc sử dụng mô hình đã được đào tạo để dự đoán các đặc điểm tính cách trên dữ liệu văn bản mới. Điều này đòi hỏi phải đưa văn bản đã được xử lý trước và được thiết kế theo đặc điểm vào mô hình phân loại hoặc dự đoán, sau đó đưa ra dự đoán dựa trên các mẫu mà nó đã học được trong quá trình đào tạo. Đầu ra của mô hình có thể ở dạng điểm xác suất cho từng đặc điểm tính cách hoặc danh mục. Sau đó, ngưỡng quyết định có thể được áp dụng cho các xác suất này để gán nhãn tính cách xác định. Cuối cùng, các đặc điểm tính cách dự đoán sau đó được đánh giá so với các đánh giá tính cách thực tế bằng cách sử dụng các số liệu đánh giá hiệu suất mô hình.

2.3.6. Đánh giá hiệu suất mô hình (Evaluation Metrics)

Đánh giá mô hình dự đoán tính cách cần kết hợp nhiều chỉ số như Accuracy, Precision, Recall, F1-score, Exact Match Ratio, Hamming Loss và Zero-One Loss để phản ánh hiệu quả trên các nhiệm vụ đa dạng. Các chỉ số này giúp hiểu rõ điểm mạnh, hạn chế của mô hình và hỗ trợ cải thiện qua phân tích như ma trận nhầm lẫn (Confusion Matrix).

3. ỨNG DỤNG DỰ ĐOÁN TÍNH CÁCH VÀ HƯỚNG NGHIÊN CỨU

3.1. Ứng dụng dự đoán tính cách

Những nghiên cứu gần đây trong dự đoán tính cách đang được ứng dụng trên một số lĩnh vực:

- Hệ thống đề xuất: Nhận dạng tính cách được sử dụng để phát triển các khuyến nghị được cá nhân hóa hơn [5, 19].
- Tuyển dụng: Các đặc điểm tính cách được sử dụng trong tuyển dụng điện tử và lựa chọn ứng viên để cải thiện sự phù hợp với công việc [3, 5, 9].
- Giáo dục: Phát hiện tính cách có thể góp phần phát triển các trải nghiệm học tập được cá nhân hóa [5].
- Phân tích phương tiện truyền thông xã hội: Tính cách suy ra từ văn bản phương tiện truyền thông xã hội giúp hiểu hành vi, sở thích và ảnh hưởng của người dùng [1, 7, 18].
- Chăm sóc sức khỏe: Dự đoán tính cách có tiềm năng dự đoán rủi ro sức khỏe tâm thần và điều chỉnh các biện pháp can thiệp [9, 19].
- An ninh mạng: Các đặc điểm tính cách có thể được sử dụng để dự đoán khả năng bị tấn công kỹ thuật xã hội [19].
- Tương tác giữa người và máy tính: Tính cách của người dùng có thể cung cấp thông tin cho việc thiết kế các giao diện và chatbot cá nhân hóa và thích ứng hơn [4, 8].
- Dự báo: Các đặc điểm tính cách có thể được đưa vào các mô hình dự báo cho các chính sách của tổ chức và quá trình ra quyết định của người tiêu dùng, với các ứng dụng trong tài chính và y tế [8].
- Chatbot được cá nhân hóa và các tác nhân đàm thoại: Có thể phát triển Chatbot với các tính cách phù hợp với sở thích của người dùng [4].

3.2. Thách thức và hướng nghiên cứu

Một số thách thức chính vẫn còn tồn tại trong lĩnh vực phát hiện tính cách từ văn bản. Trước hết, việc thu thập dữ liệu có nhãn chất lượng cao là rào cản lớn, đặc biệt trong bối cảnh dữ liệu ngôn ngữ trên mạng xã hội có tính đa ngôn ngữ cao và thiếu tính chuẩn

hóa. Việc khai thác thông tin tính cách từ nội dung đa phương tiện (hình ảnh, video, âm thanh) cũng chưa đạt hiệu quả như mong đợi [6, 18].

Ngoài ra, ngôn ngữ trực tuyến thường mang tính cảm xúc, chứa tiếng lóng và có nhiều biểu đạt phi cấu trúc, gây khó khăn trong việc suy luận chính xác đặc điểm tính cách. Thêm vào đó, tính đa dạng văn hóa và ngôn ngữ đòi hỏi mô hình phải có khả năng xử lý xuyên ngôn ngữ và thích ứng với các bối cảnh khác nhau để đảm bảo khả năng ứng dụng toàn cầu.

Trước những thách thức nêu trên, các hướng nghiên cứu hiện nay tập trung vào việc phát triển các mô hình mạnh và có khả năng khái quát tốt hơn trên nhiều miền dữ liệu, thông qua việc tích hợp dữ liệu đa phương thức và áp dụng các kỹ thuật học chuyển giao. Một hướng đầy hứa hẹn khác là ứng dụng mô hình ngôn ngữ lớn như ChatGPT trong phát hiện tính cách, nhờ khả năng hiểu ngôn ngữ sâu và linh hoạt [4, 5, 22].

Ngoài ra, việc xây dựng các mô hình có khả năng diễn giải cao giúp xác định rõ mối liên hệ giữa đặc trưng ngôn ngữ và tính cách, từ đó nâng cao độ tin cậy và tính minh bạch khi áp dụng trong thực tiễn [19]. Cuối cùng, kết hợp phát hiện tính cách với phân tích cảm xúc được kỳ vọng sẽ cung cấp cái nhìn toàn diện hơn về hành vi con người, góp phần nâng cao chất lượng phân tích văn bản trong các hệ thống thông minh [4].

4. KẾT LUẬN

Nghiên cứu cho thấy sự phát triển từ các mô hình học máy truyền thống đến các kiến trúc học sâu như CNN, RNN, và các mô hình transformer như BERT, đồng thời ghi nhận vai trò ngày càng nổi bật của các mô hình ngôn ngữ lớn như GPT và Claude trong việc nâng cao hiệu quả phát hiện tính cách. Bên cạnh các tiến bộ kỹ thuật, nghiên cứu cũng chỉ ra nhiều thách thức còn tồn tại như hạn chế về dữ liệu huấn luyện, sự phức tạp ngôn ngữ và văn hóa, cũng như yêu cầu về khả năng diễn giải và tính công bằng của mô hình. Các hướng nghiên cứu tương lai được đề xuất bao gồm: tích hợp dữ liệu đa phương thức, phát triển các mô hình giải thích được, và mở rộng khả năng khái quát hóa qua học chuyển giao và học bán giám sát.

TÀI LIỆU THAM KHẢO

[1] H. Ahmad, M. Z. Asghar, A. S. Khan, and A. Habib, "A Systematic Literature Review of Personality Trait Classification from Textual Content," *Open Computer Science*, vol. 10, no. 1, pp. 175–193, Jul. 2020, doi: 10.1515/comp-2020-0188.

[2] A. Naz, H. U. Khan, S. Alesawi, O. Ibrahim Abouola, A. Daud, and M. Ramzan, "AI Knows You: Deep Learning Model for Prediction of Extroversion Personality Trait," *IEEE Access*, vol. 12, pp. 159152–159175, 2024, doi: 10.1109/ACCESS.2024.3486578.

[3] H. Perera and L. Costa, "Personality Classification of text through Machine learning and Deep learning: A Review (2023)," 2023, Accessed: Apr. 09, 2025. [Online]. Available: https://www.authorea.com/doi/full/10.36227/techrxiv.22337746?commit=ecb1ed18b8f306f69964cf6356e99a275d497083

[4] F. Safari and A. Chalechale, "Emotion and personality analysis and detection using natural language processing, advances, challenges and future scope," *Artificial Intelligence Review*, Sep. 2023, doi: 10.1007/s10462-023-10603-3.

[5] B. Alshouha, J. Serrano-Guerrero, F. Chiclana, F. P. Romero, and J. A. Olivas, "Personality Trait Detection via Transfer Learning," *CMC*, vol. 78, no. 2, pp. 1933–1956, 2024, doi: 10.32604/cmc.2023.046711.

[6] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Information Processing & Management*, vol. 58, no. 3, p. 102532, May 2021, doi: 10.1016/j.ipm.2021.102532.

[7] J. Serrano-Guerrero, "Combining machine learning algorithms for personality trait prediction," *Egyptian Informatics Journal*, 2024.

[8] K. Yang, R. Y. K. Lau, and A. Abbasi, "Getting Personal: A Deep Learning Artifact for Text-Based Measurement of Personality," *Information Systems Research*, vol. 34, no. 1, pp. 194–222, Mar. 2023, doi: 10.1287/isre.2022.1111.

[9] A. G. Shanmukha, R. S. Shamyuktha, S. Karan, D. Gupta, and S. Palaniswamy, "Advancing Personality Detection through Word Embedments and Deep Learning: An Examination Using the MBTI Dataset," in *2024 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, May 2024, pp. 1–6. doi: 10.1109/RAICS61201.2024.10689948.

[10] Md. A. Akber, T. Ferdousi, R. Ahmed, R. Asfara, R. Rab, and U. Zakia, "Personality and emotion—A comprehensive analysis using contextual text embeddings," *Natural Language Processing Journal*, vol. 9, p. 100105, Dec. 2024, doi: 10.1016/j.nlp.2024.100105.

[11] Y. Ji *et al.*, "Personality-assisted mood modeling with historical reviews for sentiment classification," *Information Sciences*, 2023.

[12] M. Ramezani *et al.*, "Automatic personality prediction: an enhanced method using ensemble modeling," *Neural Computing and Applications*, vol. 34, no. 21, pp. 18369–18389, Nov. 2022, doi: 10.1007/s00521-022-07444-6.

[13] B. Zhan, Y. Huang, W. Cui, H. Zhang, and J. Shang, "Humanity in AI: Detecting the Personality of Large Language Models," Oct. 11, 2024, *arXiv*: arXiv:2410.08545. doi: 10.48550/arXiv.2410.08545.

[14] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2313–2339, Apr. 2020, doi: 10.1007/s10462-019-09770-z.

[15] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, and I. Pappas, "A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection," *Artif Intell Rev*, vol. 56, no. 12, pp. 15129–15215, Dec. 2023, doi: 10.1007/s10462-023-10509-0.

[16] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, no. January, p. 100026, 2023, doi: 10.1016/j.nlp.2023.100026.

[17] R. Alsini, A. Naz, H. U. Khan, A. Bukhari, A. Daud, and M. Ramzan, "Using deep learning and word embeddings for predicting human agreeableness behavior," *Sci Rep*, vol. 14, no. 1, p. 29875, Dec. 2024, doi: 10.1038/s41598-024-81506-8.

[18] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J Big Data*, vol. 8, no. 1, p. 68, May 2021, doi: 10.1186/s40537-021-00459-1.

[19] H. Lin, "A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed Gray Wolf Optimizer for feature selection," *Information Processing and Management*, 2023.

[20] K. Kelvin and Y. Utomo, "Overview of Text Based Personality Prediction Using Deep Learning," *EMACS Journal*, vol. 6, no. 2, pp. 93–100, May 2024, doi: 10.21512/emacsjournal.v6i2.11550.

[21] H. Peters and S. Matz, "Large Language Models Can Infer Psychological Dispositions of Social Media Users," Jun. 05, 2024, *arXiv*: arXiv:2309.08631. doi: 10.48550/arXiv.2309.08631.

[22] G. Serapio-García *et al.*, "Personality Traits in Large Language Models," Mar. 11, 2025, *arXiv*: arXiv:2307.00184. doi: 10.48550/arXiv.2307.00184.

[23] A. Hilliard, C. Munoz, Z. Wu, and A. S. Koshiyama, "Eliciting Personality Traits in Large Language Models," Feb. 15, 2024, *arXiv*: arXiv:2402.08341. doi: 10.48550/arXiv.2402.08341.

[24] X. Song, A. Gupta, K. Mohebbizadeh, S. Hu, and A. Singh, "Have Large Language Models Developed a Personality?: Applicability of Self-Assessment Tests in Measuring Personality in LLMs," May 24, 2023, *arXiv*: arXiv:2305.14693. doi: 10.48550/arXiv.2305.14693.

[25] S. A. Patil, S. Patil, and V. V. Tadkal, "Study on the Personality Trait Prediction Techniques Using Social Media Data," in *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Jul. 2024, pp. 1–7. doi: 10.1109/ICAIT61638.2024.10690464.

PERSONALITY DETECTION IN TEXT USING ARTIFICIAL INTELLIGENCE: FROM THEORY TO APPLICATIONS

Nguyen Thanh Nam[1*], Nguyen Cong Hao[2], Le Manh Thanh[3], Nguyen Dinh Hoa Cuong[4]

[1] Office for Administration, Hue University

[2] Department of Inspection and Legislation, Hue University

[3] Faculty of Information Technology, University of Sciences, Hue University

[3] Faculty of Technology - Business, Phu Xuan University

*Email: ntnam@hueuni.edu.vn

ABSTRACT

Personality detection from text using Artificial Intelligence (AI) is a vital interdisciplinary field that integrates natural language processing, machine learning, and psychology. It aims to identify or infer an individual's personality traits based on their written or spoken language. This paper offers a comprehensive overview of personality detection, encompassing foundational concepts and detection techniques, and presents a detailed description of the prediction model architecture, including data preprocessing, feature extraction, and classification models. The paper then highlights recent advancements in personality prediction and its applications across various domains and outlines several limitations in existing studies. Finally, it proposes future research directions to improve the effectiveness and fairness of personality detection systems.

Keywords: Deep Learning, Natural Language Processing, Machine Learning, Personality Trait Detection.

Nguyễn Thanh Nam sinh năm 1981 tại Nghệ An. Ông tốt nghiệp Cử nhân ngành Công nghệ thông tin năm 2003, Thạc sĩ ngành Khoa học máy tính năm 2009 và đang là Nghiên cứu sinh ngành Khoa học máy tính từ năm 2024 tại Trường Đại học Khoa học, Đại học Huế. Hiện nay, ông đang công tác tại Văn phòng, Đại học Huế.

Lĩnh vực nghiên cứu: ứng dụng của học máy và trí tuệ nhân tạo trong xử lý ngôn ngữ tự nhiên.

Nguyễn Công Hào sinh năm 1976 tại Thừa Thiên Huế. Ông tốt nghiệp Cử nhân ngành Toán - Tin năm 1997 tại Trường Đại học Sư phạm, Đại học Huế và Thạc sĩ ngành Tin học năm 2002 tại Trường Đại học Bách khoa Hà Nội; nhận học vị Tiến sĩ ngành Khoa học máy tính năm 2008 tại Viện Công nghệ thông tin thuộc Viện KH và Công nghệ VN. Hiện nay, ông đang công tác tại Ban Thanh tra và Pháp chế, Đại học Huế.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu mờ, Logic mờ, Các phương pháp lập luận xấp xỉ.

Lê Mạnh Thạnh sinh năm 1953 tại Quảng Trị. Ông nhận học vị Tiến sĩ ngành khoa học máy tính tại Đại học Budapest (ELTE), Hungary vào năm 1994 và trở thành Phó giáo sư tại trường Đại học Khoa học, Đại học Huế vào năm 2004. Ông hiện đang là giảng viên Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: cơ sở dữ liệu, cơ sở tri thức và lập trình logic.

Nguyễn Đình Hoa Cương đã nhận bằng Thạc sĩ Khoa học và Tiến sĩ chuyên ngành Khoa học Máy tính tại Trường Đại học Khoa học, Đại học Huế, Việt Nam và tại Đại học Khon Kaen, Thái Lan lần lượt vào các năm 2005 và 2016. Hiện nay, ông là giảng viên tại Trường Đại học Phú Xuân, Huế.

Lĩnh vực nghiên cứu: khai phá dữ liệu, khai phá văn bản, quản lý tri thức và hệ thống hỗ trợ ra quyết định