frontiers | Frontiers in Child and Adolescent Psychiatry

TYPE Original Research
PUBLISHED 09 May 2025
DOI 10.3389/frcha.2025.1519753

OPEN ACCESS

EDITED BY
Jonathan Payne,
Murdoch Children's Research Institute, Australia

REVIEWED BY
Valeska Berg,
Curtin University, Australia
Yasemin Mehmed,
The University of Melbourne, Australia

*CORRESPONDENCE
Juan Barrios
✉ juan.barrios@unige.ch

RECEIVED 30 October 2024
ACCEPTED 17 April 2025
PUBLISHED 09 May 2025

CITATION
Barrios J, Poznyak E, Lee Samson J, Rafi H, Gabay S, Cafiero F and Debbané M (2025) Detecting ADHD through natural language processing and stylometric analysis of adolescent narratives.
Front. Child Adolesc. Psychiatry 4:1519753.
doi: 10.3389/frcha.2025.1519753

Detecting ADHD through natural language processing and stylometric analysis of adolescent narratives

Juan Barrios[1]*, Elena Poznyak[1], Jessica Lee Samson[1], Halima Rafi[1], Simon Gabay[2], Florian Cafiero[3] and Martin Debbané[1,4]

[1]Developmental Clinical Psychology Research Unit, Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland, [2]Faculty of Humanities, University of Geneva, Geneva, Switzerland, [3]École Nationale des Chartes, PSL, Paris, France, [4]Research Department of Clinical, Educational and Health Psychology, University College London, London, United Kingdom

Introduction: Attention-Deficit/Hyperactivity Disorder (ADHD) significantly affects adolescents' everyday lives, particularly in emotion regulation and interpersonal relationships. Despite its high prevalence, ADHD remains underdiagnosed, highlighting the need for improved diagnostic tools. This study explores, for the first time, the potential of Natural Language Processing (NLP) and stylometry to identify linguistic markers within Self-Defining Memories (SDMs) of adolescents with ADHD and to evaluate their utility in detecting the disorder. A further novel aspect of this research is the use of SDMs as a linguistic dataset, which reveals meaningful patterns while engaging psychological processes related to identity and memory.

Method: Our objectives were to: (1) characterize linguistic features of SDMs in ADHD and control groups; (2) assess the predictive power of stylometry in classifying participants' narratives as belonging to either the ADHD or control group; and (3) conduct a qualitative analysis of key linguistic markers of each group. Sixty-six adolescents (25 diagnosed with ADHD and 41 typically developing peers) recounted SDMs in a semi-structured format; these narratives were transcribed for analysis. Stylometric features were extracted and used to train a Support Vector Machine (SVM) classifier to distinguish between narratives from the ADHD and control groups. Linguistic metrics such as wordcount, lexical diversity, lexical density, and cohesion were computed and analyzed. A qualitative analysis was also applied to examine stylistic patterns in the narratives.

Results: Adolescents with ADHD produced narratives that were shorter, less lexically diverse, and less cohesive. Stylometric analysis using an SVM classifier distinguished between ADHD and control groups with up to 100% precision. Distinct linguistic markers were identified, potentially reflecting difficulties in emotion regulation.

Discussion: These findings suggest that NLP and stylometry can enhance ADHD diagnostics by providing objective linguistic markers, thereby improving both its understanding and diagnostic procedures. Further research is needed to validate these methods in larger and more diverse populations.

KEYWORDS

ADHD, self-defining memories, emotion regulation, natural language processing, stylometry, computational linguistics

Introduction

ADHD is a neurodevelopmental disorder that affects between 5.6% and 8% of youth between 12 to 18 years (1, 2). Its causes are multifactorial: several genetic and environmental risk factors act together to increase susceptibility to this disorder and other psychiatric comorbidities (3). During adolescence, individuals with ADHD are particularly prone to challenges in emotion regulation and in interpersonal relationships (4). A significant characteristic of these difficulties is their tendency to combine and mutually reinforce. For instance, difficulties within the social domain are often exacerbated and amplified by deficiencies in emotional regulation and pragmatic language abilities (5). These issues often result in low self-esteem, social problems, increased risks of substance abuse (6), peer rejection, social isolation, or academic failure (7).

Despite its high prevalence and impact, ADHD is still relatively underdiagnosed in most countries (8). Additionally, it is well known that early detection and treatment of ADHD are effective strategies for managing its course and mitigating long-term impacts (9). Focusing on detection during adolescence offers therefore an opportunity to facilitate earlier interventions, reduce risks, and improve outcomes. However diagnosing ADHD is a complex and time-consuming process requiring a comprehensive and multidisciplinary assessment (10). This involves evaluating clinical history, using standardized rating scales, gathering school information, and applying DSM-5 or ICD-10 criteria to assess symptoms across various contexts while excluding other mental disorders with overlapping symptoms (3). Along with that, the effectiveness of ADHD standardized rating scales can be compromised by informant biases (11), where different respondents (e.g., parents, teachers, or individuals themselves) may rate the same behaviors differently due to subjective viewpoints or contextual experiences [cf. (12)]. This variability can lead to inaccurate assessments of ADHD symptoms (11) and underscores the importance of using both subjective measures (e.g., self-reports and observer ratings) and objective measures (e.g., Continuous Performance Test) to improve diagnostic accuracy and reliability (13, 14).

With the rise of Artificial Intelligence (AI) in the last decade, other objective methods to gather information have emerged like computer-based linguistic methods from the field of Natural Language Processing (NLP). These approaches offer objective measures through the analysis of text and speech features, thereby mitigating the inherent subjectivity of traditional standardized rating scales (15), with reduced implications in terms of time, cost, and infrastructure required for its deployment. For more than ten years now, NLP has been used in neuroscience and psychiatry [cf. for example, (16–19)]. In fact, the latest research in this field makes it possible not only to identify mental health risks like suicidal risk behaviour (20) but can also contribute to predict the onset of mental disorders on a linguistic basis (21–23). Considering these facts, one can say that AI and NLP techniques have emerged as robust tools for various clinical applications, demonstrating their efficacy in diverse contexts. With their ability to analyze and interpret linguistic patterns and to capture and process language use in different contexts, AI and NLP offer a direct approach to gather clinical insights that hold promise in revolutionizing the screening and diagnostic processes for ADHD and other mental-health disorders.

Research at the intersection of NLP and psychology extends beyond English to analyze various languages such as Spanish (22, 24), Chinese (25), French (26), and Korean (27, 28). Two Korean language studies analyzed the language patterns of individuals with ADHD across different age groups and contexts. The first study (27) compared language use in children diagnosed with combined ADHD with a non-clinical control group. Results showed significant between-group differences in word use and language style of both groups thereby highlighting possible distinctive linguistic markers for combined ADHD in childhood. Building on these preliminary findings, the second study (28) extended the investigation to Korean college students with ADHD symptoms and revealed the persistence of a distinct language style associated with ADHD across different developmental stages.

In the context of the present study, these findings are fundamental: if a specific language style of ADHD can be detected, then stylometric methods have the potential to detect it in other languages too and, therefore, enhance the accuracy and efficiency of preliminary screenings and/or diagnoses of ADHD (29–32) and contribute to simplify its procedure. Indeed, stylometry has its own techniques specifically oriented to detect language styles and has already proven to be very efficient in literature (33), forensic science (34) or social media studies (35). Furthermore, it also gives researchers the opportunity to examine stylometric markers in relation to clinical conditions (36).

Self-defining memories (SDMs) represent a specific type of autobiographical memory associated with the self-concept (37, 38), contributing to an individual's sense of coherence and of continuity (39). Consequently, SDMs are conceptualized as fundamental components of personal identity at cognitive, motivational, and affective levels. SDMs are particularly useful in the fields of stylometry and NLP for detecting mental health disorders because they often encapsulate significant emotional experiences, making them rich sources of data. By examining the language used in these narratives, researchers can gain insights into an individual's cognitive and emotional processes (40). In stylometry, SDMs allow for the examination of linguistic patterns and their relationship with psychological distress or well-being. Furthermore, SDMs provide a consistent and structured format for collecting data across different individuals by means of the SDMs Task and thus enhancing the reliability and validity of the research findings.

This study has three primary objectives: our first aim is to employ NLP to compare the linguistic features of SDMs in adolescents with ADHD to typically developing peers. Our second goal is to quantify the predictive power of stylometry for group classification, distinguishing between controls and ADHD individuals. Finally, our third objective is to conduct a qualitative analysis of the key linguistic markers identified. The latter is crucial for several reasons. On the one hand, qualitative analysis

offers a deeper understanding of the context and nuances of language use, which purely quantitative methods might overlook (41). On the other hand, qualitative insights can help validate and interpret the results of quantitative analyses, ensuring that identified markers are not only statistically significant but also meaningful and relevant in real-world settings (42). This can increase the potential for these markers to be used in diagnostic and screening processes.

Based on the findings of the Korean studies mentioned above (27, 28), which identified a specific narrative style associated with individuals prone to ADHD, we hypothesized that this distinct narrative style would be detectable in the self-defining narratives of adolescents. Specifically, we expected that advanced stylometric techniques would unveil narrative patterns (i.e., markers) in individuals' SDMs with ADHD.

Materials and methods

Participants

25 Adolescents with ADHD (ADHD group, 12 females and 13 males) and 41 without ADHD (control group, 22 females and 19 males) were included in the study. All adolescents were between 12 and 17 years old and fluent in French. For both groups, the exclusion criteria included a history of psychotic disorders, diagnosed personality disorder, autism spectrum disorder, or neurological disorders. In the ADHD group, these criteria were assessed during the clinical intake through an anamnestic interview conducted with the parents. In the control group, exclusion criteria were explicitly screened prior to participation through a standardized pre-task questionnaire.

ADHD group

Adolescents with ADHD were recruited as part of a research project conducted at the University of Geneva advertised in local parent associations for children with ADHD and through collaborations established with local child psychiatrists. Diagnostic criteria were investigated by detailed anamnestic interviews and confirmed using the ADHD Child Evaluation interview (43). All diagnostic assessments were conducted by experienced clinical psychologists specialized in ADHD.

Control group

Typically developing (TD) controls were recruited from the general population in Geneva through advertisements and personal referrals and received compensation for their participation. Specific inclusion criteria for this group required the absence of intellectual impairments, as assessed by the Block Design and Vocabulary subtests of the Wechsler Intelligence Scale for Children [WISC-IV; (44)].

Of note, no significant differences were found between the ADHD and control groups on the WISC-IV Block Design subtest, suggesting comparable performance between groups on these core measures of visuospatial reasoning and verbal comprehension (ADHD: $M = 10.73$, $SD = 2.74$; Control: $M = 11.85$, $SD = 2.43$; $W = 406.5$, $p = 0.141$) or the Vocabulary subtest (ADHD: $M = 12.27$, $SD = 2.49$; Control: $M = 12.31$, $SD = 3.03$; $W = 515.5$, $p = 0.957$).

Sampling

Pairwise matching

Pairwise matching was used to balance the ADHD and the control groups with respect to the means of participants' sex and age in order to make them comparable and get a more fine-grained analysis of the differences between the two groups. The two samples were matched by cardinality method (45) to find the largest matched set (in this case by age and sex) with the additional constraint that the ratio between the number of adolescents in both groups had to be equal to 1. This method minimizes between-group differences based on age or sex, while selecting the best-fitting control cases for the ADHD group with minimal loss of ADHD cases.

Final samples

In both groups the final sample meeting the pairwise inclusion criteria consisted of 24 adolescents in both groups (cf. Table 1). The mean age in the ADHD group was 15.14 (σ = 1.83) and 15.21 (σ = 1.44) for the control group. A Student two samples t-test showed that the difference was statistically not significant ($t(43.65) = -0.16$, p-value > 0.05). As a result of this pairwise matching, the total number of SDMs per group was 72, derived from 24 participants, each contributing with 3 SDMs.

Data

Self-defining memories (SDM)

The SDMs investigated in this study were collected with the SDM Task[1] (37). During this task, participants were asked to recall personal memories of events with specific attributes, which will be described next. Participants were asked to write three SDMs. They were told that SDMs refers to important life events that occurred at least one year ago and helped them to understand who they are. Other characteristics of SDMs were also given to the participants: SDMs are generally vividly represented and meaningful, they generate strong feelings (positive or negative) and are often recalled by individuals on a voluntary basis or spontaneously. While listening to the description, participants had a printed summary in front of them outlining the main points of the task. Then they were told to imagine a situation where they meet someone they are fond of during a walk to share several personal past events that powerfully convey how they have become who they are today. The participants were then given

[1]Thorne A, McLean KC. Manual for Coding Events in Self-De!ning Memories. Santa Cruz: University of California (2001).

TABLE 1 ADHD Characteristics of participants after pairwise matching.

Variable	ADHD (mean/count)	ADHD (SD/%)
Primary symptoms		
Inatention	7.61	1.2
Hyperactivity	4.61	2.78
ACE diagnosis		
ADHD-inattentive	15	62.5%
ADHD-hyperactive	1	4.2%
ADHD-combined	8	33.3%
Comorbities		
Language and/or Learning disorders	4	16.7%
Anxiety	2	8.3%
Conduct disorder	1	4.2%
Sleep disorder	1	4.2%
Medication		
On medication	9	36%

three sheets of paper and asked to write down one SDM on each sheet. For each event, participants were asked to write a one sentence summary and a sufficiently detailed description to help the imagined friend see and feel as they did in the past. Afterward, participants rated their feelings after recalling each memory using a 7-point scale from 0 (not at all) to 6 (extremely) for positive and negative affects. Finally, they estimated how much time had passed since each event in years and months, thus providing a self-reported measure of the time frame for each SDM.

Final corpus

The corpus is made up of a series of 144 SDMs, collected in two samples, one consisting of 24 adolescents with a diagnosis of ADHD and the other with 24 control participants, who all had to write a total of three SDMs each. All texts were handwritten in French, and then transcribed by the person in charge of the experiment. Spelling mistakes were corrected at the time of entry, allowing the machine to focus on the deep structure of the language, and not on the vagaries of its form.

Linguistic metrics

Among the various metrics used in NLP to analyze and quantify aspects of text in a psychological perspective, we chose word count, lexical density, lexical diversity and cohesion [cf. (40, 46–48)].

Wordcount

Is a quantitative measure of the number of tokens (: a computational approximation of the linguistic concept of "word," defined as a string of characters between two delimiters, such as punctuation or spaces) present in a given text or set of texts. Depending on the definition of a token, this means that we count the number of words, punctuation marks, and/or symbols in a text (49).

In the context of this study, we operationalize wordcount using the definition of the token of the UDPipe package (50). The delimiters are spaces, tabulations, newlines, punctuation signs, apostrophes (*J'ai* = 2 tokens) and hyphens (*beau-père* = 2 tokens). The resulting wordcount is the total count of the number of units obtained after this tokenization process.

$$\text{Wordcount} = \sum_{i=x}^{N} x \tag{1}$$

Where:

N = Total number of tokens in the text

x = 1 token.

Lexical diversity

Refers to the variety of unique lexical items (types) used in a text (51) and serves as a metric for assessing the language proficiency of individuals. In psychological research, lexical diversity has been studied across various settings. For example, (52) explored Lexical Diversity among three groups of preschool-aged children (aged 3 to 5 years): (1) children with Specific Language Impairment (SLI) (2) age-matched children with Typical Development (TD) and (3) language-matched children with typical development (LM). Watkins et al. (52) findings consistently distinguished children with SLI from their TD and LM peers. Similarly, (53) investigated lexical diversity in children older than those of Watkins study (aged 5 to 8 years) in three groups: (1) children with ADHD, (2) children with SLI, and (3) typically developing children (TD). The study found that children with SLI demonstrated lower lexical diversity compared to both the ADHD and TD groups. Conversely, no significant difference in lexical diversity was found between the ADHD and TD groups.

In this study, we use the Moving-Average Type-Token Ratio (MATTR) method to assess the lexical diversity of the two groups. The Type-Token Ratio (TTR) is a traditional measure of lexical diversity, calculated by dividing the number of unique lexical items (types) by the total number of "words" (tokens) in a given text. The Moving-Average Type-Token Ratio (MATTR) refines this measure by calculating the TTR over a sliding window of fixed size, in this case 50 tokens, applied across the entire text. This window size is frequently used because it offers an optimal balance between measurement sensitivity and statistical reliability. The average of these windowed TTRs was then computed, providing a more reliable estimate of lexical diversity that accounted for variations in text length (54). This method improves accuracy and reliability by incorporating more data points than traditional methods for computing lexical diversity.

$$\text{MATTR} = \frac{1}{N - W + 1} \sum_{i=1}^{N-W+1} \frac{V_i}{W} \tag{2}$$

Where:

N = Total number of tokens in the text

W = Window size (number of tokens in each segment)

Vi = Number of unique tokens in the i-th segment.

We take the same mathematical definition of the lexical diversity [cf. Equation 2] that (52, 53).

Lexical density

Reflects the proportion of content words (nouns, verbs, adjectives, and adverbs) to the total number of tokens in a text (55). As such, it provide a measure of the amount of information in a given text (56). Lexical density evolves during human lifespan and it is influenced by different factors like emotional state (57) and the educational background (58). To calculate this network metric, content words are extracted from the corpus by filtering according to their part-of-speech annotations, which are generated using the udpipe R package (59).

In this study, we compute the Average Lexical Density (ALD) by segmenting the text, then we compute the ALD for each segment (55), and then we average all the results for each group (cf. Equation 3), ADHD and control.

$$\mathrm{ALD} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathrm{n(CW)in}\ S_i}{\mathrm{n(TW)in}\ S_i} \times 100\right) \tag{3}$$

Where:

CW = Content words

S = Segment

TW = Total number of words in the i-th segment.

Cohesion

Analysis in NLP examines how well connected different parts of a text are connected (i.e., the use of pronouns, conjunctions, and lexical repetition), which help to maintain continuity and facilitates communication (60). High cohesion in narratives is often associated with better mental health and cognitive functioning. For example, individuals who use cohesive devices effectively tend to exhibit more organized thought processes and better narrative skills, which are indicative of a well-integrated self-concept and adaptive functioning (40). Conversely, low cohesion can signal disorganized thinking, which is often seen in various psychological disorders such as schizophrenia or severe anxiety disorder, where individuals may struggle to maintain a clear and connected narrative (61, 62). Along with this, lower cohesion is often observed in individuals with ADHD, reflecting their attentional challenges and cognitive variability.

Textual cohesion is calculated for each participant by analyzing specific cohesive devices—namely, parts of speech that are known to enhance cohesion—pronouns, adverbs, determiners, and coordinating conjunctions [cf. (63)]. Cohesive devices are counted at the participant level (3 SDMs), not the text level (1 SDM), to obtain a Cohesion Score (CS) per person (cf. Equation 4).

$$\mathrm{CS}_{\mathrm{participant}} = \frac{1}{n_i}\sum_{j=1}^{n_i} T_{i,j} \tag{4}$$

Where:

n_i = Number of cohesive devices in document i

$T_{i,j}$ = Total number of cohesive devices for participant i in part-of-speech category j.

Subsequently, the mean and standard deviation for each group are calculated based on the individual scores of the participants.

Machine learning for textual analysis

We decided to use a classical machine learning method rather than deep learning, because according to a recent survey (64) the former performs better than the latter for profiling in similar settings (short texts, boolean or few categories). Because we were confident in ADHD diagnoses of participants, we decided to rely on a supervised method—otherwise, an unsupervised method would have been more appropriate (65). Among the supervised solutions available, we choose a standard algorithm in stylometry: Support Vector Machines (SVM) and not random forest (66) or logistic regression (67, 68), as SVM allows for easy interpretation and is established as a standard method in stylometry (69–72).

All participants' SDMs are collected in a single file for each group, with the exception of those written by two people from the control group and two from the ADHD group for a final blind test.

The analyses were implemented with the SuperStyl package (73), which has been used in previous studies to build stylistic profiles with very good results (35, 69). SuperStyl use internally the SVM and other pipeline facilities from scikit-learn (74).

Training a SVM (cf. Figure 1) involves three important steps: (1) selecting features, (2) tuning hyperparameters, and (3) understanding evaluation metrics (75).

Features

We test two features:

(1) Function words (FW) are words that carry minimal semantic meaning but play a grammatical role in the sentence, including articles, prepositions, conjunctions, and auxiliary verbs (40). They are extremely valuable because (i) their large number makes them particularly useful for statistical analyses, (ii) their use is done in a more unconscious way, (iii) they are not related to the content and, for instance, allow comparisons between texts with different themes (76). Function words have been found to correlate with gender, age, emotional states and reactions to stressors (40, 77, 78).

(2) Characters 3-grams are a sequence of three consecutive characters within a string of characters (79, 80). Each

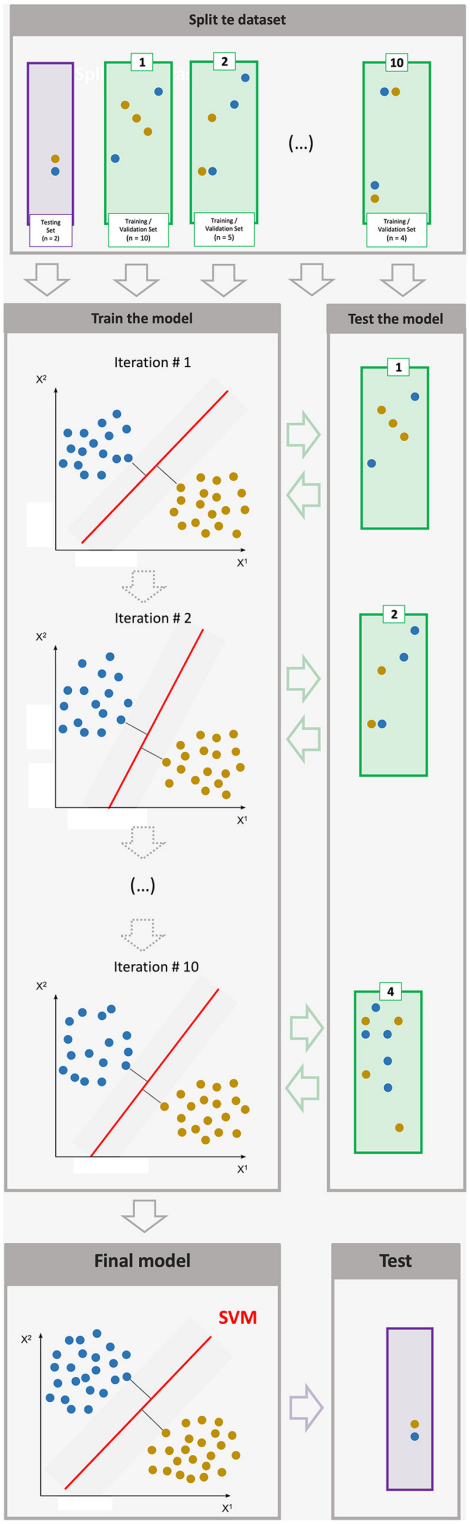

FIGURE 1
Each of the sub-corpora is divided into k samples (here 10) of similar length, so as to iteratively train the algorithm on all samples minus one, this last sample allowing each iteration to evaluate the result obtained with the k-1 others (here 9). During training, the machine statistically determines a separation boundary (a "hyperplane") between several classes (here the ADHD and control groups). In order to check whether the model produced is viable, we use a test set, which was not seen during training.

character 3-grams represents a sliding window of three characters moving through the adolescent's narrative. For example, if the latter said "During the night (. . .)," the character 3-grams will be `dur`, `uri`, `rin`, `ing`, `ng_` and so on. Character 3-grams capture subword level patterns, providing a granular view of textual features and capturing subtle stylistic nuances despite the short size of a given text (76, 81).

Hyperparameters

Because the length of SDMs varies a lot and we have more SDMs in the control group than in the ADHD groups, we test different resampling methods [undersampling the minority class, oversampling the majority class; cf. (82)] and the use of class weights. Rather than using the entire corpus, it is divided into samples of n words, and several sizes are tested (1'000, 1'250, 1'500 and 2'000). All tests are conducted with a linear kernel and a 10-fold validation, on data normalized applying Euclidean vector length normalization (L2 normalization) using z-scores for variables (33).

Evaluation metrics

Key metrics used to evaluate the performance of an SVM model are:

Precision (also known as Positive Predictive Value) is calculated as the proportion of positive predictions that were actually correct (83):

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5}$$

Recall (also known as Sensitivity or True Positive Rate) is calculated as the proportion of true positive predictions among all actual positive instances (83):

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

The F1 score is the most common measure used on imbalanced classification tasks (84). It provides a single metric balancing both precision and recall. It is calculated using the following formula:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Accuracy is computed as the proportion of correct predictions (true positives + true negatives) made by the model out of all predictions (true positives + true negatives + false positives + false negatives) (83):

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{8}$$

After running the two models (one with FW and the other with characters 3-grams), the results were examined by investigating the lexical information of the corpus and unfolding the SVM to identify its key features (i.e., the text features used by the model to make the classification).

Results

As outlined in the Methods section, we computed four linguistic metrics. The corresponding results are presented below and summarized in Table 2.

Wordcount

As defined in the *Methods*, the total wordcount was computed for each sample (see Equation 1). A Wilcoxon rank-sum test, comparing the wordcount between the ADHD group and the control group, indicated a significant difference (p-value < 0.005), with the ADHD group generating significantly shorter texts than the control group.

Lexical diversity

A Wilcoxon rank-sum test was performed to compare the distributions of Average Lexical Diversity (ALD) between the ADHD group and the control group. The results of the test indicated that the observed difference in the mean between the two groups was statistically significant (p-value < 0.0001). This indicated that the ALD values for the ADHD group were significantly lower than those of the control group.

Lexical density

A Wilcoxon rank-sum test was performed to compare the distributions of the Moving-Average Type-Token Ratio (MATTR) between the ADHD group and the control group. The results of the test indicated that the observed difference in the mean between the two groups was statistically significant (p < 0.0001). This result indicated that the MATTR values for the ADHD group were significantly lower than those of the control group.

Cohesion

A Wilcoxon rank-sum test was performed to assess the statistical significance of the difference in Cohesion Score (CS) between the ADHD and control groups. The test results demonstrated that the difference in mean CS between the two groups was highly statistically significant (p-value < 0.0001), with the ADHD group exhibiting lower CS compared to the control group.

TABLE 2 Mean and standard deviation (SD) of linguistic measures for ADHD and control groups.

Measure	Group	Mean	SD
Wordcount	ADHD	71.667	32.493
	Control	104.069	41.279
Lexical diversity (ALD)	ADHD	0.727	0.074
	Control	0.746	0.051
Lexical density (MATTR)	ADHD	0.805	0.023
	Control	0.811	0.023
Cohesion score (CS)	ADHD	66.7	29.2
	Control	97.0	30.4

Support vector machine (SVM)

The performance of the SVM model was evaluated using four key metrics (see Methods section): Precision, Recall, F1 Score, and Accuracy. The computational procedures for these metrics are detailed in Equations 5–8, respectively, and summarized in Table 3. Based on these evaluation criteria, the best results were achieved with 1,500- word samples, class weighting, and Tomek Links for Function Words (FW), and with 1,750-word samples, class weighting, and downsampling for 3-grams. Accuracy was lower with FW (85%) than with 3-grams (100%), the latter emerging as a particularly promising indicator for research in psycholinguistics. However, the results remained surprisingly good in both cases: the SVM effectively recognized the ADHD and the control groups, highlighting a distinct linguistic signal in SDMs of adolescents with ADHD. The recall for the ADHD group warranted particular attention. In a screening context, maximizing the true positive rate is essential to avoid missing individuals with ADHD. In fact, in this scenario, false positives are acceptable, as they can be excluded through further clinical assessment, whereas undetected cases are problematice, as they may go untreated. From this perspective, the 75% recall for FW was suboptimal, as it meant that approximately one-quarter of ADHD cases were missed.

Main markers of both groups

For the ADHD group, the main FW markers (see Figure 2) included the neutral pronoun ("on") combined with third-person auxiliaries and an abundance of words with syntactic function ("donc," "et," "avec"). On the other hand, the control group was found to be strongly marked by the first-person pronoun ("je," "me/m"), the adverb of quantity "plus," and the indefinite article "des."

As for 3-grams, the main markers for both groups are shown in Figure 3.

Discussion

This study targeted three main objectives. First, to use NLP to characterize the linguistic features of SDMs (i.e., wordcount, lexical diversity, lexical density and cohesion) in adolescents with ADHD compared to a control group. Second, to quantify the predictive power of stylometry for classifying ADHD vs. control individuals. And third, to qualitatively analyze key linguistic

TABLE 3 SVM results for function words (FW) and character 3-grams.

Class	Precision	Recall	F1-score	Support
Function words (FW)				
Control	0.89	0.89	0.89	9
ADHD	0.75	0.75	0.75	4
Accuracy			0.85	13
Character 3-grams				
Control	1.00	1.00	1.00	7
ADHD	1.00	1.00	1.00	3
Accuracy			1.00	10

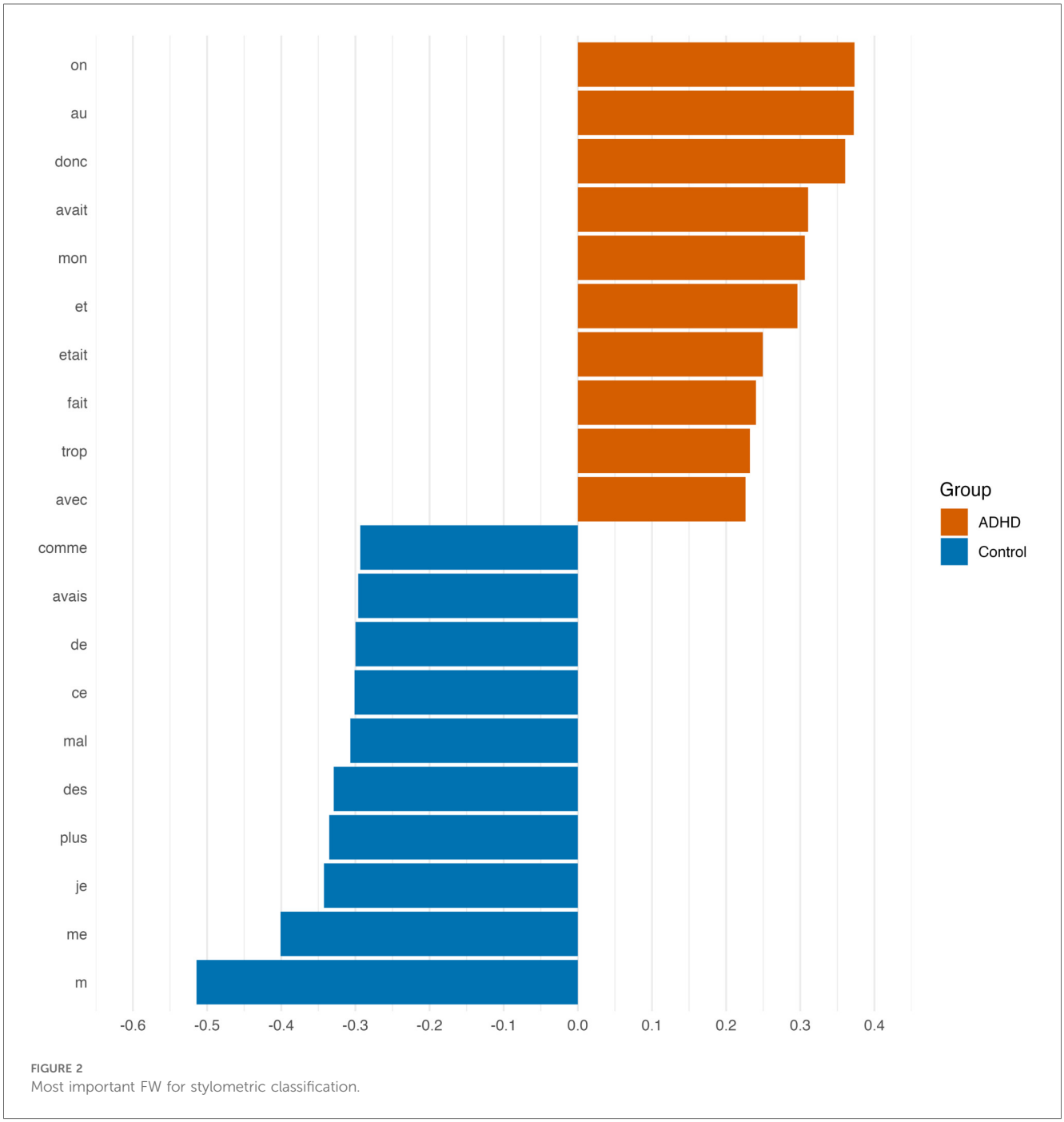

FIGURE 2
Most important FW for stylometric classification.

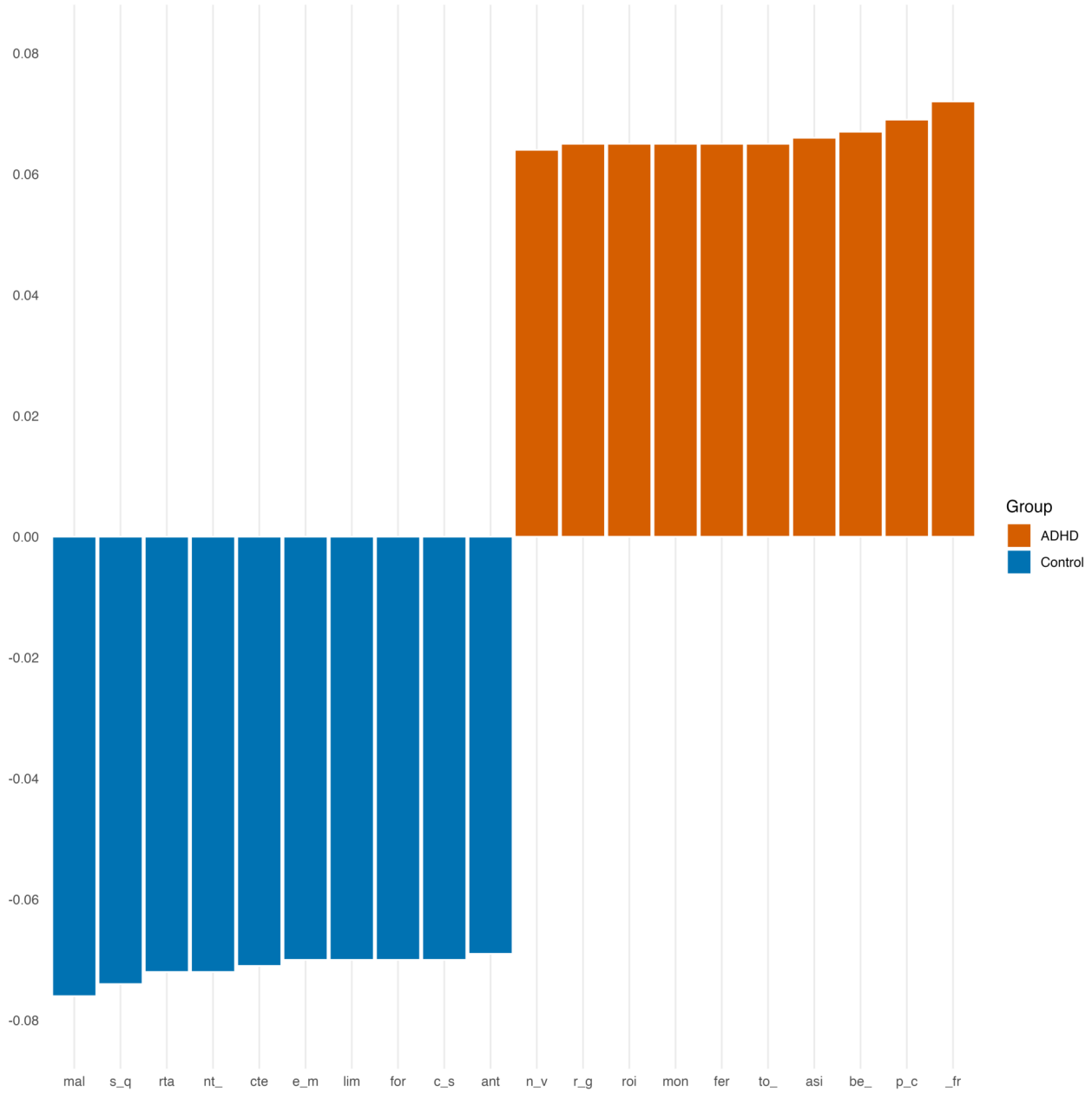

FIGURE 3
Most important 3-grams for stylometric classification.

markers, to further the scientific basis of automated text-based screening for ADHD.

For Wordcount, the ADHD group produced significantly shorter text narratives compared to the control group. This result could be explained by difficulties observed in individuals with ADHD related to their pragmatic language skills (85), and could be related with their difficulties in communicating effectively in social settings (86). However, further analysis are required to understand the relationship between wordcount and pragmatic language skills.

We also found significant differences in Lexical Diversity and Lexical Density where, in both cases, the ADHD group shows lower mean scores. Lower lexical diversity can affect communication skills and have a negative impact in social interactions and academic performance, both of which are associated with ADHD (87).

Lower lexical density in individuals with ADHD may be indicative of their cognitive and attentional profile, often characterized by difficulties in sustaining attention and deficits in executive function, which can lead to challenges in encoding, retrieving, and organizing information. These cognitive limitations can adversely affect communication by reducing language production and diversity, potentially impairing overall communication abilities and limiting social interactions and

academic performance. Although these findings suggest that lexical density is a useful metric for evaluating communication skills in populations with developmental or cognitive challenges, further analysis is required to fully understand the practical implications and underlying causes of these differences between groups.

These findings in both lexical diversity and lexical density underscore the importance of intervention programs aimed at improving language skills and self-knowledge in adolescents with ADHD. Indeed, targeted interventions in these two topics may significantly boost communication skills and foster better self-regulation, which in turn could positively impact social interactions and academic performance (86, 88, 89). However, an important point to underline here is that the focus of attention should not be on vocabulary per se, but in finding more words about the topic of the task of the SDM, i.e., personal and significant events related with narrative identity. This involves exploring which thoughts and feelings had emerged at that moment and could be done in a simple way by modifying the task of the SDMs with a semi-structured questionnaire oriented to explore and refine the contents of individual narratives.

Participants in the ADHD group also exhibit significantly lower scores in cohesion measures, indicating less structured language, which is also associated with difficulties in executive functioning (90) and working memory impairments (91). Indeed, the latter may hinder the ability to retain and manipulate information over short periods, directly impacting the coherence of verbal and written communication. Moreover, sustained attention deficits exacerbate this issue, as individuals with ADHD may lose track of the narrative, resulting in fragmented and less cohesive discourse. These findings highlight the need for targeted interventions addressing these cognitive deficits. Indeed, enhancing executive functioning through cognitive training and therapeutic strategies could improve language cohesion and overall communication skills in individuals with ADHD.

Regarding the capacity of stylometry to detect the SDMs of adolescents with ADHD at the group level, we found that the SVM classifier successfully differentiated between the autobiographical texts of adolescents with ADHD and those of the control group. The results can be considered robust, given that our lowest accuracy score was 85% using function words, while the highest accuracy score reached 100% with characters 3-grams. Of note, and for diagnostic purposes, a model with an accuracy exceeding 80% is considered to have very good performance (92). Consequently, these findings indicate that stylometric analysis is a viable method for detecting group-level differences in narrative writing between adolescents with ADHD and their non-ADHD peers. Indeed, our study confirms that adolescents with an ADHD diagnosis do have a distinct narrative style, as evidenced by significant differences in the language patterns of their autobiographical narratives compared to the control group.

Last but not least, distinct markers for each group were identified, which hold significant psychological meaning. With respect to personal pronouns used in both groups, one of the main markers for the ADHD group was the indefinite neuter pronoun "on." This result is particularly intriguing given that the SDMs task requires individuals to recall personally significant memories, which are naturally expected to evoke first-person ("I" or "me") experiences. In French, the personal pronoun "on" refers to the narrator and one or more other persons without having explicitly or necessarily mentioned them (93). Translating "on" into English typically requires choosing between "one" or "we," depending on the context [cf. (94)]. Although "one" is commonly used (Ibid.), the adolescents in this study frequently describe situations involving their peer groups in their SDMs. Consequently, "we" better captures the collective identity and shared experiences implied, making it the more appropriate translation in this case. In the control group the first singular pronoun "je" ("I") was the main marker. According to Boulard and Gauthier (95), the emergence of this personal pronoun in children's narratives enables first-person storytelling and is essential for developing self-awareness and differentiation through language. Once children move beyond the individuation stage during adolescence, individuals go through significant psychological changes as they form their identity and establish a sense of self that is distinct from their parents and peers (96). At that developmental stage, the use of personal pronouns provides information about the narrator's focus and self-perception in relation to their social environment [cf. (97)]. According to Sutin and Robins (98), in the first-person perspective, individuals see the event through their own eyes, while a reduced use of the first person may serve a distancing function helping to reduce emotional reliving and to distance the current self from the self in the recalled event. Consequently, and according to this model, the marker "on" could reveal difficulties in experiencing affects and/or a need to distance oneself from the latter reflecting difficulties in emotion regulation. Several studies have found evidence of the impact of emotional dysregulation in youth with ADHD. In fact, these children and adolescent are six times more likely to have emotional regulation difficulties than their non-affected siblings (99). These findings underscore the importance of enhancing self-focus in interventions aimed at improving self-awareness among adolescents with ADHD, thereby enhancing their ability to embody the present, which means to be focused and attentive in the "here and know" from a mind body connection perspective as is the case in mindfulness based therapy.

In conclusion, the results of this study supports the hypothesis that ADHD impacts narrative construction in measurable ways, potentially reflecting broader cognitive and linguistic differences. The use of computational linguistics, including NLP and stylometry, demonstrates the utility of machine learning approaches in psychological research, particularly for identifying subtle linguistic markers associated with mental health disorders. By identifying patterns in language use, clinicians can gain insights into the cognitive and linguistic profiles of individuals with ADHD, aiding in diagnosis and the tailoring of interventions helping to provide more targeted and effective therapeutic strategies. In this sense, these findings have implications for developing new diagnostic tools and enhancing our understanding of the cognitive and linguistic profiles of individuals with ADHD.

While the use of stylometric methods for ADHD screening and diagnosis shows promise, several methodological improvements are

needed to enhance their precision and automation. Despite these limitations, our findings highlight the potential of stylometry in both narrative identity research and in ADHD diagnostics. Stylometric analysis may offer a valuable tool for detecting subtle linguistic markers associated with mental health conditions. As these methods continue to evolve, their ability to classify narratives at both the group and individual level opens new possibilities for innovative forms of psychological assessment. However, further research is required to identify the most informative features specific to ADHD and to validate these findings in larger, more diverse samples that contain other neurodevelopmental conditions such as specific learning disorder and autism.

This study has several limitations that should be addressed in future research. Firstly, the sample size is small, which may affect the generalizability of the findings. Secondly, while 16% of participants in the ADHD group had a diagnosed learning disorder, the exclusion criteria required the absence of any clinical diagnoses for the control participants. Future studies with larger groups should assess the impact of learning disorders on the present results. Third, the methodology is limited to SDMs, while no other types of narratives have been tested. This restricts the scope of our analysis and may overlook other relevant linguistic features. Additionally, the study did not include analyses examining the influence of ADHD severity on classification performance, nor whether the identified linguistic markers were associated with general verbal abilities, ADHD subtypes, or social communication skills. This limitation is primarily due to the sample size, which may have constrained our ability to capture meaningful variation within the ADHD population. As a result, the current findings cannot definitively isolate linguistic features that are specific to ADHD, as the dataset may reflect the influence of comorbidities commonly associated with the condition.

While we cannot quantify the individual contribution of these comorbid factors to the classification outcomes, it is important to emphasize that ADHD is, by definition, a heterogeneous disorder that frequently co-occurs with other neurodevelopmental and mental health conditions. From a screening perspective, this heterogeneity does not diminish the value of our approach; rather, it underscores the potential of stylometric analysis as a broad initial filter. The ability to detect linguistic patterns characteristic of individuals within the ADHD spectrum—regardless of comorbid presentations—may serve as a valuable first step in identifying at-risk individuals who warrant further clinical assessment.

Addressing these issues is crucial to enhance the validity and utility of computational linguistic analysis in diagnosing and understanding ADHD.

Future research involving larger and more diverse samples—including comparative analyses with other clinical populations such as individuals with autism spectrum disorder or learning disabilities—will be essential to determine the specificity and diagnostic utility of these linguistic markers. Such work will be crucial for advancing this tool beyond initial screening, refining it toward diagnostic applications that align with both categorical and dimensional perspectives on ADHD.

To increase the accuracy and fine-grained detection of ADHD, several avenues for future research should be explored. First, investigating potential language marker differences based on gender is imperative. Additionally, refining detection precision to identify distinct modalities of ADHD, such as hyperactive modalities, is crucial. This entails incorporating a larger sample size, in different languages (specifically English), including more hyperactive participants, to capture broader and more representative characteristics accompanying ADHD.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

JB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. EP: Writing – review & editing. JLS: Writing – review & editing. HR: Writing – review & editing. SG: Conceptualization, Formal analysis, Methodology, Writing – review & editing. FC: Writing – review & editing. MD: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Juan Barrios was funded by the Chilean National Agency for Research and Development (ANID) through the National Doctoral Grant, folio 72190649. The PI (Prof. Martin Debbané) was funded by the Swiss National Science Foundation (Grant number 100014_179033), as well as the Marina Picasso Prize from AEMD Foundation 2018.

Acknowledgments

The authors would like to thank Pavel Blagov for the advice and reflections he generously shared with them after reading the manuscript of this article. His contributions helped to clarify some important aspects of this paper and were a stimulus in the development of our work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Ayano G, Demelash S, Gizachew Y, Tsegay L, Alati R. The global prevalence of attention deficit hyperactivity disorder in children and adolescents: an umbrella review of meta-analyses. *J Affect Disord*. (2023) 339:860–6. doi: 10.1016/j.jad.2023.07.071

2. Salari N, Ghasemi H, Abdoli N, Rahmani A, Shiri MH, Hashemian AH, et al. The global prevalence of ADHD in children and adolescents: a systematic review and meta-analysis. *Ital J Pediatr*. (2023) 49(1):48. doi: 10.1186/s13052-023-01456-1

3. APA. *Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*. Washington D.C.: American Psychiatric Association (2013).

4. Barkley RA, Murphy KR. *Attention-Deficit Hyperactivity Disorder: A Clinical Workbook*. New York: Guilford Press (2006).

5. Çiray RO, Özyurt G, Turan S, Karagöz E, Ermiş Ç, Öztürk Y, et al. The association between pragmatic language impairment, social cognition and emotion regulation skills in adolescents with ADHD. *Nord J Psychiatry*. (2022) 76(2):89–95. doi: 10.1080/08039488.2021.1938211

6. Faraone SV, Asherson P, Banaschewski T, Biederman J, Buitelaar JK, Ramos-Quiroga JA, et al. Attention-deficit/hyperactivity disorder. *Nat Rev. Dis Primers*. (2015) 1:15020. doi: 10.1038/nrdp.2015.20

7. Cuffe SP, Visser SN, Holbrook JR, Danielson ML, Geryk LL, Wolraich ML, et al. Attention-deficit/hyperactivity disorder and psychiatric comorbidity: functional outcomes in a school-based sample of children. *J Atten Disord*. (2020) 24(9):1345–54. doi: 10.1177/1087054715613437

8. Sayal K, Prasad V, Daley D, Ford T, Coghill D. ADHD in children and young people: prevalence, care pathways, and service provision. *Lancet Psychiatry*. (2018) 5(2):175–86. doi: 10.1016/S2215-0366(17)30167-0

9. Wolraich ML, Hagan Jr JF, Allan C, Chan E, Davison D, Earls M, et al. Clinical practice guideline for the diagnosis, evaluation, andtreatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics*. (2019) 144(4):e20192528. doi: 10.1542/peds.2019-2528

10. NICE. *Attention Deficit Hyperactivity Disorder: Diagnosis and Management*. London: NICE Guideline (2018). p. NG87.

11. Mikolas P, Vahid A, Bernardoni F, Süß M, Martini J, Beste C, et al. Training a machine learning classifier to identify adhd based on real-world clinical data from medical records. *Sci Rep*. (2022) 12(1):12934. doi: 10.1038/s41598-022-17126-x

12. Vergunst F, Tremblay RE, Galera C, Nagin D, Vitaro F, Boivin M, et al. Multi-rater developmental trajectories of hyperactivity-impulsivity and inattention symptoms from 1.5 to 17 years: a population-based birth cohort study. *Eur Child Adolesc Psychiatry*. (2019) 28(7):973–83. doi: 10.1007/s00787-018-1258-1

13. Emser TS, Johnston BA, Steele JD, Kooij S, Thorell L, Christiansen H. Assessing ADHD symptoms in children and adults: evaluating the role of objective measures. *Behav Brain Funct*. (2018) 14(1):11. doi: 10.1186/s12993-018-0143-x

14. Peterson BS, Trampush J, Brown M, Maglione M, Bolshakova M, Rozelle M, et al. Tools for the diagnosis of adhd in children and adolescents: a systematic review. *Pediatrics*. (2024) 153(4):e2024065854. doi: 10.1542/peds.2024-065854

15. DeSouza DD, Robin J, Gumus M, Yeung A. Natural language processing as an emerging tool to detect late-life depression. *Front Psychiatry*. (2021) 12:719125. doi: 10.3389/fpsyt.2021.719125

16. Crema C, Attardi G, Sartiano D, Redolfi A. Natural language processing in clinical neuroscience and psychiatry: a review. *Front Psychiatry*. (2022) 13:946387. doi: 10.3389/fpsyt.2022.946387

17. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, et al. Facebook language predicts depression in medical records. *Proc Natl Acad Sci USA*. (2018) 115(44):11203–8. doi: 10.1073/pnas.1802331115

18. Manabe M, Liew K, Yada S, Wakamiya S, Aramaki E. Estimation of psychological distress in japanese youth through narrative writing: text-based stylometric and sentiment analyses. *JMIR Formative Res*. (2021) 5(8):e29500. doi: 10.2196/29500

19. Pérez A, Parapar J, Barreiro A. Automatic depression score estimation with word embedding models. *Artif Intell Med*. (2022) 132:102380. doi: 10.1016/j.artmed.2022.102380

20. Tchounwou P. Environmental research and public health. *Int J Environ Res Public Health*. (2004) 1(1):1–2. doi: 10.3390/ijerph2004010001

21. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. *Nat Lang Eng*. (2017) 23(5):649–85. doi: 10.1017/S1351324916000383

22. Figueroa-Barra A, Del Aguila D, Cerda M, Gaspar PA, Terissi LD, Durán M, et al. Automatic language analysis identifies and predicts schizophrenia in first-episode of psychosis. *Schizophrenia*. (2022) 8(1):53. doi: 10.1038/s41537-022-00259-3

23. Milintsevich K, Sirts K, Dias G. Towards automatic text-based estimation of depression through symptom prediction. *Brain Inform*. (2023) 10(1):4. doi: 10.1186/s40708-023-00185-9

24. Leis A, Ronzano F, Sanz F. Detecting signs of depression in tweets in spanish: behavioral and linguistic analysis. *J Med Internet Res*. (2019) 21(6):e14199. doi: 10.2196/14199

25. Miao YL, Cheng WF, Ji YC, Zhang S, Kong YL. Aspect-based sentiment analysis in chinese based on mobile reviews for BiLSTM-CRF. *J Intell Fuzzy Syst*. (2021) 40:1–11. doi: 10.3233/JIFS-192078

26. Martin L, Muller B, Ortiz Suárez PJ, Dupont Y, Romary L, Sagot B. CamemBERT: a tasty French language model. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics (2020).

27. Kim K, Lee CH. Distinctive linguistic styles in children with ADHD. *Psychol Rep*. (2009) 105(2):365–71. doi: 10.2466/PR0.105.2.365-371

28. Kim K, Lee S, Lee C. College students with ADHD traits and their language styles. *J Atten Disord*. (2015) 19(8):687–93. doi: 10.1177/1087054713484512

29. Delavarian M, Towhidkhah F, Dibajnia P, Gharibzadeh S. Designing a decision support system for distinguishing ADHD from similar children behavioral disorders. *J Med Syst*. (2012) 36(3):1335–43. doi: 10.1007/s10916-010-9594-9

30. Duda M, Haber N, Daniels J, Wall DP. Crowdsourced validation of a machine-learning classification system for autism and ADHD. *Transl Psychiatry*. (2017) 7(5):e1133. doi: 10.1038/tp.2017.86

31. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. *Transl Psychiatry*. (2016) 6(2):e732. doi: 10.1038/tp.2015.221

32. Tachmazidis I, Chen T, Adamou M, Antoniou G. A hybrid AI approach for supporting clinical diagnosis of attention deficit hyperactivity disorder (ADHD) in adults. *Health Inf Sci Syst*. (2020) 9(1):1. doi: 10.1007/s13755-020-00123-7

33. Evert S, Proisl T, Jannidis F, Reger I, Pielström S, Schöch C, et al. Understanding and explaining Delta measures for authorship attribution. *Digit Sch Humanit*. (2017) 32(suppl_2):ii4–16. doi: 10.1093/llc/fqx023

34. Juola P. Verifying authorship for forensic purposes: a computational protocol and its validation. *Forensic Sci Int*. (2021) 325:110824. doi: 10.1016/j.forsciint.2021.110824

35. Cafiero F, Camps J-B. Who could be behind qanon? Authorship attribution with supervised machine-learning. *Digit Sch Humanit*. (2023) 38(4):1418–30. doi: 10.48550/arXiv.2303.02078

36. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language. use: our words, our selves. *Annu Rev Psychol*. (2003) 54(1):547–77. doi: 10.1146/annurev.psych.54.101601.145041

37. Singer JA, Blagov PS. *Classification System & Scoring Manual for Self-Defining Memories*. New London: Connecticut College (2022).

38. Singer JA, Salovey P. *The Remembered Self*. New York: Free Press (1994).

39. Blagov PS, Singer JA, Oost KM, Goodman JA. Self-defining memories-narrative features in relation to adaptive and maladaptive personality traits (replication and extension of Blagov & Singer, 2004). *J Pers*. (2022) 90(3):457–75. doi: 10.1111/jopy.12677

40. Pennebaker JW. *The Secret Life of Pronouns: What Our Words Say About Us*. New York: Bloomsbury Press (2011).

41. Creswell JW, Poth CN. *Qualitative Inquiry and Research Design*. 4th ed. Thousand Oaks: SAGE Publications (2017).

42. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. (2006) 3(2):77–101. doi: 10.1191/1478088706qp063oa

43. Young S. *ADHD Child Evaluation (ACE), A Diagnostic Interview of ADHD in Children*. London: Psychology Services Limited (2015).

44. Wechsler D. *Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV): Technical and Interpretive Manual*. San Antonio: The Psychological Corporation (2003).

45. Ho DE, Imai K, King G, Stuart EA. Matchit: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. (2011) 42(8):1–28. doi: 10.18637/jss.v042.i08

46. Crossley SA, Kyle K, Dascalu M. The tool for the automatic analysis of cohesion 2.0: integrating semantic similarity and text overlap. *Behav Res Methods*. (2019) 51(1):14–27. doi: 10.3758/s13428-018-1142-4

47. Graesser AC, McNamara DS, Kulikowich JM. Coh-metrix: providing multilevel analyses of text characteristics. *Educ Res*. (2011) 40(5):223–34. doi: 10.3102/0013189X11413260

48. Yoder PJ. Predicting lexical density growth rate in young children with autism spectrum disorders. *Am J Speech Lang Pathol*. (2006) 15(4):378–88. doi: 10.1044/1058-0360(2006/035)

49. Jurafsky D, Martin JH. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Pearson Prentice Hall (2009).

50. Straka M, Straková J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Hajič J, Zeman D, editors. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics (2017). p. 88–99.

51. Malvern DD, Richards BJ, Chipere N, Duran P. *Lexical Diversity and Language Development*. Gordonsville: Palgrave Macmillan (2004).

52. Watkins RV, Kelly DJ, Harbers HM, Hollis W. Measuring children's lexical diversity: differentiating typical and impaired language learners. *J Speech Lang Hear Res*. (1995) 38(6):1349–55. doi: 10.1044/jshr.3806.1349

53. Redmond SM. Conversational profiles of children with ADHD, SLI and typical development. *Clin Linguist Phon*. (2004) 18(2):107–25. doi: 10.1080/02699200310001611612

54. Covington MA, McFall JD. Cutting the gordian knot: the moving-average type–token ratio (MATTR). *J Quant Linguist*. (2010) 17(2):94–100. doi: 10.1080/09296171003643098

55. Johansson V. Lexical diversity and lexical density in speech and writing. In: *Working Papers*. Lund University, Department of Linguistics and Phonetics (2008). p. 53.

56. Halliday MAK. *Spoken and Written Language*. Oxford: Oxford University Press (1990).

57. Pennebaker JW, King LA. Linguistic styles: language use as an individual difference. *J Pers Soc Psychol*. (1999) 77(6):1296–312. doi: 10.1037/0022-3514.77.6.1296

58. Hyland K. Disciplinary interactions: metadiscourse in L2 postgraduate writing. *J Second Lang. Writ*. (2004) 13(2):133–51. doi: 10.1016/j.jslw.2004.02.001

59. Wijffels J. udpipe: tokenization, parts of speech tagging, lemmatization and dependency parsing with the "UDPipe" "NLP" toolkit. R package version 0.8.11 (2023).

60. Halliday MAK, Matthiessen C, Halliday M. *Halliday's Introduction to Functional Grammar*. 4th ed. London, England: Routledge (2013).

61. Brewin CR. Memory processes in post-traumatic stress disorder. *Int Rev Psychiatry*. (2001) 13:159–63. doi: 10.1080/09540260120074019

62. Tuval-Mashiach R, Freedman S, Bargai N, Boker R, Hadar H, Shalev AY. Coping with trauma: narrative and cognitive perspectives. *Psychiatry Interpers Biol Processes*. (2004) 67(3):280–93. doi: 10.1521/psyc.67.3.280.48977

63. Graesser AC, McNamara DS, Louwerse MM, Cai Z. Coh-metrix: analysis of text on cohesion and language. *Behav Res Methods Instrum Comput*. (2004) 36(2):193–202. doi: 10.3758/BF03195564

64. HaCohen-Kerner Y. Survey on profiling age and gender of text authors. *Expert Syst Appl*. (2022) 199:117140. doi: 10.1016/j.eswa.2022.117140

65. Cafiero F, Camps J-B. Why Molière most likely did write his plays. *Sci Adv*. (2019) 5(11):eaax5489. doi: 10.1126/sciadv.aax5489

66. Ikae C. Unine at pan-clef 2022: profiling irony and stereotype spreaders on twitter. In: *CLEF*. (2022). p. 1613–0073.

67. Modaresi P, Liebeck M, Conrad S. Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. In: *Components of an Automatic Single Document Summarization System in the News Domain*. Aachen: CEUR-WS.org (2017). p. 100–7.

68. Ward JK, Cafiero F, Fretigny R, Colgrove J, Seror V. France's citizen consultation on vaccination and the challenges of participatory democracy in health. *Social Sci Med*. (2019) 220:73–80. doi: 10.1016/j.socscimed.2018.10.032

69. Cafiero F, Camps J-B. Psyché' as a rosetta stone? Assessing collaborative authorship in the French 17th century theatre. In: *Proceedings of the Conference on Computational Humanities Research 2021*. CEUR Workshop Proceedings (2021). p. 377–91.

70. Diederich J, Kindermann J, Leopold E, Paass G. Authorship attribution with support vector machines. *Appl Intell*. (2003) 19(1/2):109–23. doi: 10.1023/A:1023824908771

71. Eder M. Rolling stylometry. *Digit Sch Humanit*. (2015) 31(3):457–69. doi: 10.1093/llc/fqv010

72. Fung G. The disputed federalist papers: SVM feature selection via concave minimization. In: *Proceedings of the 2003 Conference on Diversity in Computing*. TAPIA '03. New York: Association for Computing Machinery (2003). p. 42–6.

73. Camps J-B. *SUPERvised STYLometry*. Paris: SuperStyl (2021).

74. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. (2011) 12:2825–30. doi: 10.48550/arXiv.1201.0490

75. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification (2009).

76. Kestemont M. Function words in authorship attribution. From black magic to theory? In: Feldman A, Kazantseva A, Szpakowicz S, editors. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden: Association for Computational Linguistics (2014). p. 59–66.

77. Pennebaker JW, Lay TC. Language use and personality during crises: analyses of mayor rudolph giuliani's press conferences. *J Res Pers*. (2002) 36(3):271–82. doi: 10.1006/jrpe.2002.2349

78. Pennebaker JW, Stone LD. Words of wisdom: language use over the life span. *J Pers Soc Psychol*. (2003) 85(2):291–301. doi: 10.1037/0022-3514.85.2.291

79. Frantzeskou G, Stamatatos E, Gritzalis S, Chaski CE, Howald BS. Identifying authorship by byte-level n-grams: the source code author profile (SCAP) method. *Int J Digit Evid*. (2007) 6:1–8. https://www.researchgate.net/publication/220542545_Identifying_Authorship_by_Byte-Level_N-Grams_The_Source_Code_Author_Profile_SCAP_Method

80. Kjell B, Woods WA, Frieder O. Discrimination of authorship using visualization. *Inf Process Manage*. (1994) 30(1):141–50. doi: 10.1016/0306-4573(94)90029-9

81. Sapkota U, Bethard S, Montes M, Solorio T. Not all character n-grams are created equal: a study in authorship attribution. In: Mihalcea R, Chai J, Sarkar A, editors. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver: Association for Computational Linguistics (2015). p. 93–102.

82. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Int Res*. (2002) 16(1):321–57. doi: 10.1613/jair.953

83. Powers DMW. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation (2020).

84. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. (2009) 21(9):1263–84. doi: 10.1109/TKDE.2008.239

85. Kessler PB, Ikuta T. Pragmatic deficits in attention deficit/hyperactivity disorder: systematic review and meta-analysis. *J Atten Disord*. (2023) 27(8):812–21. doi: 10.1177/10870547231161534

86. Barkley RA, editor. *Attention-Deficit Hyperactivity Disorder*. 4th ed. New York: Guilford Publications (2014).

87. Willcutt EG. The prevalence of dsm-iv attention-deficit/hyperactivity disorder: a meta-analytic review. *Neurotherapeutics*. (2012) 9(3):490–9. doi: 10.1007/s13311-012-0135-8

88. Docking K, Munro N, Cordier R, Ellis P. Examining the language skills of children with adhd following a play-based intervention. *Child Lang Teach Ther*. (2013) 29(3):291–304. doi: 10.1177/0265659012469042

89. Westby CE, Cutler SK. Language and adhd: understanding the bases and treatment of self-regulatory deficits. *Top Lang Disord*. (1994) 14(4):58–76. doi: 10.1097/00011363-199408000-00006

90. Veraksa A, Bukhalenkova D, Kartushina N, Oshchepkova E. The relationship between executive functions and language production in 5–6-year-old children: insights from working memory and storytelling. *Behav Sci*. (2020) 10(2):52. doi: 10.3390/bs10020052

91. Martinussen R, Hayden J, Hogg-Johnson S, Rosemary T. A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*. (2005) 44(4):377–84. doi: 10.1097/01.chi.0000153228.72591.73

92. Šimundić A-M. Measures of diagnostic accuracy: basic definitions. *EJIFCC*. (2009) 19(4):203–11.

93. Éditions Larousse. Larousse.fr: encyclopédie et dictionnaires gratuits en ligne. Larousse.fr. (2023) (accessed December 26, 2023).

94. Bradley ED. Subject pronoun expression in spanish: The effects of the speech environment and of morphological regularity. *J Pragmat*. (2018) 126:67–84.

95. Boulard A, Gauthier J-M. Quand l'enfant dit "je". *Enfance*. (2012) 2(2):233–46. doi: 10.3917/enf1.122.0233

96. Erikson EH. *Identity*. New York: WW Norton (1994).

97. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol*. (2009) 29(1):24–54. doi: 10.1177/0261927X09351676

98. Sutin AR, Robins RW. When the "I" looks at the "Me": autobiographical memory, visual perspective, and the self. *Conscious Cogn*. (2008) 17(4):1386–97. doi: 10.1016/j.concog.2008.09.001

99. Anastopoulos AD, Smith TF, Garrett ME, Morrissey-Kane E, Schatz NK, Sommer JL, et al. Self-regulation of emotion, functional impairment, and comorbidity among childrenwith AD/HD. *J Atten Disord*. (2011) 15(7):583–92. doi: 10.1177/1087054710370567