PHÂN LOẠI NGƯỜI DÙNG CÓ DẤU HIỆU ADHD TỪ NGÔN NGỮ MẠNG XÃ HỘI: SO SÁNH TF-IDF VÀ DISTILBERT

● **NGUYỄN THỊ VI HÀI[1] - NGUYỄN THỊ HUYỀN TRANG[2*]**
[1]Trường Đại học Công nghệ Thành phố Hồ Chí Minh
[2]Trường Đại học Công Thương Thành phố Hồ Chí Minh
*Email: trangnthuyen@huit.edu.vn
DOI: https://doi.org/10.62831/202608011

TÓM TẮT:
Nghiên cứu khảo sát khả năng nhận diện người dùng có dấu hiệu rối loạn tăng động giảm chú ý (ADHD) thông qua ngôn ngữ trên mạng xã hội. Dữ liệu được lấy từ bộ Twitter-STMHD và xử lý ở mức người dùng, với 1.999 mẫu sau tiền xử lý. Trên tập dữ liệu này, nghiên cứu so sánh 3 mô hình gồm: Logistic Regression, Linear SVM và DistilBERT. Kết quả cho thấy, cả 3 mô hình đều phân biệt được 2 nhóm ADHD và đối chứng, trong đó Linear SVM đạt hiệu quả cao nhất với Accuracy 0.8733 và F1-score 0.8812. Kết quả gợi ý tín hiệu ngôn ngữ trên mạng xã hội có thể hỗ trợ sàng lọc ADHD, nhưng không thay thế chẩn đoán lâm sàng.

Từ khóa: ADHD, ngôn ngữ mạng xã hội, TF-IDF; DistilBERT, Linear SVM, phân loại người dùng.

1. Đặt vấn đề

Rối loạn tăng động giảm chú ý (ADHD) là một rối loạn phát triển thần kinh thường khởi phát từ thời thơ ấu nhưng có thể kéo dài đến tuổi trưởng thành. ADHD đặc trưng bởi 3 nhóm biểu hiện chính là thiếu chú ý, tăng động và xung động, đồng thời có thể gây ảnh hưởng rõ rệt đến học tập, công việc, các mối quan hệ xã hội và chất lượng sống. Các nghiên cứu dịch tễ cho thấy, đây là một trong những rối loạn phát triển thần kinh phổ biến nhất hiện nay. Trên phạm vi toàn cầu, tỷ lệ hiện mắc ADHD ở trẻ em và thanh thiếu niên được ghi nhận ở mức đáng kể [1], trong khi một phân tích tổng hợp khác ước tính tỷ lệ gộp khoảng 7,2% [2]. Ở người trưởng thành, ADHD vẫn tiếp tục hiện diện với tỷ lệ khoảng 4,4%, đi kèm mức độ đồng mắc và suy giảm chức năng không nhỏ [3].

Cùng với sự phát triển của môi trường số, mạng xã hội ngày càng trở thành nơi con người bộc lộ suy nghĩ, cảm xúc và thói quen giao tiếp trong đời sống thường ngày. Điều này khiến dữ liệu ngôn ngữ trên mạng xã hội trở thành một nguồn tư liệu có giá trị cho các nghiên cứu sức khỏe tâm thần. Một số công trình trước đây đã ghi nhận mối liên hệ giữa việc sử dụng phương tiện số với các biểu hiện liên quan đến chú ý, tăng động và xung động [4], đồng thời cho thấy mức độ sử dụng digital media cao có thể đi kèm với nguy cơ xuất hiện các triệu chứng ADHD cao

hơn theo thời gian ở thanh thiếu niên [5]. Từ góc độ phương pháp, lĩnh vực computational mental health cũng cho thấy, ngôn ngữ trực tuyến có thể phản ánh trạng thái tâm lý và cảm xúc của người dùng, từ đó hỗ trợ xây dựng các mô hình nhận diện vấn đề sức khỏe tâm thần dựa trên dữ liệu văn bản [6], [7].

Tuy nhiên, phần lớn các nghiên cứu trong hướng này tập trung vào trầm cảm, lo âu hoặc ý nghĩ tự sát, trong khi ADHD vẫn còn là một chủ đề tương đối ít được khai thác dưới góc nhìn xử lý ngôn ngữ tự nhiên. Những tiến bộ của các mô hình ngôn ngữ tiền huấn luyện như BERT đã mở ra khả năng nắm bắt ngữ cảnh văn bản tốt hơn trong nhiều bài toán NLP [8], và một số nghiên cứu gần đây cũng đã so sánh mô hình học máy truyền thống với Transformer trong phát hiện vấn đề sức khỏe tâm thần từ mạng xã hội [9]. Đối với ADHD, các công trình gần hơn với đề tài này cho thấy 2 hướng nghiên cứu nổi bật là khai thác dữ liệu người dùng ở mức hồ sơ, tiêu biểu như bộ Twitter-STMHD [10], và phân tích đặc điểm ngôn ngữ của cộng đồng ADHD trên mạng xã hội [11]. Gần đây hơn, Wiechmann và cộng sự đã đi thẳng vào bài toán phát hiện ADHD từ dữ liệu mạng xã hội và chỉ ra đây là một hướng có tiềm năng, nhưng vẫn còn nhiều vấn đề cần tiếp tục làm rõ về khả năng diễn giải và tính khái quát của mô hình [12].

Từ những cơ sở trên có thể thấy, việc nhận diện dấu hiệu ADHD qua ngôn ngữ mạng xã hội là một hướng nghiên cứu có ý nghĩa cả về mặt lý thuyết lẫn ứng dụng. Dù vậy, khoảng trống hiện nay vẫn nằm ở chỗ số lượng nghiên cứu chuyên biệt cho ADHD còn hạn chế, đặc biệt là các nghiên cứu so sánh trực tiếp giữa mô hình học máy truyền thống và mô hình học sâu trên dữ liệu ở mức người dùng. Xuất phát từ đó, nghiên cứu này tập trung vào bài toán nhận diện người dùng có dấu hiệu ADHD thông qua ngôn ngữ mạng xã hội, sử dụng dữ liệu từ Twitter-STMHD [10] và đối chiếu hiệu quả giữa các mô hình dựa trên TF-IDF với mô hình DistilBERT. Mục tiêu của nghiên cứu không nhằm thay thế chẩn đoán lâm sàng, mà hướng đến việc khảo sát khả năng khai thác tín hiệu ngôn ngữ như một nguồn thông tin hỗ trợ cho sàng lọc và nghiên cứu ADHD trong bối cảnh sức khỏe tâm thần tính toán.

2. Phương pháp nghiên cứu

2.1. Dữ liệu

Nghiên cứu sử dụng một tập con của bộ dữ liệu Twitter-STMHD [10]. Đây là bộ dữ liệu ở mức người dùng, được xây dựng cho các nghiên cứu về sức khỏe tâm thần trên mạng xã hội. Ở phiên bản gốc, Twitter-STMHD gồm 8 nhóm rối loạn tâm thần và một nhóm đối chứng; với mỗi người dùng, dữ liệu được lưu dưới 2 dạng chính là user-profile data và timeline-tweets data [10]. Trong phạm vi nghiên cứu này, chỉ 2 nhóm ADHD và đối chứng (NEG) được trích ra để xây dựng bài toán phân loại nhị phân. Từ mỗi nhóm, nghiên cứu chọn ngẫu nhiên 1000 người dùng nhằm giữ cân bằng giữa 2 lớp. Phần dữ liệu được sử dụng làm đầu vào cho mô hình là timeline-tweets, tức toàn bộ nội dung bài đăng của người dùng theo thời gian. Cách tổ chức này phù hợp với mục tiêu nhận diện dấu hiệu ADHD ở cấp độ người dùng, thay vì suy luận từ từng bài đăng riêng lẻ.

2.2. Tiền xử lý dữ liệu

Dữ liệu được xử lý ở mức người dùng. Với mỗi tài khoản, nghiên cứu trích xuất nội dung bài đăng từ tệp tweets.json để tạo dữ liệu văn bản đầu vào; đồng thời một số thông tin mô tả trong user.json, như số người theo dõi, số bạn bè và tổng số bài đăng, cũng được lưu lại để phục vụ thống kê. Văn bản sau đó được làm sạch bằng cách thay ký tự xuống dòng bằng khoảng trắng, loại bỏ URL và tên người dùng dạng @, bỏ ký hiệu # nhưng giữ lại từ khóa đi kèm, rồi chuẩn hóa khoảng trắng. Những bài đăng quá ngắn, trùng lặp hoặc không còn nội dung hợp lệ sau làm sạch được loại bỏ. Sau bước này, tối đa 200 bài đăng của mỗi người dùng được giữ lại và nối thành một văn bản duy nhất để biểu diễn cho tài khoản đó. Kết quả sau sàng lọc cho thấy, tập dữ liệu còn 1.999 người dùng, gồm 999 mẫu ADHD và 1.000 mẫu đối chứng. (Bảng 1)

2.3. Biểu diễn dữ liệu

Sau bước tiền xử lý, toàn bộ bài đăng của mỗi người dùng được nối thành một văn bản duy nhất để tạo mẫu đầu vào ở mức người dùng. Trên cơ sở đó, nghiên cứu sử dụng 2 cách biểu diễn dữ liệu tương ứng với 2 hướng tiếp cận khác nhau. Với các mô hình học máy truyền thống, văn bản được chuyển thành vector TF-IDF dựa trên unigram và bigram

Bảng 1. Thống kê mô tả tập dữ liệu sau tiền xử lý ban đầu

Nhóm	Số lượng người dùng	Số bài đăng trung bình/người dùng	Trung vị số bài đăng/người dùng	Độ dài văn bản trung bình (ký tự)	Trung vị độ dài văn bản (ký tự)
ADHD	999	195	200	17,958	16,887
NEG	1,000	194	200	15,357	13,636
Tổng	1,999	194	200	16,657	15,101

nhằm giữ lại thông tin từ vựng và các cụm từ xuất hiện thường xuyên trong dữ liệu. Lựa chọn này phù hợp với các nghiên cứu về sức khỏe tâm thần trên mạng xã hội, nơi các đặc trưng như bag-of-words, n-gram và TF-IDF vẫn được sử dụng khá phổ biến trong các bài toán phân loại văn bản [6]. Với mô hình học sâu, văn bản được mã hóa bằng tokenizer của DistilBERT trước khi đưa vào quá trình fine-tuning. Cách biểu diễn này cho phép mô hình khai thác thông tin ngữ cảnh tốt hơn nhờ kế thừa cơ chế biểu diễn ngôn ngữ của họ BERT [8]. Việc đặt 2 cách biểu diễn trong cùng một thiết lập giúp nghiên cứu so sánh trực tiếp giữa đặc trưng văn bản truyền thống và biểu diễn ngữ cảnh từ mô hình ngôn ngữ tiền huấn luyện.

2.4. Mô hình nghiên cứu

Để đánh giá hiệu quả của các hướng tiếp cận khác nhau trong bài toán nhận diện người dùng có dấu hiệu ADHD, nghiên cứu sử dụng 3 mô hình gồm Logistic Regression, Linear SVM và DistilBERT. Trong đó, Logistic Regression và Linear SVM được lựa chọn làm các mô hình cơ sở khi dữ liệu được biểu diễn bằng TF-IDF, vì đây là những mô hình có cấu trúc đơn giản, dễ triển khai và vẫn cho kết quả ổn định trong nhiều bài toán phân loại văn bản liên quan đến sức khỏe tâm thần trên mạng xã hội [6]. Bên cạnh đó, DistilBERT được sử dụng như một mô hình ngôn ngữ tiền huấn luyện thuộc họ Transformer, cho phép khai thác thông tin ngữ cảnh của văn bản tốt hơn so với các cách biểu diễn đặc trưng truyền thống [8]. Việc đặt 3 mô hình này trong cùng một thiết lập thực nghiệm giúp nghiên cứu so sánh trực tiếp giữa 2 hướng tiếp cận: mô hình học máy truyền thống dựa trên đặc trưng văn bản và mô hình học sâu dựa trên biểu diễn ngữ cảnh. Cách so sánh này cũng phù hợp với xu hướng gần đây trong các nghiên cứu phát hiện ADHD từ dữ liệu mạng xã hội, nơi hiệu quả dự đoán cần được xem xét đồng thời với khả năng diễn giải và tính phù hợp của mô hình đối với dữ liệu ngôn ngữ ở mức người dùng [12].

2.5. Cấu hình huấn luyện và đánh giá

Tập dữ liệu được chia theo tỷ lệ 70/15/15 cho 3 phần: huấn luyện, xác thực và kiểm thử, đồng thời vẫn giữ sự cân bằng tương đối giữa 2 lớp. Trong 2 mô hình Logistic Regression và Linear SVM được huấn luyện trên tập đặc trưng TF-IDF. Với DistilBERT, mô hình được fine-tune với batch size 16, learning rate 2e-5, 3 epoch và weight decay 0.01. Đây là cấu hình nằm trong khoảng thường được sử dụng cho các mô hình thuộc họ BERT trong các bài toán phân loại văn bản [8]. Mô hình tốt nhất được chọn dựa trên F1-score trên tập xác thực. Hiệu quả cuối cùng được đánh giá trên tập kiểm thử thông qua các chỉ số Accuracy, Precision, Recall và F1-score. Ngoài ra, ma trận nhầm lẫn cũng được sử dụng để quan sát rõ hơn khả năng phân biệt giữa 2 lớp.

3. Kết quả và thảo luận

3.1. Kết quả thực nghiệm

Kết quả trên tập kiểm thử cho thấy, cả 3 mô hình đều phân biệt được 2 nhóm người dùng ADHD và đối chứng, nhưng mức hiệu quả có sự khác biệt rõ rệt. Trong số đó, Linear SVM cho kết quả tốt nhất với Accuracy = 0.8733, Precision = 0.8294, Recall = 0.9400 và F1-score = 0.8812. Logistic Regression đạt kết quả khá sát với F1-score = 0.8731, trong khi DistilBERT thấp hơn với F1-score = 0.8119. Nhìn chung, kết quả này cho thấy trong thiết lập hiện tại, các mô hình truyền thống dựa trên TF-IDF vẫn hoạt động hiệu quả hơn mô hình học sâu.

Với DistilBERT, quá trình huấn luyện cho thấy train loss giảm dần, còn F1-score trên tập xác thực

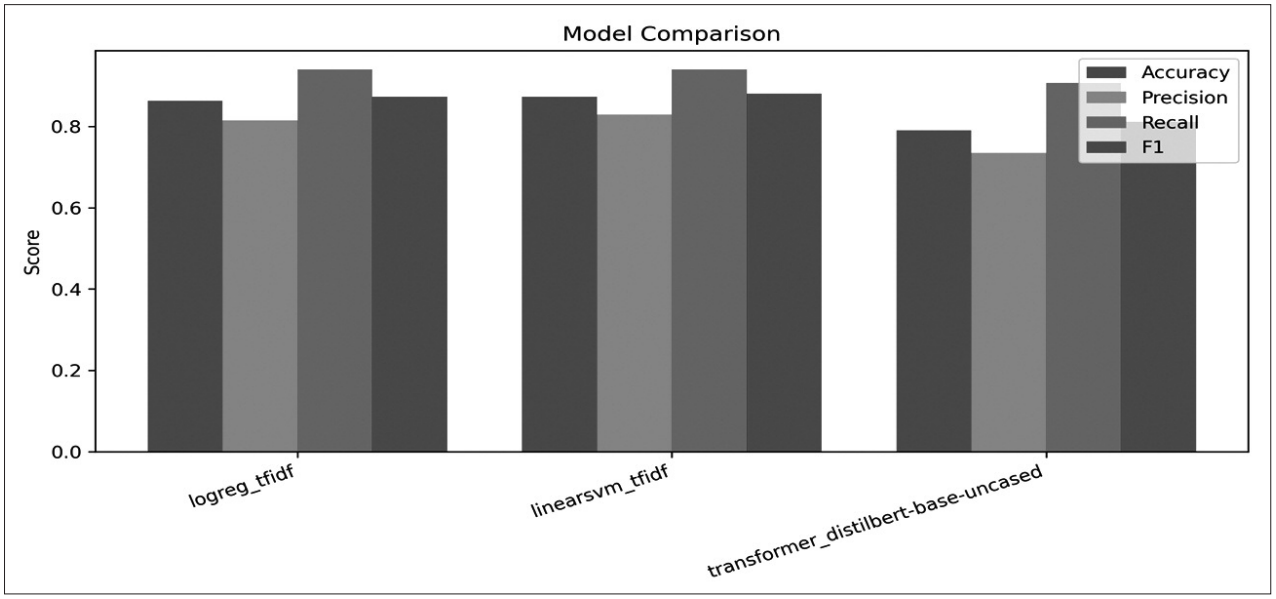
tăng qua từng epoch, chứng tỏ mô hình vẫn học được các tín hiệu phân biệt giữa 2 lớp. Tuy vậy, kết quả trên tập kiểm thử chưa vượt được 2 mô hình còn lại.

Ma trận nhầm lẫn cũng cho thấy Linear SVM và Logistic Regression giữ được sự cân bằng tốt hơn giữa 2 lớp. Trong khi đó, DistilBERT nhận diện lớp ADHD khá tốt nhưng nhầm nhiều mẫu đối chứng sang ADHD hơn, dẫn đến precision thấp hơn. Điều này cho thấy mô hình có xu hướng ưu tiên phát hiện lớp ADHD, đổi lại số trường hợp dương tính giả tăng lên.

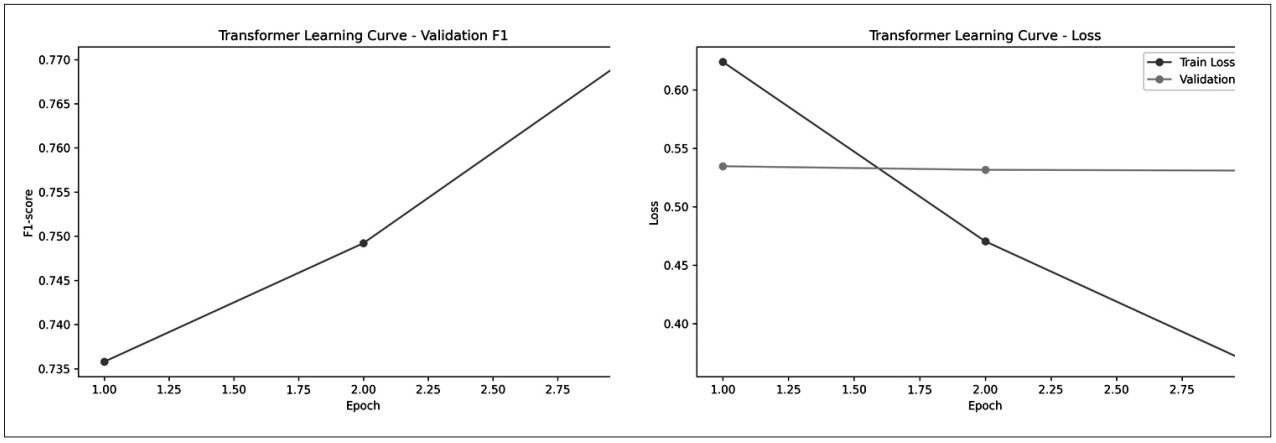
3.2. Thảo luận

Kết quả thực nghiệm cho thấy các mô hình dựa trên TF-IDF vẫn hoạt động hiệu quả trong bài toán nhận diện người dùng có dấu hiệu ADHD, đặc biệt là Linear SVM. Điều này cho thấy ở mức người dùng, các tín hiệu từ vựng và cụm từ lặp lại trong nhiều bài đăng vẫn có giá trị phân biệt rõ giữa 2 nhóm. Trong khi đó, DistilBERT cho thấy quá trình huấn luyện khá ổn định, nhưng kết quả trên tập kiểm thử vẫn thấp hơn 2 mô hình truyền thống. Từ ma trận nhầm lẫn có thể thấy mô hình này nhận diện lớp ADHD tương đối tốt, nhưng nhầm nhiều mẫu đối chứng sang ADHD hơn, làm giảm precision. Nhìn chung, kết quả cho thấy dữ liệu ngôn ngữ trên mạng xã hội có thể hỗ trợ nhận diện dấu hiệu ADHD ở cấp độ người dùng; tuy vậy, mô hình chỉ nên được xem như công cụ hỗ trợ sàng lọc, không thay thế cho chẩn đoán lâm sàng.

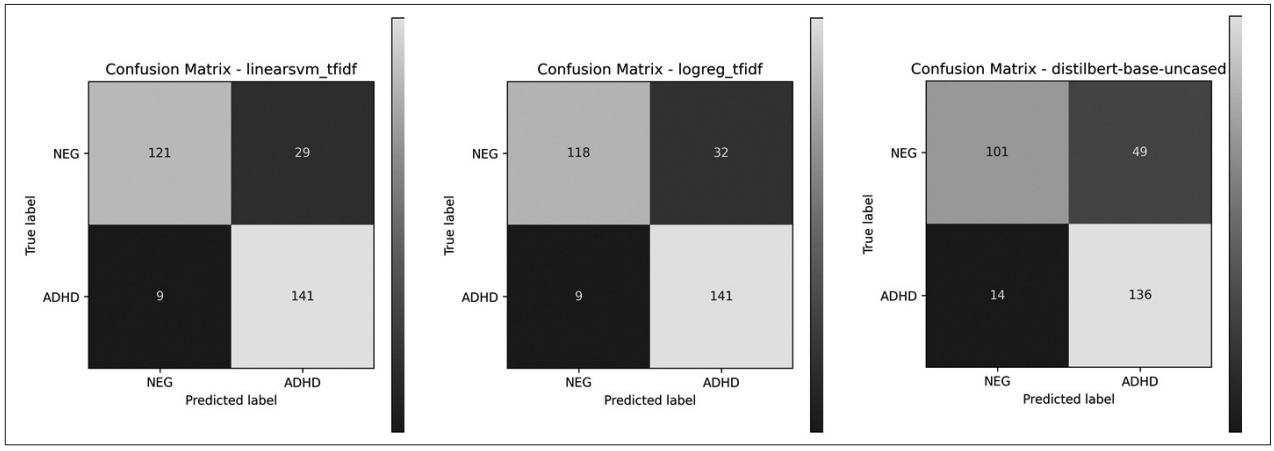
Hình 1: Kết quả so sánh các mô hình



Hình 2. Đường cong huấn luyện của DistilBERT



Hình 3. Ma trận nhầm lẫn của các mô hình



4. Kết luận

Nghiên cứu này tập trung vào bài toán nhận diện người dùng có dấu hiệu ADHD thông qua ngôn ngữ trên mạng xã hội bằng cách so sánh 3 mô hình gồm Logistic Regression, Linear SVM và DistilBERT. Kết quả thực nghiệm cho thấy, cả 3 mô hình đều có khả năng phân biệt giữa nhóm ADHD và nhóm đối chứng, trong đó Linear SVM đạt hiệu quả tốt nhất trong thiết lập hiện tại. Kết quả này cho thấy, với dữ liệu văn bản ở mức người dùng, biểu diễn TF-IDF vẫn là một lựa chọn đơn giản nhưng hiệu quả.

Bên cạnh đó, DistilBERT vẫn cho thấy khả năng học được các tín hiệu ngôn ngữ liên quan đến ADHD, dù kết quả trên tập kiểm thử chưa vượt các mô hình truyền thống. Nhìn chung, nghiên cứu cho thấy dữ liệu ngôn ngữ trên mạng xã hội có thể cung cấp thông tin hữu ích cho bài toán nhận diện dấu hiệu ADHD ở cấp độ người dùng. Tuy vậy, kết quả của mô hình nên được hiểu như một công cụ hỗ trợ sàng lọc và phân tích tín hiệu ngôn ngữ, không thay thế cho chẩn đoán lâm sàng. Trong các nghiên cứu tiếp theo, có thể mở rộng quy mô dữ liệu và thử nghiệm thêm các mô hình phù hợp hơn với văn bản dài để cải thiện hiệu quả phát hiện ■

TÀI LIỆU TRÍCH DẪN VÀ THAM KHẢO:

[1] Polanczyk G., de Lima M. S., Horta B. L. et al. (2007). The worldwide prevalence of ADHD: A systematic review and metaregression analysis. American Journal of Psychiatry, 164(6), 942-948.

[2] Thomas R., Sanders S., Doust J. et al. (2015). Prevalence of attention-deficit/hyperactivity disorder: A systematic review and meta-analysis. Pediatrics, 135(4), e994-e1001. DOI: 10.1542/peds.2014-3482.

[3] Kessler R. C., Adler L., Barkley R. et al. (2006). The prevalence and correlates of adult ADHD in the United States: Results from the National Comorbidity Survey Replication. American Journal of Psychiatry, 163(4), 716-723. DOI: 10.1176/appi.ajp.163.4.716.

[4] Beyens I., Valkenburg P. M. (2022). Children's media use and its relation to attention, hyperactivity, and impulsivity. In D. Lemish (Ed.), The Routledge international handbook of children, adolescents, and media (2nd ed., pp. 202–210). Routledge. DOI: 10.4324/9781003118824-26.

[5] Ra C. K., Cho J., Stone M. D. et al. (2018). Association of digital media use with subsequent symptoms of attention-deficit/hyperactivity disorder among adolescents. JAMA, 320(3), 255-263. DOI: 10.1001/jama.2018.8931.

[6] Chancellor S., De Choudhury M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. NPJ Digital Medicine, 3, Article 43. DOI: 10.1038/s41746-020-0233-7.

[7] Resnik P., Armstrong W., Claudino L. et al. (2015). Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 99-107). Association for Computational

Linguistics.

[8] Devlin J., Chang M. W., Lee K. et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding [Preprint]. arXiv.

[9] Bokolo B. G., Liu Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. Electronics, 12(21), 4396. DOI: 10.3390/electronics12214396.

[10] Suhavi Singh, A. K. Arora U. et al. (2022). Twitter-STMHD: An extensive user-level database of multiple mental health disorders. Proceedings of the International AAAI Conference on Web and Social Media, 16, 1182-1191.

[11] Kalantari N., Payandeh A., Zampieri M. et al. (2023). Understanding the language of ADHD and autism communities on social media. In 2023 IEEE International Conference on Big Data (BigData) (pp. 2188-2195). IEEE. DOI: 10.1109/BigData59044.2023.10386833.

[12] Wiechmann D., Kempa E., Kerz E., et al. (2024). Transparent but powerful: Explainability, accuracy, and generalizability in ADHD detection from social media data [Preprint]. arXiv.

Ngày nhận bài: 15/01/2026
Ngày phản biện đánh giá và sửa chữa: 6/02/2026
Ngày chấp nhận đăng bài: 28/02/2026

DETECTING ADHD-RELATED SIGNALS IN SOCIAL MEDIA LANGUAGE: A COMPARATIVE STUDY OF TF-IDF-BASED MODELS AND DISTILBERT

● **NGUYEN THI VI HAI**[1]
● **NGUYEN THI HUYEN TRANG**[2]
[1]Ho Chi Minh City University of Technology
[2]Ho Chi Minh City University of Industry and Trade

ABSTRACT:

This study explores the potential for identifying individuals exhibiting signs of Attention Deficit Hyperactivity Disorder (ADHD) through linguistic patterns on social media. The analysis utilizes the Twitter-STMHD dataset, comprising 1,999 pre-processed user-level samples. A comparative evaluation was conducted across three classification models: Logistic Regression, Linear Support Vector Machine (SVM), and DistilBERT. The findings demonstrate that all models effectively distinguish between ADHD and control groups, with the Linear SVM achieving the strongest performance, attaining an accuracy of 0.8733 and an F1-score of 0.8812. These results underscore the potential of language-based signals on social media as a supplementary tool for ADHD screening, while reaffirming that such approaches cannot substitute for formal clinical diagnosis.

Keywords: ADHD, social media language, TF-IDF, DistilBERT, Linear SVM, user classification.